

SUPPLEMENT TO “BOOTSTRAP WITH CLUSTER-DEPENDENCE IN TWO OR MORE DIMENSIONS”

(*Econometrica*, Vol. 89, No. 5, September 2021, 2143–2188)

KONRAD MENZEL

Department of Economics, New York University

This supplement contains extensions of the results in Menzel (2021) as well as an asymptotic theory for alternative inference procedures under multi-way clustering.

APPENDIX B: ALTERNATIVE INFERENCE PROCEDURES

THIS SECTION gives asymptotic results for alternative methods of estimating the asymptotic distribution of the sample mean \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’s (2007) pigeonhole bootstrap.

B.1. *Summary of Asymptotic Properties*

The starting point for our analysis was the 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} . The alternative versions of the bootstrap procedure proposed in Section 3 therefore either do not aim for full uniformity or are conservative.

The recommendation of 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)** Pointwise validity with respect to the variance parameters, where we allow for any of the components of σ_a^2 , σ_g^2 , σ_v^2 , σ_e^2 to be either strictly positive or zero.
- **(UNIF-1)** Uniform validity regarding clustering in means, where any of the components of σ_a^2 , σ_g^2 , σ_v^2 , σ_e^2 may be strictly positive, zero, or drifting along sequences, but $r_{NT}\sigma_v^2 \not\rightarrow 0$. 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 σ_a^2 , σ_g^2 , σ_v^2 , σ_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

Konrad Menzel: km125@nyu.edu

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

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.¹ This scenario is therefore non-generic once we allow for any type of cluster-dependence, and we find that 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 pointwise consistency is achieved by subsampling with model selection (see Proposition B.2 below), the bootstrap with model selection, and the pivotal bootstrap with model selection (Theorem 4.2 in the main text), where the pivotal bootstrap with model selection achieves refinements in the non-degenerate case (Proposition 4.2), 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 (Proposition B.1).

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 overestimates 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 B.I. In addition to the different versions of the bootstrap introduced in Section 3, BS-N, BS-S, and BS-C, this section gives asymptotic results for the following additional methods:

TABLE B.I

SUMMARY OF ESTIMATION APPROACHES FOR THE ASYMPTOTIC DISTRIBUTION OF \bar{Y}_{NT} , WHERE “CONS.” STANDS FOR “CONSERVATIVE”

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

¹More precisely, $r_{NT}\sigma_v^2 \rightarrow 0$ 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 $\text{Var}(\mathbb{E}[Y_{it}|\gamma_t])$.

- **(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 $r_{\text{NT}}\bar{Y}_{\text{NT}}$, and
- **(SUB)** inference based on the subsampling estimate for the asymptotic distribution of $r_{\text{NT}}\bar{Y}_{\text{NT}}$.

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

To highlight some of the main theoretical findings, we find that the “default” estimator from Cameron, Gelbach, and Miller (2011) for the asymptotic variance, $\hat{S}_{\text{NT,def}}^2$, is only consistent if $r_{\text{NT}}\sigma_v^2 \rightarrow 0$, whereas the modified estimator $\hat{S}_{\text{NT,sel}}^2$ is always pointwise consistent. Gaussian “plug-in” inference with a consistent estimator for the asymptotic variance is only consistent if $r_{\text{NT}}\sigma_v^2 \rightarrow 0$, 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 asymptotically valid pointwise, and the bootstrap without model selection is uniformly valid as long as $r_{\text{NT}}\sigma_v^2 \rightarrow 0$. The pigeonhole bootstrap is uniformly valid asymptotically but conservative in the degenerate case, and in addition, its pivotal version 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 $r_{\text{NT}}^{-2/3}$, 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 pointwise bootstrap, or the r_{NT}^{-2} rate for the cases for which the pivotal bootstrap yields a refinement. This comparison of theoretical properties is also illustrated in a simulation study in Section 5.

B.2. 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 $\hat{S}_{\text{NT,def}}^2$ and $\hat{S}_{\text{NT,sel}}^2$ introduced in Section 3.

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

$$Y_{it} = \alpha_i \gamma_t, \quad \alpha_i, \gamma_t \stackrel{\text{iid}}{\sim} N(0, 1).$$

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

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

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

$$\frac{T}{N} \sum_{i=1}^N (\bar{Y}_{iT} - \bar{Y}_{\text{NT}})^2 = \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \gamma_t\right)^2 \frac{1}{N} \sum_{i=1}^N (\alpha_i - \bar{\alpha}_N)^2,$$

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

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

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

where Y_1, Y_2 are independent draws from a chi-squared 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 nonzero 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 $\hat{S}_{NT,\text{def}}$ 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 $(\sqrt{q_e} Z^e + \sqrt{q_a} Z^a + \sqrt{q_g} Z^g) + \varrho V = \sqrt{1 - q_v} Z + \varrho V$, 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 Gaussian 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.

B.3. Pigeonhole Bootstrap (PGH)

We next consider Owen's (2007) "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 b th bootstrap iteration, generate random weights $M_{i,b}$ for each $i = 1, \dots, N$ ($M_{t,b}$, respectively, for $t = 1, \dots, T$) as i.i.d. draws from a binomial distri-

bution with N trials and success probability $\frac{1}{N}$ (T trials and success probability $\frac{1}{T}$, respectively).

(b) We then form the b th bootstrap mean

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

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

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

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

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

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 $\bar{Y}_{\text{NT}}^{*,\text{PG}}$ given the sample $(Y_{it})_{i,t}$ with

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

This is a modification of Owen's (2007) 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 B.1—Pigeonhole Bootstrap: *Suppose that Assumptions 2.1 and 2.2 hold. Then if $q_v = 0$, the pigeonhole bootstrap*

$$\left\| \mathbb{P}_{\text{NT}}^{*,\text{PG}}(r_{\text{NT}}(\bar{Y}_{\text{NT}}^{*,\text{PG}} - \bar{Y}_{\text{NT}})) - \mathbb{P}_{\text{NT}}(r_{\text{NT}}(\bar{Y}_{\text{NT}} - \mathbb{E}[Y_{i1}])) \right\|_{\infty} \xrightarrow{p} 0$$

uniformly, where $\mathbb{P}_{\text{NT}}^{\text{PG}}$ is the convolution of the sampling distribution for $r_{\text{NT}}(\bar{Y}_{\text{NT}} - \mathbb{E}[Y_{it}])$ 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 pointwise 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 overestimates the contribution of the average $\frac{1}{\text{NT}} \sum_{i=1}^N \sum_{t=1}^T (v_{it} + e_{it})$ to the limiting distribution by a factor of 3, 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 pointwise consistency.²

PROOF OF PROPOSITION B.1: For the b th bootstrap replication, we can decompose the mean as

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

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 $\bar{\psi}_{kT} := \frac{1}{T} \times \sum_{t=1}^T \psi_k(\gamma_t)$. We can immediately verify that for the binomial distribution, $\mathbb{E}[M_i] = 1$ and $\mathbb{E}[M_i^2] = \mathbb{E}[M_i]^2 + \text{Var}(M_i) = 2 - \frac{1}{N}$. Similarly, $\mathbb{E}[M_{2t}] = 1$ and $\mathbb{E}[M_{2t}^2] = 2 - \frac{1}{T}$, where

²Specifically, 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 overestimate the scale of the asymptotic distribution in the degenerate case by $2\sigma_e^2$.

M_{11}, \dots, M_{1N} and M_{21}, \dots, M_{2T} are also independent. Hence, conditional on e_{11}, \dots, e_{NT} ,

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

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

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

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 $\frac{1}{\sqrt{T}} \sum_{t=1}^T M_{2t,b}(\psi_k(\gamma_t) - \bar{\psi}_{kT})$.

Next, we let $(\sigma_{a,\text{NT}}^{*,\text{PG}})^2 := \frac{1}{\text{NT}-1} \sum_{i=1}^N \sum_{t=1}^T (e_{it} - \bar{e}_{\text{NT}})^2$, $(\sigma_{a,\text{NT}}^{*,\text{PG}})^2 := \frac{1}{N-1} \sum_{i=1}^N (a_i - \bar{a}_N)^2$, $(\sigma_{g,\text{NT}}^{*,\text{PG}})^2 := \frac{1}{T-1} \sum_{t=1}^T (g_t - \bar{g}_T)^2$ be the empirical variances of the projection components. Similarly, for $k = 1, 2, \dots$, we define the empirical variances $(\sigma_{\phi_k,\text{NT}}^{*,\text{PG}})^2$, $(\sigma_{\psi_k,\text{NT}}^{*,\text{PG}})^2$ and covariances $\sigma_{ak,\text{NT}}^{*,\text{PG}}$, $(\sigma_{gk,\text{NT}}^{*,\text{PG}})$ 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 $q_{s,\text{NT}}^{*,\text{PG}} := r_{\text{NT}}(\sigma_{s,\text{NT}}^{*,\text{PG}})^2$ for $s = a, g, \phi 1, \psi 1, a 1, g 1, \dots$, and

$$\mathbf{q}_{\text{NT}}^{*,\text{PG}} := (q_{e,\text{NT}}^{*,\text{PG}}, q_{a,\text{NT}}^{*,\text{PG}}, q_{g,\text{NT}}^{*,\text{PG}}, q_{\phi 1,\text{NT}}^{*,\text{PG}}, q_{\psi 1,\text{NT}}^{*,\text{PG}}, q_{a 1,\text{NT}}^{*,\text{PG}}, q_{g 1,\text{NT}}^{*,\text{PG}}, \dots).$$

We also define its population analog

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

where $q_{s,\text{NT}}^{\text{PG}} = q_{s,\text{NT}}^{*,\text{PG}}$ for each $s = a 1, g 1, \dots$, $q_{\phi_k,\text{NT}}^{\text{PG}} = q_{\psi_k,\text{NT}}^{\text{PG}} = 1$ for each $k = 1, 2, \dots$, $q_{a,\text{NT}}^{\text{PG}} = q_{a,\text{NT}}^{*,\text{PG}} + q_{v,\text{NT}}^{\text{PG}}$, $q_{g,\text{NT}}^{\text{PG}} = q_{g,\text{NT}}^{*,\text{PG}} + q_{v,\text{NT}}^{\text{PG}}$, and $q_{e,\text{NT}}^{\text{PG}} = 3q_{e,\text{NT}}^{*,\text{PG}}$.

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

Next, along any convergent sequence $\mathbf{q}_{\text{NT}}^{\text{PG}} \rightarrow \mathbf{q}^{\text{PG}}$, we can apply a CLT to obtain a Gaussian joint asymptotic distribution for any finite subset of the averages

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T M_{1i,b}(a_i + \bar{v}_i), & \quad \frac{1}{T} \sum_{t=1}^T M_{2t,b}(g_t + \bar{v}_t), \\ \frac{1}{\text{NT}} \sum_{i=1}^N \sum_{t=1}^T M_{1i,b} M_{2t,b} e_{it}, & \quad \frac{1}{N} \sum_{i=1}^N M_{1i,b}(\phi_k(\alpha_i) - \bar{\phi}_{kN}), \quad \text{and} \\ \frac{1}{T} \sum_{t=1}^T M_{2t,b}(\psi_k(\gamma_t) - \bar{\phi}_{kT}) & \quad \text{for } k = 1, 2, \dots \end{aligned}$$

Also, $\frac{N_b^*}{N}, \frac{T^*}{T} \xrightarrow{p} 1$ 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 $\mathbf{q}_{\text{NT}} \rightarrow \mathbf{q}$,

$$\left\| \mathbb{P}_{\text{NT}}^{*,\text{PG}}(r_{\text{NT}}(\bar{Y}_{\text{NT},b}^{*,\text{PG}} - \mathbb{E}[Y_{it}])) - \mathcal{L}_0(\mathbf{q}^{\text{PG}}, \mathbf{c}, \varrho) \right\|_{\infty} \rightarrow 0, \quad (\text{B.2})$$

where the simulation algorithm estimates the law $\mathbb{P}_{\text{NT}}^{*,\text{PG}}$ 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. *Q.E.D.*

B.4. 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 $m_N, m_T \rightarrow \infty$, where we assume throughout that $m_N/N, m_T/T \rightarrow 0$.
- (b) For the b th subsample, let $j(1), \dots, j(m_N)$ and $s(1), \dots, s(m_T)$ be drawn uniformly and independently without replacement from $\{1, \dots, N\}$ and $\{1, \dots, T\}$, respectively.
- (c) We then let $Y_{it,b}^{\circ} := Y_{j(i)s(t),b}$ for $i = 1, \dots, m_N$ and $t = 1, \dots, m_T$, and form the b th subsample mean $\bar{Y}_{\text{NT},b}^{\circ} := \frac{1}{m_N m_T} \sum_{i=1}^{m_N} \sum_{t=1}^{m_T} Y_{it,b}^{\circ}$.

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

$$(\hat{S}_{\text{NT},b}^{\circ})^2 := \hat{D}_a(\kappa_a) T (\hat{\sigma}_a^{\circ})^2 + \hat{D}_g(\kappa_g) T (\hat{\sigma}_g^{\circ})^2 + (\hat{\sigma}_w^{\circ})^2.$$

Here, the variance estimators $\hat{\sigma}_a^{\circ}, \hat{\sigma}_g^{\circ}, \hat{\sigma}_w^{\circ}$ are the subsample analogs of $\hat{\sigma}_a^2, \hat{\sigma}_g^2, \hat{\sigma}_w^2$, 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 $b = 1, \dots, B_{\text{NT}}^{\circ}$, where $B_{\text{NT}}^{\circ} := \binom{N}{m_N} \binom{T}{m_T}$, and denote the conditional distribution of the normalized subsample

mean given the sample $(Y_{it} : i = 1, \dots, N, t = 1, \dots, T)$ with

$$\mathbb{P}_{\text{NT}}^\circ(r_{\text{NT}}^\circ(\bar{Y}_{\text{NT}}^\circ - \bar{Y}_{\text{NT}}) \leq x) := \frac{1}{B_{\text{NT}}^\circ} \sum_{b=1}^{B_{\text{NT}}^\circ} \mathbb{1}\{r_{\text{NT}}^\circ(\bar{Y}_{\text{NT},b}^\circ - \bar{Y}_{\text{NT}}) \leq x\}.$$

Here, we denote the rate for the subsample mean with

$$(r_{\text{NT}}^\circ)^2 := m_N^{-1} \sigma_a^2 + m_T^{-1} \sigma_g^2 + (m_N m_T)^{-1} \sigma_w^2.$$

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

PROPOSITION B.2—Subsampling: *Suppose that Assumption 2.1 holds, $m_N, m_T \rightarrow \infty$, and $\frac{m_N}{N}, \frac{m_T}{T} \rightarrow 0$. Then*

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

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

It is straightforward to establish consistency for pivotal versions of subsampling, where we can use Corollary A.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 estimators 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 $\bar{Y}_{\text{NT}}^\circ - \mathbb{E}[Y_{it}]$ 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,

$$\begin{aligned} & \left\| \mathbb{P}_{\text{NT}}^\circ \left(\sqrt{m_N m_T} \left(\frac{\bar{Y}_{\text{NT},b}^\circ - \bar{Y}_{\text{NT}}}{\hat{S}_{\text{NT},b}^\circ} \right) \right) - \mathbb{P}_{\text{NT}} \left(\sqrt{\text{NT}} \left(\frac{\bar{Y}_{\text{NT}} - \mathbb{E}[Y_{it}]}{\hat{S}_{\text{NT},\text{sel}}} \right) \right) \right\|_\infty \\ &= O_p \left((r_{\text{NT}}^\circ)^{-1} + \left(\frac{r_{\text{NT}}^\circ}{r_{\text{NT}}} \right)^2 \right), \end{aligned}$$

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 $(r_{\text{NT}}^\circ)^{-1} + (r_{\text{NT}}^\circ/r_{\text{NT}})^2 = O(r_{\text{NT}}^{-2/3})$, 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 B.I in the main text.

PROOF OF PROPOSITION B.2: Define the local parameters

$$\begin{aligned} q_{a,\text{NT}}^\circ &:= (r_{\text{NT}}^\circ)^2 m_N^{-1} \sigma_a^2, & q_{g,\text{NT}}^\circ &:= (r_{\text{NT}}^\circ)^2 m_T^{-1} \sigma_g^2, \\ q_{e,\text{NT}}^\circ &:= (r_{\text{NT}}^\circ)^2 (m_N m_T)^{-1} \sigma_e^2, & q_{v,\text{NT}}^\circ &:= (r_{\text{NT}}^\circ)^2 (m_N m_T)^{-1} \sigma_v^2, \\ q_{ak,\text{NT}}^\circ &:= (r_{\text{NT}}^\circ)^2 m_N^{-1} \sigma_{ak}, & q_{gk,\text{NT}}^\circ &:= (r_{\text{NT}}^\circ)^2 m_T^{-1} \sigma_{gk}, \end{aligned} \tag{B.3}$$

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

$$\begin{aligned} q_a^\circ &:= \lim_{N,T} q_{a,m_N m_T}, & q_g^\circ &:= \lim_{N,T} q_{g,m_N m_T}, & q_e^\circ &:= \lim_{N,T} q_{e,m_N m_T}, & q_v^\circ &:= \lim_{N,T} q_{v,m_N m_T}, \\ q_{ak}^\circ &:= \lim_{N,T} q_{ak,m_N m_T}, & q_{gk}^\circ &:= \lim_{N,T} q_{gk,m_N m_T} \end{aligned}$$

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

Let $J_{\text{NT}}(x) := \mathbb{P}(r_{\text{NT}}(\bar{Y}_{\text{NT}} - \mathbb{E}[Y_{it}]) \leq x)$ and $J_{\text{NT}}^\circ(x) := \mathbb{P}_{\text{NT}}^\circ(r_{\text{NT}}^\circ(\bar{Y}_{\text{NT}}^\circ - \mathbb{E}[Y_{it}]) \leq x)$ be the respective unconditional c.d.f.s of the normalized sample mean and its subsample analog. We first check whether $J_{\text{NT}}(x)$ and $J_{\text{NT}}^\circ(x)$ 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 $J_{\text{NT}}(x)$.

For the b th subsample, rows and columns are drawn uniformly and without replacement from $\{1, \dots, N\}$ and $\{1, \dots, T\}$, respectively. Hence, the array $(Y_{it,b}^\circ : i = 1, \dots, m_N; t = 1, \dots, m_T)$ 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 $\mathbf{q}_{\text{NT}}^\circ := (q_{e,\text{NT}}^\circ, q_{a,\text{NT}}^\circ, q_{g,\text{NT}}^\circ, q_{a1,\text{NT}}^\circ, q_{g1,\text{NT}}^\circ, \dots)$, it follows from Theorem 4.1 that along any convergent sequence $\mathbf{q}_{\text{NT}}^\circ \rightarrow \mathbf{q}^\circ = (q_e^\circ, q_a^\circ, q_g^\circ, q_{a1}^\circ, q_{g1}^\circ, \dots)$, we have

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

where $\varrho^\circ := \lim_{\text{NT}} r_{\text{NT}}^\circ(\text{NT})^{-1/2}$. In particular, the respective limits of $J_{\text{NT}}(x)$ and $J_{\text{NT}}^\circ(x)$ 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 $Y_{\text{NT}}^\circ - \mathbb{E}[Y_{it}^\circ]$ 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 $J_{\text{NT}}(x)$ and $J_{\text{NT}}^\circ(x)$ 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 $q_v^\circ = q_e^\circ = 0$ and $q_a^\circ + q_g^\circ = 1$. If $q_a + q_g = 0$, then we also have $q_a^\circ + q_g^\circ = 0$. Since the subsample is a draw from the same separately exchangeable array as the initial sample, it also follows that $q_e^\circ = q_e$ and $q_v^\circ = q_v$, so that $J_{\text{NT}}(x)$ and $J_{\text{NT}}^\circ(x)$ 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 $q_v^\circ = 0$ and $q_a + q_g + q_e = q_a^\circ + q_g^\circ + q_e^\circ = 1$, so that the limiting distributions coincide. If $q_v > 0$ and $q_a + q_g > 0$, then $m_N/N \rightarrow 0$ and $m_T/T \rightarrow 0$ imply 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, $J_{\text{NT}}(x)$ and $J_{\text{NT}}^\circ(x)$ 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 $q_v = q_v^\circ$. *Q.E.D.*

APPENDIX C: 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 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 concern 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.

C.1. Multivariate Case

Another important extension concerns the case of the mean of a vector-valued array (\mathbf{Y}_{it}) , where $\mathbf{Y}_{it} = (Y_{1it}, \dots, Y_{Mit})' \in \mathbb{R}^M$, 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

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

Here, the elements of $\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_t, \boldsymbol{\varepsilon}_{it} \in \mathbb{R}^M$ are i.i.d. but different components of $f(\cdot)$ may depend on a common set of factors to induce dependence between those components.

This adds the possibility that each component of the vector Y_{it} may exhibit cluster-dependence in the mean but that certain linear combinations could in fact be degenerate. Specifically, the rate of convergence of individual components of the sample mean $\bar{\mathbf{Y}}_{\text{NT}}$ depends on whether the diagonal elements of the variance matrices of $\mathbb{E}[Y_{it}|\alpha_i]$ and $\mathbb{E}[Y_{it}|\gamma_t]$ are nonzero, and linear combinations of components are guaranteed to converge at the slower \sqrt{N} or \sqrt{T} rate only if $\text{Var}(\mathbb{E}[Y_{it}|\alpha_i])$ (or $\text{Var}(\mathbb{E}[Y_{it}|\gamma_t])$, respectively) has full rank.

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 $\mathbf{Q}_i, \mathbf{Q}_t$, and \mathbf{Q}_{it} are M -dimensional vectors whose components are defined in analogy to the scalar case. The shrinkage parameters $\hat{\lambda}_1, \dots, \hat{\lambda}_M$ are then computed component by component as in the univariate case.

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

C.2. Bootstrapping Estimators

The bootstrap procedure developed for the distribution of the sample mean $\bar{\mathbf{Y}}_{\text{NT}}$ 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 $g : \mathcal{Y} \times \Theta \rightarrow \mathbb{R}^m$. We can obtain a Z-estimator $\hat{\theta}$ for the parameter by solving m estimating equations of the form

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

where we define $\hat{g}_{\text{NT}}(\theta) := \frac{1}{\text{NT}} \sum_{i=1}^N \sum_{t=1}^T g(Y_{it}; \theta)$, and \hat{A}_{NT} is a $k \times m$ matrix which may depend on quantities estimated from the data with probability limit $\hat{A}_{\text{NT}} \xrightarrow{p} A_0$. 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 arguments³ that the estimator is asymptotically linear and satisfies the expansion

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

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

Following the proposal by [Kline and Santos \(2012\)](#), we can obtain the bootstrap analog $\hat{g}_{\text{NT}}^*(\hat{\theta}) := \frac{1}{\text{NT}} \sum_{i=1}^N g_{it}^*$ 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

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

where $\hat{G}_{\text{NT}} := \frac{1}{\text{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\theta} g(Y_{it}; \hat{\theta})$. Under regularity conditions, the proof for bootstrap consistency in Theorem 4.2 can be adapted to show that this procedure is asymptotically valid. However, it is important to note that refinements for a pivotal version of this bootstrap are generally only available if the estimating equations are linear in the parameter, so that the estimator can be written 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 $\pi : \Theta \rightarrow \mathbb{R}^M$ to the corresponding sample moments, $\frac{1}{\text{NT}} \sum_{i=1}^N \sum_{t=1}^T g(Y_{it})$. In that case, we can directly bootstrap the joint distribution of the sample moment functions via

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

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.

C.3. 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 $i = 1, \dots, N$ could be high school students, and $t = 1, \dots, T$ 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. For example, a principal may assign a more talented teacher to a classroom of “weak” students. Endogenous sorting

³See, for example, [Newey and McFadden \(1994\)](#).

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 C.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 1 if Y_{it} is observed for the dyad (i, t) , and zero otherwise. We then consider the sampling distribution of

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

conditional on W_{it} . We also let

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

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

We then make the following assumptions:

ASSUMPTION C.1: (a) As $N, T \rightarrow \infty$, sampling weights W_{it} are such that $\frac{1}{N} \sum_{i=1}^N (p_i/\bar{p})^2 \rightarrow \tau_a < \infty$ and $\frac{1}{T} \sum_{t=1}^T (p_t/\bar{p})^2 \rightarrow \tau_g < \infty$. (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 $\bar{p} \rightarrow 0$, 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 C.1, we find from elementary variance calculations that

$$r_{\text{NT},W}^{-2} := \text{Var}(\bar{Y}_{\text{NT},W})$$

$$= \frac{1}{\text{NT}\bar{p}} \left(T \bar{p} \sigma_a^2 \left[\frac{1}{N} \sum_{i=1}^N \left(\frac{p_i}{\bar{p}} \right)^2 \right] + N \bar{p} \sigma_g^2 \left[\frac{1}{T} \sum_{t=1}^T \left(\frac{p_t}{\bar{p}} \right)^2 \right] + \sigma_w^2 \right). \quad (\text{C.1})$$

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

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

We then consider the following bootstrap algorithm:

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

$$\hat{\lambda}_a := \frac{\hat{D}_a(\kappa_a) T \bar{p} \hat{\sigma}_a^2 \left[\frac{1}{N} \sum_{i=1}^N \left(\frac{p_i}{\bar{p}} \right)^2 \right]}{\hat{D}_a(\kappa_a) T \bar{p} \hat{\sigma}_a^2 \left[\frac{1}{N} \sum_{i=1}^N \left(\frac{p_i}{\bar{p}} \right)^2 \right] + \bar{p} \hat{\sigma}_w^2},$$

$$\hat{\lambda}_g := \frac{\hat{D}_g(\kappa_g) N \bar{p} \hat{\sigma}_g^2 \left[\frac{1}{T} \sum_{t=1}^T \left(\frac{p_t}{\bar{p}} \right)^2 \right]}{\hat{D}_g(\kappa_g) N \bar{p} \hat{\sigma}_g^2 \left[\frac{1}{T} \sum_{t=1}^T \left(\frac{p_t}{\bar{p}} \right)^2 \right] + \bar{p} \hat{\sigma}_w^2},$$

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 C.1, the analogous conclusions to Theorem 4.2 and Proposition 4.2 hold for the modified bootstrap distribution:

PROPOSITION C.1—Bootstrap Consistency: *Suppose that Assumptions 2.1 and C.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,*

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

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 $q_v + q_e > 0$.

PROOF: 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 take the form

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

for the vector $\phi_k := (\phi_k(\alpha_1), \dots, \phi_k(\alpha_N))'$. To characterize the limit distribution for $N\sqrt{\bar{p}}Q_k$, 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 $N\sqrt{\bar{p}}Q_k$ to $N\sqrt{\bar{p}}\tilde{Q}_k$ were given by [Götze and Tikhomirov \(1999\)](#), noting that the matrix $W + W'$ is symmetric.

Now, by Assumption C.1(a), we either have that $\sup_{i=1, \dots, N} p_i \rightarrow 0$, 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 $\sup_{i=1, \dots, N} p_i \rightarrow 0$, Corollary 2 in [Götze and Tikhomirov \(1999\)](#) implies that

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

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 $N\sqrt{\bar{p}}Q_k$ is Gaussian. By an analogous argument, we also find that the distribution of the bootstrap analog $N\sqrt{\bar{p}}Q_k^*$ converges to $N\sqrt{\bar{p}}\tilde{Q}_k$, 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 $\{1, \dots, N\}$. 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 $\phi_k(\alpha_1), \dots, \phi_k(\alpha_N), \psi_k(\gamma_1), \dots, \psi_k(\gamma_T)$ and considering the symmetric quadratic form corresponding to the $(N + T) \times (N + T)$ matrix $A = \frac{1}{2}[0, W; W'0]$. Q.E.D.

Note that the only major complication in the proof 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, $\bar{p} \rightarrow 0$, 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 Proposition 4.2 establishes refinements to the estimated percentiles for the case of non-exhaustively matched samples whenever $q_v = 0$.

C.4. 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} , $r = 1, \dots, R_{it}$. We consider inference for the pooled average

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

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 $i = 1, \dots, N$, $t = 1, \dots, T$, and $r = 1, \dots, R$, 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 C.2: (a) As $N, T \rightarrow \infty$, sampling weights R_{it} are such that $\bar{r} \rightarrow \infty$, $\frac{1}{N} \sum_{i=1}^N (r_i/\bar{r})^2 \rightarrow \varrho_a < \infty$, and $\frac{1}{T} \sum_{t=1}^T (r_t/\bar{r})^2 \rightarrow \varrho_g < \infty$. (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

$$\begin{aligned} \hat{a}_i &:= \frac{1}{\mathbb{T}_t} \sum_{t=1}^T \sum_{r=1}^{R_{it}} Y_{itr} - \bar{Y}_{NT,R}, & \hat{g}_t &:= \frac{1}{N r_t} \sum_{i=1}^N \sum_{r=1}^{R_{it}} Y_{itr} - \bar{Y}_{NT,R}, \\ \hat{v}_{it} &:= \frac{1}{R_{it}} \sum_{r=1}^{R_{it}} Y_{itr} - \hat{a}_i - \hat{g}_t + \bar{Y}_{NT,R}, & \hat{e}_{itr} &:= Y_{itr} - \hat{a}_i - \hat{g}_t - \hat{v}_{it}. \end{aligned}$$

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

(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

$$\hat{\lambda}_a := \frac{\hat{D}_a(\kappa_a) T \bar{r} \hat{\sigma}_a^2 \left[\frac{1}{N} \sum_{i=1}^N \left(\frac{r_i}{\bar{r}} \right)^2 \right]}{\hat{D}_a(\kappa_a) T \bar{r} \hat{\sigma}_a^2 \left[\frac{1}{N} \sum_{i=1}^N \left(\frac{r_i}{\bar{r}} \right)^2 \right] + \bar{r} \hat{\sigma}_w^2},$$

$$\hat{\lambda}_g := \frac{\hat{D}_g(\kappa_g) N \bar{r} \hat{\sigma}_g^2 \left[\frac{1}{T} \sum_{t=1}^T \left(\frac{r_t}{\bar{r}} \right)^2 \right]}{\hat{D}_g(\kappa_g) N \bar{r} \hat{\sigma}_g^2 \left[\frac{1}{T} \sum_{t=1}^T \left(\frac{r_t}{\bar{r}} \right)^2 \right] + \bar{r} \hat{\sigma}_w^2},$$

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 C.2, the analogous conclusions to Theorem 4.2 and Proposition 4.2 regarding bootstrap consistency and refinements hold for the modified bootstrap procedure after only minor modifications of the arguments in Theorem C.1.

C.5. Cluster-Dependence in D Dimensions

The bootstrap procedure can be immediately extended to the case of an array $(Y_{i_1 \dots i_D} : i_1 = 1, \dots, N_1, \dots, i_D = 1, \dots, N_D)$ 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 d th dimension, the “sheets” of the form $(Y_{i_1 \dots i_{d-1} j_{d+1} \dots i_D} : i_{d'} = 1, \dots, N_{d'}, d' \neq d)$ are identically distributed for each $j = 1, \dots, N_d$ and $d = 1, \dots, D$.

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

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

for some function $f(m, a_1^{(1)}, a_2^{(1)}, \dots, \alpha_{1 \dots D}^{(D)})$, where $\mu, \alpha_{1i_1}^{(1)}, \dots, \alpha_{1 \dots D i_1 \dots i_D}^{(D)}$ are i.i.d. draws from the uniform distribution for $i_d = 1, \dots, N_d$ and $d = 1, \dots, D$. As in the leading case, we consider inference with respect to the conditional mean of $Y_{i_1 \dots i_D}$ 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 $\mathbf{d} = (d_1, \dots, d_k)$. In particular, $N_{\emptyset}^{(0)} = \prod_{d=1}^D N_d$.

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

$d_1, \dots, d_k,$

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

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

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

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

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

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

For $d \in \{1, \dots, D\}$, we then draw $a_d^{(1)*}$ independently from the empirical distribution for $\hat{a}_d^{(1)}$, and for each $k = 1, \dots, D-1$ and $d_1, \dots, d_k \in \{1, \dots, D\}$, we let $a_{d_1 \dots d_k i_{d_1} \dots i_{d_k}}^{(k)*} := \hat{a}_{d_1 \dots d_k j_d^*(i_1) \dots j_d^*(i_k)}^{(k)} (\prod_{l=1}^k \omega_{d_l i_{d_l}})$ for independent draws $\omega_{d_l i_{d_l}}$ 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_d th bootstrap draw for dimension d . We then form

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

where for the bootstrap with and without model selection,

$$\hat{\lambda}_{\mathbf{d}'k} := \frac{\hat{D}_{a_{\mathbf{d}'}}^{(k)}(\kappa_{a_{\mathbf{d}'}}^{(k)}) N_{\mathbf{d}'}^{(k)} \hat{\sigma}_{a_{\mathbf{d}'}}^{2(k)}}{\hat{D}_{a_{\mathbf{d}'}}^{(k)}(\kappa_{a_{\mathbf{d}'}}^{(k)}) N_{\mathbf{d}'}^{(k)} \hat{\sigma}_{a_{\mathbf{d}'}}^{2(k)} + \hat{\sigma}_{a_{1 \dots D}}^{2(D)}}$$

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

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

Noting that the arguments behind Theorem 4.2 and Proposition 4.2 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.

C.6. 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 $(Y_{i_1 \dots i_D} : i_1, \dots, i_D = 1, \dots, N)$ whose distribution is invariant to permutations $\pi : \{1, \dots, N\} \rightarrow \{1, \dots, N\}$, that is, $Y_{i_1 \dots i_D} \stackrel{d}{=} Y_{\pi(i_1) \dots \pi(i_D)}$. Using the terminology of [Kallenberg \(2005\)](#), such an array is jointly exchangeable and can be represented as

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

and $\mu, \alpha_{i_1}^{(1)}, \alpha_{i_2}^{(1)}, \dots, \alpha_{i_1 \dots i_D}^{(D)}$ are i.i.d. uniform random variables for $i_1, \dots, i_D \in \{1, \dots, N\}$ (see [Hoover \(1979\)](#) or Theorem 7.22 in [Kallenberg \(2005\)](#)). Conditional on μ , we then consider the sampling distribution of the “ D -adic” mean

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

for N units drawn at random from a larger population (with replacement) or distribution.⁴

EXAMPLE C.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 $1, \dots, N$ 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 $R_{i_1 \dots i_D}(F)$ that equals 1 if there is an adjacency-preserving map between F and the subnetwork among the nodes i_1, \dots, i_D . We can then re-sample from the D -dimensional array with entries $Y_{i_1 \dots i_D} := R_{i_1 \dots i_D}(F)$ 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 C.5 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 Theorem 4.2 and Proposition 4.2 then go through under analogous conditions as for the original case.

REMARK C.1—U-Statistics: U- and V-statistics with a kernel of order higher than 2 are a special case of D -adic data, where $Y_{i_1, \dots, i_D} = h(X_{i_1}, \dots, X_{i_D})$ is the known kernel function $h(\cdot)$, and X_1, \dots, X_n are observed by the researcher. In that special case, the array has the simpler representation $Y_{i_1, \dots, i_D} = h(\alpha_{i_1}, \dots, \alpha_{i_D})$ in terms of a reduced number of Aldous–Hoover factors. The population analogs of $\hat{a}_{\text{di}}^{(k)}$ then correspond to terms in the usual Hoeffding expansion of the statistic. While our theoretical insights also apply to that problem, for the special case of U- or V-statistics in which the factors $\alpha_{i_1}^{(1)}$ correspond to observable data, more straightforward resampling procedures may be available.

⁴Note 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.

For example, the researcher might prefer to adapt the procedure in Arcones and Giné (1992) with an appropriate model selection step along the lines described in this paper to determine the appropriate order of degeneracy for their method. However, establishing formal properties of such a procedure is beyond the scope of this paper.

REFERENCES

- ANDREWS, D., AND P. GUGGENBERGER (2010): “Asymptotic Size and a Problem With Subsampling and With the m out of n Bootstrap,” *Econometric Theory*, 26, 426–468. [8]
- ARCONES, M., AND E. GINÉ (1992): “On the Bootstrap of U and V Statistics,” *Annals of Statistics*, 20 (2), 655–674. [20]
- CAMERON, C., J. GELBACH, AND D. MILLER (2011): “Robust Inference With Multiway Clustering,” *Journal of Business & Economic Statistics*, 29 (2), 238–249. [1,3]
- DAVEZIES, L., X. D’HAULTFÈUILLE, AND Y. GUYONVARCH (2018): “Asymptotic Results Under Multiway Clustering,” Working paper, ENSAE. [4]
- GÖTZE, F., AND N. TIKHOMIROV (1999): “Asymptotic Distribution of Quadratic Forms,” *Annals of Probability*, 27 (2), 1072–1098. [15]
- HOOVER, D. (1979): “Relations on Probability Spaces and Arrays of Random Variables,” Princeton. Working paper, Institute for Advanced Study. [17,19]
- KALLENBERG, O. (2005): *Probabilistic Symmetries and Invariance Principles*. Springer. [17,19]
- KLINE, P., AND A. SANTOS (2012): “A Score Based Approach to Wild Bootstrap Inference,” *Journal of Econometric Methods*, 1 (1), 23–41. [12]
- LOVASZ, L. (2012): *Large Networks and Graph Limits*. AMS Colloquium Publications, Vol. 60. Providence, RI: American Mathematical Society. [19]
- MENZEL, K. (2021): “Bootstrap With Cluster-Dependence in Two or More Dimensions,” Working paper, New York University. [1]
- NEWNEY, W., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” *Handbook of Econometrics*, Vol IV. Chapter 36. [12]
- OWEN, A. (2007): “The Pigeonhole Bootstrap,” *The Annals of Applied Statistics*, 1 (2), 386–411. [1,4,5]
- POLITIS, D., AND J. ROMANO (1994): “Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions,” *Annals of Statistics*, 22 (4), 2031–2050. [1,9,10]
- POLITIS, D., J. ROMANO, AND M. WOLF (1999): *Subsampling*. New York: Springer. [1]

Co-editor Ulrich K. Müller handled this manuscript.

Manuscript received 31 May, 2017; final version accepted 5 February, 2021; available online 9 February, 2021.