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. We use insights from the theory of generalized U-statistics to analyze the large-sample properties of statistics that are sample averages from the observations pooled across clusters. The asymptotic distribution of these statistics may be non-standard 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 one bootstrap procedure that is adaptive and point-wise consistent for any fixed data-generating process (DGP), and two alternative procedures that produce inference that is uniformly valid, but potentially conservative. For pivotal statistics, either procedure also provides pointwise asymptotic refinements over the Gaussian approximation when the limiting distribution is normal. Special cases and extensions discussed in the paper include 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,*

Date: November 2016 - this version: December 2018. The author thanks Matias Cattaneo, Tim Christensen, Bryan Graham, and Valentin Verdier for useful comments and gratefully acknowledges support from the NSF (SES-1459686).

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.

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 B.*

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*

[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.

the unknown distribution of X_i. The corresponding U-statistic is

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

where $w_{i_1\ldots i_D} = \mathbb{1}\{i_1 < i_2 \cdots < i_D\}$. 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 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).

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 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. However, we also consider a number of 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 long 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 B.

1.1. **Problem Description.** 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 $i = 1,\dots,N$ and $t = 1,\dots,T$ 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 $(Y_{it})_{i=1,\dots,Nt=1,\dots,T}$ 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,*

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

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

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* $\mathbb{E}[\varepsilon_{it}] = 0$, $\mathrm{Var}(\alpha_i) = \sigma_\alpha^2$, $\mathrm{Var}(\gamma_t) = \sigma_\gamma^2$, *and* $\mathrm{Var}(\varepsilon_{it}) = \sigma_\varepsilon^2$.

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

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

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

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 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. To our knowledge this analysis is new to the literature, and this paper is the first to point out that 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 correlation in second moments of Y_{it}.

[3]Since $Z_1Z_2 = \frac{1}{4}(Z_1+Z_2)^2 - \frac{1}{4}(Z_1-Z_2)^2$, where $\mathrm{Cov}(Z_1+Z_2, Z_1-Z_2) = \mathrm{Var}(Z_1) - \mathrm{Var}(Z_2) = 0$. Hence, $Z_1Z_2 = \frac{1}{2}(W_1 - W_2)$, where W_1, W_2 are independent chi-square random variables with one degree of freedom.

In order to determine what types of adaptivity and uniformity we may hope to achieve, we 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, so any inference procedure can only be uniformly valid asymptotically if it is conservative. This result includes as a special case the problem of U- and V-statistics with kernel of unknown order of degeneracy. Interestingly, all three findings do 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.

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.

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.

[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) and Aronow, Samii, and Assenova (2015) for the case of dyadic data. A more recent paper by MacKinnon, Ørregard 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 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. We show in the appendix 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 $(Y_{it} : i = 1, \ldots, N; t = 1, \ldots, T)$ 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. We furthermore give several generalizations of the main procedure and illustrate its performance using Monte Carlo simulations. Additional asymptotic results for Gaussian asymptotics, the pigeonhole bootstrap, and subsampling are given in Appendix A.

2. Representation

We assume that the sample Y_{it} for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ is embedded into a row and column (separately) 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 $\pi_1 : \{1, \ldots, \tilde{N}\} \to \{1, \ldots, \tilde{N}\}$ and $\pi_2 : \{1, \ldots, \tilde{T}\} \to \{1, \ldots, \tilde{T}\}$, we have

$$Y_{\pi_1(i)\pi_2(t)} \stackrel{d}{=} Y_{it},$$

where "$\stackrel{d}{=}$" denotes equality in distribution.

Separably exchangeable arrays can result from sampling from an infinite population of "cross-sectional" and "temporal" units ("clusters"), where the underlying double indexed array may be arbitrarily correlated, and we draw N "cross-sectional" units $i = 1, \ldots, N$ and T "temporal" units $t = 1, \ldots, T$ independently at random. Since each row and each column is drawn with the same probability, we can without loss of generality take the sample $(Y_{it} : i = 1, \ldots, N,\ t = 1, \ldots, T)$ to be the first N rows and T columns of an infinite array of the form described above.

2.1. **Exchangeable Representation.** By Theorem 1.4 in Aldous (1981) any separately exchangeable array can be represented as

$$Y_{it} = \tilde{f}(\mu, \alpha_i, \gamma_t, \varepsilon_{it})$$

for some function $f(\cdot)$, where μ, $\alpha_1, \ldots, \alpha_N$, $\gamma_1, \ldots, \gamma_T$ and $\varepsilon_{11}, \ldots, \varepsilon_{NT}$ are mutually independent, uniformly distributed random variables.[5] Similar representations are available to arrays that are jointly or separately exchangeable in more than two dimensions, see Hoover (1979) and Section 7 in Kallenberg (2005). We consider inference that is *conditional* on μ, that is conditional on the empirical distribution of Y_{it}, so that we can represent the array as

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

where $f(a, g, e) := \tilde{f}(\mu, a, g, e)$ and the factors $\alpha_1, \ldots, \alpha_N$, $\gamma_1, \ldots, \gamma_T$ and $\varepsilon_{11}, \ldots, \varepsilon_{NT}$ are the same as before.

2.2. **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.

If the relevant conditional expectations are well-defined, we can represent Y_{it} via the projection expansion

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

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

where we define $e_{it} = Y_{it} - \mathbb{E}[Y_{it}|\alpha_i, \gamma_t]$, $a_i := \mathbb{E}[Y_{i1}|\alpha_i] - \mathbb{E}[Y_{i1}]$, $g_t = \mathbb{E}[Y_{1t}|\gamma_t] - \mathbb{E}[Y_{1t}]$, $v_{it} = \mathbb{E}[Y_{it}|\alpha_i, \gamma_t] - \mathbb{E}[Y_{it}|\alpha_i] - \mathbb{E}[Y_{it}|\gamma_t] + \mathbb{E}[Y_{it}]$, and $b = \mathbb{E}[Y_{it}]$. 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 $\mathbb{E}[v_{it}|a_i, g_t] = 0$. 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 $\sigma_a^2 := \mathrm{Var}(a_i)$, $\sigma_g^2 := \mathrm{Var}(g_t)$, $\sigma_v^2 := \mathrm{Var}(v_{it})$, and $\sigma_e^2 := \mathrm{Var}(e_{it})$, respectively. We also let $w_{it} := v_{it} + e_{it}$ and denote its variance by $\sigma_w^2 = \mathrm{Var}(w_{it})$.

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.. (b) The random variables a_i/σ_a, g_t/σ_g, v_{it}/σ_v, and e_{it}/σ_e have bounded moments up to the order $4+\delta$ for some $\delta > 0$ whenever the respective variances $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2 > 0$ are non-zero. (c) We have $\sigma_a^2 + \sigma_g^2 > 0$ or $\sigma_v^2 + \sigma_e^2 > 0$.*

2.3. **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 degnerate 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.

Moreover, 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 orthnormal 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 (a) the singular values are uniformly bounded by a null sequence* $\bar{c}_k \to 0$*, that is* $c_k \leq \bar{c}_k$ *for each* $k = 1, 2, \ldots$*, and (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 C.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 appendix we compare this estimator to a "default" estimator for the asymptotic variance without a pre-test, defined as

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

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.[6]

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:

[6]Pointwise consistent model selection when a parameter relevant for the asymptotic distribution is near or at the boundary of the parameter space was first considered for the bootstrap by Andrews (2000). We also show that allowing for the case in which $\bar{v}_{NT}$ contributes to the limiting distribution, uniformly consistent estimation of the limiting distribution is not possible, neither using the bootstrap nor any alternative method.

(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 $\{1,\dots,N\}$ and $\{1,\dots,T\}$, 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 $\bar{Y}^*_{NT,1},\dots,\bar{Y}^*_{NT,B}$.

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 $\hat{S}_{NT,sel}$. For the simulation study in this paper, we implement step (c) using the two-point specification proposed by Mammen (1992) for the random variables $\omega_{1i,b}, \omega_{2t,b}$.

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* $\hat{q}_a := \max\{T\hat{\sigma}^2_a, \kappa_a\}$, $\hat{q}_g := \max\{N\hat{\sigma}^2_g, \kappa_g\}$, *and*

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

We find below that the bootstrap with model selection is consistent pointwise in $\sigma^2_a, \sigma^2_g, \sigma^2_w$, 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 $\sigma^2_a + \sigma^2_g > 0$, 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

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

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.

We first give a summary of the asymptotic properties of the bootstrap and alternative methods for estimating the asymptotic distribution, including Gaussian plug-in inference, subsampling, and Owen (2007)'s Pigeonhole bootstrap. We then establish asymptotic results for the sampling distribution and the bootstrap. Asymptotic properties for the other approaches are given in Appendix A.

4.1. **Summary of Asymptotic Properties.** The starting point of our analysis is a (novel) impossibility result in Proposition 4.1, which establishes that it is in fact not possible to achieve uniform consistency in estimating the asymptotic distribution of $\bar{Y}_{NT}$, rather uniform asymptotic validity can only be achieved by a conservative procedure.

The recommendation which inference procedure should be chosen therefore depends on the desired robustness properties, and what assumptions the researcher is willing to make regarding the underlying data generating process. We consider the following three alternative criteria, which are not nested:

- **(POINTW)** Point-wise validity with respect the variance parameters, where we allow for any of the components of $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ to be either strictly positive or zero.
- **(UNIF-1)** Uniform validity regarding clustering in means, where any of the components of $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ may be strictly positive, zero, or drifting along sequences, but $r_{NT}\sigma_v^2 \nrightarrow 0$. That is, we only exclude the degenerate case in which there is no cluster dependence in means, but cluster dependence in second moments does not vanish.
- **(UNIF-2)** Uniform validity, where we allow for any values, and drifting sequences for the components $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$.

In practice, cluster-robust methods are typically used in settings when the researcher does not know whether the data exhibit any meaningful dependence along the dimensions indexing the array $(Y_{it})_{i,t}$, but wants to guard herself against that possibility. We posit that UNIF-1 is a plausible interpretation of that idea of robustness: It only excludes the possibility that $\mathbb{E}[Y_{it}|\alpha_i, \gamma_t]$ is a random variable that has a non-degenerate distribution, but whose conditional means given α_i and γ_t happen to be close to constant.[7] This scenario is therefore non-generic once we allow for any type of cluster-dependence, and we find that

[7]More precisely, $r_{NT}\sigma_v^2 \nrightarrow 0$ would require the variance of $\text{Var}(\mathbb{E}[Y_{it}|\alpha_i, \gamma_t])$ to be of a larger order of magnitude than the variances of the conditional means given α_i or γ_t alone, $\text{Var}(\mathbb{E}[Y_{it}|\alpha_i])$ and $\text{Var}(\mathbb{E}[Y_{it}|\gamma_t])$).

extending uniformity to include this non-generic scenario (as for the third criterion) comes at the cost of a substantial power loss for the case in which observations are in fact independent within each cluster.

For criterion POINTW, we show that point-wise consistency is achieved by subsampling with model selection, the bootstrap with model selection and the pivotal bootstrap with model selection, where the pivotal bootstrap with model selection achieves refinements in the case of a Gaussian limiting distribution, and both bootstrap procedures are consistent at faster respective rates than subsampling. The non-pivotal pigeonhole bootstrap is consistent if $\sigma_a^2 + \sigma_g^2 > 0$, but conservative otherwise.

For criterion UNIF-1, uniform consistency is achieved by subsampling and the bootstrap (pivotal or not) without model selection, where again the pivotal bootstrap dominates in terms of convergence rates. Finally, under UNIF-2 only the conservative bootstrap is guaranteed to be asymptotically conservative, however at a steep price in terms of power for the degenerate cases with $r_{NT} \asymp \sqrt{NT}$ in which it over-estimates the asymptotic variance by a factor growing at the rate $\frac{\kappa_a}{T} + \frac{\kappa_g}{N}$. Proposition 4.1 implies that we cannot close this rate gap without giving up uniformity.

A full summary of the asymptotic properties of the different methods is given in Table 4.1. In addition to the different versions of the bootstrap introduced in Section 3, BS-N, BS-S, and BS-C, we also consider the following methods

- **(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 $r_{NT}\bar{Y}_{NT}$.

The pivotal versions of the different resampling procedures concern inference based on estimates for the distribution of the studentized mean, $t_{NT} := (NT)^{1/2}\, \hat{S}_{NT,def}^{-1} \bar{Y}_{NT}$ or $t_{NT} := (NT)^{1/2}\, \hat{S}_{NT,sel}^{-1} \bar{Y}_{NT}$, depending on which variance estimator is used.

To highlight some of our theoretical findings, we find that the "default" estimator from Cameron, Gelbach, and Miller (2011) for the asymptotic variance, $\hat{S}_{NT,def}^2$, is only consistent if $r_{NT}\sigma_v^2 \to 0$, whereas the modified estimator $\hat{S}_{NT,sel}^2$ is always pointwise consistent. Gaussian "plug-in" inference with a consistent estimator for the asymptotic variance is only consistent if $r_{NT}\sigma_v^2 \to 0$, subsampling inference is valid pointwise, but not uniformly, and is consistent only at a rate slower than any of the alternative procedures. The bootstrap with model selection is asympotically valid pointwise, and the bootstrap without model selection is uniformly valid as long as $r_{NT}\sigma_v^2 \to 0$. The pigeonhole bootstrap is uniformly valid asymptotically but conservative in the degenerate case, and in addition, its pivotal version

Method	Pivotal	Variance Estimator	Asymptotic Validity POINTW	UNIF-1	UNIF-2	Refinement
GAU	-	$\hat{S}^2_{NT,def}$	No	Yes	No	No
GAU	-	$\hat{S}^2_{NT,sel}$	No	No	No	No
BS-N	No	-	No	Yes	No	No
BS-N	Yes	$\hat{S}^2_{NT,def}$	No	Yes	No	Yes
BS-S	No	-	Yes	No	No	No
BS-S	Yes	$\hat{S}^2_{NT,sel}$	Yes	No	No	Yes
BS-C	No	-	Cons.	Cons.	Cons.	No
BS-C	Yes	$\hat{S}^2_{NT,sel}$	Cons.	Cons.	Cons.	(Yes)
PGH	No	-	Cons.	Cons.	No	No
PGH	Yes	$\hat{S}^2_{NT,def}$	No	Yes	No	Yes
PGH	Yes	$\hat{S}^2_{NT,sel}$	Yes	No	No	Yes
SUB	No	-	Yes	Yes	No	No
SUB	Yes	$\hat{S}^2_{NT,def}$	No	Yes	No	No
SUB	Yes	$\hat{S}^2_{NT,sel}$	Yes	No	No	No

TABLE 1. Summary of Estimation Approaches for the Asymptotic distribution of $\bar{Y}_{NT}$, where "Cons." stands for "conservative."

achieves refinements in the case of a Gaussian limiting distribution. Subsampling is consistent pointwise, but not uniformly, and approximates the asymptotic distribution at a rate no faster than $r_{NT}^{-2/3}$, assuming that subsample sizes are chosen at the respective optimal rates $m_N = O(N^{1/3}), m_T = O(T^{1/3})$. That convergence rate is slower than the r_{NT}^{-1} rate for the point-wise bootstrap, or the r_{NT}^{-2} rate for the cases for which the pivotal bootstrap yields a refinement. We also illustrate this comparison of theoretical properties in a simulation study in Section 5.

4.2. **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 $\sigma_a^2, \sigma_g^2, \sigma_e^2, \sigma_v^2$ and the covariances $\sigma_{ak} := \mathrm{Cov}(a_i, \phi_k(\alpha_i)), \sigma_{gk} := \mathrm{Cov}(g_t, \psi_k(\gamma_t))$ for $k = 1, 2, \ldots$. 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

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

an element of the sequence space ℓ^2. 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 $\mathbf{c}_{NT} := (c_{1,NT}, c_{2,NT}, \dots) \in \ell^2$ and $\mathbf{c} := (c_1, c_2, \dots) \in \ell^2$, 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 $\varrho := \lim_{N,T} \varrho_{NT}$.[8]

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

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

where $Z^e, Z_1^\phi, Z_1^\psi, Z_2^\phi, Z_2^\psi, \dots$ 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 $k = 1, 2, \dots$.

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 $\mathbf{c} = (c_1, c_2, \dots)$, we have*

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

where $\varrho := \lim_{N,T} \varrho_{NT}$ and $\|\cdot\|_\infty$ denotes the Kolmogorov metric. (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.

4.3. **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 C.1 in the appendix establishes that the component

[8] We show that without loss of generality it is sufficient to focus on convergent parameter sequences in light of arguments by Andrews and Guggenberger (2007a).

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. In particular, we establish the following impossibility result:

Proposition 4.1. ***(Estimability of Asymptotic Distribution)*** *The asymptotic distribution of $\bar{Y}_{NT}$ cannot be estimated consistently uniformly over the entire parameter space, using the bootstrap or any other method.*

See the appendix for a proof. To illustrate the problem, we re-state the counterexample underlying this impossibility result: consider the model $Y_{it} = \alpha_i \gamma_t$, where α_i, γ_t are mutually independent, with i.i.d. factors $\alpha_i \sim N(0,1)$, $\gamma_t \sim N(\mu_g, 1)$. For this model, $a_i := \mathbb{E}[Y_{it}|\alpha_i] = \alpha_i \mu_g$, $g_t := \mathbb{E}[Y_{it}|\gamma_t] = \gamma_t \mathbb{E}[\alpha_i] \equiv 0$, 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 faster 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 $\mu_g = 0$ resulting in an asymptotic variance equal to σ_v^2 from a drifting sequence $\tilde{\mu}_{T,g} := T^{-1/2} m_g$ which results in an asymptotic variance equal to $m_g^2 + \sigma_v^2$. It follows that we cannot estimate the asymptotic distribution of $\bar{Y}_{NT}$ uniformly consistently, since its variance can't be estimated consistently along this sequence.

4.4. **Bootstrap Consistency.** We now turn to the asymptotic properties for 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

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

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 $(NT)^{1/2} \hat{S}^{-1}_{NT,sel}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])$ with the distribution of its bootstrap analog, $(NT)^{1/2}(\hat{S}^*_{NT,sel})^{-1}(\bar{Y}^*_{NT} - \bar{Y}_{NT})$. Corollary C.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 C.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 C.2 in the appendix), we then obtain consistency results of the form

$$\|\mathbb{P}^*_{NT}(r_{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT})) - \mathbb{P}_{NT}(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]))\|_\infty \xrightarrow{a.s.} 0 \tag{4.2}$$

and the its pivotal analog

$$\left\| \mathbb{P}^*_{NT} \left(\sqrt{NT} \frac{\bar{Y}^*_{NT} - \bar{Y}_{NT}}{\hat{S}^*_{NT,sel}} \right) - \mathbb{P}_{NT} \left(\sqrt{NT} \frac{\bar{Y}_{NT} - \mathbb{E}[Y_{it}]}{\hat{S}_{NT,sel}} \right) \right\|_\infty \overset{a.s.}{\to} 0 \tag{4.3}$$

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

$$\|\mathbb{P}^*_{NT}(r_{NT}(\bar{Y}^*_{NT} - \bar{Y}_{NT})) - \mathcal{L}_0(\bar{\mathbf{q}}, \mathbf{c}, \varrho)\|_\infty \overset{a.s.}{\to} 0 \tag{4.4}$$

and the pivotal version of the conservative bootstrap

$$\left\| \mathbb{P}^*_{NT} \left(\sqrt{NT} \frac{\bar{Y}^*_{NT} - \bar{Y}_{NT}}{\hat{S}^*_{NT,sel}} \right) - \mathcal{L}_0(\bar{\mathbf{q}}, \mathbf{c}, \varrho) \right\|_\infty \overset{a.s.}{\to} 0 \tag{4.5}$$

Here, $\bar{\mathbf{q}} = (\bar{q}_a, \bar{q}_g, q_e, q_v, \bar{q}_{a1}, \bar{q}_{g1}, \dots)$, and $\bar{q}_a := \max\{\kappa_a/T, q_a\}$ and $\bar{q}_g := \max\{\kappa_g/N, q_g\}$, and $\bar{q}_{ak} = q_{ak}\sqrt{\bar{q}_a/q_a}$, $\bar{q}_{gk} = q_{gk}\sqrt{\bar{q}_g/q_g}$ for $k = 1, 2, \dots$, 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 (4.2) and (4.3) pointwise for any fixed $\sigma_a^2, \sigma_g^2, \sigma_e^2, \sigma_v^2$. (b) The bootstrap without model selection satisfies satisfies (4.2) and (4.3) uniformly if $q_v = 0$. (c) The conservative bootstrap satisfies (4.4) and (4.5) 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. 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 $\kappa_a/T + \kappa_g/N$.[9] 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*

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

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 using an adaptation of our bootstrap procedure to D-adic data, see Appendix B below for a discussion.

4.5. **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 the case in which the limiting distribution for the studentized mean is Gaussian. However, it is 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 $\mathcal{L}(\mathbf{q}, \mathbf{c})$, i.e. for $q_v > 0$, the studentized mean is no longer 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 first term on the right-hand side of that expression need not dominate in the limit, 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 $q_v = 0$ and the limiting distribution is Gaussian.

Hence we restrict our attention to pointwise refinements for the case of a Gaussian limiting distribution and can now state the following result:

Theorem 4.3. ***(Refinements)*** *Suppose that Assumption 2.1 holds with $\delta > 2$. Then, if $\sigma_a^2 + \sigma_g^2 > 0$ or $\sigma_v^2 = 0$ we have*

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

for the bootstrap with or without model selection, where convergence is pointwise in the distribution of the array $(Y_{it})_{i=1,\ldots,N t=1,\ldots,T}$. For the conservative bootstrap, the analogous result holds only in the nondegenerate case, $\sigma_a^2 + \sigma_g^2 > 0$.

See the appendix for a proof. Our argument uses Mammen (1992)'s result based on moment expansions of the statistic rather than the more classical approach based on Edgeworth expansions (see e.g. Liu (1988)). This allows us to include the case of a lattice distribution

for the random variables ω_{1i}, ω_{2t} in the implementation of the Wild bootstrap, including the two-point distribution described before.

5. Simulation Study

We now present simulation results to demonstrate the performance of the bootstrap procedure. 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 estimation approaches analyzed in this paper, where we consider the following alternative implementations of the bootstrap:

- **(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, 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 $\tau_\alpha^2 = \mathrm{Var}(\alpha_i)$ 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.

		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	0.085	0.076	0.072	0.063	0.077	0.072	0.063	0.072	0.088	0.071	0.141
20	20	0.070	0.069	0.066	0.057	0.068	0.066	0.056	0.067	0.071	0.067	0.097
50	50	0.059	0.059	0.059	0.051	0.059	0.058	0.051	0.060	0.060	0.059	0.074
100	100	0.056	0.057	0.056	0.051	0.056	0.056	0.051	0.056	0.058	0.057	0.070
						Design 2						
10	10	0.085	0.060	0.063	0.062	0.051	0.073	0.070	0.025	0.033	0.059	0.105
20	20	0.081	0.071	0.067	0.068	0.058	0.061	0.060	0.029	0.026	0.056	0.081
50	50	0.081	0.077	0.074	0.073	0.056	0.054	0.054	0.025	0.020	0.056	0.077
100	100	0.069	0.067	0.065	0.065	0.050	0.048	0.049	0.021	0.017	0.049	0.065
						Design 3						
10	10	0.055	0.020	0.025	0.030	0.014	0.068	0.063	0.000	0.003	0.032	0.056
20	20	0.058	0.035	0.043	0.045	0.021	0.057	0.057	0.000	0.001	0.032	0.053
50	50	0.056	0.047	0.053	0.052	0.033	0.053	0.054	0.000	0.001	0.036	0.052
100	100	0.051	0.047	0.049	0.051	0.036	0.051	0.051	0.000	0.001	0.040	0.050

TABLE 2. 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: $\sigma_a^2 = \sigma_g^2 = 0, \sigma_e^2 = 0.2$.

Results for the balanced case are given in Tables 2 and 3 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. 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 4.3. 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 3. 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 4.3 does 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.

		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	0.103	0.082	0.081	0.082	0.079	0.140	0.034	0.039	0.037	0.038	0.038	0.053
20	20	0.088	0.067	0.066	0.068	0.067	0.114	0.033	0.044	0.044	0.044	0.045	0.035
50	50	0.074	0.057	0.058	0.058	0.057	0.094	0.036	0.049	0.049	0.049	0.048	0.030
100	100	0.066	0.054	0.054	0.054	0.053	0.086	0.040	0.051	0.051	0.051	0.051	0.032
		Design 2											
10	10	0.094	0.079	0.083	0.036	0.071	0.102	0.042	0.041	0.044	0.024	0.041	0.053
20	20	0.088	0.077	0.069	0.042	0.067	0.085	0.045	0.048	0.045	0.029	0.044	0.042
50	50	0.090	0.078	0.062	0.034	0.062	0.086	0.051	0.059	0.049	0.032	0.049	0.047
100	100	0.073	0.064	0.052	0.027	0.053	0.070	0.051	0.058	0.045	0.026	0.046	0.049
		Design 3											
10	10	0.050	0.036	0.051	0.001	0.035	0.050	0.050	0.039	0.052	0.001	0.035	0.052
20	20	0.057	0.047	0.050	0.001	0.036	0.050	0.054	0.048	0.051	0.000	0.036	0.049
50	50	0.052	0.050	0.049	0.000	0.038	0.050	0.056	0.052	0.053	0.000	0.041	0.054
100	100	0.053	0.050	0.052	0.000	0.043	0.051	0.050	0.049	0.051	0.000	0.041	0.050

TABLE 3. 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: $\sigma_a^2 = \sigma_g^2 = 0, \sigma_e^2 = 0.2$.

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 report additional results for percentiles relevant for one- and two-sided tests at commonly used significance levels in the appendix.

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 4 for simulation results. While the bootstrap is not asymptotically valid if T remains fixed, results are broadly in line with those for the balanced case for the corresponding sample

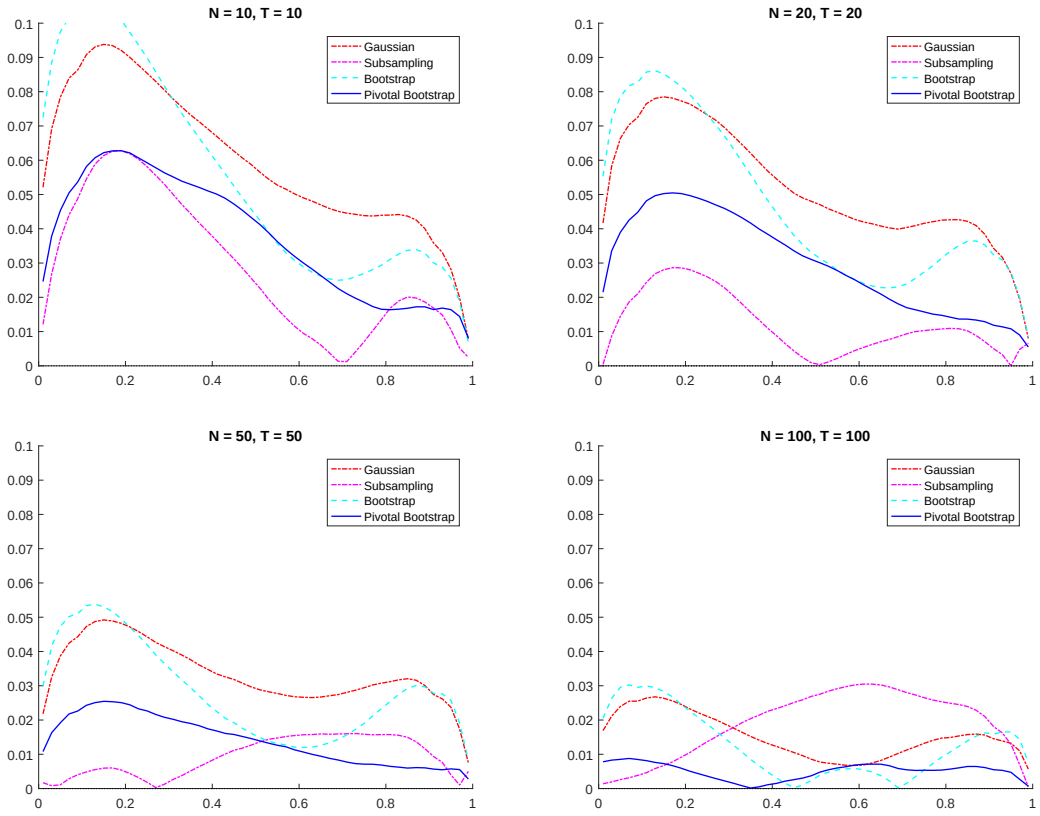

FIGURE 1. Balanced separable case: Absolute error in estimated c.d.f., plotted against nominal percentiles. Plots are based on Design 1: $\sigma_a^2 = 1, \sigma_g^2 = 0.2, \sigma_e^2 = 1$.

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 5 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

		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	0.093	0.092	0.083	0.067	0.092	0.083	0.067	0.082	0.098	0.084	0.139
20	20	0.069	0.069	0.065	0.055	0.067	0.066	0.055	0.065	0.072	0.067	0.099
50	20	0.056	0.055	0.053	0.050	0.055	0.052	0.049	0.051	0.057	0.053	0.072
100	20	0.056	0.056	0.052	0.051	0.056	0.053	0.052	0.053	0.058	0.052	0.071
						Design 2						
10	20	0.057	0.026	0.035	0.038	0.016	0.060	0.058	0.000	0.002	0.032	0.051
20	20	0.059	0.035	0.044	0.045	0.022	0.059	0.057	0.000	0.002	0.032	0.052
50	20	0.058	0.041	0.048	0.049	0.027	0.055	0.055	0.000	0.001	0.034	0.052
100	20	0.057	0.042	0.049	0.049	0.028	0.054	0.053	0.000	0.001	0.037	0.051

TABLE 4. 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

		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	0.102	0.065	0.057	0.045	0.032	0.027	0.023	0.022	0.085	0.006	0.178
20	20	0.063	0.058	0.046	0.043	0.040	0.034	0.030	0.028	0.070	0.009	0.106
50	50	0.056	0.056	0.050	0.049	0.048	0.044	0.043	0.037	0.059	0.024	0.075
100	100	0.053	0.053	0.050	0.049	0.051	0.048	0.046	0.043	0.055	0.039	0.062
∞	∞	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
						Design 2 (tests at 5 percent nominal size)						
10	10	0.105	0.040	0.044	0.025	0.004	0.004	0.004	0.003	0.018	0.001	0.110
20	20	0.100	0.055	0.056	0.051	0.002	0.002	0.002	0.001	0.011	0.000	0.089
50	50	0.103	0.068	0.069	0.067	0.002	0.002	0.001	0.001	0.007	0.000	0.083
100	100	0.111	0.082	0.083	0.083	0.001	0.001	0.001	0.001	0.006	0.000	0.089
						Design 3 (tests at 5 percent nominal size)						
10	10	0.077	0.023	0.025	0.013	0.001	0.001	0.000	0.000	0.006	0.000	0.070
20	20	0.062	0.028	0.029	0.026	0.000	0.000	0.000	0.000	0.004	0.000	0.047
50	50	0.060	0.037	0.038	0.037	0.000	0.000	0.000	0.000	0.002	0.000	0.044
100	100	0.064	0.044	0.045	0.045	0.000	0.000	0.000	0.000	0.002	0.000	0.047
∞	∞	0.065	0.050	0.050	0.050	0.000	0.000	0.000	0.000	0.000	0.000	0.050
						Design 1 (tests at 1 percent nominal size)						
10	10	0.051	0.017	0.015	0.005	0.005	0.039	0.002	0.003	0.031	0.000	0.095
20	20	0.016	0.011	0.009	0.006	0.007	0.012	0.004	0.004	0.021	0.000	0.032
50	50	0.010	0.011	0.009	0.007	0.009	0.008	0.007	0.007	0.014	0.002	0.014
100	100	0.010	0.010	0.009	0.009	0.009	0.009	0.008	0.008	0.012	0.004	0.008
∞	∞	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
						Design 2 (tests at 1 percent nominal size)						
10	10	0.061	0.007	0.008	0.001	0.000	0.000	0.000	0.000	0.005	0.000	0.045
20	20	0.051	0.008	0.009	0.005	0.000	0.000	0.000	0.000	0.001	0.000	0.024
50	50	0.051	0.012	0.014	0.011	0.000	0.000	0.000	0.000	0.001	0.000	0.019
100	100	0.058	0.016	0.018	0.017	0.000	0.000	0.000	0.000	0.000	0.000	0.021
						Design 3 (tests at 1 percent nominal size)						
10	10	0.041	0.003	0.004	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.027
20	20	0.027	0.003	0.004	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.012
50	50	0.027	0.005	0.005	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.007
100	100	0.029	0.007	0.008	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.009
∞	∞	0.032	0.010	0.010	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.010

TABLE 5. 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.

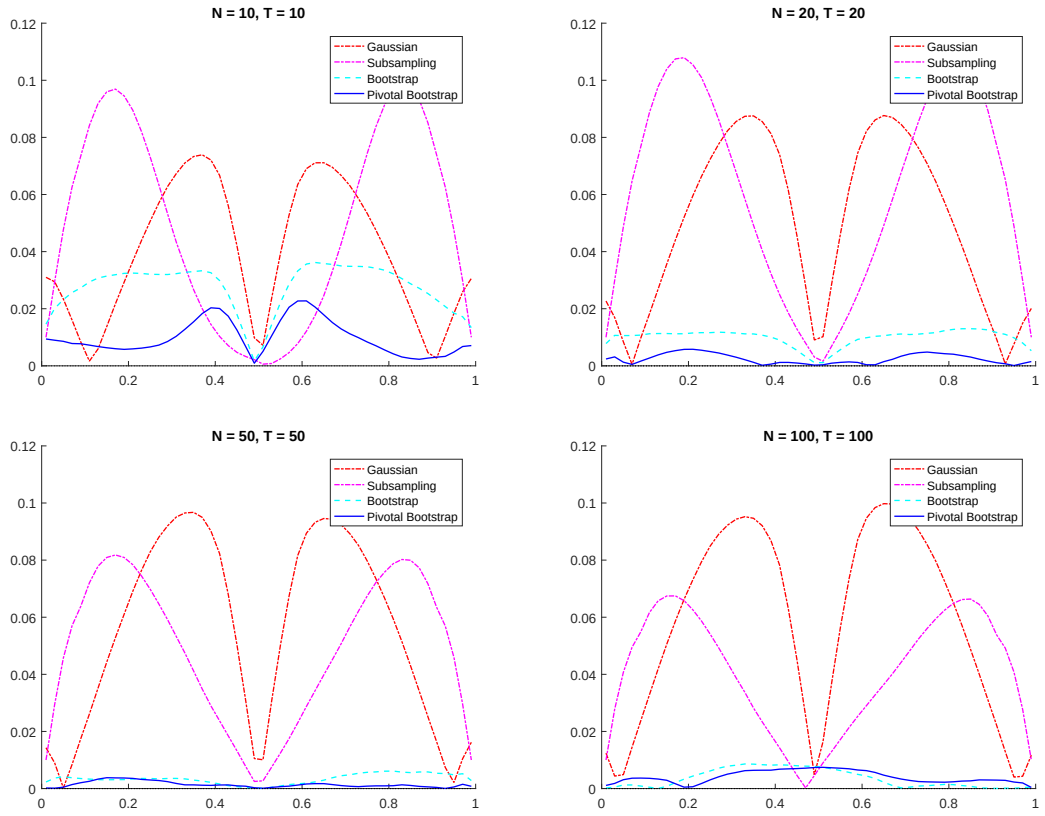

FIGURE 2. Nonseparable case: Absolute error in estimated c.d.f., plotted against nominal percentiles. Plots are based on Design 2: $\sigma_a^2 = 0.5, \sigma_g^2 = 0.5, \sigma_e^2 = 0.1$ and $\mu_a = \mu_g = 0$.

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.

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.

Andrews, D., and P. Guggenberger (2007a): "The Limit of Finite-Sample Size and a Problem with Subsampling," working paper, Yale University and UCLA.

——— (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., 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.

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

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

Götze, F., and N. Tikhomirov (1999): "Asymptotic Distribution of Quadratic Forms," *Annals of Probability*, 27(2), 1072–1098.

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.

Kline, P., and A. Santos (2012): "A Score Based Approach to Wild Bootstrap Inference," *Journal of Econometric Methods*, 1(1), 23–41.

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. Ørregard 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.

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.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," Handbook of Econometrics, Vol IV Chapter 36.

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

Politis, D., and J. Romano (1994): "Large Sample Confidence Regions based on Subsamples under Minimal Assumptions," *Annals of Statistics*, 22(4), 2031–2050.

Politis, D., J. Romano, and M. Wolf (1999): *Subsampling*. Springer, New York.

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

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. Alternative Inference Procedures

This section gives asymptotic results for alternative methods of estimating the asymptotic distribution of $\bar{Y}_{NT}$, where we consider Gaussian inference using the robust variance estimator proposed by Cameron,

Gelbach, and Miller (2011), Gaussian inference using the modified robust variance estimator $\hat{S}_{NT,sel}$ introduced in section 3, subsampling inference (Politis and Romano (1994), Politis, Romano, and Wolf (1999)), and Owen (2007)'s pigeonhole bootstrap.

A.1. **Gaussian Asymptotic Inference (GAU).** We first discuss inference using an estimator of the asymptotic variance together with quantiles of the Gaussian distribution. Specifically, we consider the two different variance estimators $\hat{S}^2_{NT,def}$ and $\hat{S}^2_{NT,def}$ introduced in Section 3.

Corollary C.1 below shows that $\hat{S}^2_{NT,sel}$ is pointwise consistent for the asymptotic variance. We now give a counterexample to show that the default estimator $\hat{S}^2_{NT,def}$ is not: Suppose that

$$Y_{it} = \alpha_i \gamma_t, \qquad \alpha_i, \gamma_t \overset{iid}{\sim} N(0,1)$$

Since α_i and γ_t are independent and have zero mean, the convergence rate of the sample mean is $r_{NT}^{-2} = (NT)^{-1}$. We can then verify that the asymptotic variance of the sample mean is

$$\text{Var}(\sqrt{NT}\bar{Y}_{NT}) = \text{Var}\left(\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\alpha_i\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\gamma_t\right]\right) = \text{Var}(\alpha_i)\text{Var}(\gamma_t) = 1$$

Plugging the model into the expression for the variance estimator and rearranging terms we find that

$$\begin{aligned}
\frac{T}{N}\sum_{i=1}^{N}(\bar{Y}_{iT} - \bar{Y}_{NT})^2 &= \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\gamma_t\right)^2 \frac{1}{N}\sum_{i=1}^{N}(\alpha_i - \bar{\alpha}_N)^2 \\
\frac{N}{T}\sum_{t=1}^{T}(\bar{Y}_{Nt} - \bar{Y}_{NT})^2 &= \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\alpha_i\right)^2 \frac{1}{T}\sum_{t=1}^{T}(\gamma_t - \bar{\gamma}_T)^2 \\
\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \bar{Y}_{NT})^2 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_i^2\gamma_t^2 - \bar{\alpha}_N^2\bar{\gamma}_T^2\right)
\end{aligned}$$

where $\bar{\alpha}_N := \frac{1}{N}\sum_{i=1}^{N}\alpha_i$ and $\bar{\gamma}_T := \frac{1}{T}\sum_{t=1}^{T}\gamma_t$. Clearly, $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\alpha_i$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\gamma_t$ converge to independent standard normal random variables, $\frac{1}{N}\sum_{i=1}^{N}(\alpha_i - \bar{\alpha}_N)^2 \xrightarrow{p} \text{Var}(\alpha_i) = 1$, and $\frac{1}{T}\sum_{t=1}^{T}(\gamma_t - \bar{\gamma}_T)^2 \xrightarrow{p} \text{Var}(\gamma_t) = 1$. Hence, by Slutsky's Lemma, it follows that

$$\hat{S}^2_{NT,def} - 1 \xrightarrow{d} Y_1 + Y_2 - 2$$

where Y_1, Y_2 are independent draws from a chi-square distribution with one degree of freedom. In particular, for this specific distribution of the array $(Y_{it})_{i,t}$, the limiting distribution on the right-hand side has zero mean and non-zero variance so that the default estimator of the asymptotic variance is unbiased but inconsistent. However, using arguments parallel to the consistency proof for the modified estimator in Proposition 4.1, the estimator $\hat{S}_{NT,def}$ remains consistent if $q_v = 0$.

Finally we turn to asymptotic validity of Gaussian inference using either variance estimator - from Theorem 4.1, the asymptotic distribution for the sample mean is $\left(\sqrt{q_e}Z^e + \sqrt{q_a}Z^a + \sqrt{q_g}Z^g\right) + \varrho V = \sqrt{1-q_v}Z + \varrho V$, where V is Wiener chaos governed by the spectral coefficients $\mathbf{c}$ and with unit variance, and Z is a random variable with a standard normal marginal distribution. Given a consistent estimator of the asymptotic variance, the Gaussian approximation assumes a limiting distribution $Z + 0 \cdot V$. Since both Z and V have zero mean and unit variance, there is no clear dominance relationship across all relevant percentiles and the tails between the true limiting distribution and the Gaussian approximation when $q_v > 0$. Hence for a given testing problem, values of $q_v > 0$ and spectral coefficients $\mathbf{c}$, Gaussian inference may or may not control size conservatively, depending on the nominal significance level and the specific distribution of Gassian chaos V.

In contrast, when $q_v = 0$, either variance estimator is consistent and Gaussian inference is asymptotically valid. However, as in the standard case of i.i.d. data, Gaussian inference does not provide higher-order refinements.

A.2. **Subsampling (SUB).** As an alternative to the bootstrap, the researcher may estimate the limiting distribution of $\bar{Y}_{NT}$ using subsampling. Specifically, we consider the following procedure:

(a) We choose subsample sizes $m_N, m_T \to \infty$, where we assume throughout that $m_N/N, m_T/T \to 0$.
(b) for the bth subsample, let $j(1), \ldots, j(m_N)$ and $s(1), \ldots, s(m_T)$ be drawn uniformly and independently without replacement from $\{1, \ldots, N\}$ and $\{1, \ldots, T\}$, respectively.
(c) We then let $Y^\circ_{it,b} := Y_{j(i)s(t),b}$ for $i = 1, \ldots, m_N$ and $t = 1, \ldots, m_T$, and form the bth subsample mean $\bar{Y}^\circ_{NT,b} := \frac{1}{m_N m_T} \sum_{i=1}^{m_N} \sum_{t=1}^{m_T} Y^\circ_{it,b}$.

For a pivotal version of subsampling, we use the variance estimator

$$(\hat{S}^\circ_{NT,b})^2 := \hat{D}_a(\kappa_a) T (\hat{\sigma}^\circ_a)^2 + \hat{D}_g(\kappa_g) T (\hat{\sigma}^\circ_g)^2 + (\hat{\sigma}^\circ_w)^2$$

Here, the variance estimators $\hat{\sigma}^\circ_a, \hat{\sigma}^\circ_g, \hat{\sigma}^\circ_w$ are the subsample analogs of $\hat{\sigma}^2_a, \hat{\sigma}^2_g, \hat{\sigma}^2_w$, the selectors $\hat{D}_a(\kappa), \hat{D}_g(\kappa)$ based on the initial sample are as defined in Section 3, and $\kappa_a, \kappa_g \geq 0$ are chosen according to whether subsampling is implemented with or without model selection.

We can enumerate the possible subsamples of this type by $b = 1, \ldots, B^\circ_{NT}$ where $B^\circ_{NT} := \binom{N}{m_N}\binom{T}{m_T}$ and denote the conditional distribution of the normalized subsample mean given the sample $(Y_{it} : i = 1, \ldots, N, t = 1, \ldots, T)$ with

$$\mathbb{P}^\circ_{NT}(r^\circ_{NT}(\bar{Y}^\circ_{NT} - \bar{Y}_{NT}) \leq x) := \frac{1}{B^\circ_{NT}} \sum_{b=1}^{B^\circ_{NT}} \mathbb{1}\left\{r^\circ_{NT}(\bar{Y}^\circ_{NT,b} - \bar{Y}_{NT}) \leq x\right\}$$

Here, we denote the rate for the subsample mean with

$$(r^\circ_{NT})^2 := m_N^{-1}\sigma_a^2 + m_T^{-1}\sigma_g^2 + (m_N m_T)^{-1}\sigma_w^2$$

We can summarize our findings for this subsampling procedure in the following proposition:

Proposition A.1. ***(Subsampling)*** *Suppose that Assumption 2.1 holds, $m_N, m_T \to \infty$, and $\frac{m_N}{N}, \frac{m_T}{T} \to 0$. Then*

$$\|\mathbb{P}^\circ_{NT}(r^\circ_{NT}(\bar{Y}^\circ_{NT} - \bar{Y}_{NT})) - \mathbb{P}_{NT}(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]))\|_\infty \xrightarrow{p} 0$$

pointwise. If in addition Assumption 2.2 holds, subsampling is consistent along drifting sequences if and only if $q_v = 0$ or $(r^\circ_{NT})^2(m_N^{-1}\sigma_a^2 + m_T^{-1}\sigma_g^2) \to 0$.

It is straightforward to establish consistency for pivotal versions of subsampling, where we can use Corollary C.1 below to show pointwise consistency for subsampling using the subsampling analog of the variance estimator with model selection, and uniform consistency regarding clustering in means (UNIF-1) without model selection.

As in the i.i.d. case, the subsampling estimator for the limiting distribution converge at a slower rate than the bootstrap, which depend on subsample sizes m_N, m_T rather than N and T, respectively. Specifically, noting that the leading terms of the decomposition of $\bar{Y}^\circ_{NT} - \mathbb{E}[Y_{it}]$ are i.i.d., we can adapt the argument in Section 2.4 of Politis and Romano (1994) to establish that for the pivotal version of subsampling

$$\left\| \mathbb{P}^\circ_{NT}\left(\sqrt{m_N m_T}\left(\frac{\bar{Y}^\circ_{NT,b} - \bar{Y}_{NT}}{\hat{S}^\circ_{NT,b}}\right)\right) - \mathbb{P}_{NT}\left(\sqrt{NT}\left(\frac{\bar{Y}_{NT} - \mathbb{E}[Y_{it}]}{\hat{S}_{NT,sel}}\right)\right)\right\|_\infty = O_P\left((r^\circ_{NT})^{-1} + \left(\frac{r^\circ_{NT}}{r_{NT}}\right)^2\right)$$

where r°_{NT} depends on the choice of the sequences m_N, m_T. We can separately check for each case with respect to the magnitudes of $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ that m_N, m_T can be chosen such that $(r^\circ_{NT})^{-1} + \left(\frac{r^\circ_{NT}}{r_{NT}}\right)^2 = O\left(r_{NT}^{-2/3}\right)$, but no faster rate can be achieved. This also gives the fastest possible rate at which subsampling can approximate the asymptotic distribution. As with subsampling of i.i.d. data, this convergence rate is the same for the pivotal as for the non-pivotal case. These findings for Gaussian asymptotic inference and subsampling are summarized in Table 4.1 in the main text.

Proof of Proposition A.1: Define the local parameters

$$\begin{aligned} q^\circ_{a,NT} &:= (r^\circ_{NT})^2 m_N^{-1}\sigma_a^2, & q^\circ_{g,NT} &:= (r^\circ_{NT})^2 m_T^{-1}\sigma_g^2 \\ q^\circ_{e,NT} &:= (r^\circ_{NT})^2 (m_N m_T)^{-1}\sigma_e^2 & q^\circ_{v,NT} &:= (r^\circ_{NT})^2 (m_N m_T)^{-1}\sigma_v^2 \\ q^\circ_{ak,NT} &:= (r^\circ_{NT})^2 m_N^{-1}\sigma_{ak} & q^\circ_{gk,NT} &:= (r^\circ_{NT})^2 m_T^{-1}\sigma_{gk} \end{aligned} \tag{A.1}$$

for $k = 1, 2, \dots$, and for given sequences m_N, m_T we denote the limits with

$$q^\circ_a := \lim_{N,T} q_{a,m_N m_T},\ q^\circ_g := \lim_{N,T} q_{g,m_N m_T}\ q^\circ_e := \lim_{N,T} q_{e,m_N m_T}\ q^\circ_v := \lim_{N,T} q_{v,m_N m_T}$$

$$q^\circ_{ak} := \lim_{N,T} q_{ak,m_N m_T},\ q^\circ_{gk} := \lim_{N,T} q_{gk,m_N m_T}$$

for $k = 1, 2, \dots$.

Let $J_{NT}(x) := \mathbb{P}(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]) \le x)$ and $J^\circ_{NT}(x) := \mathbb{P}^\circ_{NT}(r^\circ_{NT}(\bar{Y}^\circ_{NT} - \mathbb{E}[Y_{it}]) \le x)$ be the respective unconditional c.d.f.s of the normalized sample mean and its subsample analog. We first check whether $J_{NT}(x)$ and $J^\circ_{NT}(x)$ have the same limits under different assumptions on the variance components, and then give necessary and sufficient conditions for consistency of the subsampling estimator for $J_{NT}(x)$.

For the bth subsample rows and columns are drawn uniformly and without replacement from $\{1, \dots, N\}$ and $\{1, \dots, T\}$ respectively. Hence the array $\left(Y^\circ_{it,b} : i = 1, \dots, m_N; t = 1, \dots, m_T\right)$ is a draw of size $m_N \times m_T$ from the same separately exchangeable array as the initial sample with second moments $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ and spectral coefficients $\mathbf{c} = (c_1, c_2, \dots)$ for the sparse representation of $\mathbb{E}[Y_{it}|\alpha_i, \gamma_t]$.

Hence, if we let $\mathbf{q}^\circ_{NT} := (q^\circ_{e,NT}, q^\circ_{a,NT}, q^\circ_{g,NT}, q^\circ_{a1,NT}, q^\circ_{a1,NT}, q^\circ_{a2,NT}, q^\circ_{g2,NT}, \dots)$, it follows from Theorem 4.1 that along any convergent sequence $\mathbf{q}^\circ_{NT} \to \mathbf{q}^\circ = (q^\circ_e, q^\circ_a, q^\circ_g, q^\circ_{a1}, q^\circ_{g1}, q^\circ_{a2}, q^\circ_{g2}, \dots)$, we have

$$\|\mathbb{P}_{NT}(r^\circ_{NT}(\bar{Y}^\circ_{m_N m_T, b} - \mathbb{E}[Y_{it}])) - \mathcal{L}_0(\mathbf{q}^\circ, \mathbf{c}, \varrho^\circ)\|_\infty \to 0$$

where $\varrho^\circ := \lim_{NT} r^\circ_{NT}(NT)^{-1/2}$. In particular, the respective limits of $J_{NT}(x)$ and $J^\circ_{NT}(x)$ along such a subsequence are continuous and coincide if and only if $\mathbf{q}^\circ = \mathbf{q}$. Moreover, noting that the leading terms of the decomposition of $\bar{Y}^\circ_{NT} - \mathbb{E}[Y_{it}]$ are i.i.d., we can adapt the main steps of the proof of Theorem 2.1 in Politis and Romano (1994) to conclude that subsampling is consistent whenever $J_{NT}(x)$ and $J^\circ_{NT}(x)$ have same limits.

For pointwise properties of the subsampling estimator, that is whenever the variances $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ are held fixed, we need to distinguish only two cases: if $q_a + q_g > 0$ it follows that $q_v = q_e = 0$, so that $q_a + q_g = 1$. By inspection we then also have $q^\circ_v = q^\circ_e = 0$ and $q^\circ_a + q^\circ_g = 1$. If $q_a + q_g = 0$, then we also have $q^\circ_a + q^\circ_g = 0$. Since the subsample is a draw from the same separately exchangeable array as the initial sample, it also follows that $q^\circ_e = q_e$ and $q^\circ_v = q_v$, so that $J_{NT}(x)$ and $J^\circ_{NT}(x)$ have the same pointwise limits when $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ are fixed.

For drifting sequences, we can now distinguish several cases regarding the limit of the sampling distribution: If $q_v = 0$ then $q^\circ_v = 0$ and $q_a + q_g + q_e = q^\circ_a + q^\circ_g + q^\circ_e = 1$, so that the limiting distributions coincide.

If $q_v > 0$ and $q_a + q_g > 0$, then $m_N/N \to 0$ and $m_T/T \to 0$ implies that $q_a^\circ + q_g^\circ = 1$ and $q_v^\circ = 0$ so that subsampling is inconsistent along that sequence. Furthermore, for certain sequences m_N, m_T we may also have $q_a^\circ + q_g^\circ > 0$ and $q_v^\circ > 0$ when $q_a + q_g = 0$ and $q_v > 0$. Hence, $J_{NT}(x)$ and $J_{NT}^\circ(x)$ do not converge to the same limit whenever $q_v > 0$ and $q_a^\circ + q_g^\circ > 0$, so that subsampling is not consistent under these sequences. Since there is no unambiguous dominance relationship in the respective percentiles of the standard normal distribution and Wiener chaos, subsampling inference is also not guaranteed to be conservative unless $q_v = q_v^\circ$ □

A.3. **Pigeonhole Bootstrap (PGH).** We next consider Owen (2007)'s "pigeonhole" bootstrap for inference regarding $\mathbb{E}[Y_{it}]$ under multi-way clustering. Large-sample results were provided by Owen (2007) for the additively separable case, and by Davezies, D'Haultfœuille, and Guyonvarch (2018) for the asymptotic distribution at the $\sqrt{\min\{N,T\}}$ rate. We give a result at the adaptive r_{NT} rate that explicitly accounts for the non-separable case as well. To simplify derivations, we consider a slight modification of the procedure by Owen (2007), where instead of drawing units $i \in \{1,\dots,N\}$ and $t \in \{1,\dots,T\}$ with replacement, we assign each "row" i and "column" t random resampling weights M_i and M_t that are drawn i.i.d. from a fixed distribution.

Specifically, we consider the following procedure:

(a) For the bth bootstrap iteration generate random weights $M_{1i,b}$ for each $i = 1,\dots,N$ ($M_{2t,b}$, respectively, for $t = 1,\dots,T$) as i.i.d. draws from a binomial distribution with N trials and success probability $\frac{1}{N}$ (T trials and success probability $\frac{1}{T}$, respectively).

(b) We then form the bth boostrap mean

$$\bar{Y}_{NT,b}^{*,PG} := \frac{1}{N_b^* T_b^*} \sum_{i=1}^{N} \sum_{t=1}^{T} M_{1i,b} M_{2t,b} Y_{it}$$

where $N_b^* := \sum_{i=1}^N M_{1i,b}$ and $T_b^* := \sum_{t=1}^T M_{2t,b}$.

For the pivotal bootstrap we can use the modified variance estimator with or without model selection. Specifically, let

$$\begin{aligned}
\hat{s}_a^{2,*PG} &:= \frac{1}{N_b^*} \sum_{i=1}^{N} M_{1i,b} (\bar{Y}_{iT,b}^{*,PG} - \bar{Y}_{NT,b}^{*,PG})^2, \quad \hat{s}_g^{2,*PG} := T_b^* \sum_{t=1}^{T} M_{2t,b} (\bar{Y}_{Nt}^{*,PG} - \bar{Y}_{NT}^{*,PG})^2 \\
\hat{s}_w^{2,*PG} &:= \frac{1}{N_b^* T_b^*} \sum_{i=1}^{N} \sum_{t=1}^{T} M_{1i,b} M_{2t,b} (Y_{it}^{*,PG} - \bar{Y}_{NT}^{*,PG})^2
\end{aligned}$$

where we denote the row and column means $\bar{Y}_{iT,b}^{*,PG} := \frac{1}{T_b^*} \sum_{t=1}^T M_{2t,b} Y_{it}$ and $\bar{Y}_{Nt}^{*,PG} := \frac{1}{N_b^*} \sum_{i=1}^N M_{1i,b} Y_{it}$. We then form the variance estimator

$$\hat{S}_{sel,b}^{2,*PG} := T_b^* \hat{D}_a(\kappa_a) \max\left\{0, \hat{s}_a^{2,*PG} - \frac{1}{T_b^*} \hat{s}_w^{2,*PG}\right\} + N_b^* \hat{D}_g(\kappa_g) \max\left\{0, \hat{s}_g^{2,*PG} - \frac{1}{N_b^*} \hat{s}_w^{2,*PG}\right\} + \hat{s}_w^{2,*PG}$$

where the selectors $\hat{D}_a(\kappa), \hat{D}_g(\kappa)$ defined in Section 3 are evaluated for the initial sample and κ_a, κ_g are chosen according to whether we use the variance estimator with or without model selection.

We denote the conditional law of $\bar{Y}_{NT}^{*,PG}$ given the sample $(Y_{it})_{i,t}$ with

$$\mathbb{P}_{NT}^{*,PG}(r_{NT}(\bar{Y}_{NT}^{*,PG} - \bar{Y}_{NT}) \le x) := \mathbb{P}_{M_1,M_2}\left(r_{NT}(\bar{Y}_{NT,b}^{*,PG} - \bar{Y}_{NT}) \le x \middle| Y_{11},\dots,Y_{NT}\right)$$

This is a modification of Owen (2007)'s pigeonhole bootstrap with random sample size. We do not claim any theoretical advantages for this modification. Rather we only introduce it to simplify the theoretical

analysis and find that its asymptotic properties match those of the original procedure in the cases where those properties have been derived previously. The simulation study in Section 5 implements the original version proposed by Owen (2007), and results match the theoretical properties shown here for the modified version.

Specifically, we find the following:

Proposition A.2. ***(Pigeonhole Bootstrap)*** *Suppose that Assumptions 2.1 and 2.2 hold. Then if $q_v = 0$, the pigeonhole bootstrap*

$$\|\mathbb{P}^{*,PG}_{NT}(r_{NT}(\bar{Y}^{*,PG}_{NT} - \bar{Y}_{NT})) - \mathbb{P}_{NT}(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}]))\|_\infty \xrightarrow{p} 0$$

uniformly, where $\mathbb{P}^{PG}_{NT}$ is the convolution of the sampling distribution for $r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])$ with an independent Gaussian random variable with variance $2(q_e + q_v)$.

In particular, the pigeonhole bootstrap is consistent in the non-degenerate case $\sigma_a^2 + \sigma_g^2 > 0$ and asymptotically conservative for the sampling distribution under bowl-shaped loss not only point-wise but uniformly as long as $q_v = 0$. On the other hand, the pigeonhole bootstrap is not guaranteed to converge to a deterministic limit for $q_v > 0$, and it furthermore over-estimates the contribution of the average $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T (v_{it} + e_{it})$ to the limiting distribution by a factor of three, which can result in a substantial reduction in power when observations are uncorrelated or even fully independent within clusters. It is possible to show that a pivotal version of the pigeonhole that uses the two-way clustering robust variance estimator without model selection does not suffer from that power reduction in the degenerate case, but remains inconsistent when $q_v > 0$. We report simulation results for both versions of the pigeonhole bootstrap in Section 5.

One might also consider modifying the pigeonhole bootstrap using model selection along the lines of 3 in order to improve its pointwise properties at the expense of losing uniformity for $q_v > 0$. We find that in contrast to the new bootstrap procedure proposed in this paper, plausible modifications of the pigeonhole bootstrap along these lines still fail to achieve point-wise consistency.[10]

Proof of Proposition A.2: For the bth bootstrap replication, we can decompose the mean as

$$\begin{aligned}
\bar{Y}^{*,PG}_{NT,b} &= \bar{Y}_{NT} + \frac{1}{N_b^*}\sum_{i=1}^N\sum_{t=1}^T M_{1i,b}[(a_i - \bar{a}_N) + (\bar{v}_{iT} - \bar{v}_{NT})] + \frac{1}{T_b^*}\sum_{t=1}^T M_{2t,b}[(g_t - \bar{g}_T) + (\bar{v}_{Nt} - \bar{v}_{NT})] \\
&\quad + \frac{1}{N_b^* T_b^*}\sum_{i=1}^N\sum_{t=1}^T M_{1i,b}M_{2t,b}(e_{it} - \bar{e}_{NT}) \\
&\quad + \sum_{k=1}^\infty c_k \left(\frac{1}{N_b^*}\sum_{i=1}^N M_{1i,b}\left(\phi_k(\alpha_i) - \bar{\phi}_{kN}\right)\right)\left(\frac{1}{T_b^*}\sum_{t=1}^T M_{2t,b}\left(\psi_k(\gamma_t) - \bar{\psi}_{kT}\right)\right)
\end{aligned} \tag{A.12}$$

where $\bar{v}_{iT} := \frac{1}{T}\sum_{t=1}^T v_{it}$, $\bar{v}_{Nt} := \frac{1}{N}\sum_{i=1}^N v_{it}$, $\bar{\phi}_{kN} := \frac{1}{N}\sum_{i=1}^N \phi_k(\alpha_i)$, and $\bar{\psi}_{kT} := \frac{1}{T}\sum_{t=1}^T \psi_k(\gamma_k)$. We can immediately verify that for the binomial distribution, $\mathbb{E}[M_i] = 1$ and $\mathbb{E}[M_{1i}^2] = \mathbb{E}[M_{1i}]^2 + \mathrm{Var}(M_{1i}) = 2 - \frac{1}{N}$. Similarly, $\mathbb{E}[M_{2t}] = 1$ and $\mathbb{E}[M_{2t}^2] = 2 - \frac{1}{T}$, where $M_{11}, \ldots, M_{1N}$ and $M_{21}, \ldots, M_{2T}$ are also independent.

[10]Specifically, if the consistent pre-test for clustering in means fails to reject the null of no dependence, a modified bootstrap could either switch to a bootstrap that treats entries in each column or row as independent, or subtract column- or row-means from observations to eliminate a spurious correlation. We find that neither alternative is pointwise consistent, where the first proposal results in a Gaussian limit for the bootstrap distribution even when $q_v > 0$, and would therefore be inconsistent (and not necessarily conservative). The second alternative would replicate the distribution of the Wiener chaos component, but continue to over-estimate the scale of the asymptotic distirbution in the degenerate case by $2\sigma_e^2$.

Hence, conditional on $e_{11}, \ldots, e_{NT}$,

$$\begin{aligned}\mathrm{Var}_{NT}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} M_{1i,b}M_{2t,b}e_{it}\right) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbb{E}[M_{1i,b}^2]\mathbb{E}[M_{2t,b}^2] - \mathbb{E}[M_{1i,b}]^2\mathbb{E}[M_{2t,b}]^2\right)e_{it}^2 \\ &= \left(3 - \frac{2(N+T)-1}{NT}\right)\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 =: (\sigma_{e,NT}^{*,PG})^2\end{aligned}$$

noting that $M_{i,b}, M_{t,b}$ are independent. Similarly, conditional on $\alpha_1, \ldots, \alpha_N$,

$$\begin{aligned}\mathrm{Var}_{NT}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M_{1i,b}\left(\phi_k(\alpha_i) - \bar{\phi}_{kN}\right)\right) &= (\mathbb{E}[M_{1i}^2] - \mathbb{E}[M_{1i}]^2)\frac{1}{N}\sum_{i=1}^{N}\left(\phi_k(\alpha_i) - \bar{\phi}_{kN}\right)^2 \\ &= \left(1 - \frac{1}{N}\right)\frac{1}{N}\sum_{i=1}^{N}\left(\phi_k(\alpha_i) - \bar{\phi}_{kN}\right)^2 =: (\sigma_{\phi k,NT}^{*,PG})^2\end{aligned}$$

with analogous results for the variances and covariances among one-dimensional averages $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M_{1i,b}(a_i + \bar{v}_i)$, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} M_{2t,b}(g_t + \bar{v}_t)$, and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} M_{2t,b}\left(\psi_k(\gamma_t) - \bar{\psi}_{kT}\right)$.

Next, we let $(\sigma_{a,NT}^{*,PG})^2 := \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(e_{it} - \bar{e}_{NT})^2$, $(\sigma_{a,NT}^{*,PG})^2 := \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{a}_N)^2$, $(\sigma_{g,NT}^{*,PG})^2 := \frac{1}{T-1}\sum_{t=1}^{T}(g_t - \bar{g}_T)^2$, be the empirical variances of the projection components. Similarly for $k = 1, 2, \ldots$ we define the empirical variances $(\sigma_{\phi k,NT}^{*,PG})^2, (\sigma_{\psi k,NT}^{*,PG})^2$ and covariances $\sigma_{ak,NT}^{*,PG}, (\sigma_{gk,NT}^{*,PG})$ with the basis functions of the spectral representation of the conditional mean function. We can then characterize the pigeonhole bootstrap distribution in terms of the local parameters $q_{s,NT}^{*,PG} := r_{NT}(\sigma_{s,NT}^{*,PG})^2$ for $s = a, g, \phi 1, \psi 1, a1, g1, \ldots$, and

$$\mathbf{q}_{NT}^{*,PG} := (q_{e,NT}^{*,PG}, q_{a,NT}^{*,PG}, q_{g,NT}^{*,PG}, q_{\phi 1,NT}^{*,PG}, q_{\psi 1,NT}^{*,PG}, q_{a1,NT}^{*,PG}, q_{g1,NT}^{*,PG}, \ldots)$$

We also define its population analog

$$\mathbf{q}_{NT}^{PG} = (q_{e,NT}^{PG}, q_{a,NT}^{PG}, q_{g,NT}^{PG}, q_{\phi 1,NT}^{PG}, q_{\psi 1,NT}^{PG}, q_{a1,NT}^{PG}, q_{g1,NT}^{PG}, \ldots),$$

where $q_{s,NT}^{PG} = q_{s,NT}$ for each $s = a1, g1, \ldots$, $q_{\phi k,NT}^{PG} = q_{\psi k,NT}^{PG} = 1$ for each $k = 1, 2, \ldots$, $q_{a,NT}^{PG} = q_{a,NT} + q_{v,NT}$, $q_{g,NT}^{PG} = q_{g,NT} + q_{v,NT}$, and $q_{e,NT}^{PG} = 3q_{e,NT}$.

If $q_v = 0$, Lemma 3.1 together with a law of large numbers for the components corresponding to moments of the basis functions $\phi_k(\alpha_i), \psi_k(\gamma_t)$ implies that for each $K < \infty$, $\|\mathbf{q}_{NT,K}^{*,PG} - \mathbf{q}_{NT,K}^{PG}\| \xrightarrow{p} 0$ pointwise, where $\mathbf{q}_{NT,K}^{*,PG}$ and $\mathbf{q}_{NT,K}^{PG}$ denote the subvectors consisting of the first $3 + 4K$ components of $\mathbf{q}_{NT}^{*,PG}$ and $\mathbf{q}_{NT,K}^{PG}$, respectively. In particular, for the pigeonhole bootstrap all relevant variance parameters converge in probability to their corresponding population analogs, except for $q_{e,NT}^{*,PG}$ which converges to $3q_{e,NT}$ instead.

Next, along any convergent sequence $\mathbf{q}_{NT}^{PG} \to \mathbf{q}^{PG}$ we can apply a CLT to obtain a Gaussian joint asymptotic distribution for any finite subset of the averages $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} M_{1i,b}(a_i + \bar{v}_i)$, $\frac{1}{T}\sum_{t=1}^{T} M_{2t,b}(g_t + \bar{v}_t)$, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} M_{1i,b}M_{2t,b}e_{it}$, $\frac{1}{N}\sum_{i=1}^{N} M_{1i,b}\left(\phi_k(\alpha_i) - \bar{\phi}_{kN}\right)$ and $\frac{1}{T}\sum_{t=1}^{T} M_{2t,b}\left(\psi_k(\gamma_t) - \bar{\phi}_{kT}\right)$ for $k = 1, 2, \ldots$. Also, $\frac{N_b^*}{N}, \frac{T_b^*}{T} \xrightarrow{p} 1$ by a law of large numbers.

Following the truncation argument from the proof of Theorem 4.1, we can then conclude that along any convergent sequences $\mathbf{q}_{NT} \to \mathbf{q}$,

$$\|\mathbb{P}_{NT}^{*,PG}(r_{NT}(\bar{Y}_{NT,b}^{*,PG} - \mathbb{E}[Y_{it}])) - \mathcal{L}_0(\mathbf{q}^{PG}, \mathbf{c}, \varrho)\|_\infty \to 0 \tag{A.3}$$

where the simulation algorithm estimates the law $\mathbb{P}_{NT}^{*,PG}$ consistently, and $\mathbf{q}^{PG}$ coincides with $\mathbf{q}$ if and only if $q_e + q_v = 0$.

Since convergence also holds along drifting sequences, we can adapt an argument from the proof of Theorem 1 in Andrews and Guggenberger (2010) to conclude that the asymptotic properties for the pigeonhole bootstrap are in fact uniform, see the Proof for Theorem 4.2 for details □

Appendix B. Extensions

This section gives various extensions to the baseline case. We first show how to apply our results to approximate joint distributions of means in several variables and when the statistic of interest is an estimator that is defined by potentially nonlinear moment conditions. We also consider inference in regression models. We furthermore consider non-exhaustively matched data, when not all of the $N \times T$ index pairs are observed, and the case in which the (i, t) index pairs correspond to clusters of more than one unit. We finally consider clustering across D rather than two dimensions, then problems in which data concerns outcomes at the level of a dyad or larger subgroup out of a sample of N "fundamental" units. Sample averages of that type are common in the analysis of network or matching data.

B.1. **Multivariate Case.** Another important extension concerns the case of the mean of a vector-valued array $(\mathbf{Y}_{it})$, where $\mathbf{Y}_{it} = (Y_{1it}, \ldots, Y_{Mit})' \in \mathbb{R}^M$, and the joint distribution of the components of $\mathbf{Y}_{it}$ is left unrestricted. This generalization is relevant for joint tests and estimators that are defined by a vector of estimating equations described in the next subsection below.

For this case, we can consider a component-wise Aldous-Hoover representation of the array

$$\mathbf{Y}_{it} = f(\boldsymbol{\mu}, \boldsymbol{\alpha}_i, \boldsymbol{\gamma}_t, \boldsymbol{\varepsilon}_{it})$$

Here μ, $\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_t, \boldsymbol{\varepsilon}_{it} \in \mathbb{R}^M$ are i.i.d., but the individual components of the vectors $\boldsymbol{\alpha}_i$, $\boldsymbol{\gamma}_t$, and $\boldsymbol{\varepsilon}_{it}$, respectively, may be dependent in an arbitrary fashion.

We can then implement the bootstrap algorithm from the baseline case jointly in all M components of the random vector $\mathbf{Y}_{it}$, where the projections $\hat{\mathbf{a}}_i$, $\hat{\mathbf{g}}_t$ and $\hat{\mathbf{w}}_{it}$ are M-dimensional vectors whose components are defined in analogy to the scalar case. The shrinkage parameters $\hat{\lambda}_1, \ldots, \hat{\lambda}_M$ are then computed component by component as in the univariate case.

We denote the respective rates for the individual components with $\mathbf{r}_{NT} = (r_{1NT}, \ldots, r_{MNT})'$, where $r_{mNT}^2 := \mathrm{Var}(\bar{Y}_{mNT})$, the variance of the mth component of the sample average $\bar{\mathbf{Y}}_{NT}$. We also denote slowest component of $\mathbf{r}_{NT}$ with $\varrho_{NT} := \max_{m=1,\ldots,M} |r_{mNT}|$. Then using the Cramér-Wold device, it follows immediately from Theorem 4.2 that the bootstrap remains consistent for approximating the joint distribution of $\mathrm{diag}(\mathbf{r}_{NT})(\bar{\mathbf{Y}}_{NT} - \mathbb{E}[\mathbf{Y}_{it}])$ if the conditions of that theorem hold for each component $m = 1, \ldots, M$. Similarly, a refinement at the ϱ_{NT}^{-2} rate is a straightforward extension of Theorem 4.3.

B.2. **Bootstrapping Estimators.** The bootstrap procedure developed for the distribution of the sample mean $\bar{Y}_{NT}$ can be used to estimate the distribution of potentially nonlinear estimators. Specifically, suppose that the estimand of interest is a parameter θ_0 in some parameter space $\Theta \subset \mathbb{R}^k$ which satisfies moment conditions of the form

$$\mathbb{E}[g(Y_{it}; \theta_0)] = 0$$

for a known function $g : \mathcal{Y} \times \Theta \to \mathbb{R}^m$. We can obtain a Z-estimator $\hat{\theta}$ for the parameter by solving m estimating equations of the form

$$0 = \hat{A}_{NT} \hat{g}_{NT}(\hat{\theta})$$

where we define $\hat{g}_{NT}(\theta) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} g(Y_{it};\theta)$, and $\hat{A}_{NT}$ is an $k \times m$ matrix which may depend on quantities estimated from the data with probability limit $\hat{A}_{NT} \xrightarrow{p} A_0$. If we denote the Jacobian of the population moment with $G_0 := \nabla_\theta \mathbb{E}[g(Y_{it};\theta_0)]$, under regularity conditions we have from standard arguments[11] that the estimator is asymptotically linear and satisfies the expansion

$$r_{NT}(\hat{\theta} - \theta_0) = -\left(A_0 G_0\right)^{-1} r_{NT}\hat{g}_{NT}(\theta_0) + o_p(1)$$

where r_{NT} is a rate such that the distribution of $r_{NT}\hat{g}_{NT}(\theta_0)$ is asymptotically tight.

Following the proposal by Kline and Santos (2012), we can obtain the bootstrap analog $\hat{g}^*_{NT}(\hat{\theta}) := \frac{1}{NT}\sum_{i=1}^{N} g^*_{it}$ by resampling from the $N \times T \times m$ array with entries $g_{it} := g(Y_{it};\hat{\theta})$ using the (multivariate version of the) algorithm from Section 3. We can then estimate the distribution of the estimator with

$$r_{NT}(\hat{\theta}^* - \hat{\theta}) := -\left(\hat{A}_{NT}\hat{G}_{NT}\right)^{-1} r_{NT}\hat{g}^*_{NT}(\hat{\theta})$$

where $\hat{G}_{NT} := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta g(Y_{it};\hat{\theta})$. It is important to note that refinements are generally only available if the estimating equations are linear in the parameter, so that the estimator can be represented as a smooth function of sample moments.

An important special case are method of moments estimators that match model predictions as a function of the unknown parameter $\pi : \Theta \to \mathbb{R}^M$ to the corresponding sample moments, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} g(Y_{it})$. In that case, we can directly bootstrap the joint distribution of the sample moment functions via

$$\hat{g}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} g(Y_{it}) - \pi(\theta)$$

Note that the resulting estimating equations are linear in the sample moments by construction, so that the bootstrap procedure immediately inherits the asymptotic properties from the bootstrap distribution for vectors of sample means, including refinements.

B.3. **Inference in Regression Models.** In a regression model

$$y_{it} = x_i'\beta_1 + z_t'\beta_2 + u_{it}, \quad \mathbb{E}[u_{it}|x_i, z_t] = 0$$

the researcher may be interested in inference conditional on the regressors x_i, z_t. In that case the assumption that the error u_{it} be separately exchangeable conditional on (x_i, z_t) is unreasonably strong, especially under potential misspecification of the regression function. However if $u_{it}|x_i, z_t$ is a.s. continuously distributed, the conditional integral transform

$$v_{it} := F_{u|x,z}(u_{it}|x_i, z_t)$$

follows a uniform distribution conditional on x_i, z_t and may be embedded into a separately exchangeable array. This gives us the Aldous-Hoover representation

$$u_{it} = F^{-1}_{u|x,z}(v_{it}|x_i, z_t) \equiv F^{-1}_{u|x,z}(\tilde{f}(\mu, \alpha_i, \gamma_t, \varepsilon_{it})|x_i, z_t) \equiv f_\mu(\alpha_i, x_i, \gamma_t, z_t, \varepsilon_{it})$$

where $\alpha_i, \gamma_t, \varepsilon_{it}$ are i.i.d. conditional on $\mathbf{x}_N := (x_1, \ldots, x_N)$ and $\mathbf{z}_T := (z_1, \ldots, z_T)$. We can therefore find an orthogonal decomposition of u_{it} conditional on $\mathbf{x}_N, \mathbf{z}_T$ that is analogous to that for Y_{it} in the unconditional case. Under the appropriate moment conditions for x_i, z_t we can then obtain the conditional limiting distribution of $\frac{r_{NT}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_i', z_t')'u_{it}$ given $\mathbf{x}_N, \mathbf{z}_T$ via a martingale CLT using an analogous argument as in the proof of Theorem 4.1.

[11] See e.g. Newey and McFadden (1994)

The corresponding bootstrap procedure holds x_i, z_t fixed for each bootstrap replication and resamples residuals u_{it}^* from an estimate of its conditional distribution. Specifically, we let $\hat{u}_{it} := Y_{it} - x_i'\hat{\beta}_1 - z_t'\hat{\beta}_2$,

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

We also compute $\hat{\lambda}_a := \frac{\hat{D}_a(\kappa_a)T\hat{\sigma}_a^2}{\hat{D}_a(\kappa_a)T\hat{\sigma}_a^2+\hat{\sigma}_w^2}$ and $\hat{\lambda}_g := \frac{\hat{D}_g(\kappa_g)N\hat{\sigma}_g^2}{\hat{D}_g(\kappa_g)N\hat{\sigma}_g^2+\hat{\sigma}_w^2}$ for the bootstrap with or without model selection, where $\hat{\sigma}_a^2, \hat{\sigma}_g^2, \hat{\sigma}_w^2$ are defined in an analogous fashion as in Section 3. We then generate the bth bootstrap sample according $Y_{it}^* := x_i'\hat{\beta}_1 + z_t'\hat{\beta}_2 + u_{it}^*$, where

$$u_{it}^* := \sqrt{\hat{\lambda}_a}\omega_{ai,b}\hat{a}_i + \sqrt{\hat{\lambda}_g}\omega_{gt,b}\hat{g}_t + \omega_{ai,b}\omega_{gt,b}\hat{w}_{it}$$

where $\omega_{ai,b}, \omega_{gt,b}$ are i.i.d. random variables with zero mean, and second and third moments equal to unity.

A proof of asymptotic validity of this bootstrap procedure closely follows the argument the unconditional case in Lemma C.2 and Theorem 4.2. The shrinkage strategy based on the *unconditional* variance ratios yields asymptotically valid inference despite the fact that the conditional variance ratios $T\text{Var}(a_i|x_i)/(T\text{Var}(a_i|x_i)+\text{Var}(w_{it}|x_i))$ and $N\text{Var}(g_t|z_t)/(N\text{Var}(g_t|z_t)+\text{Var}(w_{it}|z_t))$ need not be constant. Specifically, for sequences under which $T\sigma_a^2$ is bounded, $T\text{Var}(a_i|x_i)$ must also be bounded with probability approaching 1. On the other hand, if $T\sigma_a^2 \to \infty$, the ratio $T\text{Var}(a_i|x_i)/(T\text{Var}(a_i|x_i) + \text{Var}(w_{it}|x_i)) \to 1$ with strictly positive probability, so that the bootstrap procedure with $\lambda_a = 1$ yields a Gaussian limit with the correct asymptotic variance. The analogous conclusions hold with respect to bounded and divergent sequences of $N\sigma_g^2$. Finally, for unbounded sequences for $T\sigma_a^2$ and $N\sigma_g^2$ which do not converge to infinity, we can partition that sequence into a divergent and one bounded subsequence along each of which the bootstrap is asymptotically valid. In particular, the bootstrap with or without model selection inherit their pointwise (uniform given $q_v = 0$, respectively) consistency properties from the unconditional case.

B.4. **Non-Exhaustively Matched Samples.** We next consider the case in which Y_{it} is observed only for some, but not all index pairs (i,t). For example, units $i = 1,\ldots,N$ could be high school students, and $t = 1,\ldots,T$ teachers, and we observe student i's test score Y_{it} after being taught by teacher t. The process for assigning students and teachers to classrooms may be "blind" to student and teacher-level characteristics α_i or γ_t, or subject to sorting. E.g. a principal may assign a more talented teacher to a classroom of "weak" students. Endogenous sorting raises additional major conceptual and practical issues for identification and estimation, so for the remainder of this section we focus exclusively on the case of "exogenous" assignment, in a sense to be made more precise in Assumption B.1 (b) below.

We can formalize such a sampling scheme by defining an $N \times T$ matrix $\mathbf{W}$ of indicator variables, where W_{it} equals one if Y_{it} is observed for the dyad (i,t), and zero otherwise. We then consider the sampling distribution of

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

conditional on W_{it}. We also let

$$p_i := \frac{1}{T}\sum_{t=1}^{T}W_{it},\ p_t := \frac{1}{N}\sum_{i=1}^{N}W_{it},\ \text{and } \bar{p} := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}W_{it} = \frac{1}{N}\sum_{i=1}^{N}p_i = \frac{1}{T}\sum_{t=1}^{T}p_t$$

We then make the following assumptions:

Assumption B.1. *(a) As $N, T \to \infty$ sampling weights W_{it} are such that $\frac{1}{N}\sum_{i=1}^{N}(p_i/\bar{p})^2 \to \tau_a < \infty$ and $\frac{1}{T}\sum_{t=1}^{T}(p_t/\bar{p})^2 \to \tau_g < \infty$. (b) The random array can be represented as $Y_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it})$ for some function $h(\cdot)$, and random variables $\alpha_i, \gamma_t, \varepsilon_{it}$ that are i.i.d. conditional on W_{it}.*

Note that part (a) does not impose any restrictions on the density/sparseness of the sampling frame, but the assumption of finite limits τ_a, τ_g amounts to a balance requirement on relative cluster sizes in either dimension. In particular we allow for the case $\bar{p} \to 0$, but rule out the existence of individual clusters that dominate in size. Part(b) can be interpreted as a "no sorting" condition that is restrictive in many contexts in which the observable (i,t) pairs are the result of matching or self-selection of economic agents. This excludes cases with assortative matching on worker and firm productivity, or samples with students and teachers that are matched according to ability.

Given Assumption B.1, we find from elementary variance calculations that

$$\begin{aligned} r_{NT,W}^{-2} &:= \mathrm{Var}(\bar{Y}_{NT,W}) \\ &= \frac{1}{NT\bar{p}}\left(T\bar{p}\sigma_a^2\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{p_i}{\bar{p}}\right)^2\right] + N\bar{p}\sigma_g^2\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{p_t}{\bar{p}}\right)^2\right] + \sigma_w^2\right) \end{aligned} \tag{B.1}$$

From this expression, we can see that clustering on α_i and γ_t matters asymptotically if and only if $N\bar{p}\sigma_g^2 + T\bar{p}\sigma_a^2$ converges to a strictly positive limit. Cluster-level variation dominates the limiting distribution if $N\bar{p}\sigma_g^2 + T\bar{p}\sigma_a^2 \to \infty$.

By Assumption B.1 (b), $\mathbb{E}[\bar{Y}_{NT,W}|\mathbf{W}] = \mathbb{E}[Y_{it}|W_{it}] = \mathbb{E}[Y_{it}]$ a.s., so that our analysis of the asymptotic distribution will focus on the studentized mean $r_{NT}(\bar{Y}_{NT,W} - \mathbb{E}[Y_{it}])$.

We then consider the following bootstrap algorithm:

(a) Generate an exhaustively matched bootstrap sample Y_{it}^*, $i = 1, \ldots, N$, $t = 1, \ldots, T$ as in the baseline case with

$$\begin{aligned} \hat{\lambda}_a &:= \frac{\hat{D}_a(\kappa_a)T\bar{p}\hat{\sigma}_a^2\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{p_i}{\bar{p}}\right)^2\right]}{\hat{D}_a(\kappa_a)T\bar{p}\hat{\sigma}_a^2\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{p_i}{\bar{p}}\right)^2\right] + \bar{p}\hat{\sigma}_w^2} \\ \hat{\lambda}_g &:= \frac{\hat{D}_g(\kappa_g)N\bar{p}\hat{\sigma}_g^2\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{p_t}{\bar{p}}\right)^2\right]}{\hat{D}_g(\kappa_g)N\bar{p}\hat{\sigma}_g^2\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{p_t}{\bar{p}}\right)^2\right] + \bar{p}\hat{\sigma}_w^2}. \end{aligned}$$

where κ_a, κ_g are chosen according to whether the bootstrap is implemented with or without model selection. For the conservative bootstrap, $\hat{\lambda}_a, \hat{\lambda}_g$ are constructed in analogy to the description in Section 3.

(b) Keep the observations for which $W_{it} = 1$ and compute the bootstrapped mean

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

We can then show that under Assumptions 2.1 and B.1, the analogous conclusions to Theorems 4.2 and 4.3 hold for the modified bootstrap distribution:

Proposition B.1. ***(Bootstrap Consistency)*** *Suppose that Assumptions 2.1 and B.1 hold. Then the sampling distribution $\mathbb{P}_{NT}(r_{NT}(\bar{Y}_{NT,W} - \mathbb{E}[Y_{it}]))$ and the bootstrap distribution $\mathbb{P}_{NT}^*(r_{NT}(\bar{Y}_{NT,W}^* - \bar{Y}_{NT,W}))$*

converge in probability to the same limit,

$$\|\mathbb{P}^*_{NT}(r_{NT}(\bar{Y}_{NT,W} - \mathbb{E}[Y_{it}])) - \mathbb{P}_{NT}(r_{NT}(\bar{Y}^*_{NT,W} - \bar{Y}_{NT,W}))\|_\infty \xrightarrow{p} 0$$

where convergence is pointwise for the bootstrap with model selection. If $q_v = 0$, convergence is uniform for the bootstrap without model selection, for $q_v > 0$ the bootstrap without selection is inconsistent. The conservative bootstrap is consistent for the case $q_v + q_e = 0$, and conservative for the case $q_v + q_e > 0$.

See Appendix C for a proof. The only major complication arises if the second-order projection term $\frac{1}{NT\bar{p}^2}\sum_{i=1}^N\sum_{t=1}^T W_{it}v_{it}$ remains relevant in the limit. In that case, the terms $\frac{1}{NT\bar{p}}\sum_{i=1}^N\sum_{t=1}^T W_{it}\phi_k(\alpha_i)\psi_k(\gamma_t)$ of the sparse representation can in general no longer be represented in terms of separate sample averages of $\phi_k(\alpha_i)$ and $\psi_k(\gamma_t)$, respectively. Instead we use results on random quadratic forms by Götze and Tikhomirov (1999) to reach the analogous conclusions. For the case of a sparse sample, $\bar{p} \to 0$, Corollary 2 in Götze and Tikhomirov (1999) furthermore implies the stronger conclusion of asymptotic normality of $r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])$ even when $q_v > 0$. Finally, a straightforward adaptation of the arguments in the proof of Theorem 4.3 establishes refinements to the estimated percentiles for the case of non-exhaustively matched samples whenever $q_v = 0$.

B.5. **Unbalanced Cluster Sizes.** Suppose that we observe R_{it} i.i.d. units in the intersection of clusters i and t, denoted by $Y_{itr},\ r = 1, \ldots, R_{it}$. We consider inference for the pooled average

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

We also define $r_i := \frac{1}{T}\sum_{t=1}^T R_{it}$, $r_t := \frac{1}{N}\sum_{i=1}^N R_{it}$, and $\bar{r} := \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T R_{it}$. Clearly, $\bar{r} = \frac{1}{N}\sum_{i=1}^N r_i = \frac{1}{T}\sum_{t=1}^T r_t$.

Note that for the case of equal-sized clusters, $R_{it} = R$, this problem is formally equivalent to clustering in three dimensions $i = 1, \ldots, N$, $t = 1, \ldots, T$, and $r = 1, \ldots, R$, where clustering in the third dimension is trivial, and the Aldous-Hoover representation is of the form

$$Y_{itr} = f(\alpha_i, \gamma_t, \varepsilon_{itr})$$

where $\alpha_i, \gamma_t, \varepsilon_{itr}$ are i.i.d. across all indices. Note that in the case of balanced cluster sizes, $R_{it} = R$ for all i, t, we can directly apply our results for the baseline case, where $Y_{it} := \frac{1}{R}\sum_{r=1}^R Y_{itr}$. The unbalanced case in which R_{it} varies across i, t requires additional assumptions under which we can adapt our approach for the case of non-exhaustively matched samples from the previous section. However, our results do not assume that R grows large.

For our results we assume that cluster size is independent of cluster effects α_i, γ_t, and that the imbalance in cluster size is bounded:

Assumption B.2. *(a) As $N, T \to \infty$ sampling weights R_{it} are such that $\bar{r} \to \infty$, $\frac{1}{N}\sum_{i=1}^N (r_i/\bar{r})^2 \to \varrho_a < \infty$ and $\frac{1}{T}\sum_{t=1}^T (r_t/\bar{r})^2 \to \varrho_g < \infty$. (b) The random array satisfies $Y_{it} = h(\alpha_i, \gamma_t, \varepsilon_{it})$, where $\alpha_i, \gamma_t, \varepsilon_{it}$ are i.i.d. conditional on R_{it}.*

Now let

$$\hat{a}_i := \frac{1}{Tr_t}\sum_{t=1}^T\sum_{r=1}^{R_{it}} Y_{itr} - \bar{Y}_{NT,R} \qquad \hat{g}_t := \frac{1}{Nr_t}\sum_{i=1}^N\sum_{r=1}^{R_{it}} Y_{itr} - \bar{Y}_{NT,R}$$

$$\hat{v}_{it} := \frac{1}{R_{it}}\sum_{r=1}^{R_{it}} Y_{itr} - \hat{a}_i - \hat{g}_t + \bar{Y}_{NT,R} \qquad \hat{e}_{itr} := Y_{itr} - \hat{a}_i - \hat{g}_t - \hat{v}_{it}$$

For our projection representation, $\hat{v}_{it}$ estimates the second projection term $\mathbb{E}[Y_{itr}|\alpha_i, \gamma_t]$, and $\hat{e}_{itr}$ may remain relevant for the limiting distribution as long as R does not grow too fast.

We then construct a bootstrap sample as follows:

(a) Generate $a_i^* := \hat{a}_{k(i)}$, $g_t^* := \hat{g}_{s(t)}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ where $k(i)$ and $s(t)$ drawn independently and uniformly at random from the index sets $\{1, \ldots, N\}$ and $\{1, \ldots, T\}$, respectively, and $v_{it}^* := \hat{v}_{k(i)s(t)}$ and $e_{itr}^* := \hat{e}_{k(i)s(t)q(r)}$ for $q(r)$ drawn independently and uniformly from $\{1, \ldots, R_{k(i),s(t)}\}$.

(b) Let $\omega_i, \omega_t, \omega_r$ be i.i.d draws from a distribution with mean zero, unit variance, and third moments equal to one for $i = 1, \ldots, N$, $t = 1, \ldots, T$, and $r = 1, \ldots, R$.

(c) Generate an $N \times T \times R$ array of bootstrap draws

$$Y_{itr}^* := \bar{Y}_{NT,R} + \sqrt{\hat{\lambda}_a} a_i^* + \sqrt{\hat{\lambda}_g} g_t^* + \omega_i \omega_t (\sqrt{\hat{\varrho}} v_{it}^* + \omega_r e_{itr}^*)$$

where $\hat{\varrho} := \frac{\bar{r}\hat{\sigma}_v^2}{\bar{r}\hat{\sigma}_v^2 + \hat{\sigma}_e^2}$ and

$$\begin{aligned} \hat{\lambda}_a &:= \frac{\hat{D}_a(\kappa_a) T \bar{r} \hat{\sigma}_a^2 \left[\frac{1}{N}\sum_{i=1}^N \left(\frac{r_i}{\bar{r}}\right)^2\right]}{\hat{D}_a(\kappa_a) T \bar{r} \hat{\sigma}_a^2 \left[\frac{1}{N}\sum_{i=1}^N \left(\frac{r_i}{\bar{r}}\right)^2\right] + \bar{r}\hat{\sigma}_w^2} \\ \hat{\lambda}_g &:= \frac{\hat{D}_g(\kappa_g) N \bar{r} \hat{\sigma}_g^2 \left[\frac{1}{T}\sum_{t=1}^T \left(\frac{r_t}{\bar{r}}\right)^2\right]}{\hat{D}_g(\kappa_g) N \bar{r} \hat{\sigma}_g^2 \left[\frac{1}{T}\sum_{t=1}^T \left(\frac{r_t}{\bar{r}}\right)^2\right] + \bar{r}\hat{\sigma}_w^2}. \end{aligned}$$

where κ_a, κ_g are chosen according to whether the bootstrap is implemented with or without model selection. For the conservative bootstrap, $\hat{\lambda}_a, \hat{\lambda}_g$ are again constructed in analogy to the description in Section 3.

Under Assumptions 2.1 and B.2, the analogous conclusions to Theorems 4.2 and 4.3 regarding bootstrap consistency and refinements hold for the modified bootstrap procedure after only minor modifications of the arguments in Theorem B.1.

B.6. **Clustering in D Dimensions.** The bootstrap procedure can be immediately extended to the case of an array $(Y_{i_1 \ldots i_D} : i_1 = 1, \ldots, N_1, \ldots, i_D = 1, \ldots, N_D)$ that may exhibit clustering in D dimensions. As in the benchmark case, we assume that the sampling units corresponding to the indices in each dimension are i.i.d. draws from a common distribution so that for the dth dimension the "sheets" of the form $\left(Y_{i_1 \ldots i_{d-1} j i_{d+1} \ldots i_D} : i_{d'} = 1, \ldots, N_{d'}, d' \neq d\right)$ are identically distributed for each $j = 1, \ldots, N_d$ and $d = 1, \ldots, D$.

Such an array is separately exchangeable, and the main result by Hoover (1979) (see also Corollary 7.23 in Kallenberg (2005)) implies that it can be represented as

$$Y_{i_1, \ldots, i_D} = f(\mu, \alpha_{1i_1}^{(1)}, \alpha_{2i_2}^{(1)}, \ldots, \alpha_{d_1 \ldots d_k i_1 \ldots i_k}^{(k)}, \ldots, \alpha_{1 \ldots D i_1 \ldots i_D}^{(D)})$$

for some function $f(m, a_1^{(1)}, a_2^{(1)} \ldots, \alpha_{1 \ldots D}^{(D)})$, where μ, $\alpha_{1i_1}^{(1)}, \ldots, a_{1 \ldots D i_1 \ldots i_D}^{(D)}$ are i.i.d. draws from the uniform distribution for $i_d = 1, \ldots, N_d$ and $d = 1, \ldots, D$. As in the leading case, we consider inference with respect to the conditional mean of $Y_{i_1 \ldots i_D}$ given μ.

This case is therefore conceptually analogous to the two-dimensional case, but we need to keep track of a larger number of terms in an orthogonal projection onto subsets of the D dimensions. For more compact notation, we let $N_{\mathbf{d}}^{(k)} := \prod_{d=1}^D N_d / \prod_{l=1}^k N_{d_l}$ for any k-variate multi-index $\mathbf{d} = (d_1, \ldots, d_k)$. In particular, $N_{()}^{(0)} = \prod_{d=1}^D N_d$.

We can then adapt the bootstrap procedure from section 3 in the following manner: For $k = 0, 1, \ldots, D$ we then recursively define projections of the array on the k dimensions $d_1, \ldots, d_k$,

$$\hat{a}^{(0)} := \frac{1}{N^{(0)}_{()}} \sum_{i_1,\ldots,i_D} Y_{i_1 \ldots i_D} =: \bar{Y}_{N_1 \ldots N_D}$$

and for any multi-indices $\mathbf{d} := (d_1, \ldots, d_k)$ and $\mathbf{i} = (i_{d_1}, \ldots, i_{d_k})$, let

$$\hat{a}^{(k)}_{\mathbf{di}} := \frac{1}{N^{(k)}_{\mathbf{d}}} \sum_{i_{d'}: d' \notin \{d_1,\ldots,d_k\}} Y_{i_1,\ldots,i_D} - \sum_{k'=0}^{k-1} \sum_{\mathbf{d}' \in D(k')} \hat{a}^{(k-1)}_{\mathbf{d}' i_{d'_1} \ldots i_{d'_{k'}}}$$

where $D(k')$ consists of the $\binom{k}{k'}$ subsets of $\{d_1, \ldots, d_k\}$ of size k'. In particular, the projection residual

$$\hat{a}^{(D)}_{i_1 \ldots i_D} := Y_{i_1 \ldots i_D} - \sum_{k=0}^{D-1} \sum_{\mathbf{d}' \in D(k)} \hat{a}^{(k-1)}_{\mathbf{d}' i_{d'_1} \ldots i_{d'_k}}$$

As in the two-dimensional case, we let $\hat{\sigma}^2_{a^{(k)}_{\mathbf{d}}}$ be the respective bias-corrected empirical variances of these components. In order to select the asymptotically relevant projection terms, for each multi-index $\mathbf{d} = (d_1, \ldots, d_k)$ we also define the selector $\hat{D}_{a^{(k)}_{\mathbf{d}}}(\kappa) := \mathbb{1}\left\{ N^{(k)}_{\mathbf{d}} \hat{\sigma}^2_{a^{(k)}_{\mathbf{d}}} \geq \kappa \right\}$ and sequences $\kappa_{a^{(k)}_{\mathbf{d}}}$ that grow to infinity at a slow rate in $\min\{N_{d_1}, \ldots, N_{d_k}\}$.

For $d \in \{1, \ldots, D\}$, we then draw $a^{(1)*}_d$ independently from the empirical distribution for $\hat{a}^{(1)}_d$, and for each $k = 1, \ldots, D-1$ and $d_1, \ldots, d_k \in \{1, \ldots, D\}$ we let $a^{(k)*}_{d_1 \ldots d_k i_{d_1} \ldots i_{d_k}} := \hat{a}^{(k)}_{d_1 \ldots d_k j^*_1(i_1) \ldots j^*_k(i_k)} \left(\prod_{l=1}^k \omega_{d_l i_{d_l}} \right)$ for independent draws $\omega_{d i_d}$ from the same distribution as in the baseline case. As before, $j^*_d(i_d)$ denotes the index of the cross-sectional unit corresponding to the i_dth bootstrap draw for dimension d. We then form

$$Y^*_{i_1 \ldots i_D} := \bar{Y}_{N_1 \ldots N_D} + \sum_{k=1}^{D} \sum_{\mathbf{d}' \in D(k)} \sqrt{\hat{\lambda}_{\mathbf{d}' k}} a^{(k-1)*}_{\mathbf{d}' i_{d'_1} \ldots i_{d'_{k'}}}$$

where for the bootstrap with and without model selection,

$$\hat{\lambda}_{\mathbf{d}' k} := \frac{\hat{D}_{a^{(k)}_{\mathbf{d}}}(\kappa_{a^{(k)}_{\mathbf{d}}}) N^{(k)}_{\mathbf{d}} \hat{\sigma}^2_{a^{(k)}_{\mathbf{d}'}}}{\hat{D}_{a^{(k)}_{\mathbf{d}}}(\kappa_{a^{(k)}_{\mathbf{d}}}) N^{(k)}_{\mathbf{d}} \hat{\sigma}^2_{a^{(k)}_{\mathbf{d}'}} + \hat{\sigma}^2_{a^{(D)}_{1 \ldots D}}}$$

is defined in analogy to the two-dimensional case. In particular, for each $\mathbf{d}^{(k)}$ we choose $\kappa_{a^{(k)}_{\mathbf{d}}}$ according to slowly increasing sequences for the bootstrap with model selection, and $\kappa_{a^{(k)}_{\mathbf{d}}} = 0$ for the bootstrap without model selection. For the conservative bootstrap, we set

$$\hat{\lambda}_{\mathbf{d}' k} := \frac{\hat{q}_{a^{(k)}_{\mathbf{d}}}}{\hat{q}_{a^{(k)}_{\mathbf{d}}} + \hat{\sigma}^2_{a^{(D)}_{1 \ldots D}}}, \quad \text{where } \hat{q}_{a^{(k)}_{\mathbf{d}}} := \max \left\{ \kappa_{a^{(k)}_{\mathbf{d}}}, N^{(k)}_{\mathbf{d}} \hat{\sigma}^2_{a^{(k)}_{\mathbf{d}'}} \right\}$$

We can then compute the bootstrapped mean $\bar{Y}^*_{N_1 \ldots N_D} := \frac{1}{N^{(0)}_{()}} \sum_{i_1 \ldots i_D} Y^*_{i_1 \ldots i_D}$ or its studentization for the pivotal bootstrap.

Noting that the arguments behind Theorems 4.2 and 4.3 do not rely on the assumption that the random array is two-dimensional, an extension of these results to the D-dimensional case requires only a few minor notational changes.

B.7. **Dyadic and D-adic Data.** The results in this paper readily extend to the case of dyadic or network data, where we observe a D-dimensional array $(Y_{i_1 \ldots i_D} : i_1, \ldots, i_D = 1, \ldots, N)$ whose distribution is invariant to permutations $\pi : \{1, \ldots, N\} \to \{1, \ldots, N\}$, that is $Y_{i_1 \ldots i_D} \stackrel{d}{=} Y_{\pi(i_1) \ldots \pi(i_D)}$. Using the terminology of

Kallenberg (2005), such an array is jointly exchangeable and can be represented as

$$Y_{i_1\ldots i_D} = f(\mu, \alpha^{(1)}_{i_1}, \alpha^{(1)}_{i_2}, \ldots, \alpha^{(k)}_{i_1\ldots i_k}, \ldots, \alpha^{(D)}_{i_1\ldots i_D})$$

and μ, $\alpha^{(1)}_{i_1}, \alpha^{(1)}_{i_2}, \ldots, \alpha^{(D)}_{i_1\ldots i_D}$ are i.i.d. uniform random variables for $i_1, \ldots, i_D \in \{1, \ldots, N\}$ (see Hoover (1979) or Theorem 7.22 in Kallenberg (2005)). Conditional on μ, we then consider the sampling distribution of the "D-adic" mean

$$\bar{Y}_{N,D} := \frac{1}{N^D} \sum_{i_1,\ldots,i_D=1}^{N} Y_{i_1\ldots i_D}$$

for N units drawn at random from a larger population (with replacement) or distribution.[12]

Example B.1. ***Subgraph Counts.*** *Suppose that the adjacency matrix with entries $G_{ij} \in \{0,1\}^{N^2}$ represents the subgraph among the set of nodes $1, \ldots, N$ drawn at random from an infinite directed graph. Then the sampling distribution for the density of network homomorphisms (adjacency-preserving maps, see Lovasz (2012)) with respect to a network F among D distinct nodes can be approximated using this bootstrap procedure in the following way: We can define an indicator $R_{i_1\ldots i_D}(F)$ that equals 1 if there is an adjacency-preserving map between F and the subnetwork among the nodes $i_1, \ldots, i_D$. We can then re-sample from the D-dimensional array with entries $Y_{i_1\ldots,i_D} := R_{i_1\ldots i_D}(F)$ using the algorithm described above, where in step (b) we draw N row identifiers with replacement at random and select columns and other dimensions of the array corresponding to the same identifiers.*

We can implement each of the three bootstrap procedures (with and without model selection and conservative bootstrap) for D-adic arrays by following the algorithm as described in Sections 3 and B.6 except that in step (b) we draw N row identifiers with replacement at random and select columns and other dimensions of the array corresponding to the same identifiers. The proofs of Theorems 4.2 and 4.3 then go through under analogous conditions as for the original case.

Appendix C. 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

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

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

[12]Note that the case in which we only include D-ads of D or fewer distinct indices in the average is nested in this formulation, potentially after rescaling the mean by a bounded sequence.

$\hat{Z}^{\phi}_{Nk'}$ ($\hat{Z}^{\psi}_{Tk}$ 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.

We can stack these sample moments

$$\hat{Z}_{NT,K} := \left(\hat{Z}^{e}_{NT}, \hat{Z}^{a}_{N}, \hat{Z}^{g}_{T}, \hat{Z}^{\phi}_{N1}, \hat{Z}^{\psi}_{T1}, \ldots, \hat{Z}^{\phi}_{NK}, \hat{Z}^{\psi}_{TK}\right).$$

Now, let the sigma-fields $\mathcal{F}_{it} := \sigma\left(\{\alpha_j, \gamma_s, \varepsilon_{js} : j = 1, \ldots, i; s = 1, \ldots, t\}\right)$, and define the filtration $\mathcal{F}_s := \mathcal{F}_{i_s,s}$ for each $s = 1, 2, \ldots$, where $i_s := [Ns/T]$. Then each component of $\hat{Z}^{e}_{NT,K}$ is a martingale adapted to $\mathcal{F}_T$, so that by a CLT for martingale difference sequences and the Cramér-Wold device,

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

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

$$r_{NT}\left(\bar{Y}_{NT,K} - \mathbb{E}[Y_{it}]\right) \quad = \quad \hat{Z}^{a}_{N} + \hat{Z}^{g}_{T} + \hat{Z}^{e}_{NT} + \varrho_{NT} \sum_{k=1}^{K} c_k \hat{Z}^{\phi}_{Nk} \hat{Z}^{\psi}_{Tk}$$

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

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

with the coefficients c_k potentially variying along the limiting sequence, $Z^{\phi}_1, Z^{\psi}_1, \ldots, Z^{\phi}_K, Z^{\psi}_K$ are i.i.d. standard normal random variables, and Z^a, Z^g are standard normal random variables with $\mathrm{Cov}(Z^a, Z^{\phi}_k) = q_{ak}/\sqrt{q_a}$, $\mathrm{Cov}(Z^g, Z^{\psi}_k) = q_{gk}/\sqrt{q_g}$, $\mathrm{Cov}(Z^a, Z^g) = \mathrm{Cov}(Z^a, Z^{\psi}_k) = \mathrm{Cov}(Z^g, Z^{\phi}_k) = 0$ for all $k = 1, 2, \ldots$.

Finally, notice 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, where the magnitude of the approximation error can be controlled uniformly under Assumption 2.2, establishing claims (a) and (b) □

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

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

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

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

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 C.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: defining $\tilde{\phi}_{ik} := \phi_k(\alpha_i) - \mathbb{E}[\phi_k(\alpha_i)]$ we can write

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

Here, $\hat{Z}_{Nkk'}^{\phi\phi} = \frac{1}{\sqrt{N}}\sum_{i=1}^N(\tilde{\phi}_{ik}\tilde{\phi}_{ik'} - \mathbb{E}[\tilde{\phi}_{ik}\tilde{\phi}_{ik'}])$, where $\mathbb{E}[\tilde{\phi}_{ik}\tilde{\phi}_{ik'}]$ 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

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

noting that by construction $\mathbb{E}\left[a_i\tilde{\phi}_{ik}\right] = 0$ 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 (C.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 C.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 $\text{Var}\left(r_{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])\right) = 1$, this corollary is an immediate consequence of the convergence rates in Lemma C.1. In particular, if $\sigma_a^2 = 0$, Lemma C.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 C.1.

Proof of Proposition 4.1: Part (b) of Lemma C.1 implies that along sequences $\sigma_a^2, \sigma_g^2, \sigma_w^2$ with $\lim_T T\sigma_a^2 = q_a$, $\lim_N \sigma_g^2 = q_g$ and $\lim_{N,T} \sigma_w^2 = q_v + q_e$. Therefore, the asymptotic variance of the sample mean

$$\lim_{N,T} \text{Var}\left(\sqrt{NT}(\bar{Y}_{NT} - \mathbb{E}[Y_{it}])\right) = \lim_{N,T}\left(T\sigma_a^2 + N\sigma_g^2 + \sigma_w^2\right) = q_a + q_g + q_v + q_e$$

cannot be estimated consistently unless $q_v = 0$ or $q_a = q_g = 0$. If the asymptotic variance cannot be estimated consistently along a particular parameter sequence, it follows in particular that the asymptotic distribution of $\bar{Y}_{NT}$ cannot be estimated uniformly consistently, establishing the claim □

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

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

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, $0 < K < \infty$,

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

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

$$\hat{Z}^*_{K,NT} := (\hat{Z}^{e,*}_{NT}, \hat{Z}^{a,*}_{N}, \hat{Z}^{g,*}_{T} \hat{Z}^{\phi,*}_{N1}, \hat{Z}^{\psi,*}_{T1}, \dots, \hat{Z}^{\phi,*}_{NK}, \hat{Z}^{\psi,*}_{TK})'$$

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 C.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: By Assumption 2.1, the third conditional moments of $\hat{a}_i, \hat{\gamma}_t, \hat{e}_{it}$ given $(Y_{it} : i = 1, \dots, N, t = 1, \dots, T)$ are almost surely bounded, so that from the same argument as in the proof of Theorem 1 in Liu (1988), the Berry-Eséen theorem together with the Cramér-Wold device implies a joint CLT for the bootstrap processes,

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

conditional on $(Y_{it})_{i=1,\dots,N t=1,\dots,T}$ almost surely. Here, $Q^*_{NT,K}$ is a $(2K+3) \times (2K+3)$ 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. For $k = 1, \dots, K$ the entries of Q corresponding to covariances between $\hat{a}_i$ and $\phi_k(\alpha_i)$ equal $q^*_{ak,NT}$, and the covariances between $\hat{g}_t$ and $\psi_k(\gamma_t)$ are equal to $q^*_{gk,NT}$. All other off-diagonal entries of $Q^*_{NT,K}$ converge almost surely to zero.

Rewriting (C.2), we obtain

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

and it follows from the joint CLT and the continuous mapping theorem 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 (C.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. By Lemma C.1 (a), $\mathbf{q}^*_{NT,K} - \mathbf{q}_{NT,K} \xrightarrow{p} 0$, 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. Hence together with the continuous mapping theorem, Lemma C.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 standard approximation arguments 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 C.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 C.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

$$\bar{\mathbf{q}}_{NT} := (q_{e,NT}, q_{a,NT} + q_{e,NT} + q_{v,NT}, q_{g,NT} + q_{e,NT} + q_{v,NT}, q_{a1,NT}, q_{g1,NT}, \dots),$$

and denote the subvector consisting of the first $2K + 3$ components of $\bar{\mathbf{q}}_{NT}$ with $\bar{\mathbf{q}}_{NT,K}$. Lemma C.1 (a) implies that $\mathbf{q}^*_{NT,K} - \bar{\mathbf{q}}_{NT,K}$ converges in probability to zero for each $K < \infty$, 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 C.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 $\|\tilde{\mathbf{c}}_{NT,K}\|^2 \leq \sum_{k=1}^{K} \tilde{c}_k^2 < \infty$ 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 Theorem 4.3. We can establish the refinements of this bootstrap procedure by verifying the conditions for part (ii) of the main theorem in chapter 5 of Mammen (1992).

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 processes

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

we have that

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

This amounts to establishing condition $DIFF_T(3, C)$ in Mammen (1992) for the process $\hat{W}^{a,*}_N$. Verifying the conditions $DIFF_S(2)$ and $VAR(2)$ follows similar steps and is more standard. Note that by inspection of the expression for $\mathbb{E}[\hat{a}_i^3]$, the conclusion does not hold in general under arbitrary drifting sequences for the second and third moments of a_i, w_{it}. Using the same arguments, we can establish conditions $DIFF_T(3, C)$, $DIFF_S(2)$, and $VAR(2)$ for $\hat{W}^{g,*}_T := \frac{1}{\sqrt{T}}\sum_{t=1}^{T} g^*_{t,b}$ at the respective rates in T.

For the analogous results for the component $\hat{W}^{e,*}_{NT} := \frac{1}{\sqrt{NT}}\sum_{t=1}^{T} e^*_{it,b}$, note that by assumption $\mathbb{E}[\omega_i^3] = \mathbb{E}[\omega_t^3] = 1$ and the draws are independent, so that that $\mathbb{E}[(\omega_i\omega_t)^3] = 1$. Hence, the third moment of e^*_{it} under the bootstrap distribution also converges in probability to the third moments of $e_{it}, \phi^k(\alpha_i), \psi^k(\gamma_t)$ under the sampling distribution, using standard arguments analogous to the previous case. In particular, conditions $DIFF_T(3, C)$, $DIFF_S(2, C)$ and $VAR(2)$ in Mammen (1992) hold for $\hat{Z}^{e,*}_{NT}$ at the respective rates in NT. Furthermore, convergence in each of finitely many components implies joint convergence of cumulants for all three components. Since we only consider pointwise convergence for cases with $q_v > 0$ the contribution of the Wiener chaos component is asymptotically negligible.

By construction, $\hat{W}^a_N$ and $\hat{W}^g_T$ and their bootstrap versions $\hat{W}^{a,*}_N$ and $\hat{W}^{g,*}_T$ are independent. Also, the components of $\hat{W}^a_N, \hat{W}^g_T, \hat{W}^e_{NT}$ as well as their bootstrap analogs are asymptotically uncorrelated. For third cumulants of weighted sums of $\hat{Z}^a_N$ and $\hat{Z}^e_{NT}$ we also need to consider the moments

$$\mathbb{E}[\hat{a}_i\hat{w}_{it}^2] = \mathbb{E}[a_i w_{it}^2](1 + O(1/N))$$

where $\mathbb{E}^*_{NT}[a_i^*(w_{it}^*)^2] - \mathbb{E}[\hat{a}_i \hat{w}_{it}^2] = O_P(N^{-1/2})$ by an analogous argument as for the third moment of $a_{i,b}^*$. By similar arguments as for the third moments of a_i and g_t, for any weights $s_1, s_2, s_3 \geq 0$, we then have

$$\mathbb{E}^*_{NT}\left[\left(s_1\hat{W}_N^{a,*} + s_2\hat{W}_T^{g,*} + s_3\hat{W}_{NT}^{e,*}\right)^3\right] - \mathbb{E}\left[\left(s_1\hat{W}_N^{a} + s_2\hat{W}_T^{g} + s_3\hat{W}_{NT}^{e,*}\right)^3\right] \quad = \quad O_P(N^{-1})$$

with the analogous conclusion for weighted sums of $\hat{Z}_T^g$ and $\hat{Z}_{NT}^e$ and their bootstrap analogs.

Since under $q_v = 0$, $\hat{\lambda}_a \xrightarrow{p} \lambda_a$ and $\hat{\lambda}_g \to \lambda_g$ pointwise for the bootstrap with or without model selection, we can combine convergence of the cumulants of the joint distribution of the individual components to verify that the conditions $DIFF_T(3, C)$, $DIFF_S(2)$, and $VAR(2)$ also hold for the weighted sums with rates in N if $\sigma_a > 0$ (T, respectively, if $\sigma_g > 0$), or NT if $\sigma_a = \sigma_g = 0$ and $\sigma_e > 0$, so that the conclusion follows from the main theorem in chapter 5 of Mammen (1992). The analogous conclusions hold for the conservative bootstrap only if $q_e + q_v = 0$ □

C.1. **Proof of Proposition B.1:** The main arguments from the Proof of Theorem 4.2 hold after a few minor modifications of the arguments for the case $q_v = 0$. The only major complication arises if the second-order projection term $\frac{1}{NT\bar{p}^2}\sum_{i=1}^N\sum_{t=1}^T W_{it}v_{it}$ is of first order as we take limits. In that case, the terms $\frac{1}{NT\bar{p}}\sum_{i=1}^N\sum_{t=1}^T W_{it}\phi_k(\alpha_i)\psi_k(\gamma_t)$ of the sparse representation can in general no longer be represented in terms of separate sample averages of $\phi_k(\alpha_i)$ and $\psi_k(\gamma_t)$, respectively.

We first consider the case of dyadic data, where the components of the second-order projection term takes the form

$$Q_k := \frac{1}{N^2\bar{p}}\sum_{i=1}^N\sum_{j=1}^N W_{it}\phi_k(\alpha_i)\phi_k(\alpha_j) = \frac{1}{N^2\bar{p}}\phi_k' W \phi_k = \frac{1}{2N^2\bar{p}}\phi_k'(W + W')\phi_k$$

for the vector $\phi_k := (\phi_k(\alpha_1), \ldots, \phi_k(\alpha_N))'$. To characterize the limit distribution for $N\sqrt{p}Q_k$, let $Z_k \sim N(0, I_N)$ and $\tilde{Q}_k := \frac{1}{2N^2\bar{p}}Z_k'(W + W')Z_k$. Conditions for convergence of $N\sqrt{p}Q_k$ to $N\sqrt{p}\tilde{Q}_k$ were given by Götze and Tikhomirov (1999), noting that the matrix $W + W'$ is symmetric.

Now, by Assumption B.1 (a), we either have that $\sup_{i=1,\ldots,N} p_i \to 0$, or that $\lim_N \bar{p} > 0$. Hence we only need to distinguish two cases regarding the asymptotic behavior of p_i. For the first case with $\sup_{i=1,\ldots,N} p_i \to 0$, Corollary 2 in Götze and Tikhomirov (1999) implies that

$$\varrho(N\sqrt{p}Q_k, N\sqrt{p}\tilde{Q}_k) \leq (\mathbb{E}|\phi_k(\alpha_i)|^3)^2 \sup_{i=1,\ldots,N}\sqrt{p_i}$$

where $\varrho(X, Y) := \sup_x |F_X(x) - F_Y(x)|$ for any two random variables X, Y with respective c.d.f.s F_X and F_Y. Furthermore, in this case the asymptotic distribution of $N\sqrt{p}Q_k$ is Gaussian. By an analogous argument, we also find that the distribution of the bootstrap analog $N\sqrt{p}Q_k^*$ converges to $N\sqrt{p}\tilde{Q}_k$, so that bootstrap consistency follows from the triangle inequality. For the second case with $\bar{p}$ bounded away from zero, p_i is bounded away from zero by a constant for at least two distinct units in $\{1, \ldots, N\}$. In that case, consistency follows instead from Theorem 3 in Götze and Tikhomirov (1999).

An extension to multilinear forms for the case in which each dimension of the random array corresponds to a different type of sampling unit can be obtained in a straightforward manner after stacking the random variates $\phi_k(\alpha_1), \ldots, \phi_k(\alpha_N), \psi_k(\gamma_1), \ldots, \psi_k(\gamma_T)$ and considering the symmetric quadratic form corresponding to the $(N + T) \times (N + T)$ matrix $A = \frac{1}{2}[0, W; W'0]$ □