Received: 2 January 2024 | Revised: 7 July 2025 | Accepted: 4 September 2025

DOI: 10.1111/sjos.70021

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Projection-based estimators for matrix/tensor-valued data

Joni Virta[1] | **Stanislav Nagy**[2] | **Klaus Nordhausen**[3,4]

[1]Department of Mathematics and Statistics, University of Turku, Turku, Finland

[2]Department of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic

[3]Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

[4]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Correspondence
Joni Virta, Department of Mathematics and Statistics, University of Turku, Turku, Finland.
Email: joni.virta@utu.fi

Funding information
Research Council of Finland, Grant/Award Numbers: 335077, 347501, 353769, 363261; Grantová Agentura České Republiky, Grant/Award Number: 24-10822S; COST Action HiTEc, Grant/Award Number: CA21163; Ministry of Education, Youth and Sport of the Czech Republic, Grant/Award Number: LL2407

Abstract

A general approach for extending estimators to matrix- and tensor-valued data is proposed. The extension is based on using random projections to project out dimensions of a tensor and then computing a multivariate estimator for each projection. The mean of the obtained set of estimates is used as the final, joint estimate. In some basic cases, the resulting estimator can be given a closed form, and particular ones are shown to coincide with existing methodology. We derive sufficient conditions for the consistency and limiting normality of the resulting estimators under weak assumptions. In particular, limiting normality is retained as soon as the number of projections grows super-linearly in the sample size, and consistency is achieved regardless of the growth rate. Comparisons with competing methods show that the extensions prove useful in extracting components for classification and yield an efficient estimator for sufficient dimension reduction.

KEYWORDS

consistency, limiting normality, matrix-valued data, resampling

1 | INTRODUCTION

1.1 | Main idea behind the projection extension

Let $\mathbf{X}_1,\ldots,\mathbf{X}_n \in \mathbb{R}^{p\times q}$ be a sample of random matrices. As with regular multivariate data, the analysis of matrix data should begin with the computation of several descriptive statistics and summary measures that allow a better understanding of the structure of the sample. For example, one might be interested in quantifying the variability/skewness/correlation structure/shape/etc. of the $\mathbf{X}_i$. And, while carrying out these tasks for regular multivariate data would be straightforward, the situation is very different for the analysis of matrix data, where none of the previous measures admit a canonical, established estimator.

The purpose of this work is to address the previous problem by devising a heuristic that allows generalizing multivariate estimators to be applicable to matrix data. This is achieved by combining two prevalent ideas used in contemporary statistics. The first one is considering the elements of the observations $\mathbf{X}_i$ not in isolation, but in an aggregate way, through the corresponding columns and rows. If a natural matrix structure for the observed data exists, focusing solely on the individual elements and their interactions would mean ignoring this structure and "forgetting" which elements originated, e.g., from the same column. This fact is often overlooked when matrix-valued data is vectorized to yield vector-valued observations, to be able to use standard multivariate methodology. Consider, e.g., medical imaging data, where either the spatial or temporal dimension of the data is in standard practice considered to consist of independent replicates to analyze the other dimension, see spatial and temporal independent component analysis (ICA) (Calhoun et al., 2003). The concept of a column-row structure is also working behind the Kronecker covariance structure $\mathbf{\Sigma} = \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$, where $\otimes$ denotes the Kronecker product. This covariance structure is the key idea behind many matrix-variate parametric distributions (Gupta & Nagar, 2018), and extensively studied in contemporary data analysis, see, e.g., Wirfält and Jansson (2014), Roś et al. (2016), Leng and Pan (2018). If a random vector $\mathbf{x} \in \mathbb{R}^{pq}$ has a covariance matrix of this form with $\mathbf{\Sigma}_1 \in \mathbb{R}^{p\times p}$ and $\mathbf{\Sigma}_2 \in \mathbb{R}^{q\times q}$, then it has a natural matrix form $\mathbf{X} \in \mathbb{R}^{p\times q}$ and $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ can be interpreted as characterizing the variabilities of the columns and the rows, respectively, of $\mathbf{X}$. The link between $\mathbf{X}$ and $\mathbf{x}$ can then be stated most succinctly using the vectorization operator as $\mathrm{vec}(\mathbf{X}) = \mathbf{x}$, where $\mathrm{vec}(\cdot)$ stacks the columns of its argument into a long vector.

The second idea we exploit is that of projections. Using projections onto low-dimensional subspaces to capture important aspects of high-dimensional data has a long tradition in statistics. In the classical projection pursuit (Friedman & Tukey, 1974; Huber, 1985) one chooses a "measure of interestingness" and finds low-dimensional projections that maximize this measure. Stahel (1981) and Donoho (1982) used projections to weigh observations in constructing robust estimators of multivariate location and scatter.

A more recent example is the random projection ensemble classification (Cannings & Samworth, 2017), where classification rules based on several random projections of the data are combined to yield a final estimate with controllable risk. Another modern example is found in the projective resampling of Li et al. (2008), where projections are used to extend a sufficient dimension reduction (SDR) method to multivariate responses. In SDR, the goal is to reduce the dimension of a predictor vector $\mathbf{x} \in \mathbb{R}^p$ without losing any information on the dependency between $\mathbf{x}$ and some univariate response $y \in \mathbb{R}$. The standard methodology does not admit to multivariate responses $\mathbf{y} \in \mathbb{R}^d$ but Li et al. (2008) circumvent this restriction by generating multiple projections $\mathbf{u}^\top\mathbf{y} \in \mathbb{R}$ of the response onto uniformly random unit vectors on the unit sphere, and using the univariate method on each projection separately, finally averaging the estimators.

Intuitively, in both of the previous examples, each of the projections captures some aspect of the structure of the original data, and the averaging combines this structural information into a single estimator. Our proposed method, motivated by the above ideas, is summarized in the next subsection, followed by an example use case in a particular context.

1.2 | Proposed method

Let S be a descriptive statistic computable from a p-variate sample. Our proposed approach for extending S to be applicable to the matrix sample $\mathbf{X}_1,\ldots,\mathbf{X}_n \in \mathbb{R}^{p\times q}$ works as follows: Denote the empirical measure of the sample by P_n and let $\mathcal{U}_2^{q-1}$ be the uniform probability distribution on the L_2 unit sphere $\mathbb{S}_2^{q-1}$ in $\mathbb{R}^q$. We first generate a random vector $\mathbf{u} \sim \mathcal{U}_2^{q-1}$ from the unit sphere and project the rows of each observation $\mathbf{X}_i$ onto $\mathbf{u}$, yielding a sample of p-variate random vectors, $\mathbf{X}_1\mathbf{u},\ldots,\mathbf{X}_n\mathbf{u} \in \mathbb{R}^p$, whose empirical distribution is denoted by $P_{n,\mathbf{u}}$ for a fixed $\mathbf{u}$. The projected observations $\mathbf{X}_i\mathbf{u}$ are weighted averages of the columns of the original observations $\mathbf{X}_i$ and the quantity $S(P_{n,\mathbf{u}})$ of the projected sample contains information also on the corresponding characteristic of the original matrix sample. Since $\mathbf{u}$ was chosen randomly, we further repeat the above procedure several times and call the average of the resulting quantities $S(P_{n,\mathbf{u}_j})$ the *column extension* of S. The name reflects the fact that the random projections act on the column space of $\mathbf{X}_i$. The corresponding row-wise extension can be derived analogously, and a generalization to tensor-valued data is also straightforward, as explained in Appendix A. Formal definition of the method is given later in Section 2.

We note that the column extension methodology is model-free in the sense that it does not presume any particular (parametric or non-parametric) probability distribution for the sample $\mathbf{X}_i$. Rather, what it estimates depends solely on the choice of the parent estimator S. In general, column extensions can be used for analogous tasks as the original S, e.g., if S can be used to form a classification rule for a vector sample, then its column extension can classify the observations $\mathbf{X}_1,\ldots,\mathbf{X}_n$ based on their column spaces (see Section 5), etc.

1.3 | A particular example scenario

To provide a concrete example, we next demonstrate the use of column extension with one particular choice of S. Assume that the sample $\mathbf{X}_1,\ldots,\mathbf{X}_n \in \mathbb{R}^{p\times q}$ is contaminated with a large amount of outliers and that we are interested in characterizing the variation of the sample in a robust, meaningful way. Numerous robust measures of variation exist for samples of random vectors, such as the M-estimates of scatter (Dümbgen et al., 2015) or the minimum covariance determinant (MCD) (Hubert et al., 2018). The latter estimator is defined as the $p \times p$ sample covariance matrix of a subset of $h < n$ observations in $\mathbb{R}^p$ with the smallest possible determinant, multiplied by a suitable consistency factor. For h proportional to $n/2$, MCD is known to enjoy a full 50% breakdown point. However, neither the M-estimates of scatter nor MCD readily generalize to samples of random matrices, and the more standard approaches, such as the maximum likelihood estimator (MLE) of the matrix normal distribution (Dutilleul, 1999), are highly non-robust. Hence, we next take S to be the MCD estimator, and use a column extension to apply it to our matrix sample.

For a uniformly random vector $\mathbf{u} \sim \mathcal{U}_2^{q-1}$ on the unit sphere, the quantity $S(P_{n,\mathbf{u}})$ now captures a particular aspect of the spread of the sample of matrices in a robust manner. Repeating this procedure for a random sample $\mathbf{u}_j \sim \mathcal{U}_2^{q-1}$, $j \in \{1,\ldots,J\}$ of unit vectors, we obtain the column

extension of the MCD as the combined estimate

$$S_J^*(P_n) = \frac{1}{J}\sum_{j=1}^{J} S\left(P_{n,\mathbf{u}_j}\right). \tag{1}$$

The number of independent projections J is supposed to grow to infinity as $n \to \infty$. As we will see, however, it is not always necessary that $J > n$. The estimator $S_J^*(P_n)$ measures the average, outlier-resistant variation of the columns of $\mathbf{X}_1,\ldots,\mathbf{X}_n$ and can be used in subsequent analyses much in the same way as the MCD is used in standard multivariate analysis, e.g., as a basis for robust principal component analysis. The same measure for the rows of the sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$, with analogous interpretations, can be obtained by applying the previous method to the transposed sample $\mathbf{X}_i^\top$.

To illustrate the above estimator in practice, we generate a sample of size $n = 200$ from the matrix normal distribution $\mathcal{N}_{4\times3}(\mathbf{0}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$, see, e.g., Gupta and Nagar (2018), where the row covariance matrix $\mathbf{\Sigma}_1 \in \mathbb{R}^{4\times4}$ and the column covariance matrix $\mathbf{\Sigma}_2 \in \mathbb{R}^{3\times3}$ are generated independently from the Wishart distributions $\mathcal{W}_4(4, \mathbf{I}_4)$ and $\mathcal{W}_3(3, \mathbf{I}_3)$, respectively. Here, of course, $\mathbf{I}_p$ stands for the $p \times p$ identify matrix. The data are then contaminated by replacing 20% of the sample with observations from $\mathcal{N}_{4\times3}(\mathbf{0}, \mathbf{\Sigma}_1^*, \mathbf{\Sigma}_2)$, where $\mathbf{\Sigma}_1^*$ is drawn from $\mathcal{W}_4(4, 100\mathbf{I}_4)$. Assume that we want to estimate $\mathbf{\Sigma}_1$ based on the contaminated sample. The scales of the parameters $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are confounded and to compare different estimators we use the distance $d(\mathbf{A}, \mathbf{B}) = \|\mathbf{A}_0 - \mathbf{B}_0\|$, where $\|\cdot\|$ denotes the Frobenius norm, and $\mathbf{A}_0 = \mathbf{A}/a_{11}$ and $\mathbf{B}_0 = \mathbf{B}/b_{11}$ have been scaled to have unit upper left corner elements.

The standard estimate for (the scaled) $\mathbf{\Sigma}_1$ is its maximum likelihood estimate $\hat{\mathbf{\Sigma}}_{1,\mathrm{MLE}}$, for which no closed-form solution exists, see Dutilleul (1999) for estimating equations and an iterative algorithm. The MLE has been implemented in the R-package `MixMatrix` (Thompson, 2019), and yields, over 100 replicates of the simulation, the average estimation error,

$$\mathrm{AVE}\{d(\hat{\mathbf{\Sigma}}_{1,\mathrm{MLE}}, \mathbf{\Sigma}_1)\} \approx 4.991.$$

On the other hand, our proposed projection approach (1) with $S = \mathrm{MCD}$ and $J = 100$ gives the estimate $\hat{\mathbf{\Sigma}}_{1,\mathrm{MCD}} \equiv S_J^*(P_n)$, with the significantly smaller average estimation error,

$$\mathrm{AVE}\{d(\hat{\mathbf{\Sigma}}_{1,\mathrm{MCD}}, \mathbf{\Sigma}_1)\} \approx 0.457.$$

Thus, $\hat{\mathbf{\Sigma}}_{1,\mathrm{MCD}}$ has approximately one-seventh of the error of $\hat{\mathbf{\Sigma}}_{1,\mathrm{MLE}}$. We also tried conducting the same experiment without the contamination, to see if $\hat{\mathbf{\Sigma}}_{1,\mathrm{MCD}}$ offers good performance also in the absence of outliers. The resulting average errors for the MLE and the MCD were 0.286 and 0.499, respectively. As expected, MLE is superior in this case (being the best unbiased estimator), but the column extension is not far behind and, contrary to the MLE, shows consistently good behavior in both scenarios (contamination or no contamination).

To summarize, with the simple cost of additional computational burden, the column extension methodology provided us with an accurate, robust estimator of matrix column variability, without the need for defining any new estimators specific to matrix data.

1.4 | Contents and the scope of the paper

The earlier example about the MCD was naturally completely arbitrary, and the projection extension technique could be applied to extend other statistics of interest S defined for a sample of

vectors to samples of matrices. Whatever the statistic, a natural question to ask is whether the estimate (1) retains some forms of convergence possessed by the parent estimator S, when $n, J \to \infty$ with suitable rates. In Section 3, we establish sufficient conditions under which convergences in probability and in distribution are preserved for $S_J^*(P_n)$. In particular, we show that no assumptions on the relative rates of n and J are needed for the convergence in probability, and that a limiting normal distribution for the estimator can be retained as soon as the number of projections J grows super-linearly in n. We also consider the special case where the parent estimator S is of such a simple form that the averaging over the J projections can be replaced by taking an expected value over $\mathcal{U}_2^{q-1}$, meaning that we can discard the parameter J altogether.

Note that the proposed way of constructing extensions is entirely heuristic and does not itself guarantee that the resulting extensions have any meaningful properties or interpretations. However, as evidenced by the examples throughout the paper, the extensions often turn out to be surprisingly efficient and have connections to existing methodology. Moreover, even if the extensions are not always perfect, the proposed methodology at the very least provides an initial guess on how a standard multivariate method might be extended to higher-order data. We also acknowledge that, for many of the example situations we consider, there exists specialized state-of-the-art methodology that can offer better (in the sense of accuracy, etc.) results than our proposal. However, our point is not to compete against those tailor-made estimators. Instead, our work shows that, even without knowing any of these specialized tools, the projection extension methodology lets one convert well-known multivariate methods into estimators that are simple to use and highly competitive against their custom-fit counterparts, which are possibly more complex, but at the same time limited in their scope solely to that singular problem.

The structure of the paper is as follows. In Section 2, we provide a formal definition for the proposed method of extending multivariate methods to arbitrary-dimensional tensor-valued data and go through short examples, applying the methodology to various familiar estimators. Section 3 develops the limiting properties of the extensions and, in particular, we give sufficient conditions on the parent method and the number of projections used to obtain consistency and a limiting normal distribution for the extension. Section 4 explores the limiting results through simulations and reveals that good results are achieved in practice already for a low number of projections. Section 5 contains three longer examples where the extensions are shown to lead into recently proposed tensor-valued methodology in the context of independent component analysis, and to estimators that outmatch their competitors in classification and sufficient dimension reduction. Finally, in Section 6 we end with a discussion. The proofs and derivations of all formulas are collected in the appendix.

2 | METHODOLOGY

2.1 | Notation and definitions

Throughout the following, we work in the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and assume that we observe an independent and identically distributed (i.i.d.) sample of random matrices, $\mathbf{X}_1, \ldots, \mathbf{X}_n$, with fixed size $p \times q$. Moreover, we assume that our interest lies in characterizing the behavior of the columns of the matrices.

Neither assumption is restrictive for the following reasons: First, if we are instead interested in the rows of $\mathbf{X}_i$, it is sufficient to apply the methodology to the transposed observations $\mathbf{X}_i^\top$. Second, if the data does not consist of random matrices but instead of random tensors of higher

order, tensor unfolding can be used to reduce the problem to the matrix case. We have provided the details in Appendix A, focusing, without loss of generality, only on samples of matrices in the main text.

By $\mathcal{D}(\mathcal{Q})$ we denote the set of all probability distributions taking values in a measurable space $\mathcal{Q}$. Let $\mathbf{X}$ be a random $p \times q$ matrix with a distribution $P \in \mathcal{D}(\mathbb{R}^{p\times q})$ and let $\mathbf{X}_1,\dots,\mathbf{X}_n$ be a random sample of matrices from P. The empirical measure of the sample is denoted by $P_n \in \mathcal{D}(\mathbb{R}^{p\times q})$. For a fixed vector $\mathbf{u} \in \mathbb{R}^q$, the measure of the random vector $\mathbf{Xu}$ is denoted by $P_{\mathbf{u}} \in \mathcal{D}(\mathbb{R}^p)$ and the empirical measure of the projected sample $\mathbf{X}_1\mathbf{u},\dots,\mathbf{X}_n\mathbf{u}$ by $P_{n,\mathbf{u}} \in \mathcal{D}(\mathbb{R}^p)$. Let $S : \mathcal{D}(\mathbb{R}^p) \to \mathcal{S}$ be the functional (estimator) that we are interested in extending to matrix data, taking values in a vector space $\mathcal{S}$. For convenience, we assume that $\mathcal{S}$ is a Euclidean space. Note that the actual dimensionality of $\mathcal{S}$ depends on the functional we are interested in. For example, in the introductory example with the MCD-estimator, we had $\mathcal{S} = \mathbb{R}^{p\times p}$.

Denote by $\mathbb{S}_2^{q-1} = \{\mathrm{x} \in \mathbb{R}^q : \|\mathrm{x}\| = 1\}$ the unit L_2-sphere and let $\mathbb{S}_1^{q-1} = \left\{\mathrm{x} = \left(x_1,\dots,x_q\right)^\top \in \mathbb{R}^q : \sum_{k=1}^q |x_k| = 1\right\}$ be the L_1-sphere in $\mathbb{R}^q$. For $\alpha = 1, 2$, also write $\mathbb{S}_{\alpha,+}^{q-1}$ for the intersection of $\mathbb{S}_\alpha^{q-1}$ with the positive orthant

$$\mathbb{S}_{\alpha,+}^{q-1} = \mathbb{S}_\alpha^{q-1} \cap \left\{\mathrm{x} \in \mathbb{R}^q : \min_{k\in\{1,\dots,q\}} x_k > 0\right\}.$$

Thus, e.g., $\mathbb{S}_{1,+}^{q-1}$ is the unit simplex,

$$\mathbb{S}_{1,+}^{q-1} = \left\{\mathrm{x} \in \mathbb{R}^q : \sum_{k=1}^q x_k = 1, \min_{k\in\{1,\dots,q\}} x_k > 0\right\}.$$

Finally, let $\mathcal{U}_\alpha^{q-1}$ and $\mathcal{U}_{\alpha,+}^{q-1}$ denote the uniform distributions on $\mathbb{S}_\alpha^{q-1}$ and $\mathbb{S}_{\alpha,+}^{q-1}$, respectively. While all the results of this paper are straightforward to generalize to L_α-spheres with any $\alpha \in (0, \infty]$, we consider here only the most important cases $\alpha = 1, 2$. Complete results facilitating such extensions, including expressions for moments and cross-moments of the uniform distributions $\mathcal{U}_\alpha^{q-1}$ and $\mathcal{U}_{\alpha,+}^{q-1}$ are collected in Appendix B.

2.2 | Projection extension

We next formulate the core of our methodology, the extension of S to samples of random matrices via averaging projections. In addition to the form introduced in Equation (1), we also consider the case where the averaging can be replaced with an expected value.

Definition 1. Let $S : \mathcal{D}(\mathbb{R}^p) \to \mathcal{S}$ and fix a projection distribution, $\mathcal{U} = \mathcal{U}_\alpha^{q-1}$ or $\mathcal{U} = \mathcal{U}_{\alpha,+}^{q-1}$, for $\alpha = 1$ or $\alpha = 2$.

1. The column extension of S is the functional $S^* : \mathcal{D}(\mathbb{R}^{p\times q}) \to \mathcal{S}$ with the action

$$S^*(P) = \mathcal{E}[S(P_{\mathbf{u}})], \tag{2}$$

where $\mathbf{u} \sim \mathcal{U}$ and the expected value $\mathcal{E}$ is taken with respect to $\mathbf{u}$.

2. The empirical column extension of S is the functional $S_J^* : \mathcal{D}(\mathbb{R}^{p\times q}) \to \mathcal{S}$ with the action

$$S_J^*(P) = \frac{1}{J}\sum_{j=1}^{J} S(P_{\mathbf{u}_j}), \tag{3}$$

where J is the number of projections and $\mathbf{u}_1,\ldots,\mathbf{u}_J$ is a random sample from $\mathcal{U}$.

While using the column extension (2) requires no computationally expensive projection resampling needed by the empirical column extension (3), the former can be computed only for the simplest of estimates, and outside of them, we must resort to the latter. The scope of the empirical column extension is wide, and there are, in general, no limitations as to what kind of estimators it can be applied to, as long as taking an average over a set of estimators makes computational sense.

Of course, the choice of the projection distribution, regardless of whether among $\mathcal{U}_\alpha^{q-1}$ and $\mathcal{U}_{\alpha,+}^{q-1}$ or not, has an impact on the properties of the column extension (despite the fact that we suppress the dependency in Definition 1). We next use several canonical examples of projection extensions, the column extensions of the mean vector and the covariance matrix, highlighting situations when choosing $\mathcal{U}_{1,+}^{q-1}$ (Example 1) and $\mathcal{U}_2^{q-1}$ (Examples 2 and 3) is quite natural.

Example 1 (mean vector). If the estimator S is sign-equivariant, in the sense that $S(P)$ changes sign when the distribution P is reflected in the origin, and $\mathbf{u} \sim \mathcal{U}_\alpha^{q-1}$, then it is easy to see that the column extension of S has to vanish. Thus, using the symmetric projection distributions $\mathcal{U}_\alpha^{q-1}$ makes little sense in the case of the mean vector $S = \mathrm{E}$ (i.e., the mean in $\mathbf{X}$), in which case $\mathcal{S} = \mathbb{R}^p$. Taking instead $\mathbf{u} \sim \mathcal{U}_{1,+}^{q-1}$, the moment expressions in Appendix B straightforwardly yield

$$S^*(P) = \mathcal{E}[\mathrm{E}(\mathbf{X}\mathbf{u})] = \frac{1}{q}\sum_{k=1}^{q} \mathrm{E}\mathbf{x}_k,$$

where $\mathbf{x}_k$ denotes the kth column of $\mathbf{X}$, $k \in \{1,\ldots,q\}$. The resulting column extension is the barycenter of the mean vectors of the columns of $\mathbf{X}$.

Example 2 (covariance matrix and principal component analysis). As variability is a sign-invariant quantity, it is more reasonable to use the projection distribution family $\mathcal{U}_\alpha^{q-1}$ for the extension of the covariance matrix. By Appendix B, the column extension of the covariance matrix $S = \mathrm{Cov}$, for which $\mathcal{S} = \mathbb{R}^{p\times p}$, is

$$\begin{aligned} S^*(P) &= \mathcal{E}\left\{\mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])\mathbf{u}\mathbf{u}^\top(\mathbf{X}-\mathrm{E}[\mathbf{X}])^\top\right]\right\} \\ &= \frac{\Gamma(3/\alpha)\Gamma(q/\alpha)}{\Gamma(1/\alpha)\Gamma((q+2)/\alpha)}\mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{X}-\mathrm{E}[\mathbf{X}])^\top\right], \end{aligned} \tag{4}$$

where, this time, the choice $\alpha = 2$ gives the best interpretable result

$$S^*(P) = \mathrm{Cov}^*(\mathbf{X}) = \frac{1}{q}\mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{X}-\mathrm{E}[\mathbf{X}])^\top\right],$$

the barycenter of the covariance matrices of the columns of $\mathbf{X}$. Let $\mathbf{U} \in \mathbb{R}^{p\times s}$ contain the s first eigenvectors of $\mathrm{Cov}^*[\mathbf{X}]$ and let $\mathbf{V} \in \mathbb{R}^{q\times t}$ contain the t first eigenvectors

of $\mathrm{Cov}^*[\mathbf{X}^\top]$, for some $s \leq p$, $t \leq q$. Then the "matrix principal components" of $\mathbf{X}$ are $\mathbf{U}^\top\mathbf{X}\mathbf{V} \in \mathbb{R}^{s\times t}$.

Recall that in principal component analysis (PCA), the observed random vector $\mathbf{x} \in \mathbb{R}^p$ is projected onto the eigenvectors of its covariance matrix corresponding to the s largest eigenvalues. The matrix extension in Example 2 is constructed by analogy: The columns (rows) of $\mathbf{X}$ are projected on the eigenvectors of the column (row) extension of the covariance matrix. The obtained matrix of principal components, $\mathbf{U}^\top\mathbf{X}\mathbf{V}$, has a total of st components and, depending on the sizes of p, q, s, t, can provide a significant dimension reduction without losing the matrix structure of the data. Note that choosing $\alpha = 1$ in Example 2 would yield equal principal components, simply changing only the scaling of the eigenvalues. The extension is not new and is equivalent to a classical tensor decomposition, the higher order singular value decomposition (HOSVD) (De Lathauwer et al., 2000), if one leaves the observation mode of the data tensor intact. The method was later rediscovered in its statistical form in Zhang and Zhou (2005) under the name $(2D)^2$PCA, in Li et al. (2010) as a preprocessing tool, and in Virta et al. (2016) under the name TPCA.

If one further standardizes the principal components, Example 2 also allows us to define standardization for matrix-valued random variables. The Mahalanobis standardization of a random vector $\mathbf{x}$ is defined as $(\mathrm{Cov}[\mathbf{x}])^{-1/2}(\mathbf{x} - \mathrm{E}[\mathbf{x}])$ where $(\mathrm{Cov}[\mathbf{x}])^{-1/2}$ is a symmetric inverse square root of $\mathrm{Cov}[\mathbf{x}]$. The analogy for random matrices is thus the two-sided standardization,

$$(\mathrm{Cov}^*[\mathbf{X}])^{-1/2}(\mathbf{X} - \mathrm{E}[\mathbf{X}])\left(\mathrm{Cov}^*[\mathbf{X}^\top]\right)^{-1/2}. \tag{5}$$

This natural extension of the Mahalanobis standardization has been previously successfully used in Virta et al. (2017), where it was introduced as a "plug-in" estimator obtained by replacing $\mathbf{x}$ with $\mathbf{X}$ in the covariance formula $\mathrm{E}[(\mathbf{x} - \mathrm{E}[\mathbf{x}])(\mathbf{x} - \mathrm{E}[\mathbf{x}])^\top]$.

The projection extensions also readily apply to moments higher than first or second. The next example, which is proven in Appendix C, discusses the fourth moment in the context of projection pursuit (Huber, 1985), where projections with maximal higher moments are used to find non-Gaussian subspaces of interest.

Example 3 (independent component analysis). Given a suitably standardized random vector $\mathbf{x} \in \mathbb{R}^p$, the classical projection pursuit, in its common usage, maximizes the following objective function over the unit sphere $\mathbb{S}_2^{p-1}$,

$$g_4(\mathbf{a}) = \mathrm{E}\left[\left(\mathbf{a}^\top\mathbf{x}\right)^4\right].$$

Then, the column extension of $S = g_4$ for the projection distribution $\mathcal{U}_2^{q-1}$ has $\mathcal{S} = \mathbb{R}$, and takes the form

$$g_4^*(\mathbf{a}) = \frac{3}{q(q+2)}\mathrm{E}\left[\left(\mathbf{a}^\top\mathbf{X}\mathbf{X}^\top\mathbf{a}\right)^2\right].$$

The interpretation behind the extension in Example 3 is that the maximizer of g_4^* "on average" optimizes the original function over all possible projections. Besides explorative data analysis, projection pursuit is a standard tool in classical independent component analysis (ICA) (Comon & Jutten, 2010; Hyvärinen & Oja, 1997), where the objective function g_4 in Example 3 has been shown to produce consistent estimates of the model parameters (Miettinen et al., 2015). Thus, its extension provides a reasonable starting point for developing projection pursuit-based tools for matrix-valued ICA.

3 | ASYMPTOTIC RESULTS

In this section, we derive the asymptotic properties of the proposed methodology. Recall that $S : \mathcal{D}(\mathbb{R}^p) \to \mathcal{S}$ is a functional into a Euclidean space $\mathcal{S}$, which we equip with the Euclidean norm, $\|\cdot\|$ (i.e., the Frobenius norm when $\mathcal{S}$ is a space of matrices). For simplicity, the following results have been derived in the case of L_2-projections, $\mathbf{u} \sim \mathcal{U}_2^{q-1}$, but the generalization to any projection distribution in $\mathcal{U}_\alpha^{q-1}$ or $\mathcal{U}_{\alpha,+}^{q-1}$ follows straightforwardly. By almost all $\mathbf{u} \in \mathbb{S}_2^{q-1}$ we mean almost all $\mathbf{u}$ with respect to the spherical Lebesgue measure on $\mathbb{S}_2^{q-1}$.

Recall further the definitions of the column extension and the empirical column extension,

$$S^*(P) = \mathcal{E}[S(P_{\mathbf{u}})],$$
$$S_J^*(P) = \frac{1}{J}\sum_{j=1}^{J} S\left(P_{\mathbf{u}_j}\right).$$

We assume that $P \in \mathcal{D}(\mathbb{R}^{p\times q})$ is such that the estimator $S(P_{\mathbf{u}})$ is well-defined for all $\mathbf{u} \in \mathbb{S}_2^{q-1}$ and integrable in $\mathbf{u}$. In particular, we assume that the sample version of the estimator $S\left(P_{n,\mathbf{u}}\right) = S\left(P_{n,\mathbf{u}}(\omega)\right)$ is (jointly) measurable, and integrable as a function of $(\mathbf{u}, \omega)^\top \in \mathbb{S}_2^{q-1} \times \Omega$, where Ω is the underlying probability space.

Two questions of interest can now be formulated. Under which conditions do the observed column extension $S^*(P_n)$ and the observed empirical column extension $S_J^*(P_n)$ converge to the population column extension $S^*(P)$, and what forms of convergence can we expect? To answer these, we first formulate a set of conditions, the combinations of which are sufficient to yield various forms of convergence. Let $\beta \in (1, 2]$ be fixed in the following.

C$_1$. *Consistency*: For almost all $\mathbf{u} \in \mathbb{S}_2^{q-1}$, the following convergence in probability holds

$$\left\|S\left(P_{n,\mathbf{u}}\right) - S(P_{\mathbf{u}})\right\| \xrightarrow[n\to\infty]{\mathrm{P}} 0.$$

C$_2$. *Integrability I-a*:

$$\mathcal{E}\sup_{n=1,2,\ldots} \mathrm{E}\left\|S\left(P_{n,\mathbf{u}}\right) - S(P_{\mathbf{u}})\right\|^\beta < \infty.$$

C$_3$. *Integrability I-b*:

$$\mathcal{E}\mathrm{E}\|S(P_{\mathbf{u}})\|^\beta < \infty.$$

C$_4$. *Marginal convergence in distribution*: There exists a measurable random process $\xi(\mathbf{u})$ indexed by $\mathbf{u} \in \mathbb{S}_2^{q-1}$ such that for all $d = 1, 2, \ldots$ and all $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{S}_2^{q-1}$ fixed,

$$\sqrt{n}\left(S\left(P_{n,\mathbf{u}_1}\right) - S\left(P_{\mathbf{u}_1}\right), \ldots, S\left(P_{n,\mathbf{u}_d}\right) - S\left(P_{\mathbf{u}_d}\right)\right)^\top \xrightarrow[n\to\infty]{D} (\xi(\mathbf{u}_1), \ldots, \xi(\mathbf{u}_d))^\top.$$

C$_5$. *Integrability II*:

$$\sup_{\mathbf{u}\in\mathbb{S}_2^{q-1}} \sup_{n=1,2,\ldots} \mathrm{E}n^{\beta/2}\left\|S\left(P_{n,\mathbf{u}}\right) - S(P_{\mathbf{u}})\right\|^\beta < \infty.$$

The rate $\sqrt{n}$ in Conditions (C_4) and (C_5) could be replaced with another rate $h_n \to \infty, n \to \infty$, giving the corresponding rate then also in Theorem 1. Note that Conditions (C_1)–(C_5) are not independent—(C_5) implies (C_1), and similarly (C_5) with (C_3) imply (C_2).

The following theorem collects the various forms of convergence that can be shown to hold under combinations of Conditions (C_1)–(C_5).

Theorem 1. *Let* $\mathbf{X} \sim P \in \mathcal{D}(\mathbb{R}^{p\times q})$. *Then the following holds true*:

i. Under (C_1) *and* (C_2)

$$\mathrm{E}\|S^*(P_n) - S^*(P)\| \xrightarrow[n\to\infty]{} 0.$$

ii. Under (C_1), (C_2), (C_3) *and for* $J \to \infty$

$$\mathcal{E}\mathrm{E}\left\|S_J^*(P_n) - S^*(P)\right\| \xrightarrow[n\to\infty]{} 0.$$

iii. Under (C_4) *and* (C_5),

$$\sqrt{n}(S^*(P_n) - S^*(P)) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{E}[\xi(\mathbf{u})]. \tag{6}$$

iv. Under (C_4), (C_5) *and* $n = o\left(J^{2-2/\beta}\right)$,

$$\sqrt{n}\left(S_J^*(P_n) - S^*(P)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{E}[\xi(\mathbf{u})]. \tag{7}$$

Part (i) of Theorem 1 states that, under the required conditions, the column extension $S^*(P_n)$ converges to its population value in mean, and thus in probability, but similar result for the almost sure convergence could also be devised using a technique analogous to that of Nagy et al. (2016, Section 5.2), under appropriate integrability conditions. By part (iii), the column extension also has a limiting distribution, at the rate $\sqrt{n}$. Part (ii) states that the empirical column extension $S_J^*(P_n)$ similarly enjoys convergence in probability to $S^*(P)$ when the number of projections $J \to \infty$ at any rate. Finally, a limiting distribution is achieved also for $S_J^*(P_n)$, if $n = o\left(J^{2-2/\beta}\right)$. That is, under $\beta = 2$, the strongest form of Condition (C_5), it is sufficient that the number of projections grows super-linearly in n, $J = \mathcal{O}(n^{1+\delta})$, for some $\delta > 0$.

The setting of Theorem 1 is quite general and makes few assumptions on the form of the estimator S. At the cost of adding an extra assumption, the conditions required by the results of Theorem 1 can be made simpler. Namely, assuming that the sample version of the estimator, $S\left(P_{n,\mathbf{u}}\right)$, follows (for fixed $\mathbf{u} \in \mathbb{S}_2^{q-1}$) a central limit theorem for i.i.d. random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, Conditions ($C_1$)–($C_5$) reduce to a simple moment condition.

Theorem 2. *Let the functional* $S : \mathcal{D}(\mathbb{R}^p) \to \mathcal{S}$ *be given by a function* $\psi : \mathbb{R}^p \to \mathcal{S}$ *so that for* $\mathbf{X} \sim P \in \mathcal{D}(\mathbb{R}^{p\times q})$ *and* $\mathbf{u} \in \mathbb{S}_2^{q-1}$ *fixed we can write*

$$S\left(P_{n,\mathbf{u}}\right) = \frac{1}{n}\sum_{i=1}^{n}\psi(\mathbf{X}_i\mathbf{u}),$$

$$S(P_{\mathbf{u}}) = \mathrm{E}\psi(\mathbf{X}\mathbf{u}).$$

Suppose that

$$\sup_{\mathbf{u}\in\mathbb{S}_2^{q-1}} \mathrm{E}\|\psi(\mathbf{X}\mathbf{u})\|^2 < \infty. \tag{8}$$

Then all (C_1)–(C_5) *hold true with* $\beta = 2$.

To conclude the section, we revisit two of the earlier estimators, the covariance matrix and the fourth moment, and use Theorem 2 to verify that they satisfy Conditions (C_1)–(C_5). Proofs of both results can be found in Appendix C.

Example 2 (continued). Let $\mathbf{X} \sim P \in \mathcal{D}(\mathbb{R}^{p\times q})$ be such that $\mathrm{E}\|\mathbf{X}\|^4 < \infty$, and let $S = \mathrm{Cov}$ be the covariance matrix. Then, Conditions (C_1)–(C_5) and, consequently, all parts of Theorem 1 hold for S with $\beta = 2$.

Example 3 (continued). Let $\mathbf{X} \sim P \in \mathcal{D}(\mathbb{R}^{p\times q})$ be such that $\mathrm{E}\|\mathbf{X}\|^8 < \infty$ and let $S = g_4$ be the fourth moment of a random vector along a fixed projection $\mathbf{a} \in \mathbb{S}_2^{p-1}$. That is, for $P_{\mathbf{u}}$, the estimator S acts as

$$S(P_{\mathbf{u}}) = \mathrm{E}\left[\left(\mathbf{a}^\top\mathbf{X}\mathbf{u}\right)^4\right].$$

Then, Conditions (C_1)–(C_5) and, consequently, all parts of Theorem 1 hold for S with $\beta = 2$.

4 | SIMULATIONS

We illustrate the asymptotic behavior and effect of the number of projections J by continuing Examples 1 and 2. In both cases, we simulate n i.i.d. samples from $\mathcal{N}_{p\times q}(\mathbf{0}, \mathbf{I}_p, \mathbf{I}_q)$, considering all combinations of $p, q \in \{3, 10\}$. The sample sizes n under consideration are 100, 200, 400, 800, and 1,600, while the number of projections J is varied across all values of $J = n$ and $J = \lceil n^{1.2} \rceil$.

To allow interpreting the simulation results, we first compute the asymptotic covariance matrices associated with the estimators in Examples 1 and 2 when the data come from our simulation setting, i.e., the matrix normal distribution. Below, $\mathbf{1}_q \in \mathbb{R}^q$ denotes the q-vector of ones and $\|\cdot\|_F$ is the Frobenius norm.

Theorem 3. *Let* $\mathbf{X} \sim \mathcal{N}_{p\times q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, *for some* $\boldsymbol{\mu} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{p\times p}$ *and* $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{q\times q}$. *Then, for S as in Example 1 and* $\mathbf{u} \sim \mathcal{U}_{1,+}^{q-1}$, *we have, for* $n = o(J)$,

$$\sqrt{n}\left(S_J^*(P_n) - S^*(P)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}, \frac{1}{q^2}\mathbf{1}_q^\top\boldsymbol{\Sigma}_2\mathbf{1}_q \cdot \boldsymbol{\Sigma}_1\right). \tag{9}$$

We note that, as shown in Theorem 1, the same limiting behavior is obtained also if the empirical column extension $S_J^*(P_n)$ is replaced with the column extension $S^*(P_n) = (1/q)\overline{\mathbf{X}}\mathbf{1}_q$. In other words, the additional variability induced by the sampling of $\mathbf{u}_1, \ldots, \mathbf{u}_J$ has no effect asymptotically, as long as the number J of sampled directions is large enough in the sense that $n = o(J)$. A result analogous to Theorem 3 holds also in the setup of Example 2.

Theorem 4. *Let* $\mathbf{X} \sim \mathcal{N}_{p\times q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, *for some* $\boldsymbol{\mu} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{p\times p}$ *and* $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{q\times q}$. *Then, for S as in Example 2 and* $\mathbf{u} \sim \mathcal{U}_2^{q-1}$, *we have, for* $n = o(J)$,

$$\sqrt{n}\,\mathrm{vec}\big(S_J^*(P_n) - S^*(P)\big) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_{p^2}\left(\mathbf{0}, \frac{1}{q^2}\|\boldsymbol{\Sigma}_2\|_F^2(\mathbf{I}_{p^2} + \mathbf{K}_{p,p})(\boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_1)\right), \tag{10}$$

where $\mathbf{K}_{p,p}$ *denotes the* $p^2 \times p^2$ *commutation matrix* (see for example Magnus & Neudecker, 1985, for details).

The idea of the simulation study is to compute m estimates of the (vectorized) quantity $b := \sqrt{n}\big(S_J^*(P_n) - S^*(P)\big)$ to estimate the corresponding covariance matrix associated with b. Those estimated covariances are then compared to the theoretical covariances given on the right-hand sides of (9) and (10). To measure the closeness of these two quantities, we use as a performance criterion the Frobenius norm of the difference between the empirical covariance matrix and the theoretical one. For a given number of repetitions m, we have therefore quantities of the form

$$\|\widehat{\mathrm{Cov}}(b_i) - \mathrm{AsCov}(b)\|_F \geq 0,$$

where the empirical covariance matrix $\widehat{\mathrm{Cov}}(b_i)$ is computed based on m independent copies $b_1, \ldots, b_m$ of the vector b defined above. AsCov(b) stands for the asymptotic covariance matrices on the right-hand sides of (9) and (10), respectively. Optimally, the parameter m should be chosen as large as possible, and, hence, in our simulations, we choose $m = 20{,}000$, as this is expected to be sufficiently large to observe the behavior of the estimators as functions of n and J.

Figures 1 and 2 present these norms as functions of J for varying sample sizes n and matrix dimensions p, q. The range of J in the figures is limited to a maximum of 800, as beyond this point, the curves stabilize and run parallel to the x-axis. Note that the scales in the subplots are not all directly comparable, as the asymptotic covariance matrices have different sizes.

The results clearly demonstrate the asymptotic properties derived in Section 3, showing that, given a sufficiently large J, the norm of the difference reaches a roughly constant, positive value (the lines turn horizontal in Figures 1 and 2), and already a J-value of 400 seems sufficient to achieve this. Our experiments (not shown here) revealed that the y-axis intercepts of these asymptotes are controlled by the number of repetitions m and approach zero as $m \to \infty$, meaning that, to obtain still better approximations, m would need to be further increased. For these two estimators, the effect of n appears negligible, which simply means that the current choices of n are large enough for the current value of m serving as the "bottleneck". The largest visual differences in the plots arise from the matrix dimensions. That is, larger matrices pose greater estimation errors, but this is expected since our criterion measures the absolute and not the relative difference between the two covariance matrices.

To demonstrate that our results are not dependent on the data being normally distributed, we present analogous figures in Appendix D, where the data is generated from a matrix-variate t_5-distribution. However, since Theorems 3 and 4 do not apply in this case, the corresponding theoretical covariance matrices were obtained via simulations based on the direct matrix extension of the respective estimators. The results are similar to those for the normal distribution, with the curves stabilizing for J-values in comparable ranges as above.

It is important to note that Examples 1 and 2 were selected because they are analytically tractable and allow for closed-form solutions for the column extensions, facilitating the comparisons in the first place. However, this also means that the example is not very practically relevant, as, if one has the closed-form solution, then there is little need to work with the corresponding

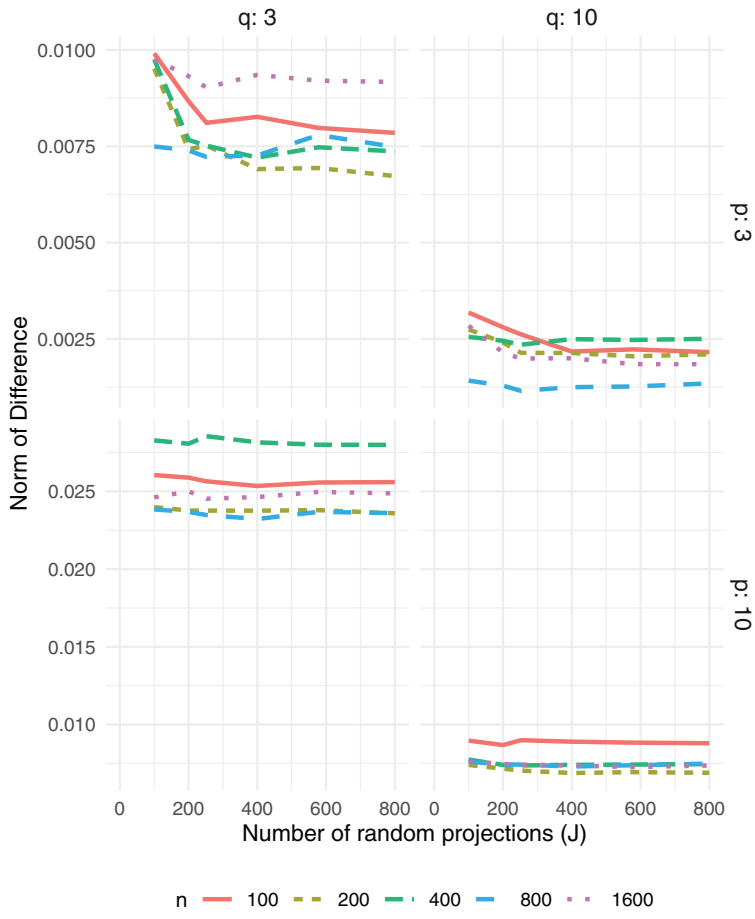

FIGURE 1 Simulation results based on $m = 20,000$ for Example 1 when $\mathbf{X} \sim \mathcal{MN}_{p\times q}(\mathbf{0}, \boldsymbol{I}_p, \boldsymbol{I}_q)$. The random projections follow the distribution $\mathcal{U}_{1,+}^{q-1}$.

empirical column extension. This is because, besides being an approximation, the latter method also carries a much higher computational overhead, as illustrated in Appendix D for Example 2, where average computation times are provided for settings similar to those considered above. Examples where our proposed methodology is more meaningful are presented in the next section.

5 | APPLICATIONS

We now provide three more detailed examples in which the proposed methodology is shown to lead to familiar established methods and yields estimators outmatching their competitors.

5.1 | Kernel principal component analysis

Kernel principal component analysis (KPCA) (Schölkopf et al., 1998) is a non-linear extension of PCA which uses the so-called kernel trick to implicitly perform principal component

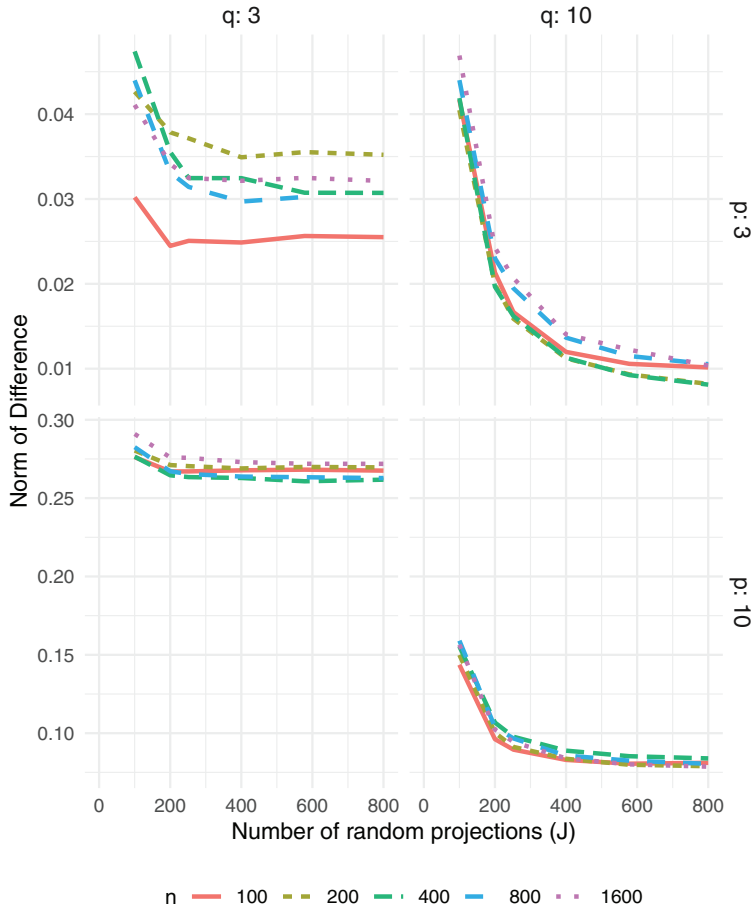

FIGURE 2 Simulation results based on $m = 20,000$ for Example 2 when $\mathbf{X} \sim \mathcal{MN}_{p\times q}(\mathbf{0}, \boldsymbol{I}_p, \boldsymbol{I}_q)$. The random projections follow the distribution $\mathcal{U}_2^{q-1}$.

analysis in a very high-dimensional, or infinite-dimensional, space. The jump to spaces of higher dimension is especially beneficial in classification and can sometimes enable linear separation in data that were not linearly separable in a lower dimension (Cristianini & Shawe-Taylor, 2000).

We next briefly go through the basic motivation behind KPCA. Assuming centered observations, $\mathbf{x}_i \in \mathbb{R}^p$, $i \in \{1,\ldots,n\}$, the sample principal components are found by projecting the observations onto the eigenbasis of the sample covariance matrix $\mathbf{S} = (1/n)\mathbf{X}^\top\mathbf{X}$, where $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^\top$. The first principal component is then $\mathbf{Xu}$ where the dominant eigenvector $\mathbf{u} \in \mathbb{R}^p$ satisfes $\mathbf{Su} = \lambda\mathbf{u}$ where λ is the dominant eigenvalue. If we premultiply the eigenequation by $\mathbf{X}$, we obtain

$$\left(\mathbf{XX}^\top\right)\mathbf{Xu} = (n\lambda)\mathbf{Xu},$$

showing that the principal components $\mathbf{Xu}$ can also be seen as eigenvectors of the kernel matrix $\mathbf{K} = \mathbf{XX}^\top = \left(\mathbf{x}_{i_1}^\top\mathbf{x}_{i_2}\right)_{i_1 i_2}$ containing the dot products of the observation vectors.

Assume then that we have some, possibly non-linear, transformation $\phi : \mathbb{R}^p \to \mathcal{M}$ of the data vectors $\mathbf{x}_i$ into a high-dimensional Hilbert space $\mathcal{M}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$. If we knew the kernel matrix $\mathbf{K}_{\mathcal{M}} = (\langle \phi(\mathbf{x}_{i_1}), \phi(\mathbf{x}_{i_2}) \rangle_{\mathcal{M}})_{i_1,i_2}$ of the data in $\mathcal{M}$, we could now obtain the principal components in $\mathcal{M}$ by simply extracting the eigenvectors of $\mathbf{K}_{\mathcal{M}}$. The idea behind the kernel trick is to approach this situation backwards: We first choose a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ and then define $\mathcal{M}$ to be a space where the inner product satisfies $\langle \phi(\mathbf{x}_{i_1}), \phi(\mathbf{x}_{i_2}) \rangle_{\mathcal{M}} = k(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$, leaving the transformation ϕ completely implicit. If k is positive semi-definite, the existence of such $\mathcal{M}$ and ϕ is guaranteed by Mercer's theorem, see, e.g., Cristianini and Shawe-Taylor (2000).

As non-linear transformations of centered observation vectors are generally not centered in the range space, one usually wants to use the centered kernel matrix $\overline{\mathbf{K}}_{\mathcal{M}} = \left(\langle \phi(\mathbf{x}_{i_1}) - \overline{\phi}, \phi(\mathbf{x}_{i_2}) - \overline{\phi} \rangle_{\mathcal{M}} \right)_{i_1,i_2}$, where $\overline{\phi} = (1/n) \sum_{i=1}^{n} \phi(\mathbf{x}_i)$. Using the properties of inner products and some simple algebra, we find that the centered kernel matrix can be expressed using the original kernel matrix as

$$\overline{\mathbf{K}}_{\mathcal{M}} = \left(\mathbf{I} - \frac{1}{n}\mathbf{J} \right) \mathbf{K}_{\mathcal{M}} \left(\mathbf{I} - \frac{1}{n}\mathbf{J} \right),$$

where $\mathbf{J} \in \mathbb{R}^{n \times n}$ has every element equal to 1 and $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix.

Having chosen a kernel function k, the column extension $\mathbf{K}^*_{\mathcal{M}}$ of the kernel matrix is simply the matrix with the (i_1, i_2)-element $\mathcal{E}\left\{ k(\mathbf{X}_{i_1}\mathbf{u}, \mathbf{X}_{i_2}\mathbf{u}) \right\}$. As the function k is usually non-linear, e.g., the Gaussian kernel $k(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \exp\{-\sigma^2 \|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\|^2\}$ is a common choice, the expected value $\mathcal{E}$ cannot generally be computed explicitly and we have to resort to the empirical column extension. Based on the heuristic suggestions in Section 2, we use the projection distribution $\mathcal{U}_2^{q-1}$.

To demonstrate this, we next applied KPCA to a simple classification example using the handwritten numeral data set available in the R-package `tensorBSS` (Virta et al., 2021). The training data set in the package consists of 7,291 instances of 16×16 matrices depicting grayscale numerals 0–9. Of those we considered, for simplicity, only the numerals 0,1,2,3, and further randomly chose 200 of those. A sample of the images is shown in Figure 3. Our goal is to use KPCA to extract a pair of principal components that provide classification information between the four groups. We distinguish three versions of KPCA, all using the Gaussian kernel with the tuning parameter $\sigma^2 = 0.05$: The empirical column extension KPCA, the empirical row extension KPCA, and the standard KPCA applied to vectorized images. For the extensions, we used $J = 100$ projection directions, and the implementation of KPCA was taken from the R-package `kernlab` (Karatzoglou et al., 2004). Since the proposed method is unsupervised, including a classification method, such as Thompson et al. (2020), as a competitor would be unfair. Rather, since the purpose of KPCA is preprocessing via dimension reduction to achieve an informative low-dimensional representation, we took, in the spirit of Samadi et al. (2017) and Billard et al. (2023), as our competing method tensorial PCA (TPCA), as defined in Zhang and Zhou (2005), Li et al. (2010), Virta et al. (2016) and implemented in the R-package `tensorBSS` (Virta et al., 2021).

The resulting bivariate scatter plots are shown in Figure 4, where both the column and row extension have nicely separated all four groups. However, this is not the case with the vectorized KPCA, whose scatter plot is largely dominated by a single group and where no clear separation between the remaining three digits is visible. This is possibly due to the large dimension of the original space, something that the column and row extensions avoid by considering not the individual elements of $\mathbf{X}$ but its rows and columns. As with the other methods, TPCA is also able to recover a projection that exactly separates the 1-digits from the others, but it leaves the remaining groups more or less overlapping. Hence, we conclude that "naturally matricial" methodology is

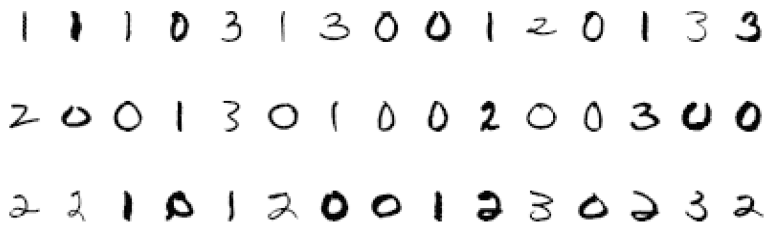

FIGURE 3 Example observations from the handwritten numeral data set.

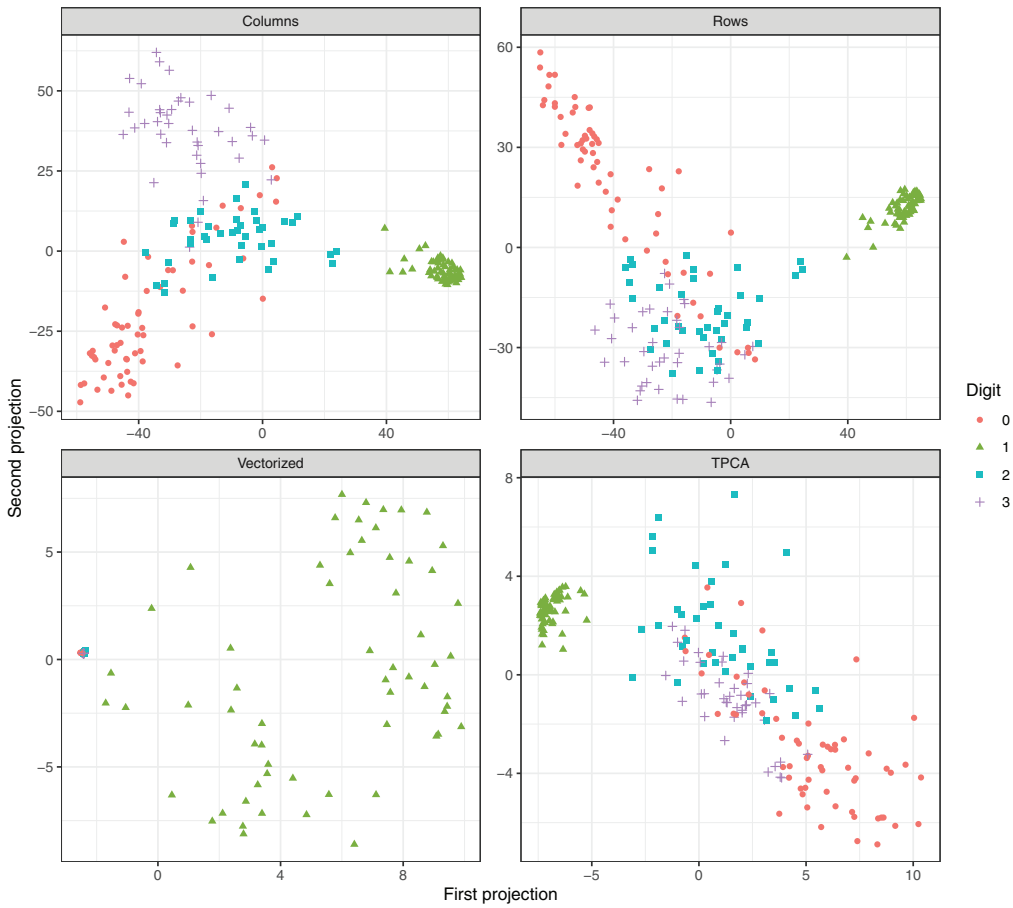

FIGURE 4 From left to right, the results of column extension KPCA, row extension KPCA, vectorization + KPCA, and TPCA.

not enough to find a good dimension reduction in the current example, and more elaborate techniques (such as non-linearity via kernels) are needed to accompany it. The proposed projection extension methodology then allows for a straightforward way of achieving this goal.

Finally, we also tried running the example with $J = 2,000$ projections instead of $J = 100$, and the resulting low-dimensional representations (not shown here) turned out to be essentially identical to Figure 4. This suggests that it is possible to obtain good approximations of the population column extensions already with $J < n$.

5.2 | Independent component analysis

Independent component analysis (ICA) is a classical branch of blind source separation (BSS) (Comon & Jutten, 2010) where we observe a random sample of centered random vectors from the independent component model:

$$\mathbf{x} = \mathbf{\Omega}\mathbf{z}, \tag{11}$$

where the invertible mixing matrix $\mathbf{\Omega} \in \mathbb{R}^{p\times p}$ is an unknown parameter and the latent vector $\mathbf{z}$ has mutually independent components and finite fourth moment, $\mathrm{E}\|\mathbf{z}\|^4 < \infty$. The objective in ICA is to estimate an inverse for the matrix $\mathbf{\Omega}$ and then recover the hidden $\mathbf{z}$.

The most basic method of solving the ICA problem is called fourth-order blind identification (FOBI) (Cardoso, 1989). The first step in FOBI is to standardize the observations as $\mathbf{x} \mapsto \mathbf{x}_{st} = (\mathrm{Cov}[\mathbf{x}])^{-1/2}\mathbf{x}$, after which it can be shown that,

$$\mathbf{x}_{st} = \mathbf{U}\mathbf{z}, \tag{12}$$

for some unknown orthogonal matrix $\mathbf{U} \in \mathbb{R}^{p\times p}$. Standardization, hence, reduces the problem of estimating an invertible $p \times p$ matrix to that of estimating only an orthogonal $p \times p$ matrix.

To estimate $\mathbf{U}$, FOBI introduces the matrix of fourth moments,

$$\mathbf{B}(\mathbf{x}) = \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top\right]. \tag{13}$$

By Miettinen et al. (2015, Theorem 7), if the kurtosis values of the components of $\mathbf{z}$ are distinct, then the orthogonal matrix $\mathbf{U}$ can be recovered from the eigendecomposition of the matrix $\mathbf{B}(\mathbf{x}_{st})$. Consequently, we obtain the solution $\mathbf{\Omega}^{-1} = \mathbf{U}^\top(\mathrm{Cov}[\mathbf{x}])^{-1/2}$.

In Virta et al. (2017), the concept of matrix-valued ICA was introduced. To extend the model (11), they separately applied mixing to the columns and rows of a matrix, reflecting the column-row paradigm discussed in Section 1. The obtained matrix independent component model reads

$$\mathbf{X} = \mathbf{\Omega}_1\mathbf{Z}\mathbf{\Omega}_2^\top, \tag{14}$$

where $\mathbf{\Omega}_1 \in \mathbb{R}^{p\times p}$, $\mathbf{\Omega}_2 \in \mathbb{R}^{q\times q}$ are invertible and the components of $\mathbf{Z}$ are mutually independent and have finite fourth moment, $\mathrm{E}\|\mathbf{Z}\|^4 < \infty$. Again, the aim is to recover the latent $\mathbf{Z}$ by estimating inverses for $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$.

By analogy to the vector-valued model, the first step towards solving (14) is to apply the column extension (5) of the Mahalanobis standardization to $\mathbf{X}$. Interestingly, this is exactly the same preprocessing step as was used in Virta et al. (2017) to formulate a method called TFOBI (for tensorial fourth-order blind identification), where its use was motivated simply by observing that it provides consistent estimates of the model parameters.

Denoting the obtained standardized random matrix by $\mathbf{X}_{st}$, Virta et al. (2017) proved that the column-row analogy of (12) holds for $\mathbf{X}_{st}$. Namely,

$$\mathbf{X}_{st} \propto \mathbf{U}_1\mathbf{Z}\mathbf{U}_2^\top,$$

for some unknown orthogonal matrices $\mathbf{U}_1 \in \mathbb{R}^{p\times p}$, $\mathbf{U}_2 \in \mathbb{R}^{q\times q}$. Next, they provided two extensions of the FOBI-matrix $\mathbf{B}$, motivated again by simply observing that they provide consistent

estimates,

$$\mathbf{B}_0(\mathbf{X}) = \frac{1}{q}\mathrm{E}\left[\mathbf{X}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\right], \quad \text{and} \quad \mathbf{B}_1(\mathbf{X}) = \frac{1}{q}\mathrm{E}\left[\|\mathbf{X}\|^2\mathbf{X}\mathbf{X}^\top\right].$$

The eigendecomposition of either can be used to estimate the missing orthogonal matrices $\mathbf{U}_1, \mathbf{U}_2$ under some additional assumptions. The next result, which is proven in Appendix C, shows that the above extensions are closely connected to the column extension of the FOBI-matrix $\mathbf{B}$.

Example 4 (fourth-order blind identification). The column extension of the matrix of fourth moments $S = \mathbf{B}$ from Equation (13) for the projection distribution $\mathcal{U}_2^{q-1}$ has $\mathcal{S} = \mathbb{R}^{p\times p}$ and takes the form

$$\mathbf{B}^*(\mathbf{X}) = \frac{1}{(q+2)}[2\mathbf{B}_0(\mathbf{X}) + \mathbf{B}_1(\mathbf{X})].$$

That is, the column extension of $\mathbf{B}$ is a linear combination of the matrices $\mathbf{B}_0$ and $\mathbf{B}_1$ used by Virta et al. (2017), and Example 4 goes to show that the methodology they constructed could have been simply obtained via the column extension of the standard FOBI.

We next provide a brief data example illustrating the column extension of ICA. As our data, we take the digit data set used earlier in Section 5.1 in the context of KPCA, but approach it this time from a different angle. We randomly take a total of $n = 400$ images of digits 1 and 7, such that the proportion $\varepsilon = 0.10$ of them are sevens and the remaining $1 - \varepsilon$ proportion ones. Our objective is to extract projections that allow detecting the outlying images (sevens), and the setting thus mimics a practical situation where the sevens have been "accidentally" included in the data set comprising otherwise of only images of the digit one, and one hopes to detect the outliers during descriptive analysis. We use two methods: The projection extension of FOBI defined in Example 4 and standard vector-valued FOBI preceded by reduction to 30 leading principal components via PCA. The preliminary reduction is necessary for the latter method, as otherwise we face an "$n < p$"-issue. Both methods were used to extract two projections, the one with the highest kurtosis and the one with the least kurtosis. Heuristically, these allow detecting possible unbalanced groups (highest kurtosis) and balanced groups (lowest kurtosis) in the data, see Corollary 2 in Peña and Prieto (2001).

The resulting bivariate scatter plots are shown in Figure 5 and reveal that the projection extension manages to identify a subspace where the one-digits take essentially a constant value, allowing a very accurate detection of the outliers. Vector ICA clearly attempts to estimate the same subspace but does not achieve a similarly minimal variation in the direction. Consequently, the detection of the seven-digit group would not be possible on the basis of the right panel of Figure 5, despite the overall similarity to the left panel. The reason for the difference in the methods' behavior is that the projection extension takes the matrix structure of the image into account in a natural way, allowing accurate modeling of the digit images that exhibit clear row-column patterns.

5.3 | Sufficient dimension reduction

A classical problem in multivariate statistics is that of *sufficient dimension reduction* (SDR), see Adragni and Cook (2009) for an overview of different classes of related methods. In SDR we observe a random vector $\mathbf{x} \in \mathbb{R}^p$ and a univariate response $y \in \mathbb{R}$ but we conjecture that the

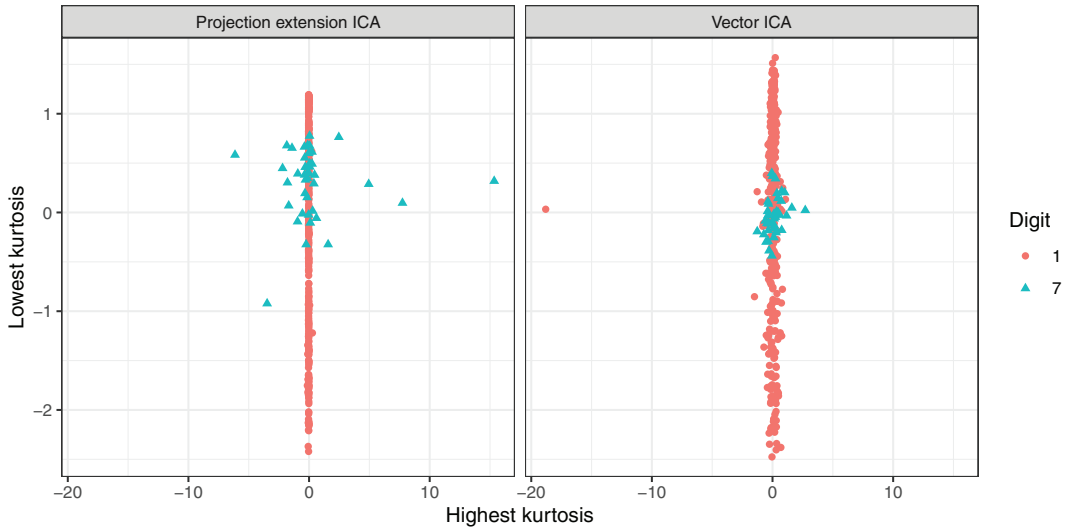

FIGURE 5 The scatter plots of the independent components with the highest and lowest kurtosis extracted from the digit data with projection extension ICA (left panel) and vector ICA (right panel). The coloring and shapes encode the two groups in the data (ones and sevens).

dependency between $\mathbf{x}$ and y is not truly p-dimensional and, instead, all predictive information in $\mathbf{x}$ can be captured with a projection onto a lower-dimensional subspace. That is, we assume there exists a matrix $\mathbf{A} \in \mathbb{R}^{p\times k}$, such that $\mathbf{x}$ and y are conditionally independent given $\mathbf{A}^\top\mathbf{x}$,

$$\mathbf{x} \perp y \quad | \mathbf{A}^\top\mathbf{x}.$$

If such a matrix is found, the subsequent analyses may then be carried out using the reduced k-dimensional $\mathbf{A}^\top\mathbf{x}$ in place of the original $\mathbf{x}$.

One of the very first and most famous SDR-methods is sliced inverse regression (SIR) (Li, 1991), where the matrix $\mathbf{A}$ is estimated simply by computing two eigendecompositions of moment-based matrices. Namely, we first standardize $\mathbf{x} \mapsto \mathbf{x}_{st} = (\mathrm{Cov}[\mathbf{x}])^{-1/2}(\mathbf{x} - \mathrm{E}[\mathbf{x}])$, and then find the first k eigenvectors $\mathbf{U} \in \mathbb{R}^{p\times k}$ of the so-called SIR matrix,

$$\mathbf{S}(\mathbf{x}_{st}; y) = \mathrm{E}\left[\mathrm{E}(\mathbf{x}_{st}|y)\mathrm{E}(\mathbf{x}_{st}|y)^\top\right].$$

The desired reduction matrix is then $\mathbf{A} = \mathbf{U}^\top(\mathrm{Cov}[\mathbf{x}])^{-1/2}$. To estimate the conditional expectations in practice, SIR resorts to an approximative procedure called "slicing" where the empirical distribution of y is discretized into h classes and the outer expected value is computed over this discrete distribution.

The ubiquitousness of matrix data caused Li et al. (2010) to introduce *dimension folding SIR*, a matrix-variate analogy of sliced inverse regression. In dimension folding, the goal is to find column and row reduction matrices $\mathbf{A} \in \mathbb{R}^{p\times k}$ and $\mathbf{B} \in \mathbb{R}^{q\times l}$ satisfying

$$\mathbf{X} \perp y \quad | \mathbf{A}^\top\mathbf{X}\mathbf{B},$$

meaning that the reduced predictor $\mathbf{A}^\top\mathbf{X}\mathbf{B}$ contains all the dependence between $\mathbf{X}$ and y compressed into just kl variables. The algorithm of Li et al. (2010) for finding the solution is

iterative in nature and requires the inversion of a $pq \times pq$ matrix, causing them to resort, for larger sized data, to a preliminary dimension reduction step which is equivalent to the column extension of PCA discussed in Section 2.

To provide a competing method to dimension folding, the standard SIR methodology is easily extended to random matrices using column extensions. First, we again carry out the standardization step as in Equation (5), producing the standardized random matrix $\mathbf{X}_{st}$. Second, the column extension of the SIR-matrix for the projection distribution $\mathcal{U}_2^{q-1}$ is straightforwardly (its structure, with respect to $\mathbf{X}$, is exactly the same as for the covariance matrix in Equation 4) seen to be

$$\mathbf{S}^*(\mathbf{X}; y) = \frac{1}{q}\mathrm{E}\left[\mathrm{E}(\mathbf{X}|y)\mathrm{E}(\mathbf{X}|y)^\top\right].$$

Thus, analogously to the vector case, we compute the k first eigenvectors $\mathbf{U} \in \mathbb{R}^{p\times k}$ of the discretized estimate of $\mathbf{S}^*(\mathbf{X}_{st}; y)$. The column reduction matrix is then $\mathbf{A} = \mathbf{U}^\top(\mathrm{Cov}^*[\mathbf{X}])^{-1/2}$, and similarly for the row reduction matrix $\mathbf{B}$. In the following, we refer to this procedure as TSIR (for "tensorial extension SIR") and to the dimension folding SIR of Li et al. (2010) as FSIR.

To compare the methods, we used TSIR to replicate the study done in Li et al. (2010) concerning an electroencephalography (EEG) data set freely available online (http://kdd.ics.uci.edu/databases/eeg/eeg.data.html), investigating genetic predisposition to alcoholism. The study contains 122 subjects, where 77 subjects are considered alcoholics and the remaining 45 subjects form a control group. Each subject was exposed to one of three different stimuli, and the voltage from 64 electrodes on the subject's scalp was sampled at 256 Hz for the duration of a single second. The measurements were replicated 120 times for each subject and stimulus, meaning that our data set is a fifth-order tensor of size $64 \times 256 \times 3 \times 120 \times 122$, where the dimensions are respectively electrodes, time, stimuli, replications, and subjects. As a preprocessing step, and to reduce the level of complexity, Li et al. (2010) considered only one of the stimuli and took the average over the replications, making their data set a third-order tensor of size $64 \times 256 \times 122$, or equivalently, 122 observations of 64×256 matrices. To make fair comparisons between the methods, we will also use this reduced form of the data.

Figure 6 shows the perspective plots of the typical recordings found in the two groups and clearly implies that differences between them exist. To condense the classification information contained by the recordings from the formidable 16,384 points per subject into something more manageable, Li et al. (2010) first used the procedure we introduced in Section 2 as the column extension of PCA to reduce the matrices to size 30×20, and then extracted a small number of components with FSIR. The performance of the method was evaluated by using leave-one-out cross-validation, where one observation was first removed to create a training set of 121 observations. Then, a quadratic discriminant analysis classification rule was fitted using two components (one row and two columns) extracted with FSIR from the training data, and this rule was used to predict the group of the removed observation. Repeating this procedure for each of the 122 observations, Li et al. (2010) correctly predicted 94 out of 122 cases using FSIR. Vectorization of the matrices followed by vector-valued SIR yielded only 92 correct predictions.

In our replication of the study, we also first reduced the observations to size 30×20 and then used TSIR to extract the first two diagonal components of $\mathbf{A}^\top\mathbf{X}\mathbf{B}$ (they should carry the most information out of any pair of components). The same cross-validation procedure then yielded us 97 correct predictions out of the total 122, putting TSIR ahead of FSIR (and on par with folded-DR, another method proposed in Li et al. 2010). The proportions of correct predictions per method have been summarized in Table 1. Additional merits of TSIR over those of FSIR are: First, it is not

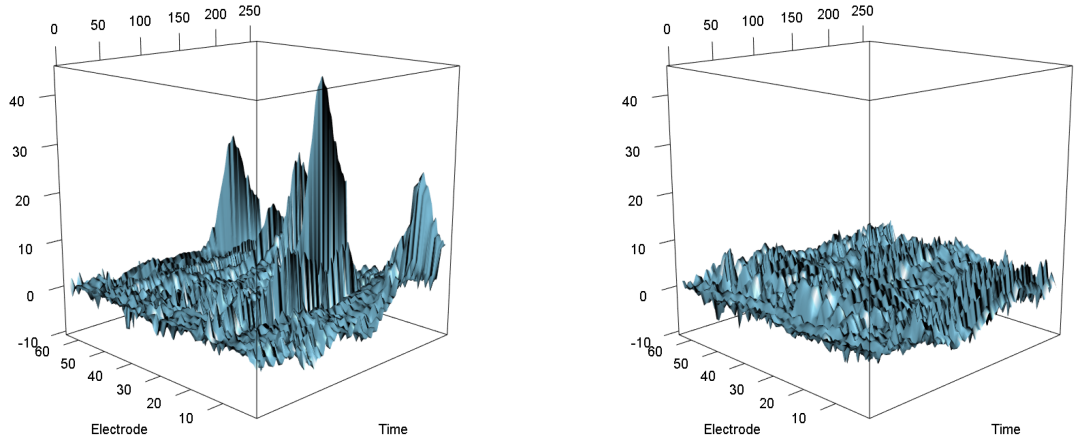

FIGURE 6 Perspective plots of the recordings of a typical subject in the alcoholic group (left) and in the control group (right).

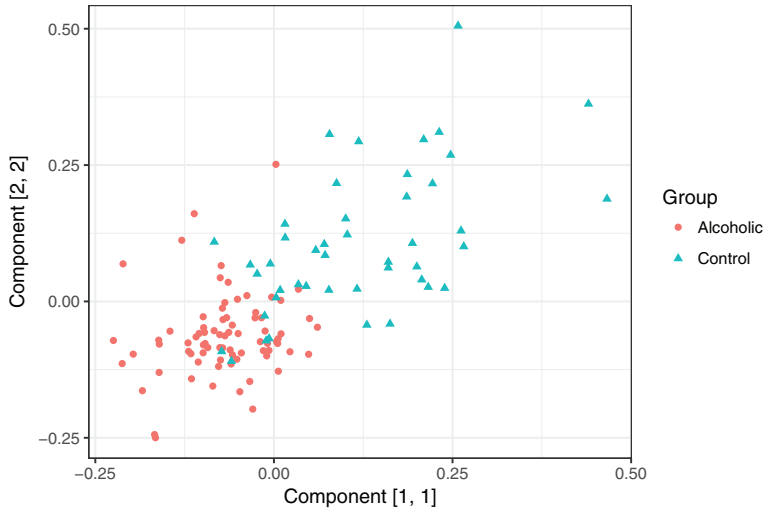

FIGURE 7 The two components found by TSIR.

iterative but instead based simply on eigendecompositions of $p \times p$ and $q \times q$ matrices. Second, it does not require the inversion of a $pq \times pq$ matrix, which, even after the preliminary dimension reduction, can be quite time-consuming and unstable. In Figure 7, we still show the scatter plot of the two components extracted by TSIR when the whole data set is used, and the plot indeed nicely exhibits the method's ability to separate the two groups.

As highlighted in Section 1.1, the use of random projections is not a novel concept in the context of SDR. In projective resampling SDR (PRSDR) of Li et al. (2008), it is assumed that the predictor $\mathbf{x} \in \mathbb{R}^p$ is vector-valued and the response $\mathbf{y} \in \mathbb{R}^d$ is also vector-valued with $d > 1$. PRSDR involves taking random projections of the multivariate response and then applying univariate SDR methods to these projections. The results are subsequently averaged across multiple projections. The corresponding SIR variant, PRSIR, is discussed in Li et al. (2008), and other

TABLE 1 The proportions of correct classifications out of 122 in the leave-one-out cross-validation study of the EEG data for four methods: FSIR (Li et al., 2010), the combination of vectorization and SIR (Li, 1991), our proposed TSIR, and folded-DR (Li et al., 2010).

Method	FSIR	Vec + SIR	TSIR	folded-DR
Accuracy	77.0%	75.4%	79.5%	79.5%

PRSDR methods employing various univariate techniques are discussed in Li et al. (2008), Chen et al. (2019), and reviewed in Dong et al. (2023).

This approach differs from ours, as we apply random projections to the predictors, which are matrix-valued, while assuming a univariate response. Nevertheless, combining our methodology with PRSDR could be an intriguing direction for future research in cases where $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{X} \in \mathbb{R}^{p \times q}$.

6 | DISCUSSION

In this paper, a general technique for extending multivariate methodology to tensor-valued random variables was proposed. The main idea behind the extension is based on the viewpoint that matrices should be approached via their rows and columns instead of the individual elements. Several examples of computing the extensions of various familiar methods were given, and the extensions were shown to have similar properties and uses as their vector-valued progenitors. Convergence properties of the extensions were studied both under a very broad class of parent estimators and for a subclass having a representation in the form of the central limit theorem, leading to simpler assumptions. Finally, applications to handwritten numeral data and EEG data showed that the extensions proved superior to some existing methods for classification purposes. Similarly, in ICA, it was shown in the example that the novel approach corresponds to an equivalent established tensorial ICA method. The beauty of the novel approach is that it allows the analysis in an appropriate way for any tensorial data, as long as a suitable corresponding multivariate method exists for vector-valued data.

ACKNOWLEDGMENTS

The authors are thankful to the two anonymous referees for their comments, which greatly helped improve the presentation and the quality of the manuscript. The authors also wish to acknowledge CSC—IT Center for Science, Finland, for computational resources. Open access publishing facilitated by Turun yliopisto, as part of the Wiley - FinELib agreement.

CONFLICT OF INTEREST STATEMENT

None of the authors have a conflict of interest to disclose.

FUNDING INFORMATION

This work was supported by Research Council of Finland (Grants 335077, 347501, 353769, and 363261), Czech Science Foundation (Grant 24-10822S), the Ministry of Education, Youth and Sport of the Czech Republic (ERC CZ grant LL2407), and COST Action HiTEc (CA21163).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Joni Virta https://orcid.org/0000-0002-2150-2769
Stanislav Nagy https://orcid.org/0000-0002-8610-4227
Klaus Nordhausen https://orcid.org/0000-0002-3758-8501

REFERENCES

Adragni, K. P., & Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*, 4385–4405.

Billard, L., Douzal-Chouakria, A., & Samadi, S. Y. (2023). Exploring dynamic structures in matrix-valued time series via principal component analysis. *Axioms*, *12*(6), 570.

Calhoun, V. D., Adali, T., Hansen, L. K., Larsen, J., & Pekar, J. J. (2003). *ICA of functional MRI data: An overview*. In *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation*. CiteSeer.

Cannings, T. I., & Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *79*(4), 959–1035.

Cardoso, J.-F. (1989). *Source separation using higher order moments*. In *IEEE International Conference on Acoustics, Speech, and Signal Processing 1989* (pp. 2109–2112). IEEE.

Chen, X., Yuan, Q., & Yin, X. (2019). Sufficient dimension reduction via distance covariance with multivariate responses. *Journal of Nonparametric Statistics*, *31*(2), 268–288.

Comon, P., & Jutten, C. (2010). *Handbook of blind source separation: Independent component analysis and applications*. Academic Press.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, *21*(4), 1253–1278.

Dong, Y., Soale, A.-N., & Power, M. D. (2023). A selective review of sufficient dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference*, *226*, 63–70.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*. Harvard University.

Dudley, R. M. (2002). *Real analysis and probability* (Cambridge Studies in Advanced Mathematics, 2nd ed.). Cambridge University Press. revised reprint of the 1989 original.

Dümbgen, L., Pauly, M., & Schweizer, T. (2015). M-functionals of multivariate scatter. *Statistics Surveys*, *9*, 32–105.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, *64*(2), 105–123.

Fang, K., Kotz, S., & Ng, K. (1990). *Symmetric multivariate and related distributions* (Monographs on Statistics and Applied Probability). Chapman and Hall.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *100*, 881–890.

Grinblat, L. Š. (1976). A limit theorem for measurable random processes and its applications. *Proceedings of the American Mathematical Society*, *61*(2), 371–376.

Gupta, A. K., & Nagar, D. K. (2018). *Matrix variate distributions*. Chapman and Hall/CRC Press.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, *13*, 435–475.

Hubert, M., Debruyne, M., & Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(3), e1421.

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483–1492.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab — An S4 package for kernel methods in R. *Journal of Statistical Software*, *11*, 1–20.

Lacko, V., & Harman, R. (2012). A conditional distribution approach to uniform sampling on spheres and balls in L_p spaces. *Metrika*, *75*(7), 939–951.

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer.

Leng, C., & Pan, G. (2018). Covariance estimation via sparse Kronecker structures. *Bernoulli*, *24*(4B), 3833–3863.

Li, B., Kim, M. K., & Altman, N. (2010). On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics*, *38*, 1094–1121.

Li, B., Wen, S., & Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, *103*, 1177–1186.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*, 316–327.

Magnus, J. R., & Neudecker, H. (1985). Matrix differential calculus with applications to simple, Hadamard, and Kronecker products. *Journal of Mathematical Psychology*, *29*(4), 474–492.

Miettinen, J., Taskinen, S., Nordhausen, K., & Oja, H. (2015). Fourth moments and independent component analysis. *Statistical Science*, *30*, 372–390.

Nagy, S., Gijbels, I., Omelka, M., & Hlubinka, D. (2016). Integrated depth for functional data: Statistical properties and consistency. *ESAIM Probability and Statistics*, *20*, 95–130.

Peña, D., & Prieto, F. J. (2001). Cluster identification using projections. *Journal of the American Statistical Association*, *96*(456), 1433–1445.

Roś, B., Bijma, F., de Munck, J. C., & de Gunst, M. C. (2016). Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure. *Journal of Multivariate Analysis*, *143*, 345–361.

Samadi, S. Y., Billard, L., Meshkani, M. R., & Khodadadi, A. (2017). Canonical correlation for principal components of time series. *Computational Statistics*, *32*(3), 1191–1212.

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.

Song, D., & Gupta, A. (1997). L_p-norm uniform distribution. *Proceedings of the American Mathematical Society*, *125*(2), 595–601.

Stahel, W. A. (1981). Breakdown of covariance estimators. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch.

Thompson, G. (2019). *MixMatrix: Classification with matrix variate normal and t distributions*. R package version 0.2.2. https://CRAN.R-project.org/package=MixMatrix

Thompson, G. Z., Maitra, R., Meeker, W. Q., & Bastawros, A. F. (2020). Classification with the matrix-variate-*t* distribution. *Journal of Computational and Graphical Statistics*, *29*(3), 668–674.

Tyler, D. E. (1982). Radial estimates and the test for sphericity. *Biometrika*, *69*(2), 429–436.

Virta, J., Koesner, C. L., Li, B., Nordhausen, K., Oja, H., & Radojicic, U. (2021). *TensorBSS: Blind source separation methods for tensor-valued observations*. R package version 0.3.8. https://CRAN.R-project.org/package=tensorBSS

Virta, J., Li, B., Nordhausen, K., & Oja, H. (2017). Independent component analysis for tensor-valued data. *Journal of Multivariate Analysis*, *162*, 172–192.

Virta, J., Taskinen, S., & Nordhausen, K. (2016). *Applying fully tensorial ICA to fMRI data*. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2016 IEEE* (pp. 1–6). IEEE.

von Bahr, B., & Esseen, C.-G. (1965). Inequalities for the *r*th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, *36*(1), 299–303.

Wirfält, P., & Jansson, M. (2014). On Kronecker and linearly structured covariance matrix estimation. *IEEE Transactions on Signal Processing*, *62*, 1536–1547.

Zhang, D., & Zhou, Z.-H. (2005). (2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, *69*(1–3), 224–231.

How to cite this article: Virta, J., Nagy, S., & Nordhausen, K. (2025). Projection-based estimators for matrix/tensor-valued data. *Scandinavian Journal of Statistics*, 1–35. https://doi.org/10.1111/sjos.70021

APPENDIX A. EXTENSION TO TENSOR DATA

We briefly describe here how the proposed column extension methodology can be applied to a sample of higher-order tensors $\mathcal{X}_i = (x^i_{\ell_1,\ldots,\ell_r}) \in \mathbb{R}^{p_1\times\cdots\times p_r}$. The main step in this is reducing the data to a sample of matrices by *unfolding* it, see De Lathauwer et al. (2000). To unfold a tensor, we first choose the *mode* or *way* of the tensor we want to inspect (a tensor of order r has r modes, analogous to a matrix having two, its columns and rows). Having fixed a mode m, we take the set of vectors

$$\left(x^i_{\ell_1,\ldots,\ell_{m-1},1,\ell_{m+1},\ldots,\ell_r},\ldots,x^i_{\ell_1,\ldots\ell_{m-1},p_m,\ell_{m+1},\ldots,\ell_r}\right),$$

for $\ell_s \in \{1,\ldots,p_s\}$, $s \in \{1,\ldots,r\}\setminus\{m\}$ and collect them into a wide matrix $\mathbf{X}_{i,(m)}$ of size $p_m \times \rho_m$ where $\rho_m = \prod_{s=1;s\neq m}^r p_s$. The matrix $\mathbf{X}_{i,(m)}$ is called the m-mode unfolding of $\mathcal{X}_i$ and the totality of our column extension theory can thus be applied to a sample of unfolded tensors to obtain estimators characterizing the mth mode of $\mathcal{X}$. Note that the order we collect the individual vectors to $\mathbf{X}_{i,(m)}$ is irrelevant (as long as it is consistent over the whole sample), as we will always project the rows of $\mathbf{X}_{i,(m)}$ onto a random vector $\mathbf{u} \in \mathbb{R}^{\rho_m}$ for which $\mathbf{u} \sim \mathbf{Pu}$ for any $\rho_m \times \rho_m$ permutation matrix $\mathbf{P}$.

APPENDIX B. UNIFORM DISTRIBUTIONS ON L_α-SPHERES AND THEIR MOMENTS

For $0 < \alpha \leq \infty$, denote by

$$\mathbb{S}_\alpha^{q-1} = \left\{\mathrm{x} \in \mathbb{R}^q : \|\mathrm{x}\|_\alpha = 1\right\},$$

the L_α unit sphere in $\mathbb{R}^q$, where the norm $\|\cdot\|_\alpha$ is defined as (for $\alpha < 1$ the triangle inequality does not hold true)

$$\|\mathrm{x}\|_\alpha = \begin{cases} \left(\sum_{k=1}^q |x_k|^\alpha\right)^{1/\alpha} & \text{for } \alpha < \infty, \\ \max_{k\in\{1,\ldots,q\}} |x_k| & \text{for } \alpha = \infty. \end{cases}$$

As in Section 2.1, by $\mathbb{S}_{\alpha,+}^{q-1}$ we denote the intersection of $\mathbb{S}_\alpha^{q-1}$ with the positive orthant

$$\mathbb{S}_{\alpha,+}^{q-1} = \mathbb{S}_\alpha^{q-1} \cap \left\{\mathrm{x} \in \mathbb{R}^q : \min_{k\in\{1,\ldots,q\}} x_k > 0\right\}.$$

For any $\alpha > 0$, let $\mathcal{U}_\alpha^{q-1}$ and $\mathcal{U}_{\alpha,+}^{q-1}$ denote the uniform distributions on $\mathbb{S}_\alpha^{q-1}$ and $\mathbb{S}_{\alpha,+}^{q-1}$, respectively. Samples from the distributions can be generated, e.g., using the algorithm of Lacko and Harman (2012).

Let $\mathbf{u} = (u_1,\ldots,u_q)^\top \sim \mathcal{U}_\alpha^{q-1}$ and $\mathbf{u}_+ = (u_{1,+},\ldots,u_{q,+})^\top \sim \mathcal{U}_{\alpha,+}^{q-1}$. By Song and Gupta (1997, Theorem 2.1) and Fang et al. (1990, Theorem 1.3), we know that for any $1 \leq k \leq q-1$ and any

non-negative integers r_i, $i = 1,\dots,k$,

$$\mathcal{E}\prod_{i=1}^{k} u_i^{r_i} = \begin{cases} 0 & \text{if some } r_i \text{ is odd,} \\ \frac{B_k(\mathbf{v}+\mathbf{r}/\alpha)}{B_k(\mathbf{v})} & \text{if all } r_i \text{ are even,} \end{cases}$$

$$\mathcal{E}\prod_{i=1}^{k} u_{i,+}^{r_i} = \frac{B_k(\mathbf{v}+\mathbf{r}/\alpha)}{B_k(\mathbf{v})},$$

where $\mathbf{v} = (1/\alpha,\dots,1/\alpha,(q-k)/\alpha)^\top \in \mathbb{R}^{k+1}$, $\mathbf{r} = (r_1,\dots,r_k,0)^\top \in \mathbb{R}^{k+1}$, and

$$B_k(\mathbf{x}) = \frac{\prod_{i=1}^{k+1}\Gamma(x_i)}{\Gamma\left(\sum_{j=1}^{k+1} x_j\right)} \quad \text{for } \mathbf{x} = (x_1,\dots,x_{k+1})^\top \in \mathbb{R}^{k+1}$$

is the multivariate beta function. In particular, we obtain that

$$\mathcal{E}u_i = 0,$$
$$\mathcal{E}u_i u_j = \delta_{ij}\frac{\Gamma(3/\alpha)\Gamma(q/\alpha)}{\Gamma(1/\alpha)\Gamma((q+2)/\alpha)},$$

where δ_{ij} is the Kronecker delta, and

$$\mathcal{E}u_{i,+} = \frac{\Gamma(2/\alpha)\Gamma(q/\alpha)}{\Gamma(1/\alpha)\Gamma((q+1)/\alpha)},$$
$$\mathcal{E}u_{i,+}u_{j,+} = \begin{cases} \frac{\Gamma(3/\alpha)\Gamma(q/\alpha)}{\Gamma(1/\alpha)\Gamma((q+2)/\alpha)} & \text{if } i = j, \\ \frac{\Gamma(2/\alpha)^2\Gamma(q/\alpha)}{\Gamma(1/\alpha)^2\Gamma((q+2)/\alpha)} & \text{if } i \neq j. \end{cases}$$

For the special cases $\alpha = 1$ and $\alpha = 2$ the preceding expressions simplify to:

- For $\alpha = 1$:

$$\mathcal{E}u_i = 0,$$

$$\mathcal{E}u_i u_j = \delta_{ij}\frac{2}{q(q+1)},$$

$$\mathcal{E}u_i u_j u_k u_l = \frac{4}{q(q+1)(q+2)(q+3)}\left(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + 3\delta_{ij}\delta_{ik}\delta_{il}\right),$$

and

$$\mathcal{E}u_{i,+} = 1/q,$$

$$\mathcal{E}u_{i,+}u_{j,+} = \frac{1+\delta_{ij}}{q(q+1)}.$$

- For $\alpha = 2$:

$$\begin{aligned}\mathcal{E}u_i &= 0,\\ \mathcal{E}u_iu_j &= \delta_{ij}\frac{1}{q},\\ \mathcal{E}u_iu_ju_ku_l &= \frac{1}{q(q+2)}\big(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}\big),\end{aligned}$$

and

$$\begin{aligned}\mathcal{E}u_{i,+} &= \frac{\Gamma(q/2)}{\sqrt{\pi}\,\Gamma((q+1)/2)},\\ \mathcal{E}u_{i,+}u_{j,+} &= \frac{1}{q}\Big(\frac{2}{\pi}\Big)^{1-\delta_{ij}}.\end{aligned}$$

APPENDIX C. PROOFS

Proof of Example 3. By definition, we have

$$g^*(\mathbf{a}) = \mathrm{E}\big[\mathbf{a}^\top\mathbf{X}\mathcal{E}\big(\mathbf{u}\mathbf{u}^\top\mathbf{X}^\top\mathbf{a}\mathbf{a}^\top\mathbf{X}\mathbf{u}\mathbf{u}^\top\big)\mathbf{X}^\top\mathbf{a}\big],$$

where the order of the expectation operators has been switched. Denoting $\mathbf{C} = \mathbf{X}^\top\mathbf{a}\mathbf{a}^\top\mathbf{X}$, the (k,l)-element of the inner expected value is $\sum_{ij} c_{ij}\mathcal{E}\big[u_iu_ju_ku_l\big]$, which then by the moment expressions in Appendix B simplifies to

$$\frac{1}{q(q+2)}(\delta_{kl}\mathrm{tr}(\mathbf{C}) + c_{kl} + c_{lk}).$$

As $\mathrm{tr}(\mathbf{C}) = \mathbf{a}^\top\mathbf{X}\mathbf{X}^\top\mathbf{a}$, the expected value has the matrix form

$$\frac{1}{q(q+2)}\big(\mathbf{a}^\top\mathbf{X}\mathbf{X}^\top\mathbf{a}\cdot\mathbf{I} + 2\mathbf{X}^\top\mathbf{a}\mathbf{a}^\top\mathbf{X}\big),$$

and plugging this in into $g^*(\mathbf{a})$ above yields the desired result. ∎

Proof of Theorem 1. *Proof of part (i).* From (C_1) and (C_2) we see that for almost all $\mathbf{u} \in \mathbb{S}_2^{q-1}$ we can write (Dudley, 2002, Theorem 10.3.6)

$$\mathrm{E}\Big\|S\big(P_{n,\mathbf{u}}\big) - S(P_{\mathbf{u}})\Big\| \xrightarrow[n\to\infty]{} 0. \tag{C1}$$

The estimator $S^*(P)$ can be for any $P \in \mathcal{D}(\mathbb{R}^{p\times q})$ written in the form

$$S^*(P) = \int_{\mathbb{S}_2^{q-1}} S(P_{\mathbf{u}})\ \mathrm{d}\ \sigma^{q-1}(\mathbf{u}), \tag{C2}$$

where by σ^{q-1} we understand the spherical Lebesgue measure in $\mathbb{R}^q$, i.e., the uniform probability measure on the unit sphere $\mathbb{S}_2^{q-1}$. By Jensen's inequality, we can bound

$$\mathrm{E}\|S^*(P_n) - S^*(P)\| \le \mathrm{E}\int_{\mathbb{S}_2^{q-1}} \left\|S\left(P_{n,\mathbf{u}}\right) - S(P_{\mathbf{u}})\right\| \,\mathrm{d}\, \sigma^{q-1}(\mathbf{u}).$$

Joint measurability of $S\left(P_{n,\mathbf{u}}\right)$ and the integrability condition (C_2) allow to use Fubini's theorem (Dudley, 2002, Theorem 4.4.5) to interchange the integrals, and (C1) together with the dominated convergence theorem (Dudley, 2002, Theorem 4.3.5) with the integrable majorant

$$\mathbf{u} \mapsto \sup_{n=1,2,\ldots} \mathrm{E}\left\|S\left(P_{n,\mathbf{u}}\right) - S(P_{\mathbf{u}})\right\|$$

to conclude.

Proof of part (ii). Write

$$\left\|S_J^*(P_n) - S^*(P)\right\| \le \left\|S_J^*(P_n) - S^*(P_n)\right\| + \|S^*(P_n) - S^*(P)\|.$$

By part (i), the second summand above vanishes in expectation as $n \to \infty$. As for the first summand, we can write

$$\begin{aligned}
\left(\mathcal{E}\mathrm{E}\left\|S_J^*(P_n) - S^*(P_n)\right\|\right)^\beta &\le \mathcal{E}\mathrm{E}\left\|\frac{1}{J}\sum_{j=1}^{J}\left(S\left(P_{n,\mathbf{u}_j}\right) - \mathcal{E}S\left(P_{n,\mathbf{u}}\right)\right)\right\|^\beta \\
&\le \frac{1}{J^\beta}\mathcal{E}\mathrm{E}\left(2J\left\|S\left(P_{n,\mathbf{u}_1}\right) - \mathcal{E}S\left(P_{n,\mathbf{u}}\right)\right\|^\beta\right) \\
&\le \frac{2}{J^{\beta-1}}\mathcal{E}\sup_{n=1,2,\ldots}\mathrm{E}\left\|S\left(P_{n,\mathbf{u}_1}\right) - \mathcal{E}S\left(P_{n,\mathbf{u}}\right)\right\|^\beta. \qquad \text{(C3)}
\end{aligned}$$

The first inequality above is Jensen's inequality (Dudley, 2002, 10.2.6) applied to the convex function $a \mapsto \|a\|^\beta$. The second inequality follows from von Bahr and Esseen (1965) who showed that in our setting

$$\mathrm{E}\mathcal{E}\left\|\sum_{j=1}^{J}\left(S\left(P_{n,\mathbf{u}_j}\right) - \mathcal{E}S\left(P_{n,\mathbf{u}}\right)\right)\right\|^\beta \le 2J\mathrm{E}\mathcal{E}\left\|S\left(P_{n,\mathbf{u}_1}\right) - \mathcal{E}S\left(P_{n,\mathbf{u}}\right)\right\|^\beta.$$

The formula above is a conditional version of Theorem 2 in von Bahr and Esseen (1965). By convexity of the function $a \mapsto \|a\|^\beta$ we also have that

$$\frac{1}{2^\beta}\|a+b\|^\beta = \left\|\frac{a+b}{2}\right\|^\beta \le \frac{\|a\|^\beta}{2} + \frac{\|b\|^\beta}{2}.$$

Now we use this formula to bound the expectation on the right-hand side of (C3)

$$\begin{aligned}
&\mathcal{E}\sup_{n=1,2,\ldots}\mathrm{E}\left\|S\left(P_{n,\mathbf{u}_1}\right) - \mathcal{E}S\left(P_{n,\mathbf{u}}\right)\right\|^\beta \\
&\le 2^{\beta-1}\left(\mathcal{E}\sup_{n=1,2,\ldots}\mathrm{E}\left\|S\left(P_{n,\mathbf{u}_1}\right) - S\left(P_{\mathbf{u}_1}\right)\right\|^\beta + \mathcal{E}\sup_{n=1,2,\ldots}\mathrm{E}\left\|S\left(P_{\mathbf{u}_1}\right) - \mathcal{E}S\left(P_{n,\mathbf{u}}\right)\right\|^\beta\right).
\end{aligned}$$

The first summand on the right-hand side is finite by (C_2). For the second summand, we have

$$\mathcal{E}\sup_{n=1,2,\ldots}\mathrm{E}\left\|S(P_{\mathbf{u}_1})-\mathcal{E}S(P_{n,\mathbf{u}})\right\|^{\beta}\le 2^{\beta-1}\left(\mathcal{E}\sup_{n=1,2,\ldots}\mathrm{E}\left\|S(P_{\mathbf{u}_1})\right\|^{\beta}+\sup_{n=1,2,\ldots}\mathrm{E}\left\|\mathcal{E}S(P_{n,\mathbf{u}})\right\|^{\beta}\right).$$

The first summand on the right-hand side is also finite, now by (C_3). In the second summand, Jensen's inequality and Fubini's theorem allow us to write

$$\sup_{n=1,2,\ldots}\mathrm{E}\left\|\mathcal{E}S(P_{n,\mathbf{u}})\right\|^{\beta}\le\mathcal{E}\sup_{n=1,2,\ldots}\mathrm{E}\left\|S(P_{n,\mathbf{u}})\right\|^{\beta},$$

where the right-hand side is finite by a combination of (C_2) and (C_3). All these derivations give that in Equation (C3) there exists a finite constant $c>0$ such that

$$\mathcal{E}\mathrm{E}\left\|S_J^*(P_n)-S^*(P_n)\right\|\le\frac{c}{J^{1-1/\beta}}. \tag{C4}$$

Because $\beta>1$ and $J\to\infty$, the conclusion of part (ii) of the theorem follows.

Proof of part (iii). The statement follows by a straightforward modification of Grinblat (1976, Theorem 3).

Proof of part (iv). Write

$$\sqrt{n}\left(S_J^*(P_n)-S^*(P_n)\right)+\sqrt{n}(S^*(P_n)-S^*(P)). \tag{C5}$$

The second summand above converges to the right-hand side of (7) by part (iii). Using (C4) from the proof of part (ii) we see that if $\sqrt{n}=o\left(J^{1-1/\beta}\right)$, then

$$\mathcal{E}\mathrm{E}\sqrt{n}\left\|S_J^*(P_n)-S^*(P_n)\right\|\xrightarrow[n\to\infty]{}0,$$

and the first summand in Equation (C5) is asymptotically negligible. The proof is finished. ∎

Proof of Theorem 2. Denote $\mathbf{Y}_i(\mathbf{u})=\psi(\mathbf{X}_i\mathbf{u})-\mathrm{E}\psi(\mathbf{X}\mathbf{u})$. First of all observe that

$$\mathrm{E}\|\mathbf{Y}_i(\mathbf{u})\|^2\le 4\mathrm{E}\|\psi(\mathbf{X}\mathbf{u})\|^2. \tag{C6}$$

To verify (C_4), one can use the basic multivariate central limit theorem. Indeed, for

$$\begin{pmatrix}S(P_{n,\mathbf{u}_1})\\ \cdots\\ S(P_{n,\mathbf{u}_d})\end{pmatrix}-\begin{pmatrix}S(P_{\mathbf{u}_1})\\ \cdots\\ S(P_{\mathbf{u}_d})\end{pmatrix}=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\psi(\mathbf{X}_i\mathbf{u}_1)\\ \cdots\\ \psi(\mathbf{X}_i\mathbf{u}_d)\end{pmatrix}-\begin{pmatrix}\mathrm{E}\psi(\mathbf{X}\mathbf{u}_1)\\ \cdots\\ \mathrm{E}\psi(\mathbf{X}\mathbf{u}_d)\end{pmatrix}=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{Y}_i(\mathbf{u}_1)\\ \cdots\\ \mathbf{Y}_i(\mathbf{u}_d)\end{pmatrix}$$

the moment condition (8), and (C6) imposed on each element $\mathbf{Y}_1(\mathbf{u}_j)$, $j=1,\ldots,d$, guarantee the validity of the multivariate central limit theorem (applied to a properly vectorized version of the tensor on the right-hand side above). Condition (C_4) is therefore satisfied.

It is easy to see that (8) implies (C_3). Further, we know that (C_5) and (C_3) imply (C_1) and (C_2). Thus, it is enough to verify (C_5).

We utilize the identity

$$\mathrm{E}\left|\sum_{i=1}^{n} y_i\right|^2 = \sum_{i=1}^{n} \mathrm{E}|y_i|^2$$

valid for real-valued independent random variables $y_1,\dots,y_n$ with $\mathrm{E}y_i = 0$ and $\mathrm{E}|y_i|^2 < \infty$ for all $i = 1,\dots,n$. Using this formula for all entries in the expression for the Frobenius norm of a matrix, one gets from the Jensen inequality and (C6)

$$\mathrm{E}n\left\|S\left(P_{n,\mathbf{u}}\right) - S(P_{\mathbf{u}})\right\|^2 = \frac{1}{n}\mathrm{E}\left\|\sum_{i=1}^{n}\mathbf{Y}_i(\mathbf{u})\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\|\mathbf{Y}_i(\mathbf{u})\|^2 \le 4\mathrm{E}\|\psi(\mathbf{X}\mathbf{u})\|^2.$$

Therefore, Condition (C_5) is satisfied for $\beta = 2$, and the theorem is proved. ∎

Proof of Example 2. Denote by $\mathbf{x}_1,\dots,\mathbf{x}_q \in \mathbb{R}^p$ the columns of the matrix $\mathbf{X}$. Then, for $\mathbf{u} = \left(u_1,\dots,u_q\right)^\top \in \mathbb{S}_2^{q-1}$,

$$S(P_{\mathbf{u}}) = \mathrm{Cov}\left[\sum_{j=1}^{q} u_j\mathbf{x}_j\right].$$

It is enough to verify (C_1)–(C_5) for centered columns of $\mathbf{X}$, i.e., when $\mathrm{E}\mathbf{x}_j = \mathbf{0}$ for all $j = 1,\dots,q$. The general case follows by application of the delta method. In our situation, the estimator $S(P_{\mathbf{u}})$ can be written as

$$S(P_{\mathbf{u}}) = \mathrm{E}\left(\sum_{j=1}^{q} u_j\mathbf{x}_j\right)\left(\sum_{k=1}^{q} u_k\mathbf{x}_k\right)^\top,$$

that is $\psi(\mathbf{X}\mathbf{u}) = (\mathbf{X}\mathbf{u})(\mathbf{X}\mathbf{u})^\top$. For any centered $\mathbf{X} \sim P \in \mathcal{D}(\mathbb{R}^{p\times q})$ and $\mathbf{u} \in \mathbb{S}_2^{q-1}$ we use convexity of $a \mapsto \|a\|^2$ to get

$$\|\psi(\mathbf{X}\mathbf{u})\|^2 = \left\|\sum_{j=1}^{q}\sum_{k=1}^{q}\left(\mathbf{x}_j u_j\right)\left(\mathbf{x}_k u_k\right)^\top\right\|^2 = \left\|\sum_{j=1}^{q}\sum_{k=1}^{q} u_j u_k \mathbf{x}_j\mathbf{x}_k^\top\right\|^2 \le q^2\sum_{j=1}^{q}\sum_{k=1}^{q}\left\|\mathbf{x}_j\mathbf{x}_k^\top\right\|^2.$$

Using Hölder's inequality we can write

$$\sup_{\mathbf{u}\in\mathbb{S}_2^{q-1}} \mathrm{E}\|\psi(\mathbf{X}\mathbf{u})\|^2 \le (pq^2)^2\mathrm{E}\|\mathbf{X}\|^4 < \infty.$$

Therefore, (8) is verified, and both Theorems 1 and 2 apply. ∎

Proof of Example 3. The proof follows using the same argument as in the proof of Example 2, and is thus omitted here. ∎

Proof of Theorem 3. As stated at the beginning of Section 3, it is straightforwardly shown that the equivalent of Theorem 1 holds in the current scenario with $\beta = 2$. Moreover, the random process $\xi(\mathbf{u})$ in Condition (C_4) is now $\mathbf{Zu}$, where the random matrix $\mathbf{Z}$ is such that $\sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{D} \mathbf{Z}$. By the standard CLT, $\mathrm{vec}(\mathbf{Z})$ has the distribution $\mathcal{N}_{pq}(\mathbf{0}, \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1)$, implying that

$$\mathcal{E}[\xi(\mathbf{u})] = \frac{1}{q}\mathbf{Z}\mathbf{1}_q = \frac{1}{q}\mathrm{vec}(\mathbf{Z}\mathbf{1}_q) = \frac{1}{q}(\mathbf{1}_q^\top \otimes \mathbf{I}_p)\mathrm{vec}(\mathbf{Z}),$$

leading to the desired limiting distribution. ∎

Proof of Theorem 4. As stated in Section 3, Theorem 1 holds in our current scenario. Furthermore, we may, without loss of generality, take $\boldsymbol{\mu} = \mathbf{0}$ and S to be the non-centered sample covariance matrix, since the effect of empirical centering on the asymptotic behavior of second moments is asymptotically negligible, see the arguing in Example 2.4.4 in Lehmann (1999). The random process $\xi(\mathbf{u})$ in Condition (C_4) is now such that

$$\mathrm{vec}\{\xi(\mathbf{u})\} = \mathbf{Z}(\mathbf{u} \otimes \mathbf{u}),$$

where $\mathbf{Z}$ satisfies

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \otimes \mathbf{X}_i - \mathrm{E}(\mathbf{X} \otimes \mathbf{X})\right) \xrightarrow[n\to\infty]{D} \mathbf{Z}.$$

Write next $\mathbf{X}_i = \boldsymbol{\Sigma}_1^{1/2}\mathbf{Y}_i\boldsymbol{\Sigma}_2^{1/2}$ where the $\mathbf{Y}_i$ are a random sample from the $p \times q$ standard matrix normal distribution. Then, by the standard CLT, $\mathrm{vec}(\mathbf{Z})$ has the p^2q^2-dimensional normal distribution with the covariance matrix

$$(\boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2})\mathrm{Cov}\{\mathrm{vec}(\mathbf{Y} \otimes \mathbf{Y})\}(\boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2}).$$

Using Lemma 3 from Magnus and Neudecker (1985), we obtain that

$$\mathrm{Cov}\{\mathrm{vec}(\mathbf{Y} \otimes \mathbf{Y})\} = (\mathbf{I}_q \otimes \mathbf{K}_{q,p} \otimes \mathbf{I}_p)\mathrm{Cov}\{\mathrm{vec}(\mathbf{Y}) \otimes \mathrm{vec}(\mathbf{Y})\}(\mathbf{I}_q \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_p),$$

where $\mathbf{K}_{p,q}$ is the (p,q)-commutation matrix. Furthermore, using Theorem 1 in Tyler (1982), we further get $\mathrm{Cov}\{\mathrm{vec}(\mathbf{Y}) \otimes \mathrm{vec}(\mathbf{Y})\} = \mathbf{I}_{p^2q^2} + \mathbf{K}_{pq,pq}$, giving

$$\mathrm{Cov}\{\mathrm{vec}(\mathbf{Y} \otimes \mathbf{Y})\} = (\mathbf{I}_q \otimes \mathbf{K}_{q,p} \otimes \mathbf{I}_p)(\mathbf{I}_{p^2q^2} + \mathbf{K}_{pq,pq})(\mathbf{I}_q \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_p).$$

Now, using the properties of commutation matrices, we obtain $\mathbf{K}_{q,p}\mathbf{K}_{p,q} = \mathbf{I}_{pq}$ and $(\mathbf{I}_q \otimes \mathbf{K}_{q,p} \otimes \mathbf{I}_p)\mathbf{K}_{pq,pq}(\mathbf{I}_q \otimes \mathbf{K}_{p,q} \otimes \mathbf{I}_p) = \mathbf{K}_{q,q} \otimes \mathbf{K}_{p,p}$, yielding the asymptotic covariance matrix of $\mathrm{vec}(\mathbf{Z})$ to be

$$(\boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2})(\mathbf{I}_{p^2q^2} + \mathbf{K}_{q,q} \otimes \mathbf{K}_{p,p})(\boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_2^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2}).$$

Finally, we have $\mathcal{E}[\mathrm{vec}\{\xi(\mathbf{u})\}] = (1/q)\mathbf{Z}\mathrm{vec}(\mathbf{I}_q)$ and arguing as in the proof of Theorem 3 shows that the desired limiting distribution is multivariate normal with the covariance matrix

$$\frac{1}{q^2}\{\mathrm{vec}(\boldsymbol{\Sigma}_2)^\top \otimes \boldsymbol{\Sigma}_1^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2}\}(\mathbf{I}_{p^2q^2} + \mathbf{K}_{q,q} \otimes \mathbf{K}_{p,p})\{\mathrm{vec}(\boldsymbol{\Sigma}_2) \otimes \boldsymbol{\Sigma}_1^{1/2} \otimes \boldsymbol{\Sigma}_1^{1/2}\}$$
$$= \frac{1}{q^2}\|\boldsymbol{\Sigma}_2\|_F^2(\mathbf{I}_{p^2} + \mathbf{K}_{p,p})(\boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_1).$$

■

Proof of Example 4. Again, by definition,

$$\mathbf{B}^*(\mathbf{X}) = \mathrm{E}\left[\mathbf{X}\mathcal{E}\left(\mathbf{u}\mathbf{u}^\top\mathbf{X}^\top\mathbf{X}\mathbf{u}\mathbf{u}^\top\right)\mathbf{X}^\top\right],$$

and, proceeding as in the proof of Example 3, we find that the inner expected value equals

$$\frac{1}{q(q+2)}\left[\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{I} + 2\mathbf{X}^\top\mathbf{X}\right].$$

Now, plugging this into $\mathbf{B}^*(\mathbf{X})$ above, we notice that $\|\mathbf{X}\|^2 = \mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\right)$ and are thus finished. ■

APPENDIX D. ADDITIONAL SIMULATION RESULTS

This appendix presents additional simulation results. Figures D1 and D2 serve as counterparts to Figures 1 and 2, illustrating the scenario where the data follow a matrix-variate t-distribution with 5 degrees of freedom (in Figures D1 and D2 we write simply $\mathbf{X} \sim t_5$). The distribution is scaled such that the row and column covariance matrices are identity matrices of the appropriate dimensions.

Since counterparts to Theorems 3 and 4 for this case are not available, the covariance matrices on the relevant right-hand sides are obtained through simulations with $s = 1{,}000{,}000$ repetitions. The results differ not much from the normal case shown in Section 4.

Figure D3 assesses the computational costs of the proposed approach for Example 2, using settings similar to those considered earlier. To this end, average computation times from 10 repetitions were measured on a Windows 10 laptop equipped with an Intel(R) Core(TM) Ultra 5 125U 1.30 GHz processor and 64 GB of RAM, utilizing the `microbenchmark` package in R 4.4.2. In this scenario, the direct estimator is well-defined, and the corresponding covariance matrix was computed using the `tensorBSS` package as a reference. As anticipated, our method—designed for cases where no direct extension of multivariate concepts to tensor-valued data exists—is slower than the direct extension. However, the additional computational costs are dependent on J. Importantly, for values such as $J = 400$ or $J = 800$, these costs remain entirely manageable.

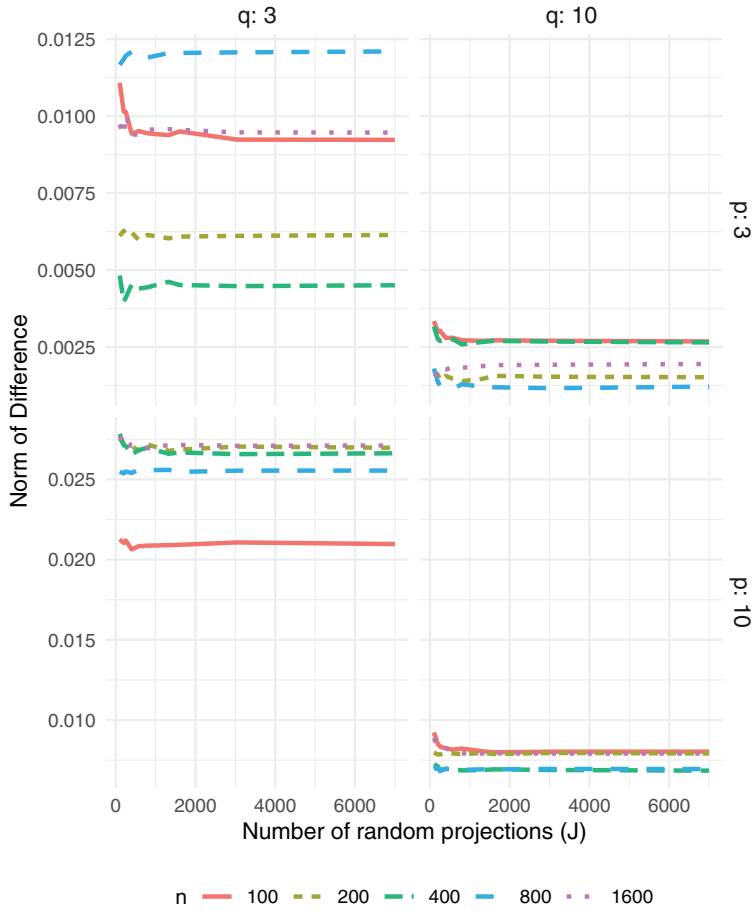

FIGURE D1 Simulation results based on $m = 20,000$ for Example 1 when $\mathbf{X} \sim t_5$ scaled such that the row and column covariance matrices are identity matrices. The random projections follow the distribution $\mathcal{U}_{1,+}^{q-1}$.

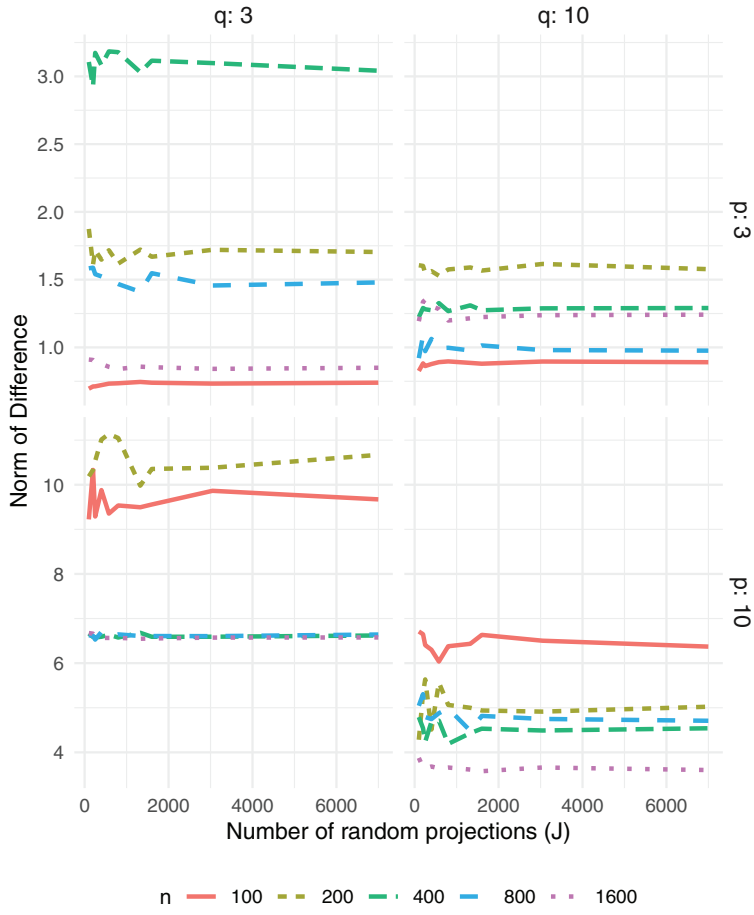

FIGURE D2 Simulation results based on $m = 20,000$ for Example 2 when $\mathbf{X} \sim t_5$ scaled such that the row and column covariance matrices are identity matrices. The random projections follow the distribution $\mathcal{U}_2^{q-1}$.

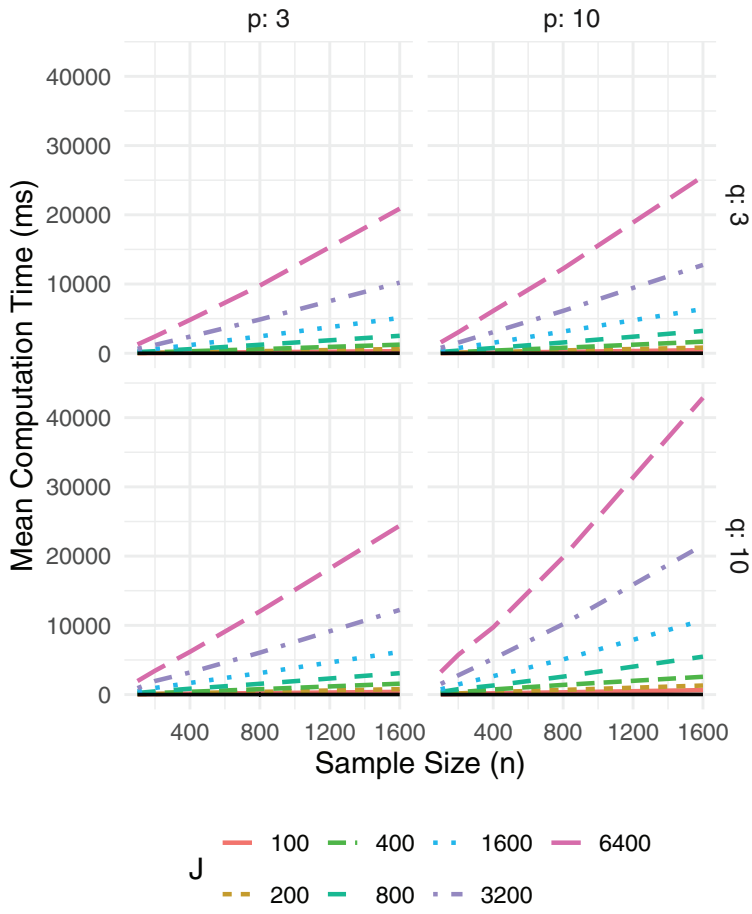

FIGURE D3 Average computation times based on 10 repetitions for various matrix dimensions (p, q), sample sizes (n), and numbers of projections (J). The black line serves as a reference, representing the direct computation of the corresponding column extension in a closed form.