Confidence Set for Group Membership*

Andreas Dzemski[†] and Ryo Okui[‡]

March 6, 2018

*The authors would like to thank Christoph Breunig, Le-Yu Chen, Eric Gautier, Hidehiko Ichimura, Hiroaki Kaido, Hiroyuki Kasahara, Kengo Kato, Toru Kitagawa, Arthur Lewbel, Artem Prokhorov, Adam Rosen, Myung Hwan Seo, Katsumi Shimotsu, Liangjun Su, Wendun Wang, Wuyi Wang, Martin Weidner, Yoon-Jae Whang and seminar participants at the Centre for Panel Data Analysis Symposium at the University of York, St Gallen, HKUST, SUFE, Sydney Econometric Reading Group, Xiamen, CUHK Workshop on Econometrics, Chinese Academy of Sciences, Asian Meeting of Econometric Society 2017, STJU, Workshop on Advances in Econometrics 2017 at Hakodate, SNU, Academia Sinica, ESEM 2017, Berlin and CFE-CMStatistics 2017 for valuable comments. Sophie Li provided excellent research assistance. A part of this research was done while Okui was at Kyoto University and Vrije Universiteit Amsterdam. This work is supported by JSPS KAKENHI Grant number 15H03329 and 16K03598.

[†]Department of Economics, University of Gothenburg, P.O. Box 640, SE-405 30 Gothenburg, Sweden. Email: andreas.dzemski@economics.gu.se

[‡]NYU Shanghai, 1555 Century Avenue, Pudong, Shanghai, China, 200122; Department of Economics, University of Gothenburg, P.O. Box 640, SE-405 30 Gothenburg, Sweden. Email: okui@nyu.edu

Abstract

We develop new procedures to quantify the statistical uncertainty from sorting units in panel data into groups using data-driven clustering algorithms. In our setting, each unit belongs to one of a finite number of latent groups and its regression curve is determined by which group it belongs to. Our main contribution is a new joint confidence set for group membership. Each element of the joint confidence set is a vector of possible group assignments for all units. The vector of true group memberships is contained in the confidence set with a pre-specified probability. The confidence set inverts a test for group membership. This test exploits a characterization of the true group memberships by a system of moment inequalities. Our procedure solves a high-dimensional one-sided testing problem and tests group membership simultaneously for all units. We also propose a procedure for identifying units for which group membership is obviously determined. These units can be ignored when computing critical values. We justify the joint confidence set under $N, T \to \infty$ asymptotics where we allow T to be much smaller than N. Our arguments rely on the theory of self-normalized sums and high-dimensional central limit theorems. We contribute new theoretical results for testing problems with a large number of moment inequalities, including an anti-concentration inequality for the quasi-likelihood ratio (QLR) statistic. Monte Carlo results indicate that our confidence set has adequate coverage and is informative. We illustrate the practical relevance of our confidence set in two applications.

Keywords: Panel data, grouped heterogeneity, clustering, confidence set, machine learning, moment inequalities, joint one-sided tests, self-normalized sums, high-dimensional CLT, anti-concentration for QLR

JEL codes: C23, C33, C38

1. Introduction

Panel data models with grouped heterogeneity have emerged as useful modeling tools to learn about heterogeneous regression curves (cf. Bonhomme and Manresa 2015; Su, Shi, and Phillips 2016; Vogt and Linton 2017). The heterogeneity can reflect unobserved characteristics (Heckman and Singer 1984) or equilibrium selection (Hahn and Moon 2010). In these models, it is assumed that the population is partitioned into a finite set of "groups." All members of a group share the same regression curve. Each unit's group membership is unobserved and has to be inferred from its behavior over time.[1] The existing literature has focused on inference with respect to the group-specific regression curves. This problem has been considered in Bonhomme and Manresa (2015), Su, Shi, and Phillips (2016), Vogt and Linton (2017), and Wang, Phillips, and Su (2016).

In the present paper, we focus on the clustering problem and study inference with respect to the group memberships. In particular, we construct confidence sets for group membership. We consider joint and unit-wise confidence sets. For a panel of N units, an element of a joint confidence set is an N-dimensional vector that states a possible group membership for each unit. Our construction guarantees that the joint confidence set contains the N-vector of true group memberships with a pre-specified probability, say 90%. For a specific unit, a unit-wise confidence set is a collection of possible group memberships. Its construction ensures that it contains the unit's true group membership at least with a pre-specified probability.

Our confidence sets are the first contribution in the econometric and statistical literature to rigorously quantify the estimation error from assigning group memberships using a data-driven clustering algorithm. If a unit's unit-wise confidence set is a singleton then the unit's group membership is clear from the data. In this case, the unit's estimated group membership is the only element in the confidence set and may be considered statistically significant. If the data does not clearly identify a unit's group membership, then the unit's confidence set contains multiple possible group memberships. Providing a joint rather than a unit-wise confidence set is important if we want to control the probability of misclassification when selecting units by group, either for a policy/program intervention or further study. In one of our empirical applications, we follow Wang, Phillips, and Su (2016) and cluster states in the U.S. into two groups. The effect of a minimum wage on unemployment is positive in one group and negative in the other. When designing a new minimum wage policy, it is important to detect the units for which group membership cannot be identified with confidence. This requires joint inference on all units.

Our unit-wise confidence sets are computed by inverting a test for group membership. The test is based on the observation that the true group membership of a specific unit satisfies a system of moment inequalities. The unit's true group membership provides a best fit to the observed behavior of a unit. Each moment inequality compares the fit of two possible group assignments.[2] We exploit the specific structure of these inequalities to recenter them so that they are binding under the null

[1] The group structure can be interpreted structurally or as an approximation to some underlying finer pattern of heterogeneity, as in Bonhomme, Lamadon, and Manresa (2016).

[2] The estimator of group membership in Bonhomme and Manresa (2015) is based on this observation.

hypothesis. It follows that testing group membership is equivalent to testing a one-sided hypothesis for a vector of moments. In particular, we test the hypothesis that the vector is the zero vector versus the alternative hypothesis that it has a positive component.

We construct our joint confidence set by combining unit-wise confidence sets. To guarantee that the joint confidence set has the desired coverage probability we use a Bonferroni-type correction. For computational reasons we do not construct our joint confidence set by inverting a joint test that tests the group memberships of all units simultaneously.[3] Note that a naive inversion of a joint test requires testing G^N possible membership configurations; this is intractable even in small panels. On the other hand, the computation of our joint confidence set is based on GN tests of group membership. This computational cost scales well in large panels. Under cross-sectional independence, which is a common assumption in panel regression, the Bonferroni correction is expected to render our joint confidence set only minimally conservative if N is large.[4]

We suggest three procedures for constructing unit-wise confidence sets, corresponding to three flavors of the underlying test of group membership. We consider two test statistics, MAX and QLR, from the literature on testing moment inequalities (cf. Rosen 2008; Andrews and Soares 2010; Romano, Shaikh, and Wolf 2014) in combination with analytical critical values derived from Gaussian approximations. The MAX statistic looks at the largest element of the tested vector of moments, while the QLR statistic minimizes a quadratic form and can be derived as the quasi-likelihood ratio test statistic of our one-sided hypothesis. We suggest two different methods to compute critical values for the MAX statistic and one way to compute critical values for the QLR statistic. To improve the coverage of the joint confidence sets in short panels we suggest adjustments of the critical values that are motivated by the finite-sample behavior of the respective test statistic under Gaussianity.

The first procedure is based on the MAX test statistic and a critical value common to all units and groups. We call it the SNS procedure. This procedure works for any correlation structure between the within-unit moments but is possibly more conservative than the other procedures. SNS stands for "self-normalized sum", referring to the theoretical justification of this procedure by the theory of self-normalized sums (de la Pena, Lai, and Shao 2009). The SNS critical value is computationally advantageous because it is not unit specific and therefore has to be computed only once. Moreover, the SNS procedure can be justified under much weaker moment conditions than the other procedures that we propose. The idea for the SNS critical value is adapted from Chernozhukov, Chetverikov, and Kato (2014). However, our critical value is defined differently from theirs. Our definition admits a finite-sample justification under an additional normality assumption.

Our second procedure combines the MAX test statistic with unit-specific critical values. We call it the MAX procedure. The critical values of the MAX procedure account for the correlation of within-unit moments. This correlation is expected to be high and the correlation structure

[3] A joint test for the group memberships of all units can be based on a system of moment inequalities that describes the group memberships of all units simultaneously.

[4] As $N \to \infty$, the Bonferroni correction inflates the theoretical coverage of a confidence set at nominal level $1 - \alpha$ by less than α^2.

may be different for different units. Theoretically, the MAX procedure is equivalent to multiplier bootstrap with Gaussian multipliers. However, to compute the Bonferroni correction in our setting, we would have to evaluate (unit-wise) bootstrap distributions at very large quantiles. This renders the computational cost of the usual Monte Carlo approximation of the bootstrap distribution prohibitive. By contrast, our proposed analytical critical values compute rapidly. The short-panel adjustment for the MAX procedure is based on the multivariate t-distribution.

Lastly, we combine the QLR test statistic with unit-specific analytical critical values. We call this the QLR procedure. The unit-specific critical values are based on a well-known approximation of the distribution of the QLR statistic under the null hypothesis by a mixture of χ^2 distributions (Kudo 1963; Wolak 1989). The short-panel adjustment for the QLR procedure is based on a mixture of F-distributions.

We also study a variation of our procedure that can increase the power of the joint confidence set. We call this approach *unit selection*. In the literature on moment inequalities, moment selection is a popular approach for increasing the power of a test. It detects inequalities that are "obviously" slack and can be disregarded when computing critical values.[5] In our setting, we recenter all inequalities to be binding under the null hypothesis and moment selection is not applicable. Nonetheless, we can still exploit the intuition that a part of the testing problem that is "obvious" should not inflate critical values. To motivate our approach, suppose the panel is split into units with low noise for which the group assignment is "obvious" and units with noisier measurements. We suggest an algorithm that learns the identities of the units in the first group and ignores these units when computing the Bonferroni adjustment for the unit-wise confidence sets for units in the second group. Our algorithm combines moment selection with iterated deletion of hypotheses. Unit selection is expected to be effective in settings with substantial heteroscedasticity.

We justify our procedures under a double asymptotic framework that sends both the number of units N and the number of time periods T to infinity. The theory allows T to be very small compared to N. For example, the SNS critical value can be justified if $T^{-1/3}(\log N) \to 0$ under some regularity conditions. Our asymptotic results establish that our confidence sets are valid uniformly over a broad class of probability measures. This class is defined in terms of bounds on the moments of covariates and error terms. These bounds restrict the heaviness of the tails of the distribution of the error term and depend on the relative magnitudes of N and T.

Our theoretical analysis relies on and extends recent results from high-dimensional statistics. A high-dimensional analysis is required since the number of simultaneously tested inequalities, $(G-1)N$, is large compared to the number of time periods T that determine the quality of the Gaussian approximation. The analysis of the SNS procedure builds on an idea in Chernozhukov, Chetverikov, and Kato (2014, Theorem 4.1). We show that their theoretical approach can be extended to accommodate our choice of critical value as well as estimation error from a preliminary

[5]Both *moment selection* and *moment recentering* address possible slackness of moment inequalities. For a comparison of the two approaches, see Allen (2017). These methods are developed in Andrews and Soares (2010), Bugni (2010), Andrews and Barwick (2012), Chernozhukov, Chetverikov, and Kato (2014), and Romano, Shaikh, and Wolf (2014).

estimation of the group-specific regression curves.

New theoretical developments are required to provide a theoretical justification of the MAX and QLR procedures. These procedures employ unit-specific critical values. This renders our approach substantially different from the high-dimensional bootstrap procedure in Chernozhukov, Chetverikov, and Kato (2014). To prove the validity of our approach, we derive a Gaussian approximation of the joint behavior of all unit-wise tests. Our assumptions about the relative magnitudes of T and N trade off increased precision of the unit-wise approximation (larger T) against a more stringent uniformity requirement (larger N). Our other results combine unit-wise finite-sample bounds with an anti-concentration inequality (Chernozhukov, Chetverikov, and Kato 2015) to argue that the unit-wise test statistics can be replaced by certain oracle test statistics. The approximation error from this replacement is controlled uniformly over all units. The oracle statistics are then jointly approximated by their normal limit using a high-dimensional central limit theorem (Chernozhukov, Chetverikov, and Kato 2016).

We contribute new theoretical results for the QLR statistic in high-dimensional one-sided testing problems. The existing results focus on testing one-sided hypotheses for finite vectors (Wolak 1991; Rosen 2008), and the underlying theoretical arguments do not extend to the high-dimensional case. Our approach uses a new approximate anti-concentration bound for the limiting distribution of the QLR statistic. We combine this anti-concentration result with a high-dimensional central limit theorem for sparse-convex sets (Chernozhukov, Chetverikov, and Kato 2016) to derive the joint limiting distribution of the unit-wise tests.

Our theoretical justification of unit selection builds on Chernozhukov, Chetverikov, and Kato (2014). Although our approach implements a different idea, we can follow the broad strokes of their argument.

For all three tests of group membership, we allow for estimated group-specific regression curves. The tested moment inequalities depend on the group-specific regression curves, and a preliminary estimator of group-specific coefficients enters the testing problem as a nuisance parameter. Provided that the estimator satisfies a weak rate condition, its effect on the distribution of the unit-wise test statistics is not of first order and can be ignored when computing critical values. We are agnostic about the specific choice of estimator of the group-specific coefficients. For example, the estimator may be based on an auxiliary training data set where group memberships are observed. Alternatively, coefficients can be estimated without information about the true group memberships. This problem has received attention in the recent econometric literature and estimators based on *kmeans* clustering (Bonhomme and Manresa 2015; Vogt and Linton 2017) or penalization (Su, Shi, and Phillips 2016; Wang, Phillips, and Su 2016) are available.

We complement our asymptotic results by Monte Carlo experiments that study the performance of our procedures in finite samples. For panels with a small number of observed time periods, our simulation results indicate that the short-panel adjustment is essential for guaranteeing correct coverage of the joint confidence set. For long panels, the procedures yield good coverage both with and without finite-sample adjustment, confirming our asymptotic results. We also demonstrate that neither the MAX nor the QLR test statistic dominates the other. In a design with substantial heteroscedasticity, we study the benefits and limits of our procedure for unit selection.

To demonstrate the usefulness of our confidence set for group membership in practice we discuss two applications. First, we follow Wang, Phillips, and Su (2016) and study heterogeneous relationships between a minimum wage and unemployment in a US state panel. For some states the relationship is positive, but for others, it is negative. We illustrate how, based on our confidence set, a policy maker can select units that belong to the positive or negative effect group while controlling the probability of misclassification. Secondly, we study the country panel data on income and democracy from Acemoglu et al. (2008). We consider the specification with group-specific trends from Bonhomme and Manresa (2015). The panel is very short ($T = 7$), which makes inference on the classification problem challenging. Our joint confidence set is still informative. In a specification with four groups, it separates the two most extreme groups.

The rest of the paper is organized as follows. Section 2 discusses the related literature and Section 3 introduces our panel model with a group structure. Section 4 motivates our approach and defines the joint and unit-wise confidence sets for group membership. Section 5 gives an asymptotic justification of our procedures and Section 6 reports our simulation results. Finally, Section 7 discusses two applications of the new methods developed in this paper to real data sets.

2. Related Literature

Classifying units into discrete groups is one of the oldest problems in statistics and statistical decision theory (Pearson 1896). Popular modeling tools are finite mixture models (McLachlan and Peel 2004). These models offer a random-effect approach to modeling discrete heterogeneity (Bonhomme, Lamadon, and Manresa 2016). In computer science, classification and clustering problems are often tackled using machine learning (Friedman, Hastie, and Tibshirani 2009). Perhaps surprisingly, we have not been able to find any research on how to conduct joint inference on the population group structure in the machine learning literature.

Algorithms in machine learning compute posterior probabilities of group membership (Murphy 2012, Chapter 5.7.2).[6] In principle, it is possible to compute unit-wise Bayesian credible sets from the posterior distribution. Although this approach is appealing in applications in computer science, it is not always a useful approach for inference in the social sciences. Consider, for example, the problem of classifying e-mail into regular mail and spam.[7] The generation of an e-mail can be modeled as a two-stage process. The first stage draws a data generating process (DGP), and the second stage generates an e-mail from this DGP. A user of an e-mail client observes new e-mail repeatedly and is interested in inference that works well in "typical" cases. In this context, it makes sense to follow the Bayesian paradigm and take the randomness of the DGP into account. In the social sciences we typically observe only one draw of the DGP and we have to ascertain that our

[6]For example, in the case of finite mixture models, posterior probabilities can be computed in the E-step of the EM algorithm (Dempster, Laird, and Rubin 1977).

[7]This example is inspired by Murphy (2012, p.5).

inference is valid for this particular DGP. Our frequentist approach is uniformly valid over a large class of DGPs and therefore fulfills this requirement.

We follow the recent econometric literature and adapt a fixed effect approach that treats the unobserved group memberships as a structural parameter. Inference in panel models with a latent group structure has been studied in Lin and Ng (2012), Bonhomme and Manresa (2015), Sarafidis and Weber (2015), Ando and Bai (2016), Vogt and Linton (2017), Wang, Phillips, and Su (2016), Lu and Su (2017), Vogt and Schmid (2017), and Gu and Volgushev (2018).[8] Previous studies address inference with respect to the group-specific regression curves. We are the first to address inference on group membership.

Our theoretical analysis relies on the theory of self-normalized sums (de la Pena, Lai, and Shao 2009) and recent results in high-dimensional statistics, particularly the central limit theorems in Chernozhukov, Chetverikov, and Kato (2016) and the anti-concentration result in Chernozhukov, Chetverikov, and Kato (2015). We contribute new theoretical results for high-dimensional testing problems.

Our confidence set is based on a characterization of the true group memberships by a system of moment inequalities. A recent review of confidence sets constructed from moment inequalities is given in Canay and Shaikh (2016). Most of the previous literature focuses on finite systems of moment inequalities. Chernozhukov, Chetverikov, and Kato (2014) provide a framework for testing high-dimensional systems of moment inequalities.[9] Our approach builds on and extends their results. To compute our joint confidence set, we solve a multiple one-sided testing problem. We provide a theoretical argument for the validity of our procedure for a diverging number of simultaneously-tested hypotheses. Romano and Wolf (2018) study a similar testing problem in a simulation experiment, but do not provide an asymptotic analysis of their approach. Even though we develop our theoretical argument in the context of a specific application, our approach can be adapted easily to other simultaneous one-sided testing problems. We expect this contribution to the theory of one-sided testing in high dimensions to be of independent interest.

3. Setting

We observe panel data (y_{it}, x_{it}), $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where y_{it} is a scalar dependent variable and x_{it} is a covariate vector. We assume that units are partitioned into a finite set of groups $\mathbb{G} = \{1, \ldots, G\}$. Group membership is unobserved. The relationship between y_{it} and x_{it} is described by a linear model. Units within the same group share the same coefficient value. Between groups, coefficient values may vary. Let $\beta_{g,t}$ denote the vector of coefficients that applies to units in group $g \in \mathbb{G}$ at time $t = 1, \ldots, T$. Unit i's true group membership is denoted g_i^0. In period t, unit i's outcome is generated according to

$$y_{it} = x_{it}' \beta_{g_i^0, t} + u_{it}, \tag{1}$$

[8]Models with a latent group structure have also been proposed for data other than panel data (Shao and Wu 2005).
[9]Estimation with many moment inequalities is examined by Menzel (2014).

where u_{it} is an error term.

This paper addresses inference with respect to the vector of latent group memberships $\{g_i^0\}_{1\leq i\leq N}$. In most practical applications, the coefficient vector is unknown and constitutes an additional source of uncertainty. We assume that an estimator $\hat{\beta}_{g,t}$ of $\beta_{g,t}$ is available. For example, estimators based on the *kmeans* algorithm (Bonhomme and Manresa 2015) or on penalization (Su, Shi, and Phillips 2016, Wang, Phillips, and Su 2016) may be used. Under a weak rate condition, our procedure controls for uncertainty from parameter estimation.

In applications, two special cases of model (1) are of particular interest.

Example 1 (Random coefficient model with a group structure)**.** *The coefficient vector is assumed to be constant over time. The model is*

$$y_{it} = x_{it}'\beta_{g_i^0} + u_{it}.$$

Estimation of this model is considered in Su, Shi, and Phillips (2016) and Wang, Phillips, and Su (2016). For this specification, we consider also an extension that adds individual fixed effects. A heuristic discussion of to apply our procedures to models with individual fixed effects is given in Section C of the Supplementary Appendix. Following Wang, Phillips, and Su (2016), we apply the random coefficient model to the analysis of heterogeneous effects of a minimum wage.

Example 2 (The group fixed effect model)**.** *The set of regressors contains a constant term. The coefficient on the constant term is group-specific and varies over time. It is called the group fixed effect. The values of the coefficients on the time-varying regressors are the same for all groups and time periods. The model is*

$$y_{it} = w_{it}'\theta + \alpha_{g_i^0,t} + u_{it},$$

where w_{it} is a vector of time-varying regressors, θ is a common slope coefficient and $\alpha_{g_i^0,t}$ is the group fixed effect. This model is developed in Bonhomme and Manresa (2015). Following their lead, we apply it to the clustering of countries according to their respective trajectories of democratization.

4. Procedure

This section discusses our approach for constructing confidence sets for group membership. First, we provide a rigorous definition of the confidence sets for group memberships discussed in this paper. We then present a characterization of the true group memberships by a system of moment inequalities. Next, we propose three procedures for computing confidence sets. We also discuss finite-sample adjustments. Lastly, we present our algorithm for unit selection.

4.1. Definition of confidence set for group membership

We consider joint confidence sets for the entire group structure as well as unit-wise confidence sets for each unit i.

A joint confidence set quantifies uncertainty about the true group structure $\{g_i^0\}_{1\le i\le N}$. It is a non-empty random subset of the set of all possible group configurations $\mathbb{G}^N$ that contains the true group structure with a pre-specified probability. Let $\mathcal{P}(\cdot)$ denote the power set of its argument. For $0<\alpha<1$, the joint confidence set $\widehat{C}_\alpha$ with confidence level $1-\alpha$ is a random element from $\mathcal{P}(\mathbb{G}^N)\setminus\{\emptyset\}$ such that

$$\liminf_{N,T\to\infty}\inf_{P\in\mathbb{P}_N} P\left(\{g_i^0\}_{1\le i\le N}\in\widehat{C}_\alpha\right)\ge 1-\alpha, \tag{2}$$

where $\mathbb{P}_N$ is a set of probability measures that satisfy certain regularity conditions. A typical element of $\widehat{C}_\alpha$ is $\{g_i\}_{1\le i\le N}$ with $g_i\in\mathbb{G}$. If $\{g_i\}_{1\le i\le N}\in\widehat{C}_\alpha$, then we cannot exclude the possibility that $\{g_i^0\}_{1\le i\le N}=\{g_i\}_{1\le i\le N}$ at a confidence level of at least $1-\alpha$.

A unit-wise confidence set for unit i is a non-empty random subset of the set of possible group memberships $\mathbb{G}$ that contains i's true group membership g_i^0 with a pre-specified probability. The unit-wise confidence set $\widehat{C}_{\alpha,i}$ at confidence level $1-\alpha$ is a random element from $\mathcal{P}(\mathbb{G})\setminus\{\emptyset\}$ such that

$$\liminf_{T\to\infty}\inf_{P\in\mathbb{P}} P\left(g_i^0\in\widehat{C}_{\alpha,i}\right)\ge 1-\alpha,$$

where $\mathbb{P}$ is a set of probability measures.

A unit-wise confidence interval quantifies the uncertainty about the group membership of one specific unit. For example, if $\widehat{C}_{\alpha,i}$ is a singleton, say $\widehat{C}_{\alpha,i}=\{1\}$, then we may conclude at confidence level $1-\alpha$ that unit i belongs to group 1. On the other hand, if $\widehat{C}_{\alpha,i}=\mathbb{G}$ then, at confidence level $1-\alpha$, the data is not informative at all about i's group membership.

4.2. Motivation of our approach

The key insight of our approach is that each unit's group membership can be characterized by a system of moment inequalities that can be used for a statistical test of the hypothesis $H_0: g_i^0=g$. Our confidence set is constructed by inverting such a test. To focus on the main idea, we assume in this section that group-specific parameters are known.

The null hypothesis $H_0: g_i^0=g$ is equivalent to

$$\mathbb{E}\left[\left(y_{it}-x_{it}'\beta_{g,t}\right)^2\right]\le\mathbb{E}\left[\left(y_{it}-x_{it}'\beta_{h,t}\right)^2\right] \tag{3}$$

for all $h\in\mathbb{G}$ and $t=1,\dots,T$. This inequality is justified under $\mathbb{E}[u_{it}\mid x_{it}]=0$, which guarantees that the true DGP minimizes a least-squares criterion. It has been used previously by Bonhomme and Manresa (2015) as a basis for their estimation procedure.

To test (3), we introduce a mean-adjusted difference between squared residuals. Let

$$d_{it}(g,h) = \frac{1}{2}\left(\left(y_{it} - x_{it}'\beta_{g,t}\right)^2 - \left(y_{it} - x_{it}'\beta_{h,t}\right)^2 + \left(x_{it}'\left(\beta_{g,t} - \beta_{h,t}\right)\right)^2\right).$$

The first two terms on the right-hand side are squared residuals. The third term ensures that $d_{it}(g,h)$ has mean zero under the null hypothesis. This can best be seen by writing

$$d_{it}(g,h) = -u_{it}x_{it}'\left(\beta_{g,t} - \beta_{h,t}\right) + \left(\beta_{g,t} - \beta_{g_i^0,t}\right)' x_{it}x_{it}'\left(\beta_{g,t} - \beta_{h,t}\right). \tag{4}$$

Here, the first term on the right-hand side has mean zero under $\mathbb{E}[u_{it} \mid x_{it}] = 0$ and the second term vanishes for $g = g_i^0$. Thus, under $g_i^0 = g$ we have

$$\mathbb{E}\left[d_{it}(g,h)\right] = 0$$

for all $h \in \mathbb{G} \setminus \{g\}$.[10] If $g_i^0 \neq g$ then there is $h \in \mathbb{G} \setminus \{g\}$ such that

$$\mathbb{E}\left[d_{it}(g,h)\right] > 0.$$

To see this, note that choosing $h = g_i^0 \in \mathbb{G} \setminus \{g\}$ guarantees that $d_{it}(g,h)$ has a strictly positive mean if $\mathbb{E}[x_{it}x_{it}']$ has full rank.

In summary, we can base a test of $H_0 : g_i^0 = g$ on the vector

$$\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[d_{it}(g,h)]\right\}_{h\in\mathbb{G}\setminus\{g\}}. \tag{5}$$

For this vector, we test equality to zero against the alternative that at least one of its components is strictly positive.

Remark 1. *The explicit mean adjustment is our solution to the problem of possibly slack moment inequalities in* (3)*. It exploits the specific structure of our problem and ensures that we test inequalities that are binding under the null hypothesis. This turns the problem of testing the moment inequalities* (3) *into a one-sided testing problem for a vector of moments. In other testing problems with moment inequalities, a similar mean adjustment is not feasible and possible slackness of the tested inequalities has to be addressed in another way. A popular solution is to use data-driven methods to detect and eliminate slack inequalities (Andrews and Soares 2010; Andrews and Barwick 2012; Romano, Shaikh, and Wolf 2014).*

[10] The assumption $\mathbb{E}[u_{it} \mid x_{it}] = 0$ implies $\mathbb{E}[d_{it}(g_i^0, h) \mid x_{it}] = 0$. The conditional version can yield a more powerful test if there is a specific alternative and a function f such that the moment $\mathbb{E}[d_{it}(g_i^0, h)f(x_{it})]$ reveals more evidence against the null hypothesis than the moment $\mathbb{E}[d_{it}(g_i^0, h)]$. In our setting, relevant alternatives are detected by large positive values of the quadratic form in (4). Therefore, we do not expect that the power of the test can be improved by using a function f to look in another direction.

4.3. Procedures for computing confidence sets

Here, we describe how to construct our confidence sets. A unit-wise confidence set is computed by inverting a test for group membership. Our joint confidence set strings together Bonferroni-corrected unit-wise confidence sets.

Let $\hat{T}_i(g)$ denote a test statistic. For a pre-specified probability α, let $c_{\alpha,1,i}(g)$ denote a critical value. Moreover, let $\hat{g}_i$ denote a point estimator of g_i^0.[11] A unit-wise confidence set for unit i is given by

$$\widehat{C}_{\alpha,i} = \left\{ g \in \mathbb{G} : \widehat{T}_i(g) \leq c_{\alpha,1,i}(g) \right\} \cup \{\hat{g}_i\}\,.$$

Adding the estimated group membership guarantees that the confidence set is never empty.[12]

A joint confidence set for all units is constructed by combining Bonferroni-corrected unit-wise confidence sets. Let $c_{\alpha,N,i}(g)$ be a Bonferroni-corrected critical value. Our joint confidence set is given by

$$\widehat{C}_{\alpha} = \bigtimes_{1 \leq i \leq N} \left\{ g \in \mathbb{G} : \widehat{T}_i(g) \leq c_{\alpha,N,i}(g) \right\} \cup \{\hat{g}_i\}\,.$$

We consider different choices for the test statistic and the critical values. For $g \in \mathbb{G}$ and $t = 1, \ldots, T$, let $\hat{\beta}_{g,t}$ denote an estimator of $\beta_{g,t}$. Define

$$\hat{d}_{it}(g,h) = \frac{1}{2}\left(\left(y_{it} - x_{it}'\hat{\beta}_{g,t} \right)^2 - \left(y_{it} - x_{it}'\hat{\beta}_{h,t} \right)^2 + \left(x_{it}' \left(\hat{\beta}_{g,t} - \hat{\beta}_{h,t} \right) \right)^2 \right).$$

The test for group membership is based on the studentized statistic

$$\hat{D}_i(g,h) = \frac{\sum_{t=1}^{T} \hat{d}_{it}(g,h)}{\sqrt{\sum_{t=1}^{T} \left(\hat{d}_{it}(g,h) - \bar{\hat{d}}_{it}(g,h) \right)^2}},$$

where $\bar{\hat{d}}_{it}(g,h) = \sum_{t=1}^{T} \hat{d}_{it}(g,h)/T$. Let $\hat{D}_i(g) = \left\{ \hat{D}_i(g,h) \right\}_{h \in \mathbb{G} \setminus \{g\}}$ denote the vector that stacks the studentized statistics for $h \in \mathbb{G} \setminus \{g\}$. We consider two test statistics to measure the distance of $\hat{D}_{it}(g)$ from zero in the direction of the positive axes: the MAX statistic and the QLR statistic. They are defined, respectively, as

$$\hat{T}_i^{\mathrm{MAX}}(g) = \max_{h \in \mathbb{G} \setminus \{g\}} \hat{D}_i(g,h),$$
$$\hat{T}_i^{\mathrm{QLR}}(g) = \min_{t \leq 0} \left(\hat{D}_i(g) - t \right)' \widehat{\Omega}_i^{-1}(g) \left(\hat{D}_i(g) - t \right),$$

with $\widehat{\Omega}_i(g) = \widehat{\Omega}_i^*(g) + \max\{\epsilon - \det(\widehat{\Omega}_i^*(g)), 0\} I_{G-1}$, where I_{G-1} is the identity matrix in $\mathcal{R}^{G-1}$, $\widehat{\Omega}_i^*(g)$

[11] Typically, such an estimator is available as part of the procedure that estimates the group-specific parameters. If not, then such an estimator can be based on inequality (3) (cf. Bonhomme and Manresa 2015).

[12] This is also required for our algorithm for unit selection to work.

is the $(G-1) \times (G-1)$ sample correlation matrix with entries

$$\left(\widehat{\Omega}_i^*(g)\right)_{h,h'} = \frac{\sum_{t=1}^T \left(\hat{d}_{it}(g,h) - \bar{\hat{d}}_{it}(g,h)\right)\left(\hat{d}_{it}(g,h') - \bar{\hat{d}}_{it}(g,h')\right)}{\sqrt{\sum_{t=1}^T \left(\hat{d}_{it}(g,h) - \bar{\hat{d}}_{it}(g,h)\right)^2 \sum_{t=1}^T \left(\hat{d}_{it}(g,h') - \bar{\hat{d}}_{it}(g,h')\right)^2}},$$

and ϵ is a positive parameter that controls the regularization of the sample correlation matrix (cf. Andrews and Barwick 2012).[13]

For the MAX test statistic we offer two different strategies for computing critical values. The *SNS critical value* is given by

$$c_{\alpha,N,i}^{\text{SNS}}(g) = c_{\alpha,N}^{\text{SNS}} = \sqrt{\frac{T}{T-1}} t_{T-1}^{-1}\left(1 - \frac{\alpha}{(G-1)N}\right),$$

where $t_{T-1}^{-1}(p)$ denotes the p-th quantile of a t-distribution with $T-1$ degrees of freedom. This critical value does not depend on any characteristics of the unit and is justified under relatively mild conditions on moments. We refer to the combination of the MAX statistic and SNS critical values as the SNS procedure. The corresponding joint confidence set is denoted by $\widehat{C}_\alpha^{\text{SNS}}$.

Our second strategy for computing critical values explicitly takes the correlation of the within-unit moments into account. Note that although the SNS critical value is robust against this correlation, it can be conservative in the presence of a strong correlation of the within-unit moments. In the literature on testing moment inequalities, the preferred way to capture correlation of the moment inequalities is to compute critical values from a bootstrap distribution that replicates the correlation (Romano, Shaikh, and Wolf 2014; Chernozhukov, Chetverikov, and Kato 2014). In our setting with unit-specific critical values, a naïve application of the bootstrap is computationally intractable.[14] Instead, we suggest an analytical critical value that is easy to compute with modern software. Even though the implementation is not based on Monte Carlo methods, our analytical critical value is mathematically equivalent to multiplier bootstrap with Gaussian multipliers; we call this the *bootstrap critical value* for the MAX statistic.

The bootstrap critical for the MAX statistic is given by

$$c_{\alpha,N,i}^{\text{MAX}}(g) = c_{\alpha,N}^{\text{MAX}}\left(\widehat{\Omega}_i(g)\right) = \Phi_{\max,\widehat{\Omega}_i(g)}^{-1}\left(1 - \frac{\alpha}{N}\right),$$

where $\Phi_{\max,V}$ denotes the distribution function of the maximal entry of a centered normal random vector with covariance matrix V. This critical value can be computed by inverting a multivariate normal probability and is straightforward to implement in modern statistical software.[15] We refer

[13] We do not study the choice of ϵ. In the simulations in Section 6 and the applications in Section 7 we follow Andrews and Barwick (2012) and set $\epsilon = 0.012$.

[14] The unit-wise critical values are large quantiles of a bootstrap distribution and are difficult to approximate accurately by unsophisticated Monte Carlo methods.

[15] For $Z \sim \mathcal{N}(0, V)$ and a scalar a, $P(\max_j Z_j \leq a) = P(Z \leq (a, \ldots, a)')$. Multivariate normal probabilities can be efficiently approximated by modern algorithms (Genz 1992). Such algorithms are implemented in the Stata package MVTNORM (Grayling and Mander 2016) and the R package mnormt (Azzalini and Genz 2016).

to the combination of the MAX statistic and the bootstrap critical values as the MAX procedure. The corresponding joint confidence set is denoted by $\widehat{C}_{\alpha}^{\mathrm{MAX}}$.

To define the critical value for the QLR test statistic, let $w(\cdot,\cdot,\cdot)$ denote the weight function defined in Kudo (1963). For a $(G-1)\times(G-1)$ covariance matrix V, define the distribution function $F_{\mathrm{QLR},V}$ by

$$F_{\mathrm{QLR},V}(t) = 1 - \sum_{j=1}^{G-1} w\left(G-1, G-1-j, V\right) P\left(\chi_j^2 > t\right), \tag{6}$$

where χ_j^2 has a χ^2-distribution with j degrees of freedom. The critical value for the QLR statistic is given by

$$c_{\alpha,N,i}^{\mathrm{QLR}}(g) = c_{\alpha,N}^{\mathrm{QLR}}\left(\widehat{\Omega}_i(g)\right) = F_{\mathrm{QLR},\widehat{\Omega}_i(g)}^{-1}\left(1-\frac{\alpha}{N}\right).$$

The weight function $w(\cdot,\cdot,\cdot)$ can be represented by a function of multivariate normal probabilities and is easily computed in statistical software (cf. footnote 15). We refer to this strategy for computing the confidence set as the QLR procedure. The corresponding joint confidence set is denoted by $\widehat{C}_{\alpha}^{\mathrm{QLR}}$.

4.4. Critical values for short panels

We suggest a heuristic correction of critical values to improve performances in short panels (i.e., panels where T is small). The critical values introduced above are based on theoretical results that allow the number of observed time periods to be very small compared to the number of units but still require $T\to\infty$ (see Section 5). It is not clear whether our asymptotic approximation is sufficiently accurate if T is small. Our heuristic correction for short panels is motivated by the SNS procedure and calibrated so that all three procedures produce the same confidence set in settings with $G=2$ groups. The SNS procedure can be justified for finite T under an additional normality assumption and does not require a short-panel adjustment.

For the MAX procedure the adjustment is based on the multivariate t-distribution. Let $F_{\max,V,T-1}^{f}$ denote the distribution function of the maximal entry of a random vector with multivariate t-distribution with scale matrix V and $T-1$ degrees of freedom. The adjusted critical value is given by

$$c_{\alpha,N,i}^{\mathrm{MAX},f} = c_{\alpha,N}^{\mathrm{MAX},f}\left(\widehat{\Omega}_i(g)\right) = \sqrt{\frac{T}{T-1}}\left(F_{\max,\widehat{\Omega}_i(g),T-1}^{f}\right)^{-1}\left(1-\frac{\alpha}{N}\right).$$

For the QLR procedure the adjustment is based on a mixture of F-distributions, as in Wolak (1987). Let

$$F_{\mathrm{QLR},\widehat{\Omega}_i(g)}^{f}(t) = 1 - \sum_{j=1}^{G-1} w\left(G-1, G-1-j, \widehat{\Omega}_i(g)\right) P\left(F_{j,T-1} > t/j\right),$$

where $F_{j,\nu}$ has an F-distribution with j and ν degrees of freedom. The adjusted critical value is given by

$$c_{\alpha,N,i}^{\text{QLR},f} = c_{\alpha,N}^{\text{QLR},f}\left(\widehat{\Omega}_i(g)\right) = \sqrt{\frac{T}{T-1}}\left(F^f_{\text{QLR},\widehat{\Omega}_i(g)}\right)^{-1}\left(1-\frac{\alpha}{N}\right).$$

All three procedures with short-panel adjustment yield the same confidence level when $G = 2$. In this case, each unit's group membership is completely described by only one moment inequality. Equivalence of the MAX procedure with short panel adjustment and the SNS procedure is immediate. For the QLR procedure, note that

$$\hat{T}_i^{\text{QLR}} = (\max(\hat{D}_i(g,h),0))^2 = \left(\hat{T}_i^{\text{SNS}}\right)^2$$

if $G = 2$ and $\hat{T}_i^{\text{SNS}} \geq 0$. The QLR statistic computes critical values from a F-distribution with 1 and $T-1$ degrees of freedom or, equivalently, a squared t_{T-1}-distribution. This establishes equivalence of the QLR procedure with short-panel adjustment and the SNS procedure.

For $G > 2$, the adjusted critical values do not reflect the finite sample distribution of the respective test statistic under an additional normality assumption. Tracking the exact distribution under normality, although desirable, is at odds with our goal of offering confidence sets that can be easily implemented and cheaply computed.

4.5. Unit selection

We propose an algorithm that detects units whose group membership is "obvious". These units can be ignored when computing the Bonferroni correction in the definition of the critical values. The algorithm combines moment selection and iterative hypothesis selection. The group membership for a unit becomes obvious if two conditions are simultaneously met. First, a test statistic that measures the difference between the left- and the right-hand side of (3) for $g = \hat{g}_i$ and $h \neq \hat{g}_i$ takes a large negative value. This corresponds to moment selection. Second, all alternative group memberships $h \neq \hat{g}_i$ are rejected. This corresponds to hypothesis selection.

The algorithm for unit selection can be combined with any of the test statistics and critical values discussed above. For $i = 1, \ldots, N$, let $\hat{T}_i^{\text{type}}$ denote a unit-wise test statistic and $c_{\alpha,N,i}^{\text{type}}$ denote a corresponding critical value, where type $=$ SNS, MAX or QLR. Our algorithm is parameterized by β, $0 \leq \beta < \alpha/3$. The larger β, the more unit selection is carried out. Setting β to zero switches off unit selection.

Moment selection is based on a counterpart to $\hat{D}_i$ which does not adjust for the mean under the null hypothesis. It is given by

$$\hat{D}_i^U(g,h) = \frac{\sum_{t=1}^T \hat{d}_{it}^U(g,h)}{\sqrt{\sum_{t=1}^T\left(\hat{d}_{it}^U(g,h) - \bar{\hat{d}}_i^U(g,h)\right)^2}},$$

where

$$\hat{d}^U_{it}(g,h) = (y_{it} - x'_{it}\hat{\beta}_{g,t})^2 - (y_{it} - x'_{it}\hat{\beta}_{h,t})^2$$

and $\bar{\hat{d}}^U_i(g,h) = \sum_{t=1}^T \hat{d}^U_{it}(g,h)/T$. For $g \in \mathbb{G}$ and $i = 1, \ldots, N$, let

$$\widehat{M}_i(g) = \left\{h \in \mathbb{G} \setminus \{g\} \mid \hat{D}^U_i(g,h) > -2c^{\text{SNS}}_{\beta,N}\right\}.$$

This set gives the selected inequalities for the hypothesis $H_0 : g^0_i = g$. Here we use the SNS critical value, but other choices may also be possible. Our algorithm proceeds as follows:

1. Set $s = 0$ and $H_i(0) = \mathbb{G}$.
2. Set $\hat{N}(s) = \sum_{i=1}^N \max_{g \in H_i(s)} \mathbf{1}\{\#\widehat{M}_i(g) \neq 0\}$.
3. Set

   $$H_i(s+1) = \left\{g \in \mathbb{G} \mid \hat{T}^{\text{type}}_i(g) \leq c^{\text{type}}_{\alpha-2\beta,\hat{N}(s),i}(g)\right\} \cup \{\hat{g}_i\}.$$

   If $H_i(s+1) = H_i(s)$ for all i then go to Step 5.
4. Set $s = s + 1$. Go to Step 3.
5. The confidence set with unit selection is given by $\widehat{C}^{\text{type}}_{\text{sel},\alpha,\beta} = \times_{1 \leq i \leq N} H_i(s+1)$.

Step 2 of the algorithm counts the number $\hat{N}(s)$ of units whose memberships are not obvious. This number is used for computing critical values. Step 3 carries out hypothesis selection. For each unit i, group memberships $g \in H_i(s+1)^c = \mathbb{G} \setminus H_i(s+1)$ are not rejected under the critical value that accounts for $\hat{N}(s)$ simultaneously tested units. We iterate moment selection (Step 2) and hypothesis selection (Step 3) until convergence. Typically, moment selection renders a unit's group membership "obvious" if the set $H_i(s+1)$ is a singleton so that $H_i(s+1) = \{\hat{g}_i\}$. Otherwise, it is likely that $\widehat{M}_i(g)$ is non-empty for some $g \neq \hat{g}_i$. Note that for hypothesis selection (Step 3) we exploit the information revealed by moment selection (Step 2) and use the critical value computed under $\hat{N}(s)$.

If there is a sufficient number of units for which group membership is "obvious" then $\widehat{C}^{\text{type}}_{\text{sel},\alpha,\beta}$ is more powerful ("smaller") than the confidence set $\widehat{C}^{\text{type}}_{\alpha}$ without moment selection. However, there is a cost of unit selection. When computing the critical value we replace α by $\alpha - 2\beta$. This adjustment controls two possible errors that each occur with probability β. The first error is estimating an incorrect group membership for a unit whose group membership is obvious "in population". The second error is classifying a non-obvious unit as obvious. Because of this cost of unit selection, confidence sets with unit selection can be more conservative ("larger") than those without if an insufficient number of units is eliminated.

Remark 2. *The unit selection procedure may be understood as a data-driven way to allocate error probability to each unit. Let α_i denote the probability that the unit-wise confidence level for unit i*

does not include the true group membership. In principle, we may distribute the total error probability α arbitrarily among the N units as long as $\sum_{i=1}^{N} \alpha_i = \alpha$. Without unit selection our procedures allocate the error probability evenly so that $\alpha_i = \alpha/N$. In our discrete testing problem, this even allocation of the failure probability can render the joint confidence set overly conservative. Each unit's marginal confidence set contains at least one group. For units that are very easy to classify, the probability that a singleton set containing only the estimated group membership does not cover the truth is less than the error probability α/N.This is a potential source of overly conservative behavior of the joint confidence set. Our algorithm for unit selection reshuffles allocated error probability from units that are easy to classify to units that are hard to classify.

Remark 3. *Our unit selection procedure builds on moment selection procedures developed by Chernozhukov, Chetverikov, and Kato (2014) and others. Allen (2017) points out that the moment recentering procedure of Romano, Shaikh, and Wolf (2014) yields a more powerful test. However, the moment recentering procedure has not been developed for settings such as ours where many moment inequalities are tested simultaneously. Note that in our setting, there is no point in doing moment selection, since the recentered inequalities are binding under the null hypothesis. Still, unit selection is possible because the estimated group memberships are always included.*

5. Asymptotic results

In this section, we establish theoretically that our procedures yield joint confidence sets that asymptotically cover the truth with a pre-specified probability, i.e., we show that (2) holds. Our results exploit recent developments in high-dimensional statistics. In particular, we rely on high-dimensional central limit theorems and anti-concentration inequalities for high-dimensional settings. We also provide new contributions to this field. All proofs are in the Appendix.

For the justification of the unit-wise confidence sets, we refer to existing results for confidence sets for finite-dimensional parameters defined by moment inequalities (Rosen 2008; Romano, Shaikh, and Wolf 2014).

5.1. Asymptotic framework and assumptions

Our asymptotic framework is of the long-panel variety and takes both the number of units N and the number of time periods T to infinity. In most panel data sets, the number of units far outstrips the number of time periods. We replicate this feature along the asymptotic sequence by allowing N to diverge at a much faster rate than T.

We introduce some assumptions. For a probability measure P, let $\mathbb{E}_P$ denote the expectation operator that integrates with respect to the measure P.

Assumption 1. *(i) The set of latent groups is enumerated as $\mathbb{G} = \{1, \ldots, G\}$. For $g, h \in \mathbb{G}$ and $g \neq h$, $\max_{1 \leq t \leq T} \|\beta_{g,t} - \beta_{h,t}\| > 0$. There exists K_β such that $\max_{g \in \mathbb{G}} \max_{1 \leq t \leq T} \|\beta_{g,t}\| \leq K_\beta$.*

(ii) *P is a probability measure such that, for $N, T \geq 1$, for each unit $i = 1, \ldots, N$, $(u_{it})_{1 \leq t \leq T}$ is an independent sequence with $\mathbb{E}_P[u_{it} \mid x_{it}] = 0$ and $\mathbb{E}_P(u_{it}^2) = \sigma_i^2$ and, for $t = 1, \ldots, T$, the matrix $\mathbb{E}_P(x_{it}x_{it}')$ is of full rank. There exists $\underline{\sigma} > 0$ such that $\mathbb{E}_P[(u_{it}/\sigma_i)^2 \mid x_{it}] \geq \underline{\sigma}^2$.*

(iii) *There exists a sequence $\gamma_{N,T,8}$ and estimators $\hat{\beta}_g$ of β_g for all $g \in \mathbb{G}$ such that*

$$P\left(\max_{g \in \mathbb{G}} \left(\frac{1}{T}\sum_{t=1}^{T} \left\|\hat{\beta}_{g,t} - \beta_{g,t}\right\|^8\right)^{1/8} > \gamma_{N,T,8}\right) \leq \xi_{N,T}$$

for a vanishing sequence $\xi_{N,T}$.

(iv) *Along the asymptotic sequence $T \leq N$ and $T^{-1/2}(\log N) \leq 1$ and, for $t = 1, \ldots, T$, the moment $\mathbb{E}_P\left[|u_{it}/\sigma_i|^8 \|x_{it}\|^8 + \|x_{it}\|^{16}/\sigma_i\right]$ exists.*

Part (i) restricts the group structure. The set of latent groups is assumed to be finite with known cardinality. Groups are unique, i.e., there are no groups that share the same coefficient values. We also assume that group-specific coefficients take values in a bounded set. This is a technical assumption that can be relaxed at the expense of a more involved statement of the asymptotic results.

Next, Part (ii) imposes assumptions on the error term. Most importantly, we assume that the innovations are independent. This rules out serial correlation. Our proofs build on recent advances in the theory of asymptotic approximations in high-dimensional settings that are currently only available for independent innovations.[16] In the future, as new results become available, it may be possible to extend our results to settings with weakly dependent observations.

Part (iii) requires existence of an estimator $\hat{\beta}_{g,t}$ that is consistent for $\beta_{g,t}$ at a certain rate. Suppose, for example, that the group-specific coefficients are estimated from an auxiliary data set with N_{aux} observations. Under some regularity conditions we can take $\gamma_{N,T,8} = O\left(N_{\text{aux}}^{-1/2}\right)$. In settings in which the coefficients are estimated without explicit knowledge about the true group memberships, rate calculations can be based on the results in Bonhomme and Manresa (2015), Su, Shi, and Phillips (2016), and Wang, Phillips, and Su (2016). These methods provide $\sqrt{NT}$ consistent estimators when the coefficients are time invariant (i.e., $\beta_{g,t} = \beta_g$).

Finally, Part (iv) is a technical assumption that guarantees the existence of all moments that enter the statements of the theorems below.

For the asymptotic analysis, it is convenient to write

$$\hat{D}_i(g,h) = \frac{T^{-1/2}\sum_{t=1}^{T} \hat{d}_{it}(g,h)/\sigma_i}{\hat{S}_{i,T}(g,h)},$$

[16] A high-dimensional CLT for possibly dependent data is proved in Zhang and Cheng (2017) for the MAX statistic. There exist some attempts to extend the SNS theory to dependent data (see, e.g., Chen et al. 2016). We are not aware of a high-dimensional anti-concentration inequality for dependent data.

where

$$\hat{S}_{i,T}^2(g,h) = \frac{1}{\sigma_i^2 T}\sum_{t=1}^{T}\left(\hat{d}_{it}(g,h) - \bar{\hat{d}}_{it}(g,h)\right)^2$$

and $\bar{\hat{d}}_{it}(g,h) = \sum_{t=1}^{T}\hat{d}_{it}(g,h)/T$. The population counterpart of $\hat{S}_{i,T}^2(g,h)$ is given by

$$s_{i,T}^2(g,h) = \frac{1}{\sigma_i^2 T}\sum_{t=1}^{T}\mathbb{E}\left(d_{it}(g,h) - \mathbb{E}[d_{it}(g,h)]\right)^2.$$

Let P denote a probability measure that satisfies Assumption 1. For a matrix A, let $\lambda_1(A)$ denote A's smallest eigenvalue. Assumptions 1(i) and (ii) imply

$$s_{i,T}^2(g_i^0,h) \geq \underline{\sigma}^2 \min_{1\leq i\leq N}\min_{h\in\mathbb{G}\setminus\{g_i^0\}}\frac{1}{T}\sum_{t=1}^{T}\lambda_1(\mathbb{E}_P(x_{it}x_{it}'))\|\beta_{g_i^0,t} - \beta_{h,t}\|^2 =: \underline{s}_{N,T}^2(P) > 0.$$

The theorems below define a class $\mathbb{P}_N$ of probability measures. This class satisfies a number of moment conditions that are defined in terms of

$$\begin{aligned}
B_{N,T,p}(P) &= \max_{1\leq t\leq T}\left(\mathbb{E}_P\left[\max_{1\leq i\leq N}\left(|u_{it}/\sigma_i|^p\,\|x_{it}\|^p + \|x_{it}\|^{2p}/\sigma_i\right)\right]/\underline{s}_{N,T}^p(P)\right)^{1/p},\\
D_{N,T,p}(P) &= \max_{1\leq i\leq N}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_P\left[|u_{it}/\sigma_i|^p\,\|x_{it}\|^p + \|x_{it}\|^{2p}/\sigma_i\right]/\underline{s}_{N,T}^p(P)\right)^{1/p}.
\end{aligned}$$

In the following, for all quantities that depend on the probability measure P, this dependence is kept implicit.

5.2. The SNS procedure

In this section, we establish validity of the joint confidence set based on the MAX test statistic with SNS critical values.

Theorem 1. *Let $\mathbb{P}_N$ denote a sequence of classes of probability measures that satisfy Assumption 1, and let*

$$\begin{aligned}
\epsilon_{1,N} &= \sup_{P\in\mathbb{P}_N}\gamma_{N,T,8}(\log N)\big(T^{-5/24}B_{N,T,8}^2\sqrt{\log N} + D_{N,T,4}\big),\\
\epsilon_{2,N} &= \sup_{P\in\mathbb{P}_N}\gamma_{N,T,8}\sqrt{T\log N}D_{N,T,2},\\
\epsilon_{3,N} &= \sup_{P\in\mathbb{P}_N}T^{-1/6}D_{N,T,3}\sqrt{\log N}.
\end{aligned}$$

and $\epsilon_N = \epsilon_{1,N} + \epsilon_{2,N} + \epsilon_{3,N} + \xi_{N,T}$*. Suppose that* $\epsilon_N \to 0$ *and*

$$\max_{P \in \mathbb{P}_N} T^{-5/24} B_{N,T,4} \sqrt{\log N} \leq 1. \tag{7}$$

Then, for each $0 < \alpha < 1$*, there is a constant* C *depending only on* α*,* G*,* K_β *and the sequence* ϵ_N *such that*

$$\sup_{P \in \mathbb{P}_N} P\left(\left\{g_i^0\right\}_{1 \leq i \leq N} \in \widehat{C}_\alpha^{\mathrm{SNS}}\right) \geq 1 - \alpha - C\epsilon_N.$$

This theorem states that the SNS confidence set contains the true group membership structure at least with probability $1 - \alpha - C\epsilon_N$. Note that the rate of convergence ϵ_N does not depend on P. Hence, convergence is uniform over $\mathbb{P}_N$.

The outline of the proof is as follows. We first replace $\hat{D}_i(g_i^0, h)$ by

$$\tilde{D}_i\left(g_i^0, h\right) := \frac{\sum_{t=1}^T d_{it}(g_i^0, h)}{\sqrt{\sum_{t=1}^T \left(d_{it}(g_i^0, h) - \bar{d}_{it}(g_i^0, h)\right)^2}}.$$

The rates $\epsilon_{1,N}$ and $\epsilon_{2,N}$ bound the rate at which $\widehat{D}_i(g_i^0, h)$ converges to $\tilde{D}_i\left(g_i^0, h\right)$. Thus, they represent the effect of estimating the group-specific coefficients. The distribution of $\tilde{D}_i\left(g_i^0, h\right)$ is approximated by a t-distribution scaled by the factor $\sqrt{T/(T-1)}$.[17] This approximation contributes $\epsilon_{3,N}$ to the overall convergence rate and relies on a Cramér-type moderate deviation inequality for self-normalized sums (Jing, Shao, and Wang 2003). Note that Condition (7) is non-essential. It is imposed to simplify the statement of the theorem. It can be relaxed at the expense of inflating $\epsilon_{1,N}$ and $\epsilon_{2,N}$.

Our result holds even if T is very small compared to N. For example, if $D_{N,T,3}$ is bounded along the asymptotic sequence then $\epsilon_{3,N}$ vanishes if $T^{-1/3}(\log N) \to 0$, allowing T to diverge to infinity at a much slower rate than N. We therefore expect that the confidence set performs well even if the panel is rather short.

Although the usefulness of the SNS theory in testing many inequalities was first discovered in Chernozhukov, Chetverikov, and Kato (2014, Theorem 4.1), our result differs in two ways. First, we use a critical value that is computed from a t-distribution, whereas their critical value is computed by transforming normal quantiles. Our approach offers an appealing symmetry between the small T setting with an additional normality assumption and the large T setting without a parametric assumption. Moreover, whereas the critical value in Chernozhukov, Chetverikov, and Kato (2014) is not defined for small T, our critical value is always computable.[18] To prove the validity of our critical value, we extend the argument in Chernozhukov, Chetverikov, and Kato (2014) by an additional approximation step. Following their argument, we first apply the Cramér-type inequality to show

[17] If the $d_{it}(g_i^0, h)$ are normally distributed, then this is the exact distribution.

[18] In our setting, the critical value in Chernozhukov, Chetverikov, and Kato (2014) is given by $\Phi^{-1}(1-\alpha/((G-1)N))/\sqrt{1-\Phi^{-1}(1-\alpha/((G-1)N))^2/T}$. If T is small, then the term inside of the square root can be negative.

that quantiles of $\tilde{D}_i(g_i^0, h)$ can be approximated by a function of normal quantiles. Our second approximation step establishes that this function of normal quantiles is well approximated by our critical value.

Second, Chernozhukov, Chetverikov, and Kato (2014, Theorem 4.1) do not consider parameter uncertainty, whereas our results quantify the effect of estimating the group-specific parameters under low-level assumptions that are easy to interpret.[19] In our proof, we reduce the problem with estimated parameters to a problem with known parameters. To this end, we bound the probability that the test rejects by the probability that the oracle statistic $\tilde{D}_i(g_i^0, h)$ exceeds the critical value associated with confidence level $1-\alpha_N$ for $\alpha_N > \alpha$. Based on a careful analysis of the tail of the t-distribution, we can show that, under the assumptions of the theorem, there is asymptotically no effect of replacing α by α_N.

5.3. The MAX procedure

In this section, we establish that the MAX procedure produces an asymptotically valid confidence set. Our result requires slightly stronger assumptions than the corresponding theorem for the SNS procedure.

We allow for strong correlation of the within-unit moment inequalities. Let $\Omega_i(g_i^0)$ denote the $(G-1)\times(G-1)$ correlation matrix with entries

$$\left(\Omega_i(g_i^0)\right)_{h,h'} = \frac{\sum_{t=1}^T \mathbb{E}\left[d_{it}(g_i^0,h)d_{it}(g_i^0,h')\right]}{\sqrt{\sum_{t=1}^T \mathbb{E}\left[d_{it}^2(g_i^0,h)\right]\sum_{t=1}^T \mathbb{E}\left[d_{it}^2(g_i^0,h')\right]}}.$$

For our theoretical result below, we assume that $\Omega_i(g_i^0)$ is nonsingular. In particular, pairs of moment inequalities are not perfectly correlated. To model strong correlation of the moment inequalities, we allow the correlation matrix to approach singularity at a controlled rate.

Theorem 2. *Suppose that there is a sequence $\omega_N > 0$ such that $\lambda_1(\Omega_i(g_i^0)) \geq \omega_N^{-1}$ for $i = 1, \ldots, N$. Let $\mathbb{P}_N$ denote a sequence of classes of probability measures that satisfy Assumption 1, and let*

$$\begin{aligned}
\epsilon_{1,N} &= \sup_{P\in\mathbb{P}_N} \gamma_{N,T,8}(\log N)\big(T^{-3/14}B_{N,T,8}^2\sqrt{\log N} + D_{N,T,4}\big),\\
\epsilon_{2,N} &= \sup_{P\in\mathbb{P}_N} \gamma_{N,T,8}\sqrt{T\log N}D_{N,T,2},\\
\epsilon_{3,N} &= \sup_{P\in\mathbb{P}_N} T^{-1/7}B_{N,T,4}\log N.
\end{aligned}$$

and

$$\epsilon_N = \big(\epsilon_{1,N}+\epsilon_{3,N}\big)(\omega_N^2 \vee 1) + \epsilon_{2,N} + \xi_{N,T}.$$

[19] Chernozhukov, Chetverikov, and Kato (2014) consider parameter uncertainty for their bootstrap procedures, but not for their SNS procedures. For their bootstrap procedure they give a high-level assumption under which parameter uncertainty can be ignored.

Suppose that $\epsilon_N \to 0$ and $T^{-1/7}(\log N) \to 0$. Then, for each $0 < \alpha < 1$ there is a constant C depending only on α, G and K_β and the sequence ϵ_N such that

$$\sup_{P \in \mathbb{P}_N} P\left(\left\{g_i^0\right\}_{1 \le i \le N} \in \widehat{C}_\alpha^{\mathrm{MAX}}\right) \ge 1 - \alpha - C\epsilon_N.$$

The theorem states that the empirical coverage probability of the MAX confidence set is at least $1 - \alpha - C\epsilon_N$. As in Theorem 1, our result establishes that the coverage probability converges uniformly over $\mathbb{P}_N$ to the nominal level.

The proof of Theorem 2 relies on two new oracle results. The first establishes that $\hat{D}_i(g_i^0, h)$ can be replaced by $D_i(g_i^0, h)$, where

$$D_i(g_i^0, h) := \frac{T^{-1/2} \sum_{t=1}^T d_{it}(g_i^0, h)/\sigma_i}{s_{i,T}(g_i^0, h)}.$$

The cost of estimating the group-specific parameters is given by $\epsilon_{1,N}$ and $\epsilon_{2,N}$. Note that, in contrast to the proof of Theorem 1, we eliminate the randomness of the denominator before deriving a distributional result. We prove this result by combining point-wise bounds with a high-dimensional anti-concentration inequality (Chernozhukov, Chetverikov, and Kato 2015, Corollary 1). Then, we approximate $D_i(g_i^0, h)$ by its normal limit using a high-dimensional central limit theorem (Chernozhukov, Chetverikov, and Kato 2016). This step of the proof contributes $\epsilon_{3,N}$ to the overall convergence rate. If the support of u_{it} and x_{it} can be bounded uniformly over i, then $\epsilon_{3,N}$ vanishes if $T^{-1/7}(\log N) \to 0$. This is a stronger condition than what is required in Theorem 1.

The second oracle result establishes that the critical value $c_{\alpha,N}^{\mathrm{MAX}}(\widehat{\Omega}_i(g_i^0))$ can be replaced by $c_{\alpha_N,N}^{\mathrm{MAX}}(\Omega_i(g_i^0))$ for some $\alpha_N \to \alpha$. Under the normal approximation of $D_i(g_i^0, h)$, $c_{\alpha_N,N}^{\mathrm{MAX}}(\Omega_i(g_i^0))$ is the critical value that gives a unit-wise confidence set with coverage $1 - \alpha_N/N$.

5.4. The QLR procedure

We now establish that the QLR confidence set has asymptotically the correct coverage. To the best of our knowledge, our formal result below represents the first theoretical analysis of the QLR statistic in a high-dimensional setting.

Theorem 3. *Suppose that there is a constant λ_1 such that $\lambda_1(\Omega_i) \ge \lambda_1 > 0$ for $i = 1, \dots, N$. Let $\mathbb{P}_N$ denote a sequence of classes of probability measures that satisfy Assumption 1, and let*

$$\begin{aligned}
\epsilon_{1,N} &= \sup_{P \in \mathbb{P}_N} \gamma_{N,T,8}(\log N)\big(T^{-3/14} B_{N,T,8}^2 \sqrt{\log N} + D_{N,T,4}\big), \\
\epsilon_{2,N} &= \sup_{P \in \mathbb{P}_N} \gamma_{N,T,8} \sqrt{T \log N} D_{N,T,2}, \\
\epsilon_{3,N} &= \sup_{P \in \mathbb{P}_N} T^{-1/7} B_{N,T,4} \log N,
\end{aligned}$$

and $\epsilon_N = \epsilon_{1,N} + \epsilon_{2,N} + \epsilon_{3,N} + \xi_{N,T}$. Suppose that $\epsilon_N \to 0$ and $T^{-1/7}(\log N) \to 0$. and that all

$P \in \mathbb{P}_N$ *impose cross-sectional independence. Then, for each* $0 < \alpha < 1$ *there is a constant* C *depending only on* α, λ_1, G, K_β *and the sequence* ϵ_N *such that*

$$\sup_{P \in \mathbb{P}_N} P\left(\left\{g_i^0\right\}_{1 \le i \le N} \in \widehat{C}_\alpha^{\text{QLR}}\right) \ge 1 - \alpha - C\epsilon_N.$$

This theorem establishes the validity of the QLR approach under similar assumptions as those imposed in Theorem 2.

The proof of Theorem 3 follows the same outline as that of Theorem 2. However, the arguments for establishing some of the steps are different and require new theoretical results. Let $D_i(g_i^0) = \{D_i(g_i^0, h)\}_{h \in \mathbb{G} \setminus \{g_i^0\}}$ and let

$$T_i^{\text{QLR}}(g_i^0) = \max_{t \le 0} \left(D_i(g_i^0) - t\right)' \Omega_i^{-1}(g_i^0) \left(D_i(g_i^0) - t\right).$$

We first apply a new anti-concentration result to justify that we can replace $\hat{T}_i^{\text{QLR}}(g_i^0)$ by $T_i^{\text{QLR}}(g_i^0)$. We then show that the set of values of $D_i(g_i^0)$ that map into rejections, i.e., that yield $T_i^{\text{QLR}}(g_i^0) > c_{\alpha,N}^{\text{QLR}}(\Omega_i)$, is a convex set in $\mathcal{R}^{G-1}$. This observation allows us to employ the central limit theorem for sparse-convex sets in Chernozhukov, Chetverikov, and Kato (2016, Proposition 3.2) from which we conclude that the oracle test statistics $\{T_i^{\text{QLR}}\}_{1 \le i \le N}$ converge *jointly* to their normal limits. For each $i = 1, \ldots, N$, the limiting distribution of $T_i^{\text{QLR}}(g_i^0)$ is described by the distribution function $F_{\text{QLR}, \Omega_i(g_i^0)}$ (Rosen 2008).

As an additional assumption, Theorem 3 imposes independence between units. We use cross-sectional independence to verify the conditions of a high-dimensional central limit theorem and to prove an anti-concentration inequality. In the first instance, cross-sectional independence can be relaxed to allow for some correlation between units at the expense of more restrictive moment conditions. In the second instance, cross-sectional independence is an essential ingredient in our proof strategy. To prove an appropriate anti-concentration result, we exploit the fact that the limiting distribution of the unit-wise test statistic $\hat{T}_i^{\text{QLR}}(g_i^0)$ has a representation as a mixture of χ^2-random variables (Kudo 1963; Nüesch 1966; Wolak 1989; Rosen 2008). Under cross-sectional independence, we can use the marginal distributions of the unit-wise tests to derive an anti-concentration result for the joint test that tests all N units simultaneously. This argument cannot be extended to a setting without cross-sectional independence.

We also deviate from the assumptions of Theorem 2 by requiring a uniform lower bound on the smallest eigenvalue of Ω_i. This bound is needed to verify the assumptions of a high-dimensional central limit theorem (Chernozhukov, Chetverikov, and Kato 2016, Proposition 3.2).

5.5. Unit selection

In this section, we provide an asymptotic justification of our algorithm for unit selection. We show that applying unit selection to any of the three procedures described above generates valid confidence sets. The following theorem gives conditions under which the coverage probability of the confidence set after unit selection converges to the nominal level. The convergence is uniform over probability measures.

Theorem 4. *Let* $\widehat{C}^{\text{type}}_{\text{sel},\alpha,\beta}$ *denote a joint confidence set, where* $\text{type} = \text{SNS}, \text{MAX}$ *or* QLR. *Suppose that* $\{\hat{g}_i\}_{1\le i\le N}$ *satisfies* $\hat{D}_i^U(\hat{g}_i, h) \le 0$ *for any* $h \in \mathbb{G}$ *and* $i = 1, \dots, N$. *Let* $\mathbb{P}_N$ *denote a sequence of classes of probability measures that satisfy the conditions in Theorem 1 if* $\text{type} = \text{SNS}$, *Theorem 2 if* $\text{type} = \text{MAX}$ *and Theorem 3 if* $\text{type} = \text{QLR}$. *In addition, suppose that*

$$\max_{P\in\mathbb{P}_N} T^{-5/36} D_{N,T,3}\sqrt{\log(N/\beta)} \le 1, \tag{8}$$

$$\max_{P\in\mathbb{P}_N} T^{-5/24} B_{N,T,4}\log(N/\beta) \le 1, \tag{9}$$

$$\max_{P\in\mathbb{P}_N} T^{2/3}\gamma_{N,T,8}\left(T^{-5/24}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right)\sqrt{\log(N/\beta)} \le 1, \tag{10}$$

$$\begin{aligned}\max_{P\in\mathbb{P}_N} & T^{1/6}\gamma^2_{N,T,8}(T^{-5/12}(\log N)B^4_{N,T,8} + D^2_{N,T,4}) \\ & \times\left(D_{N,T,1} + \sqrt{\log N} + T^{-1/4}B_{N,T,4}\log N\right)\sqrt{\log(N/\beta)} \le 1.\end{aligned} \tag{11}$$

Then, for each $0 < \alpha < 1$, *there is a constant* C *depending only on* α, G, K_β *and the sequence* ϵ_N, *defined in the theorem corresponding to the value of* type, *such that*

$$\sup_{P\in\mathbb{P}_N} P\left(\{g_i^0\}_{1\le i\le N} \in \widehat{C}^{\text{type}}_{\text{sel},\alpha,\beta}\right) \ge 1 - \alpha - C\epsilon_N - CT^{-1/6}.$$

The conditions assumed here are slightly stronger versions of the conditions required in the previous theorems. This is partly because we use an auxiliary test statistic based on moment inequalities that have not been mean-adjusted.

Although the proof strategy for Theorem 4 has been adapted from the literature on moment selection (cf. Chernozhukov, Chetverikov, and Kato 2014), details of the argument have to be modified to account for the fact that our test statistics are based on mean-adjusted moment inequalities.

For unit selection to work, it is key that our joint confidence set always includes the vector of estimated group memberships. This implies that for units whose group memberships are obvious, it suffices to control the probability that the true group membership is not the estimated one. Step 1 of our proof shows that $\hat{D}_i^U(\hat{g}_i, h) \le 0$ implies that this probability is asymptotically less than β.

The assumption $\hat{D}_i^U(\hat{g}_i, h) \le 0$ means that the estimator of group memberships is based on an empirical version of inequality (3). This assumption will be automatically satisfied for estimators based on the *kmeans* estimator such as the estimator in Bonhomme and Manresa (2015). For Theorem 1 through Theorem 3, the inclusion of estimated memberships is not required for the asymptotic validity of the confidence set and it does not matter how group memberships are estimated.

6. Monte Carlo simulations

In this section, we study the finite-sample behavior of our procedures in Monte Carlo simulations. We consider both homoscedastic and heteroscedastic designs. For all our designs, we simulate panels of $N = 50$ units that are observed over $T = 10, 20, 30, 40$ time periods. We assume that the group-specific parameters are observed and compute joint confidence sets with nominal coverage probability $1 - \alpha = 0.9$. All simulation results are based on 1000 replications.

6.1. Homoscedastic design with three groups

For our first design, we consider a model with group fixed effects and $G = 3$ groups. For unit $i = 1, \ldots, N$, the outcome in period t is given by

$$y_{it} = \alpha_{g_i^0,t} + u_{it}. \tag{12}$$

The group fixed effects $\{\alpha_{g,t}\}_{1\leq t\leq T}$ for the three groups are defined as follows. Let $\varphi_T(t) = -1/2 + 2\,|t - T/2|\,/T$. For $t = 1, \ldots, T$, $\alpha_{1,t} = 0$, $\alpha_{2,t} = \varphi_T(t) + 1$, $\alpha_{3,t} = \varphi_{T/2}(t \bmod \lceil T/2 \rceil) - 1$.[20] The time profile for the group fixed effects is plotted in Figure B.1 in the Appendix. Note that the groups can be ordered. The group fixed effect of group 2 is large in all time periods, and that of group 2 is small in all time periods. The group fixed effect of group 1 is straddled between the effects of the other two groups. This choice of group fixed effects can be viewed as a perturbation to a specification with three parallel group fixed effects.[21] All units are assigned to the same group $g^0 = 1, 2, 3$. Our specification induces strong correlation of the moment inequalities.[22]

The error terms u_{it} are *i.i.d.* draws from $\mathcal{N}(0, \sigma^2 T)$ for $\sigma = 0.25, 0.5$. Note that the variance of the error term is scaled in a way that keeps the difficulty of the classification problem constant as we increase the number of observed time periods. This makes our simulation results for different values of T informative about the accuracy of the asymptotic approximation in finite-samples.[23]

We simulate three joint confidence sets (SNS, MAX and QLR). The critical values for the QLR and MAX procedures are adjusted for short panels. For this homoscedastic design, we turn off unit selection ($\beta = 0$). Following Andrews and Barwick (2012), we set the parameter for regularizing $\widehat{\Omega}_i$ to $\epsilon = 0.012$.[24] The simulation results are summarized in Table 1, where we report simulated coverage probabilities and average cardinality of the marginal unit-wise confidence sets. For group assignments g^0 to the two "outer" groups (groups 2 and 3), the simulation results are almost identical. This is expected, since these two groups are symmetric by construction. Therefore, we

[20] $\lceil T/2 \rceil$ is the smallest integer larger than $T/2$.

[21] A specification with parallel group fixed effects induces perfectly correlated moment inequalities. This violates the assumptions under which we establish the validity of our procedures. Our perturbation is calibrated in a way that ensures that our Monte Carlo results do not reflect our particular choice for how we regularize $\widehat{\Omega}_i(g)$.

[22] For example, for $T = 40$ and $g^0 = 1$, our simulations indicate that $(\mathbb{E}\,\widehat{\Omega}_i(1))_{1,2} = -0.93$ and $(\mathbb{E}\,\widehat{\Omega}_i(2))_{1,2} = 0.98$. For $T = 40$ and $g^0 = 2$, $(\mathbb{E}\,\widehat{\Omega}_i(1))_{1,2} = -0.90$ and $(\mathbb{E}\,\widehat{\Omega}_i(2))_{1,2} = 0.98$.

[23] Note that without rescaling the variance of the error term, increasing T eventually renders the classification problem trivial. For large T, all our procedures report a confidence set that includes only the true group memberships.

[24] The results are robust to different choices of ϵ.

			empirical coverage			cardinality of CS		
g^0	σ	T	SNS	MAX	QLR	SNS	MAX	QLR
1	0.25	10	0.96	0.96	0.96	2.40	2.21	2.09
1	0.25	20	0.92	0.93	0.95	1.74	1.59	1.53
1	0.25	30	0.92	0.91	0.95	1.54	1.42	1.39
1	0.25	40	0.92	0.92	0.94	1.45	1.35	1.33
1	0.50	10	0.94	0.93	0.93	2.91	2.87	2.84
1	0.50	20	0.92	0.93	0.92	2.82	2.75	2.73
1	0.50	30	0.90	0.92	0.93	2.77	2.70	2.68
1	0.50	40	0.92	0.92	0.94	2.75	2.67	2.65
2	0.25	10	0.97	0.95	0.93	1.84	1.81	1.85
2	0.25	20	0.96	0.93	0.90	1.42	1.41	1.51
2	0.25	30	0.94	0.92	0.92	1.30	1.30	1.39
2	0.25	40	0.96	0.91	0.92	1.25	1.25	1.33
2	0.50	10	0.95	0.92	0.89	2.63	2.53	2.47
2	0.50	20	0.95	0.92	0.91	2.28	2.20	2.20
2	0.50	30	0.95	0.91	0.91	2.17	2.11	2.13
2	0.50	40	0.95	0.92	0.90	2.12	2.07	2.10
3	0.25	10	0.97	0.95	0.94	1.84	1.81	1.85
3	0.25	20	0.96	0.91	0.92	1.42	1.42	1.51
3	0.25	30	0.94	0.91	0.91	1.30	1.30	1.38
3	0.25	40	0.95	0.92	0.90	1.25	1.25	1.32
3	0.50	10	0.97	0.93	0.91	2.62	2.53	2.47
3	0.50	20	0.95	0.92	0.90	2.28	2.20	2.20
3	0.50	30	0.94	0.90	0.89	2.17	2.11	2.12
3	0.50	40	0.94	0.91	0.90	2.12	2.07	2.09

Table 1: Homoscedastic design with $G = 3$ groups. Results based on 1000 simulated joint confidence sets with $1 - \alpha = 0.9$. Critical values for MAX and QLR procedures are adjusted for short panels. "Empirical coverage" gives the simulated coverage probability of the joint confidence set. "Cardinality of CS" gives the simulated expected average cardinality of a marginal (unit-wise) confidence set.

only discuss results for $g^0 = 1, 2$.

In all simulated designs, all three procedures construct valid confidence sets, with the empirical coverage probability close to or exceeding the nominal coverage probability. Since the SNS procedure does not explicitly take into account the within-unit correlation of the moment inequalities, the SNS critical value is an upper bound to the MAX bootstrap critical value. Therefore, the SNS procedure always yields a more conservative confidence set than the MAX procedure. This is confirmed numerically in the simulations.

For $g^0 = 1$, the QLR procedure provides narrower confidence sets than the MAX procedure, despite also being more conservative. For $g^0 = 2$, the result is reversed. The MAX procedure is more powerful than the QLR procedure, despite also being more conservative. This comparison illustrates that neither of our two test statistics dominates the other.

We also simulate the QLR and MAX confidence sets without short-panel adjustment. The simulation results are given in Table B.1 in the Appendix. As expected, without short-panel adjustment the confidence set is substantially undersized in short panels. As T increases, the empirical coverage probability of the confidence set monotonically converges to the nominal level, confirming our asymptotic results. For $T = 40$ the empirical coverage is within a 5% rage of the nominal level. Since the exact rate of convergence is design dependent, we recommend always using critical values with short-panel adjustment.

Our design induces highly correlated moments. In the Supplementary Appendix, we report simulation evidence for an alternative design in which the moment inequalities are not as strongly correlated. Our procedures perform well in this alternative design.

6.2. Heteroscedastic design with two groups

We now study the finite-sample properties of our algorithm for unit selection. To make unit selection meaningful we introduce heteroscedasticity.

Again, outcomes are generated from the linear model with group fixed effects (12). There are $G = 2$ groups with time-constant group fixed effects. For all $t = 1, \ldots, T$, the group fixed effects are given by $\alpha_{1,t} = 0.5$ and $\alpha_{2,t} = -0.5$. We only simulate units with $g_i^0 = 1$. Due to the symmetry of the design this is without loss of generality.

There are two "types" of units that face different degrees of statistical noise. For the "high noise" type the error term u_{it} is an *i.i.d.* draw from $\mathcal{N}(0, \sigma^2 T)$, where $\sigma = 0.25, 0.5$. For the "low noise" type, u_{it} is an *i.i.d.* draw from $\mathcal{N}(0, (\sigma/5)^2 T)$. The type of a unit is randomized independently of everything else. Unit i is assigned to the "high noise" type with either probability 0.5 (1:1 type ratio) or with probability 0.25 (1:3 type ratio).

We only simulate SNS confidence sets. QLR and MAX with short panel adjustment give numerically identical confidence sets when $G = 2$ (see Section 4.4). We set either $\beta = 0$ (no unit selection) or $\beta = 0.01$ (unit selection).

The simulation results are reported in Table 2. In the designs with $\sigma = 0.25$, the unit selection algorithm identifies units of the "low noise" type as easy to classify and ignores them when computing

			no unit selection		with unit selection		
σ	type ratio	T	coverage	power	coverage	$\hat{N}/N$	power
0.25	1:1	10	0.95	0.59	0.95	0.52	0.67
0.25	1:1	20	0.95	0.75	0.94	0.51	0.81
0.25	1:1	30	0.95	0.80	0.92	0.51	0.85
0.25	1:1	40	0.95	0.82	0.94	0.51	0.87
0.25	1:3	10	0.98	0.59	0.95	0.28	0.78
0.25	1:3	20	0.96	0.76	0.93	0.26	0.89
0.25	1:3	30	0.97	0.80	0.92	0.26	0.90
0.25	1:3	40	0.98	0.82	0.93	0.26	0.92
0.50	1:1	10	0.96	0.10	0.96	0.90	0.09
0.50	1:1	20	0.94	0.14	0.94	0.94	0.13
0.50	1:1	30	0.95	0.15	0.97	0.96	0.14
0.50	1:1	40	0.94	0.17	0.96	0.97	0.15
0.50	1:3	10	0.97	0.10	0.97	0.85	0.09
0.50	1:3	20	0.97	0.14	0.97	0.92	0.13
0.50	1:3	30	0.98	0.15	0.98	0.94	0.14
0.50	1:3	40	0.98	0.16	0.98	0.95	0.15

Table 2: Heteroscedastic design with two groups. Results based on 1000 simulated joint confidence sets (SNS) with $1 - \alpha = 0.9$. "Coverage" gives the simulated coverage probability of the joint confidence set. "Power" gives the simulated probability of reporting a singleton marginal (unit-wise) confidence set for the "high noise" type. $\hat{N}/N$ gives the simulated expected proportion of selected units.

the Bonferroni adjustment of the critical values. Relative to the case of no unit selection, this lowers the critical values for units of the "high noise" type. Consequently, the unit-wise confidence sets for "high noise" units become more powerful and a higher proportion of singletons is reported. This effect is more pronounced in the setting with a higher proportion of "low noise" units (1:3 type ratio).

In the designs with $\sigma = 0.50$, the unit selection algorithm identifies only a small proportion of the "low noise" types as easy to classify. Relative to the case of no unit selection, the unit-wise confidence sets for the "high noise" units become *less* powerful and a *smaller* proportion of singletons is reported.

This illustrates an important trade-off in employing unit selection. It improves the power of the joint confidence set if many units are deleted, but may reduce the power if an insufficient number of units are deleted. To see why this is the case, note that unit selection affects critical values in two ways. First, it allows us to do less Bonferroni correction, which lowers the critical values. Second, it changes the nominal level of the computed confidence set from α to $\alpha - 2\beta$, which increases the critical values. Unit selection is beneficial if the first effect dominates the second effect.

7. Applications

We apply the proposed confidence sets to two empirical applications. The first studies the effect of a minimum wage, and the second studies heterogeneous trajectories of democratization.

7.1. Minimum wage and unemployment

The first application studies heterogeneity in the effect of a minimum wage on unemployment. We examine panel data of states in the US and cluster them into two groups. The effect of a minimum wage is positive in one group and negative in the other. Our confidence sets quantify the uncertainty from using a data-driven method to sort states into one of the two groups.

To estimate the group-specific effects, we replicate results from Wang, Phillips, and Su (2016). Using US panel data, they follow an approach pioneered by Neumark and Wascher (1992) and identify the effect of a minimum wage from cross-state variation. Recently, Dube, Lester, and Reich (2010) argued that the way that a local economy reacts to a minimum wage may be affected by unobserved spatial heterogeneity. Wang, Phillips, and Su (2016) address this concern by proposing a linear panel model with a group structure. They estimate the following model for state i in time period t

$$\mathtt{ue}_{it} = \beta_{g_i^0,1}\mathtt{ue}_{i(t-1)} + \beta_{g_i^0,2}\mathtt{gr}_{i(t-1)} + \beta_{g_i^0,3}\mathtt{mw}_{i(t-1)} + \mu_i + u_{it},$$

where $\mathtt{ue}_{it}$ is the unemployment rate, $\mathtt{gr}_{it}$ is the growth rate of GDP, $\mathtt{mw}_{it}$ is the real state minimum wage, μ_i is a state fixed effect and u_{it} is an error term. The coefficients that describe the linear relationship may depend on the latent group membership of state i. We estimate the grouped panel model and compute unit-wise and joint confidence sets for group membership. The presence of the individual fixed effect μ_i renders this regression model different from our canonical model (1). In Section C of the Supplementary Appendix, we explain how to apply our methods to a linear panel data model after individual fixed effects have been differenced out.

We obtain all data from the online portal of the Federal Reserve Bank of St Louis.[25] We use yearly data for all 50 states ($N = 50$) from the period 1988 to 2014 ($T = 26$). For states in which state law does not specify a minimum wage, we use instead the federally mandated minimum wage. The data is standardized so that the time series for each state have standard deviation one.

Our estimation strategy is different from that employed in Wang, Phillips, and Su (2016), but our estimates are very similar.[26] We use the CLasso estimator from Su, Shi, and Phillips (2016) to estimate the group structure. Then, we estimate the group-specific parameters by post-Lasso least squares and perform a bias correction by half-panel Jackknifing (Dhaene and Jochmans 2015).

We detect $G = 2$ groups with 26 and 24 members, respectively. Like Wang, Phillips, and Su (2016), we find that one group has a positive coefficient on the lagged minimum wage ("positive effect group"), whereas the other has a negative coefficient ("negative effect group"). The estimated coefficients are reported in Table B.2. The map in Figure 1 depicts the estimated group memberships. Significant estimates are indicated by bold colors, states with insignificant estimates are lightly

[25] The GDP data is from the US Bureau of Economic Analysis, the minimum wage and unemployment data is from the US Department of Labor, and the CPI data is from the OECD Main Economic Indicators table.

[26] Their procedure includes a post-processing step using a hierarchical clustering algorithm. The results of the procedure are sensitive to the choice of the regularization parameter that controls the intensity of the post-processing step. We choose an alternative estimation procedure for which this post-processing step is not needed.

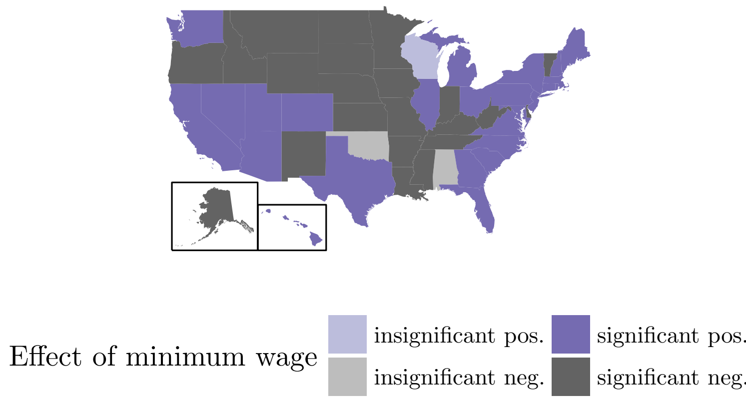

Figure 1: Estimated group memberships. The significance of the estimated group membership is based on unit-wise confidence sets at level $1 - \alpha = 0.9$.

shaded. Significance is based on unit-wise confidence sets at level $1 - \alpha = 0.9$.[27] A state's estimated group membership is significant if its unit-wise confidence set is a singleton.

The job of a policy maker who considers adjusting a minimum wage rate is complicated by heterogeneous responses to a minimum wage. An ideal policy decreases the local minimum wage in states where it has a positive effect on unemployment and increases it in states where is has a negative effect. A naïve way of assigning states to one of the two treatment groups is by estimated group membership. However, because of sampling error we may estimate the group membership of some states incorrectly and select them into the "wrong" treatment group. The naïve way of assigning states to treatments does not control this kind of misclassification. Based on our joint confidence set, we can implement an alternative procedure that controls the probability of misclassification.

We first compute a joint confidence set at confidence level $1 - \alpha$. For an illustrative calculation, we choose $\alpha = 0.33$ and compute a "1-sigma" joint confidence set.[28] In the case of two groups, all our procedures yield numerically identical results. As outlined in Section C, we remove the fixed effect by a standard fixed effect transformation and correct for bias using the half-panel Jackknife from Dhaene and Jochmans (2015). We do not employ unit selection ($\beta = 0$). The realized joint confidence set is depicted in Figure 2.[29]

To select the states for the "lower the minimum wage" treatment we select all states for which the marginal confidence set contains only the "positive effect" group. This strategy avoids uncontrolled misclassification, as these states have been found to be *jointly* significantly different from those in the "negative effect" group. Conversely, we select all states for which the marginal confidence set contains only the "negative effect" group into the "increase the minimum wage" treatment.

[27] We use critical values with a short-panel adjustment so that, in this setting with $G = 2$ groups, all our procedures compute the same confidence sets.

[28] In contrast to the pointwise case, there does not seem to be an established confidence level for joint inference over a large set. We adopt a 1-sigma confidence set ($1 - \alpha = 0.66$). This level is often used for confidence bands for impulse response functions in time-series models (cf. Stock and Watson 2001).

[29] See also Table B.3 in the Appendix.

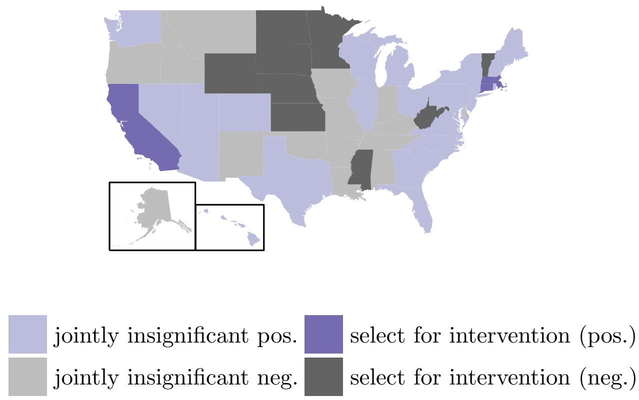

Figure 2: Joint confidence set at level $1 - \alpha = 0.66$. States with jointly significant estimated group memberships are selected into a treatment group based on the sign of their estimated effect of a minimum wage.

States for which the marginal confidence set is not a singleton are left untreated. This procedure guarantees that the probability of assigning one or several states to the wrong treatment is at most α. Based on the realized joint confidence set, we select 3 states into the "decrease the minimum wage" treatment and 9 states into the "raise the minimum wage" treatment.[30]

7.2. Paths to democracy

Our second application addresses the classification of countries based on heterogeneous trajectories of democratization. We build on the group fixed effects model proposed in Bonhomme and Manresa (2015).

Acemoglu et al. (2008) use country panel data to estimate the relationship between income (measured by GDP per capita) and democracy (measured by the Freedom House democracy index). Bonhomme and Manresa (2015) expand on this seminal study and estimate an augmented specification with group fixed effects. For country i and time period t they estimate the model

$$\texttt{democracy}_{it} = \theta_1 \texttt{democracy}_{i(t-1)} + \theta_2 \log(\texttt{gdp\_pc}_{i(t-1)}) + \alpha_{g_i^0,t} + u_{it},$$

where $\texttt{democracy}_{it}$ is the level of democracy measured by the Freedom House indicator, $\texttt{gdp\_pc}$ is GDP per capita and u_{it} is an error term. The inclusion of the group fixed effect $\alpha_{g,t}$ lends credibility to the exogeneity assumption of the linear panel model. In particular, the group fixed effect can pick up exogenous events, such as the process of decolonization, that unfold over time and impact both democratization and income growth.

We use the replication data set provided by Bonhomme and Manresa (2015). It is based on the balanced subsample from Acemoglu et al. (2008) and contains observations for $N = 90$ countries.

[30] The "decrease the minimum wage" treatment group consists of California, Connecticut and Massachusetts. The "raise the minimum wage" treatment group consists of Kansas, Minnesota, Mississippi, North Dakota, Nebraska, South Dakota, Vermont, West Virginia and Wyoming.

critical value	$\lvert\widehat{C}_{\alpha,i}\rvert = 1$	$\lvert\widehat{C}_{\alpha,i}\rvert = 2$	$\lvert\widehat{C}_{\alpha,i}\rvert = 3$	$\lvert\widehat{C}_{\alpha,i}\rvert = 4$
SNS	0	0	44	46
MAX	0	0	48	42
QLR	0	2	52	36

Table 3: Cardinality of the marginal unit-wise confidence sets for a joint confidence set at level $1 - \alpha = 0.66$.

Each country is observed every five years over the period 1970 – 2000 ($T = 7$). Details on the estimation procedure and estimates can be found in Bonhomme and Manresa (2015). Here, we focus on the pattern of grouped heterogeneity.

Bonhomme and Manresa (2015) detect $G = 4$ groups. Estimated time profiles for the group fixed effects are plotted in Figure B.2. There are two groups for which the fixed effect is almost constant over time, one with a low constant value and one with a high constant value. These are called the "low democracy" and "high democracy" groups, respectively. Then, there are two transitioning groups for which the group fixed effect starts out at about the level of the "low democracy" group, and then transitions to roughly the level of the "high democracy" group. There is an early transitioning group for which the transition starts in 1975, and a late transitioning group for which the transition starts in 1990.

As we demonstrate below, our procedures compute large, yet informative confidence sets. Based on these confidence sets, we can, for example, reject the hypotheses that all countries are "low democracy" countries or that all countries are "high democracy" countries.

We compute 1-sigma joint confidence sets, i.e., $1 - \alpha = 0.66$, based on the SNS, MAX and QLR procedures without unit selection. For the MAX and QLR procedures, we use short-panel-adjusted critical values. The cardinality of the marginal unit-wise confidence sets is reported in Table 3. All procedures generate an informative confidence set that rules out some group membership for some countries. For the MAX test statistic, taking the within-unit correlation of moment inequalities into account yields substantial power gains. With bootstrap critical values, the confidence set is uninformative about the group membership of only 36 countries, compared to 46 countries for the confidence set with the SNS critical value.

In the following we focus on the MAX joint confidence set.[31] To explore the computed confidence set further, we focus on the units that are estimated to be either "low democracy" or "high democracy" countries. These constitute 59 out of a total of 90 observed units. For the "low democracy" countries, we check whether their marginal confidence set contains the "high democracy" group. This divides the "low democracy" countries into a set of countries that is statistically separated from the group at the opposite side of the political spectrum, and a set of countries that is not. Vice versa, we check which "high democracy" countries we can rule out as members of the "low democracy" group. This characterization of the joint confidence set is reported in Table 4. For both groups, a vast majority of their estimated member countries are statistically different from the other group.

[31] We observe a moderate degree of regularization of the estimated correlation matrix. This may potentially affect the performance of the QLR statistic, whereas the MAX statistic is the more robust option.

low democracy		
Algeria(*)	Burundi(*)	Cameroon(*)
Chad(*)	China(*)	Congo, Rep.
Cote d'Ivoire(*)	Dem. Rep. Congo(*)	Egypt, Arab Rep.(*)
Gabon(*)	Guinea(*)	Indonesia
Iran(*)	Jordan(*)	Kenya(*)
Mauritania(*)	Morocco(*)	Nigeria
Paraguay(*)	Rwanda(*)	Sierra Leone
Singapore(*)	Syrian Arab Republic(*)	Togo(*)
Tunisia(*)	Uganda	
high democracy		
Australia(*)	Austria(*)	Belgium(*)
Canada(*)	Colombia	Costa Rica(*)
Cyprus	Denmark(*)	Dominican Republic
El Salvador	Finland(*)	France(*)
Guatemala	Iceland(*)	India(*)
Ireland(*)	Israel(*)	Italy(*)
Jamaica(*)	Japan(*)	Luxembourg(*)
Malaysia	Netherlands(*)	New Zealand(*)
Norway(*)	RB Venezuela(*)	Sri Lanka
Sweden(*)	Switzerland(*)	Trinidad and Tobago(*)
Turkey	United Kingdom(*)	United States(*)

Table 4: Estimated member countries for the "low democracy" and "high democracy" groups. The indicated significance of the estimated group assignments is based on a joint confidence set at level $1 - \alpha = 0.66$ (MAX procedure). Estimated "low democracy" countries with a (*) are not "high democracy" countries, and vice versa.

References

Acemoglu, Daron et al. (2008). "Income and democracy". In: *The American Economic Review* 98.3, pp. 808–842.

Allen, Roy (2017). "Testing moment inequalities: Selection versus recentering". In: *Economics Letters*. forthcoming.

Ando, Tomohiro and Jushan Bai (2016). "Panel data models with grouped factor structure under unknown group membership". In: *Journal of Applied Econometrics* 31.1, pp. 163–191.

Andrews, Donald W. K. and Panle Jia Barwick (2012). "Inference for parameters defined by moment inequalities: A recommended moment selection procedure". In: *Econometrica* 80.6, pp. 2805–2826.

Andrews, Donald WK and Gustavo Soares (2010). "Inference for parameters defined by moment inequalities using generalized moment selection". In: *Econometrica* 78.1, pp. 119–157.

Azzalini, Adelchi and Alan Genz (2016). *The R package `mnormt`: The multivariate normal and t distributions (version 1.5-5)*. URL: `http://azzalini.stat.unipd.it/SW/Pkg-mnormt`.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2016). "Discretizing unobserved heterogeneity: Approximate clustering methods for dimension reduction". Working paper.

Bonhomme, Stéphane and Elena Manresa (2015). "Grouped patterns of heterogeneity in panel data". In: *Econometrica* 83.3, pp. 1147–1184.

Bugni, Federico A (2010). "Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set". In: *Econometrica* 78.2, pp. 735–753.

Canay, Ivan and Azeem Shaikh (2016). “Practical and theoretical advances in inference for partially identified models”. Working paper.
Chen, Xiaohong et al. (2016). “Self-normalized Cramér-type moderate deviations under dependence”. In: *The Annals of Statistics* 44.4, pp. 1593–1617.
Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2014). “Testing many moment inequalities”. In: *arXiv preprint arXiv:1312.7614*.
— (2015). “Comparison and anti-concentration bounds for maxima of Gaussian random vectors”. In: *Probability Theory and Related Fields* 162.1-2, pp. 47–70.
— (2016). “Central limit theorems and bootstrap in high dimensions”. In: *arXiv preprint arXiv:1412.3661v4*.
de la Pena, Victor H., Tze Leung Lai, and Qi-Man Shao (2009). *Self-normalized processes. Probability and its Applications*. Springer-Verlag, Berlin (New York).
Dempster, Arthur, Nan Laird, and Donald Rubin (1977). “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.
Dhaene, Geert and Koen Jochmans (2015). “Split-panel jackknife estimation of fixed-effect models”. In: *The Review of Economic Studies* 82.3, pp. 991–1030.
Dube, Arindrajit, T. William Lester, and Michael Reich (2010). “Minimum wage effects across state borders: Estimates using contiguous counties”. In: *The Review of Economics and Statistics* 92.4, pp. 945–964.
Duembgen, Lutz (2010). “Bounding standard Gaussian tail probabilities”. In: *arXiv preprint arXiv:1012.2063*.
Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch (2013). *Modelling extremal events: for insurance and finance*. Vol. 33. Springer Science & Business Media.
Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2009). *The elements of statistical learning*. 2nd ed. Vol. 1. Springer.
Genz, Alan (1992). “Numerical computation of multivariate normal probabilities”. In: *Journal of Computational and Graphical Statistics* 1.2, pp. 141–149.
Grayling, Michael J and Adrian Mander (2016). “MVTNORM: Stata module to work with the multivariate normal and multivariate t distributions”. In: *Statistical Software Components*.
Gu, Jiaying and Stanislav Volgushev (2018). “Panel data quantile regression with grouped fixed effects”. Unpublished working paper.
Hahn, Jinyong and Guido Kuersteiner (2002). “Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large”. In: *Econometrica* 70.4, pp. 1639–1657.
Hahn, Jinyong and Hyungsik Roger Moon (2010). “Panel data models with finite number of multiple equilibria”. In: *Econometric Theory* 26.03, pp. 863–881.
Heckman, James and Burton Singer (1984). “A method for minimizing the impact of distributional assumptions in econometric models for duration data”. In: *Econometrica*, pp. 271–320.
Horn, Roger A and Charles R Johnson (2013). *Matrix analysis*. 2nd ed. Cambridge University Press.
Ipsen, Ilse and Rizwana Rehman (2008). “Perturbation bounds for determinants and characteristic polynomials”. In: *SIAM Journal on Matrix Analysis and Applications* 30.2, pp. 762–776.
Jing, Bing-Yi, Qi-Man Shao, and Qiying Wang (2003). “Self-normalized Cramér-type large deviations for independent random variables”. In: *The Annals of Probability* 31.4, pp. 2167–2215.
Kudo, Akio (1963). “A multivariate analogue of the one-sided test”. In: *Biometrika* 50.3/4, pp. 403–418.
Laurent, Beatrice and Pascal Massart (2000). “Adaptive estimation of a quadratic functional by model selection”. In: *Annals of Statistics*, pp. 1302–1338.

Lewis, Richard and Gregory C Reinsel (1985). "Prediction of multivariate time series by autoregressive model fitting". In: *Journal of multivariate analysis* 16.3, pp. 393–411.

Lin, Chang-Ching and Serena Ng (2012). "Estimation of panel data models with parameter heterogeneity when group membership is unknown". In: *Journal of Econometric Methods* 1.1, pp. 42–55.

Lu, Xun and Liangjun Su (2017). "Determining the number of groups in latent panel structures with an application to income and democracy". In: *Quantitative Economics* 8.3, pp. 729–760.

McLachlan, Geoffrey and David Peel (2004). *Finite mixture models*. John Wiley & Sons.

Menzel, Konrad (2014). "Consistent estimation with many moment inequalities". In: *Journal of Econometrics* 182.2, pp. 329–350.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.

Neumark, David and William Wascher (1992). "Employment effects of minimum and subminimum wages: panel data on state minimum wage laws". In: *ILR Review* 46.1, pp. 55–81.

Nüesch, Peter E (1966). "On the problem of testing location in multivariate populations for restricted alternatives". In: *The Annals of Mathematical Statistics* 37.1, pp. 113–119.

Pearson, Karl (1896). "Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia". In: *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 187, pp. 253–318.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf (2014). "A Practical Two-Step Method for Testing Moment Inequalities". In: *Econometrica* 82.5, pp. 1979–2002.

Romano, Joseph P and Michael Wolf (2018). "Multiple testing of one-sided hypotheses: combining Bonferroni and the bootstrap". In: *International Conference of the Thailand Econometrics Society*. Springer, pp. 78–94.

Rosen, Adam (2008). "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities". In: *Journal of Econometrics* 146.1, pp. 107–117.

Sarafidis, Vasilis and Neville Weber (2015). "A partially heterogeneous framework for analyzing panel data". In: *Oxford Bulletin of Economics and Statistics* 77.2, pp. 274–296.

Shao, Qing and Yuehua Wu (2005). "A consistent procedure for determining the number of clusters in regression clustering". In: *Journal of Statistical Planning and Inference* 135.2, pp. 461–476.

Soms, Andrew P (1976). "An asymptotic expansion for the tail area of the t-distribution". In: *Journal of the American Statistical Association* 71.355, pp. 728–730.

Stock, James H and Mark W Watson (2001). "Vector autoregressions". In: *The Journal of Economic Perspectives* 15.4, pp. 101–115.

Su, Liangjun, Zhentao Shi, and Peter Phillips (2016). "Identifying latent structures in panel data". In: *Econometrica* 84.6, pp. 2215–2264.

Vogt, Michael and Oliver Linton (2017). "Classification of nonparametric regression functions in heterogeneous panels". In: *Journal of the Royal Statistical Society: Series B* 79 (1), pp. 5–27.

Vogt, Michael and Matthias Schmid (2017). "Clustering with statistical error control". In: *arXiv preprint arXiv:1702.02643*.

Wainwright, Martin (2015). *Basic tail and concentration bounds 2*. URL: `https://www.stat.berkeley.edu/.../Chap2_TailBounds_Jan22_2015.pdf` (visited on 12/31/2017).

Wang, Wuyi, Peter Phillips, and Liangjun Su (2016). "Homogeneity pursuit in panel data models: theory and applications". Working paper.

Wolak, Frank A (1987). "An exact test for multiple inequality and equality constraints in the linear regression model". In: *Journal of the American Statistical Association* 82.399, pp. 782–793.

— (1989). "Testing inequality constraints in linear econometric models". In: *Journal of Econometrics* 41.2, pp. 205–235.

Wolak, Frank A (1991). "The local nature of hypothesis tests involving inequality constraints in nonlinear models". In: *Econometrica*, pp. 981–995.
Zhang, Xinyang and Guang Cheng (2017). "Guassian approximation for high dimensional vector under physical dependence". In: *Bernoulli*. forthcoming.
Zhilova, Mayya (2015). "Simultaneous likelihood-based bootstrap confidence sets for a large number of models". Working paper.

Appendix

A. Proofs of mains results

In the proofs, we drop the g argument for ease of notation and write, e.g., $d_{it}(h)$ instead of $d_{it}(g,h)$ (or $d_{it}(g_i^0,h)$). The g argument is made explicit in the statements of the lemmas. Here, we provide proofs of Theorem 1 – Theorem 3. All supporting lemmas and the proof of Theorem 4 are given in the Supplementary Appendix.

For our proof of the QLR procedure we analyze the limiting distribution of the QLR statistic, which we call the $\tilde{\chi}^2$-distribution. Let V denote a nonsingular covariance matrix, and let $X \sim \mathcal{N}(0,V)$. The $\tilde{\chi}^2(V)$ distribution is given by the distribution of the random variable

$$W = \min_{t \leq 0} (X-t)' V^{-1} (X-t).$$

The $\tilde{\chi}^2(V)$-distribution can be characterized as a mixture of χ^2-distributions (Rosen 2008) and is closely related to the $\bar{\chi}^2$-distribution (Kudo 1963, Nüesch 1966). Lemma D.13 in the Supplementary Appendix summarizes some properties of the $\tilde{\chi}^2$-distribution.

Proof of Theorem 1. We first evaluate the effect of estimation error from estimating the group-specific coefficients. Let C_1 denote the constant from Lemma D.8 and let $\zeta_{N,T}$ as defined in Lemma D.8. Let

$$\begin{aligned} a_{N,T} =& C_1 \sqrt{T} \gamma_{N,T,8} \left(T^{-5/24} B_{N,T,4} \sqrt{\log N} + D_{N,T,2} \right) \\ &+ C_1 \zeta_{N,T} \sqrt{\log N} \left(1 + T^{-1/4} B_{N,T,4} \sqrt{\log N} \right) \end{aligned}$$

Define the event

$$\mathcal{E}_{N,T,1} = \left\{ \max_{1 \leq i \leq N} \max_{h \in \mathbb{G} \setminus \{g_i^0\}} \left| \hat{D}_i(h) - \tilde{D}_i(h) \right| \leq a_{N,T} \right\}.$$

Applying Lemma D.8 with $c = 1/6$ yields

$$1 - P\left(\mathcal{E}_{N,T,1}\right) \leq N^{-1} + C_1 T^{-1/6} + C_1 \left(T^{-1/4} B_{N,T,4} / (\log N) \right)^4 \leq N^{-1} + C T^{-1/6}.$$

Note that under the assumptions of the lemma, $\zeta_{N,T} \leq 3\epsilon_{N,1}$. On $\mathcal{E}_{N,T,1}$, for $i = 1, \ldots, N$ and $h \in \mathbb{G} \setminus \{g_i^0\}$,

$$\left| \hat{D}_i(h) - \tilde{D}_i(h) \right| \leq C \left(\epsilon_{1,N} + \epsilon_{2,N} \right) / \sqrt{\log N} =: b_N.$$

Next, we discuss the contribution of the estimation error to the coverage level. Define α_N implicitly by

$$c^{\text{SNS}}_{\alpha_N,N} = c^{\text{SNS}}_{\alpha,N} - b_N.$$

To see that α_N is well-defined, note that since $c^{\text{SNS}}_{\alpha,N} \to \infty$ and $b_N \to 0$ the right-hand side of the equation is diverging, and therefore positive for large N. Moreover, $c^{\text{SNS}}_{p,N} \downarrow 0$ as $p \uparrow N/2$. This establishes the existence of α_N. Uniqueness follows from the strict monotonicity of the distribution function of the t-distribution. Let F_T denote the distribution function of a t-distributed random variable with $T-1$ degrees of freedom, and let f_T denote its density function. Let $c(\alpha) = t^{-1}_{T-1}(1-\alpha/((G-1)N))$ and $b^*_N = \sqrt{(T-1)/T}b_N$. By the mean-value theorem

$$\begin{aligned}\frac{\alpha_N}{(G-1)N} - \frac{\alpha}{(G-1)N} =& F_T(c(\alpha)) - F_T(c(\alpha_N)) \\ =& F_T(c(\alpha)) - F_T(c(\alpha) - b_N) = f_T(c^*)b^*_N,\end{aligned}$$

where c^* is a value between $c(\alpha_N)$ and $c(\alpha)$. Noting that $c(\alpha_N) < c(\alpha)$ and that f_T is decreasing on the positive axis, rearranging this equality yields

$$\begin{aligned}|\alpha_N - \alpha| \leq& f_T\left(c\left(\alpha_N\right)\right)(G-1)Nb^*_N \\ \leq& 2c\left(\alpha_N\right)\left(1 - F_T(c\left(\alpha_N\right)\right)(G-1)Nb^*_N \\ \leq& 4b^*_N\alpha_N\sqrt{\log\left((G-1)N/\alpha_N\right)} \\ \leq& 4b_N\alpha\sqrt{\log\left((G-1)N/\alpha\right)} + 4b_N\left|\alpha_N - \alpha\right|\sqrt{\log\left((G-1)N/\alpha\right)} \\ \leq& 4b_N\sqrt{\log\left((G-1)N/\alpha\right)} + o\left(\left|\alpha_N - \alpha\right|\right),\end{aligned}$$

where the second inequality follows from Lemma D.11, the third inequality follows from Lemma D.10 (with $\epsilon = 1$), and the fourth inequality follows from $b_N\sqrt{\log N} \to 0$. This recursion implies

$$|\alpha_N - \alpha| \leq 5b_N\sqrt{\log\left((G-1)N/\alpha\right)}$$

for N large enough.

We now derive an approximation based on the theory of self-normalized sums, i.e., Lemma D.12. Let $g_T : x \to x/\sqrt{1+x^2/T}$ and

$$\tilde{D}_{i,T,3}(h) = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_P\left|d_{it}(h)/(\sigma_i s_{i,T}(h))\right|^3\right)^{1/3}.$$

We apply Lemma D.12 with $\xi_t = d_{it}(h)/(\sigma_i s_{i,T}(h))$, $\nu = 1$, and $x = g_T(c^{\text{SNS}}_{\alpha_N,N})$. The lemma requires

$$g_T\left(c^{\text{SNS}}_{\alpha_N,N}\right) \leq T^{1/6}/\tilde{D}_{i,T,3}(h), \tag{13}$$

for N large enough, for all $i = 1, \ldots, N$ and $h \in \mathbb{G} \setminus \{g^0_i\}$. To prove this inequality, note that under Assumption 1 there is a constant C such that

$$\sup_{1\leq i\leq N}\sup_{h\in\mathbb{G}\setminus\{g^0_i\}}\tilde{D}_{i,T,3}(h)/D_{N,T,3} \leq C,$$

so that it is sufficient to show $T^{-1/6}c^{\text{SNS}}_{\alpha,N}D_{N,T,3} \to 0$. Setting $\epsilon = 1$ in Lemma D.10 gives

$$T^{-1/6}c^{\text{SNS}}_{\alpha,N}D_{N,T,3} \le \sqrt{T/(T-1)}T^{-1/6}2\sqrt{\log\left((G-1)N/\alpha\right)}D_{N,T,3}.$$

Under our assumptions the right-hand side vanishes and condition (13) is verified. Applying Lemma D.12 yields

$$\begin{aligned}
&\left|P\left(\tilde{D}_i(h) > c^{\text{SNS}}_{\alpha_N,N}\right) - \left(1-\Phi\left(g_T(c^{\text{SNS}}_{\alpha_N,N})\right)\right)\right| \\
=&\left|P\left(\frac{\sum_{t=1}^T d_{it}(h)/(\sigma_i s_{i,T}(h))}{\sqrt{\sum_{t=1}^T d_{it}^2(h)/(\sigma_i s_{i,T}(h))^2}} > g_T(c^{\text{SNS}}_{\alpha_N,N})\right) - \left(1-\Phi\left(g_T\left(c^{\text{SNS}}_{\alpha_N,N}\right)\right)\right)\right| \\
\le& KT^{-1/2}\tilde{D}^3_{i,T,3}\left(1+g_T(c^{\text{SNS}}_{\alpha_N,N})\right)^3\left(1-\Phi\left(g_T(c^{\text{SNS}}_{\alpha_N,N})\right)\right),
\end{aligned} \tag{14}$$

where K is the constant from Lemma D.12. For standard normal $d_{it}(h)$, we can take $\tilde{D}_{i,T,3}(h) = 2^{3/2}/\sqrt{\pi}$, and (13) is easily verified provided that $T^{-1/3}(\log N) \to 0$. As $D_{N,T,3} \ge 1$, the assumption $\epsilon_{3,N} \to 0$ requires $T^{-1/3}(\log N) \to 0$. Evaluating (14) for the special case of standard normal $d_{it}(h)$ gives

$$\begin{aligned}
&\left|\frac{\alpha_N}{(G-1)N} - \left(1-\Phi\left(g_T\left(c^{\text{SNS}}_{\alpha_N,N}\right)\right)\right)\right| \\
\le& KT^{-1/2}2^{9/2}\pi^{-3/2}\left(1+g_T(c^{\text{SNS}}_{\alpha_N,N})\right)^3\left(1-\Phi\left(g_T(c^{\text{SNS}}_{\alpha_N,N})\right)\right).
\end{aligned} \tag{15}$$

Under $T^{-1/3}(\log N) \to 0$, the right-hand side vanishes and therefore the recursive nature of the inequality implies $1-\Phi(g_T(c^{\text{SNS}}_{\alpha,N})) = \alpha_N/((G-1)N) + o(\alpha_N/((G-1)N))$. Combining inequalities (14) and (15) gives

$$\begin{aligned}
&\left|P\left(\tilde{D}_i(h) > c^{\text{SNS}}_{\alpha_N,N}\right) - \frac{\alpha_N}{(G-1)N}\right| \\
\le&\left|P\left(\tilde{D}_i(h) > c^{\text{SNS}}_{\alpha_N,N}\right) - \left(1-\Phi\left(g_T\left(c^{\text{SNS}}_{\alpha_N,N}\right)\right)\right)\right| \\
&+\left|\frac{\alpha_N}{(G-1)N} - \left(1-\Phi\left(g_T\left(c^{\text{SNS}}_{\alpha_N,N}\right)\right)\right)\right| \\
\le& KT^{-1/2}\left(\tilde{D}^3_{i,T,3} + 2^{9/2}\pi^{-3/2}\right)\left(1+g_T(c^{\text{SNS}}_{\alpha_N,N})\right)^3\left(1-\Phi\left(g_T(c^{\text{SNS}}_{\alpha_N,N})\right)\right). \\
\le& CT^{-1/2}D^3_{N,T,3}\left(1+g_T(c^{\text{SNS}}_{\alpha_N,N})\right)^3\left(1-\Phi\left(g_T(c^{\text{SNS}}_{\alpha_N,N})\right)\right) \\
\le& C\left(2T^{-1/6}D_{N,T,3}\sqrt{\log\left(\frac{(G-1)N}{\alpha}\right)}\right)^3\left(\frac{\alpha_N}{(G-1)N} + o\left(\frac{\alpha_N}{(G-1)N}\right)\right).
\end{aligned}$$

Summing up, we have

$$\begin{aligned}
&P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\hat{D}_i(h) > c^{\text{SNS}}_{\alpha,N}\right) \\
\le& P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\tilde{D}_i(h) > c^{\text{SNS}}_{\alpha,N} - b_N\right) + P\left(\mathcal{E}^c_{N,T,1}\right)
\end{aligned}$$

$$
\begin{aligned}
=&P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\tilde{D}_i(h)>c^{\mathrm{SNS}}_{\alpha_N,N}\right)+P\left(\mathcal{E}^c_{N,T,1}\right)\\
\le&\sum_{i=1}^N\sum_{h\in\mathbb{G}\setminus\{g_i^0\}}P\left(\tilde{D}_i(h)>c^{\mathrm{SNS}}_{\alpha_N,N}\right)+P\left(\mathcal{E}^c_{N,T,1}\right)\\
\le&\alpha_N+C\left(2T^{-1/6}D_{N,T,3}\sqrt{\log\left(\frac{(G-1)N}{\alpha}\right)}\right)^3+1-P\left(\mathcal{E}_{N,T,1}\right)\\
\le&\alpha+C\left(b_N\sqrt{\log N}+\epsilon_{3,N}+T^{-1/6}+N^{-1}\right).
\end{aligned}
$$

□

Proof of Theorem 2. Throughout the proof let C denote a generic constant depending only on G and K_β. Define the events

$$
\begin{aligned}
\mathcal{E}_{N,T,1}=&\left\{\max_{g\in\mathbb{G}}\left(\frac{1}{T}\sum_{t=1}^T\left\|\hat{\beta}_{g,t}-\beta_{g,t}\right\|^8\right)^{1/8}\le\gamma_{N,T,8}\right\},\\
\mathcal{E}_{N,T,2}=&\Big\{\max_{1\le i\le N}\max_{h,h'\in\mathbb{G}\setminus\{g_i^0\}}\left|(\hat{\Omega}_i)_{h,h'}-(\Omega_i)_{h,h'}\right|\le C_1(3\epsilon_{1,N}+\epsilon_{3,N})\Big\},
\end{aligned}
$$

where C_1 is the maximum of the constants from Lemma D.7 and Lemma D.8. By Assumption 1((iii)), $P(\mathcal{E}_{N,T,1})\ge 1-\epsilon_N$. Lemma D.7 and Lemma D.8 imply $P(\mathcal{E}_{N,T,2})\ge 1-CT^{-1/7}\ge 1-C\epsilon_N$. To see this, let Ω_i^* as defined in Lemma D.7 and decompose

$$
\left|(\hat{\Omega}_i)_{h,h'}-(\Omega_i)_{h,h'}\right|\le\left|(\hat{\Omega}_i)_{h,h'}-(\Omega_i^*)_{h,h'}\right|+\left|(\Omega_i^*)_{h,h'}-(\Omega_i)_{h,h'}\right|.
$$

The first term on the right-hand side is bounded by $C_1\zeta_{N,T}$ with probability more than $1-CT^{-1/7}$, where $\zeta_{N,T}$ is defined in Lemma D.8. This can be shown by applying Lemma D.8 with $c=1/7$Under the assumptions of the theorem we have

$$
\zeta_{N,T}\le\epsilon_{1,N}/(\log N)(1+\epsilon_{3,N})+(\epsilon_{1,N}/\log N)^2\le 3\epsilon_{1,N}/\log N.
$$

For $c=1/7$, Lemma D.7 controls the rate of $\left|(\Omega_i^*)_{h,h'}-(\Omega_i)_{h,h'}\right|$ and gives the upper bound

$$
C_1T^{-3/7}B^2_{N,T,4}(\log N)=C_1T^{-1/7}\epsilon^2_{3,N}/\log N\le C_1\epsilon_{3,N}/\log N.
$$

On $\mathcal{E}_{N,T,1}\cap\mathcal{E}_{N,T,2}$

$$
\|\hat{\Omega}_i-\Omega_i\|_2\le\|\hat{\Omega}_i-\Omega_i\|_F=\sqrt{\sum_{h,h'}\left|(\hat{\Omega}_i)_{h,h'}-(\Omega)_{h,h'}\right|^2}\le C(\epsilon_{1,N}+\epsilon_{3,N})/\log N.
$$

Since Ω_i is a correlation matrix we have $\|\Omega_i\|_2\le\operatorname{tr}(\Omega_i)\le G-1$ and therefore

$$
\begin{aligned}
&\|\Omega_i^{-1}\|_2\big(1\vee\|\Omega_i\|_2\|\Omega_i^{-1}\|_2\big)\|\hat{\Omega}_i^{-1}-\Omega_i^{-1}\|_2\\
\le&C\omega_N(1\vee\omega_N(G-1))(\epsilon_{1,N}+\epsilon_{2,N})/\log N\\
\le&C\|\hat{\Omega}_i-\Omega_i\|_2\big(\omega_N^2\vee 1\big)\le C_1^*\big(\omega_N^2\vee 1\big)\left(\epsilon_{1,N}+\epsilon_{3,N}\right)/\log N
\end{aligned}
$$

where C_1^* depends only on G and K_β. Lemma D.8 with $c = 1/7$ gives a lower bound on the probability of the set on which

$$\begin{aligned}\left|\hat{D}_i(h) - D_i(h)\right| \leq & C_1\Big[\sqrt{T}\gamma_{N,T,8}\left(\epsilon_{3,N} + D_{N,T,2}\right) \\ & + \left(\zeta_{N,T} + \left(\epsilon_{3,N}T^{-1/7}/\sqrt{\log N}\right)^2\right)(1+\epsilon_{3,N})\sqrt{\log N}\Big] \\ \leq & C_2^*\left(\epsilon_{1,N} + \epsilon_{2,N} + \epsilon_{3,N}\right)/\sqrt{\log N},\end{aligned}$$

where C_2^* depends only on G and K_β. Define the event

$$\mathcal{E}_{N,T,3} = \Big\{\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\left|\hat{D}_i(h) - D_i(h)\right| \leq C_2^*\left(\epsilon_{1,N} + \epsilon_{2,N} + \epsilon_{3,N}\right)/\sqrt{\log N}\Big\}.$$

By Lemma D.8,

$$P(\mathcal{E}_{N,T,3}) \geq 1 - N^{-1} - C_1\left(T^{-1/7} + \left(T^{-(1/4-1/7)}\epsilon_{N,3}\right)^4\right) \geq 1 - N^{-1} - CT^{-1/7}.$$

By Lemma D.3, there are random variables $(X_i)_{1\leq i\leq N}$ with $X_i \sim \mathcal{N}(0, \Omega_i)$ such that

$$\begin{aligned}&\sup_{(r_1,\dots,r_N)\in\mathcal{R}_{++}^N}\left|P\left(\max_{1\leq i\leq N}\left(T_i^{\text{MAX}} - r_i\right) > 0\right) - P\left(\max_{1\leq i\leq N}\left(\max_{1\leq h\leq G-1}X_{i,h} - r_i\right) > 0\right)\right| \\ \leq & C\left(T^{-1/6}B_{N,T,4}\log^{7/6}N + T^{-1/6}B_{N,T,4}^2\log N\right) \\ \leq & C\left(\epsilon_{3,N}\left(T^{-1/7}(\log N)\right)^6 + \epsilon_{3,N}^2\right)\log N \leq C_3^*\epsilon_{3,N}(\log N),\end{aligned}$$

where C_3^* depends only on G and K_β. To avoid ambiguity, denote the quantities in the statement of Lemma D.1 with a † superscript. The conclusion of the theorem follows from applying Lemma D.1 with $\epsilon_N^\dagger = (C_1^* \vee C_2^* \vee C_3^*)\epsilon_N/(\log N)$, $\hat{D}_i^\dagger = \hat{D}_i$, $D_i^\dagger = D_i$, $\hat{\Omega}_i^\dagger = \hat{\Omega}_i$ and $\Omega_i^\dagger = \Omega_i$ on the event $\mathcal{E}_{N,T,1} \cap \mathcal{E}_{N,T,2} \cap \mathcal{E}_{N,T,3}$. □

Proof of Theorem 3. Throughout the proof, let C denote a generic constant depending only on G and K_β. Let $\hat{V}_i$ denote the diagonal matrix with entries $(\hat{V}_i)_{h,h} = \hat{S}_{i,T}(h)/(\sigma_i s_{i,T}(h))$ and let

$$\begin{aligned}\Delta_i^D(h) = & \frac{1}{\sqrt{T}}\sum_{t=1}^T\frac{\hat{d}_{it}(h) - d_{it}(h)}{\sigma_i s_{i,T}(h)}, \\ \hat{\Omega}_i^V = & \hat{V}_i(g)\hat{\Omega}_i(g)\hat{V}_i(g),\end{aligned}$$

$\Delta_i^D = (\Delta_i^D(h))_{h\in\mathbb{G}\setminus\{g_i^0\}}$. Using these definitions, we may rewrite the unit-specific test statistics in a way that eliminates all random denominators

$$\hat{T}_i^{\text{QLR}} = \min_{t\leq 0}\left(\Delta_i^D + D_i - t\right)'\left[\hat{\Omega}_i^V\right]^{-1}\left(\Delta_i^D + D_i - t\right).$$

Define the events $\mathcal{E}_{N,T,1}$, $\mathcal{E}_{N,T,2}$ and $\mathcal{E}_{N,T,3}$ as in the proof of Theorem 2. Recall that on $\bigcap_{\ell=1}^3\mathcal{E}_{N,T,\ell}$ we have

$$\|\hat{\Omega}_i - \Omega_i\|_2 \leq C\epsilon_N/\log N$$

$$\|\hat{D}_i - D_i\| \leq C\epsilon_N/\sqrt{\log N}$$

and $P\left(\bigcap_{\ell=1}^{3} \mathcal{E}_{N,T,\ell}\right) \geq 1 - C\epsilon_N$. Work conditional on $\bigcap_{\ell=1}^{3} \mathcal{E}_{N,T,\ell}$. By the inequality $|\sqrt{a}-1| \leq |a-1|$,

$$\begin{aligned}
\left|(\hat{V}_i)_{h,h} - 1\right| =& \left|\sqrt{\hat{S}^2_{N,T}(g_i^0,h)/(\sigma_i^2 s^2_{i,T}(g_i^0,h))} - 1\right| \\
\leq& \left|\hat{S}^2_{N,T}(g_i^0,h)/(\sigma_i^2 s^2_{i,T}(g_i^0,h)) - 1\right| \\
\leq& \left|(\hat{\Omega}_i)_{h,h} - (\Omega)_{h,h}\right| \leq \|\hat{\Omega}_i - \Omega_i\|_2 \leq C\epsilon_N/\log N.
\end{aligned}$$

and therefore, $\|\hat{V}_i - I_{G-1}\|_2 \leq C\epsilon_N/\log N$, where I_{G-1} is the $(G-1)$ dimensional identity matrix. Write $V_i = I_{G-1}$ and decompose

$$\begin{aligned}
\hat{\Omega}_i^V - \Omega_i =& (\hat{V}_i - V_i)(\hat{\Omega}_i - \Omega_i)(\hat{V}_i - V_i) + 2V_i(\hat{\Omega}_i - \Omega_i)(\hat{V}_i - V_i) \\
&+ V_i(\hat{\Omega}_i - \Omega_i)V_i + (\hat{V}_i - V_i)\Omega_i(\hat{V}_i - V_i) + 2V\Omega_i(\hat{V}_i - V_i).
\end{aligned}$$

Noting that $\|\Omega_i\| \leq \mathrm{tr}(\Omega_i) \leq G-1$, this decomposition implies

$$\|\hat{\Omega}_i^V - \Omega_i\|_2 \leq C\left(\epsilon_N^3 + (2 + \|\Omega_i\|_2)\epsilon_N^2 + (1 + 2\|\Omega_i\|_2)\epsilon_N\right) \leq C\epsilon_N,$$

where the second-to-last inequality follows from $\epsilon_N \leq 1$ for N large enough. Therefore,

$$\begin{aligned}
&(1 \vee \|\Omega_i^{-1}\|)(\|\Omega_i\| \vee \|\Omega_i^{-1}\|)\|\hat{\Omega}_i^V - \Omega_i\|_2 \\
\leq& C(1 \vee \lambda_1^{-1})(\lambda_1^{-1} \vee (G-1))\epsilon_N/\log N \leq C_1^*\epsilon_N/\log N,
\end{aligned}$$

where C_1^* depends only on λ_1, G and K_β. Define the event

$$\mathcal{E}_{N,T,4} = \max_{1\leq i\leq N}\left\{\|D_i\| \leq 2C_1\sqrt{\log N}\right\}.$$

Taking N large enough that $\epsilon_{3,N} \leq 1$, Lemma D.7 with $c = 1/7$ yields

$$\begin{aligned}
1 - P\left(\mathcal{E}_{N,T,4}\right) \leq& \sum_{\ell=1}^{G-1} P\left(\max_{1\leq i\leq N}\|D_{i,\ell}\| > 2C_1\sqrt{\log N}\right) \\
\leq& \sum_{\ell=1}^{G-1} P\left(\max_{1\leq i\leq N}\|D_{i,\ell}\| > C_1\sqrt{\log N}\left(1 + T^{-1/4}B_{N,T,4}\sqrt{\log N}\right)\right) \\
\leq& (G-1)\left(N^{-1} + \left(T^{-(1/4-1/7)}\epsilon_{3,N}\right)^4\right) \leq C\epsilon_N,
\end{aligned}$$

where $D_{i,\ell}$ is the ℓ-th element of D_i. On $\bigcap_{\ell=1}^{4} \mathcal{E}_{N,T,\ell}$,

$$\begin{aligned}
(\|D_i\| \vee 1)\|\Omega_i^{-1}\|_2\|\hat{D}_i - D_i\| \leq& (C\sqrt{\log N} \vee 1)\lambda_1^{-1}C(\epsilon_{1,N} + \epsilon_{2,N} + \epsilon_{3,N})/\sqrt{\log N} \\
\leq& C_2^*(\epsilon_{1,N} + \epsilon_{2,N} + \epsilon_{3,N}),
\end{aligned}$$

where C_2^* is a constant that depends only on λ_1, G and K_β. By Lemma D.3, there are independent random variables $(U_i)_{1\leq i\leq N}$ with $U_i \sim \tilde{\chi}^2(\Omega_i)$ such that

$$\begin{aligned}
&\sup_{(r_1,\ldots,r_N)\in\mathcal{R}^N_{++}} \left| P\left(\max_{1\leq i\leq N}\left(T_i^{\text{QLR}} - r_i\right) > 0\right) - P\left(\max_{1\leq i\leq N}(U_i - r_i) > 0\right)\right| \\
\leq & C\left(T^{-1/6}B_{N,T,4}\log^{7/6}N + T^{-1/6}B^2_{N,T,4}\log N\right) \\
\leq & C\left(\epsilon_{3,N}\left(T^{-1/7}(\log N)\right)^6 + \epsilon^2_{3,N}\right)\log N \leq C_3^*\epsilon_{3,N}(\log N),
\end{aligned}$$

where C_3^* is a constant that depends only on λ_1, G and K_β. To avoid ambiguity, denote the quantities in the statement of Lemma D.2 with a † superscript. The conclusion of the theorem follows from applying Lemma D.2 with $\epsilon_N^\dagger = (C_1^* \vee C_2^* \vee C_3^*)\epsilon_N/\log N$, $\hat{D}_i^\dagger = D_i + \Delta_i^D$, $D_i^\dagger = D_i$, $\hat{\Omega}_i^\dagger = \hat{\Omega}_i^V$ and $\Omega_i^\dagger = \Omega_i$ on the event $\mathcal{E}_{N,T,1}\cap\mathcal{E}_{N,T,2}\cap\mathcal{E}_{N,T,3}\cap\mathcal{E}_{N,T,4}$. □

B. Figures and tables

g^0	σ	T	empirical coverage MAX	empirical coverage QLR	cardinality of CS MAX	cardinality of CS QLR
1	0.25	10	0.64	0.71	1.37	1.36
1	0.25	20	0.75	0.83	1.28	1.28
1	0.25	30	0.81	0.88	1.26	1.26
1	0.25	40	0.83	0.90	1.24	1.24
1	0.50	10	0.47	0.49	2.46	2.46
1	0.50	20	0.71	0.77	2.52	2.52
1	0.50	30	0.81	0.84	2.54	2.54
1	0.50	40	0.82	0.83	2.55	2.55
2	0.25	10	0.75	0.76	1.28	1.37
2	0.25	20	0.79	0.78	1.20	1.28
2	0.25	30	0.83	0.83	1.18	1.25
2	0.25	40	0.85	0.83	1.17	1.23
2	0.50	10	0.59	0.62	1.96	2.00
2	0.50	20	0.79	0.76	1.95	2.01
2	0.50	30	0.81	0.82	1.96	2.01
2	0.50	40	0.84	0.84	1.95	2.00
3	0.25	10	0.72	0.74	1.28	1.36
3	0.25	20	0.81	0.80	1.20	1.27
3	0.25	30	0.83	0.83	1.18	1.24
3	0.25	40	0.87	0.85	1.17	1.23
3	0.50	10	0.58	0.58	1.96	1.98
3	0.50	20	0.76	0.76	1.96	1.98
3	0.50	30	0.82	0.82	1.95	1.99
3	0.50	40	0.85	0.86	1.96	1.99

Table B.1: Homoscedastic design with $G = 3$ groups. Results based on $B = 1000$ simulated joint confidence sets with $1 - \alpha = 0.9$. Critical values for MAX and QLR procedures are *not* adjusted for short panels. "Empirical coverage" gives the simulated coverage probability of the joint confidence set. "Cardinality of CS" gives the simulated expected average cardinality of a marginal (unit-wise) confidence set.

	log(uerage)	log(gr)	log(rminwg)
Positive-effect group	0.62	-0.43	0.06
Negative-effect group	0.86	-0.18	-0.07

Table B.2: Estimated group-specific coefficients.

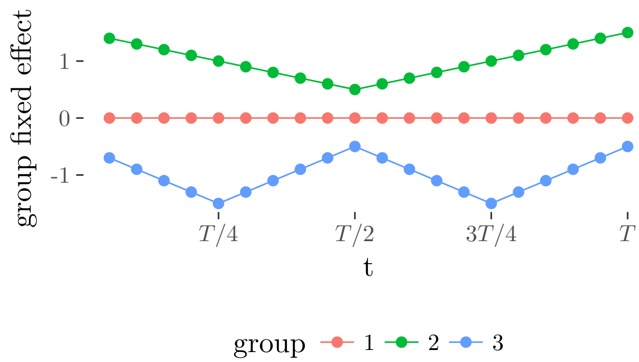

Figure B.1: Time profile of the group fixed effect for the simulation design from Section 6.1.

positive-effect group			
Arizona	California(*)	Colorado	Connecticut(*)
Florida	Georgia	Hawaii	Illinois
Massachusetts(*)	Maryland	Maine	Michigan
North Carolina	New Hampshire	New Jersey	Nevada
New York	Ohio	Pennsylvania	Rhode Island
South Carolina	Texas	Utah	Virginia
Washington	Wisconsin		
negative-effect group			
Alaska	Alabama	Arkansas	Delaware
Iowa	Idaho	Indiana	Kansas(*)
Kentucky	Louisiana	Minnesota(*)	Missouri
Mississippi(*)	Montana	North Dakota(*)	Nebraska(*)
New Mexico	Oklahoma	Oregon	South Dakota(*)
Tennessee	Vermont(*)	West Virginia(*)	Wyoming(*)

Table B.3: Estimated group memberships in minimum wage example. In a joint confidence set at significance level $1 - \alpha = 0.66$, states with a (*) have a marginal confidence set that contains only the estimated group membership. For these states, the estimated group membership is significantly different from the other group.

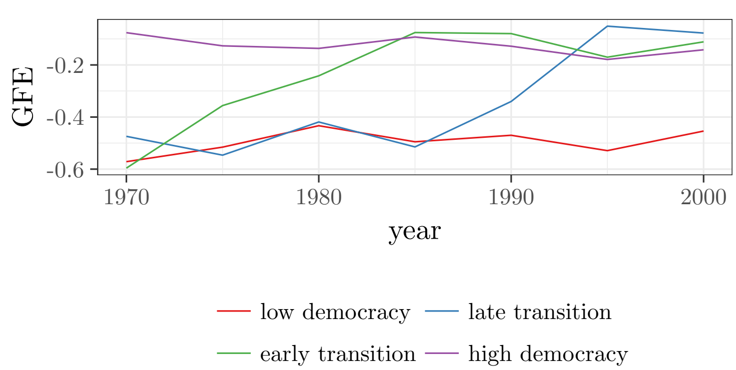

Figure B.2: Estimated time profiles for the group fixed effects.

Supplementary Appendix
for
Confidence Set for Group Membership

Andreas Dzemski and Ryo Okui

March 6, 2018

C. Extension to model with individual fixed effects

In this section, we discuss an extension of the grouped random coefficient model from Example 1 that adds individual fixed effects. As in Example 1, the group-specific coefficients are assumed to be time-invariant. We argue that our procedures can be used after applying a fixed effect transformation.

Suppose that unit i's outcome is generated from

$$y_{it} = x_{it}'\beta_g + \mu_i + u_{it},$$

where μ_i is i's fixed effect and all other quantities are defined as before. The individual fixed effect can be removed by the fixed effect transformation

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta_g + u_{it} - \bar{u}_i,$$

where $\bar{y}_i = \sum_{t=1}^{T} y_{it}/T$, $\bar{x}_i = \sum_{t=1}^{T} x_{it}/T$ and $\bar{u}_i = \sum_{t=1}^{T} u_{it}/T$.

We work with the transformed data to construct confidence sets. The natural counterpart to $\hat{d}_{it}(g,h)$ is given by

$$\begin{aligned}\hat{d}_{it}^{\mathrm{FE}}(g,h) = \frac{1}{2}\Big(&\left(y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i)'\hat{\beta}_g\right)^2 - \left(y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i)'\hat{\beta}_h\right)^2 \\ &+ \left((x_{it} - \bar{x}_i)'\left(\hat{\beta}_g - \hat{\beta}_h\right)\right)^2 \Big).\end{aligned}$$

Replacing $\hat{D}_i(g,h)$ by

$$\hat{D}_i^{\mathrm{FE}}(g,h) = \frac{\sum_{t=1}^{T} \hat{d}_{it}^{\mathrm{FE}}(g,h)}{\sum_{t=1}^{T} \left(\hat{d}_{it}^{\mathrm{FE}}(g,h) - \bar{\hat{d}}_{it}^{\mathrm{FE}}(g,h)\right)^2}, \tag{16}$$

we can follow the recipes for constructing confidence sets in Section 4.3. Strictly speaking, our asymptotic results from Section 5 do not apply here, since $\{(y_{it} - \bar{y}_i, x_{it} - \bar{x}_i)\}_{1 \le t \le T}$ is not an *i.i.d.* sequence. Heuristically, our approach is still expected to work well since

$$x_{it} - \bar{x}_i = x_{it} - \sum_{t=1}^{T} \mathbb{E}(x_{it})/T + O_p\left(T^{-1/2}\right),$$

$$u_{it}-\bar{u}_i=u_{it}+O_p\left(T^{-1/2}\right),$$

so that, asymptotically, the correlation between time periods becomes negligible.

We now discuss bias in $\hat{D}_i^{\mathrm{FE}}(g_i^0,h)$. To this end, suppose that the group-specific parameters are estimated accurately, i.e. $\hat{\beta}_g=\beta_g$ for $g\in\mathbb{G}$.[32] In the case of strictly exogenous regressors, the numerator in (16) has mean zero. With predetermined regressors, such as lagged outcomes, it may be biased. In this case, we can use the half-panel Jackknife from Dhaene and Jochmans (2015) for bias correction.[33]

To explain the adjustment based on the half-panel Jackknife, define the split sample means

$$\bar{w}_{i,1,t_0}=\sum_{t=1}^{t_0}w_{it}/t_0,\quad \bar{w}_{i,2,t_0}=\sum_{t=t_0+1}^{T}w_{it}/(T-t_0)$$

for random vectors $(w_{it})_{1\le t\le T}$. For $j=\{1,2\}$, let

$$\begin{aligned}\hat{d}_{it,j,t_0}^{\mathrm{FE}}(g,h)=&\frac{1}{2}\Big(\left(y_{it}-\bar{y}_{i,j,t_0}-(x_{it}-\bar{x}_{i,j,t_0})'\hat{\beta}_g\right)^2-\left(y_{it}-\bar{y}_{i,j,t_0}-(x_{it}-\bar{x}_{i,j,t_0})'\hat{\beta}_h\right)^2\\&+\left((x_{it}-\bar{x}_i)'\left(\hat{\beta}_g-\hat{\beta}_h\right)\right)^2\Big),\\ \hat{d}_{it,1+2}^{\mathrm{FE}}(g,h)=&\left(\hat{d}_{it,(t-1)\bmod\lfloor T/2\rfloor+1,\lfloor T/2\rfloor}^{\mathrm{FE}}(g,h)+\hat{d}_{it,(t-1)\bmod\lceil T/2\rceil+1,\lceil T/2\rceil}^{\mathrm{FE}}(g,h)\right)/2.\end{aligned}$$

The Jackknifed version of (16) is given by

$$\tilde{D}_i^{\mathrm{FE}}(g,h)=\frac{2\sum_{t=1}^{T}\hat{d}_{it}^{\mathrm{FE}}(g,h)-\sum_{t=1}^{T}\hat{d}_{it,1+2}^{\mathrm{FE}}(g,h)}{\sum_{t=1}^{T}\left(\hat{d}_{it}^{\mathrm{FE}}(g,h)-\bar{\hat{d}}_{it}^{\mathrm{FE}}(g,h)\right)^2}.$$

$\tilde{D}_i^{\mathrm{FE}}$ replaces $\hat{D}_i$ in the test statistics described in Section 4.3.

D. Lemmas

Lemma D.1 (Slutsky-type result for MAX statistic)**.** *Let α denote a constant $0<\alpha<1$. Let $\epsilon_N\ge N^{-1}$ such that*

$$\begin{aligned}\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\left|\hat{D}_i(g_i^0,h)-D_i(g_i^0,h)\right|\le&\epsilon_N\sqrt{\log N},\\ \max_{1\le i\le N}\|\Omega_i^{-1}(g_i^0)\|_2\left(1\vee\|\Omega_i^{-1}(g_i^0)\|_2\|\Omega_i(g_i^0)\|_2\right)\|\hat{\Omega}_i(g_i^0)-\Omega_i(g_i^0)\|_2\le&\epsilon_N.\end{aligned}$$

Let $(X_i)_{1\le i\le N}$ denote a collection of random vectors such that $X_i\sim N(0,\Omega_i)$ and suppose that

$$\sup_{(r_1,\dots,r_N)\in\mathbb{R}^N}\left|P\Big(\max_{1\le i\le N}T_i^{\mathrm{MAX}}(g_i^0)\le r_i\Big)-P\Big(\max_{1\le i\le N}\max_{1\le h\le G-1}X_{i,h}\le r_i\Big)\right|\le\epsilon_N(\log N).$$

[32] Our asymptotic results in Section 5 provide conditions under which parameter estimation affects only higher-order terms.

[33] An alternative approach is analytical bias correction as in Hahn and Kuersteiner (2002).

Also, suppose that

$$64(G-1)^2\log(N/\alpha)\epsilon_N \leq 1.$$

Then, there is a threshold N_0 and a constant C depending only on G and α such that for $N \geq N_0$

$$P\left(\exists i \in \{1,\dots,N\} : \hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha,N}^{\mathrm{MAX}}\big(\hat{\Omega}_i(g_i^0)\big)\right) \leq \alpha + C\epsilon_N(\log N).$$

Proof. For nonsingular covariance matrix V, write $c_{\alpha,N}(V) = c_{\alpha,N}^{\mathrm{MAX}}(V)$. Take N large enough such that

$$\log(N/\alpha) \geq \max\left\{1, \frac{\alpha^2(G-1)}{2\pi}, \frac{2^{G-1}-1}{8(G-1)^2}, \log(2(G-1))\right\},$$

If we choose N large enough, then the assumptions of the lemma imply $\epsilon_N \leq 1/2$ and thus

$$2\max_{1\leq i\leq N}\|\hat{\Omega}_i - \Omega_i\|_2\|\Omega_i^{-1}\|_2 \leq 1.$$

Therefore, we can employ Lemma D.14 to bound

$$\|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2 \leq 2\|\Omega_i^{-1}\|_2^2\|\hat{\Omega}_i - \Omega_i\|_2$$

for all $i = 1,\dots,N$, and

$$\begin{aligned}
&\|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2\left(\|\Omega_i^{-1}\|_2 \vee \|\hat{\Omega}_i^{-1}\|_2\right)\\
\leq &2\|\Omega_i^{-1}\|_2^2\|\Omega_i\|_2\|\hat{\Omega}_i - \Omega_i\|_2 + 2\left(\|\Omega_i^{-1}\|_2\|\Omega_i^{-1}\|_2^2\|\hat{\Omega}_i - \Omega_i\|_2\right)^2 \leq 4\epsilon_N.
\end{aligned} \tag{17}$$

Define

$$\alpha_N = \alpha\big(1 + 16(G-1)^2\log(N/\alpha)\max_{1\leq i\leq N}\|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2\big(\|\Omega_i\|_2 \vee \|\hat{\Omega}_i\|_2\big) \vee N^{-1}\big).$$

Note that (17) imples $\alpha \leq \alpha_N \leq 2\alpha$. First, we show that

$$c_{\alpha_N,N}(\Omega_i) \leq c_{\alpha,N}(\hat{\Omega}_i). \tag{18}$$

Let $a_N^2 = 4(G-1)\|\Omega_i\|_2\log(N/\alpha)$. Note that

$$\begin{aligned}
&\|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2\left(\|\Omega_i^{-1}\|_2 \vee \|\hat{\Omega}_i^{-1}\|_2 \vee (G-1)a_N^2\right)\\
\leq &8(G-1)^2\|\Omega_i^{-1}\|_2^2\|\Omega_i\|_2\|\hat{\Omega}_i - \Omega_i\|_2\log(N/\alpha) + 2\left(\|\Omega_i^{-1}\|_2\|\hat{\Omega}_i - \Omega_i\|_2\right)^2 \leq 1.
\end{aligned}$$

This verifies the required assumption for the application of Lemma D.18 below. For $X \sim \mathcal{N}(0,\Omega_i)$ and $\hat{X} \sim \mathcal{N}(0,\hat{\Omega}_i)$ we have

$$\begin{aligned}
&P\Big(\max_{j=1,\dots,G-1} X_j > c_{\alpha,N}(\hat{\Omega}_i)\Big)\\
\leq &P\Big(\max_{j=1,\dots,G-1} X_j > c_{\alpha,N}(\hat{\Omega}_i) \wedge \|X\|_{\max} \leq a_N\Big) + P(\|X\|_{\max} > a_N)
\end{aligned}$$

$$
\begin{aligned}
\leq & \frac{P\Big(\max_{j=1,\dots,G-1} X_j > c_{\alpha,N}(\hat{\Omega}_i) \wedge \|X\|_{\max} \leq a_N\Big)}{P\Big(\max_{j=1,\dots,G-1} \hat{X}_j > c_{\alpha,N}(\hat{\Omega}_i) \wedge \|\hat{X}\|_{\max} \leq a_N\Big)} P\Big(\max_{j=1,\dots,G-1} \hat{X}_j > c_{\alpha,N}(\hat{\Omega}_i)\Big) \\
& + P(\|X\|_{\max} > a_N) \\
\leq & \big(1 + (2^{G-1}-1)\|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2 \|\hat{\Omega}_i\|_2 + 2(G-1)a_N^2 \|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2\big)(\alpha/N) \\
& + P(\|X\|_{\max} > a_N) \\
\leq & \Bigg(1 + \|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2 (\|\Omega_i\|_2 \vee \|\hat{\Omega}_i\|_2)\big((2^{G-1}-1) + \log(N/\alpha) 8(G-1)^2\big) \\
& \quad + \frac{(\alpha/N)\sqrt{G-1}}{\sqrt{2\pi \log(N/\alpha)}}\Bigg)\frac{\alpha}{N} \leq \frac{\alpha_N}{N}.
\end{aligned}
$$

The third inequality above follows from Lemma D.18 noting that, under the assumptions of the lemma, we can take

$$
(2^{G-1}-1)\|\hat{\Omega}_i^{-1} - \Omega_i^{-1}\|_2 \|\hat{\Omega}_i\|_2 \leq (2^{G-1}-1)\epsilon_N \leq 1.
$$

The fourth inequality follows from Lemma D.17. This establishes (18). Let $(X_{i,1}, \dots, X_{i,G-1})$ denote a centered normal random vector with covariance matrix Ω_i. Next, we show that for a universal constant $\tilde{C}$ and a threshold N_0 that is independent of $(\Omega_i)_{1\leq i \leq N}$, for all $b_N > 0$,

$$
\begin{aligned}
& P\left(\Big| \max_{1\leq i\leq N}\left(\max_{1\leq h\leq G-1} X_{i,h} - c_{\alpha_N,N}(\Omega_i)\right)\Big| \leq b_N\right) \\
\leq & \tilde{C}(b_N \vee N^{-1})\sqrt{2\log(N\sqrt{G-1})}
\end{aligned}
\tag{19}
$$

for $N \geq N_0$. There exists N_0, independent of Ω_i, such that for $N \geq N_0$

$$
\sqrt{\log(N/\alpha_N)} < c_{\alpha_N,N}(\Omega_i) \leq \sqrt{2\log(G-1)} + \sqrt{2\log(N/\alpha_N)}. \tag{20}
$$

The lower bound follows from the fact that $T_i^{\mathrm{MAX}} \geq Z$ for standard normal Z in conjunction with a bound on the tail probability of a standard normal random variable (e.g., the argument in the proof of Lemma D.20 with $a = 2$). The upper bound follows from Lemma A.4 in Chernozhukov, Chetverikov, and Kato (2014). The inequality

$$
\log(N/\alpha_N) \geq \log(N/(2\alpha)) \geq \log(G-1)
$$

implies that the right-hand side of (20) can be bounded by

$$
\sqrt{2\log(N/\alpha_N)} + \sqrt{2\log(N/\alpha_N)} \leq \sqrt{8\log(N/\alpha_N)}.
$$

Therefore, to prove (19) it suffices to show

$$
\begin{aligned}
& \max_{\substack{(a_i)_{1\leq i\leq N} \\ 1\leq a_i \leq 2\sqrt{2}}} P\left(\Big| \max_{i=1,\dots,N}\left(\max_{1\leq h\leq G-1} X_{i,h} - a_i\sqrt{\log(N/\alpha_N)}\right)\Big| \leq b_N\right) \\
\leq & \tilde{C}(b_N \vee N^{-1})\sqrt{2\log(N\sqrt{G-1})}.
\end{aligned}
$$

For $N \geq 2$ we write

$$
\begin{aligned}
&\max_{\substack{(a_i)_{1\le i\le N}\\ 1\le a_i\le 8}} P\left(\left|\max_{1\le i\le N}\left(\max_{1\le h\le G-1} X_{i,h} - a_i\sqrt{\log(N/\alpha_N)}\right)\right| \le b_N\right)\\
\le &\max_{\substack{(a_i)_{1\le i\le N}\\ 1\le a_i\le 8}} \sup_{x\in\mathcal{R}} P\left(\left|\max_{1\le i\le N}\max_{1\le h\le G-1}\frac{X_{i,h}}{a_i} - x\right| \le b_N \vee N^{-1}\right)\\
\le &\tilde{C}(b_N\vee N^{-1})\sqrt{1\vee\log\left(\frac{N(G-1)}{b_N\vee N^{-1}}\right)} \le \tilde{C}(b_N\vee N^{-1})\sqrt{2\log(N\sqrt{G-1})}.
\end{aligned}
$$

The second inequality follows from Corollary 1 in Chernozhukov, Chetverikov, and Kato (2015). Collecting the results from above yields

$$
\begin{aligned}
&P\Big(\max_{1\le i\le N}\left(\hat{T}_i^{\mathrm{MAX}} - c_{\alpha,N}(\hat{\Omega}_i)\right) > 0\Big) \le P\Big(\max_{1\le i\le N}\left(\hat{T}_i^{\mathrm{MAX}} - c_{\alpha_N,N}(\Omega_i)\right) > 0\Big)\\
\le &P\Big(\max_{1\le i\le N}\left(\max_{1\le h\le G-1} X_{i,h} - c_{\alpha_N,N}(\Omega_i)\right) + \epsilon_N\sqrt{\log N} > 0\Big)\\
&+ \Big|P\Big(\max_{1\le i\le N}\left(\max_{h\le\le G-1} D_i(h) - c_{\alpha_N,N}(\Omega_i)\right) + \epsilon_N\sqrt{\log N} > 0\Big)\\
&- P\Big(\max_{1\le i\le N}\left(\max_{1\le h\le G-1} X_{i,h} - c_{\alpha_N,N}(\Omega_i)\right) + \epsilon_N\sqrt{\log N} > 0\Big)\Big|\\
\le &P\Big(\max_{1\le i\le N}\left(\max_{1\le h\le G-1} X_{i,h} - c_{\alpha_N,N}(\Omega_i)\right) > 0\Big)\\
&+ P\left(\left|\max_{1\le i\le N}\left(\max_{1\le h\le G-1} X_{i,h} - c_{\alpha_N,N}(\Omega_i)\right)\right| \le \epsilon_N\sqrt{\log N}\right) + \epsilon_N(\log N)\\
\le &\sum_{i=1}^{N}\frac{\alpha_N}{N} + \tilde{C}\epsilon_N\sqrt{\log N}\sqrt{2\log(N\sqrt{G-1})} + \epsilon_N(\log N)\\
\le &\alpha\left(1 + C\epsilon_N(\log N) + N^{-1}\right) + C\epsilon_N(\log N).
\end{aligned}
$$

The first inequality holds due to (18) and the fourth inequality holds due to (19). The last inequality holds due to

$$
\begin{aligned}
\alpha_N \le &\alpha + \alpha\left(64(G-1)^2\log\left(N/\alpha\right)\epsilon_N \vee N^{-1}\right)\\
\le &\alpha + 64\left(G-1\right)^2(\log N)\epsilon_N\left(1 - \frac{\log\alpha}{\log N}\right).
\end{aligned}
$$

□

Lemma D.2 (Slutzky-type result for QLR)**.** *Let α denote a constant $0 < \alpha < 1$. Suppose that there is a sequence ϵ_N such that*

$$
\begin{aligned}
\max_{1\le i\le N}(1\vee\|\Omega_i^{-1}(g_i^0)\|_2^2)(\|\Omega_i(g_i^0)\|_2\vee\|\Omega_i^{-1}(g_i^0)\|_2)\|\hat{\Omega}_i(g_i^0) - \Omega_i(g_i^0)\|_2 \le &\epsilon_N,\\
\max_{1\le i\le N}(\|D_i(g_i^0)\|\vee 1)\|\Omega_i^{-1}(g_i^0)\|_2\|\hat{D}_i(g_i^0) - D_i(g_i^0)\| \le &\epsilon_N(\log N),
\end{aligned}
$$

$\epsilon_N \leq 1/48$ *and*

$$32(G-1)^2\epsilon_N \log(N/\alpha) \leq 1.$$

In addition, suppose that $\max_{1\leq i\leq N}\|\hat{D}_i(g_i^0) - D_i(g_i^0)\| \leq 1$. *Let* $(U_i)_{1\leq i\leq N}$ *denote independent random variables with* $U_i \sim \tilde{\chi}^2(\Omega_i)$ *such that*

$$\sup_{(r_1,\ldots,r_N)\in\mathcal{R}^N_{++}} \left| P\left(\max_{1\leq i\leq N}(U_i - r_i) > 0\right) - P\left(\max_{1\leq i\leq N}(T_i^{\mathrm{QLR}}(g_i^0) - r_i) > 0\right)\right| \leq \epsilon_N(\log N).$$

Then, there is a constant C *and a threshold* N_0 *depending only on* α, G *and the sequence* ϵ_N *such that for* $N \geq N_0$

$$P\left(\exists i \in \{1,\ldots,N\} : \hat{T}_i^{\mathrm{QLR}}(g_i^0) > c_{\alpha,N}^{\mathrm{QLR}}\left(\hat{\Omega}_i(g_i^0)\right)\right) \leq \alpha + C\left(\epsilon_N(\log N) + N^{-1}\right).$$

Proof. To simplify notation, we fix the null hypothesis and drop the g_i^0 argument. For nonsingular covariance matrix V, we write $c_{\alpha,N}(V) = c_{\alpha,N}^{\mathrm{QLR}}(V)$. Define

$$\alpha_N = \alpha\left(1 + 96\epsilon_N \log(N/\alpha)\right)$$

and

$$\begin{aligned} b_{1,N} &= \max_{1\leq i\leq N} 2\|\Omega_i\|_2\|\Omega_i^{-1}\|_2^2\|\hat{\Omega}_i - \Omega_i\|_2, \\ b_{N,2} &= \max_{1\leq i\leq N}(2\|D_i\| + 3)\|\Omega_i^{-1}\|_2\|\hat{D}_i - D_i\|. \end{aligned}$$

Choose N large enough so that

$$32(G-1)^2 \max_{1\leq i\leq N}\|\hat{\Omega}_i - \Omega_i\|_2\left(1 \vee 2\|\Omega^{-1}\|_2^2\right)\left(\|\Omega_i\|_2 \vee \|\Omega_i^{-1}\|_2\right) \leq 1.$$

Then, Lemma D.19 gives $c_{\alpha_N,N}(\Omega_i) \leq c_{\alpha,N}(\hat{\Omega}_i)$ for N large enough. Next, we show that we can choose N_0, depending only on α, such that, for $N \geq N_0$,

$$\begin{aligned} &P\left(\left|\max_{1\leq i\leq N}\left(U_i - c_{\alpha_N,N}(\Omega_i)\right)\right| \leq b_{1,N}\max_{1\leq i\leq N}\left(c_{\alpha_N,N}(\Omega_i)\right) + b_{2,N}\right\}\right) \\ \leq &C_1\left(b_{1,N}\max_{1\leq i\leq N}\left(c_{\alpha_N,N}(\Omega_i)\right) + b_{2,N} + N^{-1}\right) + N^{-1} \end{aligned} \tag{21}$$

for a constant C_1 depending only on α and p. This follows from an application of Lemma D.9. First, we bound $c_{\alpha_N,N}(\Omega_i)$. We choose N large enough such that

$$\log N \geq \max\{-\log\alpha, 2\log\alpha\}.$$

The upper bound from Lemma D.20 implies that, for all N exceeding a threshold that depends only on α,

$$c_{\alpha_N,N}(\Omega_i) \leq c_{\alpha,N}(\Omega_i) \leq 4\log(N/\alpha) \leq 4\log N\left(1 - (\log\alpha)/(\log N)\right) \leq 8(\log N).$$

The lower bound from Lemma D.20 gives that for all N exceeding a threshold that depends only on

α

$$c_{\alpha_N,N}(\Omega_i) \geq c_{2\alpha,N}(\Omega_i) \geq \log(N/\alpha) \geq \log N\,(1-(\log\alpha)/(\log N)) \geq (1/2)(\log N).$$

These results imply that, when applying Lemma D.9, we can choose $\underline{a} = (1/2)\log N$ and $\bar{a} = 8\log N$. Next, we choose N large enough that we can take $\epsilon_N \leq 1/48$. Then,

$$b_{1,N} + b_{2,N}/(6\log N) \leq 3\epsilon_N \leq 1/16$$

and

$$b_{1,N}\max_{1\leq i\leq N} c_{\alpha_N,N}(\Omega_i) + b_{2,N} \leq 8(\log N)\,(b_{1,N} + b_{2,N}/(6\log N)) \leq \underline{a}/2\log N$$

and we can take

$$\left(b_{1,N}\max_{1\leq i\leq N} c_{\alpha_N,N}(\Omega_i) + b_{2,N}\right) \vee N^{-1} \leq \underline{a}/2\log N$$

for N large enough. Therefore, we may set $\tau = 1$ and

$$\epsilon = \left(b_{1,N}\max_{1\leq i\leq N} c_{\alpha_N,N}(\Omega_i) + b_{2,N}\right) \vee N^{-1}$$

in Lemma D.9. This proves (21). Choose N_0 such that $16p\epsilon_N(\log N) \leq 1$ for $N \geq N_0$. This is sufficient to guarantee that the assumptions of Lemma D.19 are satisfied, and therefore, $c_{\alpha,N}(\hat{\Omega}_i) \geq c_{\alpha_N,N}(\Omega_i)$. By Lemma D.5 and Lemma D.6

$$\hat{T}_i^{\text{QLR}} \leq \left(T_i^{\text{QLR}} + b_{2,N}\right)(1 + b_{1,N})$$

for all $i = 1,\ldots,N$, and therefore

$$\begin{aligned}\left\{\hat{T}_i^{\text{QLR}} > c_{\alpha,N}(\hat{\Omega}_i)\right\} &\subset \left\{\hat{T}_i^{\text{QLR}} > c_{\alpha_N,N}(\Omega_i)\right\}\\ &\subset \left\{T_i^{\text{QLR}} > c_{\alpha_N,N} - b_{1,N}c_{\alpha_N,N}(\Omega_i)/(1+b_{1,N}) - b_{2,N}\right\}\\ &\subset \left\{T_i^{\text{QLR}} > c_{\alpha_N,N} - b_{1,N}c_{\alpha_N,N}(\Omega_i) - b_{2,N}\right\}.\end{aligned}$$

Write $c_{N,i} = c_{\alpha_N,N}(\Omega_i) - b_{1,N}c_{\alpha_N,N}(\Omega_i) - b_{2,N}$. Collecting the results from above yields

$$\begin{aligned}&P\left(\exists i \in \{1,\ldots,N\} : \hat{T}_i^{\text{QLR}} > c_{\alpha,N}(\hat{\Omega}_i)\right)\\ \leq& P\left(\max_{1\leq i\leq N}\left(T_i^{\text{QLR}} - c_{\alpha_N,N}(\Omega_i) + b_{1,N}c_{\alpha_N,N}(\Omega_i) + b_{2,N}\right) > 0\right)\\ \leq& P\left(\max_{1\leq i\leq N}(U_i - c_{N,i}) > 0\right)\\ &+\left|P\left(\max_{1\leq i\leq N}(U_i - c_{N,i}) > 0\right) - P\left(\max_{1\leq i\leq N}\left(T_i^{\text{QLR}} - c_{N,i}\right) > 0\right)\right|\\ \leq& P\left(\max_{1\leq i\leq N}(U_i - c_{\alpha_N,N}(\Omega_i)) > 0\right)\end{aligned}$$

$$
\begin{aligned}
&+ P\left(\left|\max_{1\le i\le N} U_i - c_{\alpha_N,N}(\Omega_i)\right| \le \left(b_{1,N}\max_{1\le i\le N} c_{\alpha_N,N}(\Omega_i) + b_{2,N}\right)\vee N^{-1}\right)\\
&+ \epsilon_N(\log N)\\
\le&\sum_{i=1}^{N} P\left(U_i > c_{\alpha_N,N}(\Omega_i)\right) + C_1\left(b_{1,N}\max_{1\le i\le N} c_{\alpha_N,N}(\Omega_i) + b_{2,N}\right)\vee C_1 N^{-1} + N^{-1}\\
&+ \epsilon_N(\log N)\\
\le&\alpha_N + C\left(\epsilon_N(\log N) + N^{-1}\right).
\end{aligned}
$$

where C is a constant that can be chosen to depend only on C_1 and α. The fourth inequality follows from the union bound and the anti-concentration inequality (21). The conclusion follows upon noting that

$$
\begin{aligned}
\alpha_N \le \alpha + 96\epsilon_N \log(N/\alpha) \le&\alpha + 96\epsilon_N\log(N)\left(1 - \frac{\log\alpha}{\log N}\right)\\
\le&\alpha + 192\epsilon_N\log(N).
\end{aligned}
$$

□

Lemma D.3 (Large CLT for QLR statistic)**.** *Let P denote a probability measure that satisfies Assumption 1 and imposes cross-sectional independence. Let $\lambda_1 = \min_{i=1}^{N}\min_{g\in\mathbb{G}}\lambda_1\big(\Omega_i(g_i^0)\big)$ and suppose that $\lambda_1 > 0$. Then, there are random variables $(U_i)_{1\le i\le N}$ with $U_i \sim \tilde{\chi}^2(\Omega_i(g_i^0))$ such that*

$$
\begin{aligned}
&\sup_{(r_1,\ldots,r_N)\in\mathcal{R}^N_{++}}\left|P\left(\max_{1\le i\le N}\left(T_i^{\mathrm{QLR}}(g_i^0) - r_i\right) > 0\right) - P\left(\max_{1\le i\le N}(U_i - r_i) > 0\right)\right|\\
\le& C\left\{\left(\frac{GB_{N,T,4}^6\log^7((G-1)NT)}{T}\right)^{1/6} + \left(\frac{GB_{N,T,4}^6\log^3((G-1)NT)}{\sqrt{T}}\right)^{1/3}\right\},
\end{aligned}
$$

where C is a constant that depends only on λ_1 and G.

Proof. Let $t_i(x) = t_i(x_1,\ldots,x_N) = \inf_{s\le 0}(x_i - s)'\Omega_i^{-1}(x_i - s)$. We first show that, for all $r > 0$, the set $\{x\in\mathcal{R}^{N(G-1)} : t_i(x)\le r\}$ is a convex set. Let $S = \{y\in\mathcal{R}^{G-1} : y\le 0\}$. For $y\in\mathcal{R}^{G-1}$, define $\|y\|_{\Omega^{-1}} = \sqrt{y'\Omega^{-1}y}$ and $d_i(y,S) = \inf_{z\in S}\|y - z\|_{\Omega_i^{-1}}$. Convexity of S and positive definiteness of Ω_i imply that there is a unique $\hat{y}$ such that $d(y,S) = \|y - \hat{y}\|_{\Omega_i^{-1}}$. For $y_1, y_2\in\mathcal{R}^{G-1}$ and $\lambda\in[0,1]$ define $y_\lambda = \lambda y_1 + (1-\lambda)y_2$. Define also $y_\lambda^* = \lambda\hat{y}_1 + (1-\lambda)\hat{y}_2$. Then, $y_\lambda^*\in S$ and therefore, by the triangle inequality,

$$
\begin{aligned}
d(y_\lambda, S) \le \|y_\lambda - y_\lambda^*\|_{\Omega_i^{-1}} \le&\lambda\|y_1 - \hat{y}_1\|_{\Omega_i^{-1}} + (1-\lambda)\|y_2 - \hat{y}_2\|_{\Omega_i^{-1}}\\
=&\lambda d(y_1,S) + (1-\lambda)d(y_2,S).
\end{aligned}
$$

This proves that, for $r_i\in\mathcal{R}^{(G-1)}$, the set

$$
\{x\in\mathcal{R}^{N(G-1)} : t_i(x)\le r_i\} = \{x\in\mathcal{R}^{N(G-1)} : d(x_i,S)\le\sqrt{r_i}\}
$$

is convex. For $r_1, \dots, r_N \in \mathcal{R}^N_{++}$ the set

$$\bigcap_{i=1}^{N} \Big\{ x \in \mathcal{R}^{N(G-1)} : t_i(x) \le r_i \Big\}$$

is therefore a sparse-convex set, as defined in Chernozhukov, Chetverikov, and Kato (2016). Let

$$Z_{it}(h) = d_{it}(h)/(\sigma_i s_{i,T}(h)).$$

and $Z_{it} = (Z_{it}(h))_{h \in \mathbb{G} \setminus \{g_i^0\}}$. Let $\tilde{X}_t = (\tilde{X}_{1t}, \dots, \tilde{X}_{Nt})'$ with $\dim(\tilde{X}_{it}) = G-1$ for $i = 1, \dots, N$, $t = 1, \dots, T$ denote a centered normal random vector with the property that $\tilde{X}_t$ and $\tilde{X}_s$ are independent for $t \neq s$ and $\mathbb{E}_P[\tilde{X}_t(\tilde{X}_t)'] = \mathbb{E}_P[Z_t(Z_t)']$ for $i = 1, \dots, N$, $t = 1, \dots, T$. Condition (M.1") in Chernozhukov, Chetverikov, and Kato (2016) is satisfied with $b = \lambda_1$. Let v denote a vector $v = (v_j)_{j=1}^{N(G-1)}$ with $\|v_j\| = 1$ and $\|v_j\|_0 \le (G-1)$. Also, let $j(i,h) = (i-1)(G-1) + (h-1)$ and $v^{(i)} = (v_{j(i,j)})_{h \in \mathbb{G} \setminus \{g_i^0\}}$. Because of cross-sectional independence, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_P \Big(\sum_{i=1}^{N} \sum_{h \in \mathbb{G} \setminus \{g_i^0\}} v_{j(i,h)} Z_{it}(h) \Big)^2 \\
=&\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} \sum_{h \in \mathbb{G} \setminus \{g_i^0\}} \sum_{h' \in \mathbb{G} \setminus \{g_i^0\}} v_{j(i,h)} v_{j(i,h')} \mathbb{E}_P \big[Z_{it}(h) Z_{it}(h') \big] \\
=&\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} (v^{(i)})' \big[\Omega_{i,t}\big] v^{(i)} \\
=&\sum_{i=1}^{N} (v^{(i)})' \Big[\frac{1}{T}\sum_{t=1}^{T} \Omega_{i,t} \Big] v^{(i)} = \sum_{i=1}^{N} (v^{(i)})' \big[\Omega_i(g_i^0)\big] v^{(i)} \\
\ge&\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} \lambda_1 \big[\Omega_i\big] \|v^{(i)}\|^2 \ge \lambda_1 \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} \|v^{(i)}\|^2 = \lambda_1.
\end{aligned}$$

This verifies assumption (M.1") in Chernozhukov, Chetverikov, and Kato (2016). Next, by Hölder's inequality there is a constant $C_1 \ge 1$ depending only on K_β such that

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_P[|Z_{it}|^3] &\le C \big(B^4_{N,T,4}\big)^{3/4} \le C_1 G^{1/2} B^3_{N,T,4}, \\
\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_P[|Z_{it}|^4] &\le C B^4_{N,T,4} \le (C_1 G^{1/2} B^3_{N,T,4})^2.
\end{aligned}$$

This allows us to choose $C_1 G^{1/2} B^3_{N,T,4}$ as the sequence of constants in assumption (M.2) in Chernozhukov, Chetverikov, and Kato (2016). Lastly, we verify assumption (E.2) in Chernozhukov, Chetverikov, and Kato (2016). To this end, note that

$$\mathbb{E}_P \left[\max_{1 \le i \le N} \max_{h \in \mathbb{G} \setminus \{g_i^0\}} \Big| Z_{it}(h)/(G^{1/4} B^3_{N,T,4}) \Big|^4 \right] \le \sum_{h \in \mathbb{G}} \mathbb{E}_P \left[\max_{1 \le i \le N} \Big| Z_{it}(h)/(G^{1/4} B^3_{N,T,4}) \Big|^4 \right]$$

$$\leq G^2 C_1^2 B_{N,T,4}^6/(G^2 C_1^4 B_{N,T,4}^{12}) \leq 1 \leq 2,$$

where we used that $B_{N,T,4} \geq 1$. We may now apply Proposition 3.2 in Chernozhukov, Chetverikov, and Kato 2016 to deduce

$$\sup_{(r_1,\ldots,r_N)\in\mathcal{R}_{++}^N} \left| P\Big(\max_{1\leq i\leq N} \big(t_i(D_i) - r_i\big) > 0\Big) - P\Big(\max_{1\leq i\leq N} (U_i - r_i) > 0\Big)\right|$$

$$\leq \sup_{(r_1,\ldots,r_N)\in\mathcal{R}_{++}^N} \left| P\Big(\bigcap_{i=1}^N \Big\{t_i(D_i) \leq r_i\Big\}\Big) - P\Big(\bigcap_{i=1}^N \Big\{t_i\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{X}_{it}\Big) \leq r_i\Big\}\Big)\right|$$

$$\leq C\left\{\left(\frac{GB_{N,T,4}^6 \log^7((G-1)NT)}{T}\right)^{1/6} + \left(\frac{GB_{N,T,4}^6 \log^3((G-1)NT)}{\sqrt{T}}\right)^{1/3}\right\},$$

where C is a constant that depends only on λ_1, G and K_β. Next, note that

$$t_i\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{X}_{it}\Big) = \inf_{t\geq 0}\left(-\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{X}_{it} - t\right)' \Omega_i^{-1}\left(-\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{X}_{it} - t\right).$$

Since $-\sum_{t=1}^T \tilde{X}_{it}/\sqrt{T}$ is a zero-mean normal random vector with covariance matrix Ω_i the right-hand side follows a $\bar{\chi}^2(\Omega_i)$-distribution. □

Lemma D.4 (Large CLT for MAX statistic)**.** *Let P denote a probability measure satisfying Assumption 1. For $i = 1, \ldots, N$, there are centered normal random vectors X_i with $\mathbb{E}_P[X_i X_i'] = \Omega_i(g_i^0)$ such that*

$$\sup_{(r_1,\ldots,r_N)\in\mathcal{R}_{++}^N} \left| P\Big(\max_{1\leq i\leq N}\Big(\max_{h\in\mathbb{G}\setminus\{g_i^0\}} D_i(g_i^0,h) - r_i\Big) > 0\Big) \right.$$
$$\left. - P\Big(\max_{1\leq i\leq N}\Big(\max_{1\leq h\leq G-1} X_{i,h} - r_i\Big) > 0\Big)\right|$$
$$\leq C\left\{\left(\frac{GB_{N,T,4}^6 \log^7((G-1)NT)}{T}\right)^{1/6} + \left(\frac{GB_{N,T,4}^6 \log^3((G-1)NT)}{\sqrt{T}}\right)^{1/3}\right\},$$

where C is a constant depending only on G.

Let $Z_{it}(h) = d_{it}(g_i^0, h)/s_{i,T}(g_i^0, h)$ and

$$Z_t = \big((Z_{1t}(g_1^0,h))_{h\in\mathbb{G}\setminus\{g_1^0\}}, \ldots, (Z_{Nt}(g_N^0,h))_{h\in\mathbb{G}\setminus\{g_N^0\}}\big)'.$$

Let $\tilde{X}_t = (\tilde{X}_{1t}, \ldots, \tilde{X}_{Nt})'$ with $\dim(\tilde{X}_{it}) = G-1$ for $i = 1, \ldots, N$, $t = 1, \ldots, T$ denote a normal random vector with the property that $\tilde{X}_t$ and $\tilde{X}_s$ are independent for $t \neq s$ and $\mathbb{E}_P[\tilde{X}_t(\tilde{X}_t)'] = \mathbb{E}_P[Z_t(Z_t)']$ for $i = 1, \ldots, N$, $t = 1, \ldots, T$. Define $X_i = \sum_{t=1}^T \tilde{X}_{it}/\sqrt{T}$. Clearly, X_i is a normal random vector with covariance matrix Ω_i. Let $a_i = -\infty$ and $b_i = r_i$. Then we may write

$$\sup_{(r_1,\ldots,r_N)\in\mathcal{R}_{++}^N} \left| P\Big(\max_{1\leq i\leq N}\Big(\max_{h\in\mathbb{G}\setminus\{g_i^0\}} D_i(g_i^0,h) - r_i\Big) > 0\Big) \right.$$
$$\left. - P\Big(\max_{1\leq i\leq N}\Big(\max_{1\leq h\leq G-1} X_{i,h} - r_i\Big) > 0\Big)\right|$$

$$
\begin{aligned}
\leq & \sup_{(r_1,\ldots,r_N)\in\mathcal{R}^N_{++}} \Bigg| P\Bigg(\bigcap_{i=1}^{N} \bigcap_{h\in\mathbb{G}\setminus\{g_i^0\}} \{a_i < D_i(g_i^0,h) \leq b_i\} \Bigg) \\
& \qquad - P\Bigg(\bigcap_{i=1}^{N} \bigcap_{h=1}^{G-1} \{a_i < X_{i,h} \leq b_i\} \Bigg) \Bigg| \\
\leq & \sup_{(r_1,\ldots,r_N)\in\mathcal{R}^N_{++}} \Bigg| P\Bigg(\bigcap_{i=1}^{N} \bigcap_{h\in\mathbb{G}\setminus\{g_i^0\}} \left\{a_i < \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{it}(g_i^0,h) \leq b_i\right\} \Bigg) \\
& \qquad - P\Bigg(\bigcap_{i=1}^{N} \bigcap_{h=1}^{G-1} \left\{a_i < \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \tilde{X}_{it,h} \leq b_i\right\} \Bigg) \Bigg| \\
\leq & C\left\{ \left(\frac{GB_{N,T,4}^6 \log((G-1)NT)}{T} \right)^{1/6} + \left(\frac{GB_{N,T,4}^6 \log((G-1)NT)}{\sqrt{T}} \right)^{1/3} \right\}.
\end{aligned}
$$

The last inequality holds by Proposition 2.1 in Chernozhukov, Chetverikov, and Kato (2016). Their assumption (M.1) holds trivially with $b = 1$. As in the proof of Lemma D.3, their assumption (M.2) can be verified for the deterministic sequence $G^{1/4}B_{N,T,4}^3$. Then, their assumption (E.2) holds with $q = 4$.

Lemma D.5. *Suppose that $\Omega_i^{-1}(g)$ is symmetric and positive definite and*

$$
2\|\Omega_i^{-1}(g)\|_2\|\hat{\Omega}_i(g) - \Omega_i(g)\|_2 \leq 1.
$$

Then,

$$
\begin{aligned}
\hat{T}_i^{\text{QLR}}(g) \leq & \left(1 + 2\|\Omega(g)\|_2\|\Omega_i^{-1}(g)\|_2^2\|\hat{\Omega}_i(g) - \Omega_i(g)\|_2\right) \\
& \times \min_{t\leq 0} \left(\hat{D}_i(g) - t\right)' \Omega_i{}^{-1}(g) \left(\hat{D}_i(g) - t\right).
\end{aligned}
$$

Proof. For brevity, we write $\hat{D} = \hat{D}_i(g)$, $\hat{\Omega} = \hat{\Omega}_i(g)$ and $\Omega = \Omega_i(g)$ and

$$
\hat{T}_i^{\Omega} = \min_{t\leq 0}(\hat{D} - t)'\Omega^{-1}(\hat{D} - t).
$$

Let $t^* \in \mathcal{R}^p$ such that $t^* \leq 0$ and $\hat{T}_i^{\Omega} = (\hat{D} - t^*)'\Omega^{-1}(\hat{D} - t^*)$. By definition

$$
\begin{aligned}
\hat{T}_i^{\text{QLR}} \leq & \left(\hat{D} - t^*\right)' \hat{\Omega}^{-1} \left(\hat{D} - t^*\right) \\
\leq & \hat{T}_i^{\Omega} + \left|(\hat{D} - t^*)'\left[\hat{\Omega}^{-1} - \Omega^{-1}\right](\hat{D} - t^*)\right|.
\end{aligned}
$$

Let $0 < \lambda_{i,1} \leq \cdots \leq \lambda_{i,p}$ denote the eigenvalues of Ω and note that $\lambda_{i,p}^{-1} = \|\Omega\|_2^{-1}$. Let $\Omega^{-1} = P\Lambda P'$, where P is an orthogonal matrix and Λ is a diagonal matrix with diagonal entries $(\lambda_{i,j}^{-1})_{j=1}^p$. Then,

$$
\begin{aligned}
(\hat{D} - t^*)'\Omega^{-1}(\hat{D} - t^*) = & \sum_{j=1}^{p} \lambda_{i,j}^{-1} \left(P'(\hat{D} - t^*)\right)^2 \\
& \geq \min_{1\leq j\leq p}\{\lambda_{i,j}^{-1}\}\|P'(\hat{D} - t^*)\|^2 = \|\Omega\|_2^{-1}\|\hat{D} - t^*\|^2.
\end{aligned}
$$

Therefore,

$$\left|(\hat{D}-t^*)'\left[\hat{\Omega}^{-1}-\Omega^{-1}\right](\hat{D}-t^*)\right| \leq \|\hat{D}-t^*\|^2\|\hat{\Omega}^{-1}-\Omega^{-1}\|_2 \leq \|\Omega\|_2\|\hat{\Omega}^{-1}-\Omega^{-1}\|_2\hat{T}_i^{\Omega}.$$

By combining inequalities, we obtain

$$\hat{T}_i^{\text{QLR}} \leq \hat{T}_i^{\Omega}\left(1+\|\Omega\|_2\|\hat{\Omega}^{-1}-\Omega^{-1}\|_2\right) \leq \hat{T}_i^{\Omega}\left(1+2\|\Omega\|_2\|\Omega^{-1}\|_2^2\|\hat{\Omega}-\Omega\|_2\right),$$

where the last inequality holds by Lemma D.14. □

Lemma D.6. *Suppose that* $\|\hat{D}_i(g)-D_i(g)\| \leq 1$. *Then*

$$\begin{aligned}&\left|T_i^{\text{QLR}}(g)-\min_{t\leq 0}\big(\hat{D}_i(g)-t\big)'\Omega_i^{-1}(g)\big(\hat{D}_i(g)-t\big)\right|\\ \leq&(2\|D_i(g)\|+3)\|\Omega_i^{-1}(g)\|_2\|\hat{D}_i(g)-D_i(g)\|.\end{aligned}$$

Proof. For brevity, we write $D=D_i(g)$, $\hat{D}=\hat{D}_i(g)$ and $\Omega=\Omega_i(g)$ and define

$$\hat{T}_i^{\Omega}=\min_{t\leq 0}\big(\hat{D}_i(g)-t\big)'\Omega_i^{-1}(g)\big(\hat{D}_i(g)-t\big).$$

Write $\|v\|_{\Omega^{-1}}=\sqrt{v'\Omega^{-1}v}$ and note that $\|\cdot\|_{\Omega^{-1}}$ defines a vector norm and $\|v\|_{\Omega^{-1}} \leq \|v\|\cdot\|\Omega^{-1}\|_2^{1/2}$. By the triangle inequality,

$$\begin{aligned}\sqrt{\hat{T}_i^{\Omega}}=\min_{t\leq 0}\|\hat{D}-t\|_{\Omega^{-1}} \leq&\|\hat{D}-D\|_{\Omega^{-1}}+\min_{t\leq 0}\|D-t\|_{\Omega^{-1}}\\ =&\|\hat{D}-D\|_{\Omega^{-1}}+\sqrt{T_i^{\text{QLR}}} \leq \|\hat{D}-D\|\sqrt{\|\Omega^{-1}\|_2}+\sqrt{T_i^{\text{QLR}}}.\end{aligned}$$

Taking squares and using $T_i^{\text{QLR}} \leq \|\Omega^{-1}\|_2\|D\|^2$ and $\|\hat{D}-D\| \leq 1$ gives

$$\hat{T}_i^{\Omega} \leq T_i^{\text{QLR}}+(2\|D\|+1)\|\Omega^{-1}\|_2\|\hat{D}-D\|.$$

Reversing the roles of $\hat{D}$ and D gives

$$\begin{aligned}T_i^{\text{QLR}} \leq&\hat{T}_i^{\Omega}+(2\|\hat{D}\|+1)\|\Omega^{-1}\|_2\|\hat{D}-D\|\\ \leq&\hat{T}_i^{\Omega}+(2\|D\|+2\|\hat{D}-D\|+1)\|\Omega^{-1}\|_2\|\hat{D}-D\|\\ \leq&\hat{T}_i^{\Omega}+(2\|D\|+3)\|\Omega^{-1}\|_2\|\hat{D}-D\|,\end{aligned}$$

where the third inequality follows by $\|\hat{D}-D\| \leq 1$. The assertion follows by combining the inequalities. □

Lemma D.7. *Suppose that the probability measure P satisfies Assumption 1. For $h,h' \in \mathbb{G}\setminus\{g_i^0\}$ let*

$$(\Omega_i^*)_{h,h'}=\frac{T^{-1}\sum_{t=1}^{T}(d_{it}(g_i^0,h)-\bar{d}_i(g_i^0,h))(d_{it}(g_i^0,h')-\bar{d}_i(g_i^0,h'))}{\sigma_i^2 s_{i,T}(g_i^0,h)s_{i,T}(g_i^0,h')}.$$

There is a constant C depending only on K_β and G such that for $0 < c < 1$

(i)
$$P\Big(\max_{1\le i\le N}\max_{h,h'\in\mathbb{G}\setminus\{g_i^0\}}\big|(\Omega_i^*)_{h,h'} - (\Omega_i(g_i^0))_{h,h'}\big| > CT^{-(1-c)/2}(\log N)B_{N,T,4}^2\Big) \le CT^{-c},$$

(ii)
$$P\Big(T^{-1/2}\max_{1\le i\le N}\big|D_i(g_i^0,h)\big| > C\left(T^{-1/2}\sqrt{\log N} + T^{-3/4}B_{N,T,4}\log N\right)\Big) \le N^{-1} + C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4.$$

Proof.

Proof of (i) Decompose

$$\begin{aligned}&\frac{T^{-1}\sum_{t=1}^T(d_{it}(h)-\bar{d}_i(h))(d_{it}(h')-\bar{d}_i(h'))}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')} - (\Omega_i)_{h,h'}\\ =&\frac{T^{-1}\sum_{t=1}^T\left(d_{it}(h)d_{it}(h')-\mathbb{E}_P[d_{it}(h)d_{it}(h')]\right)}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')} - \left(\frac{\bar{d}_i(h)}{\sigma_i s_{i,T}(h)}\right)\left(\frac{\bar{d}_i(h')}{\sigma_i s_{i,T}(h')}\right).\end{aligned}$$

Below, we show that

$$P\Bigg(\max_{1\le i\le N}\left|\frac{T^{-1}\sum_{t=1}^T\left(d_{it}(h)d_{it}(h')-\mathbb{E}_P[d_{it}(h)d_{it}(h')]\right)}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')}\right| > C_1 B_{N,T,4}^2 T^{-(1-c)/2}\log N\Bigg) \le 2T^{-c}, \tag{22}$$

$$P\left(\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^T\frac{d_{it}(h)}{\sigma_i s_{i,T}(h)}\right| > C_2\left(T^{-1/4}\sqrt{\log N}+T^{-3/4}\log(N)\right)B_{N,T,4}\right) \le 2T^{-2} \tag{23}$$

where C_1 and C_2 are constants that depend only on K_β. For

$$x = (2C_1\vee C_2)T^{-(1-c)/2}(\log N)B_{N,T,4}^2(P) + C_2 T^{-3/2}(\log^2 N)B_{N,T,4}^2(P)$$

and a constant C_3 depending only on K_β and G

$$\begin{aligned}&P\Bigg(\left|\frac{T^{-1}\sum_{t=1}^T\left(d_{it}(h)d_{it}(h')-\mathbb{E}_P[d_{it}(h)d_{it}(h')]\right)}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')} - \left(\frac{\bar{d}_i(h)}{\sigma_i s_{i,T}(h)}\right)\left(\frac{\bar{d}_i(h')}{\sigma_i s_{i,T}(h')}\right)\right| > 2x^2\Bigg)\\ \le&P\Bigg(\left|\frac{T^{-1}\sum_{t=1}^T\left(d_{it}(h)d_{it}(h')-\mathbb{E}_P[d_{it}(h)d_{it}(h')]\right)}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')}\right| > x^2\Bigg) + \sum_{h\in\mathbb{G}}P\left(\left|\frac{\bar{d}_i(h)}{\sigma_i s_{i,T}(h)}\right| > x\right)\\ \le&C_3T^{-c},\end{aligned}$$

where the last inequality follows from (22) and (23). The assertion of the lemma follows. It remains to establish the inequalities (22) and (23). Write

$$U_{it}(h,h') = \left(d_{it}(h)d_{it}(h')-\mathbb{E}_P[d_{it}(h)d_{it}(h')]\right)/\left(\sigma_i^2 s_{i,T}(h)s_{i,T}(h')\right).$$

By

$$
\begin{aligned}
\mathbb{E}_P[U_{it}(h,h')^2] \leq & \max_{1\leq t\leq T} \mathbb{E}_P\left[\max_{1\leq i\leq N} |d_{it}(h)d_{it}(h')|^2/\sigma_i^2\right] / \left(s_{i,T}^2(h)s_{i,T}^2(h')\right) \\
\leq & \max_{1\leq t\leq T} \mathbb{E}_P\left[\max_{1\leq i\leq N} \left(|u_{it}/\sigma_i|^4 \, \|x_{it}\|^4 \|\delta_t(g_i^0,h)\|^2 \|\delta_t(g_i^0,h)\|^2\right)\right] / \underline{s}_{N,T}^4(P) \\
\leq & 16K_\beta^2 B_{N,T,4}^4(P)
\end{aligned}
$$

we have $\mathbb{E}[\max_{1\leq i\leq N}\max_{1\leq t\leq T}|U_{it}(h,h')|^2] \leq 16K_\beta^2 T B_{N,T,4}^4$ and $\mathbb{E}[U_{it}^2(h,h')] \leq 16K_\beta^2 B_{N,T,4}^4$. By Lemma A.3 in Chernozhukov, Chetverikov, and Kato (2014) there is a universal constant K such that for $C_4 = 32K_\beta^2 K$

$$
\mathbb{E}_P\left[\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^T (U_{it}(h,h') - \mathbb{E}_P[U_{it}(h,h')])\right|\right] \leq C_4 B_{N,T,4}^2 (\log N)/\sqrt{T}.
$$

Thus, by Lemma A.2 in Chernozhukov, Chetverikov, and Kato (2014) for every $r > 0$ and a universal constant K_2

$$
\begin{aligned}
& P\left(\max_{1\leq i\leq N}\left|U_{it}(h,h') - \mathbb{E}_P[U_{it}(h,h')]\right| \geq 2CB_{N,T,4}^2(\log N)/\sqrt{T} + r\right) \\
\leq & e^{-Tr^2/(48K_\beta^2 B_{N,T,4}^4)} + K_2 16K_\beta^2 r^{-2} T^{-1} B_{N,T,4}^4.
\end{aligned}
$$

Taking $r = C_1 T^{-(1-c)/2} B_{N,T,4}^2$ for $0 < c < 1$ and $C_1 = 4(\sqrt{K_2} + \sqrt{3})K_\beta \vee C$ then yields

$$
P\left(\frac{T^{-1}\sum_{t=1}^T (d_{it}(h)d_{it}(h') - \mathbb{E}_P[d_{it}(h)d_{it}(h')])}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')} > C_1 B_{N,T,4}^2 T^{-(1-c)/2}\log N\right) \leq 2T^{-c}.
$$

By Hölder's inequality

$$
\mathbb{E}_P\left[\max_{1\leq i\leq N}\max_{1\leq t\leq T}\left|\frac{d_{it}(h)}{\sigma_i s_{i,T}(h)}\right|^2\right] \leq \sqrt{\mathbb{E}_P\left(\max_{1\leq i\leq N}\max_{1\leq t\leq T}\left|\frac{d_{it}(h)}{\sigma_i s_{i,T}(h)}\right|^4\right)} \leq \sqrt{T}4K_\beta B_{N,T,4}^2.
$$

Thus, by Lemma A.3 in Chernozhukov, Chetverikov, and Kato (2014) for a universal constant K

$$
\mathbb{E}_P\left[\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^T \frac{d_{it}(h)}{\sigma_i s_{i,T}(h)}\right|\right] \leq K\left(T^{-1/2}\sqrt{\log N} + 2T^{-3/4}\sqrt{K_\beta}B_{N,T,4}\log N\right).
$$

Then, by Lemma A.2 in Chernozhukov, Chetverikov, and Kato (2014) for all $r > 0$ and a universal constant K_4

$$
\begin{aligned}
& P\Bigg(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^T \frac{d_{it}(h)}{\sigma_i s_{i,T}(h)}\right| \\
& \qquad > 2K\left(T^{-1/2}\sqrt{\log N} + 2T^{-3/4}\sqrt{K_\beta}B_{N,T,4}\log N\right) + r\Bigg) \\
\leq & e^{-Tr^2/3} + K_4 r^{-4} T^{-3} B_{N,T,4}^4.
\end{aligned} \tag{24}
$$

Now, taking $r = 2\sqrt{K_\beta}K_4^{1/4}T^{-1/4}B_{N,T,4}$ and noting that $B_{N,T,4} \geq 1$ yields

$$P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\frac{d_{it}(h)}{\sigma_i s_{i,T}(h)}\right| > C_2\left(T^{-1/2}\sqrt{\log N} + T^{-3/4}\log(N)\right)B_{N,T,4}\right) \leq 2T^{-2},$$

where C_2 is a constant that depends only on K_β.

Proof of (ii): Taking $r = 3T^{-1/2}\sqrt{\log N}$ in (24) gives

$$\begin{aligned}&P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\frac{d_{it}(h)}{\sigma_i s_{i,T}(h)}\right| > C_5\left(T^{-1/2}\sqrt{\log N} + T^{-3/4}(\log N)B_{N,T,4}\right)\right)\\ \leq &N^{-1} + \left(T^{-1/4}B_{N,T,4}/(\log N)\right)^4.\end{aligned}$$

□

Lemma D.8. *Suppose that the probability measure P satisfies Assumption 1. Then, there is a constant C depending only on K_β and G such that for $0 < c < 1$ and*

$$\begin{aligned}\zeta_{N,T} =&\gamma_{N,T,8}\big(T^{-(1-c)/4}\sqrt{\log N}B_{N,T,8}^2 + D_{N,T,4}\big)\big(1 + T^{-(1-c)/4}\sqrt{\log N}B_{N,T,4}\big)\\ &+ \gamma_{N,T,8}^2\big(T^{-(1-c)/2}(\log N)B_{N,T,8}^4 + D_{N,T,4}^2\big),\end{aligned}$$

we have

(i)
$$\begin{aligned}P\Bigg(\max_{1\leq i\leq N}\Bigg|&\frac{1}{T}\sum_{t=1}^{T}\frac{(\hat{d}_{it}(g_i^0,h) - \bar{\hat{d}}_i(g_i^0,h))(\hat{d}_{it}(g_i^0,h') - \bar{\hat{d}}_i(g_i^0,h'))}{\sigma_i^2 s_{i,T}(g_i^0,h)s_{i,T}(g_i^0,h')}\\ &- \frac{1}{T}\sum_{t=1}^{T}\frac{(d_{it}(g_i^0,h) - \bar{d}_i(g_i^0,h))(d_{it}(g_i^0,h') - \bar{d}_i(g_i^0,h'))}{\sigma_i^2 s_{i,T}(g_i^0,h)s_{i,T}(g_i^0,h')}\Bigg|\\ &> C\zeta_{N,T}\Bigg) \leq CT^{-c},\end{aligned}$$

(ii)
$$\begin{aligned}P\Bigg(T^{-1/2}\max_{1\leq i\leq N}&\left|D_i(g_i^0,h) - \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\hat{d}_i(g_i^0,h)}{\sigma_i s_{i,T}(g_i^0,h)}\right|\\ &> C\gamma_{N,T,8}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right)\Bigg) \leq CT^{-c}.\end{aligned}$$

Suppose that, additionally, $\zeta_{N,T} \vee T^{-(1-c)/4}\sqrt{\log N}B_{N,T,4} \leq 1$. Then,

(iii)
$$\begin{aligned}P\Bigg(\max_{1\leq i\leq N}\left|\hat{D}_i(h) - \tilde{D}_i(h)\right| &> C\gamma_{N,T,8}\sqrt{T}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right)\\ &\quad + C\zeta_{N,T}\left(1 + T^{-1/4}B_{N,T,4}\sqrt{\log N}\right)\sqrt{\log N}\Bigg)\\ \leq N^{-1} + CT^{-c} &+ C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4,\end{aligned}$$

(iv) $$P\Big(\max_{1\le i\le N}\Big|\hat{D}_i(h)-D_i(h)\Big| > C\sqrt{T}\gamma_{N,T,8}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N}+D_{N,T,2}\right) + C\left(\zeta_{N,T}+T^{-(1-c)/2}(\log N)B^2_{N,T,4}\right) \times\left(1+T^{-1/4}B_{N,T,4}\sqrt{\log N}\right)\sqrt{\log N}\Big) \le N^{-1}+CT^{-c}+C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4.$$

Proof. Throughout the proof, let C denote a generic constant that depends only on K_β and G.

Proof of (i): Bound as follows

$$\begin{aligned}
&\Big|\frac{1}{T}\sum_{t=1}^{T}(\hat{d}_{it}(h)-\bar{\hat{d}}_i(h))(\hat{d}_{it}(h')-\bar{\hat{d}}_i(h'))-\frac{1}{T}\sum_{t=1}^{T}(d_{it}(h)-\bar{d}_i(h))(d_{it}(h')-\bar{d}_i(h'))\Big|\\
\le&\Big|\frac{1}{T}\sum_{t=1}^{T}\big(\hat{d}_{it}(h)-d_{it}(h)-(\bar{\hat{d}}_i(h)-\bar{d}_i(h))\big)\big(\hat{d}_{it}(h')-d_{it}(h')-(\bar{\hat{d}}_i(h')-\bar{d}_i(h'))\big)\Big|\\
&+\Big|\frac{1}{T}\sum_{t=1}^{T}\big(d_{it}(h)-\bar{d}_i(h)\big)\big(\hat{d}_{it}(h')-d_{it}(h')-(\bar{\hat{d}}_i(h')-\bar{d}_i(h'))\big)\Big|\\
&+\Big|\frac{1}{T}\sum_{t=1}^{T}\big(d_{it}(h')-\bar{d}_i(h')\big)\big(\hat{d}_{it}(h)-d_{it}(h)-(\bar{\hat{d}}_i(h)-\bar{d}_i(h))\big)\Big|\\
\le&\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\hat{d}_{it}(h)-d_{it}(h)\big)^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\hat{d}_{it}(h')-d_{it}(h')\big)^2}\\
&+\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(d_{it}(h)-\bar{d}_i(h)\big)^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\hat{d}_{it}(h')-d_{it}(h')\big)^2}\\
&+\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(d_{it}(h')-\bar{d}_i(h')\big)^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\hat{d}_{it}(h)-d_{it}(h)\big)^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Big|\frac{1}{T}\sum_{t=1}^{T}\frac{(\hat{d}_{it}(h)-\bar{\hat{d}}_i(h))(\hat{d}_{it}(h')-\bar{\hat{d}}_i(h'))}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')}-\frac{1}{T}\sum_{t=1}^{T}\frac{(d_{it}(h)-\bar{d}_i(h))(d_{it}(h')-\bar{d}_i(h'))}{\sigma_i^2 s_{i,T}(h)s_{i,T}(h')}\Big|\\
\le&\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\big(\hat{d}_{it}(h)-d_{it}(h)\big)^2}{\sigma_i^2\underline{s}^2_{N,T}}}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\big(\hat{d}_{it}(h')-d_{it}(h')\big)^2}{\sigma_i^2\underline{s}^2_{N,T}}}\\
&+\max_{h,h'\in\mathbb{G}}2\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\big(d_{it}(h)-\bar{d}_i(h)\big)^2}{\sigma_i^2 s^2_{i,T}(h)}}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\big(\hat{d}_{it}(h')-d_{it}(h')\big)^2}{\sigma_i^2\underline{s}^2_{N,T}}}\\
\le&\gamma_{N,T,8}\big(T^{-(1-c)/4}\sqrt{\log N}B^2_{N,T,8}+D_{N,T,4}\big)\big(1+T^{-(1-c)/4}\sqrt{\log N}B_{N,T,4}\big)\\
&+\gamma^2_{N,T,8}\big(T^{-(1-c)/2}(\log N)B^4_{N,T,8}+D^2_{N,T,4}\big).
\end{aligned}$$

where the last inequality follows from Lemma D.7 and

$$P\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{d}_{it}(h)-d_{it}(h)}{\sigma_i s_{i,T}(h)}\right)^2 > C\gamma_{N,T,8}^2\left(T^{-(1-c)/2}B_{N,T,8}^4(\log N)+D_{N,T,4}^2\right)\right)\leq CT^{-c}. \tag{25}$$

We now prove (25). For the following calculations note that

$$\|\hat{\delta}(g_i^0,h)-\delta(g_i^0,h)\|^2\leq 2\left(\|\hat{\beta}_{g_i^0}-\beta_{g_i^0}\|^2+\|\hat{\beta}_h-\beta_h\|^2\right)\leq 4\max_{g\in\mathbb{G}}\|\hat{\beta}_g-\beta_g\|^2$$

and, since the matrix norm $\|\cdot\|_2$ is an induced norm and $\|x_{it}\|=\sqrt{x_{it}'x_{it}}$,

$$\|x_{it}x_{it}'\|_2=\sup_{\|y\|=1}\|x_{it}x_{it}'y\|\leq\frac{\|x_{it}x_{it}'x_{it}\|}{\|x_{it}\|}=\|x_{it}\|^2.$$

Decompose $\hat{d}_{it}(h)-d_{it}(h)$ as follows

$$\begin{aligned}&\hat{d}_{it}(h)-d_{it}(h)\\=&-u_{it}x_{it}'(\hat{\delta}_t(g_i^0,h)-\delta_t(g_i^0,h))\\&+(\hat{\beta}_{g_i^0,t}-\beta_{g_i^0,t})'(x_{it}x_{it}')(\hat{\delta}_t(g_i^0,h)-\delta_t(g_i^0,h))+(\hat{\beta}_{g_i^0,t}-\beta_{g_i^0,t})'(x_{it}x_{it}')\delta_t(g_i^0,h).\end{aligned}$$

By the inequality $(a+b+c)^2\leq 3(a^2+b^2+c^2)$,

$$\begin{aligned}\left(\frac{\hat{d}_{it}(h)-d_{it}(h)}{\sigma_i}\right)^2\leq&3\left|\frac{u_{it}}{\sigma_i}\right|^2\|x_{it}\|^2\|\hat{\delta}_t(g_i^0,h)-\delta_t(g_i^0,h)\|^2\\&+3\sigma_i^{-2}\|\hat{\beta}_{g_i^0,t}-\beta_{g_i^0,t}\|^2\|x_{it}\|^4\|\hat{\delta}_t(g_i^0,h)-\delta_t(g_i^0,h)\|^2\\&+3\sigma_i^{-2}\|\hat{\beta}_{g_i^0,t}-\beta_{g_i^0,t}\|^2\|x_{it}\|^4\|\delta_t(g_i^0,h)\|^2.\end{aligned}$$

Let $V_{it}=\left(|u_{it}/\sigma_i|^2\|x_{it}\|^2+\|x_{it}\|^4/\sigma_i^4\right)/\underline{s}_{N,T}^2$. Below, we show that for $0<c<1$

$$P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left(V_{it}^2-\mathbb{E}_P[V_{it}^2]\right)\right|>CT^{-(1-c)/2}B_{N,T,8}^4(\log N)\right)\leq CT^{-c}. \tag{26}$$

Now, by the Cauchy-Schwarz inequality

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{d}_{it}(h)-d_{it}(h)}{\sigma_i s_{i,T}(h)}\right)^2\\\leq&C\left\{\max_{g\in\mathbb{G}}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\beta}_{g,t}-\beta_{g,t}\right\|^4\right)^{1/2}+\max_{g\in\mathbb{G}}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\beta}_{g,t}-\beta_{g,t}\right\|^8\right)^{1/2}\right\}\\&\times\left(\frac{1}{T}\sum_{t=1}^{T}\left(\left|\frac{u_{it}}{\sigma_i}\right|^4\|x_{it}\|^4+\|x_{it}\|^8/\sigma_i^4\right)/\underline{s}_{N,T}^4\right)^{1/2}\end{aligned}$$

$$\leq C(\gamma_{N,T,8}^2 + \gamma_{N,T,8}^4)\left(\frac{1}{T}\sum_{t=1}^{T}\left(V_{it}^2 - \mathbb{E}_P\left[V_{it}^2\right]\right) + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_P\left[V_{it}^2\right]\right)^{1/2}.$$

Together with (26) this implies (25). It remains to prove (26). Note that $\mathbb{E}_P[V_{it}^2] \leq B_{N,T,8}^8$ and $\mathbb{E}_P[\max_{1\leq i\leq T}\max_{1\leq i\leq N} V_{it}^2] \leq TB_{N,T,8}^8$. By Lemma A.3 in Chernozhukov, Chetverikov, and Kato (2014) there is a universal constant K such that

$$\mathbb{E}_P\left[\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}(V_{it}^2 - \mathbb{E}_P[V_{it}^2])\right|\right] \leq KB_{N,T,8}^4\frac{\log N}{\sqrt{T}}.$$

Then, by Lemma A.2 in Chernozhukov, Chetverikov, and Kato (2014) for every $r > 0$

$$P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}(V_{it}^2 - \mathbb{E}_P[V_{it}^2])\right| > 2KB_{N,T,4}^2\log N/\sqrt{T} + r\right)$$
$$\leq e^{-Tr^2/(3B_{N,T,8}^8)} + Kr^{-2}T^{-1}B_{N,T,8}^8.$$

Then, taking $r = T^{-(1-c)/2}B_{N,T,8}^4$ for $0 < c < 1$ yields (26).

Proof of (ii): By slightly modifying the arguments above, we can prove

$$\left|\frac{1}{T}\sum_{t=1}^{T}\frac{\hat{d}_{it}(h) - d_{it}(h)}{\sigma_i s_{i,T}(h)}\right|$$
$$\leq C(\gamma_{N,T,4} + \gamma_{N,T,4}^2)\left(\frac{1}{T}\sum_{t=1}^{T}\left(V_{it} - \mathbb{E}_P\left[V_{it}\right]\right) + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_P\left[V_{it}\right]\right)^{1/2}.$$

In addition, for $0 < c < 1$,

$$P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left(V_{it} - \mathbb{E}_P[V_{it}]\right)\right| > CT^{-(1-c)/2}B_{N,T,4}^2(\log N)\right) \leq CT^{-c}$$

from whence the conclusion follows.

Proof of (iii): Define

$$S_{i,T}^{\Delta}(h) = \left(\frac{\hat{S}_{i,T}(h) - S_{i,T}(h)}{\sigma_i s_{i,T}(h)}\right)\frac{S_{i,T}(h)}{\sigma_i s_{i,T}(h)}.$$

By the inequality $|a-b| \leq |a-b|/(\sqrt{a}+\sqrt{b} \leq |a-b|/\sqrt{a}$ and (i) of the lemma we have

$$S_{i,T}^{\Delta}(h) \leq \left|\left(\hat{S}_{i,T}^2(h)/(\sigma_i s_{i,T}(h))\right)^2 - \left(S_{i,T}^2(h)/(\sigma_i s_{i,T}(h))\right)^2\right| \leq C_2\zeta_{N,T}$$

uniformly over $i = 1,\ldots,N$ on a set of probability less than CT^{-c}. By the inequality $|\sqrt{a}-1| \leq |a-1|$ and Lemma D.7 we have

$$\left|S_{i,T}^2(h)/(\sigma_i s_{i,T}(h)) - 1\right| \leq \left|\left(S_{i,T}^2(h)/(\sigma_i s_{i,T}(h))\right)^2 - 1\right| \leq C_1T^{-(1-c)/2}(\log N)B_{N,T,4}^2$$

uniformly over $i = 1, \dots, N$ on a set of probability less than CT^{-c}. By Lemma D.7(ii)

$$\left|D_i(g_i^0, h)\right| \le C\left(\sqrt{\log N} + T^{-1/4}B_{N,T,4}\log N\right)$$

uniformly over $i = 1, \dots, N$ on a set of probability less than $N^{-1} + C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4$. Now, decompose

$$\begin{aligned}
\hat{D}_i(h) - \tilde{D}_i(h) =& \frac{\sigma_i s_{i,T}(h)}{\hat{S}_{i,T}(h)}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\hat{d}_i(g_i^0, h)}{\sigma_i s_{i,T}(g_i^0, h)} - D_i(g_i^0, h)\right)\\
&- \frac{\hat{S}_{i,T} - S_{i,T}}{S_{i,T}(h)\hat{S}_{i,T}(h)}\sigma_i s_{i,T}(h)D_i(h)\\
=& \frac{S_{i,T}(h)/(\sigma_i s_{i,T}(h))}{S_{i,T}^{\Delta}(h) + S_{i,T}(h)/(\sigma_i s_{i,T}(h))}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\hat{d}_i(g_i^0, h)}{\sigma_i s_{i,T}(g_i^0, h)} - D_i(g_i^0, h)\right)\\
&- \frac{S_{i,T}^{\Delta}}{\big(S_{i,T}^{\Delta} + S_{i,T}^2(h)/(\sigma_i s_{i,T}(h))^2\big)S_{i,T}^2(h)/(\sigma_i s_{i,T}(h))}D_i(h).
\end{aligned}$$

In conjunction with part (ii) of the lemma, this decomposition implies

$$\begin{aligned}
\max_{1\le i\le N}\left|\hat{D}_i(h) - \tilde{D}_i(h)\right| \le& C\gamma_{N,T,8}\sqrt{T}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,4}^2\right)\\
&+ C\zeta_{N,T}\left(\sqrt{\log N} + T^{-1/4}B_{N,T,4}\log N\right)
\end{aligned}$$

with probability less than $CT^{-c} + N^{-1} + C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4$.

Proof of (iv): Write

$$\tilde{D}_i(h) - D_i(h) = -\left(S_{i,T}/s_{i,T} - 1\right)\left(S_{i,T}/s_{i,T}\right)^{-1}D_i(h).$$

The bounds derived in the proof of part (iii) imply

$$\max_{1\le i\le N}\left|\tilde{D}_i(h) - D_i(h)\right| \le C_1 T^{-(1-c)/2}(\log N)B_{N,T,4}^2\left(\sqrt{\log N} + T^{-1/4}B_{N,T,4}\log N\right)$$

with probability less than $CT^{-c} + N^{-1} + C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4$. The conclusion now follows from the triangle inequality and part (iii) of the lemma. □

Lemma D.9. *(Simultaneous anti-concentration) Let $\{V_i\}_{i=1}^N$ denote a collection of nonsingular $(p\times p)$-variance matrices, and let $\{W_i\}_{i=1}^N$ denote a collection of independent random variables with marginal distribution $W_i \sim \tilde{\chi}^2(V_i)$. For positive constants $\underline{a}$ and $\bar{a}$, let $\mathbb{S} = [\underline{a}\log N, \bar{a}\log N]$. Then, for each $\tau > 0$ there are constants C and N_0 that depend only on $\underline{a}$, $\bar{a}$, τ and p such that for all ϵ with $N^{-\tau} < \epsilon < \underline{a}/2(\log N)$ we have*

$$\sup_{(s_1,\dots,s_N)\in\mathbb{S}^N} P\left(|\max_{1\le i\le N} W_i - s_i| \le \epsilon\right) \le C\epsilon + N^{-1}.$$

Proof. For a collection of nonsingular $p\times p$ covariance matrices $(V_i)_{i=1}^N$ let $W_i \sim \tilde{\chi}^2(V_i)$. Let $(U_{i,j})_{j=1}^k$ denote a collection of chi-squared random variables with $U_{i,j} \sim \chi_j^2$ and $U_{i,j} \perp U_{i,k}$ for

$j \neq k$. Let $\bar{W}_i$ denote the random function $\bar{W}_i(d) = \sum_{j=1}^p 1\{d = j\}U_{i,j}$. and let $(D_i)_{i=1}^N$ denote random variables that are supported on $\{0, \ldots, p\}$ and satisfy $P(D_i = d) = w(p, p-d, V_i)$ for $d = 0, \ldots, p$. This construction ensures that $\mathcal{L}(W_i) = \mathcal{L}(\bar{W}_i(D_i))$. Let $(D_i^*)_{i=1}^N$ denote random variables that are supported on $\{1, \ldots, p\}$ and satisfy $P(D_i^* = 1) = w(p, p, V_i) + w(p, p-1, V_i)$ and $P(D_i^* = d) = w(p, p-d, V_i)$ for $d = 0, \ldots, p$ and define $W_i^* = \bar{W}_i(D_i^*)$. For $\epsilon < \underline{a}/2(\log N)$,

$$\begin{aligned} P\Big(|\max_{1\leq i\leq N} (W_i - s_i)| \leq \epsilon\Big) =& P\Big(|\max_{1\leq i\leq N} (\bar{W}_i(D_i) - s_i)| \leq \epsilon\Big) \\ \leq& P\Big(|\max_{1\leq i\leq N} (\bar{W}_i(D_i^*) - s_i)| \leq \epsilon\Big) = P\Big(|\max_{1\leq i\leq N} (W_i^* - s_i)| \leq \epsilon\Big), \end{aligned}$$

where the inequality holds since the upper bound on ϵ implies $\bar{W}_i(0) - s_i = 0 - s_i < -\epsilon$ so that units i with $D_i = 0$ do not contribute any probability mass. Then, Lemma D.16 gives C and N_0 such that for $N^{-\tau} < \epsilon < \underline{a}/2(\log N)$ and $N \geq N_0$,

$$\begin{aligned} P\left(|\max_{1\leq i\leq N} (W_i - s_i)| \leq \epsilon\right) \leq& P\left(|\max_{1\leq i\leq N} (W_i^* - s_i)| \leq \epsilon\right) \\ \leq& \sum_{(d_1,\ldots,d_N)\in\{1,\ldots,p\}^N} P\Big(D_1^* = d_1, \ldots, D_N^* = d_N\Big) \\ & \times P\left(|\max_{1\leq i\leq N} (U_{i,d_i} - s_i)| \leq \epsilon\right) \leq C\epsilon + 2N^{-1}. \end{aligned}$$

□

Lemma D.10. *Let $\nu(N) \geq 1$ denote a sequence that converges to infinity, and let $c_N(\alpha)$ denote the $(1-\alpha/N)$-quantile of the t-distribution with $\nu(N)$ degrees of freedom. Suppose that $(\log N)/\nu(N) \to 0$. For each $\epsilon > 0$ and $0 < \underline{\alpha} < 1$, there is a threshold N_0 such that for $N \geq N_0$*

$$\sup_{\underline{\alpha}\leq\alpha<1} c_N(\alpha) \leq \sqrt{2(1+\epsilon)\log(N/\alpha)}.$$

Proof. For notational convenience, write $\nu = \nu(N)$. We prove the bound for $\alpha = \underline{\alpha}$ and write $c_N = c_N(\alpha)$. Then, the uniformity follows from the monotonicity of the distribution function. Clearly, $c_N \to \infty$ so that we can take $c_N \geq 1$, provided that N is large enough. The density function of the t-distribution with ν degrees of freedom is given by $f_\nu(x) = c(\nu)\left(1 + x^2/\nu\right)^{-\frac{\nu+1}{2}}$, where

$$c(\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \to \frac{1}{\sqrt{2\pi}}$$

as $\nu \to \infty$. It follows that there is a universal constant C such that $c(\nu) \leq C$. We first show that $c_N^2/\nu = O(1)$. The proof is by contradiction. Suppose that $\limsup_{N\to\infty} c_N^2/\nu = \infty$. Applying Theorem 1 in Soms 1976 with $n = 1$ yields

$$1 - F_\nu(c_N) \leq f_\nu(c_N)\frac{1}{c_N}\left(1 + \frac{c_N^2}{\nu}\right). \tag{27}$$

This implies that

$$\frac{\alpha}{N} \leq c(\nu)\left(1 + \frac{c_N^2}{\nu}\right)^{-\frac{\nu+1}{2}}\left(1 + \frac{c_N^2}{\nu}\right) \leq C\left(1 + \frac{c_N^2}{\nu}\right)^{-\frac{\nu-1}{2}}.$$

Taking logs and rearranging gives

$$\frac{\log(N/\alpha)}{\nu} \geq \frac{1}{2}\frac{\nu-1}{\nu}\left(\log\left(1+\frac{c_N^2}{\nu}\right)-C\right).$$

The left-hand side of the inequality vanishes under the assumptions of the lemma, whereas a subsequence of the right-hand side diverges to infinity. This establishes that the inequality is impossible and therefore $c_N^2/\nu = O(1)$. This implies that there exists a constant b such that

$$1 < b \leq \left(1+\frac{c_N^2}{\nu}\right)^{\frac{\nu}{c_N^2}} \leq e,$$

so that we can take

$$\left(\left(1+\frac{c_N^2}{\nu}\right)^{\frac{\nu}{c_N^2}}\right)^{-1} \leq e^{-\frac{\nu}{\nu+1}(1+\epsilon^*/2)^{-1}}$$

for a positive ϵ^*. Then,

$$f_\nu(c_N) \leq C\left[\left(1+\frac{c_N^2}{\nu}\right)^{\frac{\nu}{c_N^2}}\right]^{-\frac{c_N^2}{2}\left[\frac{\nu+1}{\nu}\right]} \leq C\exp\left(-\frac{c_N^2}{2}(1+\epsilon^*/2)^{-1}\right).$$

Take N large enough that

$$\frac{1}{1+\epsilon^*/2} - \frac{4\log c_N}{c_N^2} > \frac{1}{1+\epsilon^*}.$$

Then, the right-hand side of (27) can be bounded by

$$\begin{aligned} C\exp\left(-\frac{c_N^2}{2}(1-\epsilon^*/2)^{-1}\right)\left(1+\frac{c_N^2}{\nu}\right) \leq& 2C\exp\left(-\frac{c_N^2}{2}\left((1+\epsilon^*/2)^{-1}-\frac{4\log c_N}{c_N^2}\right)\right) \\ \leq& 2C\exp\left(-\frac{c_N^2}{2}\left(1+\epsilon^*\right)^{-1}\right). \end{aligned}$$

Plugging in $1-F_\nu(c_N) = \alpha/N$ and taking logs gives

$$\begin{aligned} c_N^2 \leq& (1+\epsilon^*)\log\left(N/\alpha\right) + \log(2C) \\ \leq& 2(1+\epsilon^*)\log\left(N/\alpha\right)\left(1+\frac{1}{2(1+\epsilon^*)}\frac{\log(2C)}{\log(N/\alpha)}\right). \end{aligned}$$

Hence, there is a constant C such that $c_N^2 \leq C\log(N/\alpha)$. Using this inequality, we can now verify that $c_N^2/\nu \to 0$ so that

$$\left(1+\frac{c_N^2}{\nu}\right)^{\frac{\nu}{c_N^2}} \to e,$$

allowing us to take $\epsilon^* = \epsilon/2$ for sufficiently large N. Taking N large enough that

$$(1+\epsilon/2)\left(1+\frac{1}{2(1+\epsilon/2)}\frac{\log(2C)}{\log(N)}\right) \leq 1+\epsilon$$

yields $c_N^2 \leq 2(1+\epsilon)\log(N/\alpha)$. □

Lemma D.11. *For $\nu \geq 1$, let F_ν and f_ν denote the distribution and density function of a t-distributed random variable with ν degrees of freedom. For $x^2 > 2$*

$$f_\nu(x) < 2x\left(1-F_\nu(x)\right).$$

Proof. Applying Theorem 1 in Soms (1976) with $n = 2$ yields the inequality

$$1-F_\nu(x) \geq (1+x^2/\nu)\left(1-\frac{\nu}{(\nu+2)x^2}\right)f_\nu(x)/x.$$

Now, $x^2 > 2$ implies

$$1-F_\nu(x) > \left(1-\frac{1}{2}\right)f_\nu(x)/x.$$

□

Lemma D.12. *Let $\xi_1,\ldots,\xi_T$ be independent centered random variables with $E(\xi_t^2) = 1$ and $E(|\xi_t|^{2+\nu}) < \infty$ for all $1 \leq t \leq T$ where $0 < \nu \leq 1$. Let $S_T = \sum_{t=1}^T \xi_t$, $V_T^2 = \sum_{t=1}^T \xi_t^2$ and $D_{T,\nu} = (T^{-1}\sum_{t=1}^T E(|\xi_t|^{2+\nu}))^{1/(2+\nu)}$. Then uniformly in $0 \leq x \leq T^{\nu/(2(2+\nu))}/D_{T,\nu}$,*

$$\left|\frac{\Pr(S_T/V_T \geq x)}{1-\Phi(x)} - 1\right| \leq KT^{-\nu/2}D_{T,\nu}^{2+\nu}(1+x)^{2+\nu}.$$

Proof. This lemma is first proved by Jing, Shao, and Wang (2003). Here we use the version by Chernozhukov, Chetverikov, and Kato (2014, Lemma A.1), which is based on de la Pena, Lai, and Shao (2009, Theorem 7.4). □

Lemma D.13 (Properties of $\tilde{\chi}^2$-distribution)**.** *Let W denote a random variable with $\tilde{\chi}^2(V)$ distribution for a nondegenerate $p \times p$ covariance matrix. For $j = 0,\ldots,p$ let $w(p,j,V)$ denote the weight function for the $\bar{\chi}^2$-distribution defined by Kudo (1963) and Nüesch (1966).*

1. *(Weights define a probability distribution) For $j = 0,\ldots,p$, $w(p,j,V) > 0$ and*

$$\sum_{j=1}^{p} w(p,j,V) = 1.$$

   Moreover, $w(p,j,V) \leq 1/2$ for $j = 1,\ldots,N$.

2. *(Tail probabilities) Let $(U_j)_{j=1}^p$ denote chi-squared random variables, $U_j \sim \chi_j^2$. For all $c \geq 0$*

$$P(W \geq c) = \sum_{j=1}^{p} w(p,p-j,V)P(U_j \geq c).$$

3. *(Mixture representation) Let* $(U_j)_{j=1}^p$ *denote independent chi-squared random variables such that* $U_j \sim \chi_j^2$. *Let* D *denote a random variable with support in* $\{0, \ldots, p\}$ *and* $P(D = d) = w(p, p-d, V)$. *Define* $\bar{W}(d) = \sum_{j=1}^p \{d = j\} U_j$. *Then*

$$\mathcal{L}(W) = \mathcal{L}(\bar{W}(D)).$$

4. *(Calculation of weights) For subsets* $M \subset \{1, \ldots, p\}$ *let* $\bar{M}$ *denote* $\{1, \ldots, p\} \setminus M$. *For* $M_1, M_2 \subset \{1, \ldots, p\}$ *and a* $(p \times p)$*-matrix* A *let* A_{M_1, M_2} *denote* A *with the rows with indices corresponding to entries in* $\bar{M}_1$ *and the columns with indices corresponding to entries in* $\bar{M}_2$ *deleted. For* $M \neq \emptyset$ *define the normal vector* $Y_1(M) \sim N(0, V_{M,M}^{-1})$ *and the probability* $p_1(M) = P(Y_1(M) \leq 0)$. *For* $M = \emptyset$ *set* $p_1(M) = 1$. *For* $M \neq \{1, \ldots, p\}$ *define the normal vector* $Y_2(M) \sim N(0, (V^{-1})_{\bar{M},\bar{M}}^{-1})$ *and the probability* $p_2(M) = P(Y_2(M) > 0)$. *For* $M = \{1, \ldots, p\}$ *set* $p_2(M) = 1$. *The weights can be written as*

$$w(p, p-j, V) = \sum_{\substack{M \subset \{1, \ldots, p\} \\ |M| = j}} p_1(M) p_2(M).$$

Proof. (1) In the derivation of the weights (see e.g., Nüesch 1966) the weights correspond to probabilities of events that partition the sample space. To prove the asserted upper bounds use the representation from (4) and write

$$\begin{aligned} w(p, p, V) = P(Y_2(\emptyset) > 0) &\leq 1 - P(\text{there is } j = 1, \ldots, p \text{ such that } Y_{2,j}(\emptyset) \leq 0) \\ &\leq 1 - \max_{j=1,\ldots,p} P(Y_{2,j}(\emptyset) \leq 0) = \frac{1}{2}. \end{aligned}$$

For the other weights, the bound can be proved in a similar way.
(2) This can be proved analogously to the derivation of the distribution of the $\bar{\chi}^2$ statistic (see Kudo 1963; Nüesch 1966).
(3) This follows from (2) upon observing that $\tilde{\chi}^2(V)$ is supported only on the nonnegative reals and that $\{[c, \infty) : c > 0\}$ is a generating class.
(4) See Kudo (1963) and Nüesch (1966). □

Lemma D.14. *Let* $\hat{A}$ *and* A *denote nonsingular* $p \times p$ *matrices and suppose that*

$$2\|\hat{A} - A\|_2 \|A^{-1}\|_2 \leq 1.$$

Then,

$$\|\hat{A}^{-1} - A^{-1}\|_2 \leq 2\|\hat{A} - A\|_2 \|A^{-1}\|_2^2.$$

Proof. This approach is originally due to Lewis and Reinsel (1985). Like any induced norm, the $\|\cdot\|_2$-norm obeys submultiplicativity so that

$$\|\hat{A}^{-1} - A^{-1}\|_2 = \|\hat{A}^{-1}(\hat{A} - A)A^{-1}\|_2 \leq \|\hat{A} - A\|_2 \|A^{-1}\|_2 (\|A^{-1}\|_2 + \|\hat{A}^{-1} - A^{-1}\|_2).$$

Rearranging yields

$$\|\hat{A}^{-1} - A^{-1}\|_2 \leq \frac{\|\hat{A}^{-1} - A^{-1}\|_2 \|A^{-1}\|_2^2}{1 - \|\hat{A} - A\|_2 \|A^{-1}\|_2} \leq 2\|\hat{A} - A\|_2 \|A^{-1}\|_2^2.$$

□

Lemma D.15. *Let* $(\phi_i)_{i=1}^N$ *denote normal random variables such that* $\phi_i \sim N(0, I_{p_i})$ *with* $p_i \leq \bar{p}$. *Let* $\underline{a}, \bar{a} > 0$ *and let* c_N *denote a deterministic sequence. For each* $\tau > 0$ *and* $\kappa > 0$ *there exist positive constants* $\bar{C}$ *and* N_0 *such that for* $N \geq N_0$ *and all* $\epsilon > N^{-\tau}$ *we have*

$$\sup_{(a_1,\dots,a_N)\in[\underline{a},\bar{a}]^N} P\left(\Big|\max_{1\leq i\leq N} \frac{\|\phi_i\|}{a_i} - c_N\Big| \leq \epsilon\right) \leq \bar{C}_a \epsilon\sqrt{\log N} + N^{-\kappa}.$$

Proof. Let ϵ denote a generic constant satisfying $\epsilon > N^{-\tau}$. Let Γ_i denote a δ_N-covering of a sphere in $\mathbb{R}^{p_i}$ with radius a_i^{-1}, where

$$\delta_N = \frac{1}{4} N^{-\tau}\big((\kappa+1)\log N\big)^{-1/2}.$$

It is without loss of generality to assume that, for all $\gamma \in \Gamma_i$, $\|\gamma\| = a_i^{-1}$. An upper bound on $\text{card}(\Gamma_i)$ is given by

$$\text{card}(\Gamma_i) \leq b_1 N^{b_2},$$

where b_1 and b_2 depend only on κ, τ, $\underline{a}$ and $\bar{p}$. As in Zhilova (2015), note that

$$\frac{\|\phi_i\|}{a_i} = \sup_{\gamma\in\mathbb{R}^{p_i}:\|\gamma\|=a_i^{-1}} \gamma'\phi_i.$$

We employ an approximation argument based on the inequality

$$\begin{aligned} P\left(\Big|\max_{1\leq i\leq N} \frac{\|\phi_i\|}{a_i} - c_N\Big| \leq \epsilon\right) \leq & P\left(\Big|\max_{1\leq i\leq N}\max_{\gamma_j\in\Gamma_i} \gamma_j'\phi_i - c_N\Big| \leq 2\epsilon\right) \\ & + P\left(\max_{1\leq i\leq N} \sup_{\gamma\in\mathbb{R}^{p_i}:\|\gamma\|=a_i^{-1}} \min_{\gamma_j\in\Gamma_i} |(\gamma-\gamma_j)'\phi_i| > \epsilon\right) \\ \equiv & A_1 + A_2. \end{aligned}$$

To bound A_1, note that each $\gamma_j'\phi_i$ is a normal random variable with standard deviation bounded between $\bar{a}^{-1}$ and $\underline{a}^{-1}$. This follows from our assumptions about the covering Γ_i and

$$E\big[(\gamma_j'\phi_i)^2\big] = \gamma_j' E[\phi_i\phi_i']\gamma_j = \|\gamma_j\|^2 = a_i^{-2}.$$

Then, $\max_{1\leq i\leq N}\max_{\gamma_j\in\Gamma_i}|\gamma_j'\phi_i|$ is the maximum of $\sum_{i=1}^N \text{card}(\Gamma_i)$ independent normal random variables and the results for Levy concentration bounds in Chernozhukov, Chetverikov, and Kato (2015) apply. For a constant C_a depending only on $\underline{a}$ and $\bar{a}$, their Corollary 1 yields

$$A_1 \leq C_a\epsilon\sqrt{1 \vee \log\left(\frac{\sum_{i=1}^N \text{card}(\Gamma_i)}{2\epsilon}\right)} \leq C_a\epsilon\sqrt{1\vee\log\left((1/2)b_1 N^{b_2+\tau}\right)} \leq \bar{C}\epsilon\sqrt{\log N}.$$

The last inequality holds for $N \geq N_{0,a}$ and sufficiently large $\bar{C}$, where the choice of $N_{0,a}$ and $\bar{C}$ depends only on κ, τ, $\underline{a}$ and $\bar{p}$. To bound A_2 let $N_{0,b}$ be large enough such that for $N \geq N_{0,b}$ and

$t_N = \frac{1}{2}(N^{-\tau}/\delta_N)^2$ we have $t_N > \bar{p}$. For $N \geq N_{0,b}$, by the Cauchy-Schwarz inequality

$$\begin{aligned} A_2 \leq & P\left(\max_{1\leq i\leq N} \|\phi_i\|^2 > \left(\frac{\epsilon}{\delta_N}\right)^2\right) \\ \leq & P\left(\max_{1\leq i\leq N} \|\phi_i\|^2 - p_i > t_N\right) \leq N \exp\left(-\frac{t_N}{8}\right) \leq N^{-\kappa}. \end{aligned}$$

The fourth inequality follows from the fact that $\|\phi_i\|^2$ obeys the subexponential condition

$$E\left[e^{\alpha(\|\phi_i\|^2 - p_i)}\right] \leq e^{\frac{4^2\alpha^2}{2}} \qquad \text{for all } |\alpha| < \frac{1}{2\sqrt{p_i}}.$$

This implies the tail bound

$$P(\|\phi_i\|^2 - p_i > t_N) \leq e^{-\frac{t_N}{8}}$$

for $t_N > p_i$ (see, e.g., Proposition 2.2 in Wainwright 2015). The conclusion of the lemma follows by setting $N_0 = \max\{1, N_{0,a}, N_{0,b}\}$. □

Lemma D.16. *Let $(\phi_i)_{i=1}^N$ denote normal random vectors such that $\phi_i \sim N(0, I_{p_i})$ with $p_i \leq \bar{p}$. For $\underline{a}, \bar{a}, \gamma > 0$, and a positive deterministic sequence c_N such that $c_N \leq N^\gamma$ let $\mathbb{S}_N = [c_N\underline{a}, c_N\bar{a}]$. For each $\tau > 0$ and $\kappa > 0$ there exist positive constants $\bar{C}$ and N_0 such that for $N \geq N_0$ and all $\epsilon > N^{-\tau}$ we have*

$$\sup_{(s_1,\ldots,s_N)\in\mathbb{S}^N} P\left(\left|\max_{1\leq i\leq N}\left(\|\phi_i\|^2 - s_i\right)\right| \leq \epsilon\right) \leq \bar{C}\epsilon\sqrt{\frac{\log N}{c_N}} + N^{-\kappa}.$$

If the random vectors ϕ_i are independent, then we also have

$$\sup_{(s_1,\ldots,s_N)\in\mathbb{S}^N} P\left(\left|\max_{1\leq i\leq N}\left(\|\phi_i\|^2 - s_i\right)\right| \leq \epsilon\right) \leq \bar{C}\epsilon\Big(1 + \sqrt{\frac{\bar{a}c_N}{\log N}}\Big)^{-1} + N^{-\kappa}.$$

Proof. Fix $\epsilon > N^{-\tau}$ and $(s_1, \ldots, s_N) \in \mathbb{S}^N$. Let L_N denote a lower bound on $\max_{1\leq i\leq N}\|\phi_i\|$. Suppose first that the ϕ_i are independent. Then,

$$\max_{1\leq i\leq N}\|\phi_i\| \geq \max_{1\leq i\leq N}|\phi_{i,1}| \geq \max_{1\leq i\leq N}\phi_{i,1}.$$

By Example 3.5.5 in Embrechts, Klüppelberg, and Mikosch (2013)

$$\frac{\max_{1\leq i\leq N}\phi_{i,1}}{\sqrt{2\log N}} \to 1 \qquad P\text{-almost surely.}$$

Therefore, there exists a finite $N_{0,a}$ for which we may assume $N \geq N_{0,a} \Rightarrow \max_{1\leq i\leq N}\|\phi_i\| \geq \sqrt{\log N}$. This implies that, for independent ϕ_i, we may take $L_N = \sqrt{\log N}$, otherwise take $L_N = 0$. For each $i = 1, \ldots, N$ write $s_i = c_N a_i$. For $N \geq N_{0,a}$

$$P\left(|\max_{1\leq i\leq N}\left(\|\phi_i\|^2 - s_i\right)| \leq \epsilon\right)$$

$$
\begin{aligned}
\leq & P\left(\left| \max_{1\leq i\leq N} \frac{\|\phi_i\|^2}{a_i} - c_N \right| \leq \underline{a}^{-1}\epsilon \right) \\
\leq & P\left(\left(\bar{a}^{-1/2} \max_{1\leq i\leq N} \|\phi_i\| + \sqrt{c_N}\right) \left| \max_{1\leq i\leq N} \frac{\|\phi_i\|}{\sqrt{a_i}} - \sqrt{c_N} \right| \leq \underline{a}^{-1}\epsilon \right) \\
\leq & P\left(\left| \max_{1\leq i\leq N} \frac{\|\phi_i\|}{\sqrt{a_i}} - \sqrt{c_N} \right| \leq \frac{\underline{a}^{-1}\epsilon}{\bar{a}^{-1/2}L_N + \sqrt{c_N}} \right).
\end{aligned}
$$

Let

$$
\epsilon' = \frac{\underline{a}^{-1}\epsilon}{\bar{a}^{-1/2}L_N + \sqrt{c_N}}.
$$

Since $c_N < N^{\gamma}$, we can find $\tau' > 0$, depending only on $\underline{a}$, $\bar{a}$, τ and γ, such that $\epsilon' > N^{-\tau'}$. Applying Lemma D.15 with ϵ' and τ' we may now conclude that there are constants $N_{0,b}$ and $\bar{C}_a$ such that for $N \geq N_{0,b}$

$$
\begin{aligned}
P\left(\left| \max_{1\leq i\leq N} \frac{\|\phi_i\|}{\sqrt{a_i}} - \sqrt{c_N} \right| \leq \frac{\underline{a}^{-1}\epsilon}{\bar{a}^{-1/2}\sqrt{\log N} + \sqrt{c_N}} \right) \leq & \bar{C}_a \epsilon' \sqrt{\log N} + N^{-\kappa} \\
\leq & C\left(\frac{\sqrt{\log N}}{L_N + \sqrt{\bar{a} c_N}} \right) + N^{-\kappa}.
\end{aligned}
$$

The last inequality holds for conformant C. The assertion of the lemma follows by choosing $N_0 = \max\{N_{0,a}, N_{0,b}\}$ and plugging in the appropriate value of L_N. □

Lemma D.17 (Extremal bound for normal vector)**.** *Let X be a centered normal random vector of length p with covariance matrix V. Let $a > 0$. Then,*

$$
P\big(\|X\|_{\max} > \sqrt{2p\|V\|_2 \log(a)}\big) \leq P\big(\|X\|_{\max} > \sqrt{2\,\mathrm{tr}(V)\log(a)}\big) \leq \frac{\sqrt{p}}{a\sqrt{\pi \log a}}.
$$

Proof. For the first inequality, let $0 \leq \lambda_1 \leq \cdots \leq \lambda_p$ denote the eigenvalues of V. Then

$$
p\|V\|_2 = p\lambda_p \geq \sum_{j=1}^{p} \lambda_j = \mathrm{tr}(V).
$$

For the second inequality, write $c = \sqrt{2\,\mathrm{tr}(V)\log(a)}$. Then

$$
\begin{aligned}
P(\|X\|_{\max} > c) \leq \sum_{j=1}^{p} (|X_j| > c) = & 2\sum_{j=1}^{p}\left(1 - \Phi\left(\frac{c}{\sigma_j}\right)\right) \\
\leq & 2\sum_{j=1}^{p} \frac{\sigma_j}{c}\phi\left(\frac{c}{\sigma_j}\right) \\
= & \sqrt{\frac{2}{\pi}} \sum_{j=1}^{p} \frac{\sigma_j}{c} \exp\left(-\frac{c^2}{2\sigma_j^2}\right) \\
\leq & \sqrt{\frac{2}{\pi}} \exp\Big(-\frac{c^2}{2\,\mathrm{tr}(V)}\Big) \frac{1}{c}\sum_{j=1}^{p}\sigma_j \\
\leq & \sqrt{\frac{2p}{\pi}} \exp\Big(-\frac{c^2}{2\,\mathrm{tr}(V)}\Big) \frac{\sqrt{\mathrm{tr}(V)}}{c}
\end{aligned}
$$

$$=\sqrt{\frac{p}{\pi}}\frac{1}{a\sqrt{\log a}},$$

where the second inequality uses Gordon's inequality for standard normal probabilities (see, e.g., Duembgen (2010)), and the last inequality uses the inequality $\|x\|_1 \le \sqrt{p}\|x\|_2$ for vectors x of length p where $\|\cdot\|_1$ is the L_1 norm. □

Lemma D.18 (Perturbation bound for rectangular normal probabilities)**.** *Consider the centered normal p-vectors X and $\hat{X}$ with respective positive-definite covariance matrices V and $\hat{V}$. Let $x_1 = (\tilde{x}_1, \ldots, \tilde{x}_p) \in \mathcal{R}^p$ and $x = (\tilde{x}_{p+1}, \ldots, \tilde{x}_{2p}) \in \mathcal{R}^p$. Let $x_{\max} = \max_{1\le j\le 2p} |\tilde{x}|_j$ and suppose that $x_{\max} > 0$. Moreover, assume*

$$\|\hat{V}^{-1} - V^{-1}\|_2\big(\|V\|_2 \vee \|\hat{V}\|_2 \vee px_{\max}^2\big) \le 1.$$

Then, for any measurable function $g : \mathcal{R}^p \to \mathcal{R}$

$$\ell_{N,T} \le \frac{\mathbb{E}[g(\hat{X})1\{x_1 \le \hat{X} \le x_2\}]}{\mathbb{E}[g(X)1\{x_1 \le X \le x_2\}]} \le u_{N,T},$$

where

$$\begin{aligned}\ell_{N,T} &= \Big(1 + (2^p - 1)\|\hat{V}^{-1} - V^{-1}\|_2\|\hat{V}\|_2\Big)^{-1}\Big(1 + px_{\max}^2\|\hat{V}^{-1} - V^{-1}\|_2\Big)^{-1},\\ u_{N,T} &= \Big(1 + (2^p - 1)\|\hat{V}^{-1} - V^{-1}\|_2\|V\|_2\Big)\Big(1 + px_{\max}^2\|\hat{V}^{-1} - V^{-1}\|_2\Big).\end{aligned}$$

Suppose that, in addition,

$$(2^p - 1)\|\hat{V}^{-1} - V^{-1}\|_2(\|V\|_2 \vee \|\hat{V}\|_2) \le 1$$

then

$$\begin{aligned}&\Big|P(x_1 \le \hat{X} \le x_2) - P(x_1 \le X \le x_2)\Big|\\ \le&\|\hat{V}^{-1} - V^{-1}\|_2\big((2^p - 1)(\|V\|_2 \vee \|\hat{V}\|_2) + 2px_{\max}^2\big).\end{aligned}$$

Proof. Let f_X and $f_{\hat{X}}$ denote the probability densities corresponding to X and $\hat{X}$. Then,

$$\begin{aligned}&E\Big[g(\hat{X})1\{x_1 \le \hat{X} \le x_2\}\Big]\\ =&E\Big[g(X)1\{x_1 \le X \le x_2\}\frac{f_{\hat{X}}(X)}{f_X(X)}\Big]\\ =&\frac{\det(V)}{\det(\hat{V})}E\Big[g(X)1\{x_1 \le X \le x_2\}\exp\Big(-\frac{1}{2}X'\Big(\hat{V}^{-1} - V^{-1}\Big)X\Big)\Big]\\ \le&\det(\hat{V}^{-1}V)E\Big[g(X)1\{x_1 \le X \le x_2\}\exp\Big(\frac{1}{2}\|X\|^2\|\hat{V}^{-1} - V^{-1}\|_2\Big)\Big]\\ \le&\Big(1 + (2^p - 1)\|\hat{V}^{-1} - V^{-1}\|_2\|V\|_2\Big)\\ &\times E\Big[g(X)1\{x_1 \le X \le x_2\}\exp\Big(\frac{px_{\max}^2}{2}\|\hat{V}^{-1} - V^{-1}\|_2\Big)\Big]\\ \le&E\left[g(X)1\{x_1 \le X \le x_2\}\right]\Big(1 + (2^p - 1)\|\hat{V}^{-1} - V^{-1}\|_2\|V\|_2\Big)\end{aligned}$$

$$\times \left(1 + px_{\max}^2 \|\hat{V}^{-1} - V^{-1}\|_2\right).$$

The last inequality uses the inequality $\exp(x) \leq 1 + 2x$ for $x \leq 1/2$.[34] For the second inequality note that Hadamard's inequality implies (see e.g. Lemma 2.5 in Ipsen and Rehman (2008))

$$\det(\hat{V}^{-1}V) \leq \|\hat{V}^{-1}V\|_2^p \leq \left(\|I_p\|_2 + \|(\hat{V}^{-1} - V^{-1})V\|_2\right)^p$$
$$\leq 1 + (2^p - 1)\left(\|\hat{V}^{-1} - V^{-1}\|_2 \|V\|_2\right).$$

This holds since $\|I_p\|_2 = 1$ and, for $0 \leq a \leq 1$, we have

$$(1+a)^p \leq 1 + \sum_{k=1}^{p} \binom{p}{k} a^k \leq 1 + (2^p - 1)a.$$

To derive the lower bound reverse the roles of X and $\hat{X}$. □

Lemma D.19 (Perturbation bound for large quantiles)**.** *Suppose that $\hat{V}$ and V are positive definite $(p \times p)$ variance matrices for $p \geq 2$. Let $W \sim \tilde{\chi}^2(V)$ and $\widehat{W} \sim \tilde{\chi}^2(\hat{V})$. Let $\hat{c}_{\alpha,N}$ and $c_{\alpha,N}$ denote the $(1 - \alpha/N)$-quantile of $\hat{W}$ and W, respectively. Suppose that*

$$32p^2 \|\hat{V} - V\|_2 (1 \vee 2\|V^{-1}\|_2^2)(\|V\|_2 \vee \|V^{-1}\|_2) \log(N/\alpha) \leq 1.$$

There is a threshold N_0 depending only on α and p such that for $N \geq N_0$ and

$$\alpha_N = \alpha\Big(1 + 96\left(\|\hat{V} - V\|_2 (1 \vee \|V^{-1}\|_2^2)(\|V\|_2 \vee \|V^{-1}\|_2) \log(N/\alpha) + N^{-1}\right)$$

we have

$$\hat{c}_{\alpha,N} \geq c_{\alpha_N,N}.$$

Proof. Throughout the proof, we take N large enough so that

$$2\|\hat{V} - V\|_2 \|V^{-1}\|_2 \big(1 \vee 2\|V^{-1}\|_2 \|V\|_2\big) \leq 1.$$

This proof is based on the mixture representation of the $\tilde{\chi}^2$-distribution from Lemma D.13. For each $M = (m_1, \ldots, m_{|M|}) \subset \{1, \ldots, p\}$ with $m_1 < \cdots < m_{|M|}$ where $|M|$ is the cardinality of M, let S_M denote a $|M| \times p$ matrix with ones in the cells (m_k, k), $k = 1, \ldots, |M|$, and zeros in all other entries. For $M_1, M_2 \subset \{1, \ldots, p\}$ and a symmetric positive-definite matrix A, let $A_{M_1,M_2} = M_1' A M_2$. Let $\bar{M} = \{1, \ldots, p\} \setminus M$. For $M \subset \{1, \ldots, |M|\}$ and a symmetric, positive definite matrix A, we are interested in the centered normal random vector $Y_1(A, M)$ with covariance matrix

$$\Sigma_1(A, M) = \Sigma_1(A) = (A_{M,M})^{-1}$$

[34]By the series expansion of the exponential function for $0 \leq x \leq 1/2$

$$\exp(x) = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots \leq 1 + x + x\sum_{n=1}^{\infty} \frac{1}{(n+1)!}x^n = 1 + x + x\sum_{n=1}^{\infty} \frac{1}{(n+1)(n!)}x^n$$
$$\leq 1 + \frac{x}{2}(e^x - 1) \leq 1 + 2x.$$

and the centered normal random vector $Y_2(A, M)$ with covariance matrix

$$\Sigma_2(A, M) = \Sigma_2(A) = A_{\bar{M},\bar{M}} - A_{\bar{M},M} \left(A_{M,M}\right)^{-1} A_{M,\bar{M}} = \left((A^{-1})_{\bar{M},\bar{M}}\right)^{-1}.$$

We first establish some useful inequalities. By Lemma D.14

$$\|\hat{V}^{-1} - V^{-1}\|_2 \leq 2\|\hat{V} - V\|_2 \|V^{-1}\|_2^2. \tag{28}$$

In the following, let A denote a generic nonsingular, symmetric $(p \times p)$-matrix and let M denote a generic subset of $\{1, \ldots, p\}$. For any submatrix B of A, $\|B\|_2 \leq \|A\|_2$. Let $\lambda_1(A)$ denote the smallest eigenvalue of A. By the interlacing property for eigenvalues of principal submatrices (see, e.g., Theorem 4.3.28 in Horn and Johnson (2013)) and the fact that permuting a matrix does not change its eigenvalues, we have

$$\lambda_1((A)_{M,M}) \geq \lambda_1(A)$$

and therefore

$$\|(A)_{M,M}^{-1}\|_2 = \lambda_1^{-1}((A)_{M,M}) \leq \lambda_1^{-1}(A) = \|A^{-1}\|_2. \tag{29}$$

Applying this result with $A = V$ yields $\|\Sigma_1(V)\|_2 \leq \|V^{-1}\|_2$. Moreover,

$$2\|\hat{V}_{M,M}^{-1} - V_{M,M}\|_2 \|V_{M,M}^{-1}\|_2 \leq 2\|\hat{V}_{M,M} - V_{M,M}\|_2 \|\Sigma_1(V)\|_2 \leq 2\|\hat{V} - V\|_2 \|V^{-1}\|_2 \leq 1$$

so that by Lemma D.14

$$\|\Sigma_1(\hat{V}) - \Sigma_1(V)\|_2 \leq 2\|\hat{V}_{M,M} - V_{M,M}\|_2 \|\Sigma_1(V)\|_2^2 \leq 2\|\hat{V} - V\|_2 \|V^{-1}\|_2^2 \leq \|V^{-1}\|_2.$$

Then, by the triangle inequality,

$$\|\Sigma_1(\hat{V})\|_2 \leq \|\Sigma_1(V)\|_2 + \|\Sigma_1(\hat{V}) - \Sigma_1(V)\|_2 \leq 2\|V^{-1}\|_2.$$

Moreover,

$$\|\Sigma_1^{-1}(\hat{V}) - \Sigma_1^{-1}(V)\|_2 \leq \|\hat{V}_{M,M} - V_{M,M}\|_2 \leq \|\hat{V} - V\|_2.$$

By inequality (29), we have $\|\Sigma_2(V)\|_2 \leq \|V\|_2$ and therefore

$$\|(\hat{V}^{-1})_{\bar{M},\bar{M}} - (V^{-1})_{\bar{M},\bar{M}}\|_2 \|\Sigma_2(V)\|_2 \leq \|\hat{V}^{-1} - V^{-1}\|_2 \|V\|_2 \leq \frac{1}{2},$$

where the last line follows from inequality (28). Thus, by Lemma D.14

$$\begin{aligned}
\|\Sigma_2(\hat{V}) - \Sigma_2(V)\|_2 \leq& 2\|(\hat{V}^{-1})_{\bar{M},\bar{M}} - (V)_{\bar{M},\bar{M}}^{-1}\|_2 \|\Sigma_2(V)\|_2^2 \\
&\leq 4\|\hat{V}^{-1} - V^{-1}\|_2 \|\Sigma_2(V)\|_2^2 \leq 4\|\hat{V} - V\|_2 \|V^{-1}\|_2^2 \|V\|_2^2 \leq \|V\|_2.
\end{aligned}$$

By the triangle inequality,

$$\|\Sigma_2(\hat{V})\|_2 \leq \|\Sigma_2(V)\|_2 + \|\Sigma_2(\hat{V}) - \Sigma_2(V)\|_2 \leq 2\|V\|_2.$$

Moreover by inequality (28)

$$\begin{aligned}\|\Sigma_2^{-1}(\hat{V}) - \Sigma_2^{-1}(V)\|_2 \leq& \|(\hat{V}^{-1})_{\bar{M},\bar{M}} - (V^{-1})_{\bar{M},\bar{M}}\|_2 \\ \leq& \|\hat{V}^{-1} - V^{-1}\|_2 \leq 2\|\hat{V} - V\|_2 \|V^{-1}\|_2^2.\end{aligned}$$

For $(\Sigma, \hat{\Sigma}) \in \{(\Sigma_1, \hat{\Sigma}_1), (\Sigma_2, \hat{\Sigma}_2)\}$, let Y and $\hat{Y}$ denote random variables such that $Y \sim N(0, \Sigma)$ and $\hat{Y} \sim N(0, \hat{\Sigma})$. By the calculations above

$$\begin{aligned}\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_2 \leq& \|\hat{V} - V\|_2 (1 \vee 2\|V^{-1}\|_2^2), \\ \|\hat{\Sigma}\|_2 \vee \|\Sigma\|_2 \leq& 2(\|V\|_2 \vee \|V^{-1}\|_2).\end{aligned}$$

Let $a_N = \sqrt{8p(\|V\|_2 \vee \|V^{-1}\|_2)\log(N/\alpha)}$ and let N be large enough such that

$$\log(N/\alpha) \geq \max\left\{1, \frac{9(2^{2p-1})\alpha^2 p}{\pi}, \frac{2^p - 1}{8p^2}\right\}.$$

By Lemma D.17

$$P(\|Y\|_{\max} > a_N) \leq \frac{\alpha}{N^2}\sqrt{\frac{\alpha^2 p}{2\pi \log(N/\alpha)}} \leq \frac{\alpha}{3N^2 2^p}.$$

Define the probabilities

$$\begin{aligned}p(A, M) =& P\left(Y_1(A, M) \leq 0\right) P\left(Y_2(A, M) > 0\right), \\ p_N(A, M) =& P\left(Y_1(A, M) \leq 0 \wedge \|Y_1(A, M)\|_{\max} \leq a_N\right) \\ & \times P\left(Y_2(A, M) > 0 \wedge \|Y_2(A, M)\|_{\max} \leq a_N\right).\end{aligned}$$

Note that by the characterization of the $\tilde{\chi}^2$-distribution in Lemma D.13, it suffices to show that

$$\sum_{M \subset \{1,\ldots,p\}} p(V, M) P\left(U_{|M|} > \hat{c}_{\alpha,N}\right) \leq \frac{\alpha_N}{N}. \tag{30}$$

We have

$$\begin{aligned}p(V, M) \leq& p_N(V, M) + P\left(\|Y_1(V, M)\| > a_N\right) \\ & + P\left(\|Y_2(V, M)\| > a_N\right) + P\left(\|Y_1(V, M)\| > a_N\right) P\left(\|Y_2(V, M)\| > a_N\right) \\ \leq& p_N(V, M) + \frac{\alpha}{2^p N}.\end{aligned}$$

By the definition of $\hat{c}_N$,

$$\begin{aligned}\frac{\alpha}{N} =& \sum_{M \subset \{1,\ldots,p\}} p(\hat{V}, M) P\left(U_{|M|} > \hat{c}_{\alpha,N}\right) \\ \geq& \sum_{M \subset \{1,\ldots,p\}} p_N(\hat{V}, M) P\left(U_{|M|} > \hat{c}_{\alpha,N}\right).\end{aligned}$$

Hence,

$$
\begin{aligned}
&\sum_{M\subset\{1,\dots,p\}} p(V,M)P\left(U_{|M|}>\hat{c}_{\alpha,N}\right)\\
\leq&\frac{\alpha}{N}+\sum_{M\subset\{1,\dots,p\}}\left(p_N(V,M)-p_N(\hat{V},M)\right)P\left(U_{|M|}>\hat{c}_{\alpha,N}\right)+\sum_{M\subset\{1,\dots,p\}}\frac{\alpha}{2^pN^2}\\
\leq&\frac{\alpha}{N}+\sum_{M\subset\{1,\dots,p\}}\left(p_N(V,M)-p_N(\hat{V},M)\right)P\left(U_{|M|}>\hat{c}_{\alpha,N}\right)+\frac{\alpha}{N^2}\\
\leq&\frac{\alpha}{N}\left(1+N^{-1}+\sum_{M\subset\{1,\dots,p\}}\left(\frac{p_N(V,M)}{p_N(\hat{V},M)}-1\right)\right). \qquad (31)
\end{aligned}
$$

Note that, for $M\subset\{1,\dots,p\}$

$$
\begin{aligned}
&(2^p-1)\|\Sigma_1(\hat{V},M)-\Sigma_1(V,M)\|_2\|\Sigma_1(\hat{V},M)\|_2\\
\leq&2(2^p-1)\|\hat{V}-V\|_2\left(1\vee 2\|V^{-1}\|_2^2\right)\left(\|V\|_2\vee\|V^{-1}\|_2\right)\leq 1
\end{aligned}
$$

and

$$
\begin{aligned}
&\|\Sigma_1(\hat{V},M)-\Sigma_1(V,M)\|_2\left(\|\Sigma_1(V,M)\|_2\vee\|\Sigma_1(\hat{V},M)\|_2\vee 2a_N^2\right)\\
\leq&2\|\hat{V}-V\|_2\left(1\vee 2\|V^{-1}\|_2^2\right)\left(\|V\|_2\vee\|V^{-1}\|_2\right)\left(1\vee 4p^2\log\left(N/\alpha\right)\right)\leq 1.
\end{aligned}
$$

Therefore, we can apply Lemma D.18 to argue that

$$
\begin{aligned}
&\frac{P\left(Y_1\left(V,M\right)\leq 0\wedge\|Y_1\left(V,M\right)\|_{\max}\leq a_N\right)}{P\left(Y_1(\hat{V},M)\leq 0\wedge\|Y_1(\hat{V},M)\|_{\max}\leq a_N\right)}\\
\leq&1+16p^2\|\hat{V}-V\|_2\left(1\vee 2\|V^{-1}\|_2^2\right)\left(\|V\|_2\vee\|V^{-1}\|_2\right)\left(1+\frac{2(2^p-1)}{16p^2\log\left(N/\alpha\right)}\right)\log\left(N/\alpha\right)\\
\leq&1+32p^2\|\hat{V}-V\|_2\left(1\vee 2\|V^{-1}\|_2^2\right)\left(\|V\|_2\vee\|V^{-1}\|_2\right)\log\left(N/\alpha\right).
\end{aligned}
$$

Similarly, we can show that

$$
\begin{aligned}
&\frac{P\left(Y_2\left(V,M\right)\leq 0\wedge\|Y_2\left(V,M\right)\|_{\max}\leq a_N\right)}{P\left(Y_2(\hat{V},M)\leq 0\wedge\|Y_2(\hat{V},M)\|_{\max}\leq a_N\right)}\\
\leq&1+32p^2\|\hat{V}-V\|_2\left(1\vee 2\|V^{-1}\|_2^2\right)\left(\|V\|_2\vee\|V^{-1}\|_2\right)\log\left(N/\alpha\right).
\end{aligned}
$$

Under the assumptions of the lemma,

$$
32p^2\|\hat{V}-V\|_2\left(1\vee 2\|V^{-1}\|_2^2\right)\left(\|V\|_2\vee\|V^{-1}\|_2\right)\log(N/\alpha)\leq 1,
$$

so that

$$
\frac{p_N(V,M)}{p_N(\hat{V},M)}-1\leq 96p^2\|\hat{V}-V\|_2\left(1\vee 2\|V^{-1}\|_2^2\right)\left(\|V\|_2\vee\|V^{-1}\|_2\right)\log\left(N/\alpha\right).
$$

Plugging this bound into the right-hand side of (31) verifies (30) and concludes the proof. □

Lemma D.20 (Bounds on large quantiles)**.** *For a nonsingular covariance matrix V, let $c^{\text{QLR}}_{\alpha,N}(V)$ denote the $(1-\alpha/N)$-quantile of $\tilde{\chi}^2(V)$. For each $a > 1$, there is N_0 depending only on α and a, such that for $N \geq N_0$,*

$$2a^{-1}\log(N/\alpha) < c^{\text{QLR}}_{\alpha,N}(V) < 2a\log(N/\alpha).$$

Proof. Let $U_j \sim \chi^2_j$, $j = 1, \ldots, p$. For notational convenience, write $c_N = c^{\text{QLR}}_{\alpha,N}(V)$. By Lemma D.13, c_N is bounded from above by the $(1-\alpha/N)$-quantile of U_p. Lemma 1 in Laurent and Massart (2000) implies that, for each $x \geq 0$,

$$P\left(U_p - p \geq 2\sqrt{px} + 2x\right) \leq \exp(-x).$$

Suppose that $N \geq N_0 \geq \alpha^{-1}$. Choosing $x = \log(N/\alpha)$ in the above inequality yields

$$P\Big(U_p \geq p + 2\sqrt{\log(N/\alpha)}\big(\sqrt{p} + \sqrt{\log(N/\alpha)}\big)\Big) \leq \frac{\alpha}{N}.$$

For N large enough,

$$p + 2\sqrt{\log(N/\alpha)}\big(\sqrt{p} + \sqrt{\log(N/\alpha)}\big) < 2a\log(N/\alpha).$$

This establishes the upper bound on c_N. Let $x = \sqrt{2a^{-1}\log(N/\alpha)}$, and let Φ denote the distribution function and ϕ the density function of a standard normal random variable. Komatu's lower bound (see, e.g., Duembgen (2010)) is given by

$$1 - \Phi(x) > \frac{2\phi(x)}{\sqrt{4 + x^2} + x}.$$

By Lemma D.13, the distribution $\tilde{\chi}^2(V)$ has point mass $w(p, p, V) \leq 1/2$ at zero. Let $W \sim \tilde{\chi}^2(V)$ and $U_j \sim \chi^2_j$, $j = 1, \ldots, p$. Then,

$$\begin{aligned} P(W > x^2) =& \sum_{j=1}^{p} w(p, p-j, V) P(U_j > x^2) \\ >& (1 - w(p, p, V)) P(U_1 > x^2) \geq \frac{1}{2} P(U_1 > x^2). \end{aligned}$$

Suppose that N is large enough such that $\sqrt{4/x^2 + 1} \leq 2$ For a standard normal random variable Z, we have

$$\begin{aligned} \frac{1}{2}P(U_1 > x^2) = \frac{1}{2}P(|Z| > x) =& 1 - \Phi(x) \\ >& \frac{2\phi(x)}{x(1 + \sqrt{4/x^2 + 1})} \\ >& \frac{\sqrt{2}\exp\left(-\frac{x^2}{2}\right)}{3\sqrt{\pi}x} = \frac{(\alpha/N)^{a^{-1}}}{3\sqrt{\pi a^{-1}\log(N/\alpha)}} \equiv p_0^N. \end{aligned}$$

Clearly, $p_0^N/(\alpha/N) \to \infty$. For large N, this establishes x^2 as a lower bound on the $(1-\alpha/N)$-quantile of W. ☐

E. Proofs for unit selection procedures

We first introduce some additional notation. Let

$$d_{it}^U(g,h) = \frac{1}{2}[(y_{it} - x_{it}'\beta_{g,t})^2 - (y_{it} - x_{it}'\beta_{h,t})^2]$$

and

$$\tilde{D}_i^U(g,h) = \frac{\sum_{t=1}^T d_{it}^U(g,h)}{\sqrt{\sum_{t=1}^T (d_{it}^U(g,h) - \bar{d}_i^U(g,h))^2}},$$

where $\bar{d}_i^U(g,h) = \sum_{t=1}^T d_{it}^U(g,h)/T$. Let

$$(s_{i,T}^U(g,h))^2 = \frac{1}{\sigma_i^2 T}\sum_{t=1}^T Var(d_{it}^U(g,h)),$$

and

$$(S_{i,T}^U(g,h))^2 = \frac{1}{T}\sum_{t=1}^T (d_{it}^U(g,h) - \bar{d}_i^U(g,h))^2.$$

Next, we observe that moment of $d_{it}^U(g_i^0,h)$ can be bounded by terms defined for $d_{it}(g_i^0,h)$. Let

$$Z_{it}^U(h) = \frac{d_{it}^U(g_i^0,h) - \mathbb{E}_P(d_{it}^U(g_i^0,h))}{\sigma_i s_{i,T}^U(g,h)}.$$

Note that

$$\begin{aligned}
&d_{it}^U(g_i^0,h) - \mathbb{E}_P(d_{it}^U(g_i^0,h)) \\
=&\frac{1}{2}\left[u_{it}^2 - (u_{it} + x_{it}'(\beta_{g_i^0,t} - \beta_{h,t}))^2\right] - \mathbb{E}_P(x_{it}'(\beta_{g_i^0,t} - \beta_{h,t}))^2 \\
=&u_{it}x_{it}'\delta_t(h,g_i^0) + \frac{1}{2}\delta_t(g_i^0,h)'(x_{it}x_{it}' - \mathbb{E}_P(x_{it}x_{it}')\delta_t(g_i^0,h).
\end{aligned}$$

This formula indicates that

$$\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\left(\mathbb{E}_P\left(\frac{1}{T}\sum_{t=1}^T \left|Z_{it}^U(h)\right|^p\right)\right)^{1/p} \le GD_{N,T,p},$$

and

$$\max_{1\le t\le T}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\left(\mathbb{E}_P\left(\max_{1\le i\le N}\left|Z_{it}^U(h)\right|^p\right)\right)^{1/p} \le GB_{N,T,p}.$$

Moreover, we have

$$\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\left(\frac{1}{T}\sum_{t=1}^T \frac{\mathbb{E}_P\left(\left|d_{it}^U(h)\right|^p\right)}{\sigma_i^p s_{i,T}^p(h)}\right)^{1/p} \le GD_{N,T,p}.$$

Lemma E.1. *Suppose that the probability measure P satisfies Assumption 1. Then, there is a constant C depending only on K_β and G such that for $0 < c < 1$ and*

$$\begin{aligned}\zeta^U_{N,T} =& \gamma_{N,T,8}\big(T^{-(1-c)/4}\sqrt{\log N}B^2_{N,T,8} + D_{N,T,4}\big)\big(D_{N,T,2} + T^{-(1-c)/4}\sqrt{\log N}B_{N,T,4}\big)\\ &+ \gamma^2_{N,T,8}\big(T^{-(1-c)/2}(\log N)B^4_{N,T,8} + D^2_{N,T,4}\big),\end{aligned}$$

we have

$$\begin{aligned}&P\Bigg(\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g^0_i\}}\frac{1}{T}\sum_{t=1}^T\left(\frac{\hat d^U_{it}(g^0_i,h) - d^U_{it}(g^0_i,h)}{\sigma_i s^U_{i,T}(h)}\right)^2 \qquad\text{(i)}\\ &\qquad > C\gamma^2_{N,T,8}\left(T^{-(1-c)/2}B^4_{N,T,8}(\log N) + D^2_{N,T,4}\right)\Bigg) \le CT^{-c},\end{aligned}$$

$$\begin{aligned}&P\Bigg(T^{-1/2}\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g^0_i\}}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^T\frac{d^U_{it}(g^0_i,h)}{\sigma_i s^U_{i,T}(g^0_i,h)} - \frac{1}{\sqrt{T}}\sum_{t=1}^T\frac{\hat d^U_{it}(g^0_i,h)}{\sigma_i s^U_{i,T}(g^0_i,h)}\right| \qquad\text{(ii)}\\ &\qquad > C\gamma_{N,T,8}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right)\Bigg) \le CT^{-c},\end{aligned}$$

$$\begin{aligned}&P\Bigg(\max_{1\le i\le N}\max_{h\in\mathbb{G}\setminus\{g^0_i\}}\left|\frac{1}{T}\sum_{t=1}^T\frac{(\hat d^U_{it}(g^0_i,h) - \bar{\hat d}^U_i(g^0_i,h))^2}{\sigma^2_i s^2_{i,T}(g^0_i,h)} - \frac{1}{T}\sum_{t=1}^T\frac{(d^U_{it}(g^0_i,h) - \bar d^U_i(g^0_i,h))^2}{\sigma^2_i s^2_{i,T}(g^0_i,h)}\right| \qquad\text{(iii)}\\ &\qquad > C\zeta^U_{N,T}\Bigg) \le CT^{-c}.\end{aligned}$$

Suppose that, additionally, $\zeta^U_{N,T}\vee T^{-(1-c)/4}\sqrt{\log N}B_{N,T,4} \le 1$. Then

$$\begin{aligned}&P\Bigg(\max_{1\le i\le N}\left|\hat D_i(g^0_i,h) - \tilde D_i(g^0_i,h)\right| > C\gamma_{N,T,8}\sqrt{T}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right) \qquad\text{(iv)}\\ &\qquad\qquad + C\zeta^U_{N,T}\left(D_{N,T,1} + \sqrt{\log N} + T^{-1/4}B_{N,T,4}\log N\right)\Bigg)\\ \le& N^{-1} + CT^{-c} + C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4.\end{aligned}$$

Proof of Lemma E.1. **Proof of (i):** Decompose $\hat d^U_{it}(h) - d^U_{it}(h)$ as follows

$$\begin{aligned}&\hat d^U_{it}(h) - d^U_{it}(h)\\ =& -u_{it}x'_{it}(\hat\delta_t(g^0_i,h) - \delta_t(g^0_i,h))\\ &+ (x'_{it}(\beta_{g^0_i,t} - \hat\beta_{g^0_i,t}))^2/2 + (x'_{it}(\beta_{h,t} - \hat\beta_{h,t}))^2/2 - (\beta_{h,t} - \hat\beta_{h,t})(x_{it}x'_{it})\delta_t(g^0_i,h).\end{aligned}$$

By the inequality $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$,

$$\begin{aligned}\left(\frac{\hat d^U_{it}(h) - d^U_{it}(h)}{\sigma_i}\right)^2 \le& 4\left|\frac{u_{it}}{\sigma_i}\right|^2\|x_{it}\|^2\|\hat\delta_t(g^0_i,h) - \delta_t(g^0_i,h)\|^2\\ &+ 2\sigma_i^{-2}\|\hat\beta_{g^0_i,t} - \beta_{g^0_i,t}\|^4\|x_{it}\|^4\\ &+ 2\sigma_i^{-2}\|\hat\beta_{h,t} - \beta_{h,t}\|^4\|x_{it}\|^4\end{aligned}$$

$$+ 2\sigma_i^{-2}\|\hat{\beta}_{h,t} - \beta_{h,t}\|^2\|x_{it}\|^4\|\delta_t(g_i^0, h)\|^2.$$

Let $V_{it} = \left(|u_{it}/\sigma_i|^2\|x_{it}\|^2 + \|x_{it}\|^4/\sigma_i^4\right)/\underline{s}_{N,T}^2$. Now, by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\left(\frac{\hat{d}_{it}^U(h) - d_{it}^U(h)}{\sigma_i s_{i,T}(h)}\right)^2 \\
\leq &C\left\{\max_{g\in\mathbb{G}}\left(\frac{1}{T}\sum_{t=1}^T\left\|\hat{\beta}_{g,t} - \beta_{g,t}\right\|^4\right)^{1/2} + \max_{g\in\mathbb{G}}\left(\frac{1}{T}\sum_{t=1}^T\left\|\hat{\beta}_{g,t} - \beta_{g,t}\right\|^8\right)^{1/2}\right\} \\
&\times\left(\frac{1}{T}\sum_{t=1}^T\left(\left|\frac{u_{it}}{\sigma_i}\right|^4\|x_{it}\|^4 + \|x_{it}\|^8/\sigma_i^4\right)/\underline{s}_{N,T}^4\right)^{1/2} \\
\leq &C(\gamma_{N,T,8}^2 + \gamma_{N,T,8}^4)\left(\frac{1}{T}\sum_{t=1}^T\left(V_{it}^2 - \mathbb{E}_P\left[V_{it}^2\right]\right) + \frac{1}{T}\sum_{t=1}^T\mathbb{E}_P\left[V_{it}^2\right]\right)^{1/2}.
\end{aligned}$$

Here, we note that $var(d_{it}^U(h)) \geq var(d_{it}(h))$ so that $s_{i,T}^U(h) \geq s_{i,T}(h) \geq \underline{s}_{N,T}$. Together with (26), this implies the desired result.

Proof of (ii): By slightly modifying the arguments above, we can prove

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^T\frac{\hat{d}_{it}^U(h) - d_{it}^U(h)}{\sigma_i s_{i,T}^U(h)}\right| \\
\leq &C(\gamma_{N,T,4} + \gamma_{N,T,4}^2)\left(\frac{1}{T}\sum_{t=1}^T\left(V_{it} - \mathbb{E}_P\left[V_{it}\right]\right) + \frac{1}{T}\sum_{t=1}^T\mathbb{E}_P\left[V_{it}\right]\right)^{1/2}.
\end{aligned}$$

And, for $0 < c < 1$,

$$P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^T\left(V_{it} - \mathbb{E}_P[V_{it}]\right)\right| > CT^{-(1-c)/2}B_{N,T,4}^2(\log N)\right) \leq CT^{-c}.$$

Thus it follows that

$$\begin{aligned}
P\Bigg(T^{-1/2}\max_{1\leq i\leq N}&\left|\frac{1}{\sqrt{T}}\sum_{t=1}^T\frac{d_{it}^U(g_i^0, h)}{\sigma_i s_{i,T}^U(g_i^0, h)} - \frac{1}{\sqrt{T}}\sum_{t=1}^T\frac{\hat{d}_{it}^U(g_i^0, h)}{\sigma_i s_{i,T}^U(g_i^0, h)}\right| \\
&> C\gamma_{N,T,8}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right)\Bigg) \leq CT^{-c}.
\end{aligned}$$

Proof of (iii): We observe that

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^T(\hat{d}_{it}^U(h) - \bar{\hat{d}}_i^U(h))^2 - \frac{1}{T}\sum_{t=1}^T(d_{it}^U(h) - \bar{d}_i^U(h))^2\right| \\
\leq &\frac{1}{T}\sum_{t=1}^T\left(\hat{d}_{it}^U(h) - d_{it}^U(h) - (\bar{\hat{d}}_i^U(h) - \bar{d}_i^U(h))\right)^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2\Big|\frac{1}{T}\sum_{t=1}^{T}\big(d_{it}^{U}(h)\big)\big(\hat{d}_{it}^{U}(h) - d_{it}^{U}(h) - (\bar{\hat{d}}_{i}^{U}(h) - \bar{d}_{i}^{U}(h))\big)\Big| \\
\leq &\frac{1}{T}\sum_{t=1}^{T}\big(\hat{d}_{it}^{U}(h) - d_{it}^{U}(h)\big)^2 \\
&+ 2\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(d_{it}^{U}(h)\big)^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\big(\hat{d}_{it}^{U}(h) - d_{it}^{U}(h)\big)^2}.
\end{aligned}
$$

Let

$$
U_{it}^{U}(h) = \frac{\big(d_{it}^{U}(h)\big)^2 - \mathbb{E}_P(\big(d_{it}^{U}(h)\big)^2)}{\sigma_i^2 s_{i,T}^2(h)}.
$$

Note that

$$
\sum_{t=1}^{T} U_{it}^{U}(h) = \frac{1}{T}\sum_{t=1}^{T}\frac{\big(d_{it}^{U}(h)\big)^2}{\sigma_i^2 s_{i,T}^2(h)} - \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}_P(\big(d_{it}^{U}(h)\big)^2)}{\sigma_i^2 s_{i,T}^2(h)}
$$

Because, $\mathbb{E}_P\left(\max_{1\leq i\leq N}\max_{1\leq t\leq T}(U_{it}^{U}(h))^2\right) \leq CTB_{N,T,4}^4$, following the same argument as the proof of Lemma D.7 part (i) gives

$$
P\left(\max_{1\leq i\leq n}\frac{1}{T}\sum_{t=1}^{T} U_{it}^{U}(h) > CB_{N,T,4}T^{-(1-c)/2}\log N\right) \leq CT^{-c}. \tag{32}
$$

Therefore, with probability at least $1 - CT^{-c}$,

$$
\begin{aligned}
&\Big|\frac{1}{T}\sum_{t=1}^{T}\frac{(\hat{d}_{it}^{U}(h) - \bar{\hat{d}}_{i}^{U}(h))^2}{\sigma_i^2 s_{i,T}^2(h)} - \frac{1}{T}\sum_{t=1}^{T}\frac{(d_{it}^{U}(h) - \bar{d}_{i}^{U}(h))^2}{\sigma_i^2 s_{i,T}^2(h)}\Big| \\
\leq &\frac{1}{T}\sum_{t=1}^{T}\frac{\big(\hat{d}_{it}^{U}(h) - d_{it}^{U}(h)\big)^2}{\sigma_i^2 \underline{s}_{N,T}^2} \\
&+ \max_{h\in\mathbb{G}} 2\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\big(d_{it}^{U}(h)\big)^2}{\sigma_i^2 s_{i,T}^2(h)}}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\big(\hat{d}_{it}^{U}(h) - d_{it}^{U}(h)\big)^2}{\sigma_i^2 \underline{s}_{N,T}^2}} \\
\leq &\gamma_{N,T,8}\big(T^{-(1-c)/4}\sqrt{\log N}B_{N,T,8}^2 + D_{N,T,4}\big)\big(D_{N,T,2} + T^{-(1-c)/4}\sqrt{\log N}B_{N,T,4}\big) \\
&+ \gamma_{N,T,8}^2\big(T^{-(1-c)/2}(\log N)B_{N,T,8}^4 + D_{N,T,4}^2\big),
\end{aligned}
$$

where the last inequality follows from Lemma E.1 part (i) and (32).

Proof of (iv): Define

$$
S_{i,T}^{U\Delta}(h) = \left(\frac{\hat{S}_{i,T}^{U}(h) - S_{i,T}^{U}(h)}{\sigma_i s_{i,T}(h)}\right)\frac{S_{i,T}^{U}(h)}{\sigma_i s_{i,T}(h)}.
$$

By the inequality $|a-b| \le |a-b|/(\sqrt{a}+\sqrt{b}) \le |a-b|/\sqrt{a}$ and part (iii) of the lemma, we have

$$S_{i,T}^{U\Delta}(h) \le \left| \left(\hat{S}_{i,T}^2(h)/(\sigma_i s_{i,T}(h))\right)^2 - \left(S_{i,T}^2(h)/(\sigma_i s_{i,T}(h))\right)^2 \right| \le C_2 \zeta_{N,T}$$

uniformly over $i = 1, \ldots, N$ on a set of probability less than CT^{-c}. By the inequality $|\sqrt{a}-1| \le |a-1|$ and Lemma D.7 we have

$$\left|(S_{i,T}^U(h))^2/(\sigma_i s_{i,T}(h)) - 1\right| \le \left| \left((S_{i,T}^U(h))^2/(\sigma_i s_{i,T}(h))\right)^2 - 1 \right| \le C_1 T^{-(1-c)/2} (\log N) B_{N,T,4}^2$$

uniformly over $i = 1, \ldots, N$ on a set of probability less than CT^{-c}. Note that

$$\begin{aligned} \left|D_i^U(g_i^0, h)\right| &\le \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{d_{it}^U(h) - E_P(d_{it}^U(h))}{\sigma_i s_{i,T}(h)} \right| + \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{E_P(d_{it}^U(h))}{\sigma_i s_{i,T}(h)} \right| \\ &\le \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{d_{it}^U(h) - E_P(d_{it}^U(h))}{\sigma_i s_{i,T}(h)} \right| + D_{N,T,1} \end{aligned}$$

Thus, by following the same argument as that in the proof of Lemma D.7 part (ii), it holds that

$$\left|D_i^U(h)\right| \le D_{N,T,1} + C\left(\sqrt{\log N} + T^{-1/4} B_{N,T,4} \log N\right)$$

uniformly over $i = 1, \ldots, N$ on a set of probability less than $N^{-1} + C\left(T^{-1/4} B_{N,T,4}/\log(N)\right)^4$. Now, decompose

$$\begin{aligned} \hat{D}_i^U(h) - \tilde{D}_i^U(h) =& \frac{\sigma_i s_{i,T}(h)}{\hat{S}_{i,T}^U(h)} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{\hat{d}_{it}^U(h)}{\sigma_i s_{i,T}(h)} - D_i^U(h) \right) \\ &- \frac{\hat{S}_{i,T}^U(h) - S_{i,T}^U(h)}{S_{i,T}^U(h) \hat{S}_{i,T}^U(h)} \sigma_i s_{i,T}(h) D_i^U(h) \\ =& \frac{S_{i,T}^U(h)/(\sigma_i s_{i,T}(h))}{S_{i,T}^{U\Delta}(h) + S_{i,T}^U(h)/(\sigma_i s_{i,T}(h))} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{\hat{d}_{it}^U(h)}{\sigma_i s_{i,T}(h)} - D_i^U(h) \right) \\ &- \frac{S_{i,T}^{U\Delta}}{\left(S_{i,T}^{U\Delta} + (S_{i,T}^U(h))^2/(\sigma_i s_{i,T}(h))^2\right)(S_{i,T}^U(h))^2/(\sigma_i s_{i,T}(h))} D_i^U(h). \end{aligned}$$

In conjunction with part (ii) of the lemma this decomposition implies

$$\begin{aligned} \max_{1 \le i \le N} \left| \hat{D}_i^U(h) - \tilde{D}_i^U(h) \right| \le& C \gamma_{N,T,8} \sqrt{T} \left(T^{-(1-c)/4} B_{N,T,4} \sqrt{\log N} + D_{N,T,4}^2 \right) \\ &+ C \zeta_{N,T}^U \left(D_{N,T,1} + \sqrt{\log N} + T^{-1/4} B_{N,T,4} \log N \right) \end{aligned}$$

with probability less than $CT^{-c} + N^{-1} + C\left(T^{-1/4} B_{N,T,4}/\log(N)\right)^4$.

□

Proof of Theorem 4. We note that the hypothesis selection part of the procedure does not affect the theoretical analysis. This is because, here, we focus on size and thus need to consider only the behavior of the test statistics under $\{g_i^0\}_{i=1}^N$.

Let

$$J_1 = \left\{ (i,h) \mid i \in \{1,\dots,N\}, h \in \mathbb{G}\backslash\{g_i^0\}, \frac{\sqrt{T}\mathbb{E}_P(\bar{d}_i^U(g_i^0,h))}{\sigma_i s_{i,T}^U(g,h)} > -c_{\beta,N}^{\text{SNS}} \right\}$$

Roughly speaking, J_1 is the set of pairs of units and groups that are difficult to distinguish from true group membership.

In this proof, we set $c = 1/6$.

Step 1: We first prove that $P\left(\max_{(i,h)\in J_1^c} \hat{\bar{d}}_i^U(g_i^0,h) \le 0\right) > 1-\beta - CT^{-c}$.

Note that $\hat{\bar{d}}_i^U(g_i^0,h) > 0$ for some $(i,h) \in J_1^c$ implies that

$$\max_{(i,h)\in J_1} \frac{\sqrt{T}(\hat{\bar{d}}_i^U(g_i^0,h)) - \mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)} > c_{\beta,N}^{\text{SNS}}.$$

Let

$$c_{SN}(\beta) = \frac{\Phi^{-1}(1-\beta/((G-1)N))}{\sqrt{1-\Phi^{-1}(1-\beta/((G-1)N))^2/T}}.$$

Let

$$\epsilon_{N,T,1}^U = \sqrt{T}\gamma_{N,T,8}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right).$$

We have

$$\begin{aligned}
&P\left(\max_{(i,h)\in J_1} \frac{\sqrt{T}(\hat{\bar{d}}_i^U(g_i^0,h)) - \mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)} > c_{\beta,N}^{\text{SNS}}\right) \\
\le& P\left(\max_{(i,h)\in J_1} \frac{\sqrt{T}(\bar{d}_i^U(g_i^0,h)) - \mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)} > c_{\beta,N}^{\text{SNS}} - \epsilon_{N,T,1}^U\right) \\
&+ P\left(\max_{(i,h)\in J_1} \left| \frac{\sqrt{T}(\hat{\bar{d}}_i^U(g_i^0,h)) - \bar{d}_i^U(g_i^0,h))}{\sigma_i s_{i,T}^U(g,h)} \right| > \epsilon_{N,T,1}^U\right).
\end{aligned}$$

The second term on the right-hand side is bounded by CT^{-c} by Lemma E.1 part (ii). Let β_N solve $c_{\beta_N,N}^{\text{SNS}} = c_{\beta,N}^{\text{SNS}} - \epsilon_{N,T,1}^U$. As in the proof of Theorem 1, we have

$$|\beta_N - \beta| \le 4\epsilon_{N,T,1}^U\sqrt{\log((G-1)N/\beta)}.$$

Thus we have

$$\begin{aligned}
&P\left(\max_{(i,h)\in J_1} \frac{\sqrt{T}(\hat{\bar{d}}_i^U(g_i^0,h)) - \mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)} > c_{\beta,N}^{\text{SNS}}\right) \\
\le& P\left(\max_{(i,h)\in J_1} \frac{\sqrt{T}(\bar{d}_i^U(g_i^0,h)) - \mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)} > c_{\beta_N,N}^{\text{SNS}}\right) + CT^{-c}
\end{aligned}$$

$$=P\left(\max_{(i,h)\in J_1}\frac{\sqrt{T}(\bar{d}_i^U(g_i^0,h))-\mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)}>c_{SN}(c_{SN}^{-1}(c_{\beta_N,N}^{\text{SNS}}))\right)+CT^{-c}.$$

Following essentially the same argument as that in Step 1 of the proof of Theorem 4.2 of Chernozhukov, Chetverikov, and Kato (2014) shows that, under Assumptions (8) and (9),

$$P\left(\max_{(i.h)\in J_1}\frac{\sqrt{T}(\bar{d}_i^U(g_i^0,h))-\mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)}>c_{SN}(c_{SN}^{-1}(c_{\beta_N,N}^{\text{SNS}}))\right)\leq c_{SN}^{-1}(c_{\beta_N,N}^{\text{SNS}})+CT^{-c}.$$

Note that here we replace $\hat{\sigma}_j$ and σ_j in the proof of Chernozhukov, Chetverikov, and Kato (2014) with $(T^{-1}\sum_{t=1}^T(d_{it}^U(g_i^0,h)-\mathbb{E}_P(d_{it}^U(g_i^0,h))))^{1/2}$ and $\sigma_i s_{i,T}^U(g,h)$. We have

$$\begin{aligned}c_{SN}^{-1}(c_{\beta_N,N}^{\text{SNS}})=&(G-1)N\left(1-\Phi\left(\frac{c_{\beta_N,N}^{\text{SNS}}}{\sqrt{1+(c_{\beta_N,N}^{\text{SNS}})^2/T}}\right)\right)\\=&\beta+O\left(\frac{(c_{\beta_N,N}^{\text{SNS}})^3}{\sqrt{T}}\right).\end{aligned}$$

We thus have

$$\begin{aligned}&P\left(\max_{(i,h)\in J_1}\frac{\sqrt{T}(\bar{d}_i^U(g_i^0,h))-\mathbb{E}_P(d_{it}^U(g_i^0,h)))}{\sigma_i s_{i,T}^U(g,h)}>\sqrt{\frac{T}{T-1}}t_{T-1}^{-1}\left(1-\frac{\beta_N}{(G-1)N}\right)\right)\\\leq&\beta_N+O\left(\frac{(c_{\beta_N,N}^{\text{SNS}})^3}{\sqrt{T}}\right)+CT^{-c}\leq\beta+CT^{-c},\end{aligned}$$

where $(c_{\beta_N,N}^{\text{SNS}})^3/\sqrt{T}\leq CT^{-c}$ by that $(\log(N))^6/T\leq CT^{-c}$ which is implied by (8) together with $D_{N,T,3}\geq 1$ and Lemma D.10, and $\epsilon_{N,T,1}^U\sqrt{\log((G-1)N/\beta)}\leq CT^{-c}$ by assumption (10).

An implication of Step 1 is as follows. Let

$$\mathbb{N}=\left\{i\in\{1,\dots,N\}\mid\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\frac{\sqrt{T}\mathbb{E}_P(\bar{d}_i^U(g_i^0,h))}{\sigma_i s_{i,T}^U(g,h)}>-c_{\beta,N}^{\text{SNS}}\right\}.$$

Then

$$P\left(\max_{i\in\mathbb{N}^c}\max_{h\in\mathbb{G}\setminus\{g_i^0\}}\bar{\hat{d}}_i^U(g_i^0,h)\leq 0\right)>1-\beta-CT^{-c}.$$

Step 2: Next, we prove that $P(\bigtimes_{i=1}^N\hat{M}_i(g_i^0)\supseteq J_1)\geq 1-\beta-CT^{-c}$. Here, we drop the g argument for simplicity of notation when arguments are g_i^0 and h.

We note that

$$\begin{aligned}&P\left(\bigtimes_{i=1}^N\hat{M}_i(g_i^0)\not\supseteq J_1\right)\\=&P\left(\exists(i,h);\hat{D}_i^U(h)\leq -2c_{\beta_N,N}^{\text{SNS}}\text{ and }\frac{\sqrt{T}\mathbb{E}_P(\bar{d}_i^U(h))}{\sigma_i s_{i,T}^U(h)}>-c_{\beta,N}^{\text{SNS}}\right)\end{aligned}$$

$$
\begin{aligned}
\le & P\left(\exists(i,h); \tilde{D}_i^U(h) \le -2c_{\beta,N}^{\mathrm{SNS}} + \epsilon_{N,T,2}^U \text{ and } \frac{\sqrt{T}\mathbb{E}_P(\bar{d}_i^U(h))}{\sigma_i s_{i,T}^U(h)} > -c_{\beta,N}^{\mathrm{SNS}}\right) \\
& + P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g_i^0\}}\left|\hat{D}_i^U(h) - \tilde{D}_i^U(h)\right| > \epsilon_{N,T,2}^U\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\epsilon_{N,T,2}^U = & C\gamma_{N,T,8}\sqrt{T}\left(T^{-(1-c)/4}B_{N,T,4}\sqrt{\log N} + D_{N,T,2}\right) \\
& + C\zeta_{N,T}^U\left(D_{N,T,1} + \sqrt{\log N} + T^{-1/4}B_{N,T,4}\sqrt{\log N}\right).
\end{aligned}
$$

By part (iv) of Lemma E.1, noting that its condition is satisfied by (9), (10) and (11),

$$
\begin{aligned}
P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g_i^0\}}\left|\hat{D}_i^U(h) - \tilde{D}_i^U(h)\right| > \epsilon_{N,T,2}^U\right) < & N^{-1} + CT^{-c} + C\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4 \\
\le & N^{-1} + CT^{-c}.
\end{aligned}
$$

where the second inequality follows because (9) implies $\big(T^{-1/4}B_{N,T,4}/\log(N)\big)^4 \le T^{-1/6}$.

We observe

$$
\begin{aligned}
& P\left(\exists(i,h); \tilde{D}_i^U(h) \le -2c_{\beta,N}^{\mathrm{SNS}} + \epsilon_{N,T,2}^U \text{ and } \frac{\sqrt{T}\mathbb{E}_P(\bar{d}_i^U(h))}{\sigma_i s_{i,T}^U(h)} > -c_{\beta,N}^{\mathrm{SNS}}\right) \\
\le & P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g_i^0\}}\left[\sqrt{T}(E(\bar{d}_i^U(h)) - \bar{d}_i^U(h)) - (2S_{i,T}^U(h) - \sigma_i s_{i,T}^U(h))c_{\beta,N}^{\mathrm{SNS}} + 2S_{i,T}^U(h)\epsilon_{N,T,2}^U\right] > 0\right).
\end{aligned}
$$

Let

$$
(\tilde{S}_{i,T}^U(h))^2 = \frac{1}{T}\sum_{t=1}^T(d_{it}^U(h) - \mathbb{E}_P(d_{it}^U(h)))^2 - \left(\frac{1}{T}\sum_{t=1}^T(d_{it}^U(h) - \mathbb{E}_P(d_{it}^U(h)))\right)^2.
$$

We observe that

$$
\begin{aligned}
(S_{i,T}^U(h))^2 = & \frac{1}{T}\sum_{t=1}^T(d_{it}^U(h) - \mathbb{E}_P(d_{it}^U(h)))^2 \\
& + \frac{2}{T}\sum_{t=1}^T(d_{it}^U(h) - \mathbb{E}_P(d_{it}^U(h)))(\mathbb{E}_P(d_{it}^U(h)) - \mathbb{E}_P(\bar{d}_i^U(h))) \\
& - \left(\frac{1}{T}\sum_{t=1}^T(d_{it}^U(h) - \mathbb{E}_P(d_{it}^U(h)))\right)^2 \\
& + \frac{1}{T}\sum_{t=1}^T(\mathbb{E}_P(d_{it}^U(h)) - \mathbb{E}_P(\bar{d}_i^U(h)))^2 \\
\ge & (\tilde{S}_{i,T}(h))^2 + \frac{2}{T}\sum_{t=1}^T(d_{it}^U(h) - \mathbb{E}_P(d_{it}^U(h)))(\mathbb{E}_P(d_{it}^U(h)) - \mathbb{E}_P(\bar{d}_i^U(h))).
\end{aligned}
$$

If $1-\sigma_i s^U_{i,T}(h)/\tilde{S}^U_{i,T}(h)\geq -r/2$ and

$$\frac{2}{T}\sum_{t=1}^{T}(d^U_{it}(h)-\mathbb{E}_P(d^U_{it}(h)))(\mathbb{E}_P(d^U_{it}(h))-\mathbb{E}_P(\bar{d}^U_i(h)))\geq -(\tilde{S}_{i,T}(h))^2\left(\frac{r}{2}-\frac{r^2}{16}\right),$$

for some $0<r<1$, we have

$$2S^U_{i,T}(h)-\sigma_i s^U_{i,T}(h)\geq(1-r)\tilde{S}^U_{i,T}(h)$$

because

$$\begin{aligned}2S^U_{i,T}(h)-\sigma_i s^U_{i,T}(h)\geq& 2\tilde{S}^U_{i,T}(h)\left(1-\frac{r}{2}-\frac{r^2}{16}\right)^{1/2}-\sigma_i s^U_{i,T}(h)\\ =&\tilde{\sigma}_{i,h}\left(2\left(1-\frac{r}{4}\right)-\frac{\sigma_i s^U_{i,T}(h)}{\tilde{S}^U_{i,T}(h)}\right)\geq(1-r)\tilde{S}^U_{i,T}(h).\end{aligned}$$

We thus have

$$\begin{aligned}&P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g^0_i\}}\left[\sqrt{T}(E(\bar{d}^U_i(h))-\bar{d}^U_i(h))-(2S^U_{i,T}(h)-\sigma_i s^U_{i,T}(h))c^{\mathrm{SNS}}_{\beta,N}+2S^U_{i,T}(h)\epsilon^U_{N,T,2}\right]>0\right)\\ \leq&P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\frac{\sqrt{T}(E(\bar{d}^U_i(h))-\bar{d}^U_i(h))}{\tilde{S}^U_{i,T}(h)}>(1-r)c^{\mathrm{SNS}}_{\beta,N}-2\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\frac{S^U_{i,T}(h)}{\tilde{S}^U_{i,T}(h)}\epsilon^U_{N,T,2}\right)\quad(33)\\ &+P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\left|\frac{2}{T}\sum_{t=1}^{T}\tilde{a}_{it}(h)\right|>\frac{r}{2}-\frac{r^2}{16}\right)\quad(34)\\ &+P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\left|\frac{\sigma_i s^U_{i,T}(h)}{\tilde{S}^U_{i,T}(h)}-1\right|>\frac{r}{2}\right),\quad(35)\end{aligned}$$

where

$$\tilde{a}_{it}(h)=2(d^U_{it}(h)-\mathbb{E}_P(d^U_{it}(h)))(\mathbb{E}_P(d^U_{it}(h))-\mathbb{E}_P(\bar{d}^U_i(h)))/(\tilde{S}^U_{i,T}(h))^2.$$

We now take $r=T^{-(1-c)/2}B^2_{T,N,4}\log((G-1)N)$.

The first term of (33) is

$$\begin{aligned}&P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\frac{\sqrt{T}(E(\bar{d}^U_i(h))-\bar{d}^U_i(h))}{\tilde{S}^U_{i,T}(h)}>(1-r)c^{\mathrm{SNS}}_{\beta,N}-2\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\frac{S^U_{i,T}(h)}{\tilde{S}^U_{i,T}(h)}\epsilon^U_{N,T,2}\right)\\ \leq&P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\frac{\sqrt{T}(E(\bar{d}^U_i(h))-\bar{d}^U_i(h))}{\tilde{S}^U_{i,T}(h)}>(1-r)c^{\mathrm{SNS}}_{\beta,N}-C\epsilon^U_{N,T,2}\right)\\ &+P\left(\left|\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\frac{S^U_{i,T}(h)}{\tilde{S}^U_{i,T}(h)}\right|>\frac{1}{2}C\right).\end{aligned}$$

Note that we can take $C > 2$ and

$$P\left(\left|\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g(i)\}}\frac{S_{i,T}^U(h)}{\tilde{S}_{i,T}^U(h)}\right| > \frac{1}{2}C\right) < CT^{-c}$$

holds because

$$\frac{S_{i,T}^U(h)}{\tilde{S}_{i,T}^U(h)} = \frac{S_{i,T}^U(h)}{\sigma_i s_{i,T}^U(h)}\frac{\sigma_i s_{i,T}^U(h)}{\tilde{S}_{i,T}^U(h)},$$

Lemma E.1 part (iii) and following the same argument of Lemma A.5 of Chernozhukov, Chetverikov, and Kato (2014). Following the argument in the proof of Step 2 of Theorem 4.2 of Chernozhukov, Chetverikov, and Kato (2014) under (8), (9) and that (10) and (11) implies $\epsilon_{N,T,2}^U\sqrt{\log((G-1)N/\beta)} \le CT^{-1/6}$, it holds that

$$P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g(i)\}}\frac{\sqrt{T}(E(\bar{d}_i^U(h)) - \bar{d}_i^U(h))}{\tilde{S}_{i,T}^U(h)} > (1-r)c_{\beta,N}^{\text{SNS}} - C\epsilon_{N,T,2}^U\right) \le \beta + CT^{-c}.$$

For the second term (34), let $a_{it}(h) = 2(d_{it}^U(h) - \mathbb{E}_P(d_{it}^U(h)))(\mathbb{E}_P(d_{it}^U(h)) - \mathbb{E}_P(\bar{d}_i^U(h)))/(\sigma_i s_{i,T}^U(h))^2$. The second term is

$$\begin{aligned}
&P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g(i)\}}\left|\frac{1}{T}\sum_{t=1}^T \tilde{a}_{it}(h)\right| > \frac{r}{2} - \frac{r^2}{16}\right)\\
\le& P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g(i)\}}\left|\frac{1}{T}\sum_{t=1}^T a_{it}(h)\right| > \left(1-\frac{r}{2}\right)\left(\frac{r}{2} - \frac{r^2}{16}\right)\right)\\
&+ P\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g(i)\}}\left|\frac{(\tilde{S}_{i,T}^U(h))^2}{(\sigma_i s_{i,T}^U(h))^2} - 1\right| > \frac{r}{2}\right),
\end{aligned}$$

where the inequality holds because $(\tilde{S}_{i,T}^U(h))^2 \ge (1-r/2)(\sigma_i s_{i,T}^U(h))^2$ if $1 - (\tilde{S}_{i,T}^U(h))^2/(\sigma_i s_{i,T}^U(h))^2 > -r/2$. The second term is bounded by CT^{-c} by Lemma A.5 of Chernozhukov, Chetverikov, and Kato (2014) (Note that the statement of Lemma A.5 of Chernozhukov, Chetverikov, and Kato (2014) is about $\hat{\sigma}_j/\sigma_j$ (in their notation) but their proof is based on $\hat{\sigma}_j^2/\sigma_j^2$). For the first term, observe that

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_P((a_{it}(h)/T)^2) =& \frac{1}{T^2}\sum_{t=1}^T \frac{var(d_{it}^U(h))}{(\sigma_i s_{i,T}^U(h))^4}(\mathbb{E}_P(d_{it}^U(h)) - \mathbb{E}_P(\bar{d}_i^U(h)))^2\\
\le& \frac{1}{T^2}\sum_{t=1}^T \frac{(\mathbb{E}_P(d_{it}^U(h)) - \mathbb{E}_P(\bar{d}_i^U(h)))^2}{(\sigma_i s_{i,T}^U(h))^2},
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{t=1}^T E\left(\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g(i)\}}(a_{it}(h)/T)^2\right)\\
\le& \frac{1}{T^2}\sum_{t=1}^T B_{T,N,4}^2\max_{1\le i\le N}\max_{h\in\mathbb{G}\backslash\{g(i)\}}(\mathbb{E}_P(d_{it}^U(g_i^0,h)) - \mathbb{E}_P(\bar{d}_i^U(g_i^0,h)))^2/(\sigma_i s_{i,T}^U(h))^2
\end{aligned}$$

$$\leq \frac{1}{T} G B_{T,N,4}^2 D_{N,T,2}^2.$$

By Leamm A.3 of Chernozhukov, Chetverikov, and Kato (2014), we have

$$E\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\left|\frac{1}{T}\sum_{t=1}^{T}a_{it}(h)\right|\right)\leq CD_{T,N,2}\left(\frac{\sqrt{\log((G-1)N)}}{\sqrt{T}}+B_{T,N,4}\frac{\log((G-1)N)}{T}\right).$$

By Lemma A.2 of Chernozhukov, Chetverikov, and Kato (2014), we thus have

$$P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\left|\frac{1}{T}\sum_{t=1}^{T}a_{it}(h)\right|\geq CD_{T,N,2}\left(\frac{\sqrt{\log((G-1)N)}}{\sqrt{T}}+B_{T,N,4}\frac{\log((G-1)N)}{T}\right)+t\right)$$
$$\leq e^{-t^2/(3(D_{T,N,2}^2/T)}+\frac{K}{t^2}\frac{1}{T}B_{T,N,4}^2D_{T,N,2}^2,$$

for any $t>0$. Taking $t=T^{-(1-c)/2}D_{T,N,2}B_{T,N,4}$ and arranging the terms, we have

$$P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\left|\frac{1}{T}\sum_{t=1}^{T}a_{it}(h)\right|\geq CD_{T,N,2}B_{T,N,4}T^{-(1-c)/2}\log((G-1)N)\right)\leq CT^{-c}.$$

We thus have

$$P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\left|\frac{1}{T}\sum_{t=1}^{T}\tilde{a}_{it}(h)\right|>\frac{r}{2}-\frac{r^2}{16}\right)\leq CT^{-c},$$

by Assumption (9).

The third term (35) can also be analyzed by following the argument in the proof of Step 2 of Theorem 4.2 of Chernozhukov, Chetverikov, and Kato (2014) and is bounded by $\beta+CT^{-c}$ under Assumptions (8) and (9).

Summing up, we have

$$P\left(\bigtimes_{i=1}^{N}\hat{M}_i(g_i^0)\not\supseteq J_1\right)\leq\beta+CT^{-c}+N^{-1}.$$

An implication of Step 2 is as follows. Let

$$\hat{\mathbb{N}}=\left\{i\in\{1,\dots,N\}\mid M_i(g_i^0)\neq\varnothing\right\}.$$

Then

$$P\left(\hat{\mathbb{N}}\supseteq\mathbb{N}\right)\geq 1-\beta-CT^{-c}-N^{-1}.$$

Step 3: First, consider the case in which $J_1=\varnothing$. In this case, the argument in Step 1 yields that

$$P(\hat{g}_i=g_i^0,\forall i)=P\left(\max_{1\leq i\leq N}\max_{h\in\mathbb{G}\setminus\{g(i)\}}\hat{D}_i^U(g_i^0,h)\leq 0\right)>1-\beta-CT^{-c}.$$

Because $\{\hat{g}(i)\}_{i=1}^N$ is always included in the confidence set, the probability of the confidence set not including $\{g_i^0\}_{i=1}^N$ is less than $\beta+CT^{-c}<\alpha+CT^{-c}$.

Next, consider the case in which $|J_1| \geq 1$. Here, we consider the case with type = SNS. The proofs for the other two cases are similar, and therefore omitted. Observe that

$$
\begin{aligned}
&P\left(\{g_i^0\}_{i=1}^N \notin \hat{C}_{\mathrm{Sel},\alpha,\beta}^{\mathrm{SNS}}\right) \\
=&P\left(\bigcup_{i=1}^N \left(\left\{\hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha-2\beta,\hat{N}}^{\mathrm{SNS}}\right\} \cap \left\{\max_{h\in\mathbb{G}\setminus\{g_i^0\}} \hat{D}_i^U(g_i^0,h) > 0\right\}\right)\right) \\
\leq&P\left(\bigcup_{i\in\mathbb{N}} \left\{\hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha-2\beta,\hat{N}}^{\mathrm{SNS}}\right\} \cup \bigcup_{i\in\mathbb{N}^c} \left\{\max_{h\in\mathbb{G}\setminus\{g_i^0\}} \hat{D}_i^U(g_i^0,h) > 0\right\}\right) \\
\leq&P\left(\bigcup_{i\in\mathbb{N}} \left\{\hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha-2\beta,\hat{N}}^{\mathrm{SNS}}\right\}\right) \\
&+ P\left(\bigcup_{i\in\mathbb{N}^c} \left\{\max_{h\in\mathbb{G}\setminus\{g_i^0\}} \hat{D}_i^U(g_i^0,h) > 0\right\}\right).
\end{aligned}
$$

By Step 1, we have

$$
P\left(\bigcup_{i\in\mathbb{N}^c} \left\{\max_{h\in\mathbb{G}\setminus\{g_i^0\}} \hat{D}_i^U(g_i^0,h) > 0\right\}\right) \leq \beta + CT^{-c}.
$$

By Step 2, we have

$$
\begin{aligned}
&P\left(\bigcup_{i\in\mathbb{N}} \left\{\hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha-2\beta,\hat{N}}^{\mathrm{SNS}}\right\}\right) \\
\leq&P\left(\{\hat{\mathbb{N}} \supseteq \mathbb{N}\} \cap \bigcup_{i\in\mathbb{N}} \left\{\hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha-2\beta,\hat{N}}^{\mathrm{SNS}}\right\}\right) + P(\{\hat{\mathbb{N}} \not\supseteq \mathbb{N}\}) \\
\leq&P\left(\bigcup_{i\in\mathbb{N}} \left\{\hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha-2\beta,|\mathbb{N}|}^{\mathrm{SNS}}\right\}\right) + \beta + CT^{-c} + N^{-1}.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&P\left(\{g_i^0\}_{i=1}^N \notin \hat{C}_{\mathrm{Sel},\alpha,\beta}^{\mathrm{SNS}}\right) \\
\leq&P\left(\bigcup_{i\in\mathbb{N}} \left\{\hat{T}_i^{\mathrm{MAX}}(g_i^0) > c_{\alpha-2\beta,|\mathbb{N}|}^{\mathrm{SNS}}\right\}\right) + 2\beta + CT^{-c} + N^{-1}.
\end{aligned}
$$

Theorem 1 implies that

$$
P\left(\{g_i^0\}_{i=1}^N \notin \hat{C}_{\mathrm{Sel},\alpha,\beta}^{\mathrm{SNS}}\right) \leq \alpha + C\epsilon_N + CT^{-c} + N^{-1}.
$$

□

g^0	σ	T	empirical coverage			cardinality of CS		
			SNS	MAX	QLR	SNS	MAX	QLR
1	0.25	10	0.99	0.99	0.99	2.99	2.99	2.96
1	0.25	20	0.96	0.97	0.96	2.23	2.17	2.05
1	0.25	30	0.95	0.95	0.94	1.63	1.58	1.53
1	0.25	40	0.94	0.94	0.94	1.42	1.39	1.36
1	0.50	10	0.98	0.98	0.94	2.99	2.99	2.97
1	0.50	20	0.95	0.96	0.93	2.90	2.88	2.85
1	0.50	30	0.94	0.94	0.94	2.80	2.78	2.76
1	0.50	40	0.92	0.93	0.92	2.74	2.73	2.70
2	0.25	10	1.00	1.00	1.00	2.99	3.00	2.99
2	0.25	20	0.98	0.97	0.97	2.48	2.43	2.38
2	0.25	30	0.96	0.96	0.95	1.69	1.62	1.59
2	0.25	40	0.97	0.94	0.93	1.32	1.29	1.29
2	0.50	10	0.98	0.98	0.96	2.99	2.99	2.99
2	0.50	20	0.97	0.96	0.95	2.89	2.88	2.85
2	0.50	30	0.95	0.93	0.94	2.75	2.72	2.70
2	0.50	40	0.95	0.94	0.92	2.62	2.59	2.58
3	0.25	10	1.00	0.99	0.99	3.00	3.00	2.99
3	0.25	20	0.98	0.98	0.97	2.51	2.48	2.44
3	0.25	30	0.97	0.97	0.94	1.70	1.63	1.61
3	0.25	40	0.96	0.96	0.96	1.32	1.29	1.30
3	0.50	10	0.98	0.99	0.97	2.99	2.99	2.99
3	0.50	20	0.97	0.96	0.94	2.90	2.89	2.86
3	0.50	30	0.96	0.96	0.93	2.75	2.73	2.70
3	0.50	40	0.94	0.93	0.93	2.63	2.60	2.58

Table F.1: Homoscedastic design with $G = 3$ groups. Results based on $B = 1000$ simulated joint confidence sets with $1 - \alpha = 0.9$. Critical values for MAX and QLR procedures are adjusted for short panels. "Empirical coverage" gives the simulated coverage probability of the joint confidence set. "Cardinality of CS" gives the simulated expected average cardinality of a marginal (unit-wise) confidence set.

F. More simulation results

F.1. Another homoscedastic design with $G = 3$ groups

This design is defined exactly as that from Section 6.1 with the exception of defining a different set of group-specific coefficients. Let $\varphi_T^{(2)}(t) = -2 + 8\,|t - T/2|\,/T$. For $t = 1, \dots, T$, $\alpha_{1,t} = 0$, $\alpha_{2,t} = \varphi_T^{(2)}(t)$, $\alpha_{3,t} = \varphi_{T/2}^{(2)}(t \bmod \lceil T/2 \rceil)$. This specification implies moment inequalities that are less correlated than those for the design in Section 6.1. For example, for $T = 40$ and $g^0 = 1$, our simulations indicate that $(\mathbb{E}\,\widehat{\Omega}_i(1))_{1,2} = 0.00$ and $(\mathbb{E}\,\widehat{\Omega}_i(2))_{1,2} = 0.68$. For $T = 40$ and $g^0 = 2$, $(\mathbb{E}\,\widehat{\Omega}_i(1))_{1,2} = -0.00$ and $(\mathbb{E}\,\widehat{\Omega}_i(2))_{1,2} = 0.69$. We simulate $B = 1000$ joint confidence sets based on the SNS, MAX (with short-panel adjustment), and QLR (with short-panel adjustment) approach. The simulation results are reported in Table F.1.