BOOTSTRAP WITH CLUSTER-DEPENDENCE IN TWO OR MORE DIMENSIONS

KONRAD MENZEL
NEW YORK UNIVERSITY

ABSTRACT. We propose a bootstrap procedure for data that may exhibit cluster dependence in two or more dimensions. The asymptotic distribution of the sample mean or other statistics may be non-Gaussian if observations are dependent but uncorrelated within clusters. We show that there exists no procedure for estimating the limiting distribution of the sample mean under two-way clustering that achieves uniform consistency. However, we propose bootstrap procedures that achieve adaptivity with respect to different uniformity criteria. Important cases and extensions discussed in the paper include regression inference, U- and V-statistics, subgraph counts for network data, and non-exhaustive samples of matched data.

JEL Classification: C1, C12, C23, C33
Keywords: Multi-Way Cluster-Dependence, Wild Bootstrap, U-Statistics, Network Data

1. INTRODUCTION

We consider inference based on a random array (Y_{it}) that is indexed by two dimensions, where the indices $i = 1, \ldots, N$ (and $t = 1, \ldots, T$, respectively) correspond to units ("clusters") that are sampled independently at random from an infinite population, but we allow for otherwise unrestricted dependence within each row $\mathbf{Y}_{i\cdot} := (Y_{i1}, \ldots, Y_{iT})$, and within each column $\mathbf{Y}_{\cdot t} := (Y_{1t}, \ldots, Y_{Nt})$. There are various contexts in which a researcher may encounter data with cluster-dependence along multiple dimensions:

Example 1.1. ***Cluster-Dependence.*** *Cross-sectional data may be organized among multiple dimensions, e.g. a worker simultaneously pertains to a certain geographic labor market, industry, and occupation. Dependence within any of these groups may result e.g. from common economic shocks, or other group-level variables, see Moulton (1990). Cameron, Gelbach, and Miller (2011) give a more comprehensive account of settings in empirical practice for which cluster-dependence may result from sampling or other design decisions.*

Date: November 2016 - this version: February 2021. The author thanks Colin Cameron, Matias Cattaneo, Tim Christensen, Iván Fernández-Val, Joachim Freyberger, Bryan Graham, James MacKinnon, Mikkel Sølvsten, and Valentin Verdier for useful conversations, as well as the co-editor, Ulrich Müller, and four anonymous referees for useful comments and suggestions. Financial support from the NSF (SES-1459686) is also gratefully acknowledged.

Example 1.2. ***Static panels, Difference-in-differences.*** *One interpretation of this setup is a panel in which cross-sectional units are observed over time, and the outcome of interest is subject both to common aggregate shocks that are serially independent and unit-level heterogeneity.*[1] *Two-way heterogeneity of this form is a characteristic feature of classical difference-in-differences designs that aim to control for temporal shocks as well as unobserved heterogeneity. Our framework does not restrict the number of distinct shocks, or how they may interact in a generative model for the outcome variable* Y_{it}.

Example 1.3. ***Matched data.*** *For matched samples between different groups of units* $i = 1, \ldots, N$ *and* $t = 1, \ldots, T$, *respectively,* Y_{it} *measures an outcome at the level of the match. This setup includes test scores for a random sample of students and teachers, or wages (marginal product of labor) for a random sample of workers and firms. In such a setting we often observe* Y_{it} *only for a subset of the possible dyads* (i, t) *(non-exhaustively matched samples). We discuss an adaptation of our bootstrap method to non-exhaustively matched data in Appendix C in the online supplement.*

There are settings in wich the number of dimensions along which an array $(Y_{i_1 \ldots i_D})$ may be dependent could be greater than two. Our main framework can also be extended to cases in which the indices of the array pertain to the the same units in each dimension, that is the array may consist of random variables $Y_{i_1 \ldots i_D}$ with $i_d = 1, \ldots, N$ for each $d = 1, \ldots, D$. In that case we refer to the data as D-adic (dyadic if $D = 2$).

Example 1.4. ***V- and U-statistics*** *We can view V-statistics and U-statistics (see e.g. van der Vaart (1998) for definitions and a summary of classical asymptotic results) as special cases of our framework for D-adic data. For an i.i.d. random sample* $X_1, \ldots, X_N$, *a V-statistic of degree* D *with a symmetric kernel* $h(x_1, \ldots, x_D)$ *is defined as*

$$V = \frac{1}{N^D} \sum_{i_1 \ldots i_D} h(X_{i_1}, \ldots, X_{i_D})$$

which is equal to the D-fold sample average $\bar{Y}_{N,D} := \frac{1}{N^D} \sum_{i_1 \ldots i_D} Y_{i_1 \ldots i_D}$ *for the observations*

$$Y_{i_1 \ldots i_D} := h(X_{i_1}, \ldots, X_{i_D})$$

The kernel $h(\cdot)$ *is called degenerate if* $\mathbb{E}[h(x, X_2, \ldots, X_D)]$ *is constant. The asymptotic behavior of* $\bar{Y}_{N,D}$ *depends crucially on whether the kernel is degenerate, which is a feature of the unknown distribution of* X_i. *The corresponding U-statistic is*

$$U = \binom{N}{D}^{-1} \sum_{i_1 < i_2 \cdots < i_D} h(X_{i_1}, \ldots, X_{i_D}) = \binom{N}{D}^{-1} \sum_{i_1 \ldots i_D} w_{i_1 \ldots i_D} h(X_{i_1}, \ldots, X_{i_D})$$

[1] It may be possible to extend the general approach in this paper to allow for weak dependence in sampling across the time dimension, but such an extension would complicate the exposition substantially and take the focus away from the main ideas.

where $w_{i_1 \dots i_D} = \mathbb{1}\{i_1 < i_2 \cdots < i_D\}$. *Hence U-statistics can be viewed as a special case of a mean for a non-exhaustively matched sample.*

Example 1.5. ***Network data.*** *The general framework can be applied to subgraph counts or graph (homomorphism) densities in networks. Suppose that for a network with* N *nodes we observe the* $N \times N$ *adjacency matrix* $\mathbf{G}_N$ *with entries* G_{ij} *corresponding to indicators whether that network includes a directed edge from* i *to* j*, where it is usually assumed that* $G_{ii} = 0$ *for all* i *(no self-links). Following the approach in Lovasz (2012), Bickel, Chen, and Levina (2011), and Bhattacharya and Bickel (2015), we can regard* $\mathbf{G}_N$ *as a sample from an unlabeled infinite graph. For example to evaluate the extent of clustering/triadic closure in the network, we can consider triad-level subgraph counts* $T_r := \frac{6}{N(N-1)(N-2)} \sum_{i<j<k} Y_{ijk,r}$ *for* $r = 2, 3$ *where* $Y_{ijk,2} = G_{ij}G_{ik}$ *and* $Y_{ijk,3} = G_{ij}G_{ik}G_{jk}$*, so that* $Y_{ijk,3} = 0$ *whenever* i, j, k *are not distinct, and* $Y_{ijk,2} = 0$ *if* $i = j$ *or* $i = k$*. With degree heterogeneity across nodes, entries* $Y_{ijk,r}$ *exhibit dependence across each dimension of the array. This problem is a special case of the* D*-adic averages, which is discussed in the online appendix.*

Other prominent applications allowing for - not necessarily additive - dependence across several dimensions from e-commerce, biogenetics, and crop science are cited in Owen (2007).

1.1. **Problem Description.** Our main results concern the problem of bootstrapping the distribution of the sample average

$$\bar{Y}_{NT} := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}$$

The bootstrap procedure we propose in this paper is adaptive to features of the joint distribution of the random array, and approximations are as N and T grow large at the same rate. In particular, we aim to approximate the asymptotic distribution regardless of whether, or what type of cluster dependence is present. This is meant to reflect empirical practice, where the researcher aims for conclusions to be robust with respect to cluster-dependence, but without a presumption that such dependence is in fact present.

The leading case of bootstrapping the sample average already reflects the main new technical challenges arising from multi-way cluster-dependence. We also consider a number of additional practically relevant extensions and generalizations. For one, the procedure can be easily adapted for statistics that are asymptotically linear (i.e. that can be approximated via influence functions), or differentiable functions of $\bar{Y}_{NT}$. It is also conceptually straightforward to extend the procedure to settings with clustering along more than two dimensions, or D-adic data where the random array corresponds to group-level outcomes for any subset of D out of the full set of N units included in the sample. Another practically important extension concerns the case in which the variable Y_{it} is only observed for a subset of the pairs $\{(i,t) : i = 1, \dots, N, t = 1, \dots, T\}$ (non-exhaustively matched samples). For greater clarity,

the paper focusses on the leading case of cluster-dependence in two dimensions, and these generalizations are discussed in Appendix C in the online supplement.

Generally speaking, we need to distinguish three scenarios regarding the large-sample distribution of the mean: in the absence of cluster dependence, elements of the array (Y_{it}) are mutually independent, and under regularity conditions a CLT at the $(NT)^{-1/2}$ rate applies. When elements are correlated within clusters, the convergence rate of the mean is determined by the number of relevant clusters instead. Finally, in non-separable models of heterogeneity, elements within a cluster may be dependent even if they are uncorrelated. In that last case - which is specific to clustering in two or more dimensions - the asymptotic behavior of the sample mean is generally non-standard, and the conventional estimator of its asymptotic variance is not consistent. To frame ideas, we next give two stylized examples to illustrate the difference between these three cases.

Example 1.6. ***Additive Factor Model.*** *To shape ideas, consider first the case where clustering results from an additive model with cluster-level effects*

$$Y_{it} = \mu + \alpha_i + \gamma_t + \varepsilon_{it}$$

where μ is fixed and α_i, γ_t, ε_{it} are zero-mean, i.i.d. random variables for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ with bounded second moments, and $N = T$. From a standard central limit theorem we find that in the non-degenerate case with $\mathrm{Var}(\alpha_i) > 0$ or $\mathrm{Var}(\gamma_t) > 0$, the sample distribution

$$\sqrt{N}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]) \xrightarrow{d} N(0, \mathrm{Var}(\alpha_i) + \mathrm{Var}(\gamma_t)),$$

whereas in the degenerate case of no clustering, $\mathrm{Var}(\alpha_i) = \mathrm{Var}(\gamma_t) = 0$,

$$\sqrt{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]) \xrightarrow{d} N(0, \mathrm{Var}(\varepsilon_{it}))$$

where $\xrightarrow{d}$ denotes convergence in distribution.

If the marginal distributions of these three factors were known, we could simulate from the joint distribution of $(Y_{it})_{i=1,\ldots,Nt=1,\ldots,T}$ by sampling the individual components at random. A bootstrap procedure would replace these unknown distributions with consistent estimates. If the distribution of α_i is not known, an intuitively appealing estimator of α_i is

$$\hat{\alpha}_i := \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \bar{Y}_{NT}) = \alpha_i + \frac{1}{T}\sum_{t=1}^{T}(\varepsilon_{it} - \bar{\varepsilon}_{NT})$$

where $\bar{\varepsilon}_{NT} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}$. Similarly, we can estimate $\hat{\gamma}_t := \frac{1}{N}\sum_{i=1}^{N}(Y_{it} - \bar{Y}_{NT}) = \gamma_t + \frac{1}{N}\sum_{i=1}^{N}(\varepsilon_{it} - \bar{\varepsilon}_{NT})$, and $\hat{\varepsilon}_{it} := Y_{it} - \bar{Y}_{NT} - \hat{\alpha}_i - \hat{\gamma}_t$. We can then estimate the marginal distributions of $\alpha_i, \gamma_t, \varepsilon_{it}$ with the empirical distributions of $\hat{\alpha}_i, \hat{\gamma}_t$, and $\hat{\varepsilon}_{it}$, respectively.

We could then form a bootstrap sample $Y^*_{it} := \bar{Y}_{NT} + \alpha^*_i + \gamma^*_t + \varepsilon^*_{it}$ by drawing from these estimators for the marginal distributions of $\alpha_i, \gamma_t, \varepsilon_{it}$, and obtain $\bar{Y}^*_{NT} := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y^*_{it}$.

We can also verify that the conditional variances of the bootstrap distribution given the sample,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_i - \frac{1}{N}\sum_{j=1}^{N}\hat{\alpha}_j\right)^2 - \left[\mathrm{Var}(\alpha_i) + \frac{\mathrm{Var}(\varepsilon_{it})}{T}\right] \xrightarrow{p} 0$$

$$\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\gamma}_t - \frac{1}{N}\sum_{s=1}^{T}\hat{\gamma}_t\right)^2 - \left[\mathrm{Var}(\gamma_t) + \frac{\mathrm{Var}(\varepsilon_{it})}{N}\right] \xrightarrow{p} 0$$

Hence, in the non-degenerate case with $\mathrm{Var}(\alpha_i) > 0$ *or* $\mathrm{Var}(\gamma_t) > 0$*, the bootstrap distribution*

$$\sqrt{N}(\bar{Y}^*_{NT} - \bar{Y}_{NT}) \xrightarrow{d} N(0, \mathrm{Var}(\alpha_i) + \mathrm{Var}(\gamma_t))$$

converges to the same limit as the sampling distribution, so that estimation error in $\hat{\alpha}_i$ *does not affect the asymptotic variance. However, in the degenerate case of no clustering,* $\mathrm{Var}(\alpha_i) = \mathrm{Var}(\gamma_t) = 0$*, the bootstrap distribution*

$$\sqrt{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT}) \xrightarrow{d} N(0, 3\mathrm{Var}(\varepsilon_{it}))$$

asymptotically over-estimates the variance of the sampling distribution, so that this naive bootstrap procedure is inconsistent in the degenerate case.[2]

As the next example illustrates, the non-separable case has added complications from the fact that α_i, γ_t may interact. However, in either case the potential complications with the bootstrap stem entirely from the degenerate case.

Example 1.7. ***Non-Gaussian Limit Distribution.*** *For an example of non-separable heterogeneity, let*

$$Y_{it} = \alpha_i\gamma_t + \varepsilon_{it}$$

where $\alpha_i, \gamma_t, \varepsilon_{it}$ *are independently distributed, with* $\mathbb{E}[\varepsilon_{it}] = 0$*,* $\mathrm{Var}(\alpha_i) = \sigma^2_\alpha$*,* $\mathrm{Var}(\gamma_t) = \sigma^2_\gamma$*, and* $\mathrm{Var}(\varepsilon_{it}) = \sigma^2_\varepsilon$*.*

If in addition, $\mathbb{E}[\alpha_i] = \mathbb{E}[\gamma_t] = 0$*, a multivariate CLT and the continuous mapping theorem imply*

$$\begin{aligned}
\sqrt{NT}\bar{Y}_{NT} &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}(\alpha_i\gamma_t + \varepsilon_{it}) \\
&= \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\alpha_i\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\gamma_t\right) + \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it} \\
&\xrightarrow{d} \sigma_\alpha\sigma_\gamma Z_1 Z_2 + \sigma_\varepsilon Z_3
\end{aligned}$$

[2]Adaptations of the nonparametric bootstrap combining i.i.d. draws of columns and rows of the array $(Y_{it})_{i=1,\ldots,N t=1,\ldots,T}$ have been found to have similar problems, see McCullagh (2000) and Owen (2007).

where Z_1, Z_2, Z_3 are independent standard normal random variables. Since the product of two independent normal random variables is not normally distributed, $\sqrt{NT}\bar{Y}_{NT}$ is not asymptotically normal.[3] *Note also that if instead $\mathbb{E}[\alpha_i] \neq 0$ or $\mathbb{E}[\gamma_t] \neq 0$ the statistic remains asymptotically normal at the slower $\sqrt{T}$ ($\sqrt{N}$, respectively) rate.*

Non-separable heterogeneity can therefore generate dependence in second or higher moments that may contribute to the limting distribution even in the absence of correlation within clusters. Since the limiting distribution need not be Gaussian for these settings, plug-in asymptotic inference based on the normal distribution is not valid. We show below that this type of dependence in fact precludes uniformity in estimating the limiting distribution of $\bar{Y}_{NT}$. It can also be seen immediately from this example that this non-standard behavior could not be generated by a model of clustering in a single dimension, but is distinctive of the (less well-understood) case of cluster-dependence in two or more dimensions.

1.2. **Contribution.** This paper develops a theory for multi-way dependent data and proposes an inference procedure that is adaptive to the dependence structure, that is we aim to approximate the asymptotic distribution under any form of cluster dependence. In our view this type of adaptivity is crucial for common empirical practice, where the researcher aims for inference to be robust with respect to cluster-dependence, but without a presumption that such dependence is in fact present. Therefore a comprehensive analysis of the asymptotic distribution of the sample mean with multi-way clustering is needed which pays particular attention to scenarios in which observations may be uncorrelated within each cluster. We provide a comparison of the theoretical (large-sample) properties of our bootstrap procedure to those of alternative inference methods, including Gaussian "plug-in" inference, subsampling, and the "pigeonhole" bootstrap proposed by Owen (2007). We also provide simulation evidence for the most relevant cases.

To our knowledge this analysis is new to the literature, and this paper is the first to point out that even the limiting distribution for the sample average may be nonstandard in these settings. We also find that the default estimator for the asymptotic variance of the sample mean (a special case of the estimator proposed by Cameron, Gelbach, and Miller (2011)) is inconsistent due to the within-cluster dependence in second moments of Y_{it}. In order to determine what types of adaptivity and uniformity we may hope to achieve, we also establish a novel impossibility result: we find that there can be no estimator of the asymptotic distribution of the sample mean that is uniformly consistent. Instead, we provide alternative procedures - one that is pointwise consistent, and another conservative procedure that controls asymptotic size or coverage uniformly over the parameter space. As a special

[3]Since $Z_1 Z_2 = \frac{1}{4}(Z_1 + Z_2)^2 - \frac{1}{4}(Z_1 - Z_2)^2$, where $\text{Cov}(Z_1 + Z_2, Z_1 - Z_2) = \text{Var}(Z_1) - \text{Var}(Z_2) = 0$. Hence, $Z_1 Z_2 = \frac{1}{2}(W_1 - W_2)$, where W_1, W_2 are independent chi-square random variables with one degree of freedom.

case, these results apply to the problem of U- and V-statistics with kernel of unknown order of degeneracy.

Interestingly, both results (nonstandard asymptotic distribution and impossibility of uniform consistency in estimating that limit) require dependence in two or more dimensions and have no counterparts for the conventional case when observations are clustered in at most one dimension. The problem can be thought of as inference where a relevant nuisance parameter may be on, or close to, the boundary of the parameter space, resulting in a discontinuity in the pointwise asymptotic limiting distribution (see Andrews (2000), Andrews (2001), Andrews and Guggenberger (2009), and Andrews and Guggenberger (2010)). Our analysis benefits from theoretical insights and techniques developed for that abstract problem.

1.3. **Relation to the Literature.** The classical nonparametric bootstrap by Efron (1979) (see also Hall (1992), and Horowitz (2000) for an exposition) can be adapted to data that is cluster-dependent in one dimension in a straightforward manner. However with clustering in multiple dimensions, the problem of resampling is fundamentally different from the case of independent clusters, since the structure of the data no longer implies finite or weak dependence across units. In fact, McCullagh (2000) showed that there exists no scheme for resampling the raw data directly that is consistent for multi-way clustered data.[4] Our procedure combines features of the nonparametric bootstrap with those of the wild bootstrap (Wu (1986) and Liu (1988)) to achieve (pointwise) consistency in each case, as well as a conservative modification that results in uniformly valid asymptotic inference. We also establish refinements for cases in which the limiting behavior of the statistic is standard. We find that the problem of multi-way clustering has a natural connection to the theory of U- and V-statistics. For U- and V-statistics, Bretagnolle (1983) and Arcones and Giné (1992) proposed separate bootstrap procedures for the non-degenerate and degenerate case, but neither procedure is adaptive. In their analysis of weighted average derivatives under small bandwidth asymptotics, Cattaneo, Crump, and Jansson (2014) show how to adapt a naive bootstrap procedure to handle cases when the second-order U-statistic from a Hoeffding decomposition contributes to the asymptotic distribution. A recent paper by Graham (2020) gives asymptotic results for dyadic data, showing that the non-Gaussian contribution may be asymptotically negligible under sparse asymptotics.

[4]McCullagh (2000)'s argument goes as follows: there is no consistent estimator for the variance of the sample mean that is a nonnegative quadratic function of the observations Y_{it}. In particular the bootstrapped variance from any resampling scheme that draws directly from the original values of the variable of interest is a function of this type, and therefore such a bootstrap scheme cannot be consistent. We propose a hybrid scheme that does not fall under his narrower definition of the bootstrap.

Asymptotic standard errors with multi-way clustering have been proposed by Cameron, Gelbach, and Miller (2011), and can be used for "plug-in" asymptotic inference in the Gaussian limiting case - see also Cameron and Miller (2014), Aronow, Samii, and Assenova (2015), and Tabord-Meehan (2019) for the case of dyadic data. A more recent paper by MacKinnon, Nielsen, and Webb (2017) gives a condition on cluster sizes that is sufficient for asymptotic normality and consistency of these standard errors, and propose a bootstrap method for that setting. We show in the online appendix that the "pigeonhole" bootstrap proposed by Owen (2007) is asymptotically valid under non-trivial clustering in means, but conservative in the absence of clustering, and not guaranteed to achieve uniformity. A recent paper by Davezies, D'Haultfœuille, and Guyonvarch (2018) derives asymptotic properties for the pigeonhole bootstrap process for the non-degenerate case. Subsample bootstraps, including the method by Bhattacharya and Bickel (2015) for network data, adapt quite naturally to features of the data-generating process and are particularly attractive when evaluation of the statistic over the full sample is computationally very costly. The online appendix also establishes that for two-way cluster-dependent data subsampling is consistent pointwise, but not uniformly, and only at a slower rate than bootstrap alternatives.

1.4. **Notation and Overview.** Throughout the paper, we use $\mathbb{P}$ to denote the joint distribution of the array $(Y_{it})_{i,t}$, and denote drifting data-generating processes (DGP) indexed by N, T with $\mathbb{P}_{NT}$. The bootstrap distribution for (Y_{it}^*) given the realizations $(Y_{it} : i = 1, \dots, N; t = 1, \dots, T)$ is denoted $\mathbb{P}_{NT}^*$. We denote expected values under these respective distributions using $\mathbb{E}, \mathbb{E}_{NT}$, and $\mathbb{E}_{NT}^*$, respectively.

In the remainder of the paper, we first establish a representation for the array (Y_{it}) which is then used to motivate a bootstrap procedure. Formal results regarding consistency and refinements for that bootstrap procedure are given in Section 4, and we illustrate its performance using Monte Carlo simulations. Regression inference is discussed in Section 6. The online supplement Menzel (2021) provides additional asymptotic results for Gaussian asymptotics, the pigeonhole bootstrap, and subsampling as well as several generalizations of the main procedure.

2. Representation

We first consider the problem of inference on the sample mean. Here, we assume that the sample Y_{it} for $i = 1, \dots, N$ and $t = 1, \dots, T$ is embedded into a dissociated row and column exchangeable array: a *separately exchangeable array* is an infinite array $(Y_{it})_{i,t}$ such that for any integers $\tilde{N}, \tilde{T}$ and permutations $\pi_1 : \{1, \dots, \tilde{N}\} \to \{1, \dots, \tilde{N}\}$ and $\pi_2 : \{1, \dots, \tilde{T}\} \to \{1, \dots, \tilde{T}\}$, we have

$$\left(Y_{\pi_1(i)\pi_2(t)}\right)_{i,t} \overset{d}{=} (Y_{it})_{i,t},$$

where "$\stackrel{d}{=}$" denotes equality in distribution. Such an array is called *dissociated* (see Aldous (1981)) if for any $N_0, T_0 \geq 1$, $(Y_{it})_{i=1,t=1}^{i=N_0,t=T_0}$ is independent of $(Y_{it})_{i>N_0,t>T_0}$. For dyadic data we later consider dissociated, *jointly exchangeable* arrays $(Y_{ij})_{i,j}$ satisfying $\left(Y_{\pi(i)\pi(j)}\right)_{i,j} \stackrel{d}{=} (Y_{ij})_{i,j}$ for any permutation π on $\{1,\dots,\tilde{N}\}$, and for which in addition $(Y_{ij})_{i,j=1}^{N_0}$ is independent of $(Y_{ij})_{i,j>N_0}$.

We can interpret this assumption as stating that rows (and columns, respectively) correspond to units that are drawn independently from a common population, where we then observe the joint outcome for every row-column pair (or a subset of those pairs in the case of non-exhaustively matched samples). To make this more concrete, we can revisit the applications outlined in Examples 1.2-1.5:

- For difference-in-differences designs or matched data, this framework requires the units corresponding to either dimension of the sample (e.g. students and teachers, or ethnic groups and geographic districts) to represent independent draws from a common, infinite population.
- If for non-exhaustively matched data we may only observe joint outcomes for a possibly self-selected subset of unit pairs, the resulting sample may still be embedded into a (jointly or separately) exchangeable array if sample selection is also (jointly or separately) exchangeable.
- For U- and V-statistics, the kernel $Y_{i_1\dots i_D} := h(X_{i_1},\dots,X_{i_D})$ evaluated at i.i.d. observations $X_1,\dots,X_N$, forms a dissociated, jointly exchangeable array.

For network data, the graph is typically regarded as "unlabelled", i.e. node identifiers do not carry any significance for the statistical model, implying finite exchangeability. Joint ("infinite") exchangeability can be justified by regarding the sampled graph as a subgraph of an infinite graph.[5] Similarly, joint exchangeability results from network formation models that treat nodes as independent draws from a common superpopulation if the subgraph event encoded by $Y_{i_1\dots i_D}$ is fully determined by heterogeneity at the level of the polyad $(i_1,\dots,i_D)$.[6]

By Proposition 3.3 in Aldous (1981) any dissociated separately exchangeable array can be represented as

$$Y_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it}) \tag{2.1}$$

for some function $f(\cdot)$, where $\alpha_1,\dots,\alpha_N$, $\gamma_1,\dots,\gamma_T$ and $\varepsilon_{11},\dots,\varepsilon_{NT}$ are mutually independent, uniformly distributed random variables.[7] Similar representations are available to arrays

[5] See e.g. Lovasz (2012), Bickel and Chen (2009).

[6] This is in general not the case in models of strategic link formation, see e.g. Leung (2016) and Menzel (2015), which require a different approach.

[7] To be precise, Aldous (1981)'s result implies that there exists an array $(Y_{it}^* := f(\alpha_i, \gamma_t, \varepsilon_{it}))$ such that $(Y_{it}^*) \stackrel{d}{=} (Y_{it})$.

that are jointly or separately exchangeable in more than two dimensions, see Hoover (1979) and Section 7 in Kallenberg (2005).

As an important caveat, note that separate exchangeability generally does not allow for general serial or spatial dependence among the units in either dimension. By way of comparison, the main assumption in Cameron, Gelbach, and Miller (2011) requires that $\text{Cov}(Y_{it}, Y_{js}) = 0$ whenever $i \neq j$ and $s \neq t$, which allows for some forms of dependence that are not covered by our theory. For instance, the model in Example 1.6 continues to satisfy their assumption if ε_{it} (but not α_i or γ_t) are serially dependent in either dimension. Nevertheless, the negative results in this paper do also apply under that weaker condition.

2.1. **Projection.** We next show that the array $(Y_{it})_{i,t}$ permits a decomposition of the form

$$Y_{it} = b + a_i + g_t + w_{it}, \quad \mathbb{E}[w_{it}|a_i, g_t] = 0$$

where a_i and g_t are mean-zero and mutually independent, so that the joint distribution of Y_{it} can then be described in terms of the respective marginal distributions of a_i and g_t, and the conditional distribution of w_{it} given a_i, g_t. Such a representation is immediate for the leading example of the additive factor model in Example 1.6, and we now show that it is in fact without loss of generality for arrays exhibiting dependence in two or more dimensions.

We can expand Y_{it} according to

$$\begin{aligned} Y_{it} &= \mathbb{E}[Y_{it}] + (\mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}]) + (\mathbb{E}[Y_{it}|\gamma_t] - \mathbb{E}[Y_{it}]) \\ &\quad + (\mathbb{E}[Y_{it}|\alpha_i, \gamma_t] - \mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}|\gamma_t] + \mathbb{E}[Y_{it}]) + (Y_{it} - \mathbb{E}[Y_{it}|\alpha_i, \gamma_t]) \\ &=: b + a_i + g_t + v_{it} + e_{it} \end{aligned} \tag{2.2}$$

where we define $e_{it} = Y_{it} - \mathbb{E}[Y_{it}|\alpha_i, \gamma_t]$, $a_i := \mathbb{E}[Y_{i1}|\alpha_i] - \mathbb{E}[Y_{i1}]$, $g_t = \mathbb{E}[Y_{1t}|\gamma_t] - \mathbb{E}[Y_{1t}]$, $v_{it} = \mathbb{E}[Y_{it}|\alpha_i, \gamma_t] - \mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}|\gamma_t] + \mathbb{E}[Y_{it}]$, $b = \mathbb{E}[Y_{it}]$, and we assume throughout that the relevant conditional expectations are well-defined. Since temporal and cross-sectional units were drawn independently, $a_1, \ldots, a_N$ and $g_1, \ldots, g_T$ are independent of each other. Also by construction, $\mathbb{E}[e_{it}|a_i, g_t, v_{it}] = 0$ and $\mathbb{E}[v_{it}|a_i] = \mathbb{E}[v_{it}|g_t] = 0$. In particular, the terms $e_{it}, (a_i, g_t), v_{it}$ are uncorrelated.

Given this representation, we can rewrite the sample mean as

$$\bar{Y}_{NT} = b + \bar{a}_N + \bar{g}_T + \bar{v}_{NT} + \bar{e}_{NT}$$

where $\bar{a}_N := \frac{1}{N}\sum_{i=1}^N a_i$, $\bar{g}_T := \frac{1}{T}\sum_{t=1}^T g_t$, $\bar{v}_{NT} := \frac{1}{NT}\sum_{t=1}^T \sum_{i=1}^N v_{it}$, and $\bar{e}_{NT} := \frac{1}{NT}\sum_{t=1}^T \sum_{i=1}^N e_{it}$. We also denote the unconditional variances of the projections with $\sigma_a^2 := \text{Var}(a_i)$, $\sigma_g^2 := \text{Var}(g_t)$, $\sigma_v^2 := \text{Var}(v_{it})$, and $\sigma_e^2 := \text{Var}(e_{it})$, respectively. We also let $w_{it} := v_{it} + e_{it}$ and denote its variance by $\sigma_w^2 = \text{Var}(w_{it})$.

Throughout the remainder of the paper, we are going to maintain the following conditions on the distribution of the random array:

Assumption 2.1. ***(Integrability)*** *(a) Let $Y_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it})$ where $(\alpha_i)_i$, $(\gamma_t)_t$, and $(\varepsilon_{it})_{i,t}$ are random arrays whose elements are i.i.d. draws from the uniform distribution on the interval $[0,1]$. (b) The random variables a_i/σ_a, g_t/σ_g, v_{it}/σ_v, and e_{it}/σ_e are well-defined and have bounded moments up to the order $4+\delta$ for some $\delta > 0$ whenever the respective variances $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ are non-zero. (c) We have $\sigma_a^2 + \sigma_g^2 > 0$ or $\sigma_v^2 + \sigma_e^2 > 0$.*

2.2. **Low-Rank Approximation.** To understand the large sample properties of the sample mean, it is instructive to interpret the row/column projection

$$\bar{v}_{NT} \equiv \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\mathbb{E}[Y_{it}|\alpha_i,\gamma_t] - \mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}|\gamma_t] + \mathbb{E}[Y_{it}]\right) =: \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} v(\alpha_i,\gamma_t)$$

as a generalized (two-sample) U-statistic with a kernel $v(\alpha,\gamma)$ evaluated at the samples $\alpha_1,\ldots,\alpha_N$ and $\gamma_1,\ldots,\gamma_T$, respectively.

The asymptotic behavior of degenerate and non-degenerate generalized U-statistics is well-understood (see Serfling (1980) for a summary of classical results). The problem of characterizing the distribution of $\bar{Y}_{NT}$ differs from that classical problem in two major aspect: for one we also need to account for the presence of the projection error e_{it}. Furthermore the factors α_i, γ_t are not observable data, but implicitly defined by Aldous' (1981) construction. Nevertheless, these differences do not preclude us from applying general insights and techniques for U-statistics to the present problem.

Specifically, we find that we can approximate the sample and bootstrap distributions of the statistic by a function of sample averages of independent random variables. Define

$$v(\alpha,\gamma) := \mathbb{E}[Y_{it}|\alpha_i=\alpha,\gamma_t=\gamma] - \mathbb{E}[Y_{it}|\alpha_i=\alpha] - \mathbb{E}[Y_{it}|\gamma_t=\gamma] + \mathbb{E}[Y_{it}]$$

Under Assumption 2.1, the integral operator

$$S(u)(g) = \int v(a,g)u(a)F_\alpha(da)$$

and its adjoint

$$S^*(u)(a) = \int v(a,g)u(g)F_\gamma(dg)$$

are both compact, where F_α, F_γ are the marginal distributions corresponding to the joint $F_{\alpha\gamma}$ of α_i, γ_t, which are independent draws from the uniform distribution under the Aldous-Hoover representation in (2.1).

Hence, the spectral representation theorem permits the low-rank approximation

$$v(\alpha,\gamma) = \sum_{k=1}^{\infty} c_k \phi_k(\alpha)\psi_k(\gamma) \tag{2.3}$$

under the $L_2(F_{\alpha\gamma})$ norm on the space of smooth functions of $(\alpha, \gamma) \in [0,1]^2$. Here, $(c_k)_{k\geq 1}$ is a sequence of singular values with $\lim |c_k| \to 0$, and $(\phi_k(\cdot))_{k\geq 1}$ and $(\psi_k(\cdot))_{k\geq 1}$ are orthonormal bases for $L_2([0,1], F_\alpha)$ and $L_2([0,1], F_\gamma)$, respectively. Since by construction $\mathbb{E}[v(a, \gamma_t)] = \mathbb{E}[v(\alpha_i, g)] = 0$ for each $a, g \in [0,1]$, so that without loss of generality we can take $\mathbb{E}[\phi_k(\alpha_i)] = \mathbb{E}[\psi_k(\gamma_t)] = 0$ for each $k = 1, 2, \ldots$. Since the basis functions are orthonormal and α_i and γ_t independent, it follows that for any $K < \infty$ the covariance matrix of $(\phi_1(\alpha_i), \psi_1(\gamma_t), \ldots, \phi_K(\alpha_i), \psi_K(\gamma_t))$ is the $2K$-dimensional identity matrix. However, $(\phi_1(\alpha_i), \ldots, \phi_K(\alpha_i))$ may be correlated with a_i, and $(\psi_1(\gamma_t), \ldots, \psi_K(\gamma_t))$ may be correlated with g_t. Specifically, for $k = 1, 2, \ldots$ we denote

$$\sigma_{ak} := \mathrm{Cov}(a_i, \phi_k(\alpha_i)) \text{ and } \sigma_{gk} := \mathrm{Cov}(g_t, \psi_k(\gamma_t)).$$

Given this representation, we can write

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} v(\alpha_i, \gamma_t) = \sum_{k=1}^{\infty} c_k \left(\frac{1}{N}\sum_{i=1}^{N}\phi_k(\alpha_i)\right)\left(\frac{1}{T}\sum_{t=1}^{T}\psi_k(\gamma_t)\right)$$

so that the second-order projection term can also be represented as a function of countably many sample averages of i.i.d., mean-zero random variables. The limiting distribution of this term is not Gaussian, but can be represented as a linear combination of independent chi-square random variables, see e.g. Serfling (1980). Distributions of this type are known as Wiener (or Gaussian) chaos.

We find that point-wise consistency of the bootstrap does not require any additional conditions on the conditional expectation function $v(\alpha, \gamma)$ beyond Assumption 2.1. For the uniform consistency results which include the case in which the asymptotically non-Gaussian component is of first order, we need to restrict the eigenfunctions and coefficients in the spectral representation (2.3).

Assumption 2.2. *The function* $v(\alpha, \gamma) := \mathbb{E}[Y_{it}|\alpha_i = \alpha, \gamma_t = \gamma] - \mathbb{E}[Y_{it}|\alpha_i = \alpha] - \mathbb{E}[Y_{it}|\gamma_t = \gamma] + \mathbb{E}[Y_{it}]$ *admits a spectral representation*

$$v(\alpha, \gamma) = \sum_{k=1}^{\infty} c_k \phi_k(\alpha)\psi_k(\gamma)$$

under the $L_2(F_{\alpha\gamma})$ *norm, where* $(\phi_k(\alpha))$ *and* $(\psi_k(\gamma))$ *are orthonormal and orthogonal to* a_i, g_t. *Furthermore, (a) the singular values are uniformly bounded by a square summable null sequence* $\bar{c}_k$ *that is* $c_k \leq \bar{c}_k$ *for each* $k = 1, 2, \ldots$, *where* $\sum_{k=1}^{\infty} c_k^2 < \infty$. *(b) The first three moments of the eigenfunctions* $\phi_k(\alpha_i)$ *and* $\psi_k(\gamma_t)$ *are bounded by a constant* $B > 0$ *for each* $k = 1, 2, \ldots$.

Imposing common bounds on moments and singular values restricts the set of joint distributions F for the array to a uniformity class, where the sequence $\mathbf{c} := (\bar{c}_k)_{k\geq 0}$ controls the

magnitude of the error from a finite-dimensional approximation to $v(\alpha, \gamma)$, where we truncate the expansion in (2.3) after a finite number of summands $k = 1, \ldots, K$. Comparable high-level conditions on spectral approximations are commonly used to define uniformity classes in nonparameric estimation of operators, see e.g. Hall and Horowitz (2005) and Carrasco, Florens, and Renault (2007).

3. Bootstrap Procedure

The previous discussion shows that the rate of convergence and the limiting distribution of the sample mean $\bar{Y}_{NT} - \mathbb{E}[Y_{it}]$ depend crucially on the different scale parameters introduced above. For example, if observations are independent across rows and columns, then $\sqrt{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]) \xrightarrow{d} N(0, \sigma_e^2)$. If $N = T$ and within-cluster covariances are bounded away from zero in at least one dimension, then $\sqrt{N}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]) \xrightarrow{d} N(0, \sigma_a^2 + \sigma_g^2)$. Our aim is to obtain a bootstrap procedure that is adaptive for both degenerate and non-degenerate cases.

For the bootstrap procedure we can estimate the terms of the orthogonal projection in (2.2) with their sample analogs

$$\hat{a}_i := \frac{1}{T}\sum_{t=1}^{T} Y_{it} - \bar{Y}_{NT}, \quad \hat{g}_t := \frac{1}{N}\sum_{i=1}^{N} Y_{it} - \bar{Y}_{NT}, \quad \text{and } \hat{w}_{it} := Y_{it} - \hat{a}_i - \hat{g}_t - \bar{Y}_{NT}$$

For the performance of the bootstrap it is crucial at what rate(s) estimators for the different model components are consistent depending on the extent of clustering in the true DGP. Most importantly, the variance of the projection terms $\hat{a}_i$ and $\hat{g}_t$ is $\sigma_a^2 + \sigma_w^2/T$ and $\sigma_g^2 + \sigma_w^2/N$, respectively, so that the "convolution error" depending on σ_w^2 dominates in the degenerate case. In order to correct for that contribution of the row/column averages of w_{it} we would therefore want to shrink the scale of the distribution of $\hat{a}_i, \hat{g}_t$ by the variance ratios

$$\lambda_a = \frac{T\sigma_a^2}{T\sigma_a^2 + \sigma_w^2}, \text{ and } \lambda_g = \frac{N\sigma_g^2}{N\sigma_g^2 + \sigma_w^2}$$

In the bootstrap procedure we replace the unknown variances with consistent estimators in (3.1) to obtain alternative estimators for λ_a and λ_g.

To obtain the component variances, we let

$$\begin{aligned}
\hat{s}_a^2 &:= \frac{1}{N-1}\sum_{i=1}^{n}(\hat{a}_i - \bar{Y}_{NT})^2, \quad \hat{s}_g^2 := \frac{1}{T-1}\sum_{t=1}^{T}(\hat{g}_t - \bar{Y}_{NT})^2 \\
\hat{s}_w^2 &:= \frac{1}{NT - N - T}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \hat{a}_i - \hat{g}_t - \bar{Y}_{NT})^2
\end{aligned}$$

and form the estimators

$$\hat{\sigma}_a^2 = \max\left\{0, \hat{s}_a^2 - \frac{1}{T}\hat{s}_w^2\right\}, \quad \hat{\sigma}_g^2 = \max\left\{0, \hat{s}_g^2 - \frac{1}{N}\hat{s}_w^2\right\}, \quad \text{and } \hat{\sigma}_w^2 := \hat{s}_w^2 \tag{3.1}$$

We find in Lemma A.1 below that the variances σ_a^2 and σ_g^2 cannot always be estimated at a sufficiently fast rate. One of the versions of the bootstrap procedure proposed here therefore uses a consistent pre-test for the presence of cluster dependence in the first moment. To that end, we define the model selectors

$$\hat{D}_a(\kappa) := \mathbb{1}\{T\hat{\sigma}_a^2 \geq \kappa\} \text{ and } \hat{D}_g(\kappa) := \mathbb{1}\{N\hat{\sigma}_g^2 \geq \kappa\}$$

for any given value of $\kappa \geq 0$. For appropriately chosen sequences κ_a, κ_g, we then let

$$\hat{\lambda}_a := \frac{\hat{D}_a(\kappa_a)T\hat{\sigma}_a^2}{\hat{D}_a(\kappa_a)T\hat{\sigma}_a^2 + \hat{\sigma}_w^2} \text{ and } \hat{\lambda}_g := \frac{\hat{D}_g(\kappa_g)N\hat{\sigma}_g^2}{\hat{D}_g(\kappa_g)N\hat{\sigma}_g^2 + \hat{\sigma}_w^2}$$

and estimate the asymptotic variance of the sample mean with

$$\hat{S}_{NT,sel}^2 := \hat{D}_a(\kappa_a)T\hat{\sigma}_a^2 + \hat{D}_g(\kappa_g)N\hat{\sigma}_g^2 + \hat{\sigma}_w^2 \tag{3.2}$$

In the online appendix we compare this estimator to a "default" estimator for the asymptotic variance without a pre-test, defined as

$$\hat{S}_{NT,def}^2 \quad := \quad T\hat{s}_a^2 + N\hat{s}_g^2 - \hat{s}_w^2$$

Note that up to a degree of freedom correction, $\hat{S}_{NT,def}^2$ is the variance estimator from Cameron, Gelbach, and Miller (2011) for the special case of the sample mean.[8]

For the leading case of exhaustive sampling with cluster dependence in two dimensions, we then propose the following resampling algorithm to estimate the sampling distribution:

(a) For the bth bootstrap iteration, draw $a_{i,b}^* := \hat{a}_{k_b^*(i)}$ and $g_{t,b}^* := \hat{g}_{s_b^*(t)}$, where $k_b^*(i)$ and $s_b^*(t)$ are i.i.d. draws from the discrete uniform distribution on the index sets $\{1, \ldots, N\}$ and $\{1, \ldots, T\}$, respectively.

(b) Generate $w_{it,b}^* := \omega_{1i,b}\omega_{2t,b}\hat{w}_{k_b^*(i)s_b^*(t)}$, where $\omega_{1i,b}, \omega_{2t,b}$ are i.i.d. random variables with $\mathbb{E}[\omega_\cdot] = 0, \mathbb{E}[\omega_\cdot^2] = \mathbb{E}[\omega_\cdot^3] = 1$

(c) Generate a bootstrap samples of draws $Y_{it,b}^* = \bar{Y}_{NT} + \sqrt{\hat{\lambda}_a}a_{i,b}^* + \sqrt{\hat{\lambda}_g}g_{t,b}^* + w_{it,b}^*$ and obtain the bootstrapped statistic $\bar{Y}_{NT,b}^* := \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T Y_{it,b}^*$.

(d) We repeat this procedure to obtain a sample of B replications and approximate the conditional distribution of $\bar{Y}_{NT}^*$ given the sample with the empirical distribution over the bootstrap draws $\bar{Y}_{NT,1}^*, \ldots, \bar{Y}_{NT,B}^*$.

[8]The possible failure of the classical bootstrap when parameters of the asymptotic distribution are near or on the boundary of the parameter space was first demonstrated by Andrews (2000) who also proposed model selection for pointwise consistent inference. The pitfalls of post-selection inference have now been well-documented, in particular consistent model selection typically leads to failure of uniformity in asymptotic approximations, see Leeb and Pötscher (2005). We show that in the present case, this challenge is not an artifact of the approach taken in this paper. Rather, uniformly consistent estimation of the limiting distribution is not possible, neither using the bootstrap nor any alternative method, so our proposals include a procedure that achieves uniformity but is conservative.

For the pivotal bootstrap, the last step uses instead the empirical distribution of the studentized bootstrap draws to approximate the distribution of $\sqrt{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT})/\hat{S}^*_{NT,sel}$, where $\hat{S}^*_{NT,sel}$ is the bootstrap analog of the variance estimator $\hat{S}_{NT,sel}$. Typical choices for the distribution of the multiplier variables $\omega_{1i,b}, \omega_{2t,b}$ in step (b) are the Gamma distribution (with shape parameter 4 and scale parameter equal to $\frac{1}{2}$) or the two-point specification proposed by Mammen (1992).

We distinguish two versions of this bootstrap procedure:

Definition 3.1. ***(Bootstrap Procedures)***

- ***(BS-N)*** *The* bootstrap without model selection *applies steps (a)-(d) where we set* $\kappa_a = \kappa_g = 0$,
- ***(BS-S)*** *The* bootstrap with model selection *follows steps (a)-(d) where we set* κ_a, κ_g *according to increasing sequences* $\kappa_g, \kappa_a \to \infty$ *such that* $\kappa_a/T \to 0$ *and* $\kappa_g/N \to 0$.
- ***(BS-C)*** *The* conservative bootstrap *applies steps (a)-(d) where for increasing sequences* $\kappa_g, \kappa_a \to \infty$ *such that* $\kappa_a/T \to 0$ *and* $\kappa_g/N \to 0$, *we set* $\hat{q}_a := \max\{T\hat{\sigma}_a^2, \kappa_a\}$, $\hat{q}_g := \max\{N\hat{\sigma}_g^2, \kappa_g\}$, *and*

$$\hat{\lambda}_a := \frac{\hat{q}_a}{\hat{q}_a + \hat{\sigma}_w^2}\frac{\hat{q}_a}{T\hat{\sigma}_a^2}, \quad \hat{\lambda}_g := \frac{\hat{q}_g}{\hat{q}_g + \hat{\sigma}_w^2}\frac{\hat{q}_g}{N\hat{\sigma}_g^2}$$

We find below that the bootstrap with model selection is consistent pointwise in $\sigma_a^2, \sigma_g^2, \sigma_w^2$, and the bootstrap without model selection is uniformly consistent as long as the limiting distribution is Gaussian. The conservative bootstrap is consistent in the nondegerate case $\sigma_a^2 + \sigma_g^2 > 0$, but asymptotically conservative for the degenerate cases in a sense to be made more precise below. It is the only procedure discussed in this paper that is guaranteed to have uniform size control over the entire parameter space.

4. Theoretical Properties

In this section we establish large sample properties for this bootstrap procedure. The limiting behavior of the sample mean $\bar{Y}_{NT} - \mathbb{E}[Y_{it}]$ is in part determined by the variances of the components of the decomposition in (2.2). Since the rate of convergence of the sample mean depends on the component variances, we define the adaptive rate r_{NT} by

$$r_{NT}^{-2} := N^{-1}\sigma_a^2 + T^{-1}\sigma_g^2 + (NT)^{-1}\sigma_w^2 \equiv \mathrm{Var}(\bar{Y}_{NT})$$

where the last equality follows since the components in the decomposition (2.2) are uncorrelated. We maintain throughout that either $\sigma_g^2 + \sigma_a^2 > 0$ or $\sigma_w^2 > 0$, and that N and T grow at the same rate as we take limits.

4.1. Asymptotic Distribution of $\bar{Y}_{NT}$.

We now characterize the asymptotic distribution of the sample mean. To analyze which properties are uniform with respect to the joint

distribution of (Y_{it}), we also need to consider limits along any drifting sequences for the parameters $\sigma_a^2, \sigma_g^2, \sigma_e^2, \sigma_v^2$. We then parameterize the limiting distribution with the respective limits of normalized sequences

$$\begin{aligned} q_{a,NT} &:= r_{NT}^2 N^{-1}\sigma_a^2, & q_{g,NT} &:= r_{NT}^2 T^{-1}\sigma_g^2 \\ q_{e,NT} &:= r_{NT}^2 (NT)^{-1}\sigma_e^2 & q_{v,NT} &:= r_{NT}^2 (NT)^{-1}\sigma_v^2 \\ q_{ak,NT} &:= r_{NT}^2 N^{-1}\sigma_{ak} & q_{gk,NT} &:= r_{NT}^2 T^{-1}\sigma_{gk} \end{aligned} \tag{4.1}$$

for $k = 1, 2, \ldots$. We also let $\varrho_{NT} := r_{NT}(NT)^{-1/2}$. From the definition of r_{NT}, it follows that the local parameters $q_{a,NT}$, $q_{g,NT}$, $q_{e,NT}$, $q_{v,NT} \in [0,1]$ and $q_{a,NT} + q_{g,NT} + q_{e,NT} + q_{v,N} = 1$. We stack these sequences as the vector

$$\mathbf{q}_{NT} := (q_{e,NT}, q_{a,NT}, q_{g,NT}, q_{a1,NT}, q_{g1,NT}, q_{a2,NT}, q_{g2,NT}, \ldots,)$$

Similarly, we represent the singular values for the spectral decomposition (2.3) for $\mathbb{E}_{NT}[Y_{it}|\alpha_i, \gamma_t]$ and $\mathbb{E}[Y_{it}|\alpha_i, \gamma_t]$ with $\mathbf{c}_{NT} := (c_{1,NT}, c_{2,NT}, \ldots) \in \ell^2$ and $\mathbf{c} := (c_1, c_2, \ldots) \in \ell^2$, respectively.

We can summarize asymptotic properties for the various procedures in terms of these parameter sequences, where for convergent sequences $\mathbf{q}_{NT}, \mathbf{c}_{NT}, \varrho_{NT}$ we denote the limits $q_a := \lim_{N,T} q_{a,NT}$, $q_g := \lim_{N,T} q_{g,NT}$, $q_e := \lim_{N,T} q_{e,NT}$, and $q_v := \lim_{N,T} q_{v,NT}$. The limiting distribution along such a sequence will therefore depend on the parameters $\mathbf{q} := \lim_{N,T} \mathbf{q}_{NT}$, $\mathbf{c} := \lim_{N,T} \mathbf{c}_{NT}$ and $\varrho := \lim_{N,T} \varrho_{NT}$.[9]

For any fixed values of the local parameters $\mathbf{q}, \mathbf{c}$, and $\varrho \in [0,1]$ we define the law

$$\mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho) := \left(\sqrt{q_e}Z^e + \sqrt{q_a}Z^a + \sqrt{q_g}Z^g\right) + \varrho V \tag{4.2}$$

where $Z^e, Z_1^\phi, Z_1^\psi, Z_2^\phi, Z_2^\psi, \ldots$ are i.i.d. standard normal random variables,

$$V := \sum_{k=1}^{\infty} c_k Z_k^\psi Z_k^\phi$$

with the coefficients c_k potentially variying along the limiting sequence, and Z^a, Z^g are standard normal random variables with $\mathrm{Cov}(Z^a, Z_k^\phi) = q_{ak}/\sqrt{q_a}$, $\mathrm{Cov}(Z^g, Z_k^\psi) = q_{gk}/\sqrt{q_g}$, $\mathrm{Cov}(Z^a, Z^g) = \mathrm{Cov}(Z^a, Z_k^\psi) = \mathrm{Cov}(Z^g, Z_k^\phi) = 0$ for all $k = 1, 2, \ldots$.

We can now give the limit for the sampling distribution of $\bar{Y}_{NT}$:

Theorem 4.1. ***(CLT for Sampling Distribution)*** *Suppose that Assumption 2.1 holds. Then (a) along any convergent sequence $\mathbf{q}_{NT} \to \mathbf{q}$ and fixed $\mathbf{c} = (c_1, c_2, \ldots)$, we have*

$$\|\mathbb{P}_{NT}(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])) - \mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho)\|_\infty \to 0$$

[9] We show that without loss of generality it is sufficient to focus on convergent parameter sequences in light of arguments by Andrews and Guggenberger (2010).

where $\varrho := \lim_{N,T} \varrho_{NT}$, $\|\cdot\|_\infty$ *denotes the Kolmogorov metric, and the limiting distribution* $\mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho)$ *is continuous. (b) If in addition Assumption 2.2 holds, then the conclusion of (a) also holds under drifting sequences* $\mathbf{c}_{NT} \to \mathbf{c}$.

See the appendix for a proof. Note that convergence in part (a) is point-wise with respect to the conditional mean function $\mathbb{E}[Y_{it}|\alpha_i = \alpha, \gamma_t = \gamma]$, whereas part (b) gives uniform convergence within the class of distributions satisfying Assumption 2.2.

4.2. **Estimability of the Asymptotic Distribution.** The asymptotic properties of the bootstrap depend crucially on our ability to estimate the variances of the individual projection components at respective rates that are fast enough to ensure convergence of $\hat{\lambda}_a$ and $\hat{\lambda}_g$ to λ_a and λ_g, respectively. Lemma A.1 in the appendix establishes that the component variances $\sigma_a^2, \sigma_g^2, \sigma_w^2$ can be estimated consistently, but not always at a sufficiently fast rate along certain parameter sequences. We can in fact establish the stronger negative result that there exists no estimator for the asymptotic distribution that achieves consistency uniformly over the space of distributions satisfying the main assumptions of this paper.

In order to state that impossibility result formally, we first introduce some additional notation. From the Aldous-Hoover representation, the distribution of (Y_{it}) can be identified with the function $f(\alpha, \gamma, \varepsilon)$ in (2.1). We also let $\mathcal{F}$ denote the class of functions $f(\alpha, \gamma, \varepsilon)$ corresponding to distributions of (Y_{it}) satisfying Assumptions 2.1 and 2.2 for i.i.d. uniform draws $\alpha_i, \gamma_t, \varepsilon_{it}$. We then use $\mathbb{P}_{f,NT}(\cdot)$ to denote probabilities for events concerning an array of size N, T generated according to f, and $\mathrm{Var}_f(\cdot)$ for the corresponding variances. Furthermore, we use the notation $\mathbf{q}_{NT}(f) := (q_{e,NT}(f), q_{a,NT}(f), \dots,)$ for the vector of normalized variances from (4.1) given variances $\sigma_e^2(f) := \mathrm{Var}_f(e_{it})$, $\sigma_a^2(f) := \mathrm{Var}_f(a_i)$, etc., and define $\varrho_{NT}(f)$ and the singular values $\mathbf{c}_{NT}(f)$ in an analogous manner. We then have the following result:

Proposition 4.1. ***(Estimability of Asymptotic Distribution)*** *Let* $\hat{\mathcal{L}}_{NT}$ *denote an arbitrary estimator for* $\mathcal{L}_0$ *based on an array of size* N, T *from the unknown distribution. Then there exists* $\delta > 0$ *such that*

$$\liminf_{N,T\to\infty} \sup_{f\in\mathcal{F}} \mathbb{P}_{f,NT}\left(\left\|\hat{\mathcal{L}}_{NT} - \mathcal{L}_0(\mathbf{q}_{NT}(f), \mathbf{c}_{NT}(f), \varrho_{NT}(f))\right\|_\infty > \delta\right) > 0$$

Recall that Theorem 4.1 showed that the sample mean converges to a continuous limiting distribution $\mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho)$ along sequences $f_{NT} \in \mathcal{F}$ with proper limits for $\mathbf{q}_{NT}, \mathbf{c}_{NT}$. This result therefore states that we cannot estimate the asymptotic distribution uniformly consistently even when the problem is otherwise well-behaved.

The proof for this impossibility result can be found in the appendix and is based on the following counterexample: consider the model $Y_{it} = \alpha_i\gamma_t$, where α_i, γ_t are mutually independent, with i.i.d. factors $\alpha_i \sim N(0,1)$, $\gamma_t \sim N(\mu_\gamma, 1)$. Clearly, this model satisfies

Assumption 2.1, so that Theorem 4.1 implies convergence to a limiting distribution of the form (4.2). For this model, $a_i := \mathbb{E}[Y_{it}|\alpha_i] = \alpha_i\mu_\gamma$, $g_t := \mathbb{E}[Y_{it}|\gamma_t] = \gamma_t\mathbb{E}[\alpha_i] \equiv 0$, and $v_{it} = \alpha_i(\gamma_t - \mu_\gamma)$, so that $\sigma_a^2 = \mu_\gamma^2$ and $\sigma_v^2 = 1$. Clearly, μ_γ cannot be estimated from the original data at a rate faster than $T^{-1/2}$, which is the fastest possible rate at which μ_γ could be estimated from observing $\gamma_1, \ldots, \gamma_T$ directly. Hence, no test can consistently distinguish the model $\mu_\gamma = 0$ resulting in an asymptotic variance equal to σ_v^2 from a drifting sequence $\tilde{\mu}_{T,\gamma} := T^{-1/2}m_\gamma$ which results in an asymptotic variance equal to $m_\gamma^2 + \sigma_v^2$. Since the variance of the sampling distribution converges along either sequence to σ_v^2 and $m_\gamma^2 + \sigma_v^2$ respectively, it follows that it cannot be consistently estimated. The proof of Proposition 4.1 shows that this impossibility result holds not only with respect to moments of the distribution, but also in terms of weak convergence.

4.3. **Bootstrap Consistency.** We now turn to the asymptotic properties of the bootstrap described in Section 3, where we consider both a non-pivotal version, and a pivotal version based on the studentized sample mean. Specifically, consider the estimator of the asymptotic variance of the sample mean, $\hat{S}_{NT,sel}$ defined in (3.2) and its bootstrap analog

$$\hat{S}_{NT,sel}^{2*} := \hat{D}_a(\kappa_a)T\hat{\sigma}_a^{2*} + \hat{D}_g(\kappa_g)N\hat{\sigma}_g^{2*} + \hat{\sigma}_w^{2*},$$

where we hold the selectors $\hat{D}_a(\kappa_a), \hat{D}_g(\kappa_g)$ fixed at their sample values, and κ_a, κ_g are chosen according to whether the bootstrap is implemented with our without model selection.

The non-pivotal bootstrap approximates the distribution of the sample mean $r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])$ with the distribution of its bootstrap analog, $r_{NT}(\bar{Y}_{NT}^* - \bar{Y}_{NT})$. The pivotal bootstrap approximates the distribution of the studentized sample mean $(NT)^{1/2}\hat{S}_{NT,sel}^{-1}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])$ with the distribution of its bootstrap analog, $(NT)^{1/2}(\hat{S}_{NT,sel}^*)^{-1}(\bar{Y}_{NT}^* - \bar{Y}_{NT})$. Corollary A.1 in the appendix establishes that the estimator $\hat{S}_{NT,sel}$ is pointwise consistent for sequences of κ_a, κ_g increasing to infinity at a sufficiently slow rate, and its analog for $\kappa_a = \kappa_g = 0$ is uniformly consistent for $q_v = 0$. Similarly, we can use Lemma A.1 in the appendix to establish pointwise consistency of $\hat{\lambda}_a$ and $\hat{\lambda}_g$ for the bootstrap with model selection (and uniform consistency given $q_v = 0$ for the bootstrap without model selection).

Combining this with the sample CLT (Theorem 4.1) and a bootstrap CLT (Lemma A.2 in the appendix), we then obtain consistency results of the form

$$\|\mathbb{P}_{NT}^*(r_{NT}(\bar{Y}_{NT}^* - \bar{Y}_{NT})) - \mathbb{P}_{NT}(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]))\|_\infty \xrightarrow{a.s.} 0 \tag{4.3}$$

and its pivotal analog

$$\left\| \mathbb{P}_{NT}^* \left(\sqrt{NT} \frac{\bar{Y}_{NT}^* - \bar{Y}_{NT}}{\hat{S}_{NT,sel}^*} \right) - \mathbb{P}_{NT} \left(\sqrt{NT} \frac{\bar{Y}_{NT} - \mathbb{E}[Y_{it}]}{\hat{S}_{NT,sel}} \right) \right\|_\infty \xrightarrow{a.s.} 0 \tag{4.4}$$

for the bootstrap procedures with and without model selection. The conservative bootstrap generally overestimates the scale of the sampling distribution for the degenerate case, where

we obtain a convergence result of the form

$$\|\mathbb{P}^*_{NT}(r_{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT})) - \mathcal{L}_0(\bar{\mathbf{q}}, \mathbf{c}, \varrho)\|_\infty \xrightarrow{p} 0 \tag{4.5}$$

and the pivotal version of the conservative bootstrap

$$\left\| \mathbb{P}^*_{NT} \left(\sqrt{NT} \frac{\bar{Y}^*_{NT} - \bar{Y}_{NT}}{\hat{S}^*_{NT,sel}} \right) - \mathcal{L}_0(\bar{\mathbf{q}}, \mathbf{c}, \varrho) \right\|_\infty \xrightarrow{p} 0 \tag{4.6}$$

Here, $\bar{\mathbf{q}} = (q_e, \bar{q}_a, \bar{q}_g, 0, 0, \dots)$, and $\bar{q}_a := \max\{\kappa_a/T, q_a\}$ and $\bar{q}_g := \max\{\kappa_g/N, q_g\}$, which increase as $N, T \to \infty$.

Theorem 4.2. ***(Bootstrap Consistency)*** *Suppose that Assumption 2.1 holds. Then (a) the bootstrap with model selection satisfies (4.3) and (4.4) pointwise for any fixed $\sigma_a^2, \sigma_g^2, \sigma_e^2, \sigma_v^2$. (b) The bootstrap without model selection satisfies satisfies (4.3) and (4.4) uniformly if $q_v = 0$. (c) The conservative bootstrap satisfies (4.5) and (4.6) uniformly over the entire parameter space.*

See the appendix for a proof. Relating these results to the three alternative criteria stated at the beginning of this section, part (a) states that the bootstrap with model selection is pointwise valid asymptotically, which corresponds to our first criterion. The lack of uniformity of (BS-S) is not unique to this problem, but has generally been noted for inference procedures involving model selection (see Leeb and Pötscher (2005)). According to part (b), the bootstrap without model selection is valid uniformly with respect to clustering in means, but is inconsistent if $q_v > 0$, so that it is asymptotically valid according to our second criterion. The conservative bootstrap is uniformly valid without any qualifications, however in degenerate cases ($q_e + q_v > 0$) the scale of the estimated asymptotic distribution diverges at a rate $\kappa_a/T + \kappa_g/N$.[10] Comparing the respective limits for the conservative bootstrap and the sampling distribution (see Theorem 4.1), $\mathcal{L}_0(\bar{\mathbf{q}}, \mathbf{c}, \varrho)$ is a mean-preserving spread of $\mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho)$, where both distributions are symmetric about zero. In particular, estimates of percentiles from the conservative bootstrap are biased outwards (i.e. away from zero) in those cases, so that commonly used one- or two-sided hypothesis tests or confidence sets based on these estimated percentiles are asymptotically conservative.

Remark 4.1. ***U- and V-Statistics*** *Note that these results also applies to generalized (two-sample) U-statistics, which constitute a special case of our setup with $\sigma_e^2 = 0$. Specifically, the impossibility result in Proposition 4.1 implies that if the order of degeneracy of the kernel is unknown, it is not possible to estimate the distribution of a two-sample U-statistic uniformly consistently. The bootstrap procedure in this paper is pointwise adaptive with respect to the order of degeneracy of the kernel of the V-statistic. Analogous conclusions for standard (one-sample) U- and V-statistics with a kernel function of order D, can be obtained*

[10] For the choice of κ_a, κ_g implemented for the simulation study, $\kappa_a/T + \kappa_g/N \asymp \log(T) + \log(N)$.

using an adaptation of our bootstrap procedure to D-adic data, see Appendix C in the online supplement for a discussion.

4.4. **Refinements.** We next consider refinements in the approximation to the distribution of the studentized mean. We find that the bootstrap approximation provides pointwise refinements for studentized mean in the non-degenerate case $\sigma_a^2 + \sigma_g^2 > 0$. It is also important to note that refinements can in general not be obtained for certain special cases. For one, if the "Wiener chaos" term remains relevant in the limiting distribution $\mathcal{L}(\mathbf{q}, \mathbf{c}, \varrho)$, i.e. for $\varrho > 0$, the studentized mean is not asymptotically pivotal. Rather the asymptotic distribution generally depends on relative weights of the Gaussian component Z, and the spectral coefficients $\mathbf{c}$ defining the Wiener chaos component V. Hence we cannot expect the bootstrap to provide refinements for this case.

Furthermore, elementary moment calculations reveal that

$$\mathbb{E}[\hat{a}_i^3] = \mathbb{E}[a_i^3] + \frac{2}{T}\mathbb{E}[a_i w_{it}^2] + \frac{1}{T^2}\mathbb{E}[w_{it}^3]$$

where the cross-term $\mathbb{E}[a_i w_{it}^2]$ is generally non-zero unless $\mathbb{E}[w_{it}^2|a_i]$ and a_i are uncorrelated. Hence under drifting sequences for the second and third moments of a_i, the second and third terms on the right-hand side of that expression may be of the same order as the leading term, in which case the bootstrap distribution does not match the third moment of a_i under the sampling distribution. Hence, we can in general not obtain a refinement along drifting sequences even when $\varrho = 0$ and the limiting distribution is Gaussian. Hence we restrict our attention to the non-degenerate case with a Gaussian limiting distribution.

We establish refinements using now standard results on Edgeworth expansions, most importantly Theorem 5.2 in Hall (1992). To this end, we impose fairly stringent moment conditions and Cramér's condition on the characteristic functions of the marginal distributions of a_i and g_t: A random vector X with support on $\mathbb{R}^d$ is said to satisfy Cramér's condition if

$$\limsup_{\|t\| \to \infty} |\mathbb{E}[\exp\{it'X\}]| < 1 \tag{4.7}$$

where $i = \sqrt{-1}$. This condition is met whenever X has a nondegenerate, absolutely continuous component (see e.g. Hall (1992) p.65-67). We can then state the following result:

Proposition 4.2. ***(Refinements)*** *Suppose that Assumption 2.1 holds for any $0 < \delta < \infty$, and that in addition the distributions of a_i and g_t satisfy Cramér's condition (4.7). Then, if $\sigma_a^2 + \sigma_g^2 \geq C$ for some $C > 0$ we have*

$$\left\| \mathbb{P}_{NT}^* \left(\sqrt{NT} \frac{\bar{Y}_{NT}^* - \bar{Y}_{NT}}{\hat{S}_{NT,sel}^*} \right) - \mathbb{P}_{NT} \left(\sqrt{NT} \frac{\bar{Y}_{NT} - \mathbb{E}[Y_{it}]}{\hat{S}_{NT,sel}} \right) \right\|_\infty = O_P(r_{NT}^{-2} \vee (NT)^{-1/2})$$

for all three versions of the bootstrap, (BS-S), (BS-N), and (BS-C).

See the appendix for a proof. In the non-degenerate case, the limiting distribution is dominated by the components a_i, g_t. Since these are i.i.d draws from their respective distributions, we can rely on arguments from chapter 5 in Hall (1992) for the nonparametric bootstrap with i.i.d. data, after ensuring that the contribution of $v_{it} + e_{it}$ vanishes at a sufficiently fast rate. The assumption that all moments of the distribution are bounded is stronger than needed but commonly assumed in the literature (see e.g. Andrews (2002)). In general the Edgeworth expansion of the sampling distribution only requires four bounded moments, however the order of bounded moments needed for the Edgeworth expansion of the bootstrap distribution is more involved and, while finite, not stated explicitly in Theorem 5.2 in Hall (1992).

5. Simulation Study

We now present simulation results to demonstrate the performance of the bootstrap procedure for inference regarding the sample mean, where we consider balanced and unbalanced designs with additively separable and nonseparable cluster effects. Particular attention is given to the degenerate cases of uncorrelated observations, and drifting sequences. We report simulation results for each of the bootstrap approaches proposed in Sections 3 and 6:

- **(BS-S)** inference based on the bootstrap with model selection,
- **(BS-N)** inference based on the bootstrap without model selection,
- **(BS-C)** inference based on the conservative bootstrap.

In addition, we consider the following alternative inference approaches,

- **(GAU)** "plug-in" Gaussian inference using a two-way clustering robust estimator for the asymptotic variance of $\bar{Y}_{NT}$,
- **(PGH)** inference based on the Pigeonhole bootstrap estimate for the asymptotic distribution of $r_{NT}\bar{Y}_{NT}$, and
- **(SUB)** inference based on the subsampling estimate for the asymptotic distribution of $r_{NT}\bar{Y}_{NT}$.

see Appendix B in the online supplement for a precise definition and theoretical results. Our simulation designs also consider the following alternative implementations for these procedures:

- **(REG)** inference based on the asymptotic distribution of the mean, $r_{NT}\bar{Y}_{NT}$.
- **(PIV)** inference based on the asymptotic distribution of the studentized mean, where we use $t_{NT} := (NT)^{1/2}\,\hat{S}_{NT,sel}^{-1}\bar{Y}_{NT}$ for BS-N and PGH, and $t_{NT} := (NT)^{1/2}\,\hat{S}_{NT,sel}^{-1}\bar{Y}_{NT}$ for BS-S and BS-C.
- **(SYM)** symmetric inference based on the asymptotic distribution of the absolute value of the studentized mean, $|t_{NT}|$.

According to our theoretical results in Sections 4, 6, and Appendix B in the online supplement, each of these inference procedures is asymptotically valid in the non-degenerate cases, while the pivotal and symmetric bootstrap (PIV and SYM, respectively) provide refinements over their non-pivotal analogs (REG), subsampling, or Gaussian asymptotic inference. It also follows from standard arguments (see e.g. Horowitz (2000)) that theoretical refinements from SYM are of a higher order than those obtained for PIV.

5.1. **Additively Separable Designs.** For the first set of results, we generate a two-way clustered array according to the additively separable design

$$y_{it} = \sigma_a \alpha_i + \sigma_g \gamma_t + \sigma_e \varepsilon_{it}$$

where $\gamma_t, \varepsilon_{it}$ are i.i.d. standard normal. We generated $\alpha_i = (\zeta_i - \mu_\alpha)/\tau_\alpha$ for $\log \zeta_i \sim N(0,1)$, where $\mu_\alpha = \mathbb{E}[\zeta_i]$, and $\tau_\alpha^2 = \text{Var}(\alpha_i)$ were obtained using analytic formulae for the moments of the log-normal distribution. In particular, the distribution of α_i is skewed to the right.

Our simulation designs vary the relative importance of the three factors through the choice of $\sigma_a, \sigma_g, \sigma_e$. Design 1 (non-degenerate case) chooses $\sigma_a^2 = 0.5$, $\sigma_g^2 = 0.1$, and $\sigma_e^2 = 0.2$, Design 2 considers the drifting sequence $\sigma_a^2 = 5/T$, $\sigma_g^2 = 1/N$, and $\sigma_e^2 = 0.2$. Design 3 (degenerate case) sets $\sigma_a^2 = \sigma_g^2 = 0$ and $\sigma_e^2 = 0.2$. For each design in this section, simulation results were obtained from $10,000$ simulated samples with bootstrap distributions approximated using $2,000$ bootstrap draws. All rejection rates are reported as percentages.

Results for the balanced case are given in Tables 1 and 2 and largely support our theoretical claims. In particular, for all procedures rejection rates approach the nominal 0.05 significance level as N and T grow. The only exception to this is the pointwise consistent bootstrap (BS-S) under the drifting sequences in Design 2, much in line with the general concerns about post-selection inference in Leeb and Pötscher (2005). In particular, the results are consistent with the bootstrap without model selection being uniformly valid regarding clustering in means. For Design 1, the pivotal and symmetric versions of the different bootstrap procedures show marked improvements over their standard versions or Gaussian asymptotic inference, which is consistent with asymptotic refinements established in Theorem 4.2. The conservative bootstrap is consistent in the non-degerate case, but conservative under the degenerate Designs 2 and 3. Also, the pigeonhole bootstrap is consistent in its pivotal version across all designs, but the non-pivotal version is conservative in the degenerate case.

The improvements in coverage rates from asymptotic refinements are more pronounced for one-sided than two-sided rejection rates in Table 2. We can see from the simulation results that the respective biases in estimating percentiles in the lower and upper tails of the distribution via GAU have opposite signs, so that these biases partially offset each other for two-sided tests. Design 2 considers drifting sequences of DGPs for which Theorem 4.2 does

		GAU	BS-S			BS-N			BS-C	PGH		SUB
N	T	REG	REG	PIV	SYM	REG	PIV	SYM	PIV	REG	PIV	REG
						Design 1						
10	10	8.62	10.15	7.17	6.08	10.12	7.27	6.12	7.13	8.88	7.20	14.13
20	20	6.82	7.63	6.65	5.49	7.50	6.57	5.61	6.61	7.03	6.70	9.49
50	50	6.26	6.56	6.22	5.42	6.54	6.21	5.46	6.21	6.38	6.26	7.88
100	100	5.58	5.79	5.63	5.13	5.82	5.58	5.06	5.67	5.76	5.55	6.81
						Design 2						
10	10	8.58	9.52	6.36	6.04	8.38	6.55	6.30	2.21	3.57	5.90	10.44
20	20	8.25	8.82	7.04	6.86	7.22	6.09	6.19	2.65	2.60	5.85	8.24
50	50	7.54	7.78	6.89	6.74	5.77	5.33	5.23	2.39	2.03	5.16	6.98
100	100	6.70	7.02	6.26	6.41	4.83	4.50	4.60	1.90	1.60	4.55	6.36
						Design 3						
10	10	5.19	4.64	2.94	2.87	2.97	5.92	5.89	0.01	0.29	2.84	5.26
20	20	5.99	5.40	4.67	4.61	3.29	6.22	6.14	0.00	0.12	3.23	5.28
50	50	5.16	4.98	4.77	4.61	3.35	5.94	5.99	0.00	0.11	3.35	4.83
100	100	5.17	5.04	4.99	4.94	3.79	5.73	5.74	0.00	0.10	3.82	5.08

TABLE 1. Balanced separable case: false rejection rates for two-sided tests of the null $\mathbb{E}[Y_{it}] = 0$ at the 5 percent significance level. Design 1: $\sigma_a^2 = 0.5$, $\sigma_g^2 = 0.1$, $\sigma_e^2 = 0.2$; Design 2: $\sigma_a^2 = 0.5/T, \sigma_g^2 = 0.1/N, \sigma_e^2 = 0.2$; Design 3: $\sigma_a^2 = \sigma_g^2 = 0, \sigma_e^2 = 0.2$.

not predict refinements. For Design 3, our theoretical results do not imply refinements for PIV or SYM since for that specification, $y_{it} = \sigma_\varepsilon \varepsilon_{it}$ is i.i.d. Gaussian.

We also simulate the absolute error in rejection probabilities based on GAU, SUB, and BS-S (pivotal and non-pivotal) at all percentiles for Design 1. Specifically, we estimate the percentiles of the sampling distribution for each simulated sample using either method, and simulate the frequency at which the t-statistic for the sample exceeds each percentile. Figure 1 reports the absolute difference between the simulated and nominal rejection frequencies. We find that for all three methods, the absolute discrepancy between nominal and simulated rejection rates decreases as N and T grow across all percentiles. The non-pivotal bootstrap does not exhibit a clear improvement relative to plug-in asymptotic approximation, whereas rejection rates based on the pivotal bootstrap for the studentized mean are consistently closer to nominal levels.

We next assess the importance of balance in the relative sizes of N and T, as well as the relative importance of clustering in either dimension. In particular, we first consider balanced designs $T = N$ where we set $\sigma_a = 0.5, \sigma_g = 0.1$ and $\sigma_e = 0.1$. We then consider unbalanced designs where we let $N = 10, 20, 50, 100$ vary while holding $T = 20$ fixed, see Table 3 for simulation results. While the bootstrap is not asymptotically valid if T remains

		GAU	BS-S	BS-N	BS-C	PGH	SUBS	GAU	BS-S	BS-N	BS-C	PGH	SUB
		REG	PIV	PIV	PIV	PIV	REG	REG	PIV	PIV	PIV	PIV	REG
							Design 1						
10	10	10.38	7.68	7.75	7.70	7.85	14.01	3.40	3.86	3.74	3.76	3.77	5.31
20	20	8.51	6.62	6.47	6.58	6.55	10.95	3.13	4.18	4.24	4.19	4.13	3.22
50	50	7.58	6.01	5.89	5.93	5.94	9.78	3.80	5.31	5.23	5.22	5.22	3.27
100	100	6.88	5.47	5.43	5.61	5.50	8.87	3.85	4.84	4.86	4.86	4.91	3.28
							Design 2						
10	10	9.24	7.45	7.63	3.05	7.30	10.12	4.59	4.43	4.65	2.54	4.30	5.68
20	20	8.82	7.58	6.77	3.46	6.64	8.49	4.67	5.22	4.86	3.02	4.74	4.46
50	50	8.24	7.36	6.00	3.33	6.03	7.88	4.60	5.19	4.31	2.72	4.43	4.19
100	100	7.49	6.82	5.40	2.50	5.44	7.18	5.22	5.81	4.67	2.39	4.62	4.90
							Design 3						
10	10	4.82	3.14	4.87	0.04	3.09	4.56	5.16	3.35	5.49	0.04	3.52	5.14
20	20	5.23	4.30	5.06	0.00	3.41	4.68	5.47	4.60	5.40	0.01	3.69	4.97
50	50	5.21	4.79	5.26	0.00	3.78	4.84	5.40	4.98	5.47	0.00	3.81	5.13
100	100	5.18	4.98	5.54	0.00	4.26	5.04	4.70	4.58	5.07	0.00	3.80	4.58

TABLE 2. Balanced separable case: false rejection rates for one-sided tests of the null $\mathbb{E}[Y_{it}] \leq 0$ (left half of the panel) $\mathbb{E}[Y_{it}] \geq 0$ (right half of the panel) at the 5 percent significance level. Design 1: $\sigma_a^2 = 0.5$, $\sigma_g^2 = 0.1$, $\sigma_e^2 = 0.2$; Design 2: $\sigma_a^2 = 0.5/T, \sigma_g^2 = 0.1/N, \sigma_e^2 = 0.2$; Design 3: $\sigma_a^2 = \sigma_g^2 = 0, \sigma_e^2 = 0.2$.

fixed, results are broadly in line with those for the balanced case for the corresponding sample size. Overall, these results are again consistent with theoretical predictions on asymptotic validity and refinements.

5.2. **Nonseparable Designs.** Finally, we simulate a model with non-separable cluster effects, where we specify

$$y_{it} = (\alpha_i + \mu_\alpha)(\gamma_t + \mu_\gamma) - \mu_\alpha \mu_\gamma + \varepsilon_{it}$$

for i.i.d. standard normal random variables α_i, γ_t and ε_{it}. We consider one non-degenerate design with $\mu_\alpha = \mu_\gamma = 1$ (Design 1), and an alternative design with $\mu_\alpha = \mu_\gamma = 0$ for which y_{it} is not clustered in means (Design 3), as well as a design with drifting sequences (Design 2), see Table 4 for simulation results. Since 2.5th and 97.5th percentiles the Wiener chaos distribution resulting from this design differ only slightly from those of the standard normal, we also report false rejection rates for tests at the 1 percent nominal level. For an easier interpretation of the simulation results for non-Gaussian limits, we also report the theoretical limits of coverage probabilities, $N = \infty$ and $T = \infty$, in a separate row.

The point-wise consistent procedures (bootstrap with model selection and subsampling) should do well under Designs 1 and 3, where subsampling is consistent at a much slower

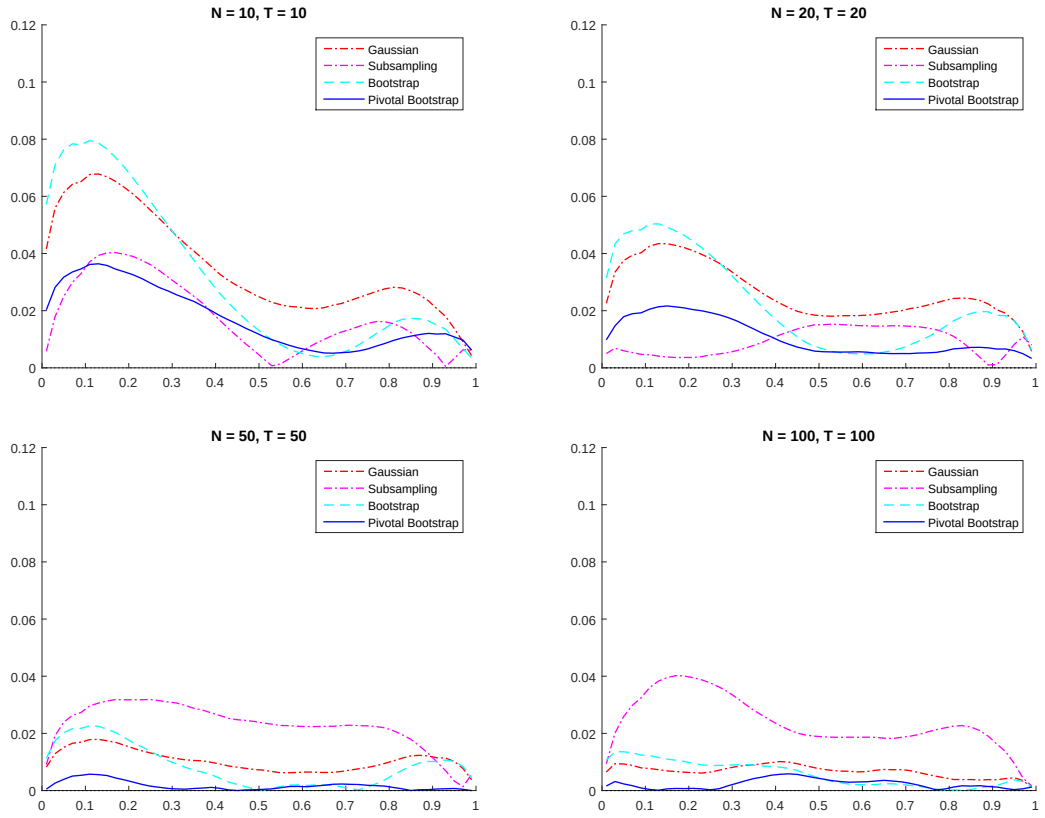

FIGURE 1. Balanced separable case: Absolute error in estimated c.d.f., plotted against nominal percentiles. Plots are based on Design 1: $\sigma_a^2 = 1, \sigma_g^2 = 0.2, \sigma_e^2 = 1$.

		GAU	BS-S			BS-N			BS-C	PGH		SUB
N	T	REG	REG	PIV	SYM	REG	PIV	SYM	PIV	REG	PIV	REG
						Design 1						
10	20	9.76	11.18	8.53	6.68	11.14	8.71	6.76	8.62	10.40	8.76	14.47
20	20	7.21	8.02	7.16	5.79	7.95	7.11	5.76	7.17	7.25	7.11	9.75
50	20	5.57	5.98	5.16	4.95	5.96	5.17	4.96	5.30	5.65	5.27	7.31
100	20	5.73	6.03	5.29	5.22	6.17	5.24	5.06	5.36	5.81	5.29	7.10
						Design 2						
10	20	5.79	5.12	3.90	3.98	3.32	6.17	6.19	0.00	0.28	3.33	5.14
20	20	5.66	5.10	4.43	4.41	3.03	5.79	5.77	0.00	0.11	3.08	4.80
50	20	6.06	5.70	5.13	5.12	3.48	6.37	6.31	0.00	0.12	3.49	5.38
100	20	5.90	5.37	5.02	5.07	3.65	6.14	6.17	0.00	0.13	3.88	5.22

TABLE 3. Unbalanced separable case: false rejection rates for two-sided tests of the null $\mathbb{E}[Y_{it}] = 0$ at the 5 percent significance level. Design 1: $\sigma_a^2 = 0.5, \sigma_g^2 = 0.1, \sigma_e^2 = 0.2$; Design 2: $\sigma_a^2 = \sigma_g^2 = 0, \sigma_e^2 = 0.2$.

rate than the bootstrap. Since none of the inference procedures is uniformly consistent, we should expect all of these to perform poorly under Design 2. However, given our theoretical results the conservative bootstrap is the only procedure that is guaranteed to be conservative across all designs.

We find that in the non-degenerate case $\mu_\alpha \neq 0$ or $\mu_\gamma \neq 0$, the bootstrap produces results that are comparable to the separable case. According to our theoretical results, all procedures are asymptotically valid, whereas PIV and SYM should produce refinements, which is consistent with the first set of simulation results.

For the degenerate case, $\mu_\alpha = \mu_\gamma = 0$, theory predicts that Gaussian inference is not asymptotically valid even when a consistent estimator of the asymptotic variance is used. We find that indeed that for the plug-in asymptotic approximation based on the Gaussian distribution rejection rates appear to converge to a value that is different from the nominal level, and based on the theoretical properties, bias in rejection rates should be expected to persist for arbitrarily large sample sizes. We do report simulated rejection rates for the corresponding limiting distribution (rows with $N = T = \infty$) which show that for the simulation designs considered here the asymptotic size distortions remain modest in magnitude, but actual rejection rates are above nominal size even in the limit for tests at the 5 percent and 1 percent level.

The bootstrap with model selection and subsampling are point-wise consistent (see Designs 1 and 3), but yield invalid inference under the drifting sequences in Design 2. The conservative bootstrap is consistent in the non-degenerate case (Design 1), but conservative under the other scenarios. Theoretical results do not indicate that the bootstrap without model selection or the pigeonhole bootstrap should be necessarily conservative in the degenerate cases (Designs 2 and 3), but the simulation results nevertheless show that rejection rates are essentially zero. Also, since the studentized mean is not asymptotically pivotal under Designs 2 and 3, theory also does not predict refinements for the pivotal or symmetric versions of either bootstrap procedure. This is reflected in the simulation results, showing no systematic difference between the alternative implementations of each bootstrap.

As for the separable case, we also simulate the absolute error in rejection probabilities based on the Gaussian, Subsampling, and bootstrap estimates with model selection (pivotal and non-pivotal) at all percentiles for the degenerate case in Design 3, which are shown in Figure 2. These results support the theoretical predictions that Gaussian plug-in inference is inconsistent for the degenerate nonseparable case, and that subsampling is consistent although at a slower rate than the bootstrap (pivotal or not) with model selection. Also, the theory does not imply asymptotic refinements for the pivotal bootstrap in this setting, so we should not expect the pivotal bootstrap to perform systematically better than its non-pivotal version.

		GAU	BS-S			BS-N			BS-C	PGH		SUB
N	T	REG	REG	PIV	SYM	REG	PIV	SYM	PIV	REG	PIV	REG
						Design 1 (tests at 5 percent nominal size)						
10	10	8.18	9.80	4.90	3.74	9.77	4.88	3.80	4.85	9.90	4.56	15.28
20	20	6.50	7.26	4.76	4.62	7.06	4.83	4.52	4.79	7.24	4.68	9.35
50	50	5.19	5.34	4.71	4.60	5.53	4.81	4.52	4.78	5.55	4.61	6.50
100	100	5.17	5.38	4.97	4.96	5.26	4.92	4.92	4.84	5.41	5.00	6.18
∞	∞	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
						Design 2 (tests at 5 percent nominal size)						
10	10	7.83	9.48	4.99	4.03	8.12	4.25	3.61	3.67	9.94	1.80	16.15
20	20	6.22	6.78	4.77	4.36	5.66	4.00	3.69	3.10	7.01	1.25	9.74
50	50	6.01	6.25	5.41	5.30	4.59	4.00	3.81	3.24	6.01	1.25	7.90
100	100	5.39	5.54	5.12	5.07	3.77	3.50	3.57	2.47	5.25	0.98	6.33
						Design 3 (tests at 5 percent nominal size)						
10	10	7.41	6.38	3.18	3.12	0.21	0.09	0.05	0.04	0.66	0.00	6.69
20	20	6.71	5.21	3.55	3.53	0.05	0.03	0.03	0.01	0.26	0.00	5.02
50	50	5.87	4.56	3.98	4.03	0.04	0.03	0.02	0.02	0.14	0.00	4.55
100	100	6.36	4.95	4.40	4.43	0.02	0.03	0.01	0.00	0.13	0.00	4.89
∞	∞	6.5	5.00	5.00	5.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00
						Design 1 (tests at 1 percent nominal size)						
10	10	2.86	3.91	1.45	0.75	3.92	1.42	0.79	1.37	4.00	1.19	9.79
20	20	1.67	2.29	0.96	0.72	2.21	1.02	0.76	0.99	2.14	0.91	4.04
50	50	1.12	1.22	0.83	0.83	1.18	0.75	0.83	0.85	1.29	0.89	1.71
100	100	1.09	1.16	0.91	0.94	1.20	0.93	0.89	0.96	1.18	0.94	1.44
∞	∞	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
						Design 2 (tests at 1 percent nominal size)						
10	10	2.01	2.98	0.96	0.56	2.20	0.63	0.30	0.58	4.14	0.14	9.57
20	20	1.29	1.58	0.76	0.57	1.05	0.51	0.45	0.49	2.30	0.05	3.27
50	50	0.92	0.97	0.75	0.60	0.75	0.58	0.51	0.47	1.69	0.04	1.37
100	100	0.91	0.90	0.68	0.65	0.37	0.31	0.27	0.19	1.19	0.01	0.47
						Design 3 (tests at 1 percent nominal size)						
10	10	4.01	1.78	0.80	0.84	0.02	0.00	0.00	0.00	0.16	0.00	2.67
20	20	2.67	0.85	0.55	0.53	0.00	0.00	0.00	0.00	0.04	0.00	1.00
50	50	2.86	0.71	0.67	0.62	0.00	0.00	0.00	0.00	0.01	0.00	0.80
100	100	2.64	0.81	0.75	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.79
∞	∞	3.2	1.00	1.00	1.00	0.000	0.000	0.000	0.000	0.000	0.000	1.00

TABLE 4. Non-separable case: false rejection rates for two-sided tests of the null $\mathbb{E}[Y_{it}] = 0$ at a nominal level of 1 percent. Design 1: $\sigma_a^2 = 0.2, \sigma_g^2 = 0.2, \sigma_e^2 = 0.2$, $\mu_a = 1$, and $\mu_g = 0$; Design 2: $\sigma_a^2 = 0.2, \sigma_g^2 = 0.2, \sigma_e^2 = 0.2$, $\mu_a = 1/\sqrt{T}$, and $\mu_g = 0$; Design 3: $\sigma_a^2 = 0.2, \sigma_g^2 = 0.2, \sigma_e^2 = 0$ and $\mu_a = \mu_g = 0$. The first two panels are for tests at a nominal level of 5 percent, the bottom panel are at 1 percent.

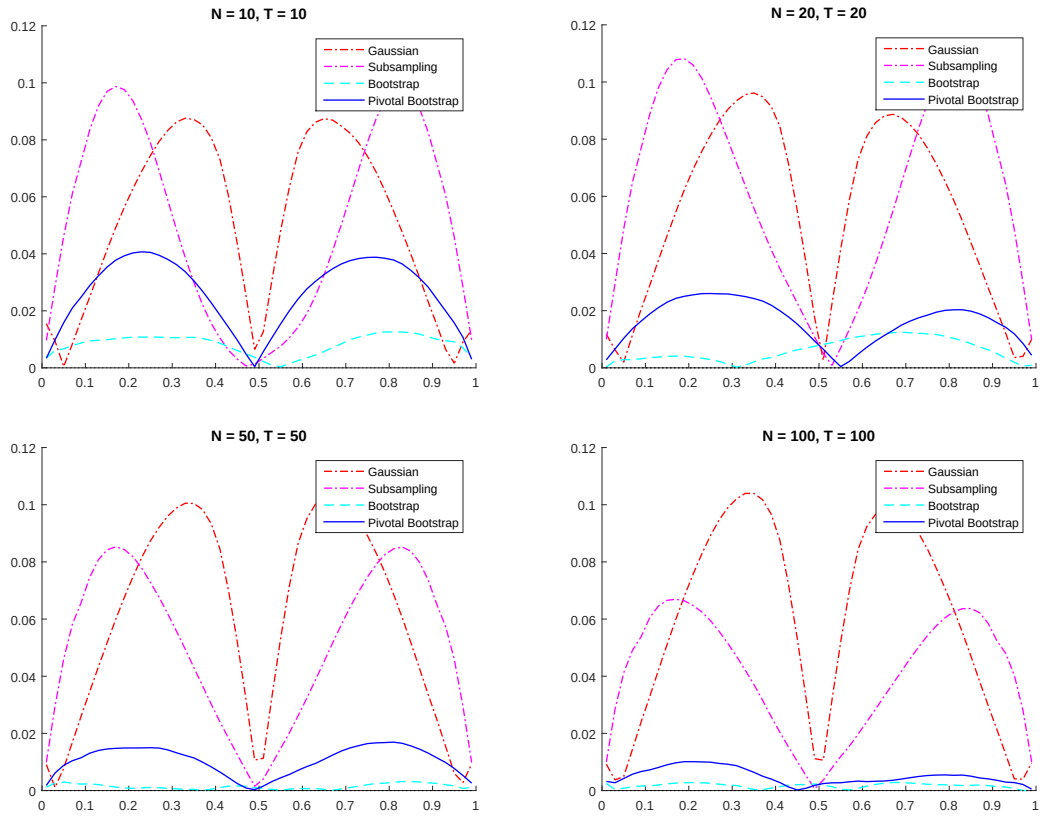

FIGURE 2. Nonseparable case: Absolute error in estimated c.d.f., plotted against nominal percentiles. Plots are based on Design 2: $\sigma_a^2 = 0.5, \sigma_g^2 = 0.5, \sigma_e^2 = 0.1$ and $\mu_a = \mu_g = 0$.

6. Inference in Regression Models

As an important application, we next discuss inference in a regression model with a scalar dependent variable y_{it} and k regressors $\mathbf{x}_{it} \in \mathbb{R}^k$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, allowing for two-way dependence in residuals. Stacking observations, we also denote $\mathbf{y} := (y_{11}, y_{21}, \ldots, y_{NT})'$ and $\mathbf{X} := (\mathbf{x}_{11}, \mathbf{x}_{21}, \ldots, \mathbf{x}_{NT})'$. For the linear projection model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}, \quad \mathbb{E}[\mathbf{x}_{it} u_{it}] = 0 \tag{6.1}$$

we then consider random-design inference regarding the coefficient $\boldsymbol{\beta}$ based on the least squares (LS) estimator[11]

$$\hat{\boldsymbol{\beta}}_{LS} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}u_{it}\right)$$

[11]As in the case of i.i.d. observations, fixed-design inference (i.e. inference conditional on $\mathbf{X}$) would require the stronger assumption that the linear regression function is correctly specified, $\mathbb{E}[u_{it}|\mathbf{X}] = 0$, and rely on conditional exchangeability considerations. If $\mathbf{x}_{it}$ has finite support, Lemma 3.3 in Crane and Towsner (2018) can be used to obtain a conditional Aldous-Hoover representation of u_{it}.

To nest this regression model into our framework, we will assume that the products $(\mathbf{x}_{it}u_{it})_{i,t}$ also constitute a dissociated, separately exchangeable array. This yields the Aldous-Hoover representation

$$\mathbf{z}_{it} := \mathbf{x}_{it}u_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it})$$

where $\alpha_i, \gamma_t, \varepsilon_{it}$ are i.i.d., and can without loss of generality be assumed to follow a uniform distribution. We can therefore find an orthogonal decomposition of $\mathbf{z}_{it}$ that is analogous to that for Y_{it} in the unconditional case. Specifically, we denote

$$\begin{aligned}
\mathbf{a}_i &:= \mathbb{E}[\mathbf{x}_{it}u_{it}|\alpha_i], \quad \mathbf{g}_t := \mathbb{E}[\mathbf{x}_{it}u_{it}|\gamma_t] \\
\mathbf{v}_{it} &:= \mathbb{E}[\mathbf{x}_{it}u_{it}|\alpha_i, \gamma_t] - \mathbf{a}_i - \mathbf{g}_t \\
\mathbf{e}_{it} &:= \mathbf{x}_{it}u_{it} - \mathbb{E}[\mathbf{x}_{it}u_{it}|\alpha_i, \gamma_t]
\end{aligned}$$

with components $\mathbf{a}_i = (a_{i1}, \ldots, a_{ik})'$, $\mathbf{g}_t = (g_{t1}, \ldots, g_{tk})'$, $\mathbf{v}_{it} = (v_{it1}, \ldots, v_{itk})'$, and $\mathbf{e}_{it} = (e_{it1}, \ldots, e_{itk})'$. We also denote the unconditional component variances with $\sigma^2_{al}, \sigma^2_{gl}, \sigma^2_{vl}, \sigma^2_{el}$ and let

$$\mathbf{w}_{it} := \mathbf{x}_{it}u_{it} - \mathbf{a}_i - \mathbf{g}_t = \mathbf{v}_{it} + \mathbf{e}_{it}$$

so that

$$\mathbf{x}_{it}u_{it} = \mathbf{a}_i + \mathbf{g}_t + \mathbf{w}_{it}$$

and $\sigma^2_{wl} := \mathrm{Var}(w_{itl}) = \sigma^2_{vl} + \sigma^2_{el}$.

Given the least squares residuals $\hat{u}_{it} := y_{it} - \mathbf{x}'_{it}\hat{\boldsymbol{\beta}}_{LS}$, we can construct an empirical analog of that decomposition, where

$$\begin{aligned}
\hat{\mathbf{a}}_i &:= \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}\hat{u}_{it} \\
\hat{\mathbf{g}}_t &:= \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{it}\hat{u}_{it}, \text{ and} \\
\hat{\mathbf{w}}_{it} &:= \mathbf{x}_{it}\hat{u}_{it} - \hat{\mathbf{a}}_i - \hat{\mathbf{g}}_t.
\end{aligned}$$

For each $l = 1, \ldots, k$ we also let $\hat{\lambda}_{al} := \frac{\hat{D}_{al}(\kappa_a)T\hat{\sigma}^2_{al}}{\hat{D}_{al}(\kappa_a)T\hat{\sigma}^2_{al} + \hat{\sigma}^2_{wl}}$ and $\hat{\lambda}_{gl} := \frac{\hat{D}_{gl}(\kappa_g)N\hat{\sigma}^2_{gl}}{\hat{D}_{gl}(\kappa_g)N\hat{\sigma}^2_{gl} + \hat{\sigma}^2_{wl}}$ for the bootstrap with or without model selection, where $\hat{\sigma}^2_{al}, \hat{\sigma}^2_{gl}, \hat{\sigma}^2_{wl}$, and $\hat{D}_{al}(\cdot), \hat{D}_{gl}(\cdot)$ are defined in an analogous fashion as in Section 3.

We can then implement the bootstrap algorithm from Section 3 as follows:

(a) For the bth bootstrap iteration, draw $\mathbf{a}^*_{i,b} := \hat{\mathbf{a}}_{k^*_b(i)}$ and $\mathbf{g}^*_{t,b} := \hat{\mathbf{g}}_{s^*_b(t)}$, where $k^*_b(i)$ and $s^*_b(t)$ are i.i.d. draws from the discrete uniform distribution on the index sets $\{1, \ldots, N\}$ and $\{1, \ldots, T\}$, respectively.

(b) Generate $\mathbf{w}^*_{it,b} := \omega_{1i,b}\omega_{2t,b}\hat{\mathbf{w}}_{k^*_b(i)s^*_b(t)}$, where $\omega_{1i,b}, \omega_{2t,b}$ are i.i.d. random variables with $\mathbb{E}[\omega_\cdot] = 0, \mathbb{E}[\omega^2_\cdot] = \mathbb{E}[\omega^3_\cdot] = 1$

(c) Simulate values of $\mathbf{z}^*_{it,b} = (z^*_{it1,b}, \ldots, z^*_{itk,b})'$ where the lth component is given by

$$z^*_{itl,b} := \sqrt{\hat{\lambda}_{al}} a^*_{il,b} + \sqrt{\hat{\lambda}_{gl}} g^*_{tl,b} + w^*_{itl,b}$$

(d) We then compute

$$\hat{\boldsymbol{\beta}}^*_{LS,b} := \hat{\boldsymbol{\beta}}_{LS} + (\mathbf{X}'\mathbf{X})^{-1} \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{z}^*_{it,b} \right)$$

for each bootstrap sample.

We can then approximate the asymptotic distribution of $r_{NT}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})$ with the simulated distribution of $r_{NT}(\hat{\boldsymbol{\beta}}^*_{LS,b} - \hat{\boldsymbol{\beta}}_{LS})$.[12] As for the sample mean, this bootstrap procedure can be implemented in three different variants: with model selection (BS-S), with no model selection (BS-N), and a conservative bootstrap (BS-C) using the alternative choices for the regularization parameters κ_a, κ_g given in Definition 3.1. For asymptotic results on the bootstrap for regression models we make the following assumptions:

Assumption 6.1. ***(Regression)*** *Assume the regression model in (6.1), where* $\mathbf{x}_{it}u_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it})$ *and* $\alpha_i, \gamma_t, \varepsilon_{it}$ *are i.i.d. uniform on* $[0,1]$*. Furthermore, (a) the matrix* $\mathbf{X}$ *has full column rank. (b) For each* $l = 1, \ldots, k$ *and some* $\delta > 0$*, the* $(4+\delta)$*th absolute moments of* x_{itl} *are bounded, and the* $(4+\delta)$*th conditional moments of each component* $a_{il}/\sqrt{\mathrm{Var}(a_{il}|\mathbf{X})}$*,* $g_{tl}/\sqrt{\mathrm{Var}(g_{tl}|\mathbf{X})}$*,* $v_{itl}/\sqrt{\mathrm{Var}(v_{itl}|\mathbf{X})}$*, and* $e_{itl}/\sqrt{\mathrm{Var}(e_{itl}|\mathbf{X})}$ *given* $\mathbf{X}$ *are bounded whenever the conditional variance of either component is strictly positive. (c) The unconditional variances satisfiy* $\mathrm{Var}(a_{il}) + \mathrm{Var}(g_{tl}) > 0$ *or* $\mathrm{Var}(w_{itl}) > 0$ *for each* $l = 1, \ldots, k$*, and (d) for each component of* $\mathbf{z}_{it} = \mathbf{x}_{it}u_{it}$ *there exists a spectral representation satisfying Assumption 2.2.*

The asymptotic properties of the bootstrap under alternative choices for $\hat{\lambda}_{al}, \hat{\lambda}_{gl}$ are then analogous to the case of the sample mean:

Proposition 6.1. ***(Regression Inference)*** *Suppose that Assumption 6.1 holds. Then (a) the estimator* $\hat{\boldsymbol{\beta}}_{LS}$ *is consistent at the* r_{NT} *rate. (b) The bootstrap with model selection satisfies (4.3) and (4.4) pointwise as* $\sigma^2_{al}, \sigma^2_{gl}, \sigma^2_{el}, \sigma^2_{vl}$ *are held fixed for all* $l = 1, \ldots, k$*, the bootstrap without model selection satisfies satisfies (4.3) and (4.4) uniformly if* $q_{vl} = 0$ *for each* $l = 1, \ldots, k$*. (c) The conservative bootstrap satisfies (4.5) and (4.6) uniformly over the entire parameter space.*

[12] In principle, the analogous procedure could be applied to the studentized estimator $r_{NT}\hat{\mathbf{V}}^{-1/2}_{LS}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})$, where we let $\hat{\mathbf{V}}_{LS}$ denote a two-way cluster-robust estimator of the asymptotic variance covariance matrix of $\hat{\boldsymbol{\beta}}_{LS}$. However, for expositional clarity we state our formal results for the non-studentized least squares estimator instead.

Since $(\mathbf{z}_{it})$ is a separately exchangeable array, this result follows from applying Theorems 4.1 and 4.2 to $Y_{it} := \mathbf{c}'\mathbf{z}_{it}$ for a vector $\mathbf{c} \in \mathbb{R}^k$. Specifically, Assumption 6.1 implies Assumptions 2.1 and 2.2 for any linear combination of that form. The conclusion then follows from the continuous mapping theorem together with the Cramér-Wold device.

One interesting question is whether the distribution of x_{it} is such that the asymptotic distribution of the LS estimator is guaranteed to be Gaussian regardless of dependence in u_{it}. Here it is possible to use Theorem 1 in de Jong (1990) to obtain sufficient condition for conditional asymptotic normality of bilinear forms $V_k := \mathbf{Z}_{1k}'\mathbf{X}\mathbf{Z}_{2k}$ of random vectors $\mathbf{Z}_{1k}, \mathbf{Z}_{2k}$ given the matrix $\mathbf{X}$. For example, using calculations analogous to those for the term $\hat{Z}_{NT}^{e,*}$ in the proof of Theorem A.2 it is possible to verify that under the conditions of this paper, V_k is asymptotically Gaussian if $\ddot{\mathbf{x}}_{it}, \ddot{\mathbf{x}}_{js}$ are mean-independent for any $(j,s) \neq (i,t)$. For difference-in-differences designs with a regressor $x_{it1} := \mathbb{1}\{t \geq T_i\}$ for a unit-specific intervention date T_i, or when $\mathbf{x}_{it} := \mathbf{x}(\xi_i, \zeta_t)$ are a non-additive function of row- and column-level attributes ξ_i and ζ_t, respectively, these conditions need not hold in general.

6.1. **Simulation Study.** We now illustrate the performance of the bootstrap for regression inference in a brief Monte Carlo study. For all simulation designs, we consider a regression model

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$$

where we generate u_{it} according to

$$u_{it} = (\mu_a + \sigma_a \alpha_i)(\mu_g + \sigma_g \gamma_t)$$

and $\alpha_i, \gamma_t \overset{iid}{\sim} N(0,1)$. As for the non-separable design for the sample mean, this design produces non-degenerate distributions of a_i and g_t if $\mu_g \neq 0$ or $\mu_a \neq 0$, and a non-Gaussian limiting distribution in the degenerate case, $\mu_a = \mu_g = 0$. For the main results, we generate x_{it} according to

$$x_{it} = \mathbb{1}\{t \geq T_i\}$$

where T_i are i.i.d. discrete uniform on $\{1, \ldots, T\}$. This setup is supposed to mimic difference-in-differences designs for which we do not expect a Gaussian limiting distribution. We also report results for other distributions for x_{it}.

Our simulation designs vary with regard to the choice of μ_a and μ_g. Design 1 (non-degenerate case) chooses $\mu_a = 5$ and $\mu_g = 0$, Design 2 considers the drifting sequence $\mu_a = 5/\sqrt{T}$ and $\mu_g^2 = 0$. Design 3 (degenerate case) sets $\mu_a = \mu_g = 0$. Throughout, we set $\sigma_a^2 = \sigma_g^2 = 0.2$. For each design we vary sample size between $N = T \in \{20, 50, 100, 200\}$. Simulation results were obtained from $10,000$ simulated samples with bootstrap distributions approximated using $2,000$ bootstrap draws. We compare the following procedures based on the least-squares estimator: Gaussian inference (GAU-LS), pivotal bootstrap with

		$H_0 : \beta = 0$				$H_1 : \beta = 0.1/\sqrt{N}$			
N	T	GAU	PIV	SYM	CONS	GAU	PIV	SYM	CONS
		Design 1 (tests at 5 percent nominal size)							
10	10	7.92	5.50	5.47	5.05	18.12	13.54	14.02	12.32
20	20	5.95	4.91	4.96	4.79	14.38	11.96	12.66	11.64
50	50	5.67	5.24	5.19	4.97	13.22	12.18	12.64	11.66
100	100	5.06	4.76	4.74	4.71	12.94	12.18	12.84	12.16
200	200	5.14	5.20	5.20	5.12	14.88	14.26	14.84	14.26
		Design 2 (tests at 5 percent nominal size)							
10	10	10.18	8.06	8.02	2.64	71.34	67.66	68.34	29.88
20	20	8.21	7.67	7.49	0.60	92.80	92.70	93.08	29.64
50	50	11.37	11.23	11.33	0.27	99.86	99.84	99.88	32.60
100	100	17.04	16.30	16.54	0.32	100.00	100.00	100.00	33.70
200	200	21.06	20.45	20.38	0.19	100.00	100.00	100.00	32.80
		Design 3 (tests at 5 percent nominal size)							
10	10	9.73	5.38	5.28	0.15	99.90	99.85	99.85	15.60
20	20	7.74	5.33	5.39	0.01	99.95	99.95	99.95	12.70
50	50	6.53	5.07	5.13	0.01	100.00	100.00	100.00	12.35
100	100	6.15	5.03	5.02	0.00	100.00	100.00	100.00	10.50
200	200	6.14	5.11	5.20	0.00	100.00	100.00	100.00	11.40

TABLE 5. Rejection rates for two-sided tests of $H_0 : \beta_1 = 0$ at a nominal level of 5 percent under the null and a local alternative. Design 1: $\sigma_a^2 = 0.2, \sigma_g^2 = 0.2$ and $\mu_a = 5$, $\mu_g = 0$; Design 2: $\sigma_a^2 = 0.2, \sigma_g^2 = 0.2$ and $\mu_a = 5/\sqrt{T}$, $\mu_g = 0$. Design 3: $\sigma_a^2 = 0.2, \sigma_g^2 = 0.2$ and $\mu_a = \mu_g = 0$.

model selection (PIV-LS), symmetric bootstrap with model selection (SYM-LS), and the conservative bootstrap (CONS-LS).

We first consider the "difference-in-differences" scenario, $x_{it} = \mathbb{1}\{t \geq T_i\}$. Table 5 reports rejection rates for a test of $H_0 : \beta_1 = 0$ at the 5 percent significance level under the null hypothesis $H_0 : \beta = 0$ (left four coliumns) and under a local alternative at the $\sqrt{N}$ rate, $H_1 : \beta = 0.1/\sqrt{N}$. As in the case of the sample mean, procedures based on LS under the drifting sequences scenario in Design 2 do not show any signs of improvement as sample size grows. Furthermore, Gaussian inference should not be expected to work in the degenerate case, and the results do indeed show that the size of the tests based on Gaussian asymptotics remains above the nominal level even for large samples. Under the local alternative, we find that all procedures have non-trivial power under each of the three designs and all except for the conservative bootstrap are consistent in the degenerate and near-degenerate designs 2 and 3.

Finally, we compare different simulation designs for the distribution of x_{it}, where for the first design $x_{it} := \mathbb{1}\{t \geq T_i\}$ is generated to mimic a "difference-in-differences" design, a

		$H_0 : \beta = 0$			
N	T	GAU	PIV	SYM	CONS
		Difference in Differences			
10	10	10.260	5.430	5.560	0.190
20	20	7.870	5.220	5.260	0.020
50	50	6.880	5.440	5.470	0.000
100	100	5.940	4.980	4.900	0.000
		i.i.d. regressors			
10	10	9.840	5.570	5.400	0.940
20	20	6.720	4.950	5.040	0.850
50	50	5.500	4.970	4.960	0.440
100	100	5.160	5.070	5.060	0.580
		Dyadic attributes			
10	10	11.340	5.430	5.300	4.040
20	20	8.110	5.010	4.850	0.480
50	50	7.930	5.560	5.500	1.250
100	100	6.950	4.940	5.040	0.240

TABLE 6. False rejection rates for two-sided tests of the null $\beta_1 = 0$ at a nominal level of 5 percent for different designs for drawing x_{it}. In all designs $\sigma_a^2 = 0.2, \sigma_g^2 = 0.2$ and $\mu_a = \mu_g = 0$.

second design looks at a regressor x_{it} that is i.i.d. across rows and columns. The third design considers regressors that are derived from attributes pertaining to the unit representing a row or a column of the array, where we choose $x_{it} := (1 - 2\mathbb{1}\{U_i \geq 0.5\})(1 - 2\mathbb{1}\{V_t \geq 0.5\})$ where $U_i, V_t \overset{iid}{\sim} U[0,1]$ ("dyadic regressors"). Generated regressors of this kind are common in matched, dyadic, or network data, where e.g. the distance between a country pair (with geographic coordinates as country-level attributes) explains trade flows in a gravity model, or an indicator whether a group of individuals share a discrete homophilous attribute (such as gender, age, or race) shifts the probability of friendship links or a clique among those individuals.

We only consider the degenerate design for u_{it} with $\mu_a = \mu_g = 0$ for this analysis. The theory predicts that the distribution of the LS coefficients should be Gaussian only under the second design with i.i.d. regressors, whereas for the other two designs we should see a non-Gaussian limiting distribution. Table 6 reports null rejection rates for a t-test of $H_0 : \beta_1 = 0$ at the nominal 5 percent significance level. The results by and large confirm the theoretical predictions, where for the second design all three procedures achieve coverage near the nominal level, whereas in the other two designs the bootstrap remains consistent, but Gaussian tests over-reject even for large samples with $N = T = 100$.

7. Conclusion

There has been great applied interest in robust inference that allows for multi-way dependence. In this paper, we provide a theoretical basis for that type of dependence, where we focus on an interpretation of the problem in which rows and columns correspond to units that are drawn independently from their respective distributions. We find that the asymptotic distribution of a sample average for an array of random variables that exhibits multi-way cluster dependence is not necessarily Gaussian, but may be nonstandard. Furthermore, there exists no uniformly adaptive procedure for estimating that asymptotic distribution.

One important practical limitation of our results is that our theory covers scenarios in which the fundamental units forming the multi-way array are drawn independently. While the negative results continue to apply for any more general setting that nests separate or joint exchangeability as a special case, the inference approach and the theory justifying it are not valid if those units are dependent. Most importantly, we are not aware of a suitable generalization of the Aldous-Hoover representation (2.1) that would allow for serial dependence among rows or columns of the array. Potential adaptations of the bootstrap procedure in this paper to accommodate scenarios of that type are beyond the scope of this paper and are left for future research.

References

Aldous, D. (1981): "Representations for Partially Exchangeable Arrays," *Journal of Multivariate Analysis*, 11, 581–598.

Andrews, D. (2000): "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68(2), 399–405.

——— (2001): "Testing when a Parameter is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69(3), 683–734.

——— (2002): "Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators," *Econometrica*, 70(1), 119–162.

Andrews, D., and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721–762.

——— (2010): "Asymptotic Size and a Problem with Subsampling and with the m out of n Bootstrap," *Econometric Theory*, 26, 426–468.

Arcones, M., and E. Giné (1992): "On the Bootstrap of U and V Statistics," *Annals of Statistics*, 20(2), 655–674.

Aronow, P., C. Samii, and V. Assenova (2015): "Cluster-Robust Variance Estimation for Dyadic Data," *Political Analysis*, 23(4), 564–577.

Bhattacharya, S., and P. Bickel (2015): "Subsampling Bootstrap of Count Features of Networks," *The Annals of Statistics*, 43(6), 2384–2411.

Bickel, P., and A. Chen (2009): "A Nonparametric View of Network Models and Newman-Girvan and other Modularities," *Proceedings of the National Academy on Sciences*, 106(50), 21068–21073.

Bickel, P., A. Chen, and E. Levina (2011): "The Method of Moments and Degree Distributions for Network Models," *Annals of Statistics*, 39(5), 2280–2301.

Bretagnolle, J. (1983): "Lois limites du bootstrap de certaines fonctionnelles," *Ann. Inst. H. Poincaré. Sec. B (N.S.)*, 3, 281–296.

Cameron, C., J. Gelbach, and D. Miller (2011): "Robust Inference With Multiway Clustering," *Journal of Business & Economic Statistics*, 29(2), 238–249.

Cameron, C., and D. Miller (2014): "Robust Inference for Dyadic Data," working paper, UC Davis and Cornell.

Carrasco, M., J. Florens, and E. Renault (2007): "Ill-Posed Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," in Heckman and Leamer (eds.): Handbook of Econometrics, Vol VI B Chapter 77.

Cattaneo, M., R. Crump, and M. Jansson (2014): "Bootstrapping Density-Weighted Average Derivatives," *Econometric Theory*, 30(1), 176–200.

Crane, H., and H. Towsner (2018): "Relatively Exchangeable Structures," *The Journal of Symbolic Logic*, 83(2), 416–442.

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2018): "Asymptotic Results under Multiway Clustering," working paper, ENSAE.

de Jong, P. (1990): "A Central Limit Theorem for Generalized Multilinear Forms," *Journal of Multivariate Analysis*, 34, 275–289.

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7(1), 1–26.

Graham, B. (2020): "Sparse Network Asymptotics for Logistic Regression," working paper, UC Berkeley.

Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hall, P., and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33(6), 2904–2929.

Hoover, D. (1979): "Relations on Probability Spaces and Arrays of Random Variables," working paper, Institute for Advanced Study, Princeton.

Horowitz, J. (2000): "The Bootstrap," Handbook of Econometrics, Vol V Chapter 52.

Kallenberg, O. (2005): *Probabilistic Symmetries and Invariance Principles*. Springer.

Kosorok, M. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer Series in Statistics. Springer, New York.

Leeb, H., and B. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Leung, M. (2016): "A Weak Law for Moments of Pairwise-Stable Networks," working paper, USC.

Liu, R. (1988): "Bootstrap Procedures Under Some Non-i.i.d. Models," *Annals of Statistics*, 16(4), 1696–1708.

Lovasz, L. (2012): "Large Networks and Graph Limits," in *AMS Colloquium Publications*, vol. 60. American Mathematical Society, Providence, RI.

MacKinnon, J., M. Nielsen, and M. Webb (2017): "Bootstrap and Asymptotic Inference with Multiway Clustering," working paper, Queen's University.

Mammen, E. (1992): *When does the Bootstrap Work: Asymptotic Results and Simulations*, vol. 77 of *Lecture Notes in Statistics*. Springer, Berlin.

McCullagh, P. (2000): "Resampling of Exchangeable Arrays," *Bernoulli*, pp. 285–301.

McLeish, D. (1974): "Dependent Central Limit Theorems and Invariance Principles," *Annals of Probability*, 2(4), 620–628.

Menzel, K. (2015): "Strategic Network Formation with Many Agents," working paper, New York University.

——— (2021): "Supplemental Material for 'Bootstrap with Cluster-Dependence in Two or More Dimensions'," .

Moulton, B. (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 72(2), 334–338.

Owen, A. (2007): "The Pigeonhole Bootstrap," *The Annals of Applied Statistics*, 1(2), 386–411.

Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley & Sons, New York.

Tabord-Meehan, M. (2019): "Inference with Dyadic Data: Asymptotic Behavior of the Dyadic-Robust t-Statistic," *Journal of Business and Economic Statistics*, 37(4), 671–680.

van der Vaart, A. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wu, C. (1986): "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14(4), 1261–1295.

Appendix A. Proofs

Proof of Theorem 4.1. Recall that the projection in (2.2) was given in terms of the variables

$$e_{it} = Y_{it} - \mathbb{E}[Y_{it}|\alpha_i, \gamma_t], \quad a_i = \mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}], \quad g_t = \mathbb{E}[Y_{it}|\gamma_t] - \mathbb{E}[Y_{it}]$$

and

$$v_{it} = \mathbb{E}[Y_{it}|\alpha_i, \gamma_t] - \mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}|\gamma_t] + \mathbb{E}[Y_{it}] = \sum_{k=1}^{\infty} c_k \psi_k(\gamma_t)\phi_k(\alpha_i)$$

where we rewrite v_{it} in terms of the low-rank representation in (2.3). Also let

$$\hat{Z}^a_N := \frac{r_{NT}}{N}\sum_{i=1}^{N} a_i, \quad \hat{Z}^g_T := \frac{r_{NT}}{T}\sum_{t=1}^{T} g_t, \quad \text{and } \hat{Z}^e_{NT} := \frac{r_{NT}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}$$

and

$$\hat{Z}^\phi_{Nk} := \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_k(\alpha_i), \qquad \hat{Z}^\psi_{Tk} := \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi_k(\gamma_t)$$

for $k = 1, 2, \ldots$. By independence of α_i and γ_t, $\hat{Z}^a_N$ and $\hat{Z}^g_T$ are uncorrelated. Since α_i and γ_t are independent, $\hat{Z}^\phi_{Nk}$ and $\hat{Z}^\psi_{Tk'}$ are uncorrelated for any pair k, k'. Also by orthogonality of the basis functions, $\hat{Z}^\phi_{Nk}$ and $\hat{Z}^\phi_{Nk'}$ ($\hat{Z}^\psi_{Tk}$ and $\hat{Z}^\psi_{Tk'}$, respectively) are uncorrelated for any $k \neq k'$. Finally by mean-independence of e_{it} and α_i, γ_t, the pairwise covariance between $\hat{Z}^e_{NT}$ and each component of $\hat{Z}^\phi_{Nk}, \hat{Z}^\psi_{Tk}, \hat{Z}^a_N, \hat{Z}^g_T$ are zero.

Central Limit Theorem: We next establish a central limit theorem for the stacked sample moments

$$\hat{Z}_{NT,K} := \left(\hat{Z}^e_{NT}, \hat{Z}^a_N, \hat{Z}^g_T, \hat{Z}^\phi_{N1}, \hat{Z}^\psi_{T1}, \ldots, \hat{Z}^\phi_{NK}, \hat{Z}^\psi_{TK}\right).$$

If a component of $\mathbf{q}_{NT}$ converges to zero, the corresponding component of $\hat{Z}_{NT,K}$ converges in probability to zero. We can therefore w.l.o.g. focus on the case in which the limit for each component $\mathbf{q}_{NT}$ is strictly positive. We then consider the process

$$\begin{aligned}\hat{W}_{NT,K} &:= (q^{-1/2}_{NT,e}\hat{Z}^e_{NT}, q^{-1/2}_{a,NT}\hat{Z}^a_N, q^{-1/2}_{g,NT}\hat{Z}^g_T, \hat{Z}^\phi_{N1}, \hat{Z}^\psi_{T1}, \ldots, \hat{Z}^\phi_{NK}, \hat{Z}^\psi_{TK})' \\ &=: (\hat{W}^e_{NT}, \hat{W}^a_N, \hat{W}^g_T, \hat{W}^\phi_{N1}, \hat{W}^\psi_{T1}, \ldots, \hat{W}^\phi_{NK}, \hat{W}^\psi_{TK})'\end{aligned}$$

To apply a martingale CLT, we choose the filtration

$$\mathcal{F}_{NT,s} := \sigma\left(\{\alpha_1, \ldots, \alpha_{\nu(s)}, \gamma_1, \ldots, \gamma_{\tau(s)}, \varepsilon_{11}, \ldots, \varepsilon_{\nu(s)T}\}\right)$$

Here we assume w.l.o.g. that $T \leq N$ and let $\tau(s) = s$ and $\nu(s) := \lceil \frac{Ns}{T} \rceil$, where $\lceil a \rceil$ denotes the largest integer smaller than or equal to a. While both N and T grow to infinity, we do not need to constrain the relative rates at which N, T grow under the asymptotic experiment.

We then seek to apply the CLT in Theorem 2.3 in McLeish (1974) to the martingale

$$\hat{M}_{NT,s} := \mathbb{E}\left[\hat{W}_{NT,K}|\mathcal{F}_{NT,s}\right]$$

with increments

$$\hat{X}_{NT,s} := \mathbb{E}\left[\hat{W}_{NT,K}|\mathcal{F}_{NT,s}\right] - \mathbb{E}\left[\hat{W}_{NT,K}|\mathcal{F}_{NT,s-1}\right]$$

Note that this CLT for martingale difference arrays allows for a triangular arrays where the row-wise distributions for the increments $\hat{X}_{NT,s}$ may change as N, T increase, and does not constrain the filtrations $\mathcal{F}_{NT,s}$ across N, T.

We next characterize the components of this martingale difference array $\hat{X}_{NT,s} = (\hat{X}^e_{NT,s}, \hat{X}^a_{N,s}, \hat{X}^g_{T,s}, \ldots, \hat{X}^\phi_{NK,s}, \hat{X}^\psi_{TK,s})$. Since $a_i = \mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}]$ is α_i-measurable and

$\alpha_1, \ldots, \alpha_N$ are independent, we have

$$\hat{X}^a_{N,s} = \frac{r_{NT}}{N\sqrt{q_{a,NT}}} \sum_{i=\nu(s-1)+1}^{\nu(s)} a_i = \frac{1}{\sqrt{N}} \sum_{i=\nu(s-1)+1}^{\nu(s)} \frac{a_i}{\sigma_a}$$

Similarly,

$$\begin{aligned}
\hat{X}^g_{T,s} &= \frac{1}{\sqrt{T}} \sum_{t=\tau(s-1)+1}^{\tau(s)} \frac{g_t}{\sigma_g} \\
\hat{X}^\phi_{NK,s} &= \frac{1}{\sqrt{N}} \sum_{i=\nu(s-1)+1}^{\nu(s)} \phi_k(\alpha_i) \\
\hat{X}^\psi_{TK,s} &= \frac{1}{\sqrt{T}} \sum_{t=\tau(s-1)+1}^{\tau(s)} \psi_k(\gamma_t)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\hat{X}^e_{NT,s} &= \frac{1}{\sqrt{NT}\sigma_e} \sum_{i=\nu(s-1)+1}^{\nu(s)} \left\{ \sum_{t=1}^{\tau(s)} e_{it} + \sum_{t=\tau(s)+1}^{T} \mathbb{E}[e_{it}|\alpha_i, \varepsilon_{it}] \right\} \\
&\quad + \frac{1}{\sqrt{NT}\sigma_e} \sum_{i=1}^{\nu(s-1)} \sum_{t=\tau(s-1)+1}^{\tau(s)} (e_{it} - \mathbb{E}[e_{it}|\alpha_i, \varepsilon_{it}])
\end{aligned}$$

Since by Assumption 2.1 the first four moments of each component of $\hat{X}_{NT,s}$ are bounded, it follows that

$$\sum_{s=1}^{T} (a'\hat{X}_{NT,s})^2 \xrightarrow{p} \sum_{s=1}^{T} a'\mathrm{Var}(\hat{X}_{NT,s})a = a'a$$

for any $a \in \mathbb{R}^{K+3}$, and furthermore we have the Lyapunov condition

$$\lim_{N,T} \frac{1}{\|a\|_2^3} \sum_{s=1}^{T} \mathbb{E}[(a'\hat{X}_{NT,s})^3] \to 0$$

which implies Condition (b) in Theorem 2.3 in McLeish (1974). It therefore follows from that theorem and the Cramér-Wold device that $r_{NT}\hat{W}_{NT}$ is asymptotically Gaussian, and therefore

$$\hat{Z}_{NT,K} \xrightarrow{d} N(0, Q)$$

where Q is a $(2K+3) \times (2K+3)$ matrix whose first three diagonal entries are q_e, q_a, and q_g, and the remaining $2K$ diagonal entries are equal to 1. For $k = 1, 2, \ldots$ the entries of Q corresponding to covariances between a_i and $\phi_k(\alpha_i)$ equal q_{ak}, and the covariances between g_t and $\psi_k(\gamma_t)$ are equal to q_{gk}. All other off-diagonal entries of Q are zero.

Truncating the expansion (2.3) at $K < \infty$, we define

$$r_{NT}\left(\bar{Y}_{NT,K} - \mathbb{E}[Y_{it}]\right) = \hat{Z}^a_N + \hat{Z}^g_T + \hat{Z}^e_{NT} + \varrho_{NT} \sum_{k=1}^{K} c_k \hat{Z}^\phi_{Nk} \hat{Z}^\psi_{Tk}$$

From the previous steps it then follows that

$$r_{NT}(\bar{Y}_{NT,K} - \mathbb{E}[Y_{it,K}]) \xrightarrow{d} \sqrt{q_a}Z_a + \sqrt{q_g}Z_g + \sqrt{q_e}Z_e + \varrho V_K$$

along each converging sequence, where

$$V_K := \sum_{k=1}^{K} c_k Z_k^{\psi} Z_k^{\phi}$$

with the coefficients c_k potentially variying along the limiting sequence, and $Z_e, Z_1^{\phi}, Z_1^{\psi}, \ldots, Z_K^{\phi}, Z_K^{\psi}$ are i.i.d. standard normal random variables, and Z^a, Z^g are standard normal random variables with $\text{Cov}(Z^a, Z_k^{\phi}) = q_{ak}/\sqrt{q_a}$, $\text{Cov}(Z^g, Z_k^{\psi}) = q_{gk}/\sqrt{q_g}$, $\text{Cov}(Z^a, Z^g) = \text{Cov}(Z^a, Z_k^{\psi}) = \text{Cov}(Z^g, Z_k^{\phi}) = 0$ for all $k = 1, 2, \ldots$.

Truncation Error: Moreover, we can show that the approximation error with respect to the distribution of $r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])$ from the truncation at $K < \infty$ can be made arbitrarily small by choosing K sufficiently large. Specifically, if we consider the truncation remainder

$$R_{K,NT} := \sum_{k=K+1}^{\infty} c_k \hat{Z}_{Nk}^{\phi} \hat{Z}_{Tk}^{\psi}$$

By Assumption 2.1 there exists $B_v < \infty$ such that $\sigma_v^2 \leq B_v$. This implies that

$$\begin{aligned}
\sum_{k=1}^{\infty} c_k^2 &= \sum_{k=1}^{\infty} c_k^2 \mathbb{E}[\phi_k(\alpha_i)]^2] \mathbb{E}[\psi_k(\gamma_t)^2] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{\infty} c_k \phi_k(\alpha_i) \psi_k(\gamma_t) \right)^2 \right] \\
&= \mathbb{E}[v_{it}^2] \leq B_v
\end{aligned}$$

where the first and second step use orthonormality of the basis functions, and independence of α_i, γ_t.
For any $\delta > 0$ we can therefore choose $K < \infty$ such that

$$\sum_{k=K+1}^{\infty} c_k^2 \leq \delta^2$$

It then follows that for the truncation remainder

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{k=K+1}^{\infty} c_k \hat{Z}_{Nk}^{\phi} \hat{Z}_{Tk}^{\psi} \right)^2 \right] &= \sum_{k=K+1}^{\infty} \mathbb{E}\left[\left(c_k \hat{Z}_{Nk}^{\phi} \hat{Z}_{Tk}^{\psi} \right)^2 \right] \\
&= \sum_{k=K+1}^{\infty} c_k^2 \mathbb{E}[(\hat{Z}_{Nk}^{\phi})^2] \mathbb{E}[(\hat{Z}_{Tk}^{\psi})^2] \\
&= \sum_{k=K+1}^{\infty} c_k^2 \leq \delta^2
\end{aligned}$$

where the first step uses orthogonality of the basis functions and independence of α_i, γ_t, the second step uses independence of α_i, γ_t, and the third step follows from the normalization of the second moments of the basis functions.
For any $\eta > 0$ we can therefore use Chebyshev's inequality to bound the probability $\mathbb{P}(|R_{K,NT}| > \eta) \leq \frac{\delta_K^2}{\eta^2}$, where δ can be made arbitrarily small by choosing $K \equiv K(\delta)$ large enough. Since the limiting distribution is continuous as shown below, $\mathcal{L}$ can also be approximated arbitrarily well as K is chosen at a suitably large value. Note furthermore that the magnitude of the approximation error can be controlled uniformly under Assumption 2.2.

Continuity of Limit Distribution: To establish claims (a) and (b), it remains to show that the limit distribution is continuous: first notice that we can verify using the convolution formula that for any continuously distributed W_1 with p.d.f. f_1 and an arbitrary random variable W_2 that is independent of W_1, the sum $W_1 + W_2$ also follows a continuous distribution with p.d.f. $f_{W_1+W_2}(s) = \mathbb{E}[f_{W_1}(s - W_2)]$. It is therefore sufficient to show that we can write the limiting distribution is that of a random variable which is a sum of independently distributed components, at least one of which has a continuous distribution.
We can then turn to the limiting distribution in (4.2),

$$\mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho) := \left(\sqrt{q_e}Z^e + \sqrt{q_a}Z^a + \sqrt{q_g}Z^g\right) + \varrho \sum_{k=1}^{\infty} c_k Z_k^{\psi} Z_k^{\phi}$$

We have by construction that $q_a + q_g + q_v + q_e = 1$, so that at least one of q_a, q_g, q_v, q_e must be strictly positive, where Assumption 2.1 guarantees that $r_{NT} > 0$ so that these are in fact well-defined. If $q_e > 0$, then $\sqrt{q_e}Z^e$ is continuously distributed, where Z^e is independent of Z^a, Z^g, and V, so that the conclusion immediately follows. If $q_v + q_e = 0$, the limiting distribution simplifies to $\sqrt{q_a}Z^a + \sqrt{q_g}Z^g$ with $q_a + q_g = 1$ and Z^a and Z^g both continuously distributed and independent of each other.
It therefore remains to consider the case $q_a + q_g > 0$ and $q_v > 0$: If $q_v > 0$, we must have that $\varrho|c_k| > 0$ for at least one value of $k \geq 1$. Since $Z^a, Z^g, Z_k^{\phi}, Z_k^{\psi}$ are jointly Gaussian, we can write $Z^a = \beta_1 Z_k^{\phi} + Z_{\perp,k}^{\phi}$ and $Z^g = \beta_2 Z_k^{\psi} + Z_{\perp,k}^{\psi}$ where $Z_k^{\phi}, Z_{\perp,k}^{\phi}, Z_k^{\psi}, Z_{k,\perp}^{\psi}$ are independent. We can then write

$$\sqrt{q_a}Z^a + \sqrt{q_g}Z^g + \varrho c_k Z_k^{\varphi} Z_k^{\psi} = \sqrt{q_a}Z_{k,\perp}^{\phi} + \sqrt{q_g}Z_{k,\perp}^{\psi} + (\beta_2\sqrt{q_g} + Z_k^{\phi})(\beta_1\sqrt{q}_a + \varrho c_k Z_k^{\psi}) - \beta_1\beta_2\sqrt{q_a q_g}$$

Since $\beta_2\sqrt{q_g} + Z_k^{\phi}$ and $\beta_1\sqrt{q}_a + \varrho c_k Z^{\psi}$ are independent and continuously distributed, then so is their product, using the formula for the density of the product of two independent random variables. Since these random variables are independent of $Z_{\perp,k}^{\phi}, Z_{k,\perp}^{\psi}, Z^e$, and $Z_{k'}^{\phi}, Z_{k'}^{\psi}$ for any $k' \neq k$, we obtain the desired conclusion □

In order to prove Theorem 4.2, we first establish rates of consistency for the estimators for the respective variances of the projection components, $\hat{\sigma}_a^2, \hat{\sigma}_g^2, \hat{\sigma}_w^2$ introduced in section 3.

Lemma A.1. *Suppose Assumption 2.1 holds. Then (a)*

$$\begin{aligned}
\hat{\sigma}_a^2 - \sigma_a^2 &= O_P\left(N^{-1/2}\left(\sigma_a + T^{-1/2}\sigma_e\right)^2 + T^{-1}\sigma_v^2\right) \\
\hat{\sigma}_g^2 - \sigma_g^2 &= O_P\left(T^{-1/2}\left(\sigma_g + N^{-1/2}\sigma_e\right)^2 + N^{-1}\sigma_v^2\right) \\
\hat{\sigma}_w^2 - \sigma_w^2 &= O_P\left((NT)^{-1/2}\sigma_e^2 + (N^{-1/2} + T^{-1/2})\sigma_v^2\right)
\end{aligned}$$

(b) There exist no estimators for σ_a^2, σ_g^2 and σ_w^2 that converge at rates faster than those given in (a). Specifically, σ_a^2 cannot be estimated at a rate faster than T^{-1} even when $\sigma_a^2 = 0$.

This lemma implies in particular that the estimators $\hat{\sigma}_a^2, \hat{\sigma}_g^2$ and $\hat{\sigma}_w^2$ are rate-optimal. Together with the continuous mapping theorem, this Lemma implies directly that $\hat{\lambda}_{NT}$ with model selection is pointwise consistent. $\hat{\lambda}_{NT}$ without model selection is uniformly consistent if $q_v = 0$, and inconsistent if $q_v > 0$.

Proof of Lemma A.1: For part (a), let $\hat{s}_a^2 := \frac{1}{N-1}\sum_{i=1}^{N}\hat{a}_i^2$, $\hat{s}_g^2 := \frac{1}{T-1}\sum_{t=1}^{T}\hat{g}_t^2$, and $\hat{s}_w^2 := \frac{1}{NT-N-T}\sum_{i=1}^{M}\sum_{t=1}^{T}\hat{w}_{it}^2$ be the empirical variances of the projection terms $\hat{a}_i, \hat{g}_t, \hat{w}_{it}$. We can also verify that $\frac{N}{N-1}\mathrm{Var}_{NT}(\hat{a}_i) = \sigma_a^2 + \sigma_w^2/T$, $\frac{T}{T-1}\mathrm{Var}_{NT}(\hat{g}_t) = \sigma_g^2 + \sigma_w^2/N$, and $\frac{NT}{NT-N-T}\mathrm{Var}_{NT}(\hat{w}_{it}) = \sigma_w^2$.

Consider first the term $\hat{s}_a^2$: We can write

$$\hat{a}_i^2 = \left(a_i + \frac{1}{T}\sum_{t=1}^{T} w_{it}\right)^2 = \left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^2 + 2\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)\frac{1}{T}\sum_{t=1}^{T} v_{it} + \left(\frac{1}{T}\sum_{t=1}^{T} v_{it}\right)^2$$

Hence we have that

$$\begin{aligned}
\hat{s}_a^2 - \left(\sigma_a^2 + \frac{1}{T}\sigma_w^2\right) &= \frac{1}{N}\sum_{i=1}^{N}\left\{\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^2 - \left(\sigma_a^2 + \frac{1}{T}\sigma_e^2\right)\right\} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)\frac{1}{T}\sum_{t=1}^{T} v_{it} + \frac{1}{N}\sum_{i=1}^{N}\left\{\left(\frac{1}{T}\sum_{t=1}^{T} v_{it}\right)^2 - \frac{1}{T}\sigma_v^2\right\} \\
&=: A_1 + A_2 + A_3
\end{aligned}$$

By independence of the rank variables $\alpha_i, \gamma_t, \varepsilon_{it}$ in the Aldous-Hoover representation and a martingale CLT, we have that

$$A_1 = O_P\left(N^{-1/2}\left(\sigma_a + T^{-1/2}\sigma_e\right)^2\right)$$

as $N \to \infty$. Next, consider the term A_3 where we can write

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T} v_{it}\right)^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{\infty} c_k \phi_{ik}\psi_{tk}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{k,k'} c_k c_{k'}\phi_{ik}\phi_{ik'}\left(\sum_{t=1}^{T}\psi_{tk}\right)\left(\sum_{t=1}^{T}\psi_{tk'}\right) \\
&= \sum_{k,k'} c_k c_{k'}\left(\frac{1}{N}\sum_{i=1}^{N}\phi_{ik}\phi_{ik'}\right)\left(\sum_{t=1}^{T}\psi_{tk}\right)\left(\sum_{t=1}^{T}\psi_{tk'}\right) \\
&=: \frac{1}{T}\sum_{k,k'}\left(\mathbb{1}\{k=k'\} + \frac{1}{\sqrt{N}}\hat{Z}^{\phi\phi}_{Nkk'}\right)\hat{Z}^{\psi}_{Tk}\hat{Z}^{\psi}_{Tk'} \qquad \text{(A.1)}
\end{aligned}$$

Here, $\hat{Z}^{\phi\phi}_{Nkk'} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\phi_{ik}\phi_{ik'} - \mathbb{E}[\phi_{ik}\phi_{ik'}])$, where $\mathbb{E}[\phi_{ik}\phi_{ik'}]$ equals 1 if $k = k'$ and zero otherwise. In particular, it follows that

$$A_3 = O_P\left(T^{-1}\sigma_v^2\right)$$

as N and T grow large. By similar calculations, we find that

$$\begin{aligned}
A_2 &= \sum_{k=1}^{\infty} c_k\left(\frac{1}{N}\sum_{i=1}^{N}\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)\phi_{ik}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\psi_{tk}\right) \\
&= O_P\left(N^{-1/2}(\sigma_a + T^{-1/2}\sigma_e)T^{-1/2}\sigma_v\right)
\end{aligned}$$

noting that by construction $\mathbb{E}\left[a_i\phi_{ik}\right] = 0$ for each $k = 1, 2, \ldots$. Aggregating the contributions of the individual terms A_1, A_2, A_3, we then obtain

$$\hat{s}_a^2 - \left(\sigma_a^2 + \frac{1}{T}\sigma_w^2\right) = O_P\left(N^{-1/2}\left(\sigma_a + T^{-1/2}\sigma_e\right)^2 + T^{-1}\sigma_v^2\right)$$

Similarly, we find that

$$\hat{s}_g^2 - \left(\sigma_g^2 + \frac{1}{N}\sigma_w^2\right) = O_P\left(T^{-1/2}\left(\sigma_g + N^{-1/2}\sigma_e\right) + N^{-1}\sigma_v^2\right)$$

Next, note that

$$\hat{\sigma}_w^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(v_{it}^2 + 2v_{it}e_{it} + e_{it}^2)$$

From calculations analogous to (A.1), we also find that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} v_{it}^2 = O_p\left(N^{-1/2} + T^{-1/2}\right)$$

Hence,

$$\hat{\sigma}_w^2 - \sigma_w^2 = O_P\left((NT)^{-1/2}\sigma_e^2 + (T^{-1/2} + N^{-1/2})\sigma_v^2\right)$$

The rates asserted in the Lemma then follow directly from the definitions of the variance estimators $\hat{\sigma}_a^2 := \max\left\{0, \hat{s}_a^2 - \frac{1}{T}\hat{s}_w^2\right\}$, $\hat{\sigma}_g^2 := \max\left\{0, \hat{s}_g^2 - \frac{1}{N}\hat{\sigma}_w^2\right\}$.
For a proof of part (b), note first that it is sufficient to find a specific family of distributions under which that rate cannot be improved upon. Specifically, consider the model

$$Y_{it} = \alpha_i\gamma_t + \varepsilon_{it}$$

where $\alpha_i, \gamma_t, \varepsilon_{it}$ are independent, $\alpha_i \sim N(\mu_a, 1)$, $\gamma_t \sim N(\mu_g, 1)$ for some $\mu_a, \mu_g \geq 0$, and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$.
To establish the rate for the contribution of terms depending on σ_v^2 to that bound, consider the case $\sigma_\varepsilon^2 = 0$ and $\mu_a = 0$. For this model, $a_i := \mathbb{E}[Y_{it}|\alpha_i] = \alpha_i\mu_g$ and $v_{it} = \alpha_i(\gamma_t - \mu_g)$, so that $\sigma_a^2 = \mu_g^2$ and $\sigma_v^2 = 1$. Clearly, μ_g cannot be estimated from the original data at a better rate than from directly observing $(\alpha_i)_{i=1}^N$ and $(\gamma_t)_{t=1}^T$. Furthermore, since $\gamma_1, \ldots, \gamma_T$ are i.i.d., there exists no consistent test for the problem $H_0 : \mu_g = 0$ against $H_1 : \mu_g = T^{-1/2}m_g$ for any arbitrary $m_g > 0$. Since under H_0, $\sigma_a^2 = 0$, whereas under H_1, $\sigma_a^2 = T^{-1}m_g^2$, there can be no estimator for σ_a^2 that is consistent at a rate faster than $T^{-1}\sigma_v^2$.
The respective contributions of terms depending on σ_a^2, σ_g^2 and σ_e^2 to the rate bound follow immediately from standard arguments for the case of i.i.d. data, which can similarly be cast in terms of pairwise testing problems between drifting DGP sequences. Finally, consistent estimation of σ_a^2 under all DGPs permitted by our framework requires simultaneously solving these pairwise testing problems that gave us the respective rate contributions depending on $\sigma_a^2, \sigma_g^2, \sigma_e^2$ and σ_v^2. Hence an upper bound is given by the slowest of these rates, which establishes the claim for the rate of consistent estimation of σ_a^2. The respective upper bounds on the rate for estimating σ_g^2 and σ_w^2 follow from analogous arguments □

From the previous result, it follows that the variance estimator $\hat{S}_{NT,sel}^2$ is pointwise consistent:

Corollary A.1. ***(Consistency of** $\hat{S}_{NT,sel}^2$**)*** *Suppose that Assumption 2.1 holds. Then for the variance estimator with model selection*

$$\left|\frac{r_{NT}^2\hat{S}_{NT,sel}^2}{NT} - 1\right| \xrightarrow{p} 0$$

pointwise for any values of $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$*. For the variance estimator without model selection convergence is uniform if* $q_v = 0$*, but the estimator is inconsistent for* $q_v > 0$*.*

Noting that $\mathrm{Var}\left(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])\right) = 1$, this corollary is an immediate consequence of the convergence rates in Lemma A.1. In particular, if $\sigma_a^2 = 0$, Lemma A.1 (a) implies that $T\hat{\sigma}_a^2 = O_p(1)$, so that for any divergent sequence $\kappa_a \to \infty$, $T\hat{\sigma}_a^2 < \kappa_a$ with probability approaching 1, in which case $\hat{D}_a(\kappa_a) = 0$. On the other hand, if $\sigma_a^2 > 0$, then $\hat{\sigma}_a^2 = \sigma_a^2 + O_p(N^{-1/2})$. Hence for the estimator with model selection, $\hat{D}_a(\kappa_a) = 1$ for any sequence κ_a such that $\kappa_a/T \to 0$. By the same reasoning, the selector $\hat{D}_g(\kappa_g) = 0$ with

probability approaching 1 if $\sigma_g^2 = 0$, and $\hat{D}_g(\kappa_g) = 1$ with probability approaching 1 if $\sigma_g^2 > 0$. The conclusions regarding estimation without model convergence are immediate given Lemma A.1.

Proof of Proposition 4.1: Consider again the model $Y_{it} = \alpha_i \gamma_t$, with $\alpha_i \sim N(0,1)$, $\gamma_t \sim N(\mu_g, 1)$, where we let $\mu_g := T^{-1/2} m_g$. Note that this model satisfies Assumptions 2.1 and 2.2, so that this counterexample is not ruled out by the conditions for the main results in this paper. Then for any finite N, T,

$$\sqrt{NT}\bar{Y}_{NT} \overset{d}{=} (m_g + Z_g) Z_a, \quad \text{where } Z_a, Z_g \overset{iid}{\sim} N(0,1)$$

In particular, taking limits along the drifting sequence for μ_g, the right-hand side expression is also the asymptotic distribution of the sample mean.
By inspection, the c.d.f. at certain quantiles of the asymptotic distribution for $m_g = 0$ is different from that for any $m_g \neq 0$. Furthermore, by the same argument as in the proof for part (b) of Lemma A.1, for any $\tilde{m}_g \neq 0$ there is no consistent test between the alternatives $m_g = \tilde{m}_g$ from $m_g = 0$. However, a uniformly consistent estimator for the asymptotic distribution would provide such a consistent test, a contradiction □

Bootstrap Distribution. In order to obtain the limit of the bootstrap distribution, we introduce some additional notation: for any array (ξ_{it}), we let the operator $\mathbb{E}^*_{NT}[\xi_{it}|\alpha_i] := \frac{1}{T}\sum_{t=1}^T \xi_{it}$ denote the row-wise average for the T observations in the ith row, $\mathbb{E}^*_{NT}[\xi_{it}|\gamma_t] := \frac{1}{N}\sum_{i=1}^N \xi_{it}$ the column-wise average for the N observations in the tth column, and $\mathbb{E}^*_{NT}[\xi_{it}] := \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \xi_{it}$ the pooled average over all NT observations. We also decompose $\hat{w}_{it} = \hat{v}_{it} + \hat{e}_{it}$ with

$$\begin{aligned} \hat{e}_{it} &= e_{it} - \mathbb{E}^*_{NT}[e_{it}|\alpha_i] - \mathbb{E}^*_{NT}[e_{it}|\gamma_t] + \mathbb{E}^*_{NT}[e_{it}] \\ \hat{v}_{it} &= v(\alpha_i, \gamma_t) = \sum_{k=1}^{\infty} c_k \psi_k(\gamma_t)\phi_k(\alpha_i) \end{aligned}$$

Given that notation we define the localized second moments of the projection terms,

$$\begin{aligned} q^*_{a,NT} &:= r_{NT}^2 N^{-1} \mathbb{E}^*_{NT}[\hat{a}_i^2] = r_{NT}^2 \frac{1}{N^2}\sum_{i=1}^N \hat{a}_i^2, & q^*_{g,NT} &:= r_{NT}^2 T^{-1} \mathbb{E}^*_{NT}[\hat{g}_t^2] = r_{NT} \frac{1}{T^2}\sum_{t=1}^T \hat{g}_t^2 \\ q^*_{e,NT} &:= r_{NT}^2 (NT)^{-1} \mathbb{E}^*_{NT}[\hat{e}_{it}^2], & q^*_{v,NT} &:= r_{NT}^2 (NT)^{-1} \mathbb{E}^*_{NT}[\hat{v}_{it}^2] \\ q^*_{ak,NT} &:= r_{NT}^2 N^{-1} \mathbb{E}^*_{NT}[\hat{a}_i \phi_k(\alpha_i)], & q^*_{gk,NT} &:= r_{NT}^2 T^{-1} \mathbb{E}^*_{NT}[\hat{g}_t \psi_k(\gamma_t)] \end{aligned}$$

for $k = 1, 2, \ldots$. We then also write

$$\mathbf{q^*_{NT}} := (q^*_{e,NT}, q^*_{a,NT}, q^*_{g,NT}, 0, 0, \ldots)$$

and $\mathbf{c}_{NT} := (c_{1,NT}, c_{2,NT}, \ldots)$, where we take the sequences $\mathbf{c}_{NT}$ and $\mathbf{q^*_{NT}}$ to be elements of ℓ^2.
We first consider convergence for a truncated version of the spectral representation for the sample mean in (2.3) at some fixed integer K, $0 < K < \infty$,

$$\begin{aligned} \bar{Y}^*_{NT,K} &:= \mathbb{E}^*_{NT}[Y_{it}] + \sqrt{\lambda_a}\frac{1}{N}\sum_{i=1}^N \hat{a}_{j(i)} + \sqrt{\lambda_g}\frac{1}{T}\sum_{t=1}^T \hat{g}_{s(t)} + \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \omega_{1i}\omega_{2t}\hat{e}_{j(i)s(t)} \\ &\quad + \frac{1}{\sqrt{NT}}\sum_{k=1}^K c_k \left[\frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_{1i}\left(\phi_k(\alpha_{j(i)}) - \mathbb{E}^*_{NT}[\phi_k(\alpha_{j(i)})]\right)\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T \omega_{2t}\left(\psi_k(\gamma_{s(t)}) - \mathbb{E}^*_{NT}[\psi_k(\gamma_{s(t)})]\right)\right] \end{aligned} \tag{A.2}$$

which is obtained by truncating the bootstrap analog of (2.3). This can be expressed in terms of the truncated bootstrap process

$$\hat{Z}^*_{NT,K} := (\hat{Z}^{e,*}_{NT}, \hat{Z}^{a,*}_N, \hat{Z}^{g,*}_T \hat{Z}^{\phi,*}_{N1}, \hat{Z}^{\psi,*}_{T1}, \dots, \hat{Z}^{\phi,*}_{NK}, \hat{Z}^{\psi,*}_{TK})'$$

where we let

$$\hat{Z}^{a,*}_{NT} := \frac{r_{NT}}{N} \sum_{i=1}^{N} \hat{a}_{j(i)}, \quad \hat{Z}^{g,*}_{NT} := \frac{r_{NT}}{T} \sum_{t=1}^{T} \hat{g}_{s(t)}, \quad \hat{Z}^{e,*}_{NT} := \frac{r_{NT}}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_{1i}\omega_{2t}\hat{e}_{j(i)s(t)}$$

and

$$\begin{aligned} \hat{Z}^{\phi,*}_{Nk} &:= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \omega_{1i} \left(\phi_k(\alpha_{j(i)}) - \mathbb{E}^*_{NT}[\phi_k(\alpha_{j(i)})] \right) \\ \hat{Z}^{\psi,*}_{Tk} &:= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \omega_{2t} \left(\psi_k(\gamma_{s(t)}) - \mathbb{E}^*_{NT}[\psi_k(\gamma_{s(t)})] \right) \end{aligned}$$

for $k = 1, \dots, K$.

To characterize the asymptotic distribution of $\bar{Y}^*_{NT,K}$, we let $\tilde{\mathbf{c}}_{NT,K} \in \ell^2$ denote the truncated version of the vector $\mathbf{c}_{NT} = (c_{1,NT}, c_{2,NT}, \dots)$ of spectral coefficients in (2.3), where the first K components of $\tilde{\mathbf{c}}_{NT,K}$ coincide with the first K components of $\mathbf{c}_{NT}$, and all remaining coordinates are set to zero. We also define the distribution

$$\mathcal{L}^*(\mathbf{c}, \mathbf{q}, \varrho, \boldsymbol{\lambda}) := \sqrt{\lambda_a q_a} Z^a + \sqrt{\lambda_g q_g} Z^g + \varrho \sum_{k=1}^{\infty} c_k Z^{\phi}_k Z^{\psi}_k + \sqrt{q_e} Z^e$$

where $\boldsymbol{\lambda} := (\lambda_a, \lambda_g)$, $Z^e, Z^{\phi}_1, Z^{\psi}_1, Z^{\phi}_2, Z^{\psi}_2, \dots$ are i.i.d. standard normal random variables, and Z^a, Z^g are random variables with a standard normal marginal distribution and covariances $\mathrm{Cov}(Z^a, Z^{\phi}_k) = q_{ak}/\sqrt{q_a}$ and $\mathrm{Cov}(Z^g, Z^{\psi}_k) = q_{gk}/\sqrt{q_g}$.

Lemma A.2. ***(Bootstrap CLT)*** *Consider the bootstrap with shrinkage parameters $\boldsymbol{\lambda}_{NT} = (\lambda_{a,NT}, \lambda_{g,NT})$ and suppose that Assumption 2.1 holds. Then for any fixed $K < \infty$ we have that*

$$\| \mathbb{P}^*_{NT}(r_{NT}(\bar{Y}^*_{NT,K} - \bar{Y}_{NT})) - \mathcal{L}^*(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}^*_{NT}, \varrho, \boldsymbol{\lambda}_{NT}) \|_{\infty} \xrightarrow{p} 0$$

Proof: First we consider the contribution of the term

$$\hat{Z}^{e,*}_{NT} := \frac{r_{NT}}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_{1i}\omega_{2t}\hat{e}_{it}$$

Notice that, holding $\hat{e}_{11}, \dots, \hat{e}_{NT}$ fixed, this sum is a bilinear form in $\omega_1 := (\omega_{11}, \dots, \omega_{1N})'$ and $\omega_2 := (\omega_{21}, \dots, \omega_{2T})'$. Theorem 1 in de Jong (1990) then gives sufficient conditions for asymptotic normality of $\hat{Z}^{e,*}_{NT}$ conditional on $\hat{e}_{11}, \dots, \hat{e}_{NT}$ and $\alpha_1, \dots, \alpha_N, \gamma_1, \dots, \gamma_T$. Specifically, we need to verify that the standardized fourth central moment, $\mathbb{E}[(\hat{Z}^{e,*}_{NT})^4] / \mathbb{E}[(\hat{Z}^{e,*}_{NT})^2]^2 \to 3$ almost surely, noting that $\mathbb{E}[\hat{Z}^{e,*}_{NT}] = 0$.

Multiplying out, the fourth power of this sum is given by

$$(r^{-1}_{NT}\hat{Z}^{e,*}_{NT})^4 = \frac{1}{(NT)^4} \sum_{i_1,\dots i_4=1}^{N} \sum_{t_1,\dots,t_4=1}^{T} \omega_{1i_1}\omega_{1i_2}\omega_{1i_3}\omega_{1i_4}\omega_{2t_1}\omega_{2t_2}\omega_{2t_3}\omega_{2t_4}\hat{e}_{i_1t_1}\hat{e}_{i_2t_2}\hat{e}_{i_3t_3}\hat{e}_{i_4t_4}$$

Since ω_{1i}, ω_{2t} are i.i.d. and mean zero, any terms in which one or more indices in $\{i_1, \dots, i_4, t_1, \dots, t_4\}$ appear as an odd number of multiples have zero expectations. When all indices appear as an even number of multiples, there are four possibilities, where for $i_1 \neq i_2, t_1 \neq t_2$ we have $\mathbb{E}[\omega^2_{1i_1}\omega^2_{1i_2}\omega^2_{2t_1}\omega^2_{2t_2}] = \mu_2^4$ (type

1), $\mathbb{E}[\omega_{1i_1}^4\omega_{2t_1}^2\omega_{2t_2}^2] = \mu_2^2\mu_4$ (type 2), $\mathbb{E}[\omega_{1i_1}^2\omega_{1i_2}^2\omega_{2t_1}^4] = \mu_2^2\mu_4$ (type 3), and $\mathbb{E}[\omega_{1i_1}^4\omega_{2t_1}^4] = \mu_2^4$ (type 4), and $\mu_2 := \mathbb{E}[\omega^2] = 1$ and $\mu_4 := \mathbb{E}[\omega^4] < \infty$.

Moreover, since $e_{i,t}$ and $e_{j,s}$ are mean-independent whenever $(i,t) \neq (j,s)$, we have that $\mathbb{E}[e_{i_1t_1}e_{i_2t_2}e_{i_3t_3}e_{i_4t_4}] = 0$ whenever at least one index pair in $\{(i_1,t_1),(i_2,t_2),(i_3,t_3),(i_4,t_4)\}$ appears exactly once. Hence the average of $\hat{e}_{i_1t_1}\hat{e}_{i_2t_2}\hat{e}_{i_3t_3}\hat{e}_{i_4t_4}$ over all tuplets in which, the index (i_s,t_s) appears exactly once converges to zero almost surely by a strong law of large numbers. Hence, the contribution of terms in which any index pair in $\{(i_1,t_1),(i_2,t_2),(i_3,t_3),(i_4,t_4)\}$ appears an odd number of times is asymptotically negligible.

There are $6N(N-1)T(T-1)$ terms of type 1 such that all index pairs appear an even number of times, $6NT(T-1)$ such terms of type 2, $6N(N-1)T$ terms of type 3, and NT terms of type 4. Since the expectations $\mathbb{E}[e_{i_1t_1}^2e_{i_2t_2}^2] =: \sigma_e^4$ and $\mathbb{E}[e_{i_1t_1}^4]$ are strictly positive and bounded by assumption regardless of the overlap between indices, the terms of types 2,3, and 4 are asymptotically negligible.

Hence, as $N,T \to \infty$,

$$\begin{aligned}\mathbb{E}[(r_{NT}^{-1}\hat{Z}_{NT}^{e,*})^4] &= \frac{6N(N-1)T(T-1)}{(NT)^4}\mathbb{E}[e_{i_1t_1}^2e_{i_2t_2}^2] + o(1)\\ &= \frac{3}{(NT)^2}\sigma_e^4 + o(1)\end{aligned}$$

where the remainder term $o(1)$ vanishes almost surely. From similar arguments, we can confirm that

$$\mathbb{E}[(r_{NT}^{-1}\hat{Z}_{NT}^{e,*})^2] = \frac{1}{NT}\sigma_e^2 + o(1)$$

so that

$$\frac{\mathbb{E}[(\hat{Z}_{NT}^{e,*})^4]}{\mathbb{E}[(\hat{Z}_{NT}^{e,*})^2]} = 3 + o(1)$$

as desired. Given the assumptions of this theorem, it then follows from Theorem 1 in de Jong (1990) that $\hat{Z}_{NT}^{e,*} \to N(0,q_e)$ conditional on $\hat{e}_{11},\ldots,\hat{e}_{NT}$ and $\alpha_1,\ldots,\alpha_N,\gamma_1,\ldots,\gamma_T$.

For the contribution of the remaining terms, note that a_i^*, g_t^* are i.i.d. draws from the empirical distribution of $\hat{a}_i, \hat{g}_t$, and ω_{1i},ω_{2t} are i.i.d. draws from the auxiliary distribution for the (Wild bootstrap) multiplier. In particular, if we let $\phi_{1ik}^* := \omega_{1i}(\phi_k(\alpha_{j(i)}) - \mathbb{E}_{NT}^*[\phi_k(\alpha_{j(i)})])$ and $\psi_{2tk}^* := \omega_{2t}(\psi_k(\gamma_{s(t)}) - \mathbb{E}_{NT}^*[\psi_k(\alpha_{s(t)})])$, then the random vectors $\zeta_i^* := (a_i^*, \phi_{1i1}^*,\ldots,\phi_{1iK}^*)'$ and $\xi_t^* := (g_t^*, \psi_{2t1}^*,\ldots,\psi_{2tK}^*)$ are i.i.d. conditional on the respective empirical distributions of $\hat{a}_i$ and $\hat{g}_t$.

By Assumption 2.1, the third conditional moments of $\hat{a}_i, \hat{g}_t$ and $(\omega_i\phi_k(\alpha_i), \omega_t\psi_k(\gamma_t))_{k\geq 1}$ given $(Y_{it} : i = 1,\ldots,N, t = 1,\ldots,T)$ are almost surely bounded, so that from the same argument as in the proof of Theorem 1 in Liu (1988), the Berry-Eséen theorem implies a CLT for the bootstrap processes $\frac{1}{\sqrt{N}}\sum_{i=1}^N \zeta_i^*$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_t^*$. Since these components are independent by construction, it follows that

$$\left\|\mathbb{P}_{NT}^*\left(\hat{Z}_{NT,K}^*\right) - N(0,Q_{NT,K}^*)\right\|_\infty = o_P(1)$$

conditional on (Y_{it}) almost surely. Here, $Q_{NT,K}^*$ is a $(2K+3)\times(2K+3)$ diagonal matrix whose first three diagonal entries are $q_{e,NT}^*$, $q_{a,NT}^*$, and $q_{g,NT}^*$, and the remaining $2K$ diagonal entries converge almost surely to 1. All other off-diagonal entries of $Q_{NT,K}^*$ converge almost surely to zero.

Finally, we can rewrite (A.2) and obtain

$$r_{NT}(\bar{Y}_{NT,K}^* - \bar{Y}_{NT}) := \hat{Z}_N^{a,*} + \hat{Z}_T^{g,*} + \hat{Z}_{NT}^{e,*} + \varrho_{NT}\sum_{k=1}^K c_k\hat{Z}_{Nk}^{\phi,*}\hat{Z}_{Tk}^{\psi,*}$$

Note that this is a Lipschitz transformation of the components of $\hat{Z}^*_{NT,K}$ with Lipschitz constant less than or equal to one. It then follows from the joint CLT and a continuous mapping theorem for the bootstrap (see e.g. Proposition 10.7 in Kosorok (2008)) that

$$\left\|\mathbb{P}^*_{NT}\left(\bar{Y}^*_{NT,K}\right) - \mathcal{L}^*(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}^*_{NT}, \varrho, \boldsymbol{\lambda}_{NT})\right\|_\infty = o_P(1)$$

establishing the claim □

Proof of Theorem 4.2. For bootstrap consistency it suffices to verify whether the limiting distributions of the sampling distribution $r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])$ and the limit of the bootstrap distribution $r_{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT})$ given the sample coincide. In what follows, we first consider the asymptotic distribution of the truncated representation of the bootstrapped mean $\bar{Y}^*_{NT,K}$ defined in (A.2) and let $\tilde{\mathbf{c}}_{NT,K} \in \ell^2$ denote the truncated version of the vector $\mathbf{c}_{NT} = (c_{1,NT}, c_{2,NT}, \dots)$ of spectral coefficients in (2.3), where the first K components of $\tilde{\mathbf{c}}_{NT,K}$ coincide with the first K components of $\mathbf{c}_{NT}$, and all remaining coordinates are set to zero.

For pointwise consistency of the bootstrap with model selection, note first that the local parameter with both $q_a + q_g > 0$ and $q_v > 0$ can only be achieved at drifting sequences, so that this case is irrelevant for point-wise convergence. In particular we can without loss of generality set $q_{ak} = q_{gk} = 0$ for each $k = 1, 2, \dots$. By Lemma A.1 (a), $\mathbf{q}^*_{NT,K} - \mathbf{q}_{NT,K} \xrightarrow{p} 0$ so that $\mathbf{q}^*_{NT,K} - \mathbf{q} \xrightarrow{p} 0$ along any converging sequence $\mathbf{q}_{NT,K} \to \mathbf{q}$, and $\hat{\lambda}_a, \hat{\lambda}_g$ are consistent for λ_a, λ_g whenever either $q_a + q_g = 0$ or $q_v = 0$, where convergence is pointwise.

Furthermore, the limit law $\mathcal{L}$ is continuous with respect to the parameters $(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}, \varrho, \lambda)$: Following the continuity argument in the proof of Theorem 4.1, the limit law $\mathcal{L}^*(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}, \varrho, \lambda)$ is a weighted sum of independent, continuously distributed random variables,

$$S^*_K := a_1 W_1 + a_2 W_2 + \cdots + a_{K+3} W_{K+3}$$

which depends on the parameters $(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}, \varrho, \lambda)$ only through the coefficients $a_1, \dots, a_{K+3}$, which are in turn square roots or products of those parameters. Furthermore, as $K \to \infty$, the variance of S^*_K converges to 1 so that $\sum_{k=1}^\infty a_k^2 = 1$. Therefore, at least one of the coefficients $a_1, \dots, a_{K+3}$ has to be strictly positive, w.l.o.g. $a_1 > 0$. The c.d.f. of S^*_K is then given by

$$F_S(z) = \int F_{W_1}\left(\frac{z - \sum_{k=2}^{K+3} a_k w_k}{a_1}\right) f_{W_2}(w_2) dw_2 \dots f_{W_{K+3}}(w_{K+3}) dw_{K+3}$$

where $F_{W_k}(\cdot), f_{W_k}(\cdot)$ denoting the c.d.f. and p.d.f. of W_k. Since W_1 is continuously distributed and $a_1 > 0$ by assumption, the c.d.f. $F_S(\cdot)$ varies continuously in $a_1, \dots, a_{K+3}$, so that $\mathcal{L}^*(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}, \varrho, \lambda)$ is indeed continuous in these parameters with respect to the KS metric.

Hence together with the continuous mapping theorem, Lemma A.2 implies that

$$\left\|\mathbb{P}^*_{NT}(r_{NT}(\bar{Y}^*_{NT,K} - \bar{Y}_{NT})) - \mathcal{L}^*(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}, \varrho, \boldsymbol{\lambda}_{NT})\right\|_\infty \xrightarrow{p} 0$$

We can then use an approximation arguments analogous to that in the proof of Theorem 4.1 to conclude that the distribution of the truncated version $\bar{Y}^*_{NT,K}$ of the bootstrap mean can be made to approximate arbitrarily closely to that of $\bar{Y}^*_{NT}$ by choosing K large enough, so that

$$\left\|\mathbb{P}^*_{NT}(r_{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT})) - \mathbb{P}^*_{NT}(r_{NT}(\bar{Y}^*_{NT,K} - \bar{Y}_{NT}))\right\|_\infty = o_P(1)$$

and

$$\left\|\mathcal{L}^*(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}, \varrho, \boldsymbol{\lambda}_{NT}) - \mathcal{L}^*(\mathbf{c}_{NT}, \mathbf{q}, \varrho, \boldsymbol{\lambda}_{NT})\right\|_\infty = o_P(1)$$

Hence pointwise convergence for the bootstrap with model selection follows from Theorem 4.1 and Lemma A.2 together with continuity of $\mathcal{L}^*(\tilde{\mathbf{c}}_{NT,K}, \mathbf{q}, \varrho, \boldsymbol{\lambda})$ in $\mathbf{q}$, and the triangle inequality. The analogous result for the pivotal bootstrap follows from Corollary A.1 together with the continuous mapping theorem.
For uniform consistency of the bootstrap without model selection, we first consider convergent drifting sequences $\mathbf{q}_{NT}, \mathbf{c}_{NT}$ with limits $\mathbf{q}$ and $\mathbf{c}$, respectively. We also let

$$\bar{\mathbf{q}}_{NT} := (q_{e,NT}, q_{a,NT} + q_{e,NT} + q_{v,NT}, q_{g,NT} + q_{e,NT} + q_{v,NT}, 0, 0, \dots).$$

Lemma A.1 (a) implies that $\mathbf{q}^*_{NT} - \bar{\mathbf{q}}_{NT}$ converges in probability to zero, and $\hat{\lambda}_a, \hat{\lambda}_g$ are consistent for λ_a and λ_g along such a sequence whenever $q_v = 0$. Convergence for the bootstrap without model selection along the convergent sequence $\mathbf{q}_{NT}$ then follows from the same arguments as for the pointwise case, noting that under Assumption 2.2 (b), the approximation error in (2.3) from truncation at $K < \infty$ can be controlled uniformly under drifting sequences for $\mathbf{c}_{NT}$.
The conservative bootstrap is identical to the bootstrap with model selection except in the event $\hat{D}_a(\kappa_a) = 0$ or $\hat{D}_g(\kappa_g) = 0$. For $\hat{D}_a(\kappa_a) = 0$ we have by inspection that $\sqrt{\frac{\hat{\lambda}_a}{N\kappa_a}} \sum_{i=1}^N a^*_{i,b} \xrightarrow{d} N(0,1)$, and for $\hat{D}_g(\kappa_g) = 0$, we have $\sqrt{\frac{\hat{\lambda}_g}{T\kappa_g}} \sum_{t=1}^T g^*_{t,b} \xrightarrow{d} N(0,1)$, whereas the other components of the bootstrap distribution coincide with their analogs for the bootstrap with model selection.
This establishes the claims of the Theorem under any convergent sequences $\mathbf{q}_{NT}, \mathbf{c}_{NT}$. To conclude the proof it remains to show that it is in fact sufficient for uniformity to consider convergent subsequences for which the appropriately normalized parameters converge to proper limits. Here we can adapt an argument from the proof of Theorem 1 in Andrews and Guggenberger (2010), noting that the limiting sequences for the truncated version spectral representation $\bar{Y}_{NT,K}$ and its bootstrap analog, $\bar{Y}^*_{NT,K}$ in the proofs of Theorem 4.1 and Lemma A.2 are both indexed by finite-dimensional subvectors of $\mathbf{c}$ and $\mathbf{q}$. Since $q_a + q_g + q_v + q_e = 1$, such a subvector of $\mathbf{q}$ can only take values in a compact set, and the norm $\|\tilde{\mathbf{c}}_{NT,K}\|^2 \leq \sum_{k=1}^K \bar{c}_k^2 < \infty$ by Assumption 2.2. Hence such a convergent subsequence for these subvectors can be extracted from $(\mathbf{q}_{NT}, \mathbf{c}_{NT})$ by the Bolzano-Weierstrass theorem, and the truncation error can then be made arbitrarily small by choosing K large enough. □

Proof of Proposition 4.2. We can establish the refinements of this bootstrap procedure by establishing separate Edgeworth expansions for the sampling and bootstrap distributions, using Theorems 2.2 and 5.1 of Hall (1992), and then showing that the first three cumulants of the bootstrap distribution converge almost surely to those of the sampling distribution.
Since by assumption $\sigma_a^2 + \sigma_g^2$ are bounded away from zero, the rate r_{NT} is no faster than $\min\{N^{-1/2}, T^{-1/2}\}$. We therefore first focus on the contribution of $\hat{Z}^{a,*}_N + \hat{Z}^{g,*}_T$, and then show that the contribution of the remaining terms is at most of the order $(NT)^{-1/2}$. Furthermore, according to our results in Lemma A.1 we have that $\lambda_a \xrightarrow{p} 1$ and $\lambda_g \xrightarrow{p} 1$ whenever $\sigma_a^2 \geq C > 0$, and $\sigma_g^2 \geq C > 0$, respectively. In what follows, we assume without loss of generality that both σ_a^2 and σ_g^2 are bounded away from zero.

Edgeworth expansions: To obtain an Edgeworth expansion for the studentized version of $\hat{Z}^{a,*}_N + \hat{Z}^{g,*}_T$ it suffices to notice that $a_1, \dots, a_N$ and $g_1, \dots, g_T$ are i.i.d. draws from their respective marginal distributions, since α_i and γ_t are i.i.d. by Assumption 2.1. Hence the smooth function model in Hall (1992) directly applies, and since bounded moments of order 4 and Cramér's condition were assumed in this Proposition, we can directly apply Theorem 2.2. in Hall (1992) to obtain the Edgeworth expansion of the sampling distribution to order $j = 2$.

For the bootstrap distribution, note that by construction of the bootstrap procedure, $a_1^*, \ldots, a_N^*$ and $g_1, \ldots, g_T^*$ are i.i.d. draws from the respective empirical distributions of $\hat{a}_i$ and $\hat{g}_t$. We can therefore directly apply Theorem 5.1 in Hall (1992), where the remaining regularity conditions are subsumed by the assumptions of this Proposition.

Comparing Moments: We next need to establish that the first three cumulants of the bootstrap distribution consistently estimate those of the sampling distribution: First note that the third moment of $\hat{a}_i$ under the sampling distribution is

$$\mathbb{E}[\hat{a}_i^3] = \left(\mathbb{E}[a_i^3] + \frac{2}{T}\mathbb{E}[a_i w_{it}^2] + \frac{1}{T^2}\mathbb{E}[w_{it}^3]\right)(1 + O(1/N))$$

where we used the fact that w_{it} is mean-independent of a_i. By the assumptions of the theorem and a central limit theorem, we then have

$$\mathbb{E}^*_{NT}[(a^*_{i,b})^3] - \mathbb{E}[a_i^3] = \frac{1}{N}\sum_{i=1}^N (\hat{a}_i^3 - \mathbb{E}[a_i^3]) = O_P(N^{-1/2}).$$

Hence, for the variables

$$\hat{W}_N^a := \frac{1}{\sqrt{N}}\sum_{i=1}^N a_i \quad \text{and } \hat{W}_N^{a,*} := \frac{1}{\sqrt{N}}\sum_{i=1}^N a^*_{i,b}$$

we have that

$$\mathbb{E}^*_{NT}\left[\left(\hat{W}_N^{a,*}\right)^3\right] - \mathbb{E}\left[\left(\hat{W}_N^a\right)^3\right] = N^{-1/2}\left(\mathbb{E}^*_{NT}[(a_i^*)^3] - \mathbb{E}[a_i^3]\right) = O_P(N^{-1})$$

Similarly, for

$$\hat{W}_N^g := \frac{1}{\sqrt{T}}\sum_{t=1}^T g_t \quad \text{and } \hat{W}_T^{g,*} := \frac{1}{\sqrt{T}}\sum_{t=1}^T g^*_{t,b}$$

we have that

$$\mathbb{E}^*_{NT}\left[\left(\hat{W}_T^{g,*}\right)^3\right] - \mathbb{E}\left[\left(\hat{W}_T^g\right)^3\right] = T^{-1/2}\left(\mathbb{E}^*_{NT}[(g_t^*)^3] - \mathbb{E}[g_t^3]\right) = O_P(T^{-1})$$

By construction, $\hat{W}_N^a$ and $\hat{W}_T^g$ and their bootstrap versions $\hat{W}_N^{a,*}$ and $\hat{W}_T^{g,*}$ are independent. For any weights s_1, s_2 we therefore have

$$\mathbb{E}^*_{NT}\left[\left(s_1\hat{W}_N^{a,*} + s_2\hat{W}_T^{g,*}\right)^3\right] - \mathbb{E}\left[\left(s_1\hat{W}_N^a + s_2\hat{W}_T^g\right)^3\right] \quad = \quad O_P(N^{-1} \vee T^{-1})$$

We can apply this in particular to the case $s_1 = \lambda_a$ and $s_2 = \lambda_g$, where $\lambda_a, \lambda_g \xrightarrow{p} 1$.
Taken together with the Edgeworth expansions of the sampling distribution and the bootstrap distribution, this implies that the bootstrap distribution $\hat{Z}_{NT}^{a,*} + \hat{Z}_{NT}^{g,*}$ approximates the sampling distribution of $\hat{Z}_N^a + \hat{Z}_T^g$ at a rate $r_{NT}^{-2} = O(N^{-1} \vee T^{-1})$ under the Kolmogorov metric.

Remainder: Finally, we need to assess the magnitude of the contribution of the remaining terms of the representation of $r_{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT})$. We already showed in the proof of Lemma A.2 that the term

$$\hat{Z}_{NT}^{e,*} + \sum_k c_k \hat{Z}_{Nk}^{\phi,*}\hat{Z}_{Tk}^{\psi,*} = O_P((NT)^{-1/2})$$

Since the limiting distribution of $\hat{Z}_{NT}^{a,*} + \hat{Z}_{NT}^{g,*}$ is Gaussian, its c.d.f. is Lipschitz-continuous, so that

$$\left\|\mathbb{P}^*_{NT}\left(\sqrt{NT}\frac{\bar{Y}^*_{NT} - \bar{Y}_{NT}}{\hat{S}^*_{NT,sel}}\right) - \mathbb{P}^*_{NT}\left(\frac{\hat{Z}_{NT}^{a,*} + \hat{Z}_{NT}^{g,*}}{\hat{S}^*_{NT,sel}}\right)\right\|_\infty = O((NT)^{-1/2})$$

as well. The analogous conclusion holds for the sampling distribution. Taken together with the rate of approximation for the leading term $\hat{Z}_N^a + \hat{Z}_T^g$ and its bootstrap analog, this establishes the claim $\square$