econometrics

MDPI

Article

Consistency of the OLS Bootstrap for Independently but Not-Identically Distributed Data: A Permutation Perspective

Alwyn Young

Department of Economics, London School of Economics, Houghton St., London WC2A 2AE, UK; a.young@lse.ac.uk

Abstract

This paper introduces a new approach to proving bootstrap consistency based upon the distribution of permutation statistics, using it to derive results covering fundamentally not-identically distributed groups of data, in which average moments do not converge to anything, with moment conditions that are less demanding than earlier results for either identically distributed or not-identically distributed data.

Keywords: bootstrap consistency; permutation distribution

1. Introduction

Data are often drawn from dissimilar environments which render the independent and identically distributed (iid) assumption that underlies many results on the bootstrap suspect.[1] This paper extends results concerning the consistency of the pairs and wild OLS bootstraps, which have mostly been derived for iid data, to general regression frameworks with independently but not-necessarily identically distributed (inid) data. Instead of considering the sampling distribution of the bootstraps, the usual approach, it notes that any permutation of the pairs bootstrap vector of sampling frequencies or the realization of the external variable used by the wild bootstrap to transform residuals is equally likely. Using results on the asymptotic distribution of permutation statistics by Wald and Wolfowitz (1944), Noether (1949), and Hoeffding (1951), these equally likely permutations can be used to characterize the bootstrap distributions conditional on the data as normal given restrictions on sample moments of the data. White's (1980a) conditions for the asymptotic normality of OLS coefficients with clustered/heteroskedastic residuals and inid data guarantee these restrictions almost surely, ensuring that the asymptotic distribution of pairs and wild bootstrapped coefficients and Wald statistics conditional on the data matches the unconditional distribution of the original OLS estimates.

While proofs of bootstrap consistency typically require the existence of at least fourth moments of the regressors with iid data, the permutation distribution allows this paper to prove consistency with no more than second regressor moments and inid data. For iid data, Mammen (1993) proved consistency of the wild OLS bootstrap coefficient and homoskedasticity-based Wald test distributions with bounded expectations of the product of the fourth power of the regressors with the squared errors and an additional Lindeberg condition. Similarly, for the pairs OLS bootstrap with iid data Freedman (1981) showed that bounded fourth moments of both regressors and errors are sufficient for consistency of the pairs bootstrap coefficient distribution[2] and that of the Wald statistic based upon the (potentially incorrect) assumption of homoskedastic errors. Stute (1990) tightened the result for the coefficient distribution alone, showing it is sufficient for the squared

check for updates

Academic Editor: Julie Le Gallo

Received: 30 June 2025

Revised: 26 September 2025

Accepted: 29 September 2025

Published: 23 October 2025

Citation: Young, A. (2025). Consistency of the OLS Bootstrap for Independently but Not-Identically Distributed Data: A Permutation Perspective. *Econometrics*, *13*(4), 41. https://doi.org/10.3390/econometrics13040041

regressors and the product of the squared regressors with the squared errors to have finite expectation. This paper proves consistency of both the coefficient and clustered heteroskedasticity robust Wald statistic distribution in a broader inid environment for both the pairs and wild bootstraps with finite expectations of only slightly more than second powers of the regressors and of the product of the second powers of the regressors with the second power of the errors. These are much less demanding assumptions than those used by Freedman and Mammen, requiring only slightly higher moments than used by Stute for the proof of only the pairs bootstrap coefficient distribution in a narrower iid environment. Moreover, when residuals are heteroskedastic or interrelated within clusters, the homoskedasticity-based Wald test is not guaranteed to be asymptotically accurate, as recognized by Freedman (1981) and Mammen (1993). In such cases, practitioners are likely to prefer clustered/heteroskedasticity robust covariance estimates and Wald statistics as these are asymptotically accurate and pivotal, respectively, ensuring the asymptotic accuracy of the conventional test and higher order accuracy and faster convergence of rejection probabilities to the nominal value in the bootstrap (Singh, 1981; Hall, 1992).

For OLS models with inid data, the salient contribution is Liu (1988), who showed that the wild bootstrap provides consistent estimates of the second central moment of a linear combination of coefficients in an OLS regression model with bounded regressors, provided the first and second moments of the wild bootstrap external variable are 0 and 1, respectively. Liu's result regarding the second central moment is easily extended to the case of the multivariate second central moments of coefficients for unbounded inid regressors without any additional restrictions on the moments of the external variable, as shown below. Our interest here, however, is in the full distribution of wild bootstrap coefficient and Wald statistic estimates, where our proof requires the existence of higher moments of the wild bootstrap external variable to ensure the convergence of higher moments to the normal. As the external variable is selected by the practitioner, and not an exogenous characteristic of the data, these additional moment conditions pose no obstacle. The two-point distribution proposed by Mammen (1993) and the Rademacher distribution, both often used in practical applications (e.g., Davidson & Flachaire, 2008), have moments of all orders.

Liu's consideration of inid data has largely not been extended, as the OLS bootstrap literature has since focused on time series dependent data, where the absence of random sampling of independent observations raises different statistical issues and the use of different bootstrap methods (see the review in Hardle et al., 2003). Djogbenou et al. (2019), who prove consistency of the wild bootstrap t-statistic distribution for independently distributed cluster groupings of data, are a notable exception. With the moment assumptions used here, plus the additional requirement of bounded slightly higher than fourth moments of the regressors, their proof allows for heterogeneity in the distribution of data across clusters. However, they limit that heterogeneity in requiring that the cross product of the regressors and the covariance matrix of coefficient estimates converge to matrices of constants, a condition that in other papers is typically motivated by an iid assumption.[3] The data-generating process examined in this paper is more fully inid in that there is no restriction that such matrices converge to anything, and the proof requires only slightly higher than second moments of the regressors. In sum, by emphasizing the permutation distribution, this paper lowers typical fourth moment restrictions on regressors to second moments, allows a fully inid data process in which average moments do not converge, and highlights the conceptual similarity between the wild and pairs bootstraps, proving results for both in a unified framework.

The paper proceeds as follows: Section 2 reviews the OLS model, White's assumptions and results regarding OLS with inid data, and pairs and wild bootstrap methods for clustered/heteroskedastic data. Section 3 presents the foundational theorems regarding the asymptotic normality of permutation distributions that motivate the results. Section 4 then combines these with White's (1980a) result to derive sufficient conditions for pairs and wild OLS bootstrap consistency with inid data and potentially cluster interdependent heteroskedastic residuals. Section 5 more fully contrasts the assumptions and results herein with those found in the papers cited above. Section 6 provides Monte Carlo evidence of the consistency of the bootstrap in a challenging environment with an inid data process where average moments do not converge, regressors have barely second moments, and residuals are bounded, varyingly skewed, sometimes bi-modal, and otherwise generally highly non-normal. The Appendix provides proofs of the main theorems, while the on-line Appendix extends the pair results to sub-sampling and provides lengthy technical proofs of otherwise minor lemmas and extensions of the theorems.

2. Framework and Notation

Our interest is in inference for the linear model where, with $i = 1 \ldots N$ observations,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ represents the $N \times 1$ matrix of observations on the dependent (outcome) variable, $\boldsymbol{X}$ the $N \times K$ matrix of observations of independent variables, $\boldsymbol{\beta}$ the $K \times 1$ vector of unobserved parameters of interest, and $\boldsymbol{\varepsilon}$ the $N \times 1$ matrix of unobserved disturbances. The ordinary least squares (OLS) estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ minimize the sum of squared estimated residuals $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$, where $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, producing the estimates

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2}$$

If the disturbances ε_i are homoskedastic with common variance $\sigma_i^2 = \sigma^2$, one can use the homoskedastic variance estimate of $\hat{\boldsymbol{\beta}}$, $(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(N-K)$, but we focus on more general inference where the ε_i are heteroskedastic and possibly interdependent within $C \leq N$ "cluster" groupings of observations, using the clustered/heteroskedasticity robust covariance estimate

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{C} \boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{3}$$

where we use the subscript g to denote the rows of matrices and vectors associated with the observations in cluster grouping g. As will be seen later, we assume that the regressors and disturbances $(\boldsymbol{X}_g, \boldsymbol{\varepsilon}_g)$ are independent across cluster groupings. When observations themselves are independent, each grouping g equals an individual observation i, $C = N$, and (3) is White's (1980a) heteroskedasticity robust covariance estimate. The clustered extension, however, is often used to allow for unspecified grouped dependence, and so we present the results within a more general framework. In describing limits, we use the subscript C, as in $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_C)$, to emphasize that the estimated coefficients and covariance estimates are functions of C realized observation groupings.

White (1980a) provided conditions for valid OLS inference when the row vector of random variables associated with each observation is independently but not necessarily identically distributed (inid). With $\boldsymbol{x}_{gj}$ denoting the jth column of $\boldsymbol{X}_g$, we extend these to allow for grouped dependence:

Theorem I (extending White, 1980a). *If there exist strictly positive finite constants γ, Δ, and η, such that*

(Ia) $(\boldsymbol{X}_g, \boldsymbol{\varepsilon}_g)$ *is a sequence of independent but not necessarily identically distributed random matrices, such that* $E(\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g) = \boldsymbol{0}_K$;

(Ib) *For all g* $E(|\boldsymbol{x}'_{gj}\boldsymbol{x}_{gk}|^{1+\gamma}) < \Delta$ *for all j, k* $= 1 \ldots K$, *and for all C sufficiently large* $\boldsymbol{M}_C = C^{-1}\sum_{g=1}^{C} E(\boldsymbol{X}'_g\boldsymbol{X}_g)$ *is non-singular with determinant* $(\boldsymbol{M}_C) > \eta$;

(Ic) *For all i,* $E(|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{x}_{gk}|^{1+\gamma}) < \Delta$ *for all j, k* $= 1 \ldots K$, *and for all C sufficiently large* $\boldsymbol{V}_C = C^{-1}\sum_{g=1}^{C} E(\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{X}_g)$ *is non-singular with determinant* $(\boldsymbol{V}_C) > \eta$;

then

(i) $\hat{\boldsymbol{\beta}}_C \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{\beta}$;

(ii) $\mathbf{V}_C^{-1/2}\boldsymbol{M}_C\sqrt{C}\left(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}\right) \overset{d(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{n}_K$;

(iii) $\boldsymbol{M}_C, \boldsymbol{V}_c$, *and their inverses are uniformly bounded for all C sufficiently large*

(iv) $C\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_C\right) - \boldsymbol{M}_C^{-1}\boldsymbol{V}_C\boldsymbol{M}_c^{-1} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{KxK}$;

(v) $(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C)^{-1}(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}) \overset{d(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \chi^2_K$;

where $\overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$ *and* $\overset{d(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$ *denote convergence almost surely and in distribution across* $(\boldsymbol{X},\boldsymbol{\varepsilon})$, *respectively,* $\boldsymbol{A}^{1/2}$ *the "square root" of symmetric positive definite matrix* $\boldsymbol{A}$,[4] $\boldsymbol{n}_K$ *the K dimensional standard normal,* χ^2_K *the central chi-squared with K degrees of freedom, and* $\mathbf{0}_K$ *and* $\mathbf{0}_{KxK}$ *vectors and matrices of zeros of the indicated dimensions.*

Remark 1. *White's covariance estimate often motivates inference with heteroskedasticity or clustering in an otherwise iid setting where each observation or cluster grouping is a draw from a fixed distribution. However,* $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C)$ *allows for asymptotically accurate inference in the much more general inid setting given the above, where* M_C, V_C, *and* $C\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C)$ *do not necessarily converge to matrices of constants, as illustrated in Monte Carlos further below.*

Remark 2. *White (1980a) used (Ia)–(Ic) to prove (i), (ii), and parts of (iii) and added the assumption* $E(|\boldsymbol{x}'_{gj}\boldsymbol{x}_{gj}\boldsymbol{x}'_{gk}\boldsymbol{x}_{gl}|^{1+\gamma})$ *to prove (iv), (v), and other results. As reviewed below, a similar fourth moment condition on the regressors is also used in prior proofs of bootstrap consistency. However, (Ia)–(Ic), with only slightly higher than second regressor moments, suffice to prove (i)–(v) and ensure bootstrap consistency, as shown in proofs and Monte Carlos below.*

Remark 3. *In practical application, moment restrictions on the data-generating process can be tested using techniques suggested by Meerschaert and Scheffler (1998), Fedotenkov (2013), and Trapani (2016), among others.*

In this paper, we examine two bootstrap techniques commonly used for OLS inference with heteroskedastic and clustered disturbances and prove the asymptotic consistency of their distributions for general inid data. Wu's (1986) external bootstrap, now commonly known as the wild bootstrap, holds the design matrix $\boldsymbol{X}$ constant and generates new realizations of the outcome vector $\boldsymbol{y}$ by multiplying the estimated residuals of each cluster grouping by an independently and identically distributed external random variable δ^w_g, so that the dependent variable for grouping g is now given by $\boldsymbol{y}^w_g = \boldsymbol{X}_g\hat{\boldsymbol{\beta}}_C + \hat{\boldsymbol{\varepsilon}}_g\delta^w_g$. Selecting $\hat{\boldsymbol{\beta}}^w_C$ so as to minimize the sum of squared residuals for this new data yields coefficient and covariance estimates expressed in terms of the original data and its estimates as

$$\hat{\boldsymbol{\beta}}^w_C = \hat{\boldsymbol{\beta}}_C + (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{C}\boldsymbol{X}'_g\hat{\boldsymbol{\varepsilon}}_g\delta^w_g\right)$$
$$\text{and } \hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^w_C) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{C}\boldsymbol{X}'_g\hat{\boldsymbol{\varepsilon}}^w_g\hat{\boldsymbol{\varepsilon}}^{w\prime}_g\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4}$$
$$\text{where } \hat{\boldsymbol{\varepsilon}}^w_g = \boldsymbol{X}_g\left(\hat{\boldsymbol{\beta}}_C - \hat{\boldsymbol{\beta}}^w_C\right) + \hat{\boldsymbol{\varepsilon}}_g\delta^w_g.$$

Repeated draws of the *Cx1* vector δ^w of iid variables are made and the resulting distribution of coefficients $\hat{\boldsymbol{\beta}}_C^w - \hat{\boldsymbol{\beta}}_C$ and Wald statistics $(\hat{\boldsymbol{\beta}}_C^w - \hat{\boldsymbol{\beta}}_C)'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C^w)^{-1}(\hat{\boldsymbol{\beta}}_C^w - \hat{\boldsymbol{\beta}}_C)$ used to evaluate the statistical significance of corresponding measures for tests of the null hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in the original sample, i.e., $\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_0$ and $(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_0)'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C)^{-1}(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_0)$. All permutations of any given realization of δ^w are equally likely, a fact that plays a prominent role in the results of this paper.

The pairs bootstrap samples with replacement C cluster groupings of "pairs" of dependent and independent variables $(\boldsymbol{y}_g, \boldsymbol{X}_g)$ from the rows of the original data $(\boldsymbol{y}, \boldsymbol{X})$, producing a new data set composed of $h = 1 \ldots C$ cluster groups of observations $(\boldsymbol{y}_h, \boldsymbol{X}_h)$, with each h corresponding to one of the original g groupings.[5] Selecting $\hat{\boldsymbol{\beta}}_C^p$ so as to minimize the sum of squared residuals for this new data, the resulting coefficient and covariance estimates can be expressed in terms of the original data, its estimates, and its indices $g = 1 \ldots C$ as

$$\begin{gathered}\hat{\boldsymbol{\beta}}_C^p = \hat{\boldsymbol{\beta}}_C + \left(\sum_{g=1}^{C} \boldsymbol{X}_g'\boldsymbol{X}_g\delta_g^p\right)^{-1}\left(\sum_{g=1}^{C} \boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\delta_g^p\right) \\ \text{and } \hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_C^p\right) = \left(\sum_{g=1}^{C} \boldsymbol{X}_g'\boldsymbol{X}_g\delta_g^p\right)^{-1}\left(\sum_{g=1}^{C} \boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g^p\hat{\boldsymbol{\varepsilon}}_g^{p\prime}\boldsymbol{X}_g\delta_g^p\right)\left(\sum_{g=1}^{C} \boldsymbol{X}_g'\boldsymbol{X}_g\delta_g^p\right)^{-1}, \\ \text{where } \hat{\boldsymbol{\varepsilon}}_g^p = \boldsymbol{X}_g\left(\hat{\boldsymbol{\beta}}_C - \hat{\boldsymbol{\beta}}_C^p\right) + \hat{\boldsymbol{\varepsilon}}_g,\end{gathered} \tag{5}$$

where δ_g^p denotes the number of times (possibly 0) cluster grouping g was drawn. Repeated bootstrap samples are made and the resulting distribution of coefficients $\hat{\boldsymbol{\beta}}_C^p - \hat{\boldsymbol{\beta}}_C$ and Wald statistics $(\hat{\boldsymbol{\beta}}_C^p - \hat{\boldsymbol{\beta}}_C)'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C^p)^{-1}(\hat{\boldsymbol{\beta}}_C^p - \hat{\boldsymbol{\beta}}_C)$ once again used to evaluate the statistical significance of corresponding measures for tests of the null hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in the original sample. As in the case of the wild bootstrap, all permutations of any given realization of the Cx1 sampling frequency vector δ^p are equally likely. Consequently, we use the common notation δ, distinguished by superscripted p or w, for seemingly dissimilar objects because these operate identically in the theorems and proofs below.

Our interest is in deriving sufficient conditions for the conditional consistency of the bootstrap distributions in an inid framework. Specifically, we show that White's (1980a) assumptions are sufficient to ensure that for the bootstrapped coefficient and clustered/heteroskedasticity robust covariance estimates, with b (both) denoting p (pairs) or w (wild),

$$\begin{gathered}\left(\sum_{g=1}^{C} \frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)\sqrt{C}(\hat{\boldsymbol{\beta}}_C^b - \hat{\boldsymbol{\beta}}_C) \overset{d(\delta^b)|as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{n}_K \\ \text{and } \sqrt{C}(\hat{\boldsymbol{\beta}}_C^b - \hat{\boldsymbol{\beta}}_C)'[C\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C^b)]^{-1}(\hat{\boldsymbol{\beta}}_C^b - \hat{\boldsymbol{\beta}}_C)\sqrt{C} \overset{d(\delta^b)|as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \chi_K^2,\end{gathered} \tag{6}$$

where $\overset{d(\delta)|as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$ denotes convergence in distribution across δ almost surely across realizations of $(\boldsymbol{X},\boldsymbol{\varepsilon})$. These results show that the asymptotic conditional distribution given the data $(\boldsymbol{X},\boldsymbol{\varepsilon})$ of the bootstrap equals the asymptotic distribution of the OLS estimates across $(\boldsymbol{X},\boldsymbol{\varepsilon})$, allowing for valid inference using the percentiles of bootstrapped coefficient estimates or Wald statistics.[6]

The key characteristic exploited in the proofs below is that any of the row permutations of the vectors δ are equally likely. Consequently, the distribution of the bootstraps can be thought of as the distribution across permutations of δ integrated across the ordered realizations of δ. Permutation theorems characterize this permutation distribution as asymptotically normal with covariance matrix $C\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C)$ provided $(\boldsymbol{X},\boldsymbol{\varepsilon})$ and δ have certain moment properties. White's (1980a) assumptions ensure that these properties hold almost surely for ($\boldsymbol{X}$,$\boldsymbol{\varepsilon}$), while the properties of the multinomial sampling frequencies δ^p and moment assumptions on the iid elements of δ^w ensure the requisite conditions on δ also hold almost surely. Consequently, almost surely conditional on the data ($\boldsymbol{X}$,$\boldsymbol{\varepsilon}$), the distributions of the bootstraps across the draws δ that determine their coefficient estimates and Wald statistics converge to the distribution of their OLS counterparts for the original sample ($\boldsymbol{X}$,$\boldsymbol{\varepsilon}$) across its data-generating process.

3. Foundational Permutation Theorems

The proofs in this paper rely on a theorem first proven by Wald and Wolfowitz (1944) and later refined by Noether (1949) and Hoeffding (1951) concerning the asymptotic distribution of root-C times the correlation of a permuted sequence with another sequence:

Theorem II: *Let* $\mathbf{z}'$ = (z_1, …, z_C) *and* $\boldsymbol{\delta}'$ = (δ_1, …, δ_C) *denote sequences of real numbers, not all equal, and* $\mathbf{d}'$ = (d_1, …, d_C) *denote any of the C! equally likely permutations of* $\boldsymbol{\delta}$*. Then, as* $C \to \infty$ *the distribution across the realizations of* $\mathbf{d}$ *of the random variable*

$$v_C = \sum_{g=1}^{C} \frac{\left[z_g - m\left(z_g\right)\right]\left[d_g - m\left(d_g\right)\right]}{s\left(z_g\right)s\left(d_g\right)C^{1/2}}, \\ \left[where\ for\ h = z\ or\ d,\ m(h_g) = \sum_{g=1}^{C} \frac{h_g}{C}\ \&\ s(h_g)^2 = \sum_{g=1}^{C} \frac{[h_g - m(h_g)]^2}{C}\right] \tag{IIa}$$

converges to that of the standard normal if for all integer $\tau > 2$

$$\lim_{C\to\infty} \frac{C^{\frac{\tau}{2}-1}\sum_{g=1}^{C}\left[z_g - m(z_g)\right]^{\tau}\sum_{g=1}^{C}\left[\delta_g - m(\delta_g)\right]^{\tau}}{\left(\sum_{g=1}^{C}\left[z_g - m(z_g)\right]^2\right)^{\tau/2}\left(\sum_{g=1}^{C}\left[\delta_g - m(\delta_g)\right]^2\right)^{\tau/2}} = 0. \tag{IIb}$$

The proof is based on showing that the moments of v_C converge to those of the standard normal. A simple multivariate extension, proven in the on-line Appendix, is

Theorem IIm: *Let* $\boldsymbol{O} = \boldsymbol{I}_{CxC} - \mathbf{1}_C\mathbf{1}_C'/C$ *denote the centering matrix,*[7] $\boldsymbol{Z}' = (z_1, \ldots, z_C)$ *a sequence of K x 1 vectors such that* $\mathbf{Z}'\boldsymbol{OZ}$ *is positive definite,* $\boldsymbol{\delta}'$ = (δ_1, …, δ_C) *a sequence of real numbers not all equal, and* $\mathbf{d}'$ = (d_1, …, d_C) *any of the C! equally likely permutations of* $\boldsymbol{\delta}$*. Then, as* $C \to \infty$ *the distribution across the realizations of d of the random variable*

$$\boldsymbol{v}_C = \left(\frac{\boldsymbol{Z}'\boldsymbol{OZ}}{C}\frac{\boldsymbol{d}'\boldsymbol{Od}}{C}\right)^{-1/2}\frac{(\boldsymbol{Z}'\boldsymbol{Od})}{\sqrt{C}} \tag{IIc}$$

converges to that of the multivariate iid standard normal if (IIb) holds for each element in the vector sequence $\boldsymbol{z}_g$*.*

Theorem II is easily extended to a probabilistic environment by noting the following result due to Ghosh (1950) that translates the almost-sure or in probability characteristics of an infinite number of moment conditions into similar statements regarding a distribution:

Theorem III: *If all the moments of the cumulative distribution function* $F_C(x)$ *converge almost surely (in probability) to those of* $F(x)$*, which possesses a density function, and for which, with* ν_{k+1} *denoting the absolute moment of order* $k + 1$*,*

$$\lim_{k\to\infty} \frac{\alpha^{k+2}\nu_{k+1}}{k+2!} = 0 \text{ for any given value of } \alpha, \tag{IIIa}$$

then $F_C(x)$ *converges almost surely (in probability) to* $F(x)$*.*

Condition (IIIa) is of course true for the normal distribution. Hoeffding (1952) generalized the result by showing that condition (IIIa) is not even needed for convergence in probability at all points of continuity of any $F(x)$ that is uniquely determined by its moments. By virtue of the Cramér–Wold device, Theorem III covers the multivariate case given in (IIc) above, as for all $\boldsymbol{\lambda}$, such that $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$, all moments of $\boldsymbol{\lambda}'\boldsymbol{v}_C$ converge to those of the standard normal. In light of Theorem III, in applying Theorem II below, we use the notation $\overset{d(\boldsymbol{d})|as(\boldsymbol{\delta},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$, i.e., almost surely across the realizations of ($\boldsymbol{\delta}$,$\boldsymbol{X}$,$\boldsymbol{\varepsilon}$), the distribution of $\boldsymbol{v}_C$ across permutations $\boldsymbol{d}$ of $\boldsymbol{\delta}$ converges to the multivariate standard normal. Theorems II and III are used to characterize the asymptotic distribution of $\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g d_g^b/C$, which appears in the expressions for the bootstrapped coefficient estimates in (4) and (5) above.

A less demanding form of Theorem II, proven in Appendix B below, provides a weaker condition under which the mean of products converges in probability across permutations to the product of means:

Theorem IV: *Let* $\boldsymbol{z}' = (z_1, \ldots, z_C)$ *and* $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_C)$ *denote sequences of real numbers, possibly all equal, and* $\boldsymbol{d}' = (d_1, \ldots, d_C)$ *any of the C! equally likely permutations of* $\boldsymbol{\delta}$*. Then, as* $C \to \infty$*, across permutations* $\boldsymbol{d}$ *of* $\boldsymbol{\delta}$*,*

$$m\left(z_g d_g\right) - m\left(z_g\right)m\left(\delta_g\right) = \sum_{g=1}^{C}\frac{z_g d_g}{C} - \sum_{g=1}^{C}\frac{z_g}{C}\sum_{g=1}^{C}\frac{\delta_g}{C} \overset{p}{\rightarrow} 0, \qquad \text{(IVa)}$$

if

$$\lim_{C\to\infty}\frac{\sum\limits_{g=1}^{C}\frac{[z_g - m(z_g)]^2}{C}\sum_{g=1}^{C}\frac{[\delta_g - m(\delta_g)]^2}{C}}{C} = 0. \qquad \text{(IVb)}$$

Theorem IV is used in proofs to make statements regarding the convergence in probability of terms such as $\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g\left(d_g^w\right)^2/C$, $\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g^p\hat{\boldsymbol{\varepsilon}}_g^{p\prime}\boldsymbol{X}_g d_g^p/C$, and $\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{X}_g d_g^p/C$, which appear in (4) and (5) above. As the satisfaction of (IVb) depends on the realized sample moments of ($\boldsymbol{X}$,$\boldsymbol{\varepsilon}$) and $\boldsymbol{\delta}$, we use the notation $\overset{p(\boldsymbol{d})|as(\boldsymbol{\delta},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$, i.e., almost surely across the realizations of ($\boldsymbol{\delta}$,$\boldsymbol{X}$,$\boldsymbol{\varepsilon}$) $m\left(z_g d_g\right)$ converges in probability across the permutations $\boldsymbol{d}$ of $\boldsymbol{\delta}$ to $m\left(z_g\right)m\left(\delta_g\right)$.

4. Results: Bootstrap Consistency with INID Data

The following result is proven in Appendix C, further below:

Theorem V: *Assume that for the wild bootstrap* $E\left[\delta_g^w\right] = 0$, $E\left[\left(\delta_g^w\right)^2\right] = 1$ *and* $E\left[\left(\delta_g^w\right)^{2(1+\theta_1)}\right] < \Delta$ *for some finite* Δ *and* $\theta_1 > 1/\gamma$*, with* γ *as given in Theorem I earlier. Assumptions (Ia)–(Ic) given in Theorem I in combination with the properties of* $\boldsymbol{\delta}$ *are sufficient to ensure that across the permutations* $\boldsymbol{d}$ *of* $\boldsymbol{\delta}^b$*, for b = p (pairs) or w (wild),*

$$\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)\left(\frac{\boldsymbol{\delta}^{b\prime}\boldsymbol{O}\boldsymbol{\delta}^b}{C}\right)^{-1/2}\sqrt{C}\left(\hat{\boldsymbol{\beta}}_C^b - \hat{\boldsymbol{\beta}}_C\right) \overset{d(\boldsymbol{d})|as(\boldsymbol{\delta}^b,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{n}_K, \qquad \text{(Va)}$$

$$C\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_C^b) - C\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_C) \overset{p(\boldsymbol{d})|as(\boldsymbol{\delta}^b,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{0}_{KxK}\,. \qquad \text{(Vb)}$$

Bounded higher moments of δ_g^w are needed to ensure that conditions (IIb) and (IVb) in Theorems II and IV are satisfied.

Let δ^* denote the ordered values of δ. Across permutations $\boldsymbol{d}$ of δ^* (Va) and (Vb) hold. These permutations, integrated across the distribution of δ^*, characterize the entire distribution of δ. Adding the result[8]

$$\frac{\delta^{p\prime}\mathbf{O}\delta^{p}}{C} \overset{p(\delta^p)}{\to} 1 \text{ and } \frac{\delta^{w\prime}\mathbf{O}\delta^{w}}{C} \overset{as(\delta^w)}{\to} 1, \tag{7}$$

implies that

$$\left(\sum_{g=1}^{C}\frac{\mathbf{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\mathbf{X}_g}{C}\right)^{-1/2}\left(\frac{\mathbf{X}'\mathbf{X}}{C}\right)\sqrt{C}(\hat{\boldsymbol{\beta}}_C^b - \hat{\boldsymbol{\beta}}_C) \overset{d(\delta^b)|as(\mathbf{X},\boldsymbol{\varepsilon})}{\to} \boldsymbol{n}_K. \tag{8a}$$

$$C\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C^b) - C\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_C) \overset{p(\delta^b)|as(\mathbf{X},\boldsymbol{\varepsilon})}{\to} \mathbf{0}_{KxK}, \tag{8b}$$

where the convergence in distribution in this case is across the bootstrap realizations of δ^b that determine the bootstrap coefficient and covariance estimates, as in (4) and (5) above. When combined with White's (1980a) result in Theorem I regarding the asymptotic distribution of OLS coefficient and covariance estimates, this establishes that almost surely the conditional (on the data) distributions of the bootstrapped coefficients and Wald statistics converge to the unconditional distributions of their OLS regression counterparts.

5. Comparison of Bootstrap Consistency Results

This section contrasts the assumptions and results of this paper with other papers on bootstrap consistency. These usually assume independent observations, with moment conditions given at that level. To simplify the comparison of moment conditions, where possible I use the i = 1 . . . N notation, taking each cluster $\boldsymbol{g}$ as composed of one observation and using the implied observational level assumptions in the theorems given above. Table 1 below summarizes key elements of the discussion that follows.

Table 1. Comparison of assumptions and results.

	Mammen (1993)	**Freedman (1981)**	**Stute (1990)**	**Liu (1988)**	**Djogbenou et al. (2019)**	**This Paper**
type of bootstrap	wild	pairs	pairs	wild	wild	both
type of data	iid	iid	iid	inid	inid	inid
bounded moments	$x_{ij}^4\varepsilon_i^2$ & x_{ij}^4	$x_{ij}^a\varepsilon_i^b$ $a+b=4$	$x_{ij}^2\varepsilon_i^2$ & x_{ij}^2	ε_i^2 & $all\ x_{ij}^n$	$(x_{ij}^2\varepsilon_i^2)^{1+\gamma}$ & $(x_{ij}^4)^{1+\gamma}$	$(x_{ij}^2\varepsilon_i^2)^{1+\gamma}$ & $(x_{ij}^2)^{1+\gamma}$
avg. moments converge	yes	yes	yes	no	yes	no
maximum cluster size					unbounded	bounded
distribution of coefficients	yes	yes	yes	no	yes	yes
. . . and Wald statistics	homo.	homo.	no	no	cl/hetero.	cl/hetero.
sub-sampling M of N		$M \to \infty$	$M \to \infty$			$M/N \to 0$ $\underset{inf}{lim}\frac{M}{N^{\gamma^*}} > 0$
moments of coefficients				yes		yes (wild)

Notes: Wald statistics based upon homoskedastic (homo.) or clustered/heteroskedasticity robust (cl/hetero.) covariance estimates.

1. Assumptions on regressors and errors

For an OLS model with iid data and potentially heteroskedastic residuals, Mammen (1993) showed that for a fixed number of regressors the wild bootstrap distributions of linear combinations of the coefficients and Wald statistics based upon the homoskedastic covariance estimate are in probability consistent, given $sup_{\|\boldsymbol{c}\|=1}E\left[(\boldsymbol{c}'\boldsymbol{x}_i)^4(1+\varepsilon_i^2)\right] < \infty$ and the Lindeberg type condition $E\left[(\boldsymbol{c}'\boldsymbol{x}_i)^2\varepsilon_i^2 I\left[(\boldsymbol{c}'\boldsymbol{x}_i)^2\varepsilon_i^2 \geq \gamma N\right]\right] \to 0$ for every fixed $\gamma > 0$. For the same model, Freedman (1981) proved almost-sure consistency of pairs bootstrap coefficients and homoskedastic-based Wald tests if the row vectors $(\boldsymbol{x}_i', y_i)$ are independently and identically distributed and $E\left[((\boldsymbol{x}_i', y_i)(\boldsymbol{x}_i', y_i)')^2\right] < \infty$. Stute (1990) tightened part of the result, showing that almost-sure convergence of the pairs bootstrap coefficients alone for iid data only requires $E(x_{ij}x_{ik})$ and $E(x_{ij}x_{ik}\varepsilon_i^2)$ to be finite. By adopting a permutation approach, this paper proves almost-sure consistency of both coefficients and Wald statistics based upon the clustered/heteroskedasticity robust covariance estimate with inid observations for both the pairs and wild bootstrap with the existence of only slightly higher moments than required by Stute (1990), i.e., $E|x_{ij}x_{ik}|^{1+\gamma} < \infty$ and $E|x_{ij}x_{ik}\varepsilon_i^2|^{1+\gamma} < \infty$ for some $\gamma > 0$. It should be noted, however, that Mammen's result was part of a broader framework that allowed for a growing number of regressors.

For inid data, Liu (1988) proved consistency in probability of the second central moment of the wild OLS bootstrap coefficient distribution with bounded regressors (with all moments) and finite second moments of ε_i. This paper proves almost-sure consistency of the wild bootstrap distribution for inid data with unbounded regressors using the additional moment conditions described above.

Djogbenou et al. (2019) prove consistency in probability of the distribution of the wild bootstrap t-statistic for within-cluster correlated but cross-cluster independent but not identically distributed data. Their assumptions on the existence of moments are those used in this paper, plus the addition of the fourth moment restriction $E|x_{ij}^4|^{1+\gamma} < \infty$ for some $\gamma > 0$. They also impose asymptotic homogeneity of the data-generating process in the form of assuming that $\boldsymbol{X}'\boldsymbol{X}/N$ converges to a matrix of constants, while for any vector $\boldsymbol{\alpha}$, such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$, there exists a finite scalar $v_{\boldsymbol{\alpha}} > 0$ and non-random sequence $\mu_N \to \infty$, such that $\mu_N\boldsymbol{\alpha}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{g=1}^{C} E(\boldsymbol{X}_g'\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_g'\boldsymbol{X}_g)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\alpha} \to v_{\boldsymbol{\alpha}}$. Thus, while papers usually use the iid assumption to motivate the convergence of key matrices to matrices of constants, Djogbenou et al. (2019) avoid the iid assumption but assume that the data nevertheless converge to such matrices. This paper, using clustered versions of White's (1980a) assumptions, requires no such convergence of the asymptotic regressor cross product and covariance matrix of coefficient estimates and as such covers more fundamentally inid data without the addition of the fourth moment condition $E|x_{ij}^4|^{1+\gamma} < \infty$.

This paper makes no explicit assumptions regarding maximum cluster size, but in practice the assumption that the expectation of vector products of the regressors are uniformly bounded for all g, i.e., $E(|\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk}|^{1+\gamma}) < \Delta$, implies that either the maximum cluster size is bounded or, as seems less likely, the expectation of individual observations shrinks with cluster size. In contrast, Djogbenou et al.'s (2019) proof of consistency allows the maximum cluster size to increase with the sample size in an unbounded fashion at a rate determined by the form of dependency (albeit unknown) within clusters. All proofs of consistency necessarily require that asymptotically individual observations or clusters exert a negligible influence on coefficient and variance estimates, although ironically it is often the strong influence of outlier observations or groupings in finite samples that makes conventional tests less accurate relative to the bootstrap (c.f. Davidson & Flachaire, 2008; Young, 2019, and the simulations below).

2. Type of consistency proven

Aside from consistency of the coefficient distribution, Freedman (1981) and Mammen (1993) prove consistency of the Wald statistic for the pairs and wild bootstrap, respectively, based upon the covariance estimate with homoskedastic errors. Djogbenou et al. (2019) prove consistency of the Wald statistic using the cluster/heteroskedasticity robust covariance estimate, which is also asymptotically accurate with homoskedastic errors. This test statistic is asymptotically pivotal and hence provides higher-order asymptotic bootstrap accuracy (Singh, 1981; Hall, 1992). This paper does the same for both the pairs and wild bootstrap using weaker moment conditions and a unified permutation framework that highlights a similarity between the two methods.

Freedman and Stute allowed for sub-sampling $M < N$ observations in the pairs bootstrap and proved convergence in distribution if M and N both go to infinity. As shown in the on-line Appendix, at the expense of complicating the proofs, the permutation-based pairs bootstrap consistency results can be extended to sub-sampling, with and without replacement, if $M/N \to 0$ and for some $\gamma^* > (1 + \gamma)^{-1}$, M is such that $\lim_{\inf} M/N^{\gamma *} > 0$. The requirement that M not fall too rapidly relative to N is needed to ensure the existence and convergence of higher moments to the normal, as the proof of Theorem II is based upon the method of moments.

Liu (1988) proves consistency of the wild bootstrap second central moment with bounded regressors. Proving such consistency with the unbounded regressors of this paper is trivial. If we assume, as did Liu (1988), that $E[\boldsymbol{\delta}^w] = \mathbf{0}_C$ and $E[\boldsymbol{\delta}^w \boldsymbol{\delta}'^w] = \boldsymbol{I}_{CxC}$ (the identity matrix), then taking the expectation with respect to this variable for a given realization of $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$, we have

$$\begin{gathered}
E[\,\hat{\boldsymbol{\beta}}^w_C|\boldsymbol{X},\boldsymbol{\varepsilon}] = \hat{\boldsymbol{\beta}}_C + (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g E[\delta^w_g] = \hat{\boldsymbol{\beta}}_C \quad \left[\text{as } \sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g = \mathbf{0}_K\right] \\
E[(\hat{\boldsymbol{\beta}}^w_C - E[\hat{\boldsymbol{\beta}}^w_C])(\hat{\boldsymbol{\beta}}^w_C - E[\hat{\boldsymbol{\beta}}^w_C])'|\boldsymbol{X},\boldsymbol{\varepsilon}] = \\
(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{C}\sum_{h=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_h'\boldsymbol{X}_h E[\delta^w_g\delta^w_h]\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad (9)\\
= (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{V}\left(\hat{\boldsymbol{\beta}}_C\right),
\end{gathered}$$

where we make use of the fact that $\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g = \boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}_K$ as the OLS estimates $\hat{\boldsymbol{\beta}}_C$ in (2) above minimize $\hat{\boldsymbol{\varepsilon}}_C'\hat{\boldsymbol{\varepsilon}}_C$. Thus, for any sample size the variance of wild bootstrap coefficient estimates equals White's clustered/heteroskedasticity robust covariance estimate for the sample. Since under White's conditions given in Theorem I, $C\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_C)$ is a consistent estimator of the asymptotic variance of $\sqrt{C}(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})$, it follows that for such general inid data the wild bootstrap coefficient variance is a consistent estimator as well, reproducing Liu's result in a more general framework.

A similar result for the pairs bootstrap is more problematic. The first two moments of the multinomial sampling frequencies ($\boldsymbol{\delta}^p$) for C draws with replacement from C cluster groups are $E[\boldsymbol{\delta}^p] = \mathbf{1}_C$ (a vector of ones) and $E[\boldsymbol{\delta}^p\boldsymbol{\delta}'^p] = \boldsymbol{I}_{CxC} - C^{-1}\mathbf{1}_C\mathbf{1}_C'$. Examining the moments of the latter half of $\hat{\boldsymbol{\beta}}^p_C - \hat{\boldsymbol{\beta}}_C = \left(\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{X}_g\delta^p_g\right)^{-1}\left(\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\delta^p_g\right)$, we see:

$$\begin{gathered}
E\left[\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\delta^p_g|\boldsymbol{X},\boldsymbol{\varepsilon}\right] = \sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g E\left[\delta^p_g\right] = \sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g = \mathbf{0}_K, \\
\&\ E\left[\left(\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\delta^p_g\right)\left(\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\delta^p_g\right)'|\boldsymbol{X},\boldsymbol{\varepsilon}\right] = \sum_{g=1}^{C}\sum_{h=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_h'\boldsymbol{X}_h E[\delta^p_g\delta^p_h] \qquad (10)\\
= \sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g - C^{-1}\sum_{g=1}^{C}\sum_{h=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_h'\boldsymbol{X}_h = \sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g = (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{V}\left(\hat{\boldsymbol{\beta}}_C\right)(\boldsymbol{X}'\boldsymbol{X}).
\end{gathered}$$

Were $\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\delta^p_g$ multiplied by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, this would prove consistency of the second central moment of pairs bootstrap coefficients, but unfortunately it is multiplied by $(\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{X}_g\delta^p_g)^{-1}$. However, it is easy to show that $(\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{X}_g\delta^p_g)^{-1}$ converges in probability to $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (see Appendix C below). Using this fact, Shao and Tu (1995) prove consistency of the second central moment using the artifice of assuming that when the minimum eigenvalue of $(\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{X}_g\delta^p_g)^{-1}$ is less than *1/2* of the minimum eigenvalue of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, an event whose probability converges to zero, $\hat{\boldsymbol{\beta}}^p_C$ is set equal to $\hat{\boldsymbol{\beta}}_C$.

It is well known that convergence in distribution does not imply convergence of moments, but the fact that the proof of Theorem II regarding the asymptotic permutation distribution of root-C correlation coefficients is based upon the method of moments (see Hoeffding, 1951 and the on-line Appendix of this paper) might lead to the erroneous conclusion that the results here imply consistency of all moments. They do not, as already implied by the discussion of the second moment of the pairs bootstrap. In Appendix C below, Theorem II is used to prove that across the equally likely permutations $\boldsymbol{d}$ of a given δ^b, for b (both) = p (pairs) or w (wild)

$$\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{\delta}'^b\boldsymbol{O}\boldsymbol{\delta}^b}{C}\right)^{-1/2}\frac{\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\delta_g^b}{\sqrt{C}}\overset{d(\boldsymbol{d})|as(\boldsymbol{\delta}^b,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}\boldsymbol{n}_K, \tag{11}$$

signifying, by the method of proof, that the moments across permutations $\boldsymbol{d}$ of $\boldsymbol{\delta}$ of the left-hand side converge to those of the multivariate standard normal. Since this is true for all $\boldsymbol{\delta}$, such that $\boldsymbol{\delta}'^b\boldsymbol{O}\boldsymbol{\delta}^b > 0$, which almost surely holds (see (L2) in Appendix C), we can equally say that across the distribution of $\boldsymbol{\delta}$, the moments of (11) converge to those of the multivariate standard normal. For the wild bootstrap $\sqrt{C}(\hat{\boldsymbol{\beta}}_C^w - \hat{\boldsymbol{\beta}}_C)$ consists of (11) multiplied by $(\boldsymbol{X}'\boldsymbol{X}/C)^{-1}(\boldsymbol{\delta}'^w\boldsymbol{O}\boldsymbol{\delta}^w/C)^{1/2}(\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g/C)^{1/2}$, and as $\boldsymbol{\delta}'^w\boldsymbol{O}\boldsymbol{\delta}^w/C\overset{as(\boldsymbol{\delta}^w)}{\rightarrow}1$, we can say that all the moments of $\sqrt{C}(\hat{\boldsymbol{\beta}}_C^w - \hat{\boldsymbol{\beta}}_C)$ converge to those of the multivariate normal with covariance matrix $C\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_C)$, although these need not be the asymptotic moments of the sample coefficients $\sqrt{C}(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})$. In the case of the pairs bootstrap, $\sqrt{C}(\hat{\boldsymbol{\beta}}_C^p - \hat{\boldsymbol{\beta}}_C)$ equals (11) multiplied by $(\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{X}_g\delta_g^p/C)^{-1}(\boldsymbol{\delta}'^b\boldsymbol{O}\boldsymbol{\delta}^b/C)^{1/2}(\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g/C)^{1/2}$, and as both $\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{X}_g\delta_g^p/C$ and $\boldsymbol{\delta}'^b\boldsymbol{O}\boldsymbol{\delta}^b/C$ are only shown to converge in probability, nothing can be said about the asymptotic moments of $\sqrt{C}(\hat{\boldsymbol{\beta}}_C^p - \hat{\boldsymbol{\beta}}_C)$ without the use of an artifice such as that of Shao and Tu (1995) mentioned above.

3. Assumptions on the wild bootstrap external variable

Liu (1988) proves the consistency of the second central moment of the wild bootstrap coefficients, assuming that the first and second moments of the wild bootstrap external variable δ_i^w are 0 and 1, respectively.[9] This paper extends the proof to consistency in distribution by additionally requiring that $E\left[(\delta_i^w)^{2(1+\theta_1)}\right] < \infty$ for $\theta_1 > 1/\gamma$, where $\gamma > 0$ is such that $E\,|\,x_{ij}x_{ik}\,|^{1+\gamma} < \infty$ and $E\,|\,x_{ij}x_{ik}\varepsilon_i^2\,|^{1+\gamma} < \infty$. As the proof of Theorem II is based on the method of moments, depending upon the existence of higher moments for the regressors higher moments on δ_i^w are needed to ensure that all moments of (11) above exist and converge to the normal. Proofs of the consistency of wild bootstrap distributions typically assume that the external variable δ_i^w comes from a particular distribution, such as the Rademacher, with moments of all orders (e.g., Mammen, 1993; Canay et al., 2021). A notable exception is Djogbenou et al. (2019), where the proof of convergence in distribution merely requires that $|\delta_i^w|^{2+\lambda} < \infty$ for some $\lambda > 0$. The wild bootstrap external variable, however, is under the control of the practitioner (i.e., not a characteristic of the given data) and, at this time, there appear to be no known advantages to using an external variable without higher moments.

6. Monte Carlo Illustration with INID Data

To illustrate the properties and consistency of the bootstraps with fully inid data, I use a data-generating process that departs strongly from the independently and identically distributed ideal. To ensure average moments do not even begin to converge in finite samples, I model underlying distributional parameters as following a random walk across the data. To stress test the theorems above, I use regressors with heavy-tailed distributions that barely satisfy the specified moment conditions. Finally, to further hinder convergence to the normal, I choose an error distribution that departs strongly from the shape of that ideal.

To begin with unclustered data, for $i = 1 \ldots N$ independent observations, I assume that:

$$\begin{gathered} y_i = \varepsilon_i, \quad \varepsilon_i = B(|a_{\varepsilon i}|, |b_{\varepsilon i}|) - \tfrac{|a_{\varepsilon i}|}{|a_{\varepsilon i}|+|b_{\varepsilon i}|}, \quad x_i = t(2.01 + B(|a_{xi}|, |b_{xi}|)) \\ a_{\varepsilon i} = a_{\varepsilon i-1} + \mathrm{U}[-.5, .5], \ \ b_{\varepsilon i} = b_{\varepsilon i-1} + \mathrm{U}[-.5, .5], \ \ a_{\varepsilon 0} = b_{\varepsilon 0} = \mathrm{U}[-.5, .5], \\ a_{xi} = a_{xi-1} + \mathrm{U}[-.5, .5], \ \ b_{xi} = b_{xi-1} + \mathrm{U}[-.5, .5],\& \ a_{x0} = b_{x0} = \mathrm{U}[-.5, .5], \end{gathered} \tag{12}$$

where $B(a, b)$ denotes an independent draw from the Beta distribution with parameters a and b (and expectation $a/(a + b)$), $t(v)$ an independent draw from the t-distribution with v degrees of freedom, and $\mathrm{U}[-.5, .5]$ an independent draw from the uniform distribution on $[-0.5, 0.5]$. The random walks a and b (separate for ε and x) with their expanding variances ensure that the moments of the data do not meaningfully converge in simulation.[10] These random walks can be thought of as an underlying population characteristic that develops, say, geographically or intertemporally. Observations, however, are drawn independently from these otherwise related distributions. Beta random variables are bounded on [0, 1] and, depending upon a and b, can be heavily skewed toward 0 or 1, symmetric unimodally around 0.5, or bimodally concentrated at both 0 and 1, to name just a few possibilities. This departs strongly from the symmetric unimodal normal distribution on the real line. Random variables drawn from a t-distribution only have finite moments up to their degrees of freedom. Thus, the regressors x_i only have moments between 2.01 and 3.01, approaching the limits of the assumptions in Theorem I.

For a clustered data-generating process, with dependence within clusters, I generate $g = 1 \ \ldots C$ clusters with cluster effects that follow (12) above (substituting g for i everywhere in those equations), and observation level data:

$$\begin{gathered} y_i = \varepsilon_i, \quad \varepsilon_i = \varepsilon_{g(i)} + B\left(\left|a_{\varepsilon g(i)}\right|, \left|b_{\varepsilon g(i)}\right|\right) - \frac{\left|a_{\varepsilon g(i)}\right|}{\left|a_{\varepsilon g(i)}\right| + \left|b_{\varepsilon g(i)}\right|}, \\ \text{and } x_i = x_{g(i)} + t(2.01 + B\left(\left|a_{xg(i)}\right|, \left|b_{xg(i)}\right|\right)), \end{gathered} \tag{13}$$

where $g(i)$ denotes the cluster to which observation i belongs. Thus, each regressor and disturbance observation within a cluster is composed of a common cluster effect plus a similarly, but independently, distributed observation effect. The estimated regression model is:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \ \text{ where } \alpha = \beta = 0. \tag{14}$$

Table 2 reports Monte Carlo results for tests of the true null of $\beta = 0$ in the OLS regression (14) for the data-generating processes described in (12) and (13) with 10, 100, 1000, 10,000, 100,000, and 1,000,000 observations or clusters (and five observations per cluster). For the conventional test, I report the p-value of the two-sided test using the squared sample t-statistic based upon the clustered/heteroskedasticity robust covariance estimate evaluated using its asymptotic chi-squared distribution. For the bootstraps, I report p-values based upon the bootstrap-c, evaluating the squared coefficient deviation from the null using the percentiles of the squared coefficient deviations of the bootstraps from the mean of their data-generating processes, and the bootstrap-t, evaluating the sample squared t-statistic using the corresponding squared bootstrap test statistics, i.e.,

$$\begin{gathered} \text{Bootstrap} - \text{c}: (\hat{\beta} - \beta)^2 \text{ evaluated using } \left(\hat{\beta}^b - \hat{\beta}\right)^2, \\ \text{Bootstrap} - \text{t}: \frac{(\hat{\beta} - \beta)^2}{\hat{V}(\hat{\beta})} \text{ evaluated using } \frac{(\hat{\beta}^b - \hat{\beta})^2}{\hat{V}(\hat{\beta}^b)}. \end{gathered} \tag{15}$$

99 draws are used for each bootstrap, and an exact test relative to the bootstrap distribution is achieved using a p-value given by $(G + (T + 1)\, U[0,1])/100$, where G and T are the number of bootstrap test statistics greater than and equal to, respectively, that of the sample and $U[0,1]$ is a draw from the uniform distribution on [0.1].[11] For the wild bootstrap, δ_c^w is drawn from the Rademacher distribution, which equals ±1 with equal probability and appears to perform better than alternatives (Davidson & Flachaire, 2008). 1000 realizations of the data-generating process are used for each specification.

As can be seen in panel (A) of the table, rejection rates using both the conventional chi-squared distribution and those of the bootstraps differ substantially from nominal value in small samples, but converge to the 0.01, 0.05, and 0.10 levels as the number of observations or clusters increases. The central 95 percentiles of the binomially distributed Monte Carlo rejection probability in 1000 independent draws are 0.004 to 0.016 at the 0.01 level, 0.037 to 0.063 at the 0.05 level, and 0.082 to 0.118 at the 0.1 level. With 1,000,000 independent observations or clusters, most rejection rates lie within those bounds. As shown in panel (B) of the table, the Kolmogorov–Smirnov test statistic for the null that the p-value distributions are uniform, i.e., the maximum absolute difference between the cumulative distribution function of each set of 1000 p-values from that of the uniform distribution, is less than or equal to 0.028 in all cases, with the p-value of the null that the distributions are uniform on [0,1] exceeding 0.41 in each instance.

Theorem V asserts conditional consistency, i.e., the asymptotic distribution of the bootstraps is normal with a covariance matrix equal to that of the conventional estimate. If so, evaluating the conventional test statistic with the full distribution of each bootstrap should asymptotically yield a p-value identical to that found by evaluating the same using the chi-squared distribution.[12] Panel (C) of Table 2 reports the correlation between the bootstrap and conventional p-values with 1,000,000 observations or clusters. As can be seen, this is at least 0.9889 in all cases. As the bootstrap p-values are based upon only a sample from their distribution, while exact relative to that distribution, they cannot be expected to equal the conventional chi-squared p-value. Their correlation with the conventional p-value, however, should be the same as that found when evaluating the conventional test statistic using 99 independent draws from the chi-squared distribution and the same formula $p = (G + (T + 1)\, U[0,1])/100$.

Panel (C) of Table 2 reports that the probability that a correlation less than or equal to that found between the bootstraps and the conventional p-value would be found when evaluating a squared t-statistic using 99 draws from the chi-squared vs. the full chi-squared distribution itself. "p-value1" evaluates the correlation using the distribution conditional on the realized conventional squared t-statistics, i.e., is a test of conditional consistency alone and does not assume consistency of the conventional test. While most p-values are very large, two of the eight are near zero, indicating that the correlation is not yet quite up to the level that would be expected from completed conditional convergence.[13] "p-value2" evaluates the correlations using the distribution across random draws of the initial conventional test statistics from the chi-squared, i.e., a joint test of convergence of the conventional test statistic and conditional consistency of the bootstraps. Here the smallest p-value is 0.045, which, given any adjustment for multiple testing, can be taken as indicating that the tests do not reject the joint null implied by the theorems above at the 0.05 level.

Table 2 illustrates the consistency of bootstrap procedures with a highly challenging data-generating process. While previous results cited above require the existence of at least fourth moments of the regressors for convergence of both coefficients and Wald statistics in environments with iid data or inid data with convergent average moments, no more than slightly higher than second regressor moments are actually sufficient for fully inid data whose moments need not converge to anything, as specified in the theorems above and illustrated in these Monte Carlos.

Table 2. Monte Carlo's illustrating consistency (1000 data sets per data-generating process, 99 bootstrap draws per data set).

	Conventional Chi-Squared			Pairs Bootstrap-c			Pairs Bootstrap-t			Wild Bootstrap-c			Wild Bootstrap-t		
	(A) empirical rejection rates at .01, .05 and .10 levels														
	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10	.01	.05	.10
observations	independent observations														
10	.108	.200	.272	.003	.038	.098	.020	.069	.126	.203	.268	.308	.084	.146	.205
100	.043	.100	.173	.012	.047	.105	.033	.082	.142	.053	.110	.178	.062	.108	.159
1000	.022	.072	.137	.008	.051	.108	.018	.067	.125	.021	.075	.141	.030	.076	.135
10,000	.015	.067	.124	.006	.050	.103	.020	.062	.116	.020	.073	.122	.024	.067	.115
100,000	.016	.063	.127	.017	.059	.123	.017	.060	.126	.011	.070	.124	.017	.068	.121
1,000,000	.012	.065	.113	.009	.050	.104	.012	.056	.111	.013	.068	.113	.014	.069	.110
clusters	independent clusters of observations														
10	.096	.169	.227	.022	.073	.126	.023	.081	.139	.149	.208	.245	.083	.127	.171
100	.030	.076	.136	.007	.045	.095	.018	.059	.104	.037	.088	.131	.037	.084	.118
1000	.015	.062	.116	.005	.045	.094	.012	.061	.109	.018	.060	.118	.020	.053	.108
10,000	.023	.070	.125	.005	.050	.103	.016	.062	.119	.023	.076	.131	.026	.069	.124
100,000	.015	.058	.110	.012	.048	.104	.014	.060	.109	.014	.055	.116	.019	.057	.109
1,000,000	.018	.053	.101	.016	.048	.093	.018	.053	.094	.015	.059	.098	.014	.055	.095

	Conventional Chi-Squared		Pairs Bootstrap-c		Pairs Bootstrap-t		Wild Bootstrap-c		Wild Bootstrap-t	
	(B) Kolmogorov–Smirnov test statistics and p-values (1,000,000 observations or clusters)									
	obs.	clusters	obs.	clusters	obs.	clusters	obs.	clusters	obs.	clusters
statistic	.021	.024	.023	.028	.020	.027	.025	.021	.022	.022
p-value	.744	.616	.678	.413	.815	.437	.573	.747	.697	.720

	Pairs Bootstrap-c		Pairs Bootstrap-t		Wild Bootstrap-c		Wild Bootstrap-t	
	(C) Correlation of p-values with conventional and p-values of said correlations (1,000,000 obs. or cl.)							
	obs.	clusters	obs.	clusters	obs.	clusters	obs.	clusters
correlation	.9907	.9889	.9905	.9895	.9901	.9910	.9891	.9900
p-value1	.466	.001	.287	.027	.092	.849	.000	.162
p-value2	.870	.045	.778	.243	.573	.940	.087	.502

Notes: Author's calculations using Stata version 18.0 code provided in on-line materials. Conventional test incorporates Stata's HC1 correction of the covariance estimate, which substantially reduces the rejection rate in samples with 10 observations. Kolmogorov–Smirnov tests are of the null that each set of 1000 p-values is drawn from the uniform distribution, with the distribution under the null calculated using 100,000 draws of 1000 uniform random variables. p-values of correlations in panel (C) are the likelihood of a smaller correlation in 100,000 instances of using 99 independent draws from the chi-squared to evaluate each of the 1000 chi-squared statistics vs. using the chi-squared distribution itself. As explained in the text, p-value1 calculates the correlation distribution conditional on the realized conventional squared t-statistics, while p-value2 calculates the correlation distribution based on conventional test statistics which are random draws from the chi-squared.

As can also be seen in Table 2, in small samples with heavy-tailed regressors the conventional clustered/heteroskedasticity robust test statistic performs poorly, with rejection probabilities far above the nominal value. In such environments, the bootstraps often perform better (Davidson & Flachaire, 2008; Young, 2019). In the simulations of Table 2, this is clearly the case for the pairs bootstrap and, to a lesser degree, with the wild bootstrap using the asymptotically pivotal t-statistic.[14] Thus, while providing the same asymptotic assurances as conventional inference methods, the bootstraps often provide a better approximation to the distribution of test statistics in small finite sample environments. It should be noted, however, that other methods also exist for improving the finite sample performance of the conventional test, such as the HC bias corrections of MacKinnon and White (1985) and the effective degrees of freedom corrections of Bell and McCaffrey (2002), Pustejovsky and Tipton (2018), and Young (2016).[15]

7. Conclusions

This paper characterizes the pairs and wild bootstraps as realizations of a permutation distribution and uses previously unexploited permutation theorems to derive less restrictive moment conditions for their conditional consistency. While prior work requires at least fourth moments of the regressors for consistency of distributions in an iid framework or inid framework where average moments converge to constants, only slightly more than second moments on the regressors are actually needed for consistency in a fully inid environment where average moments need never converge. The use of the same permutation theorems to characterize and derive new results for the asymptotic distributions of other techniques, such as bootstraps for time series, the jackknife, randomization inference, and conventional estimates on exchangeable data, is the subject of ongoing research.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/econometrics13040041/s1.

Funding: This research received no external funding.

Data Availability Statement: The Stata code used in the simulations is available as Supplementary Material, as well as on the author's website at https://personal.lse.ac.uk/YoungA/ (accessed on 26 September 2025).

Acknowledgments: I am grateful to Taisuke Otsu and anonymous referees for helpful comments.

Conflicts of Interest: The author declares no conflict of interest.

Appendices

In the proofs below, corollaries to Markov's Law of Large Numbers and the Continuous Mapping Theorem given in White (1984), as well as a Borel–Cantelli type corollary by Galambos (1987), will be useful:

Markov's Law Corollary. *Let* z_g *be a sequence of independent random variables such that* $E(|z_g|^{1+\gamma}) < \Delta < \infty$ *for some* $\gamma > 0$ *and all* g. *Then* $m(z_g) - m(E(z_g)) \overset{as}{\to} 0$.

Continuous Mapping Theorem Corollary. *Let* $g\colon \mathbb{R}^{k} \to \mathbb{R}^{l}$ *be continuous on a compact set* $D \subset \mathbb{R}^{k}$. *Suppose that* $b_C(\omega)$ *and* d_C *are kx1 vectors such that* $b_C(\omega)–d_C \overset{as}{\to} 0$*, and for all C sufficiently large,* d_C *is interior to D, uniformly in C. Then* $g\,(b_C(\omega))–g\,(d_C) \overset{as}{\to} 0$.

Borel–Cantelli Corollary. *Let* $x_1, x_2, \ldots$ *be an infinite sequence of random variables,* $F_g(x)$ *the cumulative distribution function of* x_g (i.e., *Prob*($x_g < x$)*, and* u_C *a nondecreasing sequence of real numbers such that for all* g *Prob*($x_g < \sup_C u_C$) = 1*. Then,*

$$\sum_{C=1}^{\infty} [1 - F_C(u_C)] < \infty \Rightarrow \text{Prob}(\max_{g \le C} x_g \ge u_C \text{ infinitely often}) = 0.$$

Appendix A. Proof of Theorem I

This appendix references assumptions (Ia)–(Ic) and results (i)–(v) in Theorem I and uses the notation therein. By (Ib) and (Ic) $\boldsymbol{M}_C$ and $\boldsymbol{V}_C$ are nonsingular with determinant $> \eta$ for all C sufficiently large and their elements uniformly bounded as from Jensen's Inequality:

$$\begin{aligned} E(|\boldsymbol{x}'_{gj}\boldsymbol{x}_{gk}|) &\le \left(E(|\boldsymbol{x}'_{gj}\boldsymbol{x}_{gk}|^{1+\gamma})\right)^{1/(1+\gamma)} \le \Delta^{1/(1+\gamma)} \\ E\left(\left|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{x}_{gk}\right|\right) &\le \left(E(|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{x}_{gk}|^{1+\gamma})\right)^{\frac{1}{1+\gamma}} \le \Delta^{\frac{1}{1+\gamma}}. \end{aligned} \tag{A1}$$

As the sum of the eigenvalues of a matrix equals its trace and the product its determinant, their maximum eigenvalues are less than $K\Delta^{1/(1+\gamma)}$ and their minimum eigenvalues are greater than $\eta/(K\Delta^{1/(1+\gamma)})^{K-1}$ for all C that are sufficiently large. The minimum and maximum eigenvalues of their inverses are the inverses of these. Consequently, for all sufficiently large C, the determinants of their inverses are greater than $(K\Delta^{1/(1+\gamma)})^{-K} > 0$ and, by the spectral decomposition of a real symmetric matrix, the absolute value of their elements is bounded by $(K\Delta^{1/(1+\gamma)})^{K-1}/\eta$.[16] This establishes result (iii) in Theorem I.

Using Jensen's Inequality and (Ic),

$$E(|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g|^{1+\gamma}) \leq \sqrt{E(|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{x}_{gj}|^{1+\gamma})} < \sqrt{\Delta}, \tag{A2}$$

so, by the Markov Corollary, (Ib), (Ic) and the independence of $(\boldsymbol{X}_g, \boldsymbol{\varepsilon}_g)$ across cluster groups (Ia):

$$\begin{gathered}
\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{C} - 0 = \sum_{g=1}^{C} \frac{\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g}{C} - \sum_{g=1}^{C} \frac{E\left(\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g\right)}{C} = \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{Kx1} \\
\frac{\boldsymbol{X}'\boldsymbol{X}}{C} - \boldsymbol{M}_C = \sum_{g=1}^{C} \frac{\boldsymbol{X}'_g\boldsymbol{X}_g}{C} - \sum_{g=1}^{C} \frac{E\left(\boldsymbol{X}'_g\boldsymbol{X}_g\right)}{C} \overset{as(\boldsymbol{X})}{\rightarrow} \mathbf{0}_{KxK} \\
\sum_{g=1}^{C} \frac{\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{X}_g}{C} - \boldsymbol{V}_C = \sum_{g=1}^{C} \frac{\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{X}_g}{C} - \sum_{g=1}^{C} \frac{E(\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{X}_g)}{C} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{KxK} \\
\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1} - \boldsymbol{M}_{\boldsymbol{C}}^{-1} \overset{as(\boldsymbol{X})}{\rightarrow} \mathbf{0}_{KxK},
\end{gathered} \tag{A3}$$

where the last follows from the Continuous Mapping Theorem Corollary. These results, combined with the boundedness of $\boldsymbol{M}_C^{-1}$, establish result (i) in Theorem I:

$$\hat{\boldsymbol{\beta}}_C = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1} \frac{\boldsymbol{X}'\boldsymbol{y}}{C} = \boldsymbol{\beta} + \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1} \frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{C} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{\beta}. \tag{A4}$$

$\boldsymbol{X}'\boldsymbol{\varepsilon}/\sqrt{C}$ is a vector with expectation and variance:

$$E\left(\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{\sqrt{C}}\right) = \sum_{g=1}^{C} \frac{E(\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g)}{\sqrt{C}} = 0_{Kx1}, \quad E\left(\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}}{C}\right) = \sum_{g=1}^{C} \frac{E(\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{X}_g)}{C} = \boldsymbol{V}_C. \tag{A5}$$

As noted in White (1980a, p. 829—see also White, 1980b; Hoadley, 1971), given (A5), a multivariate Liapounov Central Limit theorem implies that $\boldsymbol{V}_C^{-1/2}\boldsymbol{X}'\boldsymbol{\varepsilon}/\sqrt{C}$ is asymptotically distributed multivariate standard normal, $\boldsymbol{n}_K$, provided that for all $\boldsymbol{\kappa}$ in $\mathbb{R}^K$ and some $\delta > 0$:

$$\sum_{g=1}^{C} \frac{E(|\boldsymbol{\kappa}'\boldsymbol{V}_C^{-1/2}\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g|^{2+\delta})}{C^{(2+\delta)/2}} \rightarrow 0. \tag{A6}$$

Define $\boldsymbol{\varphi} = \boldsymbol{\kappa}'\boldsymbol{V}_C^{-1/2}$, and note that by the properties of the Rayleigh quotient $\boldsymbol{\varphi}'\boldsymbol{\varphi} \leq \boldsymbol{\kappa}'\boldsymbol{\kappa}/\lambda_{min}$, where $\lambda_{min} = \eta/(K\Delta^{1/(1+\gamma)})^{K-1}$ is the lower bound on the minimum eigenvalue of $\boldsymbol{V}_C$ $(1/\lambda_{min}$ the upper bound on the maximum eigenvalue of $\boldsymbol{V}_C^{-1})$ given earlier above. Keeping in mind then that the kth element of $\boldsymbol{\varphi}$, φ_k, is bounded, and noting that from (Ic) $E(|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{x}_{gj}|^{1+\gamma}) = E(|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g|^{2+2\gamma}) < \Delta$, we apply Minkowski's Inequality:

$$\begin{gathered}
E\left(|\boldsymbol{\kappa}'\boldsymbol{V}_C^{-1/2}\boldsymbol{X}'_g\boldsymbol{\varepsilon}_g|^{2+2\gamma}\right) = E\left(\left|\sum_{k=1}^{K} \varphi_k\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g\right|^{2+2\gamma}\right) \\
\leq \left(\sum_{k=1}^{K} \left[E(|\varphi_k\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g|^{2+2\gamma})\right]^{\frac{1}{2+2\gamma}}\right)^{2+2\gamma} < \left(\sum_{k=1}^{K} \left[|\varphi_k|^{2+2\gamma}\Delta\right]^{\frac{1}{2+2\gamma}}\right)^{2+2\gamma} < \infty,
\end{gathered} \tag{A7}$$

so (A6) holds with $\delta = 2\gamma$. Consequently, we can say that

$$\mathbf{V}_C^{-1/2}\boldsymbol{M}_C\sqrt{C}\left(\hat{\boldsymbol{\beta}}_C-\boldsymbol{\beta}\right)=\mathbf{V}_C^{-1/2}\boldsymbol{M}_C\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1}\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{\sqrt{C}}\overset{d(\boldsymbol{X},\boldsymbol{\varepsilon})}{\to}\boldsymbol{n}_K,\quad \text{as}\, \boldsymbol{M}_C\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1}\overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\to}\boldsymbol{I}_K. \tag{A8}$$

This establishes result (ii) in Theorem I.

As $\hat{\boldsymbol{\varepsilon}}_g=\boldsymbol{\varepsilon}_g+\boldsymbol{X_g}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_C)$, the jk^{th} element of $\sum_{g=1}^{C}\boldsymbol{X}_g'\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_g'\boldsymbol{X}_g/C$ equals:

$$\begin{gathered}\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g}{C}=\sum_{r=1}^{K}\sum_{s=1}^{K}\frac{(\beta_r-\hat{\beta}_r)}{C^{(\theta-1)/2}}\frac{(\beta_s-\hat{\beta}_s)}{C^{(\theta-1)/2}}\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gs}}{C^{2-\theta}}\\ +\sum_{r=1}^{K}\frac{(\beta_r-\hat{\beta}_r)}{C^{(\theta-1)/2}}\left(\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{\varepsilon}_g}{C^{1+(1-\theta)/2}}+\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gk}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gj}'\boldsymbol{\varepsilon}_g}{C^{1+(1-\theta)/2}}\right)+\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\boldsymbol{\varepsilon}_g\boldsymbol{x}_{gk}'\boldsymbol{\varepsilon}_g}{C},\end{gathered} \tag{A9}$$

where we select θ such that $\gamma/(1+\gamma) > \theta > 0$, with γ as in (Ib) and (Ic). Repeatedly applying the Cauchy–Schwarz Inequality, we have

$$\begin{gathered}\left|\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gs}}{C^{2-\theta}}\right|\le\sqrt{\sum_{g=1}^{C}\frac{\left(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\right)^2}{C^{2-\theta}}\sum_{g=1}^{C}\frac{\left(\boldsymbol{x}_{gk}'\boldsymbol{x}_{gs}\right)^2}{C^{2-\theta}}}\le\\ \sqrt{\sum_{g=1}^{C}\frac{(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj})(\boldsymbol{x}_{gr}'\boldsymbol{x}_{gr})}{C^{2-\theta}}\sum_{g=1}^{C}\frac{(\boldsymbol{x}_{gk}'\boldsymbol{x}_{gk})(\boldsymbol{x}_{gs}'\boldsymbol{x}_{gs})}{C^{2-\theta}}}\le\sqrt{\prod_{i=j,k}\underset{g\le C}{max}\frac{\boldsymbol{x}_{gi}'\boldsymbol{x}_{gi}}{C^{1-\theta}}\prod_{i=r,s}\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gi}'\boldsymbol{x}_{gi}}{C}}\\ \left|\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{\varepsilon}_g}{C^{1+(1-\theta)/2}}\right|\le\sqrt{\underset{g\le C}{max}\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj}}{C^{1-\theta}}\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gr}'\boldsymbol{x}_{gr}}{C}\sum_{g=1}^{C}\frac{(\boldsymbol{x}_{gk}'\boldsymbol{\varepsilon}_g)^2}{C}}\end{gathered} \tag{A10}$$

Using Markov's Inequality and $E(|\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk}|^{1+\gamma}) < \Delta$ in (Ib), we can state that for any $\delta > 1/(1+\gamma)$ but $< 1-\theta$

$$\sum_{C=1}^{\infty}\text{Prob}(\boldsymbol{x}_{Cj}'\boldsymbol{x}_{Cj})\ge C^{\delta})\le\sum_{C=1}^{\infty}\frac{E(|\boldsymbol{x}_{Cj}'\boldsymbol{x}_{Cj}|^{1+\gamma})}{C^{\delta(1+\gamma)}}<\sum_{C=1}^{\infty}\frac{\Delta}{C^{\delta(1+\gamma)}}<\infty. \tag{A11}$$

So, by the Borel–Cantelli Corollary, $\max_{g\le C}\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj}$ is asymptotically almost surely less than C^{δ} and hence $\max_{g\le C}\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj}/C^{1-\theta}$ almost surely converges to zero for 1-θ > $1/(1+\gamma)$, i.e., 0 < θ < $\gamma/(1+\gamma)$. Together with (A3)'s results, that $\sum_{g=1}^{C}\boldsymbol{x}_{gp}'\boldsymbol{x}_{gr}/C$ and $\sum_{g=1}^{C}(\boldsymbol{x}_{gk}'\boldsymbol{\varepsilon}_g)^2/C$ almost surely converge to bounded elements of $\boldsymbol{M}_C$ and $\boldsymbol{V}_C$, this establishes that both left-hand side terms in (A10) almost surely converge to 0. Results (i)–(iii) show that $\sqrt{C}(\beta_r-\hat{\beta}_r)$ is asymptotically normally distributed with mean zero and bounded variance less than some $\sigma^2 > 0$. Hence, asymptotically, the probability $\left|\sqrt{C}(\beta_r-\hat{\beta}_r)\right| > C^{\delta}$ for any $\delta > 0$ and $< \theta$ can be bounded by

$$\frac{2}{\sqrt{2\pi\sigma^2}}\int_{C^{\delta}}^{\infty}exp\left(\frac{-x^2}{2\sigma^2}\right)dx<\frac{2}{\sqrt{2\pi\sigma^2}}\int_{C^{\delta}}^{\infty}\frac{x}{C^{\delta}}exp\left(\frac{-x^2}{2\sigma^2}\right)dx=\frac{2\sigma^2}{\sqrt{2\pi\sigma^2}}\frac{1}{C^{\delta}}\exp\left(\frac{-C^{2\delta}}{2\sigma^2}\right), \tag{A12}$$

which is less than $1/C^{1+\delta}$ for all C that are sufficiently large. So,

$$\sum_{C=1}^{\infty}\text{Prob}(|\sqrt{C}(\beta_r-\hat{\beta}_r)|\ge C^{\delta})<\infty \tag{A13}$$

and by the Borel–Cantelli Lemma, $\sqrt{C}(\beta_r - \hat{\beta}_r)/C^{\theta} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0$. From (A3), the last $\sum_{g=1}^{C} \boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g/C$ term in (A9) almost surely converges to the jk^{th} term of $\boldsymbol{V}_{C.}$ Putting all of the above together, we see that

$$\begin{gathered}\sum_{g=1}^{C}\frac{\boldsymbol{X}'_g\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}'_g\boldsymbol{X}_g}{C} - \boldsymbol{V}_C \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{0}_{KxK} \text{ and } \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1} - \boldsymbol{M}_{\boldsymbol{C}}^{-1} \overset{as(\boldsymbol{X})}{\rightarrow} \boldsymbol{0}_{KxK}[\text{(A3) above}], \\ \text{so } C\hat{\boldsymbol{V}}\left(\hat{\boldsymbol{\beta}}_C\right) - \boldsymbol{M}_N^{-1}\boldsymbol{V}_N\boldsymbol{M}_N^{-1} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{0}_{KxK},\end{gathered} \tag{A14}$$

establishing (iv) in Theorem I. Result (v) follows from (i)–(iv).

Appendix B. Proof of Theorem IV

If either the z_g or δ_g are all identical ($z_g = z$ or $\delta_g = \delta$), Theorem IV follows immediately. Assuming this is not the case, we use the symmetry and equal likelihood of permutations to calculate the expectation of d_g and products of d_g across the row permutations $\boldsymbol{d}$ of δ:

$$\begin{gathered}E_{\boldsymbol{d}}(d_g) = \sum_{g=1}^{C}\frac{\delta_g}{C} = m(\delta_g),\ E_{\boldsymbol{d}}(d_g^2) = \sum_{g=1}^{C}\frac{\delta_g^2}{C} = m(\delta_g^2) \\ \&\ E_{\boldsymbol{d}}(d_g d_{h\neq g}) = \sum_{g=1}^{C}\sum_{h=1}^{C}\frac{\delta_g\delta_h}{C(C-1)} - \sum_{g=1}^{C}\frac{\delta_g^2}{C(C-1)} = \frac{m(\delta_g)^2C}{C-1} - \frac{m(\delta_g^2)}{C-1}.\end{gathered} \tag{B1}$$

The mean and variance of $m(z_g d_g) - m(z_g)m(d_g)$ across realizations of $\boldsymbol{d}$ are given by:

$$\begin{aligned}E_{\boldsymbol{d}}(m(z_gd_g) - m(z_g)m(d_g)) &= \sum_{g=1}^{C}\frac{z_gE_{\boldsymbol{d}}(d_g)}{C} - m(z_g)m(\delta_g) = 0, \\ &E_{\boldsymbol{d}}((m(z_gd_g) - m(z_g)m(d_g))^2) \\ &= \sum_{g\neq h=1}^{C}\frac{z_gz_hE_{\boldsymbol{d}}(d_gd_h)}{C^2} + \sum_{g=1}^{C}\frac{z_g^2E_{\boldsymbol{d}}(d_g^2)}{C^2} - m(z_g)^2m(\delta_g)^2 \\ = \left(\frac{m(\delta_g)^2C}{C-1} - \frac{m(\delta_g^2)}{C-1}\right)&\left(\sum_{g=1}^{C}\sum_{h=1}^{C}\frac{z_gz_h}{C^2} - \sum_{g=1}^{C}\frac{z_g^2}{C^2}\right) + m(\delta_g^2)\sum_{g=1}^{C}\frac{z_g^2}{C^2} - m(z_g)^2m(\delta_g)^2 \\ = \left(\frac{m(\delta_g)^2C}{C-1} - \frac{m(\delta_g^2)}{C-1}\right)&\left(m(z_g)^2 - \frac{m(z_g^2)}{C}\right) + m(\delta_g^2)\frac{m(z_g^2)}{C} - m(z_g)^2m(\delta_g)^2 \\ &= \frac{[m(z_g^2) - m(z_g)^2][m(\delta_g^2) - m(\delta_g)^2]}{C-1},\end{aligned} \tag{B2}$$

where $\Sigma_{g\neq h}$ denotes the summation across the two indices, excluding ties between them. The last line shows that if (IVb) holds, then across the permutations $\boldsymbol{d}$ of δ $m(z_gd_g) - m(z_g)m(d_g)$ converges in mean square and hence in probability to 0, as stated in Theorem IV.

Appendix C. Proof of Theorem V

We begin by noting the following Lemma, proven in Appendix D further below.

Lemma 1. *Let $\overset{as(\delta)}{\rightarrow}$ or $\overset{p(\delta)}{\rightarrow}$ denote convergence almost surely or in probability across the distribution of δ, τ any integer greater than 2, b (both) = p (pairs) or w (wild), γ > 0 be as given in Theorem I, θ_1 > 0 as in Theorem V, and η_1 and κ constants > 0. For all θ such that $\gamma/(1+\gamma) > \theta$ > 0 (pairs) or $\gamma/(1+\gamma) > \theta > 1/(1+\theta_1)$ (wild):*

$$m(\delta_g^w) \overset{as(\delta^w)}{\rightarrow} 0,\ m((\delta_g^w)^2) \overset{as(\delta^w)}{\rightarrow} 1\ \&\ C^{-\theta}m((\delta_g^w)^4) \overset{as(\delta^w)}{\rightarrow} 0 \tag{L1w}$$

$$m(\delta_g^p) = 1,\ m((\delta_g^p)^2) \overset{p(\delta^p)}{\rightarrow} 2,\ \&\ C^{-\theta}m((\delta_g^p)^2) \overset{as(\delta^p)}{\rightarrow} 0 \tag{L1p}$$

$$\textit{almost surely for all C sufficiently large} \sum_{g=1}^{C} \frac{[\delta_g^b - m(\delta_g^b)]^2}{C} > \kappa > 0 \tag{L2}$$

$$\frac{C^{(1-\theta)(\frac{\tau}{2}-1)}\sum_{g=1}^{C}[\delta_g^b - m(\delta_g^b)]^\tau}{\left(\sum_{g=1}^{C}[\delta_g^b - m(\delta_g^b)]^2\right)^{\tau/2}} \overset{as(\delta^b)}{\rightarrow} 0 \tag{L3}$$

$$\begin{gathered}\textit{almost surely for all C sufficiently large } \boldsymbol{X'X/C}, \sum_{g=1}^{C} \boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g/C, \\ \textit{and their inverses are bounded and positive definite with determinant } > \eta_1 > 0\end{gathered} \tag{L4}$$

$$\forall\ k\ \&\ \tau: \frac{C^{\theta(\frac{\tau}{2}-1)}\left|\sum_{g=1}^{C}\left(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g\right)^\tau\right|}{\left(\sum_{g=1}^{C}\left(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g\right)^2\right)^{\tau/2}} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0 \tag{L5}$$

$$\forall\ j,k: \frac{m(\left(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk}\right)^2)}{C^{1-\theta}} \overset{as(\boldsymbol{X})}{\rightarrow} 0 \tag{L6}$$

$$\forall\ j,k: \frac{m(\left(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk}\right)^4)}{C^{3-3\theta}} \overset{as(\boldsymbol{X})}{\rightarrow} 0 \tag{L7}$$

Use of θ, θ_1, and γ below follows the bounds and definitions in Lemma 1 and Theorems I and V earlier.

For a permutation $\boldsymbol{d}$ of $\boldsymbol{\delta}^w$ or $\boldsymbol{\delta}^p$, the coefficient estimates of the pairs and wild bootstrap are, following (4) and (5) in the text, given by

$$\begin{gathered}\sqrt{C}\left(\hat{\boldsymbol{\beta}}_C^p - \hat{\boldsymbol{\beta}}_C\right) = \boldsymbol{A}^{-1}\boldsymbol{a} \text{ and } \sqrt{C}(\hat{\boldsymbol{\beta}}_C^w - \hat{\boldsymbol{\beta}}_C) = (\boldsymbol{X'X}/C)^{-1}\boldsymbol{a}, \\ \text{where } \boldsymbol{A} = \sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\boldsymbol{X}_g}{C}d_g \text{ and } \boldsymbol{a} = \sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g}{\sqrt{C}}d_g.\end{gathered} \tag{C1}$$

Our objective is to describe the distribution of these objects across permutations $\boldsymbol{d}$ given the realization of conditions on $\boldsymbol{\delta}$, $\boldsymbol{X}$, and $\boldsymbol{\varepsilon}$. When results apply to both bootstraps, we use the notation $\boldsymbol{d}$; when they apply to only one bootstrap, we use the notation $\boldsymbol{d}^w$ or $\boldsymbol{d}^p$.

Regarding the jk^{th} element of $\mathbf{A}$, given by $\sum_{g=1}^{C}\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk}d_g^p/C$, we can apply Theorem IV with $z_g = \boldsymbol{x}_{gj}'\boldsymbol{x}_{gk}$. Condition IVb in this case requires that:

$$\left[\frac{m(\left(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk}\right)^2) - m(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk})^2}{C^{1-\theta}}\right]\left[\frac{m((\delta_g^p)^2) - m(\delta_g^p)^2}{C^\theta}\right] \overset{as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0, \tag{C2}$$

which is guaranteed by (L1p), (L4), and (L6) above. So,

$$\boldsymbol{A} - \frac{\boldsymbol{X'X}}{C}\underbrace{m(\delta_g^p)}_{=1} \overset{p(\boldsymbol{d})|as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{KxK}, \tag{C3}$$

and by the corollary to the Continuous Mapping Theorem given above, $\boldsymbol{A}^{-1}$ converges in probability to bounded positive definite $(\boldsymbol{X'X}/C)^{-1}$ (as in L4).

Regarding $\boldsymbol{a}$, we apply Theorem IIm in the text with $\boldsymbol{Z}' = \{z_1, \ldots, z_C\}$ and $z_g = \boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g$. Since $\boldsymbol{Z}'\mathbf{1}_C = \boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} = \mathbf{0}_K$, the mean of z_g is zero, and so we have $\boldsymbol{Z}'\boldsymbol{O}\boldsymbol{Z} = \sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g$ and $\boldsymbol{Z}'\boldsymbol{O}\boldsymbol{d} = \sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g d_g$. From (L2) we know that almost surely $\boldsymbol{d}'\boldsymbol{O}\boldsymbol{d}/C = \boldsymbol{\delta}'^b\boldsymbol{O}\boldsymbol{\delta}^b/C > \kappa > 0$, while (L4) ensures that $\boldsymbol{Z}'\boldsymbol{O}\boldsymbol{Z}/C$ is positive definite with determinant > η_1 > 0. Hence, following Theorems II and III, the distribution across $\boldsymbol{d}$ of

$$\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{d}'\boldsymbol{O}\boldsymbol{d}}{C}\right)^{-1/2}\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g d_g}{\sqrt{C}} \tag{C4}$$

converges almost surely (across δ^b,$\boldsymbol{X}$,$\boldsymbol{\varepsilon}$) to that of the iid multivariate standard normal, as by (L3) and (L5) condition (IIb) in Theorem II, holds for all elements z_{gk} in $\boldsymbol{z}_g$.

Using (L4) and the fact that $\boldsymbol{\delta}'^b\boldsymbol{O}\boldsymbol{\delta}^b/C = \boldsymbol{d}'\boldsymbol{O}\boldsymbol{d}/C$ is a scalar, we then have:

$$\begin{gathered}
\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)\left(\frac{\boldsymbol{\delta}^p{\prime}\boldsymbol{O}\boldsymbol{\delta}^p}{C}\right)^{-1/2}\sqrt{C}\left(\hat{\boldsymbol{\beta}}_C^p-\hat{\boldsymbol{\beta}}_C\right) \\
=\underbrace{\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)\boldsymbol{A}^{-1}\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{1/2}}_{\overset{p(\boldsymbol{d})|as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}\boldsymbol{I}_{KxK}} \\
*\underbrace{\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{d}^p{\prime}\boldsymbol{O}\boldsymbol{d}^p}{C}\right)^{-1/2}\boldsymbol{a}}_{\overset{d(\boldsymbol{d})|as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}\boldsymbol{n}_K}\overset{d(\boldsymbol{d})|as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}\boldsymbol{n}_K, \\
\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)\left(\frac{\boldsymbol{\delta}^w{\prime}\boldsymbol{O}\boldsymbol{\delta}^w}{C}\right)^{-1/2}\sqrt{C}\left(\hat{\boldsymbol{\beta}}_C^w-\hat{\boldsymbol{\beta}}_C\right) \\
=\left(\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\right)^{-1/2}\left(\frac{\boldsymbol{d}^w{\prime}\boldsymbol{O}\boldsymbol{d}^w}{C}\right)^{-1/2}\boldsymbol{a}\overset{d(\boldsymbol{d})|as(\delta^w,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}\boldsymbol{n}_K,
\end{gathered}\tag{C5}$$

thereby establishing the claim in (Va).

Regarding the wild bootstrap clustered/heteroskedasticity robust covariance estimates, for a permutation $\boldsymbol{d}^w$ of $\boldsymbol{\delta}^w$, we have

$$C\hat{V}\left(\hat{\boldsymbol{\beta}}_C^w\right)=\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1}\boldsymbol{W}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1},\text{ where }\boldsymbol{W}=\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g^w\hat{\boldsymbol{\varepsilon}}_g^{w\prime}\boldsymbol{X}_g}{C}.\tag{C6}$$

Using $\hat{\boldsymbol{\varepsilon}}_g^w=\boldsymbol{X}_g\left(\hat{\boldsymbol{\beta}}_C-\hat{\boldsymbol{\beta}}_C^w\right)+\hat{\boldsymbol{\varepsilon}}_g d_g^w$, from (4) in the text, the jk^{th} element of $\boldsymbol{W}$ is given by:

$$\begin{gathered}
\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g^w\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g^w}{C}=\underbrace{m\left(\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g\left(d_g^w\right)^2\right)}_{''a''}+\sum_{r=1}^{K}\sum_{s=1}^{K}\frac{\boldsymbol{\delta}^w{\prime}\boldsymbol{O}\boldsymbol{\delta}^w}{C^{1+\theta}}\hat{\eta}_r\hat{\eta}_s\underbrace{m\left(\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gs}}{C^{1-\theta}}\right)}_{''c''} \\
-\sum_{r=1}^{K}\sqrt{\frac{\boldsymbol{\delta}^w{\prime}\boldsymbol{O}\boldsymbol{\delta}^w}{C^{1+\theta}}}\hat{\eta}_p\left[\underbrace{m\left(\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g d_g^w}{C^{(1/2)(1-\theta)}}\right)+m\left(\frac{\boldsymbol{x}_{gk}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g d_g^w}{C^{(1/2)(1-\theta)}}\right)}_{''b''}\right] \\
\left[\text{where }\hat{\boldsymbol{\eta}}=\left(\frac{\boldsymbol{d}{\prime}\boldsymbol{O}\boldsymbol{d}}{C}\right)^{-1/2}\sqrt{C}(\hat{\boldsymbol{\beta}}_C^w-\hat{\boldsymbol{\beta}}_C)\right].
\end{gathered}\tag{C7}$$

For "*a*", we note that $\left(d_g^w\right)^2$ is the permutation of $\left(\delta_g^w\right)^2$ and apply Theorem IV with $z_g=\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g$. Condition (IVb) requires that:

$$\left[\frac{m((\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2)-m(\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2}{C^{1-\theta}}\right]\left[\frac{m((\delta_g^w)^4)-m((\delta_g^w)^2)^2}{C^{\theta}}\right]\overset{as(\delta^w,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}0.\tag{C8}$$

From (L1w) and (L4), we know that $[m((\delta_g^w)^4) - m((\delta_g^w)^2)^2]/C^\theta$ and $m(\boldsymbol{x}'_{gj}\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2/C^{1-\theta} \overset{as(\delta^w,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0$. Applying the Cauchy–Schwarz Inequality (here, and frequently below),

$$\frac{m((\boldsymbol{x}'_{gj}\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2)}{C^{1-\theta}} = \sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gj}\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2}{C^{2-\theta}} \le \sqrt{\prod_{i=j,k}\sum_{g=1}^{C}\frac{\left(\boldsymbol{x}'_{gi}\hat{\boldsymbol{\varepsilon}}_g\right)^4}{C^{2-\theta}}} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0, \tag{C9}$$

where the last is guaranteed by (L4) and (L5) as

$$\sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gi}\hat{\boldsymbol{\varepsilon}}_g)^4}{C^{2-\theta}} = \frac{\overbrace{C^{\theta(\frac{4}{2}-1)}\sum_{g=1}^{C}(\boldsymbol{x}'_{gi}\hat{\boldsymbol{\varepsilon}}_g)^4}^{\overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}0\text{ (L5 with }\tau=4)}}{\left(\sum_{g=1}^{C}(\boldsymbol{x}'_{gi}\hat{\boldsymbol{\varepsilon}}_g)^2\right)^{4/2}}\overbrace{\left(\sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gi}\hat{\boldsymbol{\varepsilon}}_g)^2}{C}\right)^2}^{as\text{ bounded (L4)}} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0. \tag{C10}$$

So, by Theorem IV,

$$\text{“}a\text{''} : m(\boldsymbol{x}'_{gj}\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g\left(d_g^w\right)^2) - m(\boldsymbol{x}'_{gj}\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)\underbrace{m((\delta_g^w)^2)}_{\overset{as(\delta^w)}{\rightarrow}1\text{ (L1w)}} \overset{p(\boldsymbol{d})|as(\delta^w,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0. \tag{C11}$$

For “b”, we apply Theorem IV with $z_g = \boldsymbol{x}'_{gj}\hat{\boldsymbol{x}}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g/C^{(1/2)(1-\theta)}$, so condition (IVb) requires that

$$\frac{m((\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2/C^{1-\theta}) - m(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g/C^{(1/2)(1-\theta)})^2}{C^{1-\theta}}\left[\frac{m(\delta_i^{w2}) - m(\delta_i^w)^2}{C^\theta}\right] \overset{as(\delta^w,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0. \tag{C12}$$

Using (L1w) and

$$\left|m\left(\frac{\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g}{C^{(1/2)(1-\theta)}}\right)\right| = \left|\sum_{g=1}^{C}\frac{\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g}{C^{1+(1/2)(1-\theta)}}\right| \le \underbrace{\sqrt{\sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr})^2}{C^{2-\theta}}}}_{as(\boldsymbol{X},\boldsymbol{\varepsilon})\rightarrow 0\text{ (L6)}}\underbrace{\sqrt{\sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2}{C}}}_{as\text{ bounded (L4)}} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0, \tag{C13}$$

$$\frac{m((\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2/C^{1-\theta})}{C^{1-\theta}} = \sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2}{C^{3-2\theta}} \le \underbrace{\sqrt{\sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr})^4}{C^{4-3\theta}}}}_{as(\boldsymbol{X},\boldsymbol{\varepsilon})\rightarrow 0\text{ (L7)}}\underbrace{\sqrt{\sum_{g=1}^{C}\frac{\left(\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g\right)^4}{C^{2-\theta}}}}_{as(\boldsymbol{X},\boldsymbol{\varepsilon})\rightarrow 0\text{ (C10)}} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0, \tag{C14}$$

we see that condition (IVb) is met and by Theorem IV we then have

$$b'' : m\left(\frac{\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g d_g^w}{C^{(1/2)(1-\theta)}}\right) - \underbrace{m\left(\frac{\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g}{C^{(1/2)(1-\theta)}}\right)}_{as(\boldsymbol{X},\boldsymbol{\varepsilon})\rightarrow 0\text{(C13)}}\underbrace{m(\delta_i^w)}_{as(\delta^w)\rightarrow 0\text{ (L1w)}} \overset{p(\boldsymbol{d})|as(\delta^w,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0. \tag{C15}$$

Finally, for “c”, we note that

$$\left|m\left(\frac{\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\boldsymbol{x}_{gs}}{C^{1-\theta}}\right)\right| \le \sqrt{\frac{m(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr})^2}{C^{1-\theta}}\frac{m(\boldsymbol{x}'_{gk}\boldsymbol{x}_{gs})^2}{C^{1-\theta}}}\underbrace{\overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0}_{\text{by (L6)}}. \tag{C16}$$

From the above, we see that the $\hat{\eta}_r$ in (C7) are multiplied by $\sqrt{\delta^{w\prime}\boldsymbol{O}\delta^{w}/C^{1+\theta}}$ which from (L1w) converges almost surely (across δ^{w}) to 0, "c" terms which almost surely (across $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$) converge to 0, and "b" terms which also almost surely (across δ^{w}, $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$) converge in probability across permutations $\boldsymbol{d}^{w}$ to zero. As the $\hat{\eta}_r$, from (L4) and (C5) almost surely (across δ^{w}, $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$) converge in distribution across permutations $\boldsymbol{d}^{w}$ of δ^{w} to normal variables with bounded variance, it follows that when so multiplied they converge in probability across permutations $\boldsymbol{d}^{w}$ to zero. This leaves only the "a" term, and consequently, using (C10), we see that

$$\boldsymbol{W}-\sum_{g=1}^{C}\frac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g}{C}\overset{p(\boldsymbol{d})|as(\delta^{w},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}0_{KxK}\text{ and so }C\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_C^{w}\right)-C\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_C\right)\overset{p(\boldsymbol{d})|as(\delta^{w},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}\mathbf{0}_{KxK}, \tag{C17}$$

which establishes (Vb) for the wild bootstrap.

For the pairs bootstrap clustered/heteroskedasticity robust covariance estimates, for a permutation $\boldsymbol{d}^{p}$ of δ^{p}, we have from (5),

$$\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_C^{p}\right)=\boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{A}^{-1},\boldsymbol{A}=\textstyle\sum_{g=1}^{C}\dfrac{\boldsymbol{X}_g'\boldsymbol{X}_g}{C}d_g^{p}\text{and }\boldsymbol{P}=\textstyle\sum_{g=1}^{C}\dfrac{\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g^{p}\hat{\boldsymbol{\varepsilon}}_g^{p\prime}\boldsymbol{X}_g}{C}d_g^{p}. \tag{C18}$$

Using $\hat{\boldsymbol{\varepsilon}}_g^{p}=\boldsymbol{X}_g\left(\hat{\boldsymbol{\beta}}_C-\hat{\boldsymbol{\beta}}_C^{p}\right)+\hat{\boldsymbol{\varepsilon}}_g$ given in (5) earlier, the jk^{th} element of $\boldsymbol{P}$ is given by

$$\begin{gathered}\sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g^{p}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g^{p}d_g^{p}}{C}=\underbrace{m\left(\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g d_g^{p}\right)}_{''d''}+\sum_{r=1}^{K}\sum_{s=1}^{K}\frac{\delta^{p\prime}\boldsymbol{O}\delta^{p}}{C^{1+\theta}}\hat{\eta}_r\hat{\eta}_s\underbrace{m\left(\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gs}d_g^{p}}{C^{1-\theta}}\right)}_{''f''}\\ -\sum_{r=1}^{K}\sqrt{\frac{\delta^{p\prime}\boldsymbol{O}\delta^{p}}{C^{1+\theta}}}\hat{\eta}_r\left[\underbrace{m\left(\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g d_g^{p}}{C^{(1/2)(1-\theta)}}\right)+m\left(\frac{\boldsymbol{x}_{gk}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g d_g^{p}}{C^{(1/2)(1-\theta)}}\right)}_{''e''}\right]\\ \left[\text{where }\hat{\boldsymbol{\eta}}=\left(\frac{\boldsymbol{d}'\boldsymbol{O}\boldsymbol{d}}{C}\right)^{-1/2}\sqrt{C}(\hat{\boldsymbol{\beta}}_C^{p}-\hat{\boldsymbol{\beta}}_C)\right].\end{gathered} \tag{C19}$$

For "d", we apply Theorem IV with $z_g=\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g$ and, as by (L1p), (L4), and (C10) $[m((\delta_g^{p})^2)-m(\delta_g^{p})^2]/C^{\theta}$, $m(\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2/C^{1-\theta}$ and $m((\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2)/C^{1-\theta}$ all $\overset{as(\delta^{p},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$ 0, condition (IVb) is met, so

$$\text{"}d\text{"}:m(\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g d_g^{p})-m(\boldsymbol{x}_{gj}'\hat{\boldsymbol{\varepsilon}}_g\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)\underbrace{m(\delta_g^{p})}_{=1\text{ (L1p)}}\overset{p(\boldsymbol{d})|as(\delta^{p},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}0. \tag{C20}$$

For "e", we apply Theorem IV with $z_g=\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g/C^{(1/2)(1-\theta)}$ and, as by (L1p), (C13), and (C14) $[m((\delta_g^{p})^2)-m(\delta_g^{p})^2]/C^{\theta}$, $m(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g/C^{(1/2)(1-\theta)})^2/C^{1-\theta}$ and $m((\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2/C^{1-\theta})/C^{1-\theta}$ all $\overset{as(\delta^{p},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$ 0, condition (IVb) is met, so

$$\text{"}e\text{"}:m(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g d_g^{p})-m(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)\underbrace{m(\delta_g^{p})}_{=1\text{ (L1p)}}\overset{p(\boldsymbol{d})|as(\delta^{p},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}0. \tag{C21}$$

For "f ", we apply Theorem IV with $z_g=\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gs}/C^{1-\theta}$ and see that condition (IVb) holds as by (L1p) and (C16) $[m((\delta_g^{p})^2)-m(\delta_g^{p})^2]/C^{\theta}$ and $m(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gr}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gs}/C^{1-\theta})^2/C^{1-\theta}$ $\overset{as(\delta^{p},\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow}$ 0, while by the Cauchy–Schwarz Inequality and (L7),

$$\frac{m((\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\boldsymbol{x}_{gs})^2/C^{2(1-\theta)})}{C^{1-\theta}} \leq \sqrt{\sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr})^4}{N^{4-3\theta}}\sum_{g=1}^{C}\frac{(\boldsymbol{x}'_{gk}\boldsymbol{x}_{gs})^4}{N^{4-3\theta}}} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0, \tag{C22}$$

so

$$\text{“}f\text{”} : m(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\boldsymbol{x}_{gs}d_g^p) - m(\boldsymbol{x}'_{gj}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\boldsymbol{x}_{gs})\underbrace{m(\delta_g^p)}_{=1\text{ (L1p)}} \overset{p(\boldsymbol{d})|as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} 0. \tag{C23}$$

Similar to the case of the wild bootstrap, the $\hat{\eta}_r$ in (C19), which from (L4) and (C5) almost surely (across δ^p,$\boldsymbol{X}$,$\boldsymbol{\varepsilon}$) converge in distribution across permutations $\boldsymbol{d}^p$ of δ^p to normal variables with bounded variance, are multiplied by $\sqrt{\delta^{p\prime}\boldsymbol{O}\delta^p/C^{1+\theta}}$, which from (L1p) converges almost surely (across δ^p) to 0 and “e” and “f” terms, which almost surely (across δ^p,$\boldsymbol{X}$,$\boldsymbol{\varepsilon}$) converge in probability across permutations $\boldsymbol{d}^p$ to zero, and hence, when so multiplied, converge in probability across permutations $\boldsymbol{d}^p$ to zero. This leaves only the “d” term, and so, using (C3) earlier,

$$\boldsymbol{P} - \sum_{g=1}^{C}\frac{\boldsymbol{X}'_g\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}'_g\boldsymbol{X}_g}{C} \overset{p(\boldsymbol{d})|as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{0}_{KxK} \;\&\; \boldsymbol{A}^{-1} - \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{C}\right)^{-1} \overset{p(\boldsymbol{d})|as(\delta^p,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{0}_{KxK},$$
$$\text{and hence } C\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_C^p\right) - C\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}_C\right) \overset{p(\boldsymbol{d})|as(\delta^b,\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \boldsymbol{0}_{KxK}, \tag{C24}$$

which establishes (Vb) for the pairs bootstrap.

Appendix D. Proof of Lemma 1 in Appendix B

(L1), (L2), (L3): We prove these for the wild bootstrap, placing the more involved proofs for the pairs in the on-line appendix. From the assumptions $E\left[\delta_g^w\right] = 0$ & $E\left[(\delta_g^w)^2\right] = 1$ (Theorem V) and the Strong Law of Large Numbers, we know that $m(\delta_g^w) \overset{as(\delta^w)}{\rightarrow} 0$ and $m((\delta_g^w)^2) \overset{as(\delta^w)}{\rightarrow} 0$. Markov's Inequality, $E\left[(\delta_g^w)^{2(1+\theta_1)}\right] < \Delta$ (Theorem V), and $\theta > 1/(1+\theta_1)$ (Lemma 1) imply that there exists a v in $(1/(1+\theta_1),\theta)$ such that

$$\sum_{C=1}^{\infty}\text{Prob}((\delta_C^w)^2 \geq C^v) \leq \sum_{C=1}^{\infty}\frac{E((\delta_C^w)^{2(1+\theta_1)})}{C^{v(1+\theta_1)}} < \sum_{C=1}^{\infty}\frac{\Delta}{C^{v(1+\theta_1)}} < \infty, \tag{D1}$$

and thus, by the Borel–Cantelli Corollary given above, $max_{g\leq C}(\delta_g^w)^2/C^\theta \overset{as(\delta^w)}{\rightarrow} 0$, and so,

$$\frac{m((\delta_g^w)^4)}{C^\theta} = \sum_{g=1}^{C}\frac{(\delta_g^w)^4}{C^{1+\theta}} \leq \underset{g\leq C}{max}\frac{(\delta_g^w)^2}{C^\theta}m((\delta_g^w)^2) \overset{as(\delta^w)}{\rightarrow} 0. \tag{D2}$$

This establishes (L1w). As $\delta'^w\boldsymbol{O}\delta^w/C \overset{as(\delta^w)}{\rightarrow} 1$, for all C that are sufficiently large, $\delta'^w\boldsymbol{O}\delta^w/C$ is almost surely greater than some $\kappa > 0$, as stated in (L2). Regarding (L3), as for $\tau > 2$,

$$\left|\sum_{g=1}^{C}[\delta_g^w - m(\delta_g^w)]^\tau\right| \leq \sum_{g=1}^{C}\left|\delta_g^w - m\left(\delta_g^w\right)\right|^\tau$$
$$\leq \left(\underset{g\leq C}{max}[\delta_g^w - m(\delta_g^w)]^2\right)^{\frac{\tau}{2}-1}\sum_{g=1}^{C}[\delta_g^w - m(\delta_g^w)]^2, \tag{D3}$$

we have

$$0 < \frac{C^{(1-\theta)(\frac{\tau}{2}-1)}\left|\sum_{g=1}^{C}[\delta_g^w - m(\delta_g^w)]^\tau\right|}{\left(\sum_{g=1}^{C}[\delta_g^w - m(\delta_g^w)]^2\right)^{\tau/2}} \leq \left(\frac{\max\limits_{g\leq C}\frac{[\delta_g^w - m(\delta_g^w)]^2}{C^\theta}}{\sum_{g=1}^{C}\frac{[\delta_g^w - m(\delta_g^w)]^2}{C}}\right)^{\frac{\tau}{2}-1}. \tag{D4}$$

From the above, we know the denominator of the last almost surely converges to 1, while as for the numerator, using (L1w) and the result from (D1) $max_{g\leq C}(\delta_g^w)^2/C^\theta \overset{as(\delta^w)}{\rightarrow} 0$:

$$\max_{g\leq C}\frac{[\delta_g^w - m(\delta_g^w)]^2}{C^\theta} \leq \max_{g\leq C}\frac{(\delta_g^w)^2}{C^\theta} + 2\left|\frac{m(\delta_g^w)}{C^{(1/2)\theta}}\right|\max_{g\leq C}\left(\frac{(\delta_g^w)^2}{C^\theta}\right)^{1/2} + \frac{m(\delta_g^w)^2}{C^\theta} \overset{as(\delta^w)}{\rightarrow} 0. \tag{D5}$$

Consequently, (D4) almost surely converges to 0 for $\theta > 1/(1+\theta_1)$, proving (L3).

(L4): In the proof of Theorem I in Appendix A, we saw that $\boldsymbol{X}'\boldsymbol{X}/C - \boldsymbol{M}_C \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{KxK}$ and $\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g/C - \boldsymbol{V}_C \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{KxK}$, where the determinants of $\boldsymbol{M}_C$ and $\boldsymbol{V}_C$ are $> \eta > 0$ for all sufficiently large C and the absolute values of their elements are uniformly bounded by $\Delta^{1/(1+\gamma)}$. By the Continuous Mapping Theorem Corollary given above, $(\boldsymbol{X}'\boldsymbol{X}/C)^{-1} - \boldsymbol{M}_C^{-1} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{KxK}$ and $(\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g/C)^{-1} - \boldsymbol{V}_C^{-1} \overset{as(\boldsymbol{X},\boldsymbol{\varepsilon})}{\rightarrow} \mathbf{0}_{KxK}$, where for all C sufficiently large the determinants of $\boldsymbol{M}_C^{-1}$ and $\boldsymbol{V}_C^{-1}$ are greater than $(K\Delta^{1/(1+\gamma)})^{-K} > 0$ and the absolute values of their elements bounded by $(K\Delta^{1/(1+\gamma)})^{K-1}/\eta$, as proven earlier. It follows that almost surely for all C sufficiently large, $\boldsymbol{X}'\boldsymbol{X}/C$, $\sum_{g=1}^{C}\boldsymbol{X}_g'\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g'\boldsymbol{X}_g/C$ and their inverses have the same properties.

(L5), (L6), and (L7): Following the same logic used in (D3) and (D4) and using the Cauchy–Schwarz Inequality, we note that:

$$\frac{C^{\theta(\frac{\tau}{2}-1)}\left|\sum_{g=1}^{C}(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^\tau\right|}{\left(\sum_{g=1}^{C}(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2\right)^{\tau/2}} \leq \frac{C^{\theta(\frac{\tau}{2}-1)}\sum_{g=1}^{C}\left|(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^\tau\right|}{\left(\sum_{g=1}^{C}(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2\right)^{\tau/2}} \leq \left(\frac{\max\limits_{g\leq C}(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2/C^{1-\theta}}{\sum_{g=1}^{C}(\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2/C}\right)^{\frac{\tau}{2}-1} \tag{D6a}$$

$$\sum_{g=1}^{C}\frac{(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk})^2}{C^{2-\theta}} \leq \sum_{g=1}^{C}\frac{\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gk}}{C^{2-\theta}} \leq \frac{\max\limits_{g\leq C}\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj}}{C^{1-\theta}}m(\boldsymbol{x}_{gk}'\boldsymbol{x}_{gk}) \tag{D6b}$$

$$\sum_{g=1}^{C}\frac{(\boldsymbol{x}_{gj}'\boldsymbol{x}_{gk})^4}{C^{4-3\theta}} \leq \left(\frac{\max\limits_{g\leq C}\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj}}{C^{1-\theta}}\right)^2\frac{\max\limits_{g\leq C}\boldsymbol{x}_{gk}'\boldsymbol{x}_{gk}}{C^{1-\theta}}m\left(\boldsymbol{x}_{gk}'\boldsymbol{x}_{gk}\right). \tag{D6c}$$

So, to prove (L5)–(L7) it is sufficient to show that the right-hand sides of these inequalities converge to zero. In Appendix A, we already showed that almost surely $m(\boldsymbol{x}_{gk}'\boldsymbol{x}_{gk})$ is bounded and $\max_{g\leq C}\boldsymbol{x}_{gj}'\boldsymbol{x}_{gj}/C^{1-\theta}$ converges to 0, which establishes this for (D6b) and (D6c).

Turning to the right-hand side of (D6a), as shown in Appendix A $m((\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2)$ almost surely converges to the diagonal element of $\mathbf{V}_C$ in Theorem (Ic), whose smallest eigenvalue is greater than $\eta/(K\Delta^{1/(1+\gamma)})^{K-1}$ for all sufficiently large C. From the Schur–Horn Theorem, we know that the smallest diagonal element of $\mathbf{V}_C$ is greater than or equal to its smallest eigenvalue, and hence the term $m((\boldsymbol{x}_{gk}'\hat{\boldsymbol{\varepsilon}}_g)^2)$ in the denominator of (D6a) is almost surely greater than $\eta/(K\Delta^{1/(1+\gamma)})^{K-1}$> 0 for all C sufficiently large. Regarding the max term in the numerator, using $\hat{\boldsymbol{\varepsilon}}_g = \boldsymbol{\varepsilon}_g + \boldsymbol{X}_g(\boldsymbol{\beta}\text{-}\hat{\boldsymbol{\beta}}_C)$ and the Cauchy–Schwarz Inequality we have

$$\frac{(\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2}{C^{1-\theta}} = \frac{(\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g)^2}{C^{1-\theta}} + \sum_{r=1}^{K}\sum_{s=1}^{K}\frac{(\beta_r-\hat{\beta}_r)}{C^{(1/2)(\theta-1)}}\frac{(\beta_s-\hat{\beta}_s)}{C^{(1/2)(\theta-1)}}\frac{\boldsymbol{x}'_{gk}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\boldsymbol{x}_{gs}}{C^{2-2\theta}}$$

$$+2\sum_{r=1}^{K}\frac{(\beta_r-\hat{\beta}_r)}{C^{(1/2)(\theta-1)}}\frac{\boldsymbol{x}'_{gk}\boldsymbol{x}_{gr}\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g}{C^{(3/2)(1-\theta)}} \leq \frac{\left(\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g\right)^2}{C^{1-\theta}} + \tag{D7}$$

$$\sum_{r=1}^{K}\sum_{s=1}^{K}\left|\frac{\beta_r-\hat{\beta}_r}{C^{(1/2)(\theta-1)}}\right|\left|\frac{\beta_s-\hat{\beta}_s}{C^{(1/2)(\theta-1)}}\right|\prod_{i=j,k,p,q}\sqrt{\frac{\boldsymbol{x}'_{gi}\boldsymbol{x}_{gi}}{C^{1-\theta}}} + 2\sum_{r=1}^{K}\left|\frac{\beta_r-\hat{\beta}_r}{C^{(1/2)(\theta-1)}}\right|\sqrt{\frac{(\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g)^2}{C^{1-\theta}}}\prod_{i=k,r}\sqrt{\frac{\boldsymbol{x}'_{gi}\boldsymbol{x}_{gi}}{C^{1-\theta}}}.$$

It was shown in Appendix A that $\max_{g\leq C}\boldsymbol{x}'_{gj}\boldsymbol{x}_{gj}/C^{1-\theta} \overset{as(\mathbf{X},\boldsymbol{\varepsilon})}{\rightarrow} 0$ and $\sqrt{C}(\beta_r - \hat{\beta}_r)/C^{\delta}$ is asymptotically almost surely less than 1 for all $\delta > 0$, so that $(\beta_r - \hat{\beta}_r)/C^{(1/2)(\theta-1)} \overset{as(\mathbf{X},\boldsymbol{\varepsilon})}{\rightarrow} 0$. Consequently, to prove that $\max_{g\leq C}(\boldsymbol{x}'_{gk}\hat{\boldsymbol{\varepsilon}}_g)^2/C^{1-\theta}$ converges almost surely to zero, it is sufficient to show that $\max_{g\leq C}(\boldsymbol{x}'_{gk}\boldsymbol{\varepsilon}_g)^2/C^{1-\theta}$ converges almost surely to zero. However, $E(|\boldsymbol{x}'_{gj}\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}'_g\boldsymbol{x}_{gk}|^{1+\gamma}) < \Delta$ in Theorem I (Ic), by the same argument used in (A11) above, ensures that this is the case for $0 < \theta < \gamma/(1+\gamma)$. In sum, (D6a)–(D6c) converge to 0 for all θ in $(0,\gamma/(1+\gamma))$, proving (L5)–(L7). As $\theta_1 > 1/\gamma$ in Theorem V, the condition $\theta > 1/(1+\theta_1)$ for the wild bootstrap in Lemma 1 and the proof of (L3) above can also be met without contradiction.

Notes

[1] As examples: (i) Thornton (2008) used a randomized experiment to investigate the demand for and effects of learning HIV status across north, central, and south Malawi, which differ systematically in their ethnicity and religion. (ii) Cai et al. (2009) investigated saliency by randomly assigning restaurant arrivals in China to tables with different menu setups; not surprisingly, the total bill paid varies systematically with the time of day.

[2] With independent homoskedastic errors, the bootstrap resampling of estimated residuals (rather than the data itself) always yields consistent estimates of the coefficient distribution for a fixed number of OLS regressors (Bickel & Freedman, 1983).

[3] Canay et al. (2021), who examine wild bootstrap consistency when the number of independent cluster groupings is fixed, similarly allow for heterogeneity across clusters while assuming convergence of the full sample cross-product and covariance matrices to matrices of constants and, additionally, convergence of the projection of regressors on each other within each cluster to a common matrix.

[4] With $\boldsymbol{E}$ equal to the matrix of eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix of eigenvalues of $\boldsymbol{A}$, $\boldsymbol{A}^{1/2} = \boldsymbol{E}\boldsymbol{\Lambda}^{1/2}\boldsymbol{E}'$, where $\boldsymbol{\Lambda}^{1/2}$ is the diagonal matrix with entries equal to the square root of those of $\boldsymbol{\Lambda}$.

[5] The on-line Appendix proves consistency for sub-sampling, with and without replacement, $M < C$ groupings.

[6] Although, as noted by Cavaliere and Georgiev (2020), even when conditional consistency does not hold, valid inference using the bootstrap is still possible if the unconditional limit distribution of the sample test statistic equals the average of the random limit distribution of the bootstrap given the data.

[7] Where $\mathbf{1}c$ denotes a $Cx1$ vector of ones and $\boldsymbol{I}_{CxC}$ the CxC identity matrix.

[8] For the wild bootstrap, (7) follows immediately from the assumptions on moments. The proof for the pairs bootstrap is lengthy and is given in the on-line Appendix.

[9] Liu (1988) also advocated selecting $E((\delta_g{}^w)^3) = 1$ to correct for skewness in the Edgeworth expansion. However, Monte Carlos find that forms of $\delta_g{}^w$ that make this assumption perform less well than those that do not (Davidson & Flachaire, 2008; MacKinnon, 2015)

[10] To illustrate, the average skewness (i.e., standardized third central moment) of the first 1,000,000 Beta data-generating processes in 1000 runs of (12) ranges from −0.41 to 0.43, with a standard deviation of 0.14. A single run of 2 billion observations also shows no sign of converging, as the average skewness of the first 200, 400, 600, . . . , 2000 million Beta data-generating processes is .018, .021, .018, .014, .013, .024, .028, .031, .029, and .026, respectively.

[11] Hope (1968) noted that with k an integer and M draws from a continuous bootstrap distribution, an exact test relative to that distribution at level $\alpha = k/(M + 1)$ is achieved when the null is rejected if $k-1$ or fewer draws are greater than the sample test statistic. Jockel (1986) showed the same is true for draws from an arbitrary distribution, if $(G + (T + 1)\,U)$ is less than $\alpha(M + 1)$. As ties in these samples are exceedingly rare and αx100 is an integer, in the Table U plays no role and the test rejects when $G = 0$, ≤ 4 or ≤ 9 at the 0.01, 0.05 and 0.1 levels, respectively.

[12] This should be apparent for the bootstrap-t in (13), while in the case of the bootstrap-c, evaluating $\left(\hat{\beta}-\beta\right)^2$ using $\left(\hat{\beta}^b-\hat{\beta}\right)^2$ is identical to evaluating $\left(\hat{\beta}-\beta\right)^2/V(\hat{\beta})$ using $\left(\hat{\beta}^b-\hat{\beta}\right)^2/V(\hat{\beta})$.

[13] These instances are for the bootstrap-c, which is not based upon an asymptotically pivotal test statistic and hence does not provide the higher-order asymptotic accuracy of the bootstrap-t (Singh, 1981; Hall, 1992).

[14] The performance of the wild bootstrap is considerably improved if one imposes the null in the estimation of the residuals in the initial OLS regression (Davidson & Flachaire, 2008; Djogbenou et al., 2019). However, the same can be said for the conventional test, which raises the question of the correct benchmark comparison.

[15] The estimates in Table 1 incorporate Stata's default HC1 correction of the conventional covariance estimate, which reduces the rejection rate in the smallest samples. Without this correction, rejection rates at the 0.01 level are 0.134 and 0.107 with 10 observations or clusters, respectively.

[16] Let $\boldsymbol{E}$ denote the eigenvectors, $\boldsymbol{\Lambda}$ the diagonal matrix of eigenvalues, $\lambda_{\max}$ the maximum eigenvalue of the symmetric positive definite matrix $\boldsymbol{A}$, aij the ijth element, and $\boldsymbol{\alpha} i$ a vector of 0s with a 1 in the ith position. By the Cauchy–Schwarz Inequality and properties of the Rayleigh quotient, $a_{ij}^2 = (\boldsymbol{\alpha}_i'\boldsymbol{E\Lambda E}'\boldsymbol{\alpha}_j)^2 \leq (\boldsymbol{\alpha}_i'\boldsymbol{E\Lambda E}'\boldsymbol{\alpha}_i)(\boldsymbol{\alpha}_j'\boldsymbol{E\Lambda E}'\boldsymbol{\alpha}_j) \leq \lambda_{max}^2$.

References

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*, 169–181.

Bickel, P. J., & Freedman, D. A. (1983). Bootstrapping regression models with many parameters. In P. J. Bickel, K. A. Doksum, & J. L. Hodges (Eds.), *A festschrift for Erich L. Lehmann in honor of his sixty-fifth birthday*. Wadsworth.

Cai, H., Chen, Y., & Fang, H. (2009). Observational learning: Evidence from a randomized natural field experiment. *American Economic Review*, *99*, 864–882. [CrossRef]

Canay, I. A., Santos, A., & Shaikh, A. M. (2021). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics*, *103*, 346–363. [CrossRef]

Cavaliere, G., & Georgiev, I. (2020). Inference under random limit bootstrap measures. *Econometrica*, *88*, 2547. [CrossRef]

Davidson, R., & Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, *146*, 162–169. [CrossRef]

Djogbenou, A. A., MacKinnon, J. G., & Nielsen, M. O. (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics*, *212*, 393–412. [CrossRef]

Fedotenkov, I. (2013). A bootstrap method to test for the existence of finite moments. *Journal of Nonparametric Statistics*, *25*, 315–322. [CrossRef]

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, *9*, 1218–1228. [CrossRef]

Galambos, J. (1987). *The asymptotic theory of extreme order statistics* (2nd ed.). Robert E. Krieger Publishing Co.

Ghosh, M. N. (1950). Convergence of random distribution functions. *Bulletin of the Calcutta Mathematical Society*, *42*, 217–226.

Hall, P. (1992). *The bootstrap and edgeworth expansion*. Springer.

Hardle, W., Horowitz, J., & Kreiss, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review*, *71*, 435–459. [CrossRef]

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics*, *42*, 1977–1991. [CrossRef]

Hoeffding, W. (1951). A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, *22*, 558–566. [CrossRef]

Hoeffding, W. (1952). The large sample power of tests based upon permutations of observations. *The Annals of Mathematical Statistics*, *23*, 169–192. [CrossRef]

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B (Methodological)*, *30*, 582–598. [CrossRef]

Jockel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *The Annals of Statistics*, *14*, 336–347. [CrossRef]

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, *16*, 1696–1708. [CrossRef]

MacKinnon, J. G. (2015). Wild bootstrap cluster confidence intervals. *L'Actualité Économique*, *9*, 11–33. [CrossRef]

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 305–325. [CrossRef]

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, *21*, 255–285. [CrossRef]

Meerschaert, M. M., & Scheffler, H.-P. (1998). A simple robust estimation method for the thickness of heavy tails. *Journal of Statistical Planning and Inference*, *71*, 19–34. [CrossRef]

Noether, G. E. (1949). On a theorem by Wald and Wolfowitz. *The Annals of Mathematical Statistics*, *20*, 455–458. [CrossRef]

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Business and Economic Statistics*, *36*, 672–683. [CrossRef]

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. Springer.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, *9*, 1187–1195. [CrossRef]

Stute, W. (1990). Bootstrap of the linear correlation model. *Statistics: A Journal of Theoretical and Applied Statistics*, *21*, 433–436.

Thornton, R. L. (2008). The demand for, and impact of, learning HIV status. *American Economic Review*, *98*, 1829–1863. [CrossRef]

Trapani, L. (2016). Testing for (in)finite moments. *Journal of Econometrics*, *191*, 57–68. [CrossRef]

Wald, A., & Wolfowitz, J. (1944). Statistical tests based on permutations of the observations. *The Annals of Mathematical Statistics*, *15*, 358–372. [CrossRef]

White, H. (1980a). A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838. [CrossRef]

White, H. (1980b). Nonlinear regression on cross-section data. *Econometrica*, *48*, 721–746. [CrossRef]

White, H. (1984). *Asymptotic theory for econometricians*. Academic Press.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*, 1261–1295. [CrossRef]

Young, A. (2016). *Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections*. Working paper, January 2016. Available online: https://personal.lse.ac.uk/YoungA/ (accessed on 26 September 2025).

Young, A. (2019). Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, *134*, 557–598. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.