INFERENCE FOR HIGH-DIMENSIONAL EXCHANGEABLE ARRAYS

HAROLD D. CHIANG, KENGO KATO, AND YUYA SASAKI

Abstract. We consider inference for high-dimensional exchangeable arrays where the dimension may be much larger than the cluster sizes. Specifically, we consider separately and jointly exchangeable arrays that correspond to multiway clustered and polyadic data, respectively. Such exchangeable arrays have seen a surge of applications in empirical economics. However, both exchangeability concepts induce highly complicated dependence structures, which poses a significant challenge for inference in high dimensions. In this paper, we first derive high-dimensional central limit theorems (CLTs) over the rectangles for the exchangeable arrays. Building on the high-dimensional CLTs, we develop novel multiplier bootstraps for the exchangeable arrays and derive their finite sample error bounds in high dimensions. The derivations of these theoretical results rely on new technical tools such as Hoeffding-type decomposition and maximal inequalities for the degenerate components in the Hoeffiding-type decomposition for the exchangeable arrays. We illustrate applications of our bootstrap methods to robust inference in demand analysis, robust inference in extended gravity analysis, uniform confidence bands for density estimation with network data, and penalty choice for ℓ_1-penalized regression under multiway cluster sampling.

1. Introduction

In empirical studies in economics, we often employ data of volumes and attributes of flows of resources and commodities that are affected by supply shocks from the origin of the flow and demand shocks from the destination of the flow. Although supply and demand shocks are essential in economic analysis, a proper treatment of data generated by these shocks requires non-standard econometric methods due to the two-dimensional clustered dependence induced by these shocks.

When the set of agents generating the supply and the set of agents generating the demand are different, the data is *two-way clustered*. Leading examples are market share data that is two-way clustered by products and markets, where shares of a product are dependent across markets due to a common supply shock by the identical producer and shares of multiple products within a market are dependent due to a common demand shock by consumers in the identical market.

When the set of agents generating the supply and the set of agents generating the demand are the same, the data is *dyadic*. Leading examples are international trade data, where volumes of exports from an exporter are dependent across importers due to a common supply shock and volumes of imports to an importer are dependent across exporters due to a common demand shock.

Both of these types of data naturally entail complex dependence structures through common supply shocks by agents from an identical origin on agents across multiple destinations and common demand shocks by agents from an identical destination on agents across multiple origins. As

Date: First arXiv version: September 10, 2020. This version: January 1, 2021.

Key words and phrases. Bootstrap, exchangeable array, multiway clustering, network data, polyadic data.

H.D. Chiang is supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. K. Kato is partially supported by NSF grants DMS-1952306 and DMS-2014636.

such, standard microeconometric methods that presume cross-sectional random sampling are not applicable to either of these two types of data.

Starting with the seminal papers by Fafchamps and Gubert (2007) for dyadic data and Cameron et al. (2011) for multiway clustering, the recent econometrics literature develops methods and theories of how to deal with these types of dependent data – see below for a more comprehensive literature review. The existing literature, however, does not cover a method of high-dimensional inference, even though a number of robust identification strategies for structural economic models entail high-dimensionality in inference – see the next paragraph for examples. In this light, we develop a method of high-dimensional inference under general multiway clustering and polyadic sampling in this paper. For two-way clustered data $\{(X_{ij}^1,\dots,X_{ij}^p)^T : 1 \le i \le N_1, 1 \le j \le N_2\}$ of random vectors with high dimensions $p \gg \min\{N_1,N_2\}$, we develop a method and theory for bootstrap approximation of the distribution of the sample mean $N_1^{-1}N_2^{-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(X_{ij}^1,\dots,X_{ij}^p)^T$. Similarly, for dyadic data $\{(X_{ij}^1,\dots,X_{ij}^p)^T : 1 \le i,j \le n, i \ne j\}$ of random vectors with high dimension $p \gg n$, we develop a method and theory for bootstrap approximation of the distribution of the sample mean $n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{j\ne i}(X_{ij}^1,\dots,X_{ij}^p)^T$. We also generalize our results for these cases of two-way clustering and dyadic data to the cases of general multiway clustering and polyadic data, respectively.

Our proposed method applies to a number of important robust identification approaches for structural economic models. For demand analysis with a two-way clustered data consisting of N_1 products and N_2 markets, Gandhi et al. (2020) derive many moment inequalities of the of form

$$N_1^{-1}N_2^{-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(X_{ij}^1(\theta),\dots,X_{ij}^p(\theta))^T \ge 0, \tag{1.1}$$

where $(X_{ij}^1(\theta),\dots,X_{ij}^p(\theta))^T$ denotes a p-dimensional vector-valued random function of structural parameters θ. While most existing studies on demand analyses do not account for statistical dependence within a product i or within a market j, robust inference can be achieved by accounting for the two-way dependence – see Chiang et al. (2019). Similarly, for extended gravity analysis with a two-way clustered data consisting of N_1 firms and N_2 countries, Morales et al. (2019) derive many moment inequalities of the form (1.1). With our theory of approximating the distribution of the sample mean $N_1^{-1}N_2^{-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(X_{ij}^1(\theta),\ \dots,\ X_{ij}^p(\theta))^T$, inverting the Kolmogorov-Smirnov test allows for inference about the structural parameters θ similarly to Chernozhukov et al. (2019a). Again, while existing studies do not account statistical dependence within an exporter i or within an importer j, robust inference can be achieved by accounting for the two-way dependence. See Sections 4.1 and 4.2 ahead for details of these two applications to demand analysis and extended gravity analysis, respectively. As another useful application, our proposed technology allows for drawing uniform confidence bands for "the densities of migration across states, trade across nations, liabilities across banks, or minutes of telephone conversation among individuals" (Graham et al., 2019, 2020). To our knowledge, this paper is the first to provide a valid method for construction of simultaneous confidence intervals for kernel type estimator under dyadic clustering. In practice, dyadic data often has a point mass at zero. Our proposed method also allows for such a mixture distribution. See Section 4.3 for the application of dyadic kernel density estimation. Finally, our proposed technology also allows for selection of theoretically valid choice of a penalty for implementing ℓ_1-regularized regression (Lasso) under multiway clustering. To our knowledge, there is no existing theoretically justified method for Lasso penalty selection under an exchangeable sampling setting. See Section 4.4 for the application of Lasso penalty selection.

The two sampling frameworks of interest in this paper, namely multiway clustering and polyadic sampling, can be formulated as exchangeable random arrays. Specifically, a natural stochastic framework for modeling of mutliway clustering is that of *separately exchangeable* arrays (MacKinnon et al., 2020). For network/dyadic data, on the other hand, Bickel and Chen (2009) propose the use of *jointly exchangeable* arrays, which has since become a popular model for such data structures, see Graham (2019) and Graham and de Paula (2019) for recent reviews as well as the issue edited by Abbring and de Paula (2017). While formal definitions of these exchangeability concepts are postponed until Sections 2 and 3, it is worth noting that the exchangeable structures arise naturally in many economic applications. For example, in the context of modeling dynamic oligopoly with investment, Athey and Schmutzler (2001) indicate that the assumption of firms' profit functions being exchangeable is consistent with models of Cournot oligopoly, vertical product differentiation, and differentiated product models where the firms have identical cross-price effects. In these contexts, exchangeability imposes the symmetry in the identities of firms such that each firm cares only about the actions and state variables of its rivals, but not about the match between a competitor's identity and actions/state variables. They also point out the close link between exchangeability and the notation of anonymity in cooperative game theory and social choice theory (e.g. Moulin, 1988). Another such example is from the analysis of supply and demand in differentiated products markets. Berry et al. (1995) point out that both the demand and the cost functions for a product are exchangeable in vectors of characteristics of all other products. This emerges when the cost functions depend only on own-product characteristics, and is true for differentiated products demand system in which the demand for a product is independent of the ordering of competitors' products but only on their characteristics. They also observe that a unique Nash equilibrium implies several forms of exchangeability in the observed and unobserved random variables in their demand model (Berry et al., 1995, Section 5.1). Furthermore, Menzel (2016) observes that exchangeability of a certain form is a standard feature in almost all commonly used empirical specification for game-theoretic models with more than two players.

1.1. **Relation to the Literature.** High-dimensional central limit theorems (CLTs) and bootstraps over rectangles with the "$p \gg n$" regime are studied by Chernozhukov et al. (2013a, 2014, 2015, 2016, 2017a), Deng and Zhang (2020), Chernozhukov et al. (2019b), Kuchibhotla et al. (2020), and Fang and Koike (2020) for the independent case, by Chen (2018), Chen and Kato (2020, 2019) for U-statistics and processes, and by Zhang and Wu (2017), Zhang and Cheng (2018), Chernozhukov et al. (2019a), Koike (2019) for time series dependence. To the best of our knowledge, there is no result that considers extensions to exchangeable arrays in this literature. This paper builds on and complements those references by providing high-dimensional CLTs and bootstrap methods for exchangeable arrays.

Regression models with common shocks has been investigated by Andrews (2005) under exchangeability with one-dimensional index. Standard errors under multiway clustering (or separately exchangeable arrays) are proposed by Cameron et al. (2011) for parametric models, such as linear and nonlinear regression models – also see Cameron and Miller (2015, Section V) for a survey. Uniform asymptotic theory under multiway clustering is studied by Menzel (2017), covering both degenerate and non-degenerate cases. Focusing on the non-degenerate cases, Davezies et al. (2018, 2020) develop functional limit theorems for Donsker classes under multiway clustering. See also Chiang and Sasaki (2019), Chiang et al. (2019), MacKinnon (2019), and MacKinnon et al. (2020) for some other extensions and applications. To our best knowledge, no existing theory in this literature permits increasing or high-dimensional inference.

Theory of finite dimensional asymptotics (with fixed dimensions) for polyadic data (or jointly exchangeable arrays) is well-studied, see, e.g., Silverman (1976) and Eagleson and Weber (1978). Standard errors under dyadic data are first proposed by Fafchamps and Gubert (2007) and further studied by Cameron and Miller (2014), Aronow et al. (2015), and Tabord-Meehan (2019). Davezies et al. (2020) develop functional limit theorems for Donsker classes under polyadic sampling. To the best of our knowledge, no existing theory in this literature permits increasing or high-dimensional inference.

Methodologically, this paper is also related to the recent literature on high-dimensional U-statistics, such as Chen (2018), Chen and Kato (2020, 2019), among others. Under suitable assumptions, the data of our interest can be written as U-statistic-like latent structure (in distribution) via the Aldous-Hoover-Kallenberg representation (Aldous, 1981; Hoover, 1979; Kallenberg, 2006), i.e. the data can be written as a kernel function of some latent independent random variables. However, unlike the case with U-statistics, neither the kernel nor the latent independent random variables is known to us. In addition, we need to cope with the existence of extra idiosyncratic shocks in the latent structure. Both of these aspects present extra challenges.

The identification-robust inference applications considered in this paper are also related to the extensive literature of testing conditional moment inequalities, which includes, but are not limited to, Andrews and Shi (2013), Chernozhukov et al. (2013b), Lee et al. (2013), Armstrong (2014), Armstrong and Chan (2016), Andrews and Shi (2017), Chetverikov (2018), Lee et al. (2018), Bai et al. (2019) and Chernozhukov et al. (2019a). To the best of our knowledge, no theory that permits multiway clustered or polyadic data has been developed in this literature.

Regarding our bootstraps, McCullagh (2000) shows that no resampling scheme for the raw data is consistent for variance of a sample mean under multiway clustering. A Pigeonhole bootstrap is subsequently proposed by Owen (2007) and its different variants are further investigated in Owen and Eckles (2012), Menzel (2017) and Davezies et al. (2018, 2020). Whether the pigeonhole bootstrap works for increasing or high-dimensional test statistics remains unknown to us. We therefore develop a novel bootstrap method in this paper which we argue works for high-dimensional data.

Finally, we develop novel Hoeffding-type decompositions for both separately and jointly exchangeable arrays and establish symmetrization inequalities for Hoeffding-type projection terms in both cases. This allows us to obtain several new maximal inequalities that lead to sharp rates for degenerate components in Hoeffding-type decompositions in both cases. Such symmetrization and maximal inequalities play a crucial role in establishing the high-dimensional CLTs as well as the validity of the bootstrap methods. These technical results are of independent interest and would be useful for other analyses of multiway clustering and polyadic data. The proofs of these technical results are highly nontrivial and indeed more involved than the U-statistic case due to the unknown (and, in jointly exchangeable case, index-dependent) nature of kernel functions and the presence of the extra unobserved shocks. For example, the proof of the symmetrization inequality for multiway clustering involves a careful induction argument (see Lemma 3 in the Appendix), combined with a repeated conditioning argument. Also, the proof of the maximal inequality for polyadic data involves a delicate conditioning argument, combined with the symmetrization inequalities for

U-statistics with index-dependent kernels (cf. de la Peña and Giné, 1999). In comparison, the empirical process results in Davezies et al. (2020) rely on substantially different symmetrization inequalities. Specifically, symmetrization inequalities developed in Davezies et al. (2020) are applied to the whole empirical process and do not lead to correct orders for degenerate components in Hoeffding-type decompositions (indeed, Davezies et al. (2020) do not derive Hoeffding-type decompositions), thereby not powerful enough to derive our results; see Remarks 10 and 11 in the Appendix for details.

In the present paper, we focus on the case where the sample mean is non-degenerate, where the approximating distribution is Gaussian. In the univariate case, Menzel (2017) develops inference methods robust to degenerate situations, where the limit distribution may have a Gaussian chaos component, similarly to U-statistics. In the high-dimensional case with $p \gg n$, existing techniques used in the Gaussian approximation, such as a Slepian-Stein method and the anti-concentration inequality (cf. Chernozhukov et al., 2013a, 2014, 2017a), can not be directly extended to non-Gaussian approximating distributions such as Gaussian chaos distributions. Indeed, there have been no results concerning high-dimensional non-Gaussian approximations (by high-dimension we mean $p \gg n$), including a simpler setting of degenerate U-statistics. Extensions of the results of the present paper to degenerate cases are left to future research. That said, non-degenerate sample means are natural in multivariate applications. This is because non-constant coordinates of multi-dimensional random vectors are often i-specific and j-specific as is the case in the aforementioned applications (Gandhi et al., 2020; Morales et al., 2019), and these i- and j-specific non-constant coordinates induce non-degeneracy. For this reason, we believe our focus on non-degenerate cases in fact will not significantly narrow the scope of applicability.

1.2. **Notations and Organization.** Let $\mathbb{N}$ denote the set of positive integers. We use $\|\cdot\|, \|\cdot\|_0, \|\cdot\|_1$, and $\|\cdot\|_\infty$ to denote the Euclidean, ℓ_0, ℓ_1, and ℓ^∞-norms for vectors, respectively (precisely, $\|\cdot\|_0$ is not a norm but a seminorm). For two real vectors $\boldsymbol{a} = (a_1, \dots, a_p)^T$ and $\boldsymbol{b} = (b_1, \dots, b_p)^T$, the notation $\boldsymbol{a} \le \boldsymbol{b}$ means that $a_j \le b_j$ for all $1 \le j \le p$. Let $\mathrm{supp}(\boldsymbol{a})$ denote the support of $\boldsymbol{a} = (a, \dots, a_p)^T$, i.e., $\mathrm{supp}(\boldsymbol{a}) = \{j : a_j \ne 0\}$. We denote by $\odot$ the Hadamard (element-wise) product, i.e., for $\boldsymbol{i} = (i_1, \dots, i_K)$ and $\boldsymbol{j} = (j_1, \dots, j_K)$, $\boldsymbol{i} \odot \boldsymbol{j} = (i_1 j_1, \dots, i_K j_K)$. For any $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$. For $0 < \beta < \infty$, let ψ_β be the function on $[0, \infty)$ defined by $\psi_\beta(x) = e^{x^\beta} - 1$. Let $\|\cdot\|_{\psi_\beta}$ denote the associated Orlicz norm, i.e., $\|\xi\|_{\psi_\beta} = \inf\{C > 0 : \mathbb{E}[\psi_\beta(|\xi|/C)] \le 1\}$ for a real-valued random variable ξ. For $\beta \in (0, 1)$, $\|\cdot\|_{\psi_\beta}$ is not a norm but a quasi-norm, i.e., there exists a constant C_β depending only on β such that $\|\xi_1 + \xi_2\|_{\psi_\beta} \le C_\beta(\|\xi_1\|_{\psi_\beta} + \|\xi_2\|_{\psi_\beta})$. Let $U[0, 1]$ denote the uniform distribution on $[0, 1]$. "Constants" refer to nonstochastic and finite positive numbers.

The rest of the paper is organized as follows . In Section 2, we develop a high-dimensionl CLT (over the rectangles) and a bootstrap method for multiway clustering (seperately exchangeable arrays). In Section 3, we develop analogous results to polyadic sampling. We illustrate four applications in Section 4, present simulation results in Section 5, and demonstrate an empirical application in Section 6. We defer all the technical proofs to the Appendix.

2. Multiway Clustering

In this section, we consider separately exchangeable arrays that correspond to multiway clustered data. Pick any $K \in \mathbb{N}$. With $\boldsymbol{i} = (i_1, \dots, i_K) \in \mathbb{N}^K$, we consider a K-array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^K}$ consisting of random vectors in $\mathbb{R}^p$. We denote by $X_{\boldsymbol{i}}^j$ the j-th coordinate of $\boldsymbol{X}_{\boldsymbol{i}}$: $\boldsymbol{X}_{\boldsymbol{i}} = (X_{\boldsymbol{i}}^1, \dots, X_{\boldsymbol{i}}^p)^T$. We say that the array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^K}$ is *separately exchangeable* if the following condition is satisfied (cf. Kallenberg, 2006, Section 3.1).

Definition 1 (Separate exchangeability)**.** *A K-array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^K}$ is called separately exchangeable if for any K permutations $\pi_1,\dots,\pi_K$ of $\mathbb{N}$, the arrays $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^K}$ and $(\boldsymbol{X}_{(\pi_1(i_1),\dots,\pi_K(i_K))})_{\boldsymbol{i}\in\mathbb{N}^K}$ are identically distributed in the sense that their finite dimensional distributions agree.*

From the Aldous-Hoover-Kallenberg representation (see Kallenberg, 2006, Corollary 7.23), any separately exchangeable array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^K}$ is generated by the structure

$$\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{f}((U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\boldsymbol{e}\in\{0,1\}^K}),\ \boldsymbol{i}\in\mathbb{N}^K,\quad \{U_{\boldsymbol{i}\odot\boldsymbol{e}} : \boldsymbol{i}\in\mathbb{N}^K, \boldsymbol{e}\in\{0,1\}^K\} \overset{i.i.d.}{\sim} U[0,1]$$

for some Borel measurable map $\mathfrak{f} : [0,1]^{2^K} \to \mathbb{R}^p$. For example, when $K = 2$, then $\boldsymbol{X}_{\boldsymbol{i}}$ is generated as $\boldsymbol{X}_{(i_1,i_2)} = \mathfrak{f}(U_{(0,0)}, U_{(i_1,0)}, U_{(0,i_2)}, U_{(i_1,i_2)})$.

The latent variable $U_{\boldsymbol{0}}$ appears commonly in all $\boldsymbol{X}_{\boldsymbol{i}}$'s. In the present paper, as in Andrews (2005) and Menzel (2017), we consider inference conditional on $U_{\boldsymbol{0}}$ and treat it as fixed. In the rest of Section 2, we will assume (without further mentioning) that the array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^K}$ has mean zero (conditional on $U_{\boldsymbol{0}}$) and is generated by the structure

$$\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{g}((U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\boldsymbol{e}\in\{0,1\}^K\setminus\{\boldsymbol{0}\}}),\ \boldsymbol{i}\in\mathbb{N}^K, \tag{2.1}$$

where $\mathfrak{g}$ is now a map from $[0,1]^{2^K-1}$ into $\mathbb{R}^p$.

Suppose that we observe $\{\boldsymbol{X}_{\boldsymbol{i}} : \boldsymbol{i}\in[\boldsymbol{N}]\}$ with $\boldsymbol{N} = (N_1,\dots,N_K)$ and $[\boldsymbol{N}] = \prod_{k=1}^K\{1,\dots,N_k\}$. We are interested in approximating the distribution of the sample mean

$$\boldsymbol{S}_{\boldsymbol{N}} = \frac{1}{\prod_{k=1}^K N_k}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}\boldsymbol{X}_{\boldsymbol{i}}$$

in the high-dimensional setting where the dimension p is allowed to entail $p \gg \min\{N_1,\dots,N_K\}$.

Example 1 (Empirical process indexed by function class with increasing cardinality)**.** Our setting covers the following situation: let $\{Y_{\boldsymbol{i}} : \boldsymbol{i}\in\mathbb{N}^K\}$ be random variables taking values in an abstract measurable space $(S,\mathcal{S})$, and suppose that they are generated as

$$Y_{\boldsymbol{i}} = \breve{\mathfrak{g}}((U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\boldsymbol{e}\in\{0,1\}^K\setminus\{\boldsymbol{0}\}}).$$

Let $f_j : S \to \mathbb{R}$ for $1 \le j \le p$ be measurable functions, and define $X_{\boldsymbol{i}}^j = f_j(Y_{\boldsymbol{i}}) - \mathbb{E}[f_j(Y_{\boldsymbol{i}})]$. In this case, the sample mean $\boldsymbol{S}_{\boldsymbol{N}}$ can be regarded as the empirical process $f \mapsto (\prod_{k=1}^K N_k)^{-1}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(f(Y_{\boldsymbol{i}}) - \mathbb{E}[f(Y_{\boldsymbol{i}})])$ indexed by the function class $\mathcal{F} = \{f_1,\dots,f_p\}$. Allowing $p\to\infty$ as $\min_{1\le k\le K} N_k \to \infty$ enables us to cover empirical processes indexed by function classes with increasing cardinality.

For later convenience, we fix some additional notations. Let $n = \min_{1\le k\le K} N_k$ and $\overline{N} = \max_{1\le k\le K} N_k$ denote the minimum and maximum cluster sizes, respectively. For $1 \le k \le K$, denote by $\mathcal{E}_k = \{\boldsymbol{e} = (e_1,\dots,e_K) \in \{0,1\}^K : \sum_{k=1}^K e_k = k\}$ the set of vectors in $\{0,1\}^K$ whose support has cardinality k. Let $\boldsymbol{e}_k \in \mathbb{R}^K$ denote the vector such that the k-th coordinate of $\boldsymbol{e}_k$ is 1 and the other coordinates are 0. For a given $\boldsymbol{e}\in\{0,1\}^K$, define

$$I_{\boldsymbol{e}}([\boldsymbol{N}]) = \{\boldsymbol{i}\odot\boldsymbol{e} : \boldsymbol{i}\in[\boldsymbol{N}]\} \subset \mathbb{N}_0^K \quad \text{with } \mathbb{N}_0 = \mathbb{N}\cup\{0\}.$$

The following decomposition of the sample mean $\boldsymbol{S}_{\boldsymbol{N}}$ will play a fundamental role in our analysis, which is reminiscent of the Hoeffding decomposition for U-statistics (Lee, 1990; de la Peña and Giné, 1999).

Lemma 1 (Hoeffding decomposition of separately exchangeable array)**.** *For any $\boldsymbol{i} \in \mathbb{N}^K$, define recursively*

$$\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}_k} = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{\boldsymbol{i}\odot\boldsymbol{e}_k}],\ k = 1,\dots,K,$$

$$\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}} = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{e}'\leq\boldsymbol{e}}] - \sum_{\substack{\boldsymbol{e}'\leq\boldsymbol{e}\\ \boldsymbol{e}'\neq\boldsymbol{e}}} \hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'}, \quad \boldsymbol{e} \in \bigcup_{k=2}^{K} \mathcal{E}_k.$$

Then, we have

$$\boldsymbol{X}_{\boldsymbol{i}} = \sum_{\boldsymbol{e}\in\{0,1\}^K\setminus\{\boldsymbol{0}\}} \hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}}.$$

Consequently, we can decompose the sample mean $\boldsymbol{S}_{\boldsymbol{N}} = (\prod_{k=1}^{K} N_k)^{-1} \sum_{\boldsymbol{i}\in[\boldsymbol{N}]} \boldsymbol{X}_{\boldsymbol{i}}$ as

$$\boldsymbol{S}_{\boldsymbol{N}} = \sum_{k=1}^{K} \sum_{\boldsymbol{e}\in\mathcal{E}_k} \frac{1}{\prod_{k'\in\mathrm{supp}(\boldsymbol{e})} N_{k'}} \sum_{\boldsymbol{i}\in I_{\boldsymbol{e}}([\boldsymbol{N}])} \hat{\boldsymbol{X}}_{\boldsymbol{i}}. \tag{2.2}$$

The proof of this lemma can be found in Appendix B.1.

Example 2 ($K = 3$ case)**.** For instance, if $K = 3$, then for $\boldsymbol{i} = (i_1, i_2, i_3) \in \mathbb{N}^3$,

$$\hat{\boldsymbol{X}}_{(i_1,0,0)} = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,0)}],\ \hat{\boldsymbol{X}}_{(0,i_2,0)} = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(0,i_2,0)}],\ \hat{\boldsymbol{X}}_{(0,0,i_3)} = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(0,0,i_3)}],$$

$$\hat{\boldsymbol{X}}_{(i_1,i_2,0)} = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,0)}, U_{(0,i_2,0)}, U_{(i_1,i_2,0)}] - \hat{\boldsymbol{X}}_{(i_1,0,0)} - \hat{\boldsymbol{X}}_{(0,i_2,0)},\ \text{etc.},$$

$$\hat{\boldsymbol{X}}_{(i_1,i_2,i_3)} = \boldsymbol{X}_{\boldsymbol{i}} - \hat{\boldsymbol{X}}_{(i_1,i_2,0)} - \hat{\boldsymbol{X}}_{(0,i_2,i_3)} - \hat{\boldsymbol{X}}_{(i_1,0,i_3)} - \hat{\boldsymbol{X}}_{(i_1,0,0)} - \hat{\boldsymbol{X}}_{(0,i_2,0)} - \hat{\boldsymbol{X}}_{(0,0,i_3)}.$$

Remark 1 (Hoeffding decomposition)**.** The reason that we call (2.2) the Hoeffding decomposition comes from the fact that if the dimension p is fixed, for each fixed $k = 1,\dots,K$ and $\boldsymbol{e} \in \mathcal{E}_k$, the component

$$\frac{1}{\prod_{k'\in\mathrm{supp}(\boldsymbol{e})} N_{k'}} \sum_{\boldsymbol{i}\in I_{\boldsymbol{e}}([\boldsymbol{N}])} \hat{\boldsymbol{X}}_{\boldsymbol{i}}$$

scales as $(\prod_{k'\in\mathrm{supp}(\boldsymbol{e})} N_{k'})^{-1/2} = O(n^{-k/2})$ with $n = \min_{1\leq k'\leq K} N_{k'}$ under moment conditions. See Corollary 3 in Appendix A. This is completely analogous to the Hoeffiding decomposition of U-statistics and from this analogy we shall call (2.2) the Hoeffding decomposition.

The leading term in the decomposition (2.2) is

$$\sum_{\boldsymbol{e}\in\mathcal{E}_1} \frac{1}{\prod_{k'\in\mathrm{supp}(\boldsymbol{e})} N_{k'}} \sum_{\boldsymbol{i}\in I_{\boldsymbol{e}}([\boldsymbol{N}])} \hat{\boldsymbol{X}}_{\boldsymbol{i}} = \sum_{k=1}^{K} N_k^{-1} \sum_{i_k=1}^{N_k} \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(0,\dots,0,i_k,0,\dots,0)}],$$

which we call the Hájek projection of $\boldsymbol{S}_{\boldsymbol{N}}$. Define

$$\boldsymbol{W}_{k,i_k} = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(0,\dots,0,i_k,0,\dots,0)}],\ k = 1,\dots,K,$$

$$\boldsymbol{S}_{\boldsymbol{N}}^{W} = \sum_{k=1}^{K} N_k^{-1} \sum_{i_k=1}^{N_k} \boldsymbol{W}_{k,i_k},\ \text{and}\ \ \Sigma_{W_k} = \mathbb{E}[\boldsymbol{W}_{k,1}\boldsymbol{W}_{k,1}^{T}],\ k = 1,\dots,K.$$

Since $\boldsymbol{S}_{\boldsymbol{N}}^{W}$ is the sum of independent random vectors, it is expected that the distribution of $\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}}$ can be approximated by $N(\boldsymbol{0}, \Sigma)$, where

$$\Sigma = \sum_{k=1}^{K} (n/N_k)\Sigma_{W_k},$$

as long as the remainder term is negligible. This suggests the following multiplier bootstrap for multiway clustering.

2.1. **Multiplier bootstrap for multiway clustering.** Let $\{\xi_{1,i_1}\}_{i_1=1}^{N_1},\dots,\{\xi_{K,i_K}\}_{i_K=1}^{N_K}$ be independent $N(0,1)$ random variables independent of the data. Ideally, we want to make use of the bootstrap statistic

$$\sum_{k=1}^{n} N_k^{-1}\sum_{i_k=1}^{N_k}\xi_{k,i_k}(\boldsymbol{W}_{k,i_k}-\boldsymbol{S}_{\boldsymbol{N}}).$$

However, this bootstrap is infeasible as $\boldsymbol{W}_{k,i_k}=\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}}\mid U_{(0,\dots,i_k,\dots,0)}]$ are unknown to us. Estimation of $\boldsymbol{W}_{k,i_k}$ is nontrivial as $U_{(0,\dots,i_k,\dots,0)}$ is a latent variable. To gain an insight into how to estimate $\boldsymbol{W}_{k,i_k}$, consider the case where $K=2$. Then $\boldsymbol{W}_{1,i_1}=\mathbb{E}[\boldsymbol{X}_{(i_1,i_2)}\mid U_{(i_1,0)}]=\mathbb{E}[\mathfrak{g}(U_{(i_1,0)},V_{(i_1,i_2)})\mid U_{(i_1,0)}]$ with $V_{(i_1,i_2)}=(U_{(0,i_2)},U_{(i_1,i_2)})$. Since $U_{(i_1,0)}$ and $V_{(i_1,i_2)}$ are independent and the latter variable is independent across i_2, we see that $\boldsymbol{W}_{1,i_1}$ can be estimated by taking the average of $\boldsymbol{X}_{(i_1,i_2)}$ over i_2.

Building on this intuition, in general, we propose to estimate each $\boldsymbol{W}_{k,i_k}$ by

$$\overline{\boldsymbol{X}}_{k,i_k}=\frac{1}{\prod_{k'\neq k}N_{k'}}\sum_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_K}\boldsymbol{X}_{\boldsymbol{i}},\ i_k=1,\dots,N_k;k=1,\dots,K,$$

i.e., the sample mean taken over all indices but i_k. Then, we apply the multiplier bootstrap to $\overline{\boldsymbol{X}}_{k,i_k}$ in place of $\boldsymbol{W}_{k,i_k}$

$$\boldsymbol{S}_{\boldsymbol{N}}^{MB}=\sum_{k=1}^{K}N_k^{-1}\sum_{i_k=1}^{N_k}\xi_{k,i_k}(\overline{\boldsymbol{X}}_{k,i_k}-\boldsymbol{S}_{\boldsymbol{N}}).$$

To the best of our knowledge, this multiplier bootstrap for multiway clustering is new in the literature. We will formally study the validity of this multiplier bootstrap for high-dimensional multiway clustered data with $p\gg n$ in the next two subsections.

2.2. **High-dimensional CLT for multiway clustering.** We first establish a high-dimensional CLT for $\boldsymbol{S}_{\boldsymbol{N}}$ over the class of rectangles,

$$\mathcal{R}=\left\{\prod_{j=1}^{p}[a_j,b_j]:-\infty\le a_j\le b_j\le\infty,\ 1\le j\le p\right\}.$$

This high-dimensional CLT will be a building block for establishing the validity of the multiplier bootstrap considered in the preceding section.

We start with discussing regularity conditions. Denote by $\mathbf{1}=(1,\dots,1)$ the vector of ones. Let $D_{\boldsymbol{N}}\ge 1$ be a given constant that may depend on the cluster sizes $\boldsymbol{N}$, and let $\underline{\sigma}>0$ be another given constant independent of the cluster sizes $\boldsymbol{N}$. We will assume *either* of the following moment conditions.

$$\max_{1\le j\le p}\|X_{\mathbf{1}}^{j}\|_{\psi_1}\le D_{\boldsymbol{N}},\quad\text{or}\tag{2.3}$$

$$\mathbb{E}[\|\boldsymbol{X}_{\mathbf{1}}\|_\infty^q]\le D_{\boldsymbol{N}}^q\quad\text{for some } q\in(4,\infty).\tag{2.4}$$

We will also assume *both* of the following conditions.

$$\max_{1\le j\le p;1\le k\le K}\mathbb{E}[|W^j_{k,1}|^{2+\kappa}]\le D^\kappa_{\boldsymbol{N}},\ \kappa=1,2, \tag{2.5}$$

$$\min_{1\le j\le p;1\le k\le K}\mathbb{E}[|W^j_{k,1}|^{2}]\ge \underline{\sigma}^2. \tag{2.6}$$

Condition (2.3) requires that each coordinate of $\boldsymbol{X_1}$ is sub-exponential. By Jensen's inequality, Condition (2.3) implies that

$$\max_{1\le j\le p;1\le k\le K}\|W^j_{k,1}\|_{\psi_1}\le D_{\boldsymbol{N}}.$$

Condition (2.4) is an alternative moment condition on $\boldsymbol{X_1}$. Condition (2.4) is satisfied for example under the following situation: Suppose that $\boldsymbol{X_i}$ is given by $\boldsymbol{X_i}=\varepsilon_{\boldsymbol{i}}\boldsymbol{Z_i}$ where $\varepsilon_{\boldsymbol{i}}$ is a scalar "error" variable while $\boldsymbol{Z}$ is a vector of "covariates". If each coordinate of $\boldsymbol{Z_i}$ is bounded by a constant $\overline{D}_{\boldsymbol{N}}$ (that may depend on $\boldsymbol{N}$) and $\varepsilon_{\boldsymbol{i}}$ has finite q-th moment, then $\mathbb{E}[\|\boldsymbol{X_i}\|^q_\infty]\le\overline{D}^q_{\boldsymbol{N}}\mathbb{E}[|\varepsilon_{\boldsymbol{i}}|^q]$. Again, by Jensen's inequality, Condition (2.4) implies that

$$\max_{1\le k\le K}\mathbb{E}[\|\boldsymbol{W}_{k,1}\|^q_\infty]\le D^q_{\boldsymbol{N}}.$$

Condition (2.5) requires the maximum of third (respectively, fourth) moment across coordinates to be increasing at speed no faster than the first (respectively, second) power of $D_{\boldsymbol{N}}$. By Jensen's inequality, Condition (2.5) is satisfied if $\max_{1\le j\le p}\mathbb{E}[|X^j_{\boldsymbol{1}}|^{2+\kappa}]\le D^\kappa_{\boldsymbol{N}}$ for $\kappa=1,2$. Condition (2.6) guarantees that the Hájek projection is nondegenerate.

Let $\gamma=N(\boldsymbol{0},\Sigma)$.

Theorem 1 (High-dimensional CLT for multiway clustering)**.** *Suppose that either Condition (2.3) or (2.4) holds, and further that both Conditions (2.5) and (2.6) hold. Then, there exists a constant C such that*

$$\sup_{R\in\mathcal{R}}|\mathbb{P}(\sqrt{n}\boldsymbol{S_N}\in R)-\gamma_\Sigma(R)|$$
$$\le\begin{cases} C\left(\frac{D^2_{\boldsymbol{N}}\log^7(p\overline{N})}{n}\right)^{1/6} & \textit{if Condition (2.3) holds,}\\ C\left[\left(\frac{D^2_{\boldsymbol{N}}\log^7(p\overline{N})}{n}\right)^{1/6}+\left(\frac{D^2_{\boldsymbol{N}}\log^3(p\overline{N})}{n^{1-2/q}}\right)^{1/3}\right] & \textit{if Condition (2.4) holds,}\end{cases}$$

where the constant C depends only on $\underline{\sigma}$ and K if Condition (2.3) holds, while C depends only on $q,\underline{\sigma}$, and K if Condition (2.4) holds.

Remark 2 (Refinement under subgaussianity)**.** The recent paper of Chernozhukov et al. (2019b) provides some improvements on convergence rate of Gaussian approximation under the subgaussian tail assumption for the sample mean of independent random vectors. With this new technique, if we strengthen Condition (2.3) by replacing the ψ_1-norm $\|\cdot\|_{\psi_1}$ with the ψ_2-norm $\|\cdot\|_{\psi_2}$ (i.e., each coordinate $\boldsymbol{X_1}$ is sub-Gaussian), the bound $C\left(n^{-1}D^2_{\boldsymbol{N}}\log^7(p\overline{N})\right)^{1/6}$ in Theorem 1 can be improved to $C\left(n^{-1}D^2_{\boldsymbol{N}}\log^5(p\overline{N})\right)^{1/4}$.

2.3. **Validity of multiplier bootstrap for multiway clustering.** We are now in position to establish the validity of the proposed multiplier bootstrap for multiway clustered data. Let $\mathbb{P}_{|\boldsymbol{X}_{[\boldsymbol{N}]}}$ denote the law conditional on the data $\boldsymbol{X}_{[\boldsymbol{N}]}=(\boldsymbol{X_i})_{\boldsymbol{i}\in[\boldsymbol{N}]}$. Define

$$\hat{\Delta}_W=\max_{1\le j\le p;1\le k\le K}\frac{1}{N_k}\sum_{i_k=1}^{N_k}(\overline{X}^j_{k,i_k}-W^j_{k,i_k})^2,$$

which accounts for the estimation error of $\overline{\boldsymbol{X}}_{k,i_k}$ for $\boldsymbol{W}_{k,i_k}$. Also, let $\overline{\sigma} = \max_{1\le j\le p; 1\le k\le K} \sqrt{\mathbb{E}[|W_{k,1}^j|^2]}$. The following theorem shows that as soon as $\hat{\Delta}_W$ is sufficiently small (i.e, $\overline{\sigma}^2\hat{\Delta}_W \log^4 p = o_P(1)$), then the multiplier bootstrap is consistent over the rectangles under mild conditions on the dimension p.

Theorem 2 (Validity of multiplier bootstrap for multiway clustering)**.** *Consider the following two cases.*

(i). Conditions (2.3), (2.5), and (2.6) hold, and there exist constants C_1 and $\zeta_1,\ \zeta_2 \in (0,1)$ such that

$$\mathbb{P}\left(\overline{\sigma}^2\hat{\Delta}_W \log^4 p > C_1 n^{-\zeta_2}\right) \le C_1 n^{-1} \quad \textit{and} \tag{2.7}$$

$$\frac{D_{\boldsymbol{N}}^2(\log^2 n)\log^5(p\overline{N})}{n} \le C_1 n^{-\zeta_1}. \tag{2.8}$$

(ii). Conditions (2.4), (2.5), and (2.6) hold, and there exist constants C_1 and $\zeta_1,\ \zeta_2 \in (0,1)$ such that Condition (2.7) holds and

$$\frac{D_{\boldsymbol{N}}^2\log^5(pn)}{n}\bigvee\left(\frac{D_{\boldsymbol{N}}^2\log^3 p}{n^{1-4/q}}\right)^2 \le C_1 n^{-\zeta_1}. \tag{2.9}$$

Then, under either Case (i) or (ii), there exists a constant C such that

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}_{|\boldsymbol{X}_{[\boldsymbol{N}]}}(\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}}^{MB}\in R) - \gamma_\Sigma(R)\right| \le Cn^{-(\zeta_1\wedge\zeta_2)/4}$$

with probability at least $1 - Cn^{-1}$, where the constant C depends only on $\underline{\sigma}, K$, and C_1 under Case (i), while C depends only on $q, \underline{\sigma}, K$, and C_1 under Case (ii).

Remark 3 (Discussion on Conditions (2.7)–(2.9))**.** Conditions (2.7)–(2.9) are placed to guarantee that the error bound for our multiplier bootstrap decreases at a polynomial rate in n. If we are to show a weaker result, namely,

$$\sup_{R\in\mathcal{R}}|\mathbb{P}_{|\boldsymbol{X}_{[\boldsymbol{N}]}}(\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}}^{MB}\in R) - \gamma_\Sigma(R)| = o_P(1) \tag{2.10}$$

as $n\to\infty$ (with the understanding that $p, \overline{\sigma}, D_{\boldsymbol{N}}$, and $\overline{N}$ are functions of n), then Conditions (2.7)–(2.9) can be weakened to $\overline{\sigma}\hat{\Delta}_W\log^4 p = o_P(1), D_{\boldsymbol{N}}^2\log^5(p\overline{N}) = o(n)$, and $(n^{-1}D_{\boldsymbol{N}}^2\log^5(pn))\vee(n^{1-2/q}D_{\boldsymbol{N}}^2\log^3 p) = o(1)$, respectively. (The critical case $q = 4$ is allowed for (2.10); note that the high-dimensional CLT (Theorem 1) also holds with $q = 4$.)

Condition (2.7) is a high-level condition on the estimation accuracy of $\overline{\boldsymbol{X}}_{k,i_k}$ for $\boldsymbol{W}_{k,i_k}$. We provide primitive sufficient conditions for Condition (2.7) to hold in the following proposition.

Proposition 1 (Primitive sufficient conditions for Condition (2.7))**.** *Consider the following two cases.*

(i') Conditions (2.3), (2.5), and (2.6) hold, and there exist constants C_1 and $\zeta\in(0,1)$ such that

$$\frac{\overline{\sigma}^2D_{\boldsymbol{N}}^2\log^7 p}{n} \le C_1 n^{-\zeta}. \tag{2.11}$$

(ii') Conditions (2.4), (2.5), and (2.6) hold, and there exist constants C_1 and $\zeta \in (2/q, 1)$ such that

$$\frac{\overline{\sigma}^2 D_{\boldsymbol{N}}^2 \log^5 p}{n} \le C_1 n^{-\zeta}. \tag{2.12}$$

Under Case (i'), for any $\nu \in (1/\zeta, \infty)$, there exists a constant C depending only on ν, K, and C_1 such that

$$\mathbb{P}\left(\overline{\sigma}^2 \hat{\Delta}_W \log^4 p > C n^{-\zeta + 1/\nu}\right) \le C n^{-1}.$$

Under Case (ii'), there exists a constant C depending only on q, K, and C_1 such that

$$\mathbb{P}\left(\overline{\sigma}^2 \hat{\Delta}_W \log^4 p > C n^{-\zeta + 2/q}\right) \le C n^{-1}.$$

Remark 4 (Discussion on Conditions (2.11) and (2.12))**.** If we are to follow Remark 3 and to show a sufficient condition for $\overline{\sigma}\hat{\Delta}_W \log^4 p = o_P(1)$, then Conditions (2.11) and (2.12) can be weakened to $\overline{\sigma}^2 D_{\boldsymbol{N}}^2 \log^7 p = o(n)$ and $\overline{\sigma}^2 D_{\boldsymbol{N}}^2 \log^5 p = o(n^{1-2/q})$, respectively.

In practice, we often normalize the coordinates of the sample mean by estimates of the standard deviations, so that each coordinate is approximately distributed as $N(0,1)$. In view of the high-dimensional CLT, the approximate variance of the j-th coordinate of $\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}}$ is given by $\sigma_j^2 = \mathrm{Var}(\sqrt{n} S_{\boldsymbol{N}}^{W,j})$, where $S_{\boldsymbol{N}}^{W,j}$ is the j-th coordinate of $\boldsymbol{S}_{\boldsymbol{N}}^W$. This can be estimated by

$$\hat{\sigma}_j^2 = \sum_{k=1}^K \frac{n}{N_k^2} \sum_{i_k=1}^{N_k} (\overline{X}_{k,i_k}^j - S_{\boldsymbol{N}}^j)^2.$$

Let $\Lambda = \mathrm{diag}\{\sigma_1^2, \dots, \sigma_p^2\}$ and $\hat{\Lambda} = \mathrm{diag}\{\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2\}$. We consider to approximate the distribution of $\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}}$ by $\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}}^{MB}$.

Corollary 1. *Consider Cases (i) and (ii) in Theorem 2. In Case (i), assume further that*

$$\frac{D_{\boldsymbol{N}}^2 \log^7(p\overline{N})}{n} \le C_1 n^{-3(\zeta_1 \wedge \zeta_2)/2},$$

while in Case (ii) assume further that

$$\frac{D_{\boldsymbol{N}}^2 \log^7(p\overline{N})}{n} \bigvee \left(\frac{D_{\boldsymbol{N}}^2 \log^3(p\overline{N})}{n^{1-2/q}}\right)^2 \le C_1 n^{-3(\zeta_1 \wedge \zeta_2)/2}.$$

Then, there exists a constant C such that for $\boldsymbol{Y} \sim N(\boldsymbol{0}, \Sigma)$,

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \in R) - \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) \right| \le C n^{-(\zeta_1 \wedge \zeta_2)/4} \quad \textit{and}$$
$$\mathbb{P}\left\{ \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\boldsymbol{X}_{[\boldsymbol{N}]}}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}}^{MB} \in R) - \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) \right| \le C n^{-(\zeta_1 \wedge \zeta_2)/4} \right\} \ge 1 - C n^{-1},$$

where the same convention on the constant C as in Theorem 2 applies.

3. Polyadic Data

In this section, we consider another class of exchangeable arrays, namely, jointly exchangeable arrays, which correspond to polyadic data. The notations in the current section are independent from those in Section 2 unless otherwise noted. Joint exchangeability induces a more complex dependence structure on arrays than separate exchangeability, but still we are able to develop analogous results to the preceding section for jointly exchangeable arrays as well. It should be noted, however, that we do require a different bootstrap and technical tools (cf. Appendix C) to accommodate a specific dependence structure induced from joint exchangeability.

Pick any $K \in \mathbb{N}$. For a given positive integer $n \geq K$, let $I_{n,K} = \{(i_1, \ldots, i_K) : 1 \leq i_1, \ldots, i_K \leq n \text{ and } i_1, \ldots, i_K \text{ are distinct}\}$. Also let $I_{\infty,K} = \bigcup_{n=K}^{\infty} I_{n,K}$. For any $\boldsymbol{i} = (i_1, \ldots, i_K) \in \mathbb{N}^K$, let $\{\boldsymbol{i}\}^+$ denote the set of distinct *nonzero* elements of $(i_1, \ldots, i_K)$. For example, $\{(2,0,1,2)\}^+ = \{1,2\}$.

In this section, we consider a K-array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ consisting of random vectors in $\mathbb{R}^p$. We say that the array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ is *jointly exchangeable* if the following condition is satisfied (cf. Kallenberg, 2006, Section 3.1).

Definition 2 (Joint exchangeability). *A K-array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ is called jointly exchangeable if for any permutation π of $\mathbb{N}$, the arrays $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ and $(\boldsymbol{X}_{(\pi(i_1), \ldots, \pi(i_K))})_{\boldsymbol{i} \in I_{\infty,K}}$ are identically distributed.*

From the Aldous-Hoover-Kallenberg representation (see Kallenberg, 2006, Theorem 7.22), any jointly exchangeable array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ is generated by the structure

$$\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{f}((U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \{0,1\}^K}), \ \boldsymbol{i} \in I_{\infty,K}, \quad \{U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} : \boldsymbol{i} \in I_{\infty,K}, \boldsymbol{e} \in \{0,1\}^K\} \overset{i.i.d.}{\sim} U[0,1]$$

for some Borel measurable map $\mathfrak{f} : [0,1]^{2^K} \to \mathbb{R}^p$. For example, when $K = 2$, then $\boldsymbol{X}_{(i_1,i_2)}$ is generated as $\boldsymbol{X}_{(i_1,i_2)} = \mathfrak{f}(U_\varnothing, U_{i_1}, U_{i_2}, U_{\{i_1,i_2\}})$. (We will write $U_{i_k} = U_{\{i_k\}}$ for the notational convenience.) Here the coordinates of the vector $(U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \{0,1\}^K}$ are understood to be properly ordered, so that, e.g., when $K = 2$, $\boldsymbol{X}_{(i_1,i_2)} = \mathfrak{f}(U_\varnothing, U_{i_1}, U_{i_2}, U_{\{i_1,i_2\}})$ and $\boldsymbol{X}_{(i_2,i_1)} = \mathfrak{f}(U_\varnothing, U_{i_2}, U_{i_1}, U_{\{i_1,i_2\}})$ differ (although they have the identical distribution).

As in the separately exchangeable case, we consider inference conditional on $U_\varnothing$, and in what follows, we will assume that the array $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ has mean zero (conditional on $U_\varnothing$) and is generated by the structure

$$\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{g}((U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \{0,1\}^K \setminus \{\boldsymbol{0}\}}), \ \boldsymbol{i} \in I_{\infty,K}, \tag{3.1}$$

where $\mathfrak{g}$ is now a map from $[0,1]^{2^K - 1}$ into $\mathbb{R}^p$.

Suppose that we observe $\{\boldsymbol{X}_{\boldsymbol{i}} : \boldsymbol{i} \in I_{n,K}\}$ with $n \geq K$ and are interested in distributional approximation of the polyadic sample mean

$$\boldsymbol{S}_n := \frac{(n-K)!}{n!} \sum_{\boldsymbol{i} \in I_{n,K}} \boldsymbol{X}_{\boldsymbol{i}}.$$

in the high-dimensional setting where the dimension p is allowed to entail $p \gg n$.

As in Section 2, define $\mathcal{E}_k = \{\boldsymbol{e} = (e_1, \ldots, e_K) \in \{0,1\}^K : \sum_{k=1}^K e_k = k\}$ for $1 \le k \le K$. The analysis of the polyadic sample mean relies on the following decomposition of $\boldsymbol{X}_{\boldsymbol{i}}$:

$$\boldsymbol{X}_{\boldsymbol{i}} = \sum_{k=1}^K \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{i_k}] + \left(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{i_1}, \ldots, U_{i_K}] - \sum_{k=1}^K \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{i_k}] \right)$$
$$+ \sum_{k=2}^K \left(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \cup_{r=1}^k \mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \cup_{r=1}^{k-1} \mathcal{E}_r}] \right).$$

This leads to the decomposition

$$\begin{aligned}\boldsymbol{S}_n = \frac{1}{n} \sum_{j=1}^n \mathbb{E}\left[\frac{(n-K)!}{(n-1)!} \sum_{k=1}^K \sum_{\boldsymbol{i} \in I_{n,K} : i_k = j} \boldsymbol{X}_{\boldsymbol{i}} \;\middle|\; U_j \right] \\ + \frac{(n-K)!}{n!} \sum_{\boldsymbol{i} \in I_{n,K}} \left(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{i_1}, \ldots, U_{i_K}] - \sum_{k=1}^K \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{i_k}] \right) \\ + \sum_{k=2}^K \frac{(n-K)!}{n!} \sum_{\boldsymbol{i} \in I_{n,K}} \left(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \cup_{r=1}^k \mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \cup_{r=1}^{k-1} \mathcal{E}_r}] \right).\end{aligned} \tag{3.2}$$

The second term on the right-hand side of (3.2) is a degenerate U-statistic and thus negligible compared with the first term under moment conditions (this term can be expanded into $K-1$ terms each of which scales as $O(n^{-k/2})$ if p is fixed for $k = 2, \ldots, K$). The analysis of the third term is more complicated but it will be shown that the k-th term inside the first summation scales as $O(n^{-k/2})$ if the dimension p is fixed, so that the third term on the right-hand side of (3.2) is also negligible compared with the first term. See Appendix C for details. Applying the Hoeffding decomposition to the second term on the right-hand side of (3.2), combining it with the third term on the right-hand side of (3.2), and aligning the terms according to their orders, we can obtain a Hoeffding-type decomposition for jointly exchangeable arrays. As in the multiway clustering case, we call the first term on the right-hand side of (3.2) the Hájek projection of $\boldsymbol{S}_n$.

Defining $h_k(u) = \mathbb{E}[\boldsymbol{X}_{(1,\ldots,K)} \mid U_k = u]$ for $k = 1, \ldots, K$, we can simplify the Hájek projection into

$$\boldsymbol{S}_n^W = \frac{1}{n} \sum_{j=1}^n \boldsymbol{W}_j, \quad \text{with} \quad \boldsymbol{W}_j = \sum_{k=1}^K h_k(U_j).$$

Since $\{\boldsymbol{W}_j\}_{j=1}^n$ are i.i.d., we can expect that $\sqrt{n}\boldsymbol{S}_n^W$ can be approximated (in distribution) by $N(\mathbf{0}, \Sigma)$, where

$$\Sigma = \mathbb{E}\left[\boldsymbol{W}_1 \boldsymbol{W}_1^T\right].$$

This suggests the following version of multiplier bootstrap for polyadic data.

3.1. **Multiplier bootstrap for polyadic data.** Let $\{\xi_j\}_{j=1}^n$ be independent $N(0,1)$ random variables independent of the data. Ideally, we want to make use of the multiplier bootstrap statistic

$$\frac{1}{n} \sum_{j=1}^n \xi_j (\boldsymbol{W}_j - K\boldsymbol{S}_n).$$

This is infeasible, however, as the projections $\boldsymbol{W}_j$ are unknown. As an alternative, we replace each $\boldsymbol{W}_j$ by its estimate

$$\hat{\boldsymbol{W}}_j = \frac{(n-K)!}{(n-1)!} \sum_{k=1}^{K} \sum_{\boldsymbol{i} \in I_{n,K}: i_k = j} \boldsymbol{X}_{\boldsymbol{i}},$$

and apply the multiplier bootstrap to $\hat{\boldsymbol{W}}_j$, i.e.,

$$\boldsymbol{S}_n^{MB} := \frac{1}{n} \sum_{j=1}^{n} \xi_j (\hat{\boldsymbol{W}}_j - K \boldsymbol{S}_n)$$

For example, when $K = 2$ (dyadic), this mulitplier bootstrap simplifies into

$$\boldsymbol{S}_n^{MB} = \frac{1}{n} \sum_{j=1}^{n} \xi_j \left\{ \frac{1}{(n-1)} \sum_{i'=1; i' \neq j}^{n} (\boldsymbol{X}_{(i',j)} + \boldsymbol{X}_{(j,i')}) - 2\boldsymbol{S}_n \right\},$$

which coincides with the multiplier bootstrap statistic considered in Section 3.2 of Davezies et al. (2020). However, Davezies et al. (2020) do not consider the extension to general K arrays, and focus on the empirical process indexed by a Donsker class, which excludes the high-dimensional sample mean. We will study the validity of this multiplier bootstrap for general polyadic data in the following two subsections.

3.2. **High-dimensional CLT for polyadic data.** We consider to approximate the distribution of $\sqrt{n}\boldsymbol{S}_n$ by a Gaussian distribution on the set of rectangles $\mathcal{R}$ as defined in Section 2.

Let $D_n \geq 1$ be a given constant that may depend on n, and $\underline{\sigma} > 0$ be another given constant independent of n. We will assume either of the following moment conditions.

$$\max_{1 \leq \ell \leq p} \|X^{\ell}_{(1,\dots,K)}\|_{\psi_1} \leq D_n, \quad \text{or} \tag{3.3}$$

$$\mathbb{E}[\|\boldsymbol{X}_{(1,\dots,K)}\|_\infty^q] \leq D_n^q \quad \text{for some } q \in (4,\infty). \tag{3.4}$$

We will also assume both of the following conditions.

$$\max_{1 \leq \ell \leq p} \mathbb{E}[|W_1^{\ell}|^{2+k}] \leq D_n^k, \ \kappa = 1, 2, \tag{3.5}$$

$$\min_{1 \leq \ell \leq p} \mathbb{E}[|W_1^{\ell}|^2] \geq \underline{\sigma}^2. \tag{3.6}$$

The conditions required here are similar to those in the case of multiway clustering in Section 2. The main difference is that Conditions (3.5) and (3.6) are now imposed on $\boldsymbol{W}_1$.

Let $\gamma_\Sigma = N(\mathbf{0}, \Sigma)$.

Theorem 3 (High-dimensional CLT for polyadic data)**.** *Suppose that either Condition (3.3) or (3.4) holds, and further both Conditions (3.5) and (3.6) hold. Then, there exists a constant C such that*

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\boldsymbol{S}_n \in R) - \gamma_\Sigma(R) \right|$$
$$\leq \begin{cases} C \left(\frac{D_n^2 \log^7(pn)}{n} \right)^{1/6} & \textit{if Condition (3.3) holds,} \\ C \left[\left(\frac{D_n^2 \log^7(pn)}{n} \right)^{1/6} + \left(\frac{D_n^2 \log^3(pn)}{n^{1-2/q}} \right)^{1/3} \right] & \textit{if Condition (3.4) holds,} \end{cases}$$

where the constant C depends only on $\underline{\sigma}$ and K if Condition (3.3) holds, while C depends only on $q, \underline{\sigma}$, and K if Condition (3.4) holds.

Remark 5 (Comparison with Silverman (1976))**.** Theorem 3 is a high-dimensional extension of Theorem A in Silverman (1976) that establishes a CLT for jointly exchangeable arrays with fixed p. The covariance matrix of the limiting Gaussian distribution in Silverman (1976) has a different expression than our Σ, but we will verify below that two expressions are indeed the same. The covariance matrix given in Corollary to Theorem A in Silverman (1976) reads as follows: Let $\check{\boldsymbol{X}}_{(i_1,\dots,i_K)}$ be the symmetrized version of $\boldsymbol{X}_{(i_1,\dots,i_K)}$, i.e., $\check{\boldsymbol{X}}_{(i_1,\dots,i_K)} = (K!)^{-1}\sum_{(i_1',\dots,i_K')}\boldsymbol{X}_{(i_1',\dots,i_K')}$ where the summation is taken over all permutations of $(i_1,\dots,i_K)$. The covariance matix given in Silverman (1976) is $\Sigma_S = K^2\mathbb{E}[\check{\boldsymbol{X}}_{(1,\dots,K)}\check{\boldsymbol{X}}_{(1,K+1,\dots,2K)}]$. On the other hand,

$$\sum_{k=1}^K \mathbb{E}[\boldsymbol{X}_{(1,\dots,K)} \mid U_k = u] = \sum_{k=1}^K \mathbb{E}[\check{\boldsymbol{X}}_{(1,\dots,K)} \mid U_k = u] = K\mathbb{E}[\check{\boldsymbol{X}}_{(1,\dots,K)} \mid U_1 = u],$$

so that

$$\Sigma = K^2\mathbb{E}\left[\mathbb{E}[\check{\boldsymbol{X}}_{(1,\dots,K)} \mid U_1]\mathbb{E}[\check{\boldsymbol{X}}_{(1,\dots,K)} \mid U_1]\right] = K^2\mathbb{E}[\check{\boldsymbol{X}}_{(1,\dots,K)}\check{\boldsymbol{X}}_{(1,K+1,\dots,2K)}] = \Sigma_S,$$

as claimed.

3.3. **Validity of multiplier bootstrap for polyadic data.** Let $\mathbb{P}_{|\boldsymbol{X}_{I_{n,K}}}$ denote the law conditional on the data $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i}\in I_{n,K}}$. Define

$$\hat{\Delta}_{W,1} = \max_{1\le \ell\le p}\frac{1}{n}\sum_{j=1}^n(\hat{W}_j^\ell - W_j^\ell)^2.$$

In addition, let $\overline{\sigma} = \max_{1\le\ell\le p}\sqrt{\mathbb{E}[|W_1^\ell|^2]}$.

Theorem 4 (Validity of multiplier bootstrap for polyadic data)**.** *Consider the following two cases.*

(i). Conditions (3.3), (3.5), and (3.6) hold, and there exist constants C_1 and ζ_1, $\zeta_2 \in (0,1)$ such that

$$\mathbb{P}\left(\overline{\sigma}^2\hat{\Delta}_{W,1}\log^4 p > C_1 n^{-\zeta_2}\right) \le C_1 n^{-1} \quad \textit{and} \tag{3.7}$$

$$\frac{D_n^2(\log^2 n)\log^5(pn)}{n} \le C_1 n^{-\zeta_1}. \tag{3.8}$$

(ii). Conditions (3.4), (3.5), and (3.6) hold, and there exist constants C_1 and ζ_1, $\zeta_2 \in (0,1)$ such that Condition (3.7) holds and

$$\frac{D_n^2\log^5(pn)}{n}\bigvee\left(\frac{D_n^2\log^3 p}{n^{1-4/q}}\right)^2 \le C_1 n^{-\zeta_1}. \tag{3.9}$$

Then, under either Case (i) or (ii), there exists a constant C such that

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}_{|\boldsymbol{X}_{I_{n,K}}}(\sqrt{n}\boldsymbol{S}_n^{MB}\in R) - \gamma_\Sigma(R)\right| \le Cn^{-(\zeta_1\wedge\zeta_2)/4}.$$

with probability at least $1 - Cn^{-1}$, where the constant C depends only on $\underline{\sigma}, K$, and C_1 under Case (i), while C depends only on $q, \underline{\sigma}, K$, and C_1 under Case (ii).

The following proposition provides primitive sufficient conditions for Condition (3.7) to hold.

Proposition 2 (Primitive sufficient conditions for Condition (3.7))**.** *Consider the following two cases.*

- *(i') Conditions (3.3), (3.5), and (3.6) hold, and there exist constants C_1 and $\zeta \in (0,1)$ such that*
$$\frac{\overline{\sigma}^2 D_n^2 \log^7 p}{n} \le C_1 n^{-\zeta}.$$
- *(ii') Conditions (3.4), (3.5), and (3.6) hold, and there exist constants C_1 and $\zeta \in (2/q,1)$ such that*
$$\frac{\overline{\sigma}^2 D_n^2 \log^5 p}{n} \le C_1 n^{-\zeta}.$$

Under Case (i'), for any $\nu \in (1/\zeta, \infty)$, there exists a constant C depending only on ν, K, and C_1 such that
$$\mathbb{P}\left(\overline{\sigma}^2 \hat{\Delta}_{W,1} \log^4 p > C n^{-\zeta+1/\nu}\right) \le C n^{-1}.$$
Under Case (ii'), there exists a constant C depending only on q, K, and C_1 such that
$$\mathbb{P}\left(\overline{\sigma}^2 \hat{\Delta}_{W,1} \log^4 p > C n^{-\zeta+2/q}\right) \le C n^{-1}.$$

Finally, we consider normalized sample means for polyadic data. In light of the high-dimensional CLT for polyadic data, the approximate variance of the ℓ-th coordinate of $\sqrt{n}\boldsymbol{S}_n$ is given by $\sigma_\ell^2 = \mathrm{Var}(W_1^\ell)$, which can be estimated by
$$\hat{\sigma}_\ell^2 = \frac{1}{n}\sum_{k=1}^n (\hat{W}_k^\ell - K S_n^\ell)^2.$$
Let $\Lambda = \mathrm{diag}\{\sigma_1^2, \dots, \sigma_p^2\}$ and $\hat{\Lambda} = \mathrm{diag}\{\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2\}$. We consider to approximate the distribution of $\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_n$ by $\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_n^{MB}$.

Corollary 2. *Consider Cases (i) and (ii) in Theorem 4. In Case (i), assume further that*
$$\frac{D_n^2 \log^7(pn)}{n} \le C_1 n^{-3(\zeta_1 \wedge \zeta_2)/2},$$
while in Case (ii) assume further that
$$\frac{D_n^2 \log^7(pn)}{n} \bigvee \left(\frac{D_n^2 \log^3(pn)}{n^{1-2/q}}\right)^2 \le C_1 n^{-3(\zeta_1 \wedge \zeta_2)/2}.$$
Then, there exists a constant C such that for $\boldsymbol{Y} \sim N(\boldsymbol{0}, \Sigma)$,
$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_n \in R) - \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) \right| \le C n^{-(\zeta_1 \wedge \zeta_2)/4} \quad \textit{and}$$
$$\mathbb{P}\left\{ \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\boldsymbol{X}_{I_{n,K}}}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_n^{MB} \in R) - \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) \right| \le C n^{-(\zeta_1 \wedge \zeta_2)/4} \right\} \ge 1 - C n^{-1},$$
where the same convention on the constant C as in Theorem 4 applies.

The proof is analogous to Corollary 1 and thus omitted.

Remark 6 (Discussion on rate conditions for polyadic data)**.** Similar to Remark 3 and 4, if one is interested only in bootstrap consistency, Conditions (3.7)–(3.9 can be weakened to $\overline{\sigma}\hat{\Delta}_W \log^4 p = o_P(1), D_n^2 \log^5(pn) = o(n)$, and $(n^{-1} D_n^2 \log^5(pn)) \vee (n^{1-2/q} D_n^2 \log^3 p) = o(1)$, respectively. In addition, to show $\overline{\sigma}\hat{\Delta}_W \log^4 p = o_P(1)$, the two rate conditions in Proposition 2 can be weakened to $\overline{\sigma}^2 D_n^2 \log^7 p = o(n)$ and $\overline{\sigma}^2 D_n^2 \log^5 p = o(n^{1-2/q})$, respectively.

4. Applications

In this section, we illustrate four applications of our proposed methods and theories. Section 4.1 presents robust inference in demand analysis under differentiated products markets with market share data. Section 4.2 presents robust inference in extended gravity analysis with trade data. Section 4.3 presents how to construct confidence bands for densities of flows in polyadic data. Section 4.4 presents penalty choice for the Lasso and the performance of its corresponding estimate.

4.1. **Robust inference in demand analysis with market share data.** Market share data used for demand analysis under differentiated products markets naturally exhibit two-way clustering due to the economic structure of supply and demand. Typical market share data are double-indexed by products i and markets j. Observations are generally dependent across markets j due to a common supply shock generated by the producer of product i. Observations are also generally dependent across products i due to a common demand shock in market j. We illustrate an application of our proposed theory to the frontier approach of robust identification for demand models using this type of data.

Following Berry (1994) and Berry et al. (1995), consider a model of demand for N_1 products indexed by $i = 1, \ldots, N_1$ with an outside option $i = 0$ in N_2 markets. Consumer c derives utility $u_{cij} = \delta_{ij} + \epsilon_{cij}$ for product i in market j, where δ_{ij} is the mean utility and ϵ_{cij} denotes idiosyncratic shock with the type-I extreme value distribution. The mean utility is in turn modeled by $\delta_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\theta} + \eta_{ij}$, where $\boldsymbol{X}_{ij}$ is a vector of observed product and market characteristics and η_{ij} denotes unobserved characteristics. Suppose that each consumer c in market j chooses the product i yielding the highest utility, i.e., $s_{cij} = \mathbb{1}\{u_{ij} \geq u_{i'j} \text{ for all } i = 0, 1, \ldots, N_1\}$. Aggregation yields the product share $\pi_{ij} = \mathbb{E}[s_{cij} \mid \delta_{1j}, \ldots, \delta_{N_1 j}]$. The standard market share inversion in turn yields the mean utility $\delta_{ij} = \log \pi_{ij} - \log \pi_{0j}$. Suppose that we obtain instrumental variables $\boldsymbol{z}_{ij}$ that is mean orthogonal to the unobserved characteristics η_{ij}.

In this setup, the standard econometric approach uses the generalized method of moments (GMM) with the mean orthogonality condition $\mathbb{E}[\eta_{ij} \mid \boldsymbol{z}_{ij}] = 0$. However, due to zero and/or near-zero market shares in actual market share data, this standard approach is known to suffer from unbounded moments of moment functions. In this light, Gandhi et al. (2020) propose a robust identification approach. Specifically, they derive upper and lower bounds of mean utility functions, denoted by δ_{ij}^u and δ_{ij}^ℓ, respectively, and propose a family of moment inequalities of the form

$$H_0(\boldsymbol{\theta}) : \begin{cases} \mathbb{E}[(\boldsymbol{X}_{ij}^T\boldsymbol{\theta} - \delta_{ij}^u)g(\boldsymbol{z}_{ij})] \leq 0 \\ \mathbb{E}[-(\boldsymbol{X}_{ij}^T\boldsymbol{\theta} - \delta_{ij}^\ell)g(\boldsymbol{z}_{ij})] \leq 0 \end{cases}$$

for all $g \in \mathcal{G}$ in an infinite set $\mathcal{G}$ of non-negative instrumental functions – see Gandhi et al. (2020). While most existing studies (including Gandhi et al., 2020) on demand analyses do not account for statistical dependence within a product i or within a market j, robust inference can be achieved by accounting for the two-way dependence – see Chiang et al. (2019).

Applying our proposed method in Section 2, we may conduct inference for the utility parameters $\boldsymbol{\theta}$ under two-way clustered market share data in the following manner. Define the p-dimensional random vector $\boldsymbol{X}_{ij}(\boldsymbol{\theta})$ by

$$\boldsymbol{X}_{ij}(\boldsymbol{\theta}) = \big((\boldsymbol{X}_{ij}^T\boldsymbol{\theta} - \delta_{ij}^u)g_1(\boldsymbol{z}_{ij}), (\delta_{ij}^\ell - \boldsymbol{X}_{ij}^T\boldsymbol{\theta})g_1(\boldsymbol{z}_{ij}), \ldots, (\boldsymbol{X}_{ij}^T\boldsymbol{\theta} - \delta_{ij}^u)g_{p/2}(\boldsymbol{z}_{ij}), (\delta_{ij}^\ell - \boldsymbol{X}_{ij}^T\boldsymbol{\theta})g_{p/2}(\boldsymbol{z}_{ij})\big)^T$$

for an increasing number $p/2$ of instrumental functions $\{g_1, \dots, g_{p/2}\} \subset \mathcal{G}$. Define the test statistic $T_N(\boldsymbol{\theta}) = \max\{\boldsymbol{S_N}(\boldsymbol{\theta})\}$ (or its normalized version), where $\boldsymbol{S_N}(\boldsymbol{\theta}) = (N_1N_2)^{-1}\sum_{(i,j)\in[\boldsymbol{N}]} \boldsymbol{X}_{ij}(\boldsymbol{\theta})$. To approximate the distribution of $T_N(\boldsymbol{\theta})$, let $\hat{\boldsymbol{W}}_{1,i}(\boldsymbol{\theta}) = N_2^{-1}\sum_{j=1}^{N_2} \boldsymbol{X}_{ij}(\boldsymbol{\theta}) - \boldsymbol{S_N}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{W}}_{2,j}(\boldsymbol{\theta}) = N_1^{-1}\sum_{i=1}^{N_1} \boldsymbol{X}_{ij}(\boldsymbol{\theta}) - \boldsymbol{S_N}(\boldsymbol{\theta})$. Construct the multiplier process $\boldsymbol{S}_{\boldsymbol{N}}^{MB}(\boldsymbol{\theta}) = N_1^{-1}\sum_{i=1}^{N_1} \xi_{1,i}\hat{\boldsymbol{W}}_{1,i}(\boldsymbol{\theta}) + N_2^{-1}\sum_{j=1}^{N_2} \xi_{2,j}\hat{\boldsymbol{W}}_{2,j}(\boldsymbol{\theta})$, where $\{\xi_{1,i}\}$ and $\{\xi_{2,j}\}$ are independent $N(0,1)$ random variables independent of the data. Let $c(1-\alpha;\boldsymbol{\theta})$ denote the conditional $(1-\alpha)$-quantile of $\max\{\boldsymbol{S}_{\boldsymbol{N}}^{MB}(\boldsymbol{\theta})\}$. Our test rejects the null hypothesis $H_0(\boldsymbol{\theta})$ if $T_N(\boldsymbol{\theta}) > c(1-\alpha;\boldsymbol{\theta})$. Inverting this test provides a confidence region for the utility parameters $\boldsymbol{\theta}$.

4.2. **Robust inference in extended gravity analysis with trade data.** Trade data used for gravity analysis naturally exhibit two-way clustering due to the economic structure of supply and demand. Typical trade data are double-indexed by exporters i and importers j. Observations are generally dependent across importers j due to a common supply shock generated by the exporter i. Observations are also generally dependent across importers i due to a common demand shock in the destination j. We illustrate an application of our proposed theory to the frontier approach of robust identification in extended gravity analysis using this type of data.

Morales et al. (2019) introduces an extended gravity model with an implied static profit random function $\pi_{ijt}(\cdot)$ that firm i receives from exporting to country j in year t, where $\pi_{ijt}(\cdot)$ takes structural parameters $\boldsymbol{\theta}$ as arguments. Write $\pi_{ijj't}(\boldsymbol{\theta}) = \pi_{ijt}(\boldsymbol{\theta}) - \pi_{ij't}(\boldsymbol{\theta})$. We assumed to know the set $\mathcal{A}_{ijt}$ of all the countries j' that share the same cost structure with country j from the viewpoint of firm i in year t. Let d_{ijt} denote the indicator that firm i exports to country j in year t, let $\boldsymbol{z}_{ijt}$ denote a vector of variables including components of costs that depend on gravity and extended gravity variables, and let δ denote the rate of future discounting. In this setting and with these notations, Morales et al. (2019) propose a family of moment inequalities of the form

$$H_0(\boldsymbol{\theta}) : \mathbb{E}\left[\sum_{j'\in\mathcal{A}_{ijt}} g(\boldsymbol{z}_{ijt}, \boldsymbol{z}_{ij't})d_{ijt}(1-d_{ijt})(\pi_{ijj't}(\boldsymbol{\theta}) + \delta\pi_{ijj'(t+1)}(\boldsymbol{\theta}))\right] \geq 0$$

for all $g \in \mathcal{G}$ of non-negative functions satisfying certain restrictions – see Morales et al. (2019).

Define the p-dimensional random vector $\boldsymbol{X}_{ij}(\boldsymbol{\theta})$ by

$$\boldsymbol{X}_{ij}(\boldsymbol{\theta}) = \begin{pmatrix} -\sum_{j'\in\mathcal{A}_{ijt}} g_1(\boldsymbol{z}_{ijt}, \boldsymbol{z}_{ij't})d_{ijt}(1-d_{ijt})(\pi_{ijj't}(\boldsymbol{\theta}) + \delta\pi_{ijj'(t+1)}(\boldsymbol{\theta})) \\ \vdots \\ -\sum_{j'\in\mathcal{A}_{ijt}} g_p(\boldsymbol{z}_{ijt}, \boldsymbol{z}_{ij't})d_{ijt}(1-d_{ijt})(\pi_{ijj't}(\boldsymbol{\theta}) + \delta\pi_{ijj'(t+1)}(\boldsymbol{\theta})) \end{pmatrix}$$

for an increasing number p of instrumental functions $\{g_1, \dots, g_p\} \subset \mathcal{G}$. Define the test statistic $T_N(\boldsymbol{\theta}) = \max\{\boldsymbol{S_N}(\boldsymbol{\theta})\}$ (or its normalized version), where $\boldsymbol{S_N}(\boldsymbol{\theta}) = (N_1N_2)^{-1}\sum_{(i,j)\in[\boldsymbol{N}]} \boldsymbol{X}_{ij}(\boldsymbol{\theta})$. Then, confidence regions for $\boldsymbol{\theta}$ can be constructed as in the preceding section.

4.3. **Confidence bands for truncated densities of flows in dyadic data.** Researchers are often interested in "the densities of migration across states, trade across nations, liabilities across banks, or minutes of telephone conversation among individuals" (Graham et al., 2019). Densities of these flow measures use polyadic data. We illustrate an application of our proposed theory in Section 3 to constructing a confidence band for such density functions.

Following Graham et al. (2019), we suppose we observe the dyadic data $\{Y_{ij} : 1 \le i \ne j \le n\}$ that admits the structure

$$Y_{ij} = \mathfrak{g}(U_i, U_j, U_{\{i,j\}}) \tag{4.1}$$

where $\mathfrak{g}$ is symmetric in the first two arguments and hence $Y_{ij} = Y_{ji}$. We are interested in inference on the density of Y_{ij}. However, in certain empirical applications, such as international trade (see Head and Mayer, 2014)), a proportion of the variable of interest is zero. Hence we assume that Y_{ij} has a probability mass at zero, i.e. Y_{ij} is such that $\mathbb{P}(Y_{ij} \ne 0) = a \in (0, 1]$, and $Y_{ij} \sim f$ when $Y_{ij} \ne 0$, where f is a density function on $\mathbb{R}$. Let $b(y) = af(y)$ denote the scaled density. We may estimate $f(\cdot) = b(\cdot)/a$ by $\hat{f}(\cdot) = \hat{b}(\cdot)/\hat{a}$, where

$$\hat{a} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathbb{1}(Y_{ij} \ne 0), \quad \hat{b}(y) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} K_h(y - Y_{ij}) \mathbb{1}(Y_{ij} \ne 0).$$

Here $K : \mathbb{R} \to \mathbb{R}$ is a kernel function (a function that integrates to one), $K_h(\cdot) := h^{-1}K(\cdot/h)$, and $h = h_n \to 0$ is a bandwidth.

We consider to construct simultaneous confidence intervals (bands) for f over the set of design points $y_1, \dots, y_p$, where $p = p_n \to \infty$ is allowed. Define

$$\hat{X}^{\ell}_{ij} = \left\{ \frac{K_h(y_\ell - Y_{ij})}{\hat{a}} - \frac{\hat{b}(y_\ell)}{\hat{a}^2} \right\} \mathbb{1}(Y_{ij} \ne 0), \quad 1 \le i < j \le n, \quad \hat{X}^{\ell}_{ij} = \hat{X}^{\ell}_{ji}, \quad 1 \le j < i \le n,$$

for $\ell = 1, \dots, p$. Then, the multiplier bootstrap statistic is given by

$$\boldsymbol{S}^{MB}_n = \frac{1}{n} \sum_{i=1}^{n} \xi_j (\hat{\boldsymbol{W}}_j - 2\boldsymbol{S}_n), \text{ where } \boldsymbol{S}_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \hat{\boldsymbol{X}}_{ij} \quad \text{and} \quad \hat{W}^{\ell}_i = \frac{1}{n-1} \sum_{j \in \{1,\dots,n\} \setminus \{i\}} 2\hat{X}^{\ell}_{ij}.$$

For a given $\alpha \in (0, 1)$, consider the $(1 - \alpha)$-simultaneous confidence intervals defined by

$$\mathcal{I}(1 - \alpha) := \prod_{\ell=1}^{p} \left[\hat{f}(y_\ell) \pm \frac{\hat{c}(1 - \alpha)}{\sqrt{n}} \right] \quad \text{and} \quad \mathcal{I}^N(1 - \alpha) := \prod_{\ell=1}^{p} \left[\hat{f}(y_\ell) \pm \frac{\hat{\sigma}_\ell \hat{c}^N(1 - \alpha)}{\sqrt{n}} \right],$$

where $\hat{\sigma}^2_\ell = n^{-1} \sum_{k=1}^{n} (\hat{W}^{\ell}_k - 2S^{\ell}_n)^2$, $\hat{\Lambda} = \operatorname{diag}(\hat{\sigma}^2_1, \dots, \hat{\sigma}^2_p)$, $\hat{c}(1-\alpha)$ is the conditional $(1-\alpha)$-quantile of $\|\sqrt{n}\boldsymbol{S}^{MB}_n\|_\infty$, and $\hat{c}^N(1 - \alpha)$ is the conditional $(1 - \alpha)$-quantile of $\|\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}^{MB}_n\|_\infty$. The first method $\mathcal{I}(1 - \alpha)$ is a constant-length confidence band, while the second method $\mathcal{I}^N(1 - \alpha)$ is a variable-length confidence band based on Studentization.

The following proposition establishes asymptotic validity of the confidence bands. We will assume that there exists a conditional density of $Y_{i,j}$ given U_i, denoted by $f_{Y_{12}|U_1}(y \mid u)$. Let $\overline{f}_h(y) = \int K_h(y - z)f(z)dz$ denote the surrogate density. Recall that a kernel K is an r-th order kernel for some $r \ge 2$ if $\int y^t K(y)dy = 0$ for $t = 1, \dots, r - 1$ and $\int |y^r K(y)|dy < \infty$.

Proposition 3. *Suppose that: (i) the data is generated following Equation (4.1) with point mass at zero, $\mathbb{P}(Y_{ij} \ne 0) = a \in (0, 1]$ and $Y_{ij} \sim f$ with probability $(1 - a)$, and a is independent of n; (ii) the conditional density $f_{Y_{12}|U_1}$ is bounded by some constant independent of n; (iii) for the set of non-zero design points $\{y_1, \dots, y_p\} \subset \mathbb{R}$, $\min_{1 \le \ell \le p} \operatorname{Var}(f_{Y_{12}|U_1}(y_\ell \mid U_1))$ is greater than some positive constant independent of n; (iv) the kernel K is a bounded r-th order kernel for some $r \ge 2$; (iv) the bandwidth satisfies $h \to 0, nh^2 \to \infty$ as $n \to \infty$ and $\log^7(pn) = o(nh^2)$. Then we have*

$$\mathbb{P}\left((\overline{f}_h(y_\ell))_{\ell=1}^{p} \in \mathcal{I}(1 - \alpha) \right) \to (1 - \alpha) \quad \textit{and} \quad \mathbb{P}\left((\overline{f}_h(y_\ell))_{\ell=1}^{p} \in \mathcal{I}^N(1 - \alpha) \right) \to (1 - \alpha).$$

In addition, if f is at least r-continuously differentiable, $\|f^{(r)}\|_\infty < \infty$, and $nh^{2r}\log p = o(1)$, then

$$\mathbb{P}\left((f(y_\ell))_{\ell=1}^p \in \mathcal{I}(1-\alpha)\right) \to (1-\alpha) \quad \textit{and} \quad \mathbb{P}\left((f(y_\ell))_{\ell=1}^p \in \mathcal{I}^N(1-\alpha)\right) \to (1-\alpha).$$

Some comments on the proposition are in order.

Remark 7. (i) The assumption that $\mathfrak{g}$ in Equation (4.1) being symmetric in its first two arguments can in fact be relaxed. In such case, the conclusions in Proposition 3 continue to hold under the additional assumption that $f_{Y_{12}|U_2}$ is bounded by a constant that is independent of n. Also, when $a = 1$ and $r = 2$, the proposed dyadic kernel density estimator reduces to the estimator of Graham et al. (2020). The proposition complements Graham et al. (2020) by providing valid simultaneous confidence intervals for their dyadic kernel density estimator.

(ii) In some applications, such as in our empirical illustration in Section 6, the object of interest is $b(\cdot)$. For such case, one can simply omit the estimation of a by setting $\hat{a} = 1$ while keeping $\hat{b}(\cdot)$ unaltered. The conclusions in Proposition 3 continue to hold with this modification.

(iii) The proof of Proposition 3 does not follow directly from the results of Section 3, as we have to handle the estimation errors of $\hat{a}$ and $\hat{b}(\cdot)$, which involves additional substantial work.

4.4. **Penalty choice for Lasso under multiway clustering.** Consider a regression model

$$Y_{\boldsymbol{i}} = f(\boldsymbol{Z}_{\boldsymbol{i}}) + \varepsilon_{\boldsymbol{i}}, \quad \mathbb{E}[\varepsilon_{\boldsymbol{i}} \mid \boldsymbol{Z}_{\boldsymbol{i}}] = 0, \quad \boldsymbol{i} \in [\boldsymbol{N}],$$

where $Y_{\boldsymbol{i}}$ is a scalar outcome variable, $\boldsymbol{Z}_{\boldsymbol{i}} \in \mathbb{R}^d$ is a d-dimensional vector of covariates, $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function of interest, and $\varepsilon_{\boldsymbol{i}}$ is an error term. We approximate f by a linear combination of technical controls $\boldsymbol{X}_{\boldsymbol{i}} = P(\boldsymbol{Z}_{\boldsymbol{i}})$ for some transformation $P : \mathbb{R}^d \to \mathbb{R}^p$, i.e.,

$$f(\boldsymbol{Z}_{\boldsymbol{i}}) = \boldsymbol{X}_{\boldsymbol{i}}^T \beta_0 + r_{\boldsymbol{i}}, \, \boldsymbol{i} \in [\boldsymbol{N}],$$

where $r_{\boldsymbol{i}}$ is a bias term. The dimension p can be much larger than the cluster sizes $\boldsymbol{N}$, but we assume that the vector $\beta_0 \in \mathbb{R}^p$ is sparse in the sense that $\|\beta_0\|_0 = s \ll n$. Suppose that the array $\left((Y_{\boldsymbol{i}}, \boldsymbol{Z}_{\boldsymbol{i}}^T)^T\right)_{\boldsymbol{i} \in \mathbb{N}^K}$ is separately exchangeable and generated as

$$(Y_{\boldsymbol{i}}, \boldsymbol{Z}_{\boldsymbol{i}}^T)^T = \mathfrak{g}((U_{\boldsymbol{i} \odot \boldsymbol{e}})_{\boldsymbol{e} \in \{0,1\}^K \setminus \{\boldsymbol{0}\}}), \; \boldsymbol{i} \in \mathbb{N}^K, \quad \{U_{\boldsymbol{i} \odot \boldsymbol{e}} : \boldsymbol{i} \in \mathbb{N}^K, \boldsymbol{e} \in \{0,1\}^K \setminus \{\boldsymbol{0}\}\} \overset{i.i.d.}{\sim} U[0,1],$$

for some Borel measurable map $\mathfrak{g} : [0,1]^{2^K - 1} \to \mathbb{R}^{1+d}$.

Arguably, one of the most popular estimation methods for such a high-dimensional regression problem is the Lasso (Tibshirani, 1996); we refer to Bühlmann and van de Geer (2011); Giraud (2015); Wainwright (2019) as standard references on high-dimensional statistics. Let $N = \prod_{k=1}^K N_k$ denote the total sample size. The Lasso estimate for β_0 is defined by

$$\hat{\beta}^\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{N} \sum_{\boldsymbol{i} \in [\boldsymbol{N}]} (Y_{\boldsymbol{i}} - \boldsymbol{X}_{\boldsymbol{i}}^T \beta)^2 + \lambda \|\beta\|_1 \right\},$$

where $\lambda > 0$ is a penalty level. We estimate the vector $\boldsymbol{f} = (f_{\boldsymbol{i}})_{\boldsymbol{i} \in [\boldsymbol{N}]} = (f(\boldsymbol{Z}_{\boldsymbol{i}}))_{\boldsymbol{i} \in [\boldsymbol{N}]}$ by $\hat{\boldsymbol{f}}^\lambda = (\boldsymbol{X}_{\boldsymbol{i}}^T \hat{\beta}^\lambda)_{\boldsymbol{i} \in [\boldsymbol{N}]}$. Let $\|\boldsymbol{t}\|_{N,2}^2 = N^{-1} \sum_{\boldsymbol{i} \in [\boldsymbol{N}]} t_{\boldsymbol{i}}^2$ for $\boldsymbol{t} = (t_{\boldsymbol{i}})_{\boldsymbol{i} \in [\boldsymbol{N}]}$.

In what follows, we discuss the statistical performance of the Lasso estimate. Following Bickel et al. (2009), we say that Condition RE(s, c_0) holds (RE refers to "restricted eigenvalue") if, for a given positive constant $c_0 \geq 1$, the inequality

$$\kappa(s, c_0) = \min_{\substack{J \subset \{1,\dots,p\} \\ 1 \leq |J| \leq s}} \inf_{\substack{\theta \in \mathbb{R}^p, \, \theta \neq 0 \\ \|\theta_{J^c}\|_1 \leq c_0 \|\theta_J\|_1}} \frac{\sqrt{s N^{-1} \sum_{\boldsymbol{i} \in [\boldsymbol{N}]} (\theta^T \boldsymbol{X}_{\boldsymbol{i}})^2}}{\|\theta_J\|_1} > 0$$

holds with $J^c = \{1,\dots,p\} \setminus J$. Here for $\theta = (\theta_1,\dots,\theta_p)^T$ and $J \subset \{1,\dots,p\}$, $\theta_J = (\theta_j)_{j\in J}$. Keep in mind that as the covariates are random, the restricted eigenvalue $\kappa(s,c_0)$ is random as well.

Theorem 1 of Belloni and Chernozhukov (2013) implies that if, for a given $c > 1$,

- $\lambda \ge 2c\|\boldsymbol{S}_{\boldsymbol{N}}\|_\infty$ with $\boldsymbol{S}_{\boldsymbol{N}} = N^{-1}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}\varepsilon_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}}$ and
- Condition RE(s,c_0) holds with $c_0 = (c+1)/(c-1)$,

then the following nonasymptotic bounds hold with $\kappa = \kappa(s,c_0)$:

$$\|\hat{\boldsymbol{f}}^\lambda - \boldsymbol{f}\|_{N,2} \le 3\|\boldsymbol{r}\|_{N,2} + \left(1 + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{\kappa}.$$

To ensure that $\lambda \ge 2c\|\boldsymbol{S}_{\boldsymbol{N}}\|_\infty$ with high probability, say $1-\eta$ for some small $\eta > 0$, we will chose λ to be an estimate of the $(1-\eta)$-quantile of $2c\|\boldsymbol{S}_{\boldsymbol{N}}\|_\infty$. To this end, we first estimate the error terms $\varepsilon_{\boldsymbol{i}}$ by pre-estimating β_0 by the preliminary Lasso estimate $\tilde{\beta} = \hat{\beta}^{\lambda_0}$ with penalty $\lambda^0 = \tau_n(n^{-1}\log p)^{1/2}$ for some slowing growing sequence $\tau_n \to \infty$. In the following, we take $\tau_n = \log n$ for the sake of simplicity but other choices also work. We apply the multiplier bootstrap to $\tilde{\boldsymbol{S}}_{\boldsymbol{N}} = N^{-1}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}\tilde{\varepsilon}_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}}$ instead of $\boldsymbol{S}_{\boldsymbol{N}}$.

We note that Hájek projection to $\boldsymbol{S}_{\boldsymbol{N}}$ is given by $\sum_{k=1}^K N_k^{-1}\sum_{k=1}^{N_k}\boldsymbol{V}_{k,i_k}$, where

$$\boldsymbol{V}_{k,i_k} = \mathbb{E}[\varepsilon_{(1,\dots,1,i_k,1,\dots,1)}\boldsymbol{X}_{(1,\dots,1,i_k,1,\dots,1)} \mid U_{(0,\dots,0,i_k,0,\dots,0)}].$$

We estimate $\boldsymbol{V}_{k,i_k}$ by

$$\tilde{\boldsymbol{V}}_{k,i_k} = \Big(\prod_{k'\neq k} N_{k'}\Big)^{-1}\sum_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_K}\tilde{\varepsilon}_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}}.$$

Let $\{\xi_{1,i_1}\}_{i_1=1}^{N_1},\dots,\{\xi_{K,i_K}\}_{i_K=1}^{N_K}$ be i.i.d. $N(0,1)$ variables independent of the data, and consider

$$\Lambda_{\boldsymbol{N}}^{\xi} = \left\|\sum_{k=1}^K \frac{1}{N_k}\sum_{i_k=1}^{N_k}\xi_{k,i_k}(\tilde{\boldsymbol{V}}_{k,i_k} - \tilde{\boldsymbol{S}}_{\boldsymbol{N}})\right\|_\infty.$$

We propose to choose λ as

$$\lambda = \lambda(\eta) = 2c\Lambda_{\boldsymbol{N}}^{\xi}(1-\eta),$$

where $\Lambda_{\boldsymbol{N}}^{\xi}(1-\eta)$ denotes the conditional $(1-\eta)$-quantile of $\Lambda_{\boldsymbol{N}}^{\xi}$. We allow η to decrease with n, i.e, $\eta = \eta_n \to 0$.

The following proposition establishes the asymptotic validity of our choice of λ (as $n \to \infty$) under multiway clustering. In what follows, we understand that $s, p, \boldsymbol{N}, \eta$ are functions of n while other parameters such as $c, q, \underline{\kappa}$ are independent of n.

Proposition 4 (Penalty choice for the Lasso under multiway clustering)**.** *Suppose that: (i) there exist some constants $q \in [4,\infty)$ independent of n and $D_{\boldsymbol{N}}$ that may depend on $\boldsymbol{N}$ (and thus on n) such that $\mathbb{E}[|\varepsilon_{\boldsymbol{1}}|^{2q}] \vee \mathbb{E}[\|\boldsymbol{X}_{\boldsymbol{1}}\|_\infty^{2q}] \le D_{\boldsymbol{N}}^q$ and $\max_{1\le j\le p}\max_{1\le k\le K}\mathbb{E}[|V_{k,1}^j|^{2+\ell}] \le D_{\boldsymbol{N}}^\ell$ for $\ell = 1, 2$; (ii) $\mathbb{E}[|V_{k,1}^j|^2]$ is bounded and bounded away from zero uniformly in $1 \le j \le p$ and $1 \le k \le K$; (iii) there exists a positive constant $\underline{\kappa}$ independent of n such that $\kappa(s,c_0) \ge \underline{\kappa}$ with probability $1 - o(1)$; (iv) as $n \to \infty$, $\|\boldsymbol{r}\|_{N,2} = O(\sqrt{(s\log p)/n})$ and*

$$\frac{s\overline{N}^{1/q}D_{\boldsymbol{N}}^3\log^7(p\overline{N})}{n}\bigvee\frac{D_{\boldsymbol{N}}^2\log^5(pn)}{n^{1-2/q}} = o(1).$$

Then, we have $\lambda \geq 2c\|\boldsymbol{S}_{\boldsymbol{N}}\|_\infty$ *with probability* $1-\eta-o(1)$. *Further, we have*

$$\lambda = O_P\left(\sqrt{\frac{\log p}{n}} \bigvee \sqrt{\frac{\log(1/\eta)}{n}}\right).$$

Consequently, if we take $\eta = \eta_n \to 0$, *we have*

$$\|\hat{\boldsymbol{f}}^\lambda - \boldsymbol{f}\|_{N,2} = O_P\left(\sqrt{\frac{s\log p}{n}} \bigvee \sqrt{\frac{s\log(1/\eta)}{n}}\right).$$

The proof of Proposition 4 does not follow directly from the results of Section 2, as we have to take care of the estimation error of the preliminary Lasso estimate $\tilde{\beta}$, which requires extra work.

Condition (iii) in the preceding proposition is a high-level condition on the sample gram matrix. The following proposition provides primitive sufficient conditions for Condition (iii) to hold in the two-way clustering case, i.e., $K=2$.

Proposition 5 (RE condition under two-way clustering $K=2$)**.** *Consider* $K=2$ *and let* $B_{\boldsymbol{N}} = \sqrt{\mathbb{E}[\max_{\boldsymbol{i}\in[\boldsymbol{N}]}\|\boldsymbol{X}_{\boldsymbol{i}}\|_\infty^2]}$. *Suppose that the eigenvalues of* $\mathbb{E}[\boldsymbol{X}_{\boldsymbol{1}}\boldsymbol{X}_{\boldsymbol{1}}^T]$ *are bounded and bounded away from zero, and* $sB_{\boldsymbol{N}}^2\log^4(p\overline{N}) = o(n)$. *Then, there exists a positive constant* $\underline{\kappa}$ *independent of* n *such that* $\kappa(s,c_0) \geq \underline{\kappa}$ *with probability* $1-o(1)$.

Under Condition (i) of Proposition 4, $B_{\boldsymbol{N}} \leq \overline{N}^{1/q}D_{\boldsymbol{N}}$, so that $sB_{\boldsymbol{N}}^2\log^4(p\overline{N}) = o(n)$ reduces to $s\overline{N}^{1/q}D_{\boldsymbol{N}}\log^4(p\overline{N}) = o(n)$, which is implied by Condition (iv) of Proposition 4.

The proof of Proposition 5 relies on Lemma 2.7 in Lecué and Mendelson (2017) and an extension of Lemma P.1 in Belloni et al. (2018), whose proof in turn relies on the techniques in Rudelson and Vershynin (2008), from the i.i.d. case to two-way clustering.

Remark 8 (Column standardization)**.** For intepretability of the Lasso estimate, in practice, we often rescale the penalty by the weighted ℓ_1-norm (as in Belloni and Chernozhukov (2011) in the quantile regression case) to make sure that the coefficients are penalized in a comparable manner. All the results in this section continue to hold under this practice as the conditions assumed in Proposition 4 guarantee the sample second moment of each covariate is consistent uniformly over the coordinates.

5. Simulation Studies

5.1. **Uniform Coverage under Multiway Clustering.** In this section, we present simulation studies to evaluate finite sample performance of the proposed multiplier bootstrap method for multiway clustering. For simulation designs, we use two-way and three-way clustered sampling. With $\Sigma_{\boldsymbol{Z}}$ denoting the $p\times p$ covariance matrix consisting of elements of the form $4^{-|r-c|}$ in its (r,c)-th position, two-way clustered samples are generated according to

$$\boldsymbol{X}_{\boldsymbol{i}} = \frac{1}{4}\left(\boldsymbol{Z}_{(i_1,0)} + \boldsymbol{Z}_{(0,i_2)}\right) + \frac{1}{2}\boldsymbol{Z}_{(i_1,i_2)}.$$

where (i) $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}} \sim N(\boldsymbol{0},\Sigma_{\boldsymbol{Z}})$ independently for $\boldsymbol{i}\in\{(i_1,i_2)\in\mathbb{N}^2 : 1\leq i_1\leq N_1, 1\leq i_2\leq N_2\}$ and $\boldsymbol{e}\in\{0,1\}^2$ in one design, and (ii) $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}} \sim BN(\boldsymbol{0},\Sigma_{\boldsymbol{Z}}) + (1-B)N(\boldsymbol{0},2\Sigma_{\boldsymbol{Z}})$ and $B\sim$ Bernoulli(0.5) independently for $\boldsymbol{i}\in\{(i_1,i_2)\in\mathbb{N}^2 : 1\leq i_1\leq N_1, 1\leq i_2\leq N_2\}$ and $\boldsymbol{e}\in\{0,1\}^2$ in the other design. Likewise, three-way clustered samples are generated according to

$$\boldsymbol{X}_{\boldsymbol{i}} = \frac{1}{12}\left(\boldsymbol{Z}_{(i_1,0,0)} + \boldsymbol{Z}_{(0,i_2,0)} + \boldsymbol{Z}_{(0,0,i_3)} + \boldsymbol{Z}_{(i_1,i_2,0)} + \boldsymbol{Z}_{(i_1,0,i_3)} + \boldsymbol{Z}_{(0,i_2,i_3)}\right) + \frac{1}{2}\boldsymbol{Z}_{(i_1,i_2,i_3)},$$

where (i) $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}} \sim N(\boldsymbol{0}, \Sigma_{\boldsymbol{Z}})$ independently for $\boldsymbol{i} \in \{(i_1, i_2, i_3) \in \mathbb{N}^3 : 1 \leq i_1 \leq N_1, 1 \leq i_2 \leq N_2, 1 \leq i_3 \leq N_3\}$ and $\boldsymbol{e} \in \{0,1\}^3$ in one design, and (ii) $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}} \sim BN(\boldsymbol{0}, \Sigma_{\boldsymbol{Z}}) + (1-B)N(\boldsymbol{0}, 2\Sigma_{\boldsymbol{Z}})$ and $B \sim \text{Bernoulli}(0.5)$ independently for $\boldsymbol{i} \in \{(i_1, i_2, i_3) \in \mathbb{N}^3 : 1 \leq i_1 \leq N_1, 1 \leq i_2 \leq N_2, 1 \leq i_3 \leq N_3\}$ and $\boldsymbol{e} \in \{0,1\}^3$ in the other design. For each of these data generating designs, we run 2,500 Monte Carlo iterations to compute the uniform coverage frequencies of $\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}}]$ for the nominal probabilities of 80%, 90% and 95% using our proposed multiplier bootstrap for multiway clustering with 2,500 bootstrap iterations.

Tables 1 and 2 show simulation results for two-way cluster sampled data and three-way cluster sampled data, respectively. The columns consist of the dimension p of $\boldsymbol{X}$, and the two-way sample size (N_1, N_2) or the three-way sample size (N_1, N_2, N_3). The displayed numbers indicate the simulated uniform coverage frequencies for the nominal probabilities of 80%, 90% and 95%. For each dimension $p \in \{25, 50, 100\}$, sample sizes vary as $(N_1, N_2) \in \{(25,25), (50,50), (100,100)\}$ in Table 1, and sample sizes vary as $(N_1, N_2, N_3) \in \{(25,25,25), (50,50,50), (100,100,100)\}$ in Table 2. Observe that, for each nominal probability, the uniform coverage frequencies approach the nominal probability as the sample size increases. These results support the theoretical property of our multiplier bootstrap method. We ran many other sets of simulations with various designs and sample sizes not presented here, but this observed pattern to support our theory remains invariant across all the different sets of simulations.

5.2. **Uniform Coverage under Polyadic Data.** In this section, we present simulation studies to evaluate finite sample performance of the proposed multiplier bootstrap method for polyadic data. We shall focus on the the most common case in practice, the dyadic data, i.e. $K = 2$. With $\Sigma_{\boldsymbol{Z}}$ denoting the $p \times p$ covariance matrix consisting of elements of the form $4^{-|r-c|}$ in its (r,c)-th position, dyadic samples are generated according to

$$\boldsymbol{X}_{i,j} = \frac{1}{4}\left(\boldsymbol{Z}_{(i,0)} + \boldsymbol{Z}_{(j,0)}\right) + \frac{1}{2}\boldsymbol{Z}_{(i,j)},$$

where (i) $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}} \sim N(\boldsymbol{0}, \Sigma_{\boldsymbol{Z}})$ independently for $\boldsymbol{i} \in \{(i,j) \in \mathbb{N}^2 : 1 \leq i,j \leq n, i \neq j\}$ and $\boldsymbol{e} \in \{1\} \times \{0,1\}$ in one design, and (ii) $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}} \sim BN(\boldsymbol{0}, \Sigma_{\boldsymbol{Z}}) + (1-B)N(\boldsymbol{0}, 2\Sigma_{\boldsymbol{Z}})$ and $B \sim \text{Bernoulli}(0.5)$ independently for $\boldsymbol{i} \in \{(i,j) \in \mathbb{N}^2 : 1 \leq i,j \leq n, i \neq j\}$ and $\boldsymbol{e} \in \{1\} \times \{0,1\}$ in the other design. We run 2,500 Monte Carlo iterations to compute the uniform coverage frequencies of $\boldsymbol{S}_n$ for the nominal probabilities of 80%, 90% and 95% using our proposed multiplier bootstrap for multiway clustering with 2,500 bootstrap iterations.

Table 3 shows simulation results. The columns consist of the dimension p of $\boldsymbol{X}$, and the polyadic sample size N. The displayed numbers indicate the simulated uniform coverage frequencies for the nominal probabilities of 80%, 90% and 95%. For each dimension $p \in \{25, 50, 100\}$, sample sizes vary as $n \in \{50, 100, 200\}$. Observe that, for each nominal probability, the uniform coverage frequencies approach the nominal probability as the sample size increases. These results support the theoretical property of our multiplier bootstrap method. We ran many other sets of simulations with various designs and sample sizes not presented here, but this observed pattern to support our theory remains invariant across all the different sets of simulations.

5.3. **Uniform Confidence Band for Densities of Dyadic Data.** In this section, we present simulation studies to evaluate finite sample performance of the proposed uniform confidence bands for probability density functions of dyadic data that is presented in Section 4.3. Dyadic data are

Distribution of $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}}$	(i) Gaussian								
Normalization	No								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
Sample Sizes: N_1, N_2	25	50	100	25	50	100	25	50	100
80% Coverage	0.834	0.834	0.807	0.838	0.829	0.794	0.864	0.815	0.813
90% Coverage	0.928	0.921	0.909	0.935	0.925	0.906	0.943	0.916	0.910
95% Coverage	0.973	0.964	0.955	0.973	0.963	0.954	0.976	0.962	0.960
Normalization	Yes								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
Sample Sizes: N_1, N_2	25	50	100	25	50	100	25	50	100
80% Coverage	0.753	0.776	0.788	0.740	0.783	0.793	0.698	0.758	0.791
90% Coverage	0.876	0.889	0.895	0.860	0.882	0.900	0.834	0.876	0.896
95% Coverage	0.933	0.943	0.947	0.921	0.938	0.947	0.902	0.936	0.948

Distribution of $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}}$	(ii) Mixture								
Normalization	No								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
Sample Sizes: N_1, N_2	25	50	100	25	50	100	25	50	100
80% Coverage	0.824	0.817	0.803	0.859	0.841	0.814	0.864	0.828	0.814
90% Coverage	0.927	0.908	0.905	0.942	0.931	0.919	0.943	0.910	0.917
95% Coverage	0.967	0.954	0.956	0.976	0.968	0.960	0.973	0.957	0.962
Normalization	Yes								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
Sample Sizes: N_1, N_2	25	50	100	25	50	100	25	50	100
80% Coverage	0.747	0.772	0.785	0.716	0.768	0.783	0.711	0.776	0.789
90% Coverage	0.861	0.882	0.891	0.848	0.878	0.887	0.841	0.884	0.888
95% Coverage	0.925	0.939	0.940	0.912	0.938	0.944	0.914	0.941	0.942

TABLE 1. Simulation results for two-way ($K = 2$) cluster sampled data. Displayed are the dimension p of $\boldsymbol{X}$, the two-way sample size (N_1, N_2) with $N_1 = N_2$, and the simulated uniform coverage frequencies for the nominal probabilities of 80%, 90% and 95%.

generated according to

$$Y_{i,j} = \frac{1}{4}(U_{i,0} + U_{j,0}) + \frac{1}{2}U_{i,j},$$

where (i) $U_{\boldsymbol{i}\odot\boldsymbol{e}} \sim N(0,1)$ independently for $\boldsymbol{i} \in \{(i,j) \in \mathbb{N}^2 : 1 \leq i,j \leq n, i \neq j\}$ and $\boldsymbol{e} \in \{1\} \times \{0,1\}$ in one design, and (ii) $U_{\boldsymbol{i}\odot\boldsymbol{e}} \sim Logistic(0,1)$ independently for $\boldsymbol{i} \in \{(i,j) \in \mathbb{N}^2 : 1 \leq i,j \leq n, i \neq j\}$ and $\boldsymbol{e} \in \{1\} \times \{0,1\}$ in the other design.

We use the Epanechnikov kernel function K for estimation and inference for the probability density functions f of $Y_{i,j}$. We use the $n^{1/5}$-undersmoothed version of two Silverman's rules of thumb, i.e., (a) $h_n^1 = 1.06\hat{\sigma}_{Y_{i,j}} n^{-2/5}$ and (b) $h_n^2 = 0.9 \min\left\{\hat{\sigma}_{Y_{i,j}}, \widehat{IQR}_{Y_{i,j}}/1.34\right\} n^{-2/5}$ where $\hat{\sigma}_{Y_{i,j}}$ and $\widehat{IQR}_{Y_{i,j}}$ are the sample standard deviation and the sample interquartile range of $Y_{i,j}$, respectively. Confidence bands for f are constructed on the interval $[-2, 2]$ with the grid size of 201. We run 2,500 Monte Carlo iterations to compute the uniform coverage frequencies of f on this grid

Distribution of $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}}$	(i) Gaussian								
Normalization	No								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
Sample Sizes: N_1, N_2, N_3	25	50	100	25	50	100	25	50	100
80% Coverage	0.819	0.808	0.805	0.834	0.812	0.808	0.843	0.817	0.813
90% Coverage	0.912	0.912	0.910	0.932	0.914	0.908	0.929	0.918	0.902
95% Coverage	0.952	0.958	0.951	0.971	0.958	0.956	0.973	0.962	0.956
Normalization	Yes								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
Sample Sizes: N_1, N_2, N_3	25	50	100	25	50	100	25	50	100
80% Coverage	0.777	0.780	0.792	0.768	0.789	0.785	0.732	0.768	0.797
90% Coverage	0.879	0.884	0.892	0.874	0.890	0.888	0.852	0.878	0.898
95% Coverage	0.938	0.936	0.953	0.939	0.944	0.935	0.925	0.935	0.945

Distribution of $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}}$	(ii) Mixture								
Normalization	No								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
Sample Sizes: N_1, N_2, N_3	25	50	100	25	50	100	25	50	100
80% Coverage	0.829	0.823	0.810	0.824	0.822	0.810	0.852	0.818	0.803
90% Coverage	0.921	0.916	0.904	0.923	0.915	0.908	0.946	0.913	0.908
95% Coverage	0.964	0.958	0.952	0.960	0.958	0.956	0.974	0.959	0.958
Normalization	Yes								
Dimension of $\boldsymbol{X}_{\boldsymbol{i}}$: p	25	25	25	50	50	50	100	100	100
80% Coverage	0.777	0.786	0.776	0.779	0.767	0.789	0.741	0.763	0.796
90% Coverage	0.887	0.891	0.890	0.885	0.880	0.895	0.859	0.878	0.894
95% Coverage	0.940	0.943	0.940	0.939	0.938	0.943	0.924	0.939	0.946

Table 2. Simulation results for three-way ($K = 3$) cluster sampled data. Displayed are the dimension p of $\boldsymbol{X}$, the three-way sample size (N_1, N_2, N_3) with $N_1 = N_2 = N_3$, and the simulated uniform coverage frequencies for the nominal probabilities of 80%, 90% and 95%.

for the nominal probabilities of 80%, 90% and 95% using our proposed multiplier bootstrap for inference about the probability density functions of dyadic data with 2,500 bootstrap iterations.

Table 4 shows simulation results. The columns consist of the dyadic sample sizes $n \in \{250, 500\}$. The displayed numbers indicate the simulated uniform coverage frequencies for the nominal probabilities of 80%, 95% and 95%. Observe that, for each nominal probability and for each data generating design, the uniform coverage frequencies approach the nominal probability as the sample size increases. These results support the theoretical property of our multiplier bootstrap method for constructing uniform confidence bands for probability density functions of dyadic data.

6. Empirical Illustration

In this section, we present an empirical application of our proposed method in Section 4.3 to constructing uniform confidence bands for the density functions of bilateral trade volumes in the international trade, with a similar motivation to that stated in Graham et al. (2019, 2020). Recall

Distribution of $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}}$	(i) Gaussian								
Normalization	No								
Dimension of $\boldsymbol{X}_{i,j}$: p	25	25	25	50	50	50	100	100	100
Sample Size: n	50	100	200	50	100	200	50	100	200
80% Coverage	0.791	0.782	0.800	0.792	0.798	0.795	0.803	0.805	0.801
90% Coverage	0.902	0.898	0.901	0.909	0.898	0.905	0.909	0.906	0.912
95% Coverage	0.953	0.954	0.950	0.958	0.951	0.956	0.956	0.954	0.957
Normalization	Yes								
Dimension of $\boldsymbol{X}_{i,j}$: p	25	25	25	50	50	50	100	100	100
Sample Size: n	50	100	200	50	100	200	50	100	200
80% Coverage	0.713	0.744	0.780	0.664	0.736	0.770	0.621	0.718	0.768
90% Coverage	0.837	0.869	0.889	0.806	0.854	0.886	0.780	0.845	0.876
95% Coverage	0.918	0.928	0.946	0.887	0.923	0.943	0.867	0.915	0.942

Distribution of $\boldsymbol{Z}_{\boldsymbol{i}\odot\boldsymbol{e}}$	(ii) Mixture								
Normalization	No								
Dimension of $\boldsymbol{X}_{i,j}$: p	25	25	25	50	50	50	100	100	100
Sample Size: n	50	100	200	50	100	200	50	100	200
80% Coverage	0.777	0.781	0.786	0.778	0.798	0.786	0.794	0.801	0.797
90% Coverage	0.884	0.902	0.894	0.904	0.908	0.892	0.911	0.899	0.899
95% Coverage	0.948	0.953	0.952	0.960	0.958	0.950	0.957	0.953	0.954
Normalization	Yes								
Dimension of $\boldsymbol{X}_{i,j}$: p	25	25	25	50	50	50	100	100	100
Sample Size: n	50	100	200	50	100	200	50	100	200
80% Coverage	0.697	0.762	0.763	0.659	0.734	0.756	0.615	0.720	0.746
90% Coverage	0.824	0.870	0.878	0.807	0.863	0.870	0.773	0.857	0.870
95% Coverage	0.901	0.927	0.941	0.884	0.928	0.925	0.870	0.921	0.933

Table 3. Simulation results for dyadic data. Displayed are the dimension p of $\boldsymbol{X}$, the dyadic sample size n, and the simulated uniform coverage frequencies for the nominal probabilities of 80%, 90% and 95%.

that our method extends those by Graham et al. (2019) in that we can draw uniform confidence bands as opposed to point-wise confidence intervals. From this analysis, we can learn about the evolution of the distributions of international trade volumes over time.

We employ the international trade data used in Head and Mayer (2014), that come from the Direction of Trade Statistics (DoTS). This data set contains information about bilateral trade flows among 208 economies for 59 years from 1948 to 2006. In this analysis, we will focus on the relatively recent years, 1990, 1995, 2000 and 2005. Our measure of the bilateral trade volume Y_{ij} is defined as the logarithm of the sum of the trade flow from economy i to economy j and the trade flow from economy j to economy i. We use the same software code as that used for our simulation analysis presented in Section 5.3 to draw confidence bands of the probability density function of Y_{ij}. Since there is a probability mass at zero in the international trade volumes, what we estimate is precisely the Lebesgue-Radon-Nikodym derivative of the continuous part of the distribution, rather than the probability density function. Specifically, we use $\hat{b}(y)$ defined in Section 4.3 for estimation,

Distribution of $U_{\boldsymbol{i}\odot\boldsymbol{e}}$	(i) Gaussian					
Bandwidth Rule	(a) h_n^1			(b) h_n^2		
Sample Sizes: n	250	500	1000	250	500	1000
80% Coverage	0.712	0.788	0.790	0.678	0.778	0.787
90% Coverage	0.835	0.908	0.906	0.813	0.889	0.913
95% Coverage	0.902	0.953	0.962	0.880	0.949	0.959

Distribution of $U_{\boldsymbol{i}\odot\boldsymbol{e}}$	(ii) Logistic					
Bandwidth Rule	(a) h_n^1			(b) h_n^2		
Sample Sizes: n	250	500	1000	250	500	1000
80% Coverage	0.792	0.817	0.799	0.781	0.809	0.794
90% Coverage	0.906	0.916	0.914	0.899	0.914	0.908
95% Coverage	0.955	0.962	0.962	0.951	0.958	0.961

Table 4. Simulation results for uniform confidence bands on $[-2, 2]$ of probability density functions of dyadic data. Displayed are the dyadic sample sizes n and the simulated uniform coverage frequencies for the nominal probabilities of 80%, 90% and 95%.

and confidence bands are constructed by setting $\hat{a} = 1$. That said, we shall call it a density for conciseness.

Figures 1 and 2 illustrate estimates and confidence bands of the density functions of Y_{ij} in each of the years 1990, 1995, 2000 and 2005. Each panel of these figures displays the kernel density estimates in a solid curve and the 95% uniform confidence bands in a gray shade. In addition, we also display the proportion of zero bilateral trade volumes to the left of the kernel density plots so we can get an idea of the complementary proportion that consists the density of the continuously distributed part of the distribution. Although we treat Y_{ij} as the logarithm of the bilateral trade volumes in estimation and inference, we use the original scale (as opposed to the logarithm) on the horizontal axis for ease of reading the graphs.

Observe that the proportion of the zero trade volume is decreasing over time, and the density function is accordingly moving upward over time. Despite this pattern of the changes over time, the shapes of the density functions are rather similar in the middle of the distribution across time. This observation entails a high level of confidence given the reasonably tight confidence bands. However, notice that the right tail of the distribution becomes fatter as time progresses, implying that there is an increasing number of bilateral pairs with very large trade volumes. Again, this observation entails a high level of confidence given the tight confidence bands.

7. Summary

Empirical data in use for economic analysis are often clustered in two or more ways, where one source of dependence across units of demand is the common supply shock, and the the other source of dependence across units of supply is the common demand shock. When the set of agents generating the supply and the set of agents generating the demand are different, then such data is separately exchangeable or two-way clustered. Examples include market share data. When the set

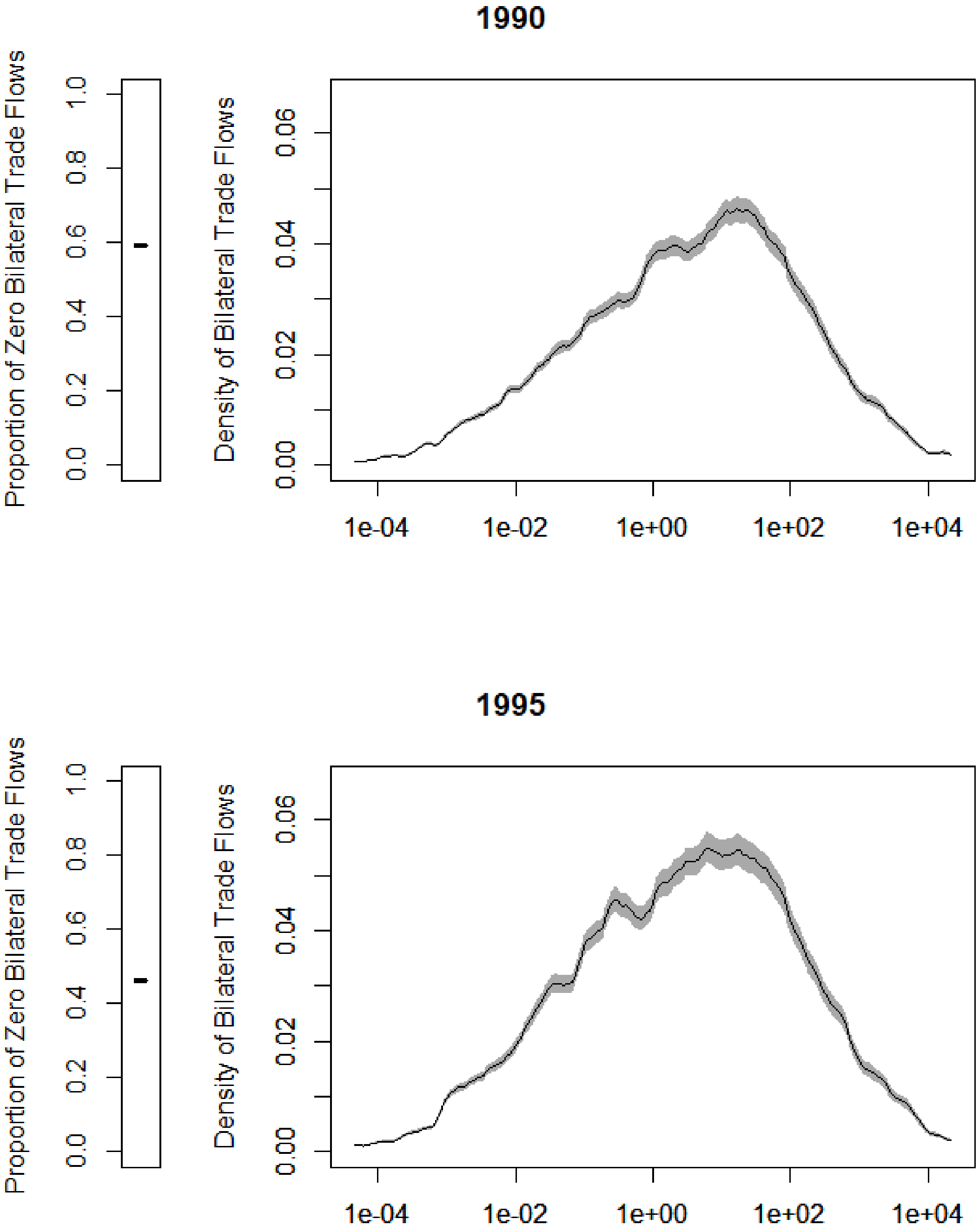

Figure 1. The kernel density estimates (solid curve) and the 95% uniform confidence bands (gray shade) of the bilateral trade volumes in 1990 and 1995.

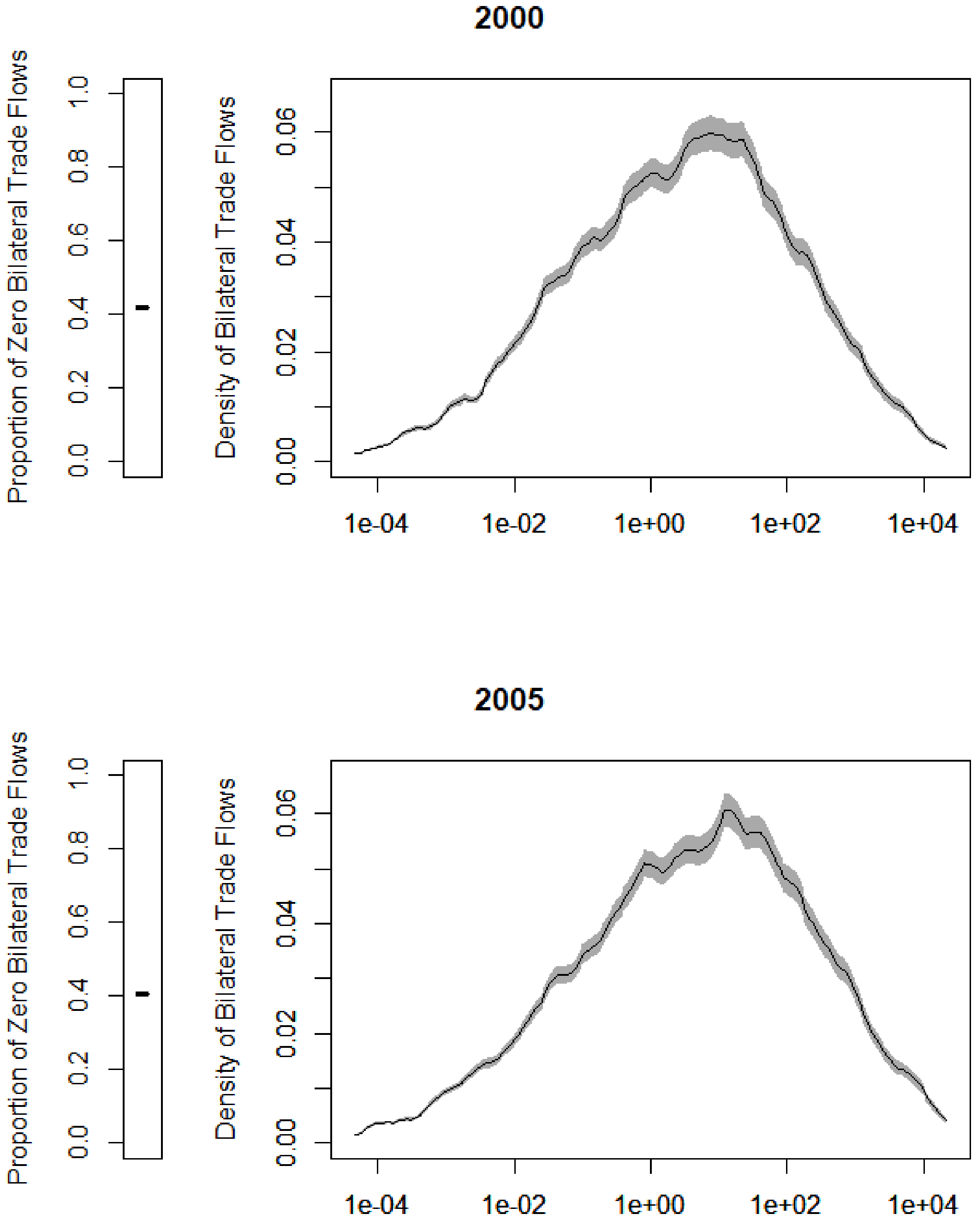

Figure 2. The kernel density estimates (solid curve) and the 95% uniform confidence bands (gray shade) of the bilateral trade volumes in 2000 and 2005.

of agents generating the supply and the set of agents generating the demand are the same, then such data is jointly exchangeable or dyadic. Examples include international trade data.

In this paper, for both separately exchangeable data and jointly exchangeable data, we develop methods and theories for inference about multi-dimensional, increasing-dimensional and high-dimensional parameters. Based on non-asymptotic Gaussian approximation error bounds for the test-statistic on hyper-rectangles, we propose bootstrap methods and establish their finite sample validity. Simulation studies support the theoretical properties of the method.

Four applications of the proposed method are illustrated. For demand analysis with a two-way clustered data consisting of N_1 products and N_2 markets, Gandhi et al. (2020) derive high-dimensional moment inequalities. Similarly, for extended gravity analysis with a two-way clustered data consisting of N_1 firms and N_2 countries, Morales et al. (2019) derive high-dimensional moment inequalities. With our theory of approximating the distribution of a multiway sample mean of a high-dimensional random vector, inverting the Kolmogorov-Smirnov test allows for inference about the structural parameters in these two settings. Third, extending Graham et al. (2019), our method was demonstrated to apply to construction of uniform confidence bands for probability density functions of dyadic data. Finally, we also demonstrate an application of our proposed method to penalty tuning parameter choice for ℓ_1-penalized regression under multiway cluster sampling. As such, our basic theory paves the way for a variety of applications to analyses of multiway-clustered and dyadic/polyadic data in econometrics.

Appendix

Appendix A. Maximal Inequalities for Multiway Clustering

In this section, we shall develop maximal inequalities for separately exchangeable arrays. As in Section 2, let $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^K}$ be a K-array consisting of random vectors in $\mathbb{R}^p$ with mean zero generated by the structure (2.1), i.e., $\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{g}((U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\boldsymbol{e}\in\{0,1\}^K\setminus\{\boldsymbol{0}\}})$ for $\boldsymbol{i} \in \mathbb{N}^K$. We will follow the notations used in Section 2. The following theorem is fundamental.

Theorem 5. *Pick any $1 \le k \le K$ and $\boldsymbol{e} \in \mathcal{E}_k$. Then, for any $q \in [1,\infty)$, we have*

$$\left(\mathbb{E}\left[\left\|\sum_{\boldsymbol{i}\in I_{\boldsymbol{e}}([\boldsymbol{N}])}\hat{\boldsymbol{X}}_{\boldsymbol{i}}\right\|_\infty^q\right]\right)^{1/q} \le C(\log p)^{k/2}\left(\mathbb{E}\left[\max_{1\le j\le p}\left(\sum_{\boldsymbol{i}\in I_{\boldsymbol{e}}([\boldsymbol{N}])}|\hat{X}_{\boldsymbol{i}}^j|^2\right)^{q/2}\right]\right)^{1/q},$$

where C is a constant that depends only on q and K.

The following corollary is immediate from Jensen's inequality.

Corollary 3 (Global maximal inequality)**.** *For any $1 \le k \le K, \boldsymbol{e} \in \mathcal{E}_k$, and $q \in [1,\infty)$, we have*

$$\left(\mathbb{E}\left[\left\|\sum_{\boldsymbol{i}\in I_{\boldsymbol{e}}([\boldsymbol{N}])}\hat{\boldsymbol{X}}_{\boldsymbol{i}}\right\|_\infty^q\right]\right)^{1/q} \le C(\log p)^{k/2}\sqrt{\prod_{k'\in\operatorname{supp}(\boldsymbol{e})}N_{k'}}(\mathbb{E}[\|\hat{\boldsymbol{X}}_{\boldsymbol{1}\odot\boldsymbol{e}}\|_\infty^{q\vee 2}])^{1/(q\vee 2)}, \qquad \text{(A.1)}$$

where C is a constant that depends only on q and K.

Remark 9. By Jensen's inequality, $\mathbb{E}[\|\hat{\boldsymbol{X}}_{\boldsymbol{1}\odot\boldsymbol{e}}\|_\infty^{q\vee 2}]$ on the right-hand side of (A.1) can be replaced by $\mathbb{E}[\|\boldsymbol{X}_{\boldsymbol{1}}\|_\infty^{q\vee 2}]$ by adjusting the constant C.

The proof of Theorem 5 relies on the following symmetrization inequality. Recall that a Rademacher random variable is a random variable taking ±1 with equal probability.

Lemma 2 (Symmetrization)**.** *Pick any $1 \le k \le K$. Let $\{\epsilon_{1,i_1}\}, \dots, \{\epsilon_{k,i_k}\}$ be independent Rademacher random variables independent of the U-variables. Then, for any nondecreasing convex function $\Phi : [0, \infty) \to [0, \infty)$, we have*

$$\mathbb{E}\left[\Phi\left(\left\|\sum_{i_1,\dots,i_k} \hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right)\right] \le \mathbb{E}\left[\Phi\left(2^k\left\|\sum_{i_1,\dots,i_k} \epsilon_{1,i_1}\cdots\epsilon_{k,i_k}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right)\right].$$

The proof of Lemma 2 in turn relies on the following result.

Lemma 3. *Let $\boldsymbol{i} \in \mathbb{N}^K$. Pick any $1 \le k \le K$ and let $\boldsymbol{e} \in \mathcal{E}_k$. Then, for any $\ell \in \operatorname{supp}(\boldsymbol{e})$, conditionally on $(U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_\ell}$, the vector $\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}}$ has mean zero.*

Proof of Lemma 3. For illustration, consider first the $K = 3$ case and $\boldsymbol{e} = (1, 1, 1)$. Then

$$\hat{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{X}_{\boldsymbol{i}} - \hat{\boldsymbol{X}}_{(i_1,i_2,0)} - \hat{\boldsymbol{X}}_{(0,i_2,i_3)} - \hat{\boldsymbol{X}}_{(i_1,0,i_3)} - \hat{\boldsymbol{X}}_{(i_1,0,0)} - \hat{\boldsymbol{X}}_{(0,i_2,0)} - \hat{\boldsymbol{X}}_{(0,0,i_3)}.$$

Given $(U_{(i_1,0,0)}, U_{(0,i_2,0)}, U_{(i_1,i_2,0)})$, we have

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{X}}_{(0,i_2,i_3)} \mid U_{(i_1,0,0)}, U_{(0,i_2,0)}, U_{(i_1,i_2,0)}] &= \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(0,i_2,0)}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(0,i_2,0)}] = 0, \\
\mathbb{E}[\hat{\boldsymbol{X}}_{(i_1,0,i_3)} \mid U_{(i_1,0,0)}, U_{(0,i_2,0)}, U_{(i_1,i_2,0)}] &= \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,0)}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,0)}] = 0.
\end{aligned}$$

Conclude that

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}} \mid U_{(i_1,0,0)}, U_{(0,i_2,0)}, U_{(i_1,i_2,0)}] &= \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,0)}, U_{(0,i_2,0)}, U_{(i_1,i_2,0)}] \\
&\quad - (\hat{\boldsymbol{X}}_{(i_1,i_2,0)} + \hat{\boldsymbol{X}}_{(i_1,0,0)} + \hat{\boldsymbol{X}}_{(0,i_2,0)}) \\
&= 0.
\end{aligned}$$

The proof for the general case is by induction on k. The conclusion is trivial when $k = 1$. Suppose that the lemma is true up to $k - 1$. Then,

$$\begin{aligned}
&\mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_\ell}] \\
&\quad = \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_\ell}] - \hat{\boldsymbol{X}}_{\boldsymbol{i}\odot(\boldsymbol{e}-\boldsymbol{e}_\ell)} \\
&\qquad - \sum_{\substack{\boldsymbol{e}'\le\boldsymbol{e} \\ \boldsymbol{e}'\ne\boldsymbol{e},\boldsymbol{e}-\boldsymbol{e}_\ell}} \mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}-\boldsymbol{e}_\ell}] \quad \text{(by the definition of } \hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}}\text{)} \\
&\quad = \sum_{\substack{\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_\ell \\ \boldsymbol{e}'\ne\boldsymbol{e}-\boldsymbol{e}_\ell}} \hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} - \sum_{\substack{\boldsymbol{e}'\le\boldsymbol{e} \\ \boldsymbol{e}'\ne\boldsymbol{e},\boldsymbol{e}-\boldsymbol{e}_\ell}} \mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}-\boldsymbol{e}_\ell}] \quad \text{(by plugging in the expansion of } \hat{\boldsymbol{X}}_{\boldsymbol{i}\odot(\boldsymbol{e}-\boldsymbol{e}_\ell)}\text{)} \\
&\quad = \sum_{\substack{\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_\ell \\ \boldsymbol{e}'\ne\boldsymbol{e}-\boldsymbol{e}_\ell}} \mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}-\boldsymbol{e}_\ell}] - \sum_{\substack{\boldsymbol{e}'\le\boldsymbol{e} \\ \boldsymbol{e}'\ne\boldsymbol{e},\boldsymbol{e}-\boldsymbol{e}_\ell}} \mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}-\boldsymbol{e}_\ell}] \\
&\quad = - \sum_{\substack{\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_{\ell'},\ell'\ne\ell \\ \ell\in\operatorname{supp}(\boldsymbol{e}'),\ell'\in\operatorname{supp}(\boldsymbol{e})}} \mathbb{E}\left[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}-\boldsymbol{e}_\ell}\right].
\end{aligned}$$

Here, we have used the fact that $\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'}$ is $\sigma((U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}'})$-measurable, so that $\mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}-\boldsymbol{e}_\ell}] = \hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'}$ as long as $\operatorname{supp}(\boldsymbol{e}') \subset \operatorname{supp}(\boldsymbol{e} - \boldsymbol{e}_\ell)$. For any $\boldsymbol{e}' \le \boldsymbol{e} - \boldsymbol{e}_{\ell'}$ with $\ell' \ne \ell, \ell \in \operatorname{supp}(\boldsymbol{e}')$, and

$\ell' \in \mathrm{supp}(\boldsymbol{e})$, we have

$$\mathbb{E}\left[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}-\boldsymbol{e}_\ell}\right] = \mathbb{E}\left[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}'} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}''})_{\boldsymbol{e}''\le\boldsymbol{e}'-\boldsymbol{e}_\ell}\right] = 0$$

by the induction hypothesis. Conclude that $\mathbb{E}[\hat{\boldsymbol{X}}_{\boldsymbol{i}\odot\boldsymbol{e}} \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_\ell}] = 0$. □

Proof of Lemma 2. Let $\boldsymbol{e} = (\underbrace{1,\dots,1}_{k},0,\dots,0)$. Given $(U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{i}\in[\boldsymbol{N}],\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_1}$, $\{\sum_{i_2,\dots,i_k}\hat{\boldsymbol{X}}_{(i_1,i_2\dots,i_k,0,\dots,0)} : i_1 = 1,\dots,N_1\}$ are independent with mean zero (the latter follows from Lemma 3). Hence, applying the symmetrization inequality (van der Vaart and Wellner (1996), Lemma 2.3.6) conditionally on $(U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{i}\in[\boldsymbol{N}],\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(\left\|\sum_{i_1,\dots,i_k}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right) \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{i}\in[\boldsymbol{N}],\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_1}\right] \\
&= \mathbb{E}\left[\Phi\left(\left\|\sum_{i_1}\left(\sum_{i_2,\dots,i_k}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right)\right\|_\infty\right) \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{i}\in[\boldsymbol{N}],\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_1}\right] \\
&\le \mathbb{E}\left[\Phi\left(2\left\|\sum_{i_1}\epsilon_{1,i_1}\left(\sum_{i_2,\dots,i_k}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right)\right\|_\infty\right) \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{i}\in[\boldsymbol{N}],\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_1}\right] \\
&= \mathbb{E}\left[\Phi\left(2\left\|\sum_{i_1,\dots,i_k}\epsilon_{1,i_1}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right) \mid (U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{i}\in[\boldsymbol{N}],\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_1}\right]
\end{aligned}$$

By Fubini's theorem, we have

$$\mathbb{E}\left[\Phi\left(\left\|\sum_{i_1,\dots,i_k}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right)\right] \le \mathbb{E}\left[\Phi\left(2\left\|\sum_{i_1,\dots,i_k}\epsilon_{1,i_1}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right)\right].$$

Next, given $\{\epsilon_{1,i_1}\} \cup \{U_{\boldsymbol{i}\odot\boldsymbol{e}'}\}_{\boldsymbol{i}\in[\boldsymbol{N}],\boldsymbol{e}'\le\boldsymbol{e}-\boldsymbol{e}_2}$, $\{\sum_{i_1,i_3,\dots,i_K}\epsilon_{1,i_1}\hat{\boldsymbol{X}}_{(i_1,i_2\dots,i_K,0,\dots,0)} : i_2 = 1,\dots,N_2\}$ are independent with mean zero, so that by the symmetrization inequality and Fubini's theorem, we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(2\left\|\sum_{i_1,\dots,i_k}\epsilon_{1,i_1}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right)\right] \\
&= \mathbb{E}\left[\Phi\left(2\left\|\sum_{i_2}\left(\sum_{i_1,i_3\dots,i_k}\epsilon_{1,i_1}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right)\right\|_\infty\right)\right] \\
&\le \mathbb{E}\left[\Phi\left(4\left\|\sum_{i_2}\epsilon_{2,i_2}\left(\sum_{i_1,i_3\dots,i_k}\epsilon_{1,i_1}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right)\right\|_\infty\right)\right] \\
&= \mathbb{E}\left[\Phi\left(4\left\|\sum_{i_1,\dots,i_k}\epsilon_{1,i_1}\epsilon_{2,i_2}\hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)}\right\|_\infty\right)\right].
\end{aligned}$$

The conclusion of the lemma follows from repeating this procedure. □

We are now in position to prove Theorem 5.

Proof of Theorem 5. In this proof, the notation $\lesssim$ means that the left-hand side is less than the right-hand side up to a constant that depends only on q and K. We may assume without loss of generality $\boldsymbol{e} = (\underbrace{1, \dots, 1}_{k}, 0, \dots, 0)$. In view of Lemma 2, it suffices to show that

$$\mathbb{E}\left[\left\| \sum_{i_1,\dots,i_k} \epsilon_{1,i_1} \cdots \epsilon_{k,i_k} \hat{\boldsymbol{X}}_{(i_1,\dots,i_k,0,\dots,0)} \right\|_\infty^q \right] \lesssim (\log p)^{qk/2} \mathbb{E}\left[\max_{1 \le j \le p} \left(\sum_{i_1,\dots,i_k} |\hat{X}^j_{(i_1,\dots,i_k,0,\dots,0)}|^2 \right)^{q/2} \right].$$

By conditioning and Lemma 2.2.2 in van der Vaart and Wellner (1996), together with the fact that that the L^q-norm is bounded from above by the $\psi_{2/k}$-norm up to some constant that depends only on (q,k), the problem boils down to proving that, for any constants $a_{i_1,\dots,i_k}$,

$$\left\| \sum_{i_1,\dots,i_k} \epsilon_{1,i_1} \cdots \epsilon_{k,i_k} a_{i_1,\dots,i_k} \right\|_{\psi_{2/k}} \lesssim \sqrt{\sum_{i_1,\dots,i_k} a^2_{i_1,\dots,i_k}},$$

but this follows from Corollary 3.2.6 in de la Peña and Giné (1999). Indeed, let

$$(\epsilon'_1, \epsilon'_2, \dots) = (\epsilon_{1,1}, \dots, \epsilon_{1,N_1}, \epsilon_{2,1}, \dots, \epsilon_{K,N_K}),$$

and define correspondingly

$$b_{j_1 \dots j_K} = \begin{cases} a_{i_1 \dots i_K} & \text{if } j_1 = i_1, j_2 = N_1 + i_2, \dots, j_K = \prod_{k=1}^{K-1} N_k + i_K, \\ 0 & \text{otherwise} \end{cases}$$

for $i_k = 1, \dots, N_k, k = 1, \dots, K$. Then,

$$\sum_{i_1,\dots,i_K} \epsilon_{1,i_1} \cdots \epsilon_{K,i_K} a_{i_1 \dots i_K} = \sum_{j_1 < \cdots < j_K} \epsilon'_{j_1} \dots \epsilon'_{j_K} b_{j_1 \dots j_K}.$$

Corollary 3.2.6 in de la Peña and Giné (1999) implies that the $\psi_{2/k}$-norm of the right-hand side is $\lesssim \sqrt{\sum_{j_1 < \cdots < j_K} b^2_{j_1 \dots j_K}} = \sqrt{\sum_{i_1,\dots,i_K} a^2_{i_1 \dots i_K}}$. □

Remark 10 (Comparison with Davezies et al. (2020))**.** Lemma S2 of Davezies et al. (2020) derives a symmetrization inequality for the empirical process of an separately exchangeable array. Their symmetrization inequality is substantially different from the maximal inequalities developed in this section, in the sense that their symmetrization inequality is applied to the whole sample mean and does not lead to correct orders to degenerate components of the Hoeffding decomposition. Indeed, Davezies et al. (2020) do not derive a Hoeffding-type decomposition for separately exchangeable arrays.

Appendix B. Proofs for Section 2

B.1. **Proof of Lemma 1.** The lemma follows from the fact that $\mathbb{E}[\boldsymbol{X_i} \mid (U_{\boldsymbol{i} \odot \boldsymbol{e}})_{\boldsymbol{e} \le \boldsymbol{1}}] = \boldsymbol{X_i}$, so that $\boldsymbol{X_i} = \hat{\boldsymbol{X}}_{\boldsymbol{i}} + \sum_{\boldsymbol{e} \le \boldsymbol{1}, \boldsymbol{e} \ne \boldsymbol{1}} \hat{\boldsymbol{X}}_{\boldsymbol{i} \odot \boldsymbol{e}} = \sum_{\boldsymbol{e} \in \{0,1\}^K \setminus \{\boldsymbol{0}\}} \hat{\boldsymbol{X}}_{\boldsymbol{i} \odot \boldsymbol{e}}$. □

B.2. **Proof of Theorem 1.** We will assume Condition (2.3). The proof under Condition (2.4) is similar and thus omitted. In this proof, let C denote a generic constant that depends only on $\underline{\sigma}$ and K. We divide the proof into two steps.

Step 1. We first prove the following bound for the Hájek projection

$$\sup_{R\in\mathcal{R}} |\mathbb{P}(\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}}^{W} \in R) - \gamma_{\Sigma}(R)| \leq C\left(\frac{D_{\boldsymbol{N}}^2 \log^7(p\overline{N})}{n}\right)^{1/6}.$$

For the notational convenience, we assume $K = 2$; the proof for the general case is completely analogous. Let $\overline{\boldsymbol{W}}_k = N_k^{-1}\sum_{i_k} \boldsymbol{W}_{k,i_k}$. By Proposition 2.1 in Chernozhukov et al. (2017a), we have

$$\sup_{R\in\mathcal{R}} |\mathbb{P}(\sqrt{N_k}\overline{\boldsymbol{W}}_k \in R) - \gamma_{\Sigma_{W_k}}(R)| \leq C\left(\frac{D_{\boldsymbol{N}}^2 \log^7(p\overline{N})}{n}\right)^{1/6}, \ k = 1, 2.$$

For any rectangle $R = \prod_{j=1}^{p}[a_j, b_j]$, vector $\boldsymbol{w} = (w_1, \ldots, w_p)^T \in \mathbb{R}^P$, and scalar $c > 0.$, we use the notation $[cR + \boldsymbol{w}] = \prod_{j=1}^{p}[ca_j + w_j, cb_j + w_j]$, which is still a rectangle. With this in mind, observe that for any rectangle $R \in \mathcal{R}$,

$$\mathbb{P}(\sqrt{n}(\overline{\boldsymbol{W}}_1 + \overline{\boldsymbol{W}}_2) \in R) = \mathbb{E}\left[\mathbb{P}\left(\sqrt{N_1}\overline{\boldsymbol{W}}_1 \in [\sqrt{N_1/n}R - \sqrt{N_1}\overline{\boldsymbol{W}}_2] \mid \overline{\boldsymbol{W}}_2\right)\right]$$

Since $\overline{\boldsymbol{W}}_1$ and $\overline{\boldsymbol{W}}_2$ are independent, the right-hand side is bounded by

$$\mathbb{E}\left[\gamma_{\Sigma_{W_1}}([\sqrt{N_1/n}R - \sqrt{N_1}\overline{\boldsymbol{W}}_2])\right] + C\left(\frac{D_{\boldsymbol{N}}^2 \log^7(p\overline{N})}{n}\right)^{1/6}.$$

For $\boldsymbol{Y}_1 \sim N(\boldsymbol{0}, \Sigma_{W_1})$ independent of $\overline{\boldsymbol{W}}_2$, we have

$$\gamma_{\Sigma_{W_1}}([\sqrt{N_1/n}R - \sqrt{N_1}\overline{\boldsymbol{W}}_2]) = \mathbb{P}(\boldsymbol{Y}_1 \in [\sqrt{N_1/n}R - \sqrt{N_1}\overline{\boldsymbol{W}}_2] \mid \overline{\boldsymbol{W}}_2),$$

so that

$$\begin{aligned}\mathbb{E}\left[\gamma_{\Sigma_{W_1}}([\sqrt{N_1/n}R - \sqrt{N_1}\overline{\boldsymbol{W}}_2])\right] &= \mathbb{P}(\boldsymbol{Y}_1 \in [\sqrt{N_1/n}R - \sqrt{N_1}\overline{\boldsymbol{W}}_2]) \\ &= \mathbb{P}(\sqrt{N_2}\overline{\boldsymbol{W}}_2 \in [\sqrt{N_2/n}R - \sqrt{N_2/N_1}\boldsymbol{Y}_1]) \\ &= \mathbb{E}\left[\mathbb{P}(\sqrt{N_1}\overline{\boldsymbol{W}}_2 \in [\sqrt{N_2/n}R - \sqrt{N_2/N_2}\boldsymbol{Y}_1] \mid \boldsymbol{Y}_1)\right].\end{aligned}$$

Since $\boldsymbol{Y}_1$ and $\overline{\boldsymbol{W}}_2$ are independent, the far right-hand side is bounded by

$$\mathbb{E}\left[\gamma_{\Sigma_{W_2}}([\sqrt{N_2/n}R - \sqrt{N_2/N_2}\boldsymbol{Y}_1])\right] + C\left(\frac{D_{\boldsymbol{N}}^2 \log^7(p\overline{N})}{n}\right)^{1/6}.$$

For $\boldsymbol{Y}_2 \sim N(\boldsymbol{0}, \Sigma_{W_2})$ independent of $\boldsymbol{Y}_1$, the first term can be written as $\mathbb{P}(\sqrt{n/N_1}\boldsymbol{Y}_1 + \sqrt{n/N_2}\boldsymbol{Y}_2 \in R) = \gamma_{\Sigma}(R)$. Conclude that

$$\mathbb{P}(\sqrt{n}(\overline{\boldsymbol{W}}_1 + \overline{\boldsymbol{W}}_2) \in R) \leq \gamma_{\Sigma}(R) + C\left(\frac{D_{\boldsymbol{N}}^2 \log^7(p\overline{N})}{n}\right)^{1/6}.$$

The reverse inequality follows similarly.

Step 2. We will prove the conclusion of the theorem. Recall the decomposition:

$$\boldsymbol{S}_{\boldsymbol{N}} = \boldsymbol{S}_{\boldsymbol{N}}^{W} + \boldsymbol{R}_{\boldsymbol{N}} \quad \text{with} \quad \boldsymbol{R}_{\boldsymbol{N}} = \sum_{k=2}^{K}\sum_{\boldsymbol{e}\in\mathcal{E}_k} \frac{1}{\prod_{k'\in\mathrm{supp}(\boldsymbol{e})} N_{k'}} \sum_{\boldsymbol{i}\in I_{\boldsymbol{e}}([\boldsymbol{N}])} \hat{\boldsymbol{X}}_{\boldsymbol{i}}.$$

By Corollary 3, we have

$$\mathbb{E}[\|\boldsymbol{R}_{\boldsymbol{N}}\|_\infty] \le C \sum_{k=2}^{K} n^{-k/2} (\log p)^{k/2+1} D_{\boldsymbol{N}} \le C n^{-1} D_{\boldsymbol{N}} (\log p)^2.$$

For $R = \prod_{j=1}^{p} [a_j, b_j]$ with $\boldsymbol{a} = (a_1, \dots, a_p)^T$ and $\boldsymbol{b} = (b_1, \dots, b_p)^T$, we have

$$\begin{aligned}
&\mathbb{P}(\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}} \in R) = \mathbb{P}(\{-\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}} \le -\boldsymbol{a}\} \cap \{\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}} \le \boldsymbol{b}\}) \\
&\le \mathbb{P}(\{-\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}} \le -\boldsymbol{a}\} \cap \{\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}} \le \boldsymbol{b}\} \cap \{\|\sqrt{n}\boldsymbol{R}_N\|_\infty \le t\}) + \mathbb{P}(\|\sqrt{n}\boldsymbol{R}_N\|_\infty > t) \\
&\le \mathbb{P}(\{-\sqrt{n}\boldsymbol{S}^W_{\boldsymbol{N}} \le -\boldsymbol{a} - t\} \cap \{\sqrt{n}\boldsymbol{S}^W_{\boldsymbol{N}} \le \boldsymbol{b} + t\}) + C t^{-1} n^{-1/2} D_{\boldsymbol{N}} (\log p)^2 \\
&\le \gamma_\Sigma(\{\boldsymbol{y} \in \mathbb{R}^p : -\boldsymbol{y} \le -\boldsymbol{a} + t, \boldsymbol{y} \le \boldsymbol{b} + t\}) \\
&\qquad + C \left(\frac{D^2_{\boldsymbol{N}} \log^7(p\overline{N})}{n} \right)^{1/6} + C t^{-1} n^{-1/2} D_{\boldsymbol{N}} (\log p)^2 \\
&\le \gamma_\Sigma(R) + C t \sqrt{\log p} + C \left(\frac{D^2_{\boldsymbol{N}} \log^7(p\overline{N})}{n} \right)^{1/6} + C t^{-1} n^{-1/2} D_{\boldsymbol{N}} (\log p)^2,
\end{aligned}$$

where the last line follows from Nazarov's inequality – see Lemma 7 in Appendix F. Choosing $t = n^{-1/4} D^{1/2}_{\boldsymbol{N}} (\log^3 p)^{1/4}$, we have

$$\begin{aligned}
\mathbb{P}(\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}} \in R) &\le \gamma_\Sigma(R) + C \left(\frac{D^2_{\boldsymbol{N}} \log^7(p\overline{N})}{n} \right)^{1/6} + C \left(\frac{D^2_{\boldsymbol{N}} \log^5 p}{n} \right)^{1/4} \\
&\le \gamma_\Sigma(R) + C \left(\frac{D^2_{\boldsymbol{N}} \log^7(p\overline{N})}{n} \right)^{1/6}.
\end{aligned}$$

The reverse inequality follows similarly. □

B.3. **Proof of Theorem 2.** We separately prove the theorem under Cases (i) and (ii).

Case (i). Let C denote a generic constant that depends only on $\underline{\sigma}, K$, and C_1. Also the notation $\lesssim$ means that the left-hand side is bounded by the right-hand side up to a constant that depends only on $\underline{\sigma}, K$, and C_1.

Conditionally on $\boldsymbol{X}_{[\boldsymbol{N}]}$, we have $\sqrt{n}\boldsymbol{S}^{MB}_{\boldsymbol{N}} \sim N(\boldsymbol{0}, \hat{\Sigma})$, where

$$\hat{\Sigma} = \sum_{k=1}^{K} \frac{n}{N_k^2} \sum_{i_k=1}^{N_k} (\overline{\boldsymbol{X}}_{k,i_k} - \boldsymbol{S}_{\boldsymbol{N}})(\overline{\boldsymbol{X}}_{k,i_k} - \boldsymbol{S}_{\boldsymbol{N}})^T.$$

Hence, to obtain a bound on $\sup_{R \in \mathcal{R}} |\mathbb{P}_{|\boldsymbol{X}_{[\boldsymbol{N}]}}(\sqrt{n}\boldsymbol{S}^{MB}_{\boldsymbol{N}} \in R) - \gamma_\Sigma(R)|$, it suffices to bound $\|\hat{\Sigma} - \Sigma\|_\infty$ in view of Lemma 8 in Appendix F. We note that

$$\|\hat{\Sigma} - \Sigma\|_\infty \le \sum_{k=1}^{K} \underbrace{\max_{1 \le j, \ell \le p} \left| \frac{n}{N_k^2} \sum_{i_k=1}^{N_k} \overline{X}^j_{k,i_k} \overline{X}^\ell_{k,i_k} - \frac{n}{N_k} S^j_{\boldsymbol{N}} S^\ell_{\boldsymbol{N}} - \frac{n}{N_k} \mathbb{E}[W^j_{k,1} W^\ell_{k,1}] \right|}_{=:\hat{\Delta}_{W,k}}.$$

We will focus on bounding $\hat{\Delta}_{W,1}$ as similar bounds hold for $\hat{\Delta}_{W,k}$ with $k \in \{2, \dots, K\}$.

Observe that

$$\frac{n}{N_1^2}\sum_{i_1=1}^{N_1}\overline{X}_{1,i_1}^{j}\overline{X}_{1,i_1}^{\ell} = \frac{n}{N_1^2}\sum_{i_1=1}^{N_1}(\overline{X}_{1,i_1}^{j}-W_{1,i_1}^{j})(\overline{X}_{1,i_1}^{\ell}-W_{1,i_1}^{\ell}) + \frac{n}{N_1^2}\sum_{i_1=1}^{N_1}(\overline{X}_{1,i_1}^{j}-W_{1,i_1}^{j})W_{1,i_1}^{\ell}$$
$$+\frac{n}{N_1^2}\sum_{i_1=1}^{N_1}W_{1,i_1}^{j}(\overline{X}_{1,i_1}^{\ell}-W_{1,i_1}^{\ell}) + \frac{n}{N_1^2}\sum_{i_1=1}^{N_1}W_{1,i_1}^{j}W_{1,i_1}^{\ell}.$$

By the Cauchy-Schwarz inequality and the definition of n, we obtain

$$\begin{aligned}\hat{\Delta}_{W,1} \le{}& \underbrace{\max_{1\le\ell\le p}\frac{1}{N_1}\sum_{i_1=1}^{N_1}(\overline{X}_{1,i_1}^{\ell}-W_{1,i_1}^{\ell})^2}_{=:\hat{\Delta}_{W,1,1}} + 2\hat{\Delta}_{W,1,1}^{1/2}\sqrt{\max_{1\le\ell\le p}\frac{1}{N_1}\sum_{i_1=1}^{N_1}|W_{1,i_1}^{\ell}|^2} \\ &+ \underbrace{\max_{1\le j,\ell\le p}\left|\frac{1}{N_1}\sum_{i_1}^{N_1}(W_{1,i_1}^{j}W_{1,i_1}^{\ell}-\mathbb{E}[W_{1,1}^{j}W_{1,1}^{\ell}])\right|}_{=:\hat{\Delta}_{W,1,2}} + \max_{1\le\ell\le p}|S_{\boldsymbol{N}}^{\ell}|^2.\end{aligned} \tag{B.1}$$

For the second term on the right-hand side, we have

$$\frac{1}{N_1}\sum_{i_1=1}^{N_1}|W_{1,i_1}^{\ell}|^2 \le \mathbb{E}[|W_{1,i_1}^{\ell}|^2] + \frac{1}{N_1}\sum_{i_1=1}^{N_1}(|W_{1,i_1}^{\ell}|^2-\mathbb{E}[|W_{1,i_1}^{\ell}|^2]) \le \overline{\sigma}^2 + \hat{\Delta}_{W,1,2}. \tag{B.2}$$

Further, since $S_{\boldsymbol{N}}^{\ell} = N_1^{-1}\sum_{i_1=1}^{N_1}(\overline{X}_{1,i_1}^{\ell}-W_{1,i_1}^{\ell}) + N_1^{-1}\sum_{i_1=1}^{N_1}W_{1,i_1}^{\ell}$, we have

$$\max_{1\le\ell\le p}|S_{\boldsymbol{N}}^{\ell}|^2 \lesssim \hat{\Delta}_{W,1,1} + \hat{\Delta}_{W,1,3}^2, \tag{B.3}$$

where $\hat{\Delta}_{W,1,3} = \max_{1\le\ell\le p}|N_1^{-1}\sum_{i_1=1}^{N_1}W_{1,i_1}^{\ell}|$. Combining (B.1)–(B.3), we have

$$\hat{\Delta}_{W,1} \lesssim \hat{\Delta}_{W,1,1} + \overline{\sigma}\hat{\Delta}_{W,1,1}^{1/2} + \hat{\Delta}_{W,1,2} + \hat{\Delta}_{W,1,3}^2.$$

It remains to find bounds on the four terms on the right-hand side.

First, by Condition (2.7), we have $\hat{\Delta}_{W,1,1}\log^4 p \le C_1 n^{-\zeta_1}$ and $\overline{\sigma}\hat{\Delta}_{W,1,1}^{1/2}\log^2 p \le Cn^{-\zeta_2/2}$ with probability at least $1-Cn^{-1}$. Second, we note that

$$\mathbb{E}\left[\max_{1\le i_1\le N_1}\max_{1\le\ell\le p}|W_{1,i_1}^{\ell}|^4\right] \lesssim (\log p\overline{N})^4\max_{1\le\ell\le p}\||W_{1,1}^{\ell}|^4\|_{\psi_{1/4}} = (\log p\overline{N})^4\underbrace{\max_{1\le\ell\le p}\|W_{1,1}^{\ell}\|_{\psi_1}^4}_{\le D_{\boldsymbol{N}}^4}.$$

Applying Lemma 8 in Chernozhukov et al. (2015), we have

$$\begin{aligned}\mathbb{E}[\hat{\Delta}_{W,1,2}] &\lesssim N_1^{-1}\sqrt{(\log p)\max_{1\le j,\ell\le p}\sum_{i_1=1}^{N_1}\mathbb{E}[|W_{1,i_1}^{j}W_{1,i_1}^{\ell}|^2]} + N_1^{-1}\sqrt{\mathbb{E}\left[\max_{1\le i_1\le N_1}\max_{1\le\ell\le p}|W_{1,i_1}^{\ell}|^4\right]}\log p \\ &\lesssim N_1^{-1/2}D_{\boldsymbol{N}}\log^{1/2}p + N_1^{-1}D_{\boldsymbol{N}}^2(\log p)\log^2(p\overline{N}) \\ &\lesssim n^{-1/2}D_{\boldsymbol{N}}\log^{1/2}p + n^{-1}D_{\boldsymbol{N}}^2\log^3(p\overline{N}).\end{aligned}$$

Now, applying Lemma E.2 in Chernozhukov et al. (2017a) with $\eta = 1$ and $\beta = 1/2$, together with the fact that

$$\left\| \max_{1 \le i_1 \le N_1} \max_{1 \le \ell \le p} |W^{\ell}_{1,i_1}|^2 \right\|_{\psi_{1/2}} = \left\| \max_{1 \le i_1 \le N_1} \max_{1 \le \ell \le p} |W^{\ell}_{1,i_1}| \right\|_{\psi_1}^2 \lesssim (\log pN_1)^2 D_{\boldsymbol{N}}^2,$$

we have

$$\mathbb{P}\left(\hat{\Delta}_{W,1,2} \ge 2\mathbb{E}[\hat{\Delta}_{W,1,2}] + t \right) \le \exp\left(-\frac{nt^2}{3D_{\boldsymbol{N}}^2} \right) + 3\exp\left\{ -\left(\frac{nt}{CD_{\boldsymbol{N}}^2 \log^2(p\overline{N})} \right)^{1/2} \right\}.$$

Setting $t = \{Cn^{-1}D_{\boldsymbol{N}}^2 \log n\}^{1/2} \vee \{Cn^{-1}D_{\boldsymbol{N}}^2 (\log^2 n)\log^3(p\overline{N})\}$, we conclude that

$$\mathbb{P}\left(\hat{\Delta}_{W,1,2} \ge C\{(n^{-1}D_{\boldsymbol{N}}^2 \log^{1/2}(pn) + n^{-1}D_{\boldsymbol{N}}^2 (\log n)^2 \log^3(p\overline{N})\} \right) \le Cn^{-1}.$$

Condition (2.8) then guarantees that $\hat{\Delta}_{W,1,2}\log^2 p \le Cn^{-\zeta_1/2}$ with probability at least $1 - Cn^{-1}$.

Finally, since $\overline{\sigma}^2 \le (\max_{1\le\ell\le p}\mathbb{E}[|W^{\ell}_{1,1}|^3])^{2/3} \le 1 + \max_{1\le\ell\le p}\mathbb{E}[|W^{\ell}_{1,1}|^3] \lesssim D_{\boldsymbol{N}}$, using Lemma 8 in Chernozhukov et al. (2015), we have

$$\mathbb{E}[\hat{\Delta}_{W,1,3}] \lesssim (n^{-1}D_{\boldsymbol{N}}\log p)^{1/2} + n^{-1}D_{\boldsymbol{N}}\log(p\overline{N}).$$

Applying Lemma E.2 in Chernozhukov et al. (2017a) with $\eta = 1$ and $\beta = 1$, we have

$$\hat{\Delta}_{W,1,3}^2 \log^2 p \le \underbrace{C\{n^{-1}D_{\boldsymbol{N}}(\log^2 p)\log(pn) + n^{-2}D_{\boldsymbol{N}}^2(\log^2 n)(\log^2 p)\log^2(p\overline{N})\}}_{\le Cn^{-\zeta_1}}$$

with probability at least $1 - Cn^{-1}$. Conclude that $\hat{\Delta}_{W,1}\log^2 p \le Cn^{-(\zeta_1\wedge\zeta_2)/2}$ with probability at least $1 - Cn^{-1}$. The desired result then follows from Lemma 8 in Appendix F.

Case (ii). The proof is similar to the previous case. We only point out required modifications. Let C denote a generic constant that depends only on $q, \underline{\sigma}, K$, and C_1. The similar modification applies to $\lesssim$. In view of the previous case, we only have to find bounds on $\hat{\Delta}_{W,1,2}$ and $\hat{\Delta}_{W,1,3}$.

Applying Lemma 8 in Chernozhukov et al. (2015), we have

$$\begin{aligned} \mathbb{E}[\hat{\Delta}_{W,1,2}] &\lesssim N_1^{-1}\sqrt{(\log p)\max_{1\le j,\ell\le p}\sum_{i_1=1}^{N_1}\mathbb{E}[|W^j_{1,i_1}W^{\ell}_{1,i_1}|^2]} + N_1^{-1}\sqrt{\mathbb{E}\left[\max_{1\le i_1\le N_1}\max_{1\le\ell\le p}|W^{\ell}_{1,i_1}|^4\right]}\log p \\ &\lesssim N_1^{-1/2}D_{\boldsymbol{N}}\log^{1/2} p + N_1^{-1+2/q}D_{\boldsymbol{N}}^2\log p \\ &\lesssim n^{-1/2}D_{\boldsymbol{N}}\log^{1/2} p + n^{-1+2/q}D_{\boldsymbol{N}}^2\log p. \end{aligned}$$

Applying the Fuk-Nagaev inequality (Lemma E.2 in Chernozhukov et al. (2017a)) with $s = q/2$, we have

$$\begin{aligned} \mathbb{P}\left(\hat{\Delta}_{W,1,2} \ge 2\mathbb{E}[\hat{\Delta}_{W,1,2}] + t \right) &\le \exp\left(-\frac{N_1 t^2}{3D_{\boldsymbol{N}}^2} \right) + \frac{CN_1 D_{\boldsymbol{N}}^q}{N_1^{q/2}t^{q/2}} \\ &\le \exp\left(-\frac{nt^2}{3D_{\boldsymbol{N}}^2} \right) + \frac{CD_{\boldsymbol{N}}^q}{n^{q/2-1}t^{q/2}}. \end{aligned}$$

Setting $t = (Cn^{-1}D_{\boldsymbol{N}}^2\log n)^{1/2}\bigvee(Cn^{-1+4/q}D_{\boldsymbol{N}}^2)$, we have

$$\mathbb{P}\left(\hat{\Delta}_{W,1,2} \ge C\{(n^{-1}D_{\boldsymbol{N}}^2\log(pn))^{1/2} + n^{-1+4/q}D_{\boldsymbol{N}}^2\log p\} \right) \le Cn^{-1}.$$

Condition (2.9) then guarantees that $\hat{\Delta}_{W,1,2}\log^2 p \le Cn^{-\zeta_1/2}$ with probability at least $1-Cn^{-1}$. A bound for $\hat{\Delta}_{W,1,3}$ can be obtained similarly. Using Lemma 8 in Chernozhukov et al. (2015), we have

$$\mathbb{E}[\hat{\Delta}_{W,1,3}] \lesssim (n^{-1}D_{\boldsymbol{N}}\log p)^{1/2} + n^{-1+1/q}D_{\boldsymbol{N}}\log p.$$

Applying Lemma E.2 in Chernozhukov et al. (2017a) with $s=q$, we have

$$\mathbb{P}\left(\hat{\Delta}_{W,1,3} \ge 2\mathbb{E}[\hat{\Delta}_{W,1,3}] + t\right) \le \exp\left(-\frac{nt^2}{3D_{\boldsymbol{N}}}\right) + \frac{CD^q_{\boldsymbol{N}}}{n^{q-1}t^q}.$$

Setting $t = (Cn^{-1}D_{\boldsymbol{N}}\log n)^{1/2}\bigvee(Cn^{-1+2/q}D_{\boldsymbol{N}})$, we conclude that

$$\hat{\Delta}^2_{W,1,3}\log^2 p \le \underbrace{C\{n^{-1}D_{\boldsymbol{N}}(\log^2 p)\log(pn) + n^{-2+4/q}\log^4 p\}}_{\le Cn^{-\zeta_1}}$$

with probability at least $1-Cn^{-1}$. □

B.4. **Proof of Proposition 1.** We separately prove the theorem under Cases (i') and (ii').

Case (i'). Let the notation $\lesssim$ mean that the left-hand side is bounded by the right-hand side up to a constant that depends only on ν, K, and C_1. We will show that $\mathbb{P}(\overline{\sigma}^2\hat{\Delta}_{W,1,1}\log^4 p > n^{-\zeta+1/\nu}) \lesssim n^{-1}$, where

$$\hat{\Delta}_{W,1,1} = \max_{1\le\ell\le p}\frac{1}{N_1}\sum_{i_1=1}^{N_1}(\overline{X}^{\ell}_{1,i_1} - W^{\ell}_{1,i_1})^2.$$

Similar bounds hold for $\max_{1\le\ell\le p} N_k^{-1}\sum_{i_k=1}^{N_k}(\overline{X}^{\ell}_{k,i_k} - W^{\ell}_{k,i_k})^2$ with $k\in\{2,\dots,K\}$.

We first note that

$$\hat{\Delta}_{W,1,1} = \max_{1\le\ell\le p}\frac{1}{N_1}\sum_{i_1=1}^{N_1}(\overline{X}^{\ell}_{1,i_1} - W^{\ell}_{1,i_1})^2 \le \frac{1}{N_1}\sum_{i_1=1}^{N_1}\|\overline{\boldsymbol{X}}_{1,i_1} - \boldsymbol{W}_{1,i_1}\|^2_\infty.$$

Pick any $i_1\in\mathbb{N}$. For each $\boldsymbol{i}_{-1} = (i_2,\dots,i_K)\in\mathbb{N}^{K-1}$ and $\boldsymbol{e}\in\{0,1\}^{K-1}$, define the vector

$$V_{\boldsymbol{i}_{-1}\odot\boldsymbol{e}} = (U_{(0,\boldsymbol{i}_{-1}\odot\boldsymbol{e})}, U_{(i_1,\boldsymbol{i}_{-1}\odot\boldsymbol{e})}).$$

With this notation, we can rewrite $\boldsymbol{X}_{\boldsymbol{i}}$ with $\boldsymbol{i} = (i_1,\boldsymbol{i}_{-1})$ as

$$\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{g}\big(U_{(i_1,0,\dots,0)}, (V_{\boldsymbol{i}_{-1}\odot\boldsymbol{e}})_{\boldsymbol{e}\in\{0,1\}^{K-1}\setminus\{\boldsymbol{0}\}}\big).$$

From this expression, we see that, conditionally on $U_{(i_1,0,\dots,0)}$, the $(K-1)$-array $(\boldsymbol{X}_{(i_1,\boldsymbol{i}_{-1})})_{\boldsymbol{i}_{-1}\in\mathbb{N}^{K-1}}$ is separately exchangeable with mean vector $\boldsymbol{W}_{1,i_1}$ generated by $\{V_{\boldsymbol{i}_{-1}\odot\boldsymbol{e}} : \boldsymbol{i}_{K-1}\in\mathbb{N}^{K-1}, \boldsymbol{e}\in\{0,1\}^{K-1}\setminus\{\boldsymbol{0}\}\}$. Applying Corollary 3 conditionally on $U_{(i_1,0,\dots,0)}$ (the fact that $U_{\boldsymbol{i}\odot\boldsymbol{e}}$ are uniform on $[0,1]$ is not crucial in the proof of Corollary 3) combined with Jensen's inequality, we have

$$\mathbb{E}[\|\overline{\boldsymbol{X}}_{1,i_1} - \boldsymbol{W}_{1,i_1}\|^{2\nu}_\infty \mid U_{(i_1,0\dots,0)}] \lesssim \underbrace{\left(\sum_{k=1}^{K-1} n^{-k/2}(\log p)^{k/2}\right)^{2\nu}}_{\lesssim(n^{-1}\log p)^\nu}\mathbb{E}[\|\boldsymbol{X}_{(i_1,1,\dots,1)}\|^{2\nu}_\infty \mid U_{(i_1,\dots,0)}],$$

so that by Fubini's theorem

$$\mathbb{E}[\|\overline{\boldsymbol{X}}_{1,i_1} - \boldsymbol{W}_{1,i_1}\|^{2\nu}_\infty] \lesssim (n^{-1}\log p)^\nu\mathbb{E}[\|\boldsymbol{X}_{(i_1,1,\dots,1)}\|^{2\nu}_\infty] \lesssim (n^{-1}D^2_{\boldsymbol{N}}\log^3 p)^\nu.$$

This implies that $\mathbb{E}[(\overline{\sigma}^2 \hat{\Delta}_{W,1,1} \log^4 p)^{\nu}] \lesssim n^{-\zeta\nu}$ under our assumption. By Markov's inequality, we conclude that

$$\mathbb{P}\left(\overline{\sigma}^2 \hat{\Delta}_{W,1,1} \log^4 p > n^{-\zeta+1/\nu}\right) \lesssim n^{-1}.$$

This completes the proof.

Case (ii'). The proof is similar to the previous case. We only point out required modifications. Set $\nu = q/2$ in the previous case.Under Case (ii'), we have

$$\mathbb{E}[\|\overline{\boldsymbol{X}}_{1,i_1} - \boldsymbol{W}_{1,i_1}\|_\infty^q] \lesssim (n^{-1}\log p)^{\nu} \underbrace{\mathbb{E}[\|\boldsymbol{X}_{(i_1,1,\dots,1)}\|_\infty^q]}_{\le D_{\boldsymbol{N}}^q},$$

which implies that $\mathbb{E}[(\overline{\sigma}^2 \hat{\Delta}_{W,1,1} \log^4 p)^{q/2}] \lesssim n^{-\zeta q/2}$. Markov's inequality yields the desired result. □

B.5. **Proof of Corollary 1.** We only prove the corollary under Case (i). The proof for Case (ii) is similar. Let C denote a generic constant that depends only on $\underline{\sigma}, K$, and C_1. We first note that from the proof of Theorem 2, we have

$$\max_{1 \le j \le p} \left|\sigma_j^2/\hat{\sigma}_j^2 - 1\right| \le C n^{-(\zeta_1 \wedge \zeta_2)/2}/\log^2 p$$

with probability at least $1 - Cn^{-1}$. By Theorem 1, we have

$$\sup_{R \in \mathcal{R}} \left|\mathbb{P}(\sqrt{n}\Lambda^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \in R) - \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R)\right| \le C n^{-(\zeta_1 \wedge \zeta_2)/4}.$$

By the Borell-Sudakov-Tsirel'son inequality and the fact $\mathbb{E}[\|\Lambda^{-1/2}\boldsymbol{Y}\|_\infty] \le C\sqrt{\log p}$, which is implied by the Gaussianity of $\Lambda^{-1/2}\boldsymbol{Y}$, we have

$$\mathbb{P}\left(\|\Lambda^{-1/2}\boldsymbol{Y}\|_\infty > C\sqrt{\log(pn)}\right) \le n^{-1},$$

Combining the high-dimensional CLT, we see that

$$\mathbb{P}\left(\|\sqrt{n}\Lambda^{-1/2}\boldsymbol{S}_{\boldsymbol{N}}\|_\infty > C\sqrt{\log(pn)}\right) \le C n^{-(\zeta_1 \wedge \zeta_2)/4}.$$

Since $\frac{n^{-(\zeta_1 \wedge \zeta_2)/2}}{\log^2 p} \times \sqrt{\log(pn)} \le \frac{Cn^{-(\zeta_1 \wedge \zeta_2)/4}}{\log^{3/2} p}$, we have

$$\mathbb{P}\left(\|\sqrt{n}(\hat{\Lambda}^{-1/2} - \Lambda^{-1/2})\boldsymbol{S}_{\boldsymbol{N}}\|_\infty > t_n\right) \le C n^{-(\zeta_1 \wedge \zeta_2)/4}.$$

with $t_n = \frac{Cn^{-(\zeta_1 \wedge \zeta_2)/4}}{\log^{3/2} p}$.

Now, for $R = \prod_{j=1}^p [a_j, b_j]$ with $\boldsymbol{a} = (a_1, \dots, a_p)^T$ and $\boldsymbol{b} = (b_1, \dots, b_p)^T$, we have

$$\begin{aligned}
\mathbb{P}\left(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \in R\right) \le& \mathbb{P}\left(\{-\sqrt{n}\Lambda^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \le -\boldsymbol{a} + t_n\} \cap \{\sqrt{n}\Lambda^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \le \boldsymbol{b} + t_n\}\right) \\
&+ \mathbb{P}\left(\|\sqrt{n}(\hat{\Lambda}^{-1/2} - \Lambda^{-1/2})\boldsymbol{S}_{\boldsymbol{N}}\|_\infty > t_n\right) \\
\le& \mathbb{P}\left(\{-\Lambda^{-1/2}\boldsymbol{Y} \le -\boldsymbol{a} + t_n\} \cap \{\Lambda^{-1/2}\boldsymbol{Y} \le \boldsymbol{b} + t_n\}\right) + C n^{-(\zeta_1 \wedge \zeta_2)/4} \\
\le& \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) + C n^{-(\zeta_1 \wedge \zeta_2)/4},
\end{aligned}$$

where the last inequality follows from Lemma 7 together with the fact that

$$t_n\sqrt{\log p} \le C n^{-(\zeta_1 \wedge \zeta_2)/4}/\log p \le C n^{-(\zeta_1 \wedge \zeta_2)/4}.$$

Thus, we have

$$\mathbb{P}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \in R) \le \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) + Cn^{-(\zeta_1 \wedge \zeta_2)/4}.$$

Likewise, we have

$$\mathbb{P}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \in R) \ge \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) - Cn^{-(\zeta_1 \wedge \zeta_2)/4}.$$

Conclude that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}_{\boldsymbol{N}} \in R) - \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) \right| \le Cn^{-(\zeta_1 \wedge \zeta_2)/4}.$$

Similarly, using Theorem 2 and following similar arguments, we conclude that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\boldsymbol{X}_{[\boldsymbol{N}]}}(\sqrt{n}\hat{\Lambda}^{-1/2}\boldsymbol{S}^{MB}_{\boldsymbol{N}} \in R) - \mathbb{P}(\Lambda^{-1/2}\boldsymbol{Y} \in R) \right| \le Cn^{-(\zeta_1 \wedge \zeta_2)/4}$$

with probability at least $1 - Cn^{-1}$. □

Appendix C. Maximal Inequalities for Polyadic Data

In this section, we shall develop maximal inequalities for jointly exchangeable arrays. As in Section 3, let $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ be a K-array consisting of random vectors in $\mathbb{R}^p$ with mean zero generated by the structure (3.1), i.e., $\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{g}((U_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \{0,1\}^K \setminus \{\boldsymbol{0}\}})$. We will follow the notations used in Section 3. Recall that $I_{n,K} = \{(i_1, \dots, i_K) : 1 \le i_1, \dots, i_K \le n \text{ and } i_1, \dots, i_K \text{ are distinct}\}$.

We first point out that when analyzing the sample mean $\boldsymbol{S}_n$, it is without loss of generality to assume that $\boldsymbol{X}_{\boldsymbol{i}}$ is symmetric in the components of $\boldsymbol{i}$, i.e.,

$$\boldsymbol{X}_{(i_1,\dots,i_K)} = \boldsymbol{X}_{(i'_1,\dots,i'_K)} \tag{C.1}$$

for any permutation $(i'_1, \dots, i'_K)$ of $(i_1, \dots, i_K)$. This is because even if $\boldsymbol{X}_{\boldsymbol{i}}$ is not symmetric in the components of $\boldsymbol{i}$, we can instead work with its symmetrized version

$$\check{\boldsymbol{X}}_{(i_1,\dots,i_K)} = \frac{1}{K!} \sum_{(i'_1,\dots,i'_K)} \boldsymbol{X}_{(i'_1,\dots,i'_K)},$$

where the summation is taken over all permutations of $(i_1, \dots, i_K)$. It is not difficult to see that the array $(\check{\boldsymbol{X}}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{\infty,K}}$ continues to be jointly exchangeable and satisfies that

$$\boldsymbol{S}_n = \frac{(n-K)!}{n!} \sum_{\boldsymbol{i} \in I_{n,K}} \check{\boldsymbol{X}}_{\boldsymbol{i}} = \binom{n}{K}^{-1} \sum_{1 \le i_1 < \dots < i_K \le n} \check{\boldsymbol{X}}_{\boldsymbol{i}}.$$

Henceforth, in this section, we will maintain Condition (C.1).

In the decomposition (3.2), the second term on the right-hand side

$$\mathbb{U}_n = \binom{n}{K}^{-1} \sum_{1 \le i_1 < \dots < i_K \le n} \left(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{i_1}, \dots, U_{i_K}] - \sum_{k=1}^{K} \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_{i_k}] \right)$$

is a degenerate U-statistic (with a symmetric kernel) of degree K. Indeed, if we define $\mathfrak{t}(u_1, \dots, u_K) = \mathbb{E}[\boldsymbol{X}_{(1,\dots,K)} \mid U_1 = u_1, \dots, U_K = u_K] - \sum_{k=1}^{K} \mathbb{E}[\boldsymbol{X}_{(1,\dots,K)} \mid U_k = u_k]$, then $\mathfrak{t}$ is symmetric and

$$\mathbb{U}_n = \binom{n}{K}^{-1} \sum_{1 \le i_1 < \dots < i_K \le n} \mathfrak{t}(U_{i_1}, \dots, U_{i_K}).$$

The kernel $\mathfrak{t}$ is degenerate as

$$\mathbb{E}[\mathfrak{t}(u, U_2, \dots, U_K)] = \mathbb{E}[\boldsymbol{X}_{(1....,K)} \mid U_1 = u] - \mathbb{E}[\boldsymbol{X}_{(1,...,K)} \mid U_1 = u] = 0.$$

Applying Corollary 5.6 in Chen and Kato (2020), we obtain the following lemma.

Lemma 4. *For any $q \in [1, \infty)$, we have*

$$(\mathbb{E}[\|\mathbb{U}_n\|_\infty^q])^{1/q} \le C \sum_{k=2}^{K} n^{-k/2} (\log p)^{k/2} (\mathbb{E}[\|\boldsymbol{X}_{(1,\dots,K)}\|_\infty^{q\vee 2}])^{1/(q\vee 2)},$$

where C is a constant that depends only on q and K.

We turn to the analysis of the third term on the right-hand side of (3.2)

$$\sum_{k=2}^{K} \frac{(n-K)!}{n!} \sum_{\boldsymbol{i} \in I_{n,K}} \Big(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k}\mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k-1}\mathcal{E}_r}] \Big)$$
$$= \sum_{k=2}^{K} \binom{n}{K}^{-1} \sum_{1\le i_1<\dots<i_K\le n} \Big(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k}\mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k-1}\mathcal{E}_r}] \Big)$$

where the quality follows from Condition (C.1).

Lemma 5. *For any $k = 2, \dots, K$ and $q \in [1, \infty)$, we have*

$$\left(\mathbb{E}\left[\left\| \binom{n}{K}^{-1} \sum_{1\le i_1<\dots<i_K\le n} \Big(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k}\mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k-1}\mathcal{E}_r}] \Big) \right\|_\infty^q \right] \right)^{1/q}$$
$$\le C n^{-k/2} (\log p)^{1/2} (\mathbb{E}[\|\boldsymbol{X}_{(1,\dots,K)}\|_\infty^{q\vee 2}]^{1/(q\vee 2)},$$

where C is a constant that depends only on q and K.

Before the formal proof of Lemma 5, which is somewhat involved, we shall look at the case with $k = K = 2$ to understand the bound. If $k = K = 2$, then the term in question is

$$\binom{n}{2}^{-1} \sum_{1\le i<j\le n} (\mathbb{E}[\boldsymbol{X}_{(i,j)} \mid U_i, U_j, U_{\{i,j\}}] - \mathbb{E}[\boldsymbol{X}_{(i,j)} \mid U_i, U_j]).$$

Conditionally on U_i's, this is the sum of independent random vectors with mean zero, so the bound in the lemma can be deduced from applying the symmetrization inequality (van der Vaart and Wellner, 1996, Lemma 2.3.6) conditionally on U_i's and then Lemma 2.2.2 in van der Vaart and Wellner (1996) to the weighted sum of Rademacher variables conditionally on all U-variables. The general case is more involved and we will apply the symmetrization inequality for U-statistics with index-dependent kernels; cf. Theorem 3.5.3 in de la Peña and Giné (1999) and the remark after the theorem.

Proof of Lemma 5. In this proof, the notation $\lesssim$ means that the left-hand side is bounded by the right-hand side up to a constant that depends only on q and K. Fix any $k = 2, \dots, K$. Conditionally on $\mathcal{U}_{k-1} = \{U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+} : \boldsymbol{e} \in \cup_{r=1}^{k-1}\mathcal{E}_r, \boldsymbol{i} \in I_{\infty,K}\}$, the component

$$\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k}\mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k-1}\mathcal{E}_r}]$$

is a function of $(U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\mathcal{E}_k}$ with mean zero

$$\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k}\mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k-1}\mathcal{E}_r}] = \mathfrak{h}_{(\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\in\mathcal{E}_k}}((U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\mathcal{E}_k}).$$

The function $\mathfrak{h}_{(\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\in\mathcal{E}_k}}$ implicitly depends on $(U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k-1}\mathcal{E}_r}$, so that it is indexed by $(\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\in\mathcal{E}_k}$ (the vector $(\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\cup_{r=1}^{k-1}\mathcal{E}_r}$ is uniquely determined by $(\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\in\mathcal{E}_k}$ so it is enough to index the function by $(\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\in\mathcal{E}_k}$). Define

$$\mathcal{J}_{n,k} = \{(\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\in\mathcal{E}_k} : 1 \le i_1 < \cdots < i_K \le n\}.$$

This is a collection of vectors of sets where each vector contains $m_k = \binom{K}{k}$ sets. We denote a generic element of $\mathcal{J}_{n,k}$ by $J = (J_1, \dots, J_{m_k})$ by ordering the elements of $\mathcal{E}_k$. We will also write $U_J = (U_{J_1}, \dots, U_{J_{m_k}})$. Then we arrive at the expression

$$\sum_{1\le i_1<\cdots<i_K\le n} \Big(\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k}\mathcal{E}_r}] - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid (U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^{k-1}\mathcal{E}_r}]\Big) = \sum_{J\in\mathcal{J}_{n,k}} \mathfrak{h}_J(U_J).$$

We will apply Theorem 3.5.3 in de la Peña and Giné (1999) to bound the q-th moment of the ℓ^∞-norm of the right-hand side. Let $\{\epsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+} : \boldsymbol{e}\in\mathcal{E}_k, 1 \le i_1 < \cdots < i_K \le n\}$ be independent Rademacher random variables independent of everything else. We first note that conditionally on $\mathcal{U}_{k-1}$, $\sum_{J\in\mathcal{J}_{n,k}} \mathfrak{h}_J(U_J)$ can be seen as a U-statistic with index-dependent kernels by adding zero kernels. In view of Remark 3.5.4 ii) in de la Peña and Giné (1999), to apply their Theorem 3.5.3, we need to verify that $\mathfrak{h}_J = \mathfrak{h}_{(J_1,\dots,J_{m_k})}$ is symmetric in the sense that

$$\mathfrak{h}_{(J_1,\dots,J_{m_k})}(u_{J_1},\dots,u_{J_{m_k}}) = \mathfrak{h}_{(J_1',\dots,J_{m_k}')}(u_{J_1'},\dots,u_{J_{m_k}'})$$

for any permutation $(J_1',\dots,J_{m_k}')$ of $(J_1,\dots,J_{m_k})$. But this follows from the definition of $\mathfrak{h}_J$ and Condition (C.1). Now, applying Theorem 3.5.3 in de la Peña and Giné (1999), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{J\in\mathcal{J}_{n,k}} \mathfrak{h}_J(U_J)\right\|_\infty^q \mid \mathcal{U}_{k-1}\right] &\lesssim \mathbb{E}\left[\left\|\sum_{J\in\mathcal{J}_{n,k}} \epsilon_{J_1}\mathfrak{h}_J(U_J)\right\|_\infty^q \mid \mathcal{U}_{k-1}\right] \\ &= \mathbb{E}\left[\left\|\sum_{J_1}\epsilon_{J_1}\left(\sum_{J_2,\dots,J_{m_k}} \mathfrak{h}_J(U_J)\right)\right\|_\infty^q \mid \mathcal{U}_{k-1}\right].\end{aligned}$$

Here the summation $\sum_{J_1}\sum_{J_2,\dots,J_{m_k}}$ is understood as

$$\sum_{\substack{J_1:\exists(J_2,\dots,J_{m_k}) \\ \text{such that } (J_1,J_2,\dots,J_{m_k})\in\mathcal{J}_{n,k}}} \quad \sum_{(J_2,\dots,J_{m_k}):(J_1,J_2,\dots,J_{m_k})\in\mathcal{J}_{n,k}}.$$

Conditioning on U_J's and applying Lemma 2.2.2 in van der Vaart and Wellner (1996), we have

$$\mathbb{E}\left[\left\|\sum_{J_1}\epsilon_{J_1}\left(\sum_{J_2,\dots,J_{m_k}} \mathfrak{h}_J(U_J)\right)\right\|_\infty^q \mid \mathcal{U}_{k-1}\right] \lesssim (\log p)^{q/2}\mathbb{E}\left[\left(\sum_{J_1}\left\|\sum_{J_2,\dots,J_{m_k}} \mathfrak{h}_J(U_J)\right\|_\infty^2\right)^{q/2} \mid \mathcal{U}_{k-1}\right].$$

Observe that given J_1, the number of $(J_2,\dots,J_{m_k})$ such that $(J_1,J_2,\dots,J_{m_k})\in\mathcal{J}_{n,k}$ is

$$\binom{n-k}{K-k} = O(n^{K-k}).$$

To see this, observe that $J = (J_1,\dots,J_{m_k}) \in \mathcal{J}_{n,k}$ is of the form $J = (\{\boldsymbol{i}\odot\boldsymbol{e}\}^+)_{\boldsymbol{e}\in\mathcal{E}_k}$ for some $(i_1,\dots,i_K)$ such that $1 \le i_1 < \cdots < i_K \le n$. Fixing J_1 corresponds to fixing k elements of $i_1,\dots,i_K$, so the number of possible $(J_2,\dots,J_{m_k})$ coincides with the number of ways to choose remaining $K-k$ elements from $n-k$ integers.

Thus, by the Cauchy-Schwarz inequality, we have

$$\sum_{J_1} \left\| \sum_{J_2,\dots,J_{m_k}} \mathfrak{h}_J(U_J) \right\|_\infty^2 \lesssim n^{K-k} \sum_J \|\mathfrak{h}_J(U_J)\|_\infty^2.$$

Combining Fubini and the fact that the size of $\mathcal{J}_{n,k}$ is $\binom{n}{K} = O(n^K)$, we have

$$\mathbb{E}\left[\left\| \sum_{J \in \mathcal{J}_{n,k}} \mathfrak{h}_J(U_J) \right\|_\infty^q \right] \lesssim n^{(K-k/2)q} (\log p)^{q/2} \mathbb{E}\left[\left(|\mathcal{J}_{n,k}|^{-1} \sum_J \|\mathfrak{h}_J(U_J)\|_\infty^2 \right)^{q/2} \right].$$

Using Jensen's inequality and the definition of $\mathfrak{h}_J$, we conclude that

$$\left(\mathbb{E}\left[\left\| \sum_{J \in \mathcal{J}_{n,k}} \mathfrak{h}_J(U_J) \right\|_\infty^q \right] \right)^{1/q} \lesssim n^{K-k/2} (\log p)^{1/2} (\mathbb{E}[\|\boldsymbol{X}_{(1,\dots,K)}\|_\infty^{q\vee 2}])^{1/(q\vee 2)}.$$

This completes the proof. □

Remark 11 (Comparison with Davezies et al. (2020))**.** Lemma A.1 in Davezies et al. (2020) derives a symmetrization inequality for the empirical process of a jointly exchangeable array. Essentially, the same comparison made in Remark 10 applies to the comparison of their Lemma A.1 with the maximal inequalities developed in this section. Lemma S3 in Davezies et al. (2020) covers the degenerate case but focuses only on the $K = 2$ case. As seen in the proof of Lemma 5 above, however, handling the degenerate components in $K > 2$ cases is highly nontrivial.

Appendix D. Proofs for Section 3

D.1. **Proof of Theorem 3.** Given Lemmas 4 and 5, the proof is almost identical to that of Theorem 1. We omit the details for brevity. □

D.2. **Proof of Theorem 4.** Conditionally on $(\boldsymbol{X}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_{n,K}}$, we have $\sqrt{n}\boldsymbol{S}_n^{MB} \sim N(\boldsymbol{0}, \hat{\Sigma})$, where

$$\hat{\Sigma} = \frac{1}{n} \sum_{j=1}^n (\hat{\boldsymbol{W}}_j - K\boldsymbol{S}_n)(\hat{\boldsymbol{W}}_j - K\boldsymbol{S}_n)^T$$

As in the proof of Theorem 2, the desired result follows from bounding $\hat{\Delta}_W = \|\hat{\Sigma} - \Sigma\|_\infty$.

We first note that

$$\hat{\Delta}_W = \max_{1 \le \ell, \ell' \le p} \left| \frac{1}{n} \sum_{j=1}^n (\hat{W}_j^\ell - K S_n^\ell)(\hat{W}_j^{\ell'} - K S_n^{\ell'}) - \mathbb{E}[W_1^\ell W_1^{\ell'}] \right|.$$

For every $\ell, \ell' \in \{1, \dots, p\}$,

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n (\hat{W}_j^\ell - K S_n^\ell)(\hat{W}_j^{\ell'} - K S_n^{\ell'}) &= \frac{1}{n} \sum_{j=1}^n \hat{W}_j^\ell \hat{W}_j^{\ell'} - K^2 S_n^\ell S_n^{\ell'} \\ &= \frac{1}{n} \sum_{j=1}^n (\hat{W}_j^\ell - W_j^\ell)(\hat{W}_j^{\ell'} - W_j^{\ell'}) + \frac{1}{n} \sum_{j=1}^n (\hat{W}_j^\ell - W_j^\ell) W_j^{\ell'} \\ &\quad + \frac{1}{n} \sum_{j=1}^n W_j^\ell (\hat{W}_j^{\ell'} - W_j^{\ell'}) + \frac{1}{n} \sum_{k=1}^n W_j^\ell W_j^{\ell'} - K^2 S_n^\ell S_n^{\ell'}. \end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$\hat{\Delta}_W \le \underbrace{\max_{1\le \ell \le p} \frac{1}{n}\sum_{j=1}^n (\hat{W}_j^\ell - W_j^\ell)^2}_{=:\Delta_{W,1}} + 2\hat{\Delta}_{W,1}^{1/2}\sqrt{\max_{1\le \ell\le p}\frac{1}{n}\sum_{j=1}^n |W_j^\ell|^2}$$

$$+ \underbrace{\max_{1\le \ell,\ell'\le p}\left|\frac{1}{n}\sum_{j=1}^n (W_j^\ell W_j^{\ell'} - \mathbb{E}[W_1^\ell W_1^{\ell'}])\right|}_{=\hat{\Delta}_{W,2}} + K^2 \max_{1\le \ell\le p}|S_n^\ell|^2.$$

For the second term on the right-hand side, we have

$$\frac{1}{n}\sum_{k=1}^n |W_j^\ell|^2 \le \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n |W_j^\ell|^2\right] + \frac{1}{n}\sum_{j=1}^n (|W_j^\ell|^2 - \mathbb{E}[|W_1^\ell|^2]) \le \overline{\sigma}^2 + \hat{\Delta}_{W,2}.$$

Further, since $KS_n^\ell = n^{-1}\sum_{j=1}^n (\hat{W}_j^\ell - W_j^\ell) + n^{-1}\sum_{j=1}^n W_j^\ell$, we have

$$K^2 \max_{1\le \ell\le p}|S_n^\ell|^2 \le 2\hat{\Delta}_{W,1} + 2\hat{\Delta}_{W,3}^2,$$

where $\hat{\Delta}_{W,3} = \max_{1\le \ell\le p}|n^{-1}\sum_{j=1}^n W_j^\ell|$. Conclude that

$$\hat{\Delta}_W \lesssim \hat{\Delta}_{W,1} + \overline{\sigma}\hat{\Delta}_{W,1}^{1/2} + \hat{\Delta}_{W,2} + \hat{\Delta}_{W,3}$$

up to a universal constant. The rest is completely analogous to the latter part of the proof of Theorem 2. We omit the details for brevity. □

D.3. **Proof of Proposition 2.** We only prove the proposition under Case (i'). The proof for Case (ii') is similar (cf. the proof of Proposition 1). In this proof, the notation $\lesssim$ means that the left-hand side is bounded by the right-hand side up to a constant that depends only on ν, K, and C_1. Recall that $\boldsymbol{W}_j$ can can be written as

$$\boldsymbol{W}_j = \mathbb{E}\left[\frac{(n-K)!}{(n-1)!}\sum_{k=1}^K \sum_{\boldsymbol{i}\in I_{n,K}: i_k = j} \boldsymbol{X}_{\boldsymbol{i}} \;\Big|\; U_j\right].$$

We have

$$\hat{\Delta}_{W,1} = \max_{1\le \ell\le p}\frac{1}{n}\sum_{j=1}^n (\hat{W}_j^\ell - W_j^\ell)^2 \le \frac{1}{n}\sum_{j=1}^n \|\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j\|_\infty^2$$

$$\lesssim \sum_{k=1}^K \frac{1}{n}\sum_{j=1}^n \left\|\frac{(n-K)!}{(n-1)!}\sum_{\boldsymbol{i}\in I_{n,K}: i_k = j}(\boldsymbol{X}_{\boldsymbol{i}} - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_j])\right\|_\infty^2.$$

Consider the $k = 1$ term. Pick any $j \in \mathbb{N}$. Let $I_{\infty,K-1}^{-j} = \{(i_2,\dots,i_K) \in (\mathbb{N}\setminus\{j\})^{K-1} : i_2,\dots,i_K \text{ are distinct}\}$. Given U_j, for each $\boldsymbol{i}_{-1} = (i_2,\dots,i_K) \in I_{\infty,K}^{-j}$ and $\boldsymbol{e} \in \{0,1\}^{K-1}$, define the vector

$$V_{\{\boldsymbol{i}_{-1}\odot \boldsymbol{e}\}^+} = (U_{\{\boldsymbol{i}_{-1}\odot\boldsymbol{e}\}^+}, U_{\{(j,\boldsymbol{i}_{-1}\odot\boldsymbol{e})\}^+}).$$

With this notation, we can rewrite $\boldsymbol{X}_{\boldsymbol{i}}$ with $\boldsymbol{i} = (j, \boldsymbol{i}_{-1})$ as

$$\boldsymbol{X}_{\boldsymbol{i}} = \mathfrak{g}\big(U_j, (V_{\{\boldsymbol{i}_{-1} \odot \boldsymbol{e}\}^+})_{\boldsymbol{e} \in \{0,1\}^{K-1} \setminus \{\boldsymbol{0}\}}\big).$$

From this expression, we see that, conditionally on U_j, the array $(\boldsymbol{X}_{(j,\boldsymbol{i}_{-1})})_{\boldsymbol{i}_{-1} \in I_{\infty,K-1}^{-j}}$ is jointly exchangeable with mean vector $\mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_j]$. Applying Lemmas 4 and 5 conditionally on U_j (the fact that U-variables are uniform on $(0,1)$ is not crucial in the proofs), we have

$$\mathbb{E}\left[\left\| \frac{(n-K)!}{(n-1)!} \sum_{\boldsymbol{i} \in I_{n,K} : i_k = j} (\boldsymbol{X}_{\boldsymbol{i}} - \mathbb{E}[\boldsymbol{X}_{\boldsymbol{i}} \mid U_j]) \right\|_\infty^{2\nu} \mid U_j \right] \lesssim \underbrace{\left(\sum_{k=1}^{K-1} n^{-k/2} (\log p)^{k/2} \right)^{2\nu}}_{\lesssim (n^{-1} \log p)^\nu} \mathbb{E}[\|\boldsymbol{X}_{(j,\boldsymbol{i}_{-1})}\|_\infty^{2\nu} \mid U_j],$$

where $\boldsymbol{i}_{-1} \in I_{\infty,K-1}^{-j}$ is arbitrary. By Fubini's theorem, the expectation of the left-hand side can be bounded as

$$\lesssim (n^{-1} \log p)^\nu \mathbb{E}[\|\boldsymbol{X}_{(j,\boldsymbol{i}_{-1})}\|_\infty^{2\nu}] \lesssim (n^{-1} D_n^2 \log^3 p)^\nu.$$

Similar bounds hold for other k. Conclude that $\mathbb{E}[(\overline{\sigma}^2 \hat{\Delta}_{W,1} \log^4 p)^\nu] \lesssim n^{-\zeta\nu}$ under our assumption. Together with Markov's inequality, we obtain

$$\mathbb{P}\left(\overline{\sigma}^2 \hat{\Delta}_{W,1} \log^4 p > n^{-\zeta + 1/\nu} \right) \lesssim n^{-1}.$$

This completes the proof. □

Appendix E. Proof for Section 4

E.1. **Proof of Proposition 3.**

Proof. In this proof, the notation $\lesssim$ means that the left-hand side is less than the right-hand side up to an n independent constant. Also, $\sum_{j \neq i}$ is understood as $\sum_{j \in \{1,\dots,n\} \setminus \{i\}}$. We will establish the validity of multiplier bootstrap for the non-normalized test statistic as it implies the result for normalized test statistic in view of Corollary 2 under the rate condition of this proposition.

First, let us derive the Hájek projection for the test statistic. Observe that

$$\begin{aligned}
&\mathbb{E}[K_h(y - Y_{ij}) \mathbb{1}(Y_{ij} \neq 0) \mid U_i] = a \int K(z) f_{Y_{12}|U_1}(y + zh \mid U_i) dz = O(1), \\
&\mathbb{E}[|K_h(y - Y_{ij})| \mathbb{1}(Y_{ij} \neq 0) \mid U_i, U_j] \vee \mathbb{E}[K_h^2(y - Y_{ij}) \mathbb{1}(Y_{ij} \neq 0)] = O(h^{-1}) \\
&\mathbb{E}[\hat{b}(y)] = \mathbb{E}[K_h(y - Y_{ij}) \mathbb{1}(Y_{ij} \neq 0)] = a \int K_h(y - z) f(z) dz = a \overline{f}_h(y) =: \overline{b}_h(y).
\end{aligned}$$

Then the Hoeffding type decomposition (3.2) along with Lemma 5 yield $|\hat{a} - a| = O_P(n^{-1/2})$ and $\max_{1 \le \ell \le p} |\hat{b}(y_\ell) - \overline{b}_h(y_\ell)| = O_P\left(\sqrt{n^{-1} \log p}\right)$. Linearization yields that uniformly over $y \in \{y_1, \dots, y_p\}$,

$$\sqrt{n} \left(\frac{\hat{b}(y)}{\hat{a}} - \frac{b_h(y)}{a} \right) = \sqrt{n} \left[\frac{\hat{b}(y) - \overline{b}_h(y)}{a} - \frac{(\hat{a} - a) \overline{b}_h(y)}{a^2} \right] + O_P\left(\frac{\log p}{\sqrt{n}} \right).$$

The leading term on the right-hand side can be written as

$$\frac{\sqrt{n}}{n(n-1)} \sum_{1 \le i < j \le n} 2 \left\{ \frac{K_h(y - Y_{ij})}{a} - \frac{\overline{b}_h(y)}{a^2} \right\} \mathbb{1}(Y_{ij} \neq 0).$$

Note that the summands are centered. Let us define

$$X_{ij}^{\ell} = 2\left\{\frac{K_h(y - Y_{ij})}{a} - \frac{\overline{b}_h(y)}{a^2}\right\}\mathbb{1}(Y_{ij} \neq 0),$$

then we can write

$$\sqrt{n}(\hat{f}(y_\ell) - \overline{f}_h(y_\ell)) = \sqrt{n}(S_n^{\ell} - \mathbb{E}[S_n^{\ell}]) + O_P(\log p/\sqrt{n})$$

uniformly over ℓ. Note a term of the leading component's Hájek projection is given by

$$W_i^{\ell} = \mathbb{E}\left[X_{ij}^{\ell} \mid U_i\right] = 2\left\{\frac{\overline{f}_h(y_\ell \mid U_i)}{a} - \frac{\overline{b}_h(y)}{a^2}\right\}\mathbb{P}\left(Y_{ij} \neq 0 \mid U_i\right).$$

Let us now show the first statement in the proposition. Denote by $\tilde{c}(1-\alpha)$ the conditional $(1-\alpha)$-th quantile of $\|\sqrt{n}\tilde{\boldsymbol{S}}_n^{MB}\|_\infty$, where

$$\tilde{\boldsymbol{S}}_n^{MB} = \frac{1}{n}\sum_{i=1}^{n}\xi_j(\tilde{\boldsymbol{W}}_j - 2\tilde{\boldsymbol{S}}_n), \quad \tilde{W}_i^{\ell} = \frac{1}{n-1}\sum_{j\neq i}2\left\{\frac{K_h(y_\ell - Y_{ij})}{a} - \frac{\overline{b}_h(y_\ell)}{a^2}\right\}\mathbb{1}(Y_{ij} \neq 0),$$

and $\tilde{\boldsymbol{S}}_n = (n(n-1))^{-1}\sum_{1\le i<j\le n}\boldsymbol{X}_{ij}$. In addition, denote by $\tilde{\mathcal{I}}(1-\alpha)$ the infeasible confidence interval

$$\tilde{\mathcal{I}}(1-\alpha) = \prod_{\ell=1}^{p}\left[\hat{f}(y_\ell) \pm \frac{\tilde{c}(1-\alpha)}{\sqrt{n}}\right]$$

Observe that $\|\boldsymbol{X}_{ij}\|_\infty \lesssim h^{-1}$ and thus for $D_n = Ch^{-1}$ with some appropriate constant $C > 0$, once h is small enough, $\max_{1\le\ell\le p}\|X_{12}^{\ell}\|_{\psi_1} \le D_n$, $\max_{1\le\ell\le p}\mathbb{E}[|X_{12}^{\ell}|^{2+\kappa}] \lesssim h^{-(1+\kappa)} \lesssim D_n^{\kappa}$ for $\kappa = 1, 2$. This verifies the conditions required for Theorem 3 and Corollary 2 under Condition (3.3) and Remark 6. For $\boldsymbol{Y} \sim N(\boldsymbol{0}, \Sigma)$ with $\Sigma = 2\mathbb{E}[\boldsymbol{W}_1\boldsymbol{W}_1^T]$, we now have

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}_{|\boldsymbol{X}_{I_{n,2}}}\left(\sqrt{n}\tilde{\boldsymbol{S}}_n^{MB} \in R\right) - \gamma_\Sigma(R)\right| = o_P(1).$$

Observe that conditional on $\boldsymbol{X}_{I_{n,2}}$, we have

$$S_n^{MB} \sim N(\boldsymbol{0}, \hat{\Sigma}), \text{ where } \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{W}}_i - \boldsymbol{S}_n)(\hat{\boldsymbol{W}}_i - \boldsymbol{S}_n)^T.$$

In view of Lemma 8, it suffices to show $\|\hat{\Sigma} - \Sigma\|_\infty = o_P((\log p)^{-2})$, as this implies

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}_{|\boldsymbol{X}_{I_{n,2}}}\left(\sqrt{n}\hat{\boldsymbol{S}}_n^{MB} \in R\right) - \gamma_\Sigma(R)\right| = o_P(1),$$

which in turn gives the desired result. Now, using a similar decomposition as in the proof of Theorem 4, it holds that

$$\begin{aligned}
&\max_{1\le\ell,\ell'\le p}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{(\hat{W}_i^{\ell} - 2S_n^{\ell})(\hat{W}_i^{\ell'} - 2S_n^{\ell'}) - (\tilde{W}_i^{\ell} - 2\tilde{S}_n^{\ell})(\tilde{W}_i^{\ell'} - 2\tilde{S}_n^{\ell'})\right\}\right| \\
&= \max_{1\le\ell,\ell'\le p}\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{W}_i^{\ell}\hat{W}_i^{\ell'} - \tilde{W}_i^{\ell}\tilde{W}_i^{\ell'})\right| + 4\max_{1\le\ell,\ell'\le p}\left|(\tilde{S}_n^{\ell}\tilde{S}_n^{\ell'} - S_n^{\ell}S_n^{\ell'})\right| = I + II.
\end{aligned}$$

First let us consider I. Using the algebraic fact that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{W}_i^{\ell}\hat{W}_i^{\ell'} - \tilde{W}_i^{\ell}\tilde{W}_i^{\ell'}) = \frac{1}{n}\sum_{i=1}^{n}\left\{(\hat{W}_i^{\ell} - \tilde{W}_i^{\ell})(\hat{W}_i^{\ell'} - \tilde{W}_i^{\ell'}) + (\hat{W}_i^{\ell} - \tilde{W}_i^{\ell})\tilde{W}_i^{\ell'} + \tilde{W}_i^{\ell}(\hat{W}_i^{\ell'} - \tilde{W}_i^{\ell'})\right\},$$

we have

$$I \le \underbrace{\max_{1\le\ell\le p} \frac{1}{n}\sum_{i=1}^{n}(\hat{W}_i^\ell - \tilde{W}_i^\ell)^2}_{=:III} + \underbrace{2\Delta_1^{1/2}\sqrt{\max_{1\le\ell\le p}\frac{1}{n}\sum_{i=1}^{n}|\tilde{W}_i^\ell|^2}}_{=:IV}.$$

Now let us consider III. Note that as a is bounded away from zero, with probability $1-o(1)$, it holds for all i that

$$\begin{aligned}|\hat{W}_i^\ell - \tilde{W}_i^\ell|^2 =& \left|\frac{2}{n-1}\sum_{j\ne i}\left\{\frac{a-\hat{a}}{a\hat{a}}K_h(y_\ell - Y_{ij}) + \frac{\hat{a}^2\overline{b}_h(y_\ell) - a^2\hat{b}(y_\ell)}{a^2\hat{a}^2}\right\}\right|^2 \\ \lesssim&(a-\hat{a})^2\left|\frac{1}{n-1}\sum_{j\ne i}K_h(y_\ell - Y_{ij})\right|^2 + (\hat{a}-a)^2 \vee |\overline{b}_h(y_\ell) - \hat{b}(y_\ell)|^2.\end{aligned}$$

To obtain a bound for the first term on the right-hand side, note that for any $j \ne i$,

$$\begin{aligned}\max_{1\le\ell\le p}\left|\frac{1}{n-1}\sum_{j\ne i}K_h(y_\ell - Y_{ij})\right|^2 \lesssim& \max_{1\le\ell\le p}\left|\frac{1}{n-1}\sum_{j\ne i}\{K_h(y_\ell - Y_{ij}) - \mathbb{E}[K_h(y_\ell - Y_{ij}) \mid U_i]\}\right|^2 \\ &+ \max_{1\le\ell\le p}|\mathbb{E}[K_h(y_\ell - Y_{ij}) \mid U_i]|^2.\end{aligned}$$

Conditional on U_i, Theorem 2.14.1 in van der Vaart and Wellner (1996) with $p=2$ yields

$$\begin{aligned}&\mathbb{E}\left[\max_{1\le\ell\le p}\left|\frac{1}{n-1}\sum_{j\ne i}\{K_h(y_\ell - Y_{ij}) - \mathbb{E}[K_h(y_\ell - Y_{ij}) \mid U_i]\}\right|^2 \mid U_i\right] \\ &\lesssim \frac{\log p}{n}\cdot\mathbb{E}\left[\max_{1\le\ell\le p}(K_h(y_\ell - Y_{ij}))^2 \mid U_i\right]. \qquad \text{(E.1)}\end{aligned}$$

Thus by Fubini, we have

$$\mathbb{E}\left[\max_{1\le\ell\le p}\left|\frac{1}{n-1}\sum_{j\ne i}\{K_h(y_\ell - Y_{ij}) - \mathbb{E}[K_h(y_\ell - Y_{ij}) \mid U_i]\}\right|^2\right] = O\left(\frac{\log p}{nh^2}\right) = o(1).$$

On the other hand, $\max_{1\le\ell\le p}|\mathbb{E}[K_h(y_\ell - Y_{ij}) \mid U_i]| = O(1)$ as $f_{Y_{12}|U_1}$ and K are bounded. Hence

$$\begin{aligned}III \le& \frac{1}{n}\sum_{i=1}^{n}\max_{1\le\ell\le p}|\hat{W}_i^\ell - \tilde{W}_i^\ell|^2 \\ =&O_P\left(\mathbb{E}\left[\max_{1\le\ell\le p}\left|\frac{(a-\hat{a})}{n-1}\sum_{j\ne i}K_h(y_\ell - Y_{ij})\right|^2\right] + (\hat{a}-a)^2 \vee \max_{1\le\ell\le p}|\overline{b}_h(y_\ell) - \hat{b}(y_\ell)|^2\right) \\ =&O_P\left(\frac{1}{n} \vee \frac{\log p}{n}\right) = O_P\left(\frac{\log p}{n}\right).\end{aligned}$$

Next let us consider IV. Observe that

$$\max_{1\le\ell\le p}\frac{1}{n}\sum_{i=1}^{n}|\tilde{W}_i^\ell|^2 \lesssim \frac{1}{n}\sum_{i=1}^{n}\max_{1\le\ell\le p}\left|\tilde{W}_i^\ell - \mathbb{E}[\tilde{W}_i^\ell \mid U_i]\right|^2 + \frac{1}{n}\sum_{i=1}^{n}\max_{1\le\ell\le p}\left|\mathbb{E}[\tilde{W}_i^\ell \mid U_i]\right|^2 = O_P(1)$$

To see this, observe that since $\mathbb{E}[\tilde{\boldsymbol{W}}_i \mid U_i] = \mathbb{E}[\boldsymbol{X}_{12} \mid U_1]$ for all i, by Fubini,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\max_{1\le\ell\le p}\left|\mathbb{E}[\tilde{W}_i^\ell \mid U_i]\right|^2\right] \le \mathbb{E}\left[\max_{1\le\ell\le p}\left|\mathbb{E}[X_{12}^\ell \mid U_1]\right|^2\right] = O(1).$$

Furthermore, conditional on U_i,

$$\mathbb{E}\left[\max_{1\le\ell\le p}\left|\tilde{W}_i^\ell - \mathbb{E}[\tilde{W}_i^\ell \mid U_i]\right|^2 \mid U_i\right] = \mathbb{E}\left[\max_{1\le\ell\le p}\left|\tilde{W}_i^\ell - \mathbb{E}[X_{ij} \mid U_i]\right|^2 \mid U_i\right].$$

Similar to Equation (E.1), conditional on U_i,

$$\mathbb{E}\left[\max_{1\le\ell\le p}\left|\tilde{W}_i^\ell - \mathbb{E}[X_{ij} \mid U_i]\right|^2 \mid U_i\right] \lesssim \frac{\log p}{n}\mathbb{E}[\|\boldsymbol{X}_{12}\|_\infty^2 \mid U_1].$$

By Fubini, we have

$$\mathbb{E}\left[\max_{1\le\ell\le p}\left|\tilde{W}_i^\ell - \mathbb{E}[X_{ij} \mid U_i]\right|^2\right] \lesssim \frac{\log p}{nh^2} = o(1).$$

Thus we have,

$$IV = O_P(III^{1/2}) \cdot O_P(1) = O_P\left(\sqrt{\frac{\log p}{n}}\right).$$

Now, for II, since $\|\tilde{\boldsymbol{S}}_n\|_\infty = O_P(\sqrt{\log p/n})$ following the Gaussian approximation of Theorem 3, using the fact that

$$\hat{S}_n^\ell \hat{S}_n^{\ell'} - \tilde{S}_n^\ell \tilde{S}_n^{\ell'} = (\hat{S}_n^\ell - \tilde{S}_n^\ell)(\hat{S}_n^{\ell'} - \tilde{S}_n^{\ell'}) + \tilde{S}_n^\ell(\hat{S}_n^{\ell'} - \tilde{S}_n^{\ell'}) + (\hat{S}_n^\ell - \tilde{S}_n^\ell)\tilde{S}_n^{\ell'},$$

we have $II = O_p\left(|\hat{a} - a| \vee \max_{1\le\ell\le p}|\hat{b}(y_\ell) - \overline{b}_h(y_\ell)|\right) = O_P((n^{-1}\log p)^{1/2})$. Combining the results, we have

$$\|\hat{\Sigma} - \Sigma\|_\infty \log^2 p = O_P\left(\sqrt{\frac{\log^5 p}{n}}\right) = o_P(1).$$

For the second statement of this proposition, note that the bias can be controlled uniformly over $y \in \{y_1, \ldots, y_p\}$ by

$$|\overline{f}_h(y) - f(y)| \le \frac{h^r}{r!}\|f^{(r)}\|_\infty \int |z^r K(z)|dz = O(h^r).$$

Thus, by Lemma 7, we have

$$\begin{aligned}&\left|\mathbb{P}\left((\overline{f}_h(y_\ell))_{\ell=1}^p \in \mathcal{I}(1-\alpha)\right) - \mathbb{P}\left((f(y_\ell))_{\ell=1}^p \in \mathcal{I}(1-\alpha)\right)\right| \\ &\lesssim \sqrt{n\log p}\cdot \max_{1\le\ell\le p}|\overline{f}_h(y_\ell) - f(y_\ell)| = O(h^r\sqrt{n\log p}).\end{aligned}$$

The argument here follows similar steps as in Corollary 3 in Kato and Sasaki (2018). We omit the detail for brevity. □

E.2. **Proof of Proposition 4.** In this proof, the notation $\lesssim$ means that the left-hand side is bounded by the right-hand side up to a constant independent of n.

By Theorem 1 (use Condition (2.4)), we have

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(\|\sqrt{n}\boldsymbol{S}_{\boldsymbol{N}}\|_\infty \le t) - \mathbb{P}(\|\boldsymbol{G}\|_\infty \le t)| \to 0,$$

where $\boldsymbol{G} \sim N(\boldsymbol{0},\Sigma)$ with $\Sigma = \sum_{k=1}^K (n/N_k)\mathbb{E}[\boldsymbol{V}_{k,1}\boldsymbol{V}_{k,1}^T]$. Conditionally on $((Y_{\boldsymbol{i}}, \boldsymbol{Z}_{\boldsymbol{i}}^T)^T)_{\boldsymbol{i}\in[\boldsymbol{N}]}$, we have

$$\sum_{k=1}^K \frac{\sqrt{n}}{N_k}\sum_{i_k=1}^{N_k} \xi_{k,i_k}(\tilde{\boldsymbol{V}}_{k,i_k} - \tilde{\boldsymbol{S}}_N) \sim N(\boldsymbol{0},\tilde\Sigma), \quad \text{where } \tilde\Sigma = \sum_{k=1}^K (n/N_k^2)\sum_{i_k=1}^{N_k}(\tilde{\boldsymbol{V}}_{k,i_k}-\tilde{\boldsymbol{S}}_{\boldsymbol{N}})(\tilde{\boldsymbol{V}}_{k,i_k}-\tilde{\boldsymbol{S}}_{\boldsymbol{N}})^T.$$

Thus, in view of Lemma 8, it suffices to show that $\|\tilde\Sigma - \Sigma\|_\infty \log^2 p = o_P(1)$ (the bound on λ follows from the Gaussian concentration). Further, Proposition 1 and the proof of Theorem 2 under polynomial moment conditions (see also Remark 3) imply that $\|\hat\Sigma - \Sigma\|_\infty = o_P((\log p)^{-2})$, where $\hat\Sigma = \sum_{k=1}^K (n/N_k^2)\sum_{i_k=1}^{N_k}(\hat{\boldsymbol{V}}_{k,i_k} - \boldsymbol{S}_{\boldsymbol{N}})(\hat{\boldsymbol{V}}_{k,i_k} - \boldsymbol{S}_{\boldsymbol{N}})^T$ and $\hat{\boldsymbol{V}}_{k,i_k} = (\prod_{k'\neq k} N_{k'})^{-1}\sum_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_K} \varepsilon_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}}$. Thus, it suffices to show that $\|\tilde\Sigma - \hat\Sigma\|_\infty = o_P((\log p)^{-2})$.

Recall that $\lambda^0 = (\log n)(n^{-1}\log p)^{1/2}$. We note that

$$\mathbb{E}[\|\boldsymbol{G}\|_\infty] \lesssim \max_{j,k}\sqrt{\mathbb{E}[(V_{k,1}^j)^2]\log p} \lesssim \sqrt{\log p},$$

so that $\lambda^0 \ge 2c\|\boldsymbol{S}_{\boldsymbol{N}}\|_\infty$ with probability $1-o(1)$. By assumption, $\kappa(s,c_0)$ is bounded away from zero with probability $1-o(1)$. Thus, Theorem 1 in Belloni and Chernozhukov (2013) implies that

$$\sqrt{\frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(\boldsymbol{X}_{\boldsymbol{i}}^T(\tilde\beta - \beta_0))^2} = O_P\left(\sqrt{\frac{s\log^3(p\overline{N})}{n}}\right).$$

Observe that

$$\|\tilde\Sigma - \hat\Sigma\|_\infty \le \sum_{k=1}^K \underbrace{\max_{1\le j,\ell\le p}\left|\frac{1}{N_k}\sum_{i_k=1}^{N_k}(\tilde V_{k,i_k}^j\tilde V_{k,i_k}^\ell - \hat V_{k,i_k}^j \hat V_{k,i_k}^\ell)\right|}_{=:(I_k)} + K\underbrace{\max_{1\le j,\ell\le p}\left|\tilde S_{\boldsymbol{N}}^j\tilde S_{\boldsymbol{N}}^\ell - S_{\boldsymbol{N}}^j S_{\boldsymbol{N}}^\ell\right|}_{=:(II)}.$$

We first consider the term (I_k). We shall focus on $k=1$ as similar bounds hold for other k. Observe that

$$\begin{aligned}\frac{1}{N_1}\sum_{i_1=1}^{N_1}(\tilde V_{1,i_1}^j\tilde V_{1,i_1}^\ell - \hat V_{1,i_1}^j\hat V_{1,i_1}^\ell) =& \frac{1}{N_1}\sum_{i_1=1}^{N_1}(\tilde V_{1,i_1}^j - \hat V_{1,i_1}^j)(\tilde V_{1,i_1}^\ell - \hat V_{1,i_1}^\ell) + \frac{1}{N_1}\sum_{i_1=1}^{N_1}(\tilde V_{1,i_1}^j - \hat V_{1,i_1}^j)\hat V_{1,i_1}^\ell \\ &+ \frac{1}{N_1}\sum_{i_1=1}^{N_1}\hat V_{1,i_1}^j(\tilde V_{1,i_1}^\ell - \hat V_{1,i_1}^\ell).\end{aligned}$$

By Cauchy-Schwarz, we have

$$(I_1) \le \underbrace{\max_{1\le j\le p}\frac{1}{N_1}\sum_{i_1=1}^{N_1}(\tilde V_{1,i_1}^j - \hat V_{1,i_1}^j)^2}_{=:(III)} + 2(III)^{1/2}\sqrt{\underbrace{\max_{1\le \ell\le p}\frac{1}{N_1}\sum_{i_1=1}^{N_1}|\hat V_{1,i_1}^\ell|^2}_{=:(IV)}}.$$

To bound (IV), we note that

$$\max_{1\le \ell\le p} \frac{1}{N_1}\sum_{i_1=1}^{N_1} |\hat{V}^{\ell}_{1,i_1}|^2 \le \frac{1}{N_1}\sum_{i_1=1}^{N_1} \left\| \hat{\boldsymbol{V}}_{1,i_1} - \mathbb{E}[\hat{\boldsymbol{V}}_{1,i_1} \mid U_{(i_1,0,\dots,0)}]\right\|_\infty^2 + \frac{1}{N_1}\sum_{i_1=1}^{N_1} \left\|\mathbb{E}[\hat{\boldsymbol{V}}_{1,i_1} \mid U_{(i_1,0,\dots,0)}]\right\|_\infty^2$$

Since $\mathbb{E}[\hat{\boldsymbol{V}}_{1,i_1} \mid U_{(i_1,0,\dots,0)}] = \mathbb{E}[\varepsilon_{\boldsymbol{1}} X_{\boldsymbol{1}} \mid U_{(1,0,\dots,0)}]$ for all i_1, by Fubini and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N_1}\sum_{i_1=1}^{N_1}\left\|\mathbb{E}[\hat{\boldsymbol{V}}_{1,i_1} \mid U_{(i_1,0,\dots,0)}]\right\|_\infty^2\right] &\le \left(\mathbb{E}\left[\left\|\mathbb{E}[\varepsilon_{\boldsymbol{1}} X^{\ell}_{\boldsymbol{1}} \mid U_{(1,0,\dots,0)}]\right\|_\infty^q\right]\right)^{2/q} \\ &\le \left(\mathbb{E}\left[\max_{1\le\ell\le p} |\varepsilon_{\boldsymbol{1}} X^{\ell}_{\boldsymbol{1}}|^q\right]\right)^{2/q} \le D^2_{\boldsymbol{N}}.\end{aligned}$$

Conditionally on $U_{(i_1,0,\dots,0)}$,

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{V}}_{1,i_1} - \mathbb{E}[\hat{\boldsymbol{V}}_{1,i_1} \mid U_{(i_1,0,\dots,0)}]\right\|_\infty^2 \mid U_{(i_1,0,\dots,0)}\right] \le \left(\mathbb{E}\left[\left\|\hat{\boldsymbol{V}}_{1,i_1} - \mathbb{E}[\varepsilon_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,\dots,0)}]\right\|_\infty^q \mid U_{(i_1,0,\dots,0)}\right]\right)^{2/q}.$$

As in the proof of Proposition 1, conditionally on $U_{(i_1,0,\dots,0)}$, the array $(\varepsilon_{(i_1,\boldsymbol{i}_{-1})}\boldsymbol{X}_{(i_1,\boldsymbol{i}_{-1})})_{\boldsymbol{i}_{-1}\in\mathbb{N}^{K-1}}$ is separately exchangeable with mean vector $\mathbb{E}[\varepsilon_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,\dots,0)}]$. By Corollary 3, we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{V}}_{1,i_1} - \mathbb{E}[\varepsilon_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,\dots,0)}]\right\|_\infty^q \mid U_{(i_1,0,\dots,0)}\right] \lesssim n^{-q/2}(\log p)^{q/2}\mathbb{E}[\|\varepsilon_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}}\|_\infty^q \mid U_{(i_1,0,\dots,0)}].$$

By Fubini, we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{V}}_{1,i_1} - \mathbb{E}[\varepsilon_{\boldsymbol{i}}\boldsymbol{X}_{\boldsymbol{i}} \mid U_{(i_1,0,\dots,0)}]\right\|_\infty^q\right] \lesssim n^{-q/2}(\log p)^{q/2} D^q_{\boldsymbol{N}}.$$

Conclude that $|(IV)| = O_P(D^2_{\boldsymbol{N}})$.

Next, we shall bound the term (III). Observe that by Cauchy-Schwarz,

$$\begin{aligned}|\tilde{V}^j_{1,i_1} - \hat{V}^j_{1,i_1}| &= \left|\frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K} X^j_{\boldsymbol{i}}(\boldsymbol{X}^T_{\boldsymbol{i}}(\tilde{\beta}-\beta_0) + r_{\boldsymbol{i}})\right| \\ &\le \sqrt{\frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K}(X^j_{\boldsymbol{i}})^2}\left(\sqrt{\frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K}(\boldsymbol{X}^T_{\boldsymbol{i}}(\tilde{\beta}-\beta_0))^2} + \sqrt{\frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K} r^2_{\boldsymbol{i}}}\right),\end{aligned}$$

so that the term (III) is bounded as

$$\begin{aligned}&\lesssim \max_j \frac{1}{N}\sum_{i_1=1}^{N_1}\left(\frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K}(X^j_{\boldsymbol{i}})^2\right)\left(\frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K}(\boldsymbol{X}^T_{\boldsymbol{i}}(\tilde{\beta}-\beta_0))^2 + \frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K} r^2_{\boldsymbol{i}}\right) \\ &\le \left(\max_{j,i_1}\frac{1}{\prod_{k\ne 1} N_k}\sum_{i_2,\dots,i_K}(X^j_{\boldsymbol{i}})^2\right)\underbrace{\left(\frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(\boldsymbol{X}^T_{\boldsymbol{i}}(\tilde{\beta}-\beta_0))^2 + \frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]} r^2_{\boldsymbol{i}}\right)}_{=O_P\left(\frac{s\log^3(p\overline{N})}{n}\right)}.\end{aligned}$$

Observe that

$$\begin{aligned}
&\mathbb{E}\left[\max_{j,i_1}\frac{1}{\prod_{k\neq 1}N_k}\sum_{i_2,\dots,i_K}(X_{\boldsymbol{i}}^j)^2\right] \\
&\le \mathbb{E}\left[\max_{j,i_1}\left|\frac{1}{\prod_{k\neq 1}N_k}\sum_{i_2,\dots,i_K}\{(X_{\boldsymbol{i}}^j)^2-\mathbb{E}[(X_{(i_1,1,\dots,1)}^j)^2\mid U_{(i_1,0,\dots,0)}]\}\right|\right] \\
&\quad + \mathbb{E}\left[\max_{j,i_1}\mathbb{E}[(X_{(i_1,1,\dots,1)}^j)^2\mid U_{(i_1,0,\dots,0)}]\right].
\end{aligned}$$

By Hölder's inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{j,i_1}\mathbb{E}[(X_{(i_1,1,\dots,1)}^j)^2\mid U_{(i_1,0,\dots,0)}]\right] &\le \mathbb{E}\left[\max_{i_1}\mathbb{E}[\|\boldsymbol{X}_{(i_1,1,\dots,1)}\|_\infty^2\mid U_{(i_1,0,\dots,0)}]\right] \\
&\le \mathbb{E}\left[\max_{i_1}\left(\mathbb{E}[\|\boldsymbol{X}_{(i_1,1,\dots,1)}\|_\infty^{2q}\mid U_{(i_1,0,\dots,0)}]\right)^{1/q}\right] \\
&\le \left(\mathbb{E}\left[\sum_{i_1}\mathbb{E}[\|\boldsymbol{X}_{(i_1,1,\dots,1)}\|_\infty^{2q}\mid U_{(i_1,0,\dots,0)}]\right]\right)^{1/q} \\
&\le \overline{N}^{1/q}D_{\boldsymbol{N}}.
\end{aligned}$$

Applying Corollary 3 conditionally on $U_{(i_1,0,\dots,0)}$ (cf. the proof of Proposition 1), we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{j}\left|\frac{1}{\prod_{k\neq 1}N_k}\sum_{i_2,\dots,i_K}\{(X_{\boldsymbol{i}}^j)^2-\mathbb{E}[(X_{(i_1,1,\dots,1)}^j)^2\mid U_{(i_1,0,\dots,0)}]\}\right|^q\mid U_{(i_1,0,\dots,0)}\right] \\
&\lesssim n^{-q/2}(\log p)^{q/2}\mathbb{E}[\|\boldsymbol{X}_{(i_1,1,\dots,1)}\|_\infty^{2q}\mid U_{(i_1,0,\dots,0)}].
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{j,i_1}\left|\frac{1}{\prod_{k\neq 1}N_k}\sum_{i_2,\dots,i_K}\{(X_{\boldsymbol{i}}^j)^2-\mathbb{E}[(X_{(i_1,1,\dots,1)}^j)^2\mid U_{(i_1,0,\dots,0)}]\}\right|\right] \\
&\le \left(\sum_{i_1}\mathbb{E}\left[\max_{j}\left|\frac{1}{\prod_{k\neq 1}N_k}\sum_{i_2,\dots,i_K}\{(X_{\boldsymbol{i}}^j)^2-\mathbb{E}[(X_{(i_1,1,\dots,1)}^j)^2\mid U_{(i_1,0,\dots,0)}]\}\right|^q\right]\right)^{1/q} \\
&\lesssim \overline{N}^{1/q}n^{-1/2}(\log p)^{1/2}D_{\boldsymbol{N}}.
\end{aligned}$$

Conclude that $(III)=O_P\left(\{n^{-1}s\overline{N}^{1/q}D_{\boldsymbol{N}}\log^3(p\overline{N})\}^{1/2}\right)$ and consequently

$$|(I_1)|=O_P\left(\{n^{-1}s\overline{N}^{1/q}D_{\boldsymbol{N}}^3\log^3(p\overline{N})\}^{1/2}\right).$$

Finally, to bound $|(II)|$, observe that

$$\begin{aligned}
\tilde{S}_{\boldsymbol{N}}^j\tilde{S}_{\boldsymbol{N}}^\ell - S_{\boldsymbol{N}}^jS_{\boldsymbol{N}}^\ell = (\tilde{S}_{\boldsymbol{N}}^j-S_{\boldsymbol{N}}^j)(\tilde{S}_{\boldsymbol{N}}^\ell-S_{\boldsymbol{N}}^\ell)+S_{\boldsymbol{N}}^j(\tilde{S}_{\boldsymbol{N}}^\ell-S_{\boldsymbol{N}}^\ell) \\
+(\tilde{S}_{\boldsymbol{N}}^j-S_{\boldsymbol{N}}^j)S_{\boldsymbol{N}}^\ell.
\end{aligned}$$

Then, we have

$$|(II)| \leq \max_{1\leq j\leq p}\left|\frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(\tilde{\varepsilon}_{\boldsymbol{i}}-\varepsilon_{\boldsymbol{i}})X_{\boldsymbol{i}}^{j}\right|^2 + 2\|\boldsymbol{S}_{\boldsymbol{N}}\|_\infty\cdot\max_{1\leq j\leq p}\left|\frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(\tilde{\varepsilon}_{\boldsymbol{i}}-\varepsilon_{\boldsymbol{i}})X_{\boldsymbol{i}}^{j}\right|.$$

By Cauchy-Schwarz, we have

$$\begin{aligned}\max_{1\leq j\leq p}\left|\frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(\tilde{\varepsilon}_{\boldsymbol{i}}-\varepsilon_{\boldsymbol{i}})X_{\boldsymbol{i}}^{j}\right| &\leq \max_{1\leq j\leq p}\sqrt{\frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(X_{\boldsymbol{i}}^{j})^2}\left(\sqrt{\frac{1}{N}\sum_{\boldsymbol{i}\in[\boldsymbol{N}]}(\boldsymbol{X}_{\boldsymbol{i}}^{T}(\beta_0-\tilde{\beta}))^2}+\|\boldsymbol{r}\|_{N,2}\right)\\ &= O_P\left(\sqrt{\frac{sD_{\boldsymbol{N}}\log^3(\overline{N}p)}{n}}\right),\end{aligned}$$

so that $|(II)| = O_P\left(n^{-1}sD_{\boldsymbol{N}}^3\log^3(p\overline{N}) + \{n^{-2}sD_{\boldsymbol{N}}(\log p)(\log^3(p\overline{N}))\}^{1/2}\right)$.

Combining the above bounds, we have $\|\tilde{\Sigma}-\hat{\Sigma}\|_\infty = O_P\left(\{n^{-1}s\overline{N}^{1/q}D_{\boldsymbol{N}}^3\log^3(p\overline{N})\}^{1/2}\right)$. This implies that $\|\tilde{\Sigma}-\hat{\Sigma}\|_\infty\log^2 p = o_P(1)$, as required. □

E.3. **Proof of Proposition 5.** Recall that $K=2$. We write $\boldsymbol{X}_{i,j}$ instead of $\boldsymbol{X}_{(i,j)}$ for the notational simplicity. Define the $N\times p$ matrix $\mathbb{X} = (\boldsymbol{X}_{1,1},\dots,\boldsymbol{X}_{N_1,1},\boldsymbol{X}_{2,1},\dots,\boldsymbol{X}_{N_1,N_2})^T$. The s-sparse eigenvalue with $1\leq s\leq p$ for $\mathbb{X}$ is defined by

$$\phi_{\min}(s) = \min_{\|\theta\|_0\leq s,\|\theta\|=1}\|\mathbb{X}\theta\|_{N,2}.$$

By Lecué and Mendelson (2017, Lemma 2.7), if $\phi_{\min}(s)\geq\phi_1$, then for $2\leq s\leq p$, we have

$$\|\mathbb{X}\theta\|_{N,2}^2 \geq \phi_1^2\|\theta\|^2 - \frac{\|\theta\|_1^2}{s-1}\times\underbrace{\max_{1\leq\ell\leq p}\sum_{(i,j)\in[\boldsymbol{N}]}(X_{i,j}^{\ell})^2/N}_{=:\hat{\rho}}$$

for all $\theta\in\mathbb{R}^p$. We can then deduce that for $s_1\leq(s-1)\phi_1^2/(2(1+c_0)^2\hat{\rho})$, we have

$$\kappa(s_1,c_0)\geq\phi_1/\sqrt{2}.$$

Lemma 6 below implies that $\phi_{\min}(s)$ is bounded away from zero with probability $1-o(1)$. Further, observe that

$$\hat{\rho}\leq\max_{1\leq\ell\leq p}\mathbb{E}[(X_{1,1}^{\ell})^2] + \max_{1\leq\ell\leq p}\left|N^{-1}\sum_{(i,j)\in[\boldsymbol{N}]}\{(X_{i,j}^{\ell})^2-\mathbb{E}[(X_{1,1}^{\ell})^2]\}\right|.$$

The first term on the right-hand side is $O(1)$, while the second term is $o_P(1)$ (which follows from Lemma 6 below with $s=1$), so that $\hat{\rho}=O_P(1)$. The conclusion of the proposition follows from rescaling s. □

Lemma 6 (Sparse eigenvalues for two-way clustering)**.** *Suppose that* $(\boldsymbol{X}_{i,j})_{(i,j)\in[\boldsymbol{N}]}$ *with* $[\boldsymbol{N}] = \{1,\dots,N_1\}\times\{1,\dots,N_2\}$ *is sampled from a separately exchangeable array* $(\boldsymbol{X}_{i,j})_{(i,j)\in\mathbb{N}^2}$ *generated as* $\boldsymbol{X}_{i,j} = \mathfrak{g}(U_{i,0},U_{0,j},U_{i,j})$ *for some Borel measurable map* $\mathfrak{g}:[0,1]^3\to\mathbb{R}^p$ *and i.i.d.* $U[0,1]$

variables $U_{i,0}, U_{j,0}, U_{i,j}$. Pick any $1 \le s \le p \wedge n$. Let $B = \sqrt{\mathbb{E}[M^2]}$ with $M = \max_{(i,j)\in[\boldsymbol{N}]} \|\boldsymbol{X}_{i,j}\|_\infty$. Define

$$\delta_{\boldsymbol{N}} = \sqrt{s}B\left(\frac{1}{\sqrt{n}}\left\{\log^{1/2} p + (\log s)(\log^{1/2}\overline{N})(\log^{1/2} p)\right\} \bigvee \frac{1}{\sqrt{\overline{N}}}\left\{\log p + (\log \overline{N})(\log p)\right\}\right).$$

Then, we have

$$\mathbb{E}\left[\sup_{\|\theta\|_0\le s, \|\theta\|=1}\left|\frac{1}{N}\sum_{(i,j)\in[\boldsymbol{N}]}\{(\theta^T\boldsymbol{X}_{i,j})^2 - \mathbb{E}[(\theta^T\boldsymbol{X}_{1,1})^2]\}\right|\right] \lesssim \delta_{\boldsymbol{N}}^2 + \delta_{\boldsymbol{N}} \sup_{\|\theta\|_0\le s,\, \|\theta\|=1}\sqrt{\mathbb{E}[(\theta^T\boldsymbol{X}_{1,1})^2]}$$

up to a universal constant. In addition, we have $\delta_{\boldsymbol{N}} \lesssim \{n^{-1}sB^2\log^4(p\overline{N})\}^{1/2}$ up to a universal constant.

Proof of Lemma 6. In this proof, the notation $\lesssim$ means that the left-hand side is bounded by the right-hand side up to a universal constant.

Let $\Theta_s = \cup_{|T|=s}\{\theta \in \mathbb{R}^p : \|\theta\| = 1, \operatorname{supp}(\theta) \subset T\}$. Further, let $Z_{i,j}(\theta) = (\theta^T\boldsymbol{X}_{i,j})^2 - \mathbb{E}[(\theta^T\boldsymbol{X}_{1,1})^2]$. Then, for each θ, $Z_{i,j}(\theta)$ is a centered random variable. Consider the decomposition

$$Z_{i,j}(\theta) = \mathbb{E}[Z_{i,1}(\theta) \mid U_{i,0}] + \mathbb{E}[Z_{1,j}(\theta) \mid U_{0,j}] + \underbrace{Z_{i,j}(\theta) - \mathbb{E}[Z_{i,1}(\theta) \mid U_{i,0}] - \mathbb{E}[Z_{1,j}(\theta) \mid U_{0,j}]}_{=:\hat{Z}_{i,j}(\theta)}.$$

We divide the rest of the proof into two steps.

Step 1. Consider first the term $\sum_{i,j}\mathbb{E}[Z_{i,j}(\theta) \mid U_{i,0}] = N_2\sum_{i=1}^{N_1}\mathbb{E}[Z_{i,1}(\theta) \mid U_{i,0}]$, which consists of i.i.d. variables. Observe that $\mathbb{E}[Z_{i,1}(\theta) \mid U_{i,0}]$ has mean 0 and by symmetrization

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i=1}^{N_1}\mathbb{E}[Z_{i,1}(\theta) \mid U_{i,0}]\right|\right] &= \mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i=1}^{N_1}\left(\theta^T\mathbb{E}[\boldsymbol{X}_{i,1}\boldsymbol{X}_{i,1}^T \mid U_{i,0}]\theta - \mathbb{E}[(\theta^T\boldsymbol{X}_{1,1})^2]\right)\right|\right]\\ &\le 2\mathbb{E}\left[\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i=1}^{N_1}\epsilon_i\left(\theta^T\mathbb{E}[\boldsymbol{X}_{i,1}\boldsymbol{X}_{i,1}^T \mid U_{i,0}]\theta\right)\right| \mid \boldsymbol{X}_{[\boldsymbol{N}]}\right]\right]\\ &\le 2\mathbb{E}\left[\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i=1}^{N_1}\epsilon_i(\theta^T\boldsymbol{X}_{i,1})^2\right| \mid \boldsymbol{X}_{[\boldsymbol{N}]}\right]\right],\end{aligned}$$

where $(\epsilon_i)_{i=1}^{N_1}$ is a sequence of independent Rademacher random variables that are independent of $(\boldsymbol{X}_{i,j})_{(i,j)\in[\boldsymbol{N}]}$, and the second inequality follows from Jensen's inequality. Now, the following bound can be obtained by following the proof of Lemma P.1. in Belloni et al. (2018) with $\mathcal{U}$ set to be a singleton set:

$$\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i=1}^{N_1}\epsilon_i(\theta^T\boldsymbol{X}_{i,1})^2\right| \mid \boldsymbol{X}_{[\boldsymbol{N}]}\right] \lesssim \sqrt{s}MR_1(\log^{1/2} p + (\log s)(\log^{1/2}\overline{N})(\log^{1/2} p)),$$

where $R_1 = \sup_{\theta\in\Theta_s}\left(\sum_{i=1}^{N_1}(\theta^T\boldsymbol{X}_{i,1})^2\right)^{1/2}$.

Choosing $\delta_{\boldsymbol{N},1} = BN_1^{-1/2}\sqrt{s}\{\log^{1/2} p + (\log s)(\log^{1/2} \overline{N})(\log^{1/2} p)\}$, by Cauchy-Schwarz, we have

$$I := \mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i=1}^{N_1}\epsilon_i(\theta^T\boldsymbol{X}_{i,1})^2\right|\right] \lesssim \frac{\delta_{\boldsymbol{N},1}\mathbb{E}[MR_1]}{B\sqrt{N_1}} \le \left(\frac{\delta_{\boldsymbol{N},1}}{B}\right)\left(\frac{\mathbb{E}[M^2]\mathbb{E}[R_1^2]}{N_1}\right)^{1/2}$$
$$\le \delta_{\boldsymbol{N},1}(\mathbb{E}[R_1^2/N_1])^{1/2} \lesssim \delta_{\boldsymbol{N},1}\left(I + \sup_{\theta\in\Theta_s}\mathbb{E}[(\theta^T\boldsymbol{X}_{1,1})^2]\right)^{1/2}.$$

Using the algebraic fact that $a^2 \le \delta^2 a + \delta^2 b$ implies $a \le \delta^2 + a^{-1}\delta^2 b$, we have

$$I \lesssim \delta_{\boldsymbol{N},1}^2 + \delta_{\boldsymbol{N},1}\sqrt{\sup_{\theta\in\Theta_s}\mathbb{E}[(\theta^T\boldsymbol{X}_{1,1})^2]}.$$

The same bound holds for $\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|N_2^{-1}\sum_{j=1}^{N_2}\mathbb{E}[Z_{1,j}(\theta)\mid U_{0,j}]\right|\right]$. Conclude that

$$\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\frac{1}{N}\sum_{i,j}(\mathbb{E}[Z_{i,j}(\theta)\mid U_{i,0}] + \mathbb{E}[Z_{i,j}(\theta)\mid U_{0,j}])\right|\right] \lesssim \delta_{\boldsymbol{N},2}^2 + \delta_{\boldsymbol{N},2}\sqrt{\sup_{\theta\in\Theta_s}\mathbb{E}[(\theta^T\boldsymbol{X}_{1,1})^2]},$$

where $\delta_{\boldsymbol{N},2} = Bn^{-1/2}\sqrt{s}\{\log^{1/2} p + (\log s)(\log^{1/2}\overline{N})(\log^{1/2} p)\} \lesssim Bn^{-1/2}\sqrt{s}\log^2(p\overline{N})$.

Step 2. Now, to obtain a bound on $\mathbb{E}[\sup_{\theta\in\Theta_s}|N^{-1}\sum_{i,j}\hat{Z}_{i,j}(\theta)|]$, by Lemma 2, we have the following symmetrization inequality

$$\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i,j}\hat{Z}_{i,j}(\theta)\right|\right] \le 4\mathbb{E}\left[\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i,j}\epsilon_i\epsilon_j'\hat{Z}_{i,j}(\theta)\right| \mid \boldsymbol{X}_{[\boldsymbol{N}]}\right]\right]$$
$$\lesssim \mathbb{E}\left[\mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i,j}\epsilon_i\epsilon_j'(\theta^T\boldsymbol{X}_{i,j})^2\right| \mid \boldsymbol{X}_{[\boldsymbol{N}]}\right]\right],$$

where (ϵ_i) and (ϵ_i') are independent copies of Rademacher random variables independent of $(\boldsymbol{X}_{i,j})_{(i,j)\in[\boldsymbol{N}]}$, and the second inequality follows from Jensen's inequality. Conditionally on $(\boldsymbol{X}_{i,j})_{(i,j)\in[\boldsymbol{N}]}$, $\sum_{i,j}\epsilon_i\epsilon_j'(\theta^T\boldsymbol{X}_{i,j})^2$ is a Rademacher chaos of degree 2 (cf. the proof of Theorem 5). Hence, Corollary 5.1.8 in de la Peña and Giné (1999) yields that

$$II := \mathbb{E}\left[\sup_{\theta\in\Theta_s}\left|\sum_{i,j}\epsilon_i\epsilon_j'(\theta^T\boldsymbol{X}_{i,j})^2\right| \mid \boldsymbol{X}_{[\boldsymbol{N}]}\right] \lesssim \left\|\sup_{\theta\in\Theta_s}\left|\sum_{i,j}\epsilon_i\epsilon_j'(\theta^T\boldsymbol{X}_{i,j})^2\right|\right\|_{\psi_1|\boldsymbol{X}}$$
$$\lesssim \int_0^{\mathrm{diam}(\Theta_s)}\log N(\Theta_s,\rho_{\boldsymbol{X}},t)dt,$$

where $\|\cdot\|_{\psi_1|\boldsymbol{X}}$ is the ψ_1-norm evaluated conditionally on $(\boldsymbol{X}_{i,j})_{(i,j)\in[\boldsymbol{N}]}$, $\rho_{\boldsymbol{X}}$ is a pseudometric on Θ_s defined by $\rho_{\boldsymbol{X}}(\theta,\overline{\theta}) = \left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\{(\theta^T\boldsymbol{X}_{i,j})^2 - (\overline{\theta}^T\boldsymbol{X}_{i,j})^2\}^2\right)^{1/2}$, and $\mathrm{diam}(\Theta_s)$ is the $\rho_{\boldsymbol{X}}$-diameter of Θ_s. Now, for any two θ, $\bar{\theta} \in \Theta_s$,

$$\begin{aligned}
\rho_{\boldsymbol{X}}(\theta, \bar{\theta}) &= \left(\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left\{ (\theta^T \boldsymbol{X}_{i,j})^2 - (\bar{\theta}^T \boldsymbol{X}_{i,j})^2 \right\}^2 \right)^{1/2} \\
&\le \left(\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left\{ (\theta^T \boldsymbol{X}_{i,j}) + (\bar{\theta}^T \boldsymbol{X}_{i,j}) \right\}^2 \right)^{1/2} \max_{(i,j)\in[\boldsymbol{N}]} |(\theta - \bar{\theta})^T \boldsymbol{X}_{i,j}| \\
&\le \sqrt{2} R_2 \|\theta - \bar{\theta}\|_{\boldsymbol{X}},
\end{aligned}$$

where $R_2 = \sup_{\theta \in \Theta_s} \left(\sum_{(i,j)\in[\boldsymbol{N}]} (\theta^T \boldsymbol{X}_{i,j})^2 \right)^{1/2}$ and $\|\theta\|_{\boldsymbol{X}} = \max_{(i,j)\in[\boldsymbol{N}]} |\theta^T \boldsymbol{X}_{i,j}|$. Thus, we have

$$\begin{aligned}
\int_0^{\mathrm{diam}(\Theta_s)} \log N(\Theta_s, \rho_{\boldsymbol{X}}, t) dt &\le \int_0^{2\sqrt{2s}MR_2} \log N\left(\Theta_s/\sqrt{s}, \|\cdot\|_{\boldsymbol{X}}, t/(\sqrt{2s}R_2)\right) dt \\
&= 2\sqrt{2s} R_2 \int_0^M \log N\left(\Theta_s/\sqrt{s}, \|\cdot\|_{\boldsymbol{X}}, t\right) dt.
\end{aligned}$$

Lemma 3.9 and Equation (3.10) in Rudelson and Vershynin (2008) yield that for some universal constant A,

$$\begin{aligned}
&\int_0^M \log N\left(\Theta_s/\sqrt{s}, \|\cdot\|_{\boldsymbol{X}}, t\right) dt \\
&\le \int_0^{M/\sqrt{s}} \log\left(\binom{p}{s} (1 + 2M/t)^s \right) dt + \int_{M/\sqrt{s}}^M \log\left((2p)^{At^{-2}M^2 \log N} \right) dt \\
&\le \frac{M}{\sqrt{s}} \log \binom{p}{s} + \sqrt{s} \int_0^{M/\sqrt{s}} \log(1 + 2M/t) dt + AM^2 (\log N)(\log(2p)) \int_{M/\sqrt{s}}^M \frac{dt}{t^2} \\
&\lesssim M\sqrt{s} \log p + M(1 + 2\sqrt{s}) \log\left(1 + \frac{1}{2\sqrt{s}}\right) + A\sqrt{s} M (\log N)(\log(2p)) \\
&\lesssim \sqrt{s} M \left(\log p + (\log \overline{N})(\log p) \right),
\end{aligned}$$

where the second term follows from integration by parts

$$\begin{aligned}
\sqrt{s} \int_0^{M/\sqrt{s}} \log(1 + 2M/t) dt &\le \sqrt{s} t \log\left(1 + \frac{2M}{t}\right) \Big|_0^{M/\sqrt{s}} + \sqrt{s} 2M \log(t + 2M) \Big|_0^{M/\sqrt{s}} \\
&\lesssim M(1 + 2\sqrt{s}) \log\left(1 + \frac{1}{2\sqrt{s}}\right).
\end{aligned}$$

Hence, we have $II \lesssim sR_2 M \left\{ \log p + (\log \overline{N})(\log p) \right\}$.

Setting $\delta_{\boldsymbol{N},3} = sN^{-1/2} B \left(\log p + (\log \overline{N})(\log p) \right)$, we have

$$\begin{aligned}
III := \mathbb{E}\left[\sup_{\theta \in \Theta_s} \left| \sum_{i,j} \epsilon_i \epsilon'_j (\theta^T \boldsymbol{X}_{i,j})^2 \right| \right] &\lesssim \frac{\delta_{\boldsymbol{N},3} \mathbb{E}[MR_2]}{B\sqrt{N}} \le \left(\frac{\delta_{\boldsymbol{N},3}}{B} \right) \left(\frac{\mathbb{E}[M^2]\mathbb{E}[R_2^2]}{N} \right)^{1/2} \\
&\le \delta_{\boldsymbol{N},3} \left(\frac{\mathbb{E}[R_2^2]}{N} \right)^{1/2} \lesssim \delta_{\boldsymbol{N},3} \left(III + \sup_{\theta \in \Theta_s} \mathbb{E}[(\theta^T \boldsymbol{X}_{1,1})^2] \right)^{1/2}.
\end{aligned}$$

Using the same algebraic fact as in Step 1 yields that $III \lesssim \delta_{\boldsymbol{N},3}^2 + \delta_{\boldsymbol{N},3} \sqrt{\sup_{\theta \in \Theta_s} \mathbb{E}[(\theta^T \boldsymbol{X}_{1,1})^2]}$.

Finally, since $n \leq \sqrt{N}$ and $s \leq n$, we have

$$\frac{sB}{\sqrt{N}}\left(\log p + (\log \overline{N})(\log p)\right) \lesssim \frac{sB}{n}\left(\log p + (\log \overline{N})(\log p)\right) \lesssim \frac{\sqrt{s}B}{\sqrt{n}} \log^2(p\overline{N}).$$

This completes the proof. □

Appendix F. Technical Tools

Lemma 7 (Nazarov's inequality)**.** *Let $\boldsymbol{Y} = (Y^1, \dots, Y^p)^T$ be a centered Gaussian random vector in $\mathbb{R}^p$ such that $\mathbb{E}[|Y^j|^2] \geq \underline{\sigma}^2$ for all $1 \leq j \leq p$ and some constant $\underline{\sigma} > 0$. Then for every $\boldsymbol{y} \in \mathbb{R}^p$ and $\delta > 0$,*

$$\mathbb{P}(\boldsymbol{Y} \leq \boldsymbol{y} + \delta) - \mathbb{P}(\boldsymbol{Y} \leq \boldsymbol{y}) \leq \frac{\delta}{\underline{\sigma}}(\sqrt{2\log p} + 2).$$

Proof. This is Lemma A.1 in Chernozhukov et al. (2017a); see Chernozhukov et al. (2017b) for its proof. □

Lemma 8 (Gaussian comparison over rectangles)**.** *Let $\boldsymbol{Y}$ and $\boldsymbol{W}$ be centered Gaussian random vectors in $\mathbb{R}^d$ with covariance matrices $\Sigma^Y = (\Sigma^Y_{j,k})_{1 \leq j,k \leq d}$ and $\Sigma^W = (\Sigma^W_{j,k})_{1 \leq j,k \leq d}$, respectively, and let $\Delta = \|\Sigma^Y - \Sigma^W\|_\infty$. Suppose that $\min_{1 \leq j \leq d} \Sigma^Y_{j,j} \bigvee \min_{1 \leq j \leq d} \Sigma^W_{j,j} \geq \underline{\sigma}^2$ for some constant $\underline{\sigma} > 0$. Then*

$$\sup_{R \in \mathcal{R}} |\mathbb{P}(\boldsymbol{Y} \in R) - \mathbb{P}(\boldsymbol{W} \in R)| \leq C(\Delta \log^2 d)^{1/2},$$

where C is a constant that depends only on $\underline{\sigma}$.

Proof. See Corollary 5.1 in Chernozhukov et al. (2019b). □

References

Abbring, J. H. and A. de Paula (2017): "Special issue on econometrics of networks," *The Econometrics Journal*, 20, SI–SII.

Aldous, D. J. (1981): "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, 11, 581–598.

Andrews, D. W. (2005): "Cross-section regression with common shocks," *Econometrica*, 73, 1551–1585.

Andrews, D. W. and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666.

——— (2017): "Inference based on many conditional moment inequalities," *Journal of Econometrics*, 196, 275–287.

Armstrong, T. B. (2014): "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, 181, 92–116.

Armstrong, T. B. and H. P. Chan (2016): "Multiscale adaptive inference on conditional moment inequalities," *Journal of Econometrics*, 194, 24–43.

Aronow, P. M., C. Samii, and V. A. Assenova (2015): "Cluster–robust variance estimation for dyadic data," *Political Analysis*, 23, 564–577.

Athey, S. and A. Schmutzler (2001): "Investment and market dominance," *RAND Journal of Economics*, 32, 1–26.

Bai, Y., A. Santos, and A. Shaikh (2019): "A practical method for testing many moment inequalities," *University of Chicago, Becker Friedman Institute for Economics Working Paper.*

Belloni, A. and V. Chernozhukov (2011): "ℓ_1-penalized quantile regression in high-dimensional sparse models," *Annals of Statistics*, 39, 82–130.

——— (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018): "Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework," *Annals of Statistics*, 46, 3643–3675.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," *RAND Journal of Economics*, 25, 242–262.

Bickel, P. J. and A. Chen (2009): "A nonparametric view of network models and Newman–Girvan and other modularities," *Proceedings of the National Academy of Sciences*, 106, 21068–21073.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, 37, 1705–1732.

Bühlmann, P. and S. van de Geer (2011): *Statistics for High-Dimensional Data*, Springer Series in Statistics, Springer, Heidelberg, methods, theory and applications.

Cameron, A. C. and D. L. Miller (2014): "Robust inference for dyadic data," *Unpublished manuscript, University of California-Davis.*

Cameron, C. A., J. B. Gelbach, and D. L. Miller (2011): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

Cameron, C. A. and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 50, 317–372.

Chen, X. (2018): "Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications," *Annals of Statistics*, 46, 642–678.

Chen, X. and K. Kato (2019): "Randomized incomplete U-statistics in high dimensions," *Annals of Statistics*, 47, 3127–3156.

——— (2020): "Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications," *Probability Theory and Related Fields*, 176, 1097–1163.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013a): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *Annals of Statistics*, 41, 2786–2819.

——— (2014): "Gaussian approximation of suprema of empirical processes," *Annals of Statistics*, 42, 1564–1597.

——— (2015): "Comparison and anti-concentration bounds for maxima of Gaussian random vectors," *Probability Theory and Related Fields*, 162, 47–70.

——— (2016): "Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings," *Stochastic Processes and their Applications*, 126, 3632–3651.

——— (2017a): "Central limit theorems and bootstrap in high dimensions," *Annals of Probability*, 45, 2309–2352.

——— (2017b): "Detailed proof of Nazarov's inequality," *arXiv:1711.10696.*

——— (2019a): "Inference on causal and structural parameters using many moment inequalities," *Review of Economic Studies*, 86, 1867–1900.

Chernozhukov, V., D. Chetverikov, K. Kato, and Y. Koike (2019b): "Improved Central Limit Theorem and bootstrap approximations in high dimensions," *arXiv:1912.10529.*

Chernozhukov, V., S. Lee, and A. M. Rosen (2013b): "Intersection bounds: Estimation and inference," *Econometrica*, 81, 667–737.

Chetverikov, D. (2018): "Adaptive tests of conditional moment inequalities," *Econometric Theory*, 34, 186–227.

Chiang, H. and Y. Sasaki (2019): "Lasso under multi-way clustering: Estimation and post-selection inference," *arXiv:1905.02107.*

Chiang, H. D., K. Kato, Y. Ma, and Y. Sasaki (2019): "Multiway cluster robust double/debiased machine learning," *arXiv:1909.03489.*

Davezies, L., X. D'Haultfoeuille, and Y. Guyonvarch (2018): "Asymptotic results under multiway clustering," *arXiv:1807.07925.*

——— (2020): "Empirical process results for exchangeable arrays," *Annals of Statistics*, forthcoming.

de la Peña, V. and E. Giné (1999): *Decoupling: From Dependence to Independence*, Springer.

Deng, H. and C.-H. Zhang (2020): "Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors," *Annals of Statistics*, forthcoming.

Eagleson, G. K. and N. C. Weber (1978): "Limit theorems for weakly exchangeable arrays," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, vol. 84, 123–130.

Fafchamps, M. and F. Gubert (2007): "The formation of risk sharing networks," *Journal of Development Economics*, 83, 326–350.

Fang, X. and Y. Koike (2020): "High-dimensional central limit theorems by Stein's method," *arXiv:2001.10917.*

Gandhi, A., Z. Lu, and X. Shi (2020): "Estimating demand for differentiated products with zeroes in market share data," *SSRN 3503565.*

Giraud, C. (2015): *Introduction to High-Dimensional Statistics*, vol. 139 of *Monographs on Statistics and Applied Probability*, CRC Press, Boca Raton, FL.

Graham, B. and A. de Paula (2019): *The Econometric Analysis of Network Data*, Academic Press.

Graham, B. S. (2019): "Network data," Tech. rep., National Bureau of Economic Research.

Graham, B. S., F. Niu, and J. L. Powell (2019): "Kernel density estimation for undirected dyadic data," *arXiv:1907.13630.*

——— (2020): "Minimax risk and uniform convergence rates for nonparametric dyadic regression," *arXiv:2012.08444.*

Head, K. and T. Mayer (2014): "Gravity equations: Workhorse, toolkit, and cookbook," in *Handbook of International Economics*, Elsevier, vol. 4, 131–195.

Hoover, D. (1979): "Relations on probability spaces and arrays of random variables," Working paper.

Kallenberg, O. (2006): *Probabilistic Symmetries and Invariance Principles*, Springer Science & Business Media.

Kato, K. and Y. Sasaki (2018): "Uniform confidence bands in deconvolution with unknown error distribution," *Journal of Econometrics*, 207, 129–161.

Koike, Y. (2019): "Gaussian approximation of maxima of Wiener functionals and its application to high-frequency data," *Annals of Statistics*, 47, 1663–1687.

Kuchibhotla, A. K., S. Mukherjee, and D. Banerjee (2020): "High-dimensional CLT: Improvements, non-uniform extensions and large deviations," *Bernoulli*, forthcoming.

Lecué, G. and S. Mendelson (2017): "Sparse recovery under weak moment assumptions," *Journal of the European Mathematical Society*, 19, 881–904.

Lee, A. J. (1990): *U-Statistics: Theory and Practice*, CRC Press.

Lee, S., K. Song, and Y.-J. Whang (2013): "Testing functional inequalities," *Journal of Econometrics*, 172, 14–32.

——— (2018): "Testing for a general class of functional inequalities," *Econometric Theory*, 34, 1018–1064.

MacKinnon, J. G. (2019): "How cluster-robust inference is changing applied econometrics," *Canadian Journal of Economics*, 52, 851–881.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2020): "Wild bootstrap and asymptotic inference with multiway clustering," *Journal of Business and Economic Statistics*, forthcoming.

McCullagh, P. (2000): "Resampling and exchangeable arrays," *Bernoulli*, 6, 285–301.

Menzel, K. (2016): "Inference for games with many players," *Review of Economic Studies*, 83, 306–337.

——— (2017): "Bootstrap with clustering in two or more dimensions," *arXiv:1703.03043.*

Morales, E., G. Sheu, and A. Zahler (2019): "Extended Gravity," *Review of Economic Studies*, 86, 2668–2712.

Moulin, H. (1988): *Axioms of Cooperative Game Theory*, New York: Cambridge University Press.

Owen, A. B. (2007): "The pigeonhole bootstrap," *Annals of Applied Statistics*, 1, 386–411.

Owen, A. B. and D. Eckles (2012): "Bootstrapping data arrays of arbitrary order," *Annals of Applied Statistics*, 6, 895–927.

Rudelson, M. and R. Vershynin (2008): "On sparse reconstruction from Fourier and Gaussian measurements," *Communications on Pure and Applied Mathematics*, 61, 1025–1045.

Silverman, B. W. (1976): "Limit theorems for dissociated random variables," *Advances in Applied Probability*, 8, 806–819.

Tabord-Meehan, M. (2019): "Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic," *Journal of Business & Economic Statistics*, 37, 671–680.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Wainwright, M. J. (2019): *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Zhang, D. and W. B. Wu (2017): "Gaussian approximation for high dimensional time series," *Annals of Statistics*, 45, 1895–1919.

Zhang, X. and G. Cheng (2018): "Gaussian approximation for high dimensional vector under physical dependence," *Bernoulli*, 24, 2640–2675.

(H. D. Chiang) Department of Economics, University of Wisconsin-Madison, William H. Sewell Social Science Building, 1180 Observatory Drive, Madison, WI 53706, USA.

Email address: hdchiang@wisc.edu

(K. Kato) Department of Statistics and Data Science, Cornell University, 1194 Comstock Hall, Ithaca, NY 14853, USA.

Email address: kk976@cornell.edu

(Y. Sasaki) Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA.

Email address: yuya.sasaki@vanderbilt.edu