# Rephrasing the lengthy and involved proof of Kristof's theorem: A tutorial with some new findings

**Preprint** · March 2024

DOI: 10.13140/RG.2.2.20026.98243

**1 author:**

Haruhiko Ogasawara
Otaru University of Commerce
**163** PUBLICATIONS **922** CITATIONS

# Rephrasing the lengthy and involved proof of Kristof's theorem: A tutorial with some new findings

March 4, 2024

Haruhiko Ogasawara*

*Professor Emeritus, Otaru University of Commerce, Otaru 047-8501 Japan; Email: emt-hogasa@emt.otaru-uc.ac.jp; Web: https://www.otaru-uc.ac.jp/~emt-hogasa/

Abstract: Kristof's theorem gives the global maximum and minimum of the trace of some matrix products without using calculus or Lagrange multipliers with various applications in psychometrics and multivariate analysis. However, the underutilization has been seen irrespective of its great use in practice. This may partially be due to the lengthy and involved proof of the theorem. In this tutorial, some known or new lemmas are rephrased or provided to understand the essential points in the proof. ten Berge's generalized Kristof theorem is also addressed. Then, the modified Kristof and ten Berge theorems using parent orthonormal matrices are shown, which may be of use to see the properties of the Kristof and ten Berge theorems.

Keywords: von Neumann's trace inequality, generalized Kristof theorem, suborthonormal, semiorthonormal, singular value decomposition.

## 1. Introduction

Kristof's (1969/1970) theorem for the global maximum of the trace of matrix products gives simple derivations of the least square solutions for various problems in psychometrics and multivariate analysis. A special but basic case of the theorem for two sets of matrix products yielding a bilinear form was given by von Neumann (1937, Theorem 1) known as his trace inequality, which was introduced in psychometrics by Green (1969, p. 317) based on the comment of Ingram Olkin.

In spite of its great use, von Neumann's derivation using sophisticated mathematics was not easy for applied researchers to follow. Simplifications or elementary derivations of the theorem has been given by e.g., Kristof (1970) and Mirsky (1975). It is to be noted that these two authors also gave extensions of von Neumann's trace inequality to those with more than two sets of matrix products in different forms.

While the proof by Kristof of his theorem is based on elementary linear algebra using mostly self-contained materials, the proof is long and involved. This may be one of the reasons for the relatively small frequency of citations as commented by Levin (1979, p. 109) "Kristof's theorem has not received its due attention in the psychometric literature", which was also cited by Waller (2018, Introduction), who also stated that "Underutilization of this method likely stems, in part, to the mathematical complexity of Kristof's (1964, 1970) writings" (Abstract).

One of the purposes of this tutorial is to break down Kristof's long and involved derivation using independent lemmas to provide a transparent structure of the proof. Note that the lemmas may also be of interest as general results in elementary linear algebra. The second purpose is to introduce a short derivation of von Neumann's trace inequality obtained by Mirsky (1975) as mentioned earlier, where Fan's (1951) lemma with a self-contained didactic proof is introduced. Note that von Neumann's trace inequality and its extensions have wide applications in various fields e.g., applied linear algebra, mathematical physics and the hyperelasticity of isotropic materials as well as psychology as reviewed by Miranda and Thompson (1993), who cited Kristof (1970) among associated references.

The remainder of this article is organized as follows. In Section 2, some lemmas are introduced for Kristof's theorem followed by a didactic derivation of von Neumann's trace inequality. Section 3 gives didactic proofs of Kristof's theorem for the 3-fold or tri-linear case and the general case. In Section 4, ten Berge's generalized theorem and modifications of Kristof and ten Berge theorems are presented. Some applications of these theorems are shown in Section 5. Section 6 gives discussions. In the appendix, technical details are provided.

## 2. Lemmas for Kristof's theorem and a didactic derivation of von Neumann's trace inequality

In this section, six lemmas and a theorem with a didactic derivation of von Neumann's trace inequality in line with the later derivation of Kristof's theorem will be shown. Lemma 1 gives the maximum of the sum of products of two quantities required for von Neumann's trace inequality, followed by Lemma 1A corresponding to Kristof (1970, Lemma 1) for the similar maximum of the sum of products of more than two quantities for the derivation of Kristof's theorem. Lemma 2 shows the same ranges of the traces irrespective of their absolute values of associated diagonal elements with permutation, which corresponds to Kristof (1970, Lemma 2).

Lemma 3 is a new independent lemma corresponding to the symmetric condition for a maximized trace in Kristof (1970, (iv) of the proof of Theorem (first version)). Lemma 4 is the second independent lemma for the property that symmetric $\mathbf{AD}$ and $\mathbf{DA}$ with $\mathbf{D}$ being diagonal make $\mathbf{A}$ diagonal (Kristof, 1970, (iv) of the proof of Theorem (first version)).

Lemma 5 is the third independent lemma when we have two products of orthonormal and diagonal matrices (a special case of Kristof, 1970, (iv) of the proof of Theorem (first version)) for von Neumann's trace inequality, which was provided to understand the inequality as a special case of Kristof's theorem. Theorem 1 is for von Neumann's trace inequality with the derivation similar to the later one for Kristof's general theorem.

**Lemma 1: The maximum of the sum of products of two quantities (Hardy, Littlewood & Pólya, 1934/1952, Subsection 10.2; von Neumann, 1937, Theorem 1; Simon, 2005, Lemma 1.8)**. *For two sets of m numbers with* $a_1 \geq \cdots \geq a_m \geq 0$ *and*

$b_1 \geq \cdots \geq b_m \geq 0$, *let* $a_1^*,...,a_m^*$ *and* $b_1^*,...,b_m^*$ *be arbitrary cases in each set of m! permutations including possibly the same ones. Then, the maximum of* $\sum_{i=1}^m a_i^* b_i^*$ *over the permutations is given by* $\sum_{i=1}^m a_i b_i$ .

Proof. Without loss of generality, consider the maximum of $\sum_{i=1}^m a_i b_i^*$ . Suppose that $b_1^* \neq b_1$ . Then, exchanging $b_1^*$ and $b_k^* = b_1 (k \neq 1)$ in the permutation, $\sum_{i=1}^m a_i b_i^*$ increases if $b_1^* \neq b_k^*$ and is unchanged if $b_1^* = b_k^*$ since $(a_i - a_j)(b_i - b_j) \geq 0$ and consequently $a_i b_i + a_j b_j - (a_i b_j + a_j b_i) \geq 0$ $(1 \leq i \leq j \leq m)$ . Using this possibly exchanged permutation, redefine $b_1^*,...,b_m^*$ . Then, when $b_2^* \neq b_2$ , exchange $b_2^*$ and $b_k^* = b_2 (k > 2)$ . Repeating this process until the possible exchange of $b_{m-1}^*$ and $b_m^* = b_{m-1}$ when $b_{m-1}^* \neq b_{m-1}$ . The final permutation gives $\sum_{i=1}^m a_i b_i^* = \sum_{i=1}^m a_i b_i$ , which is the maximum since no permutation $b_1^*,...,b_m^*$ using pairwise exchanges after the final one increases $\sum_{i=1}^m a_i b_i^*$ . Q.E.D.

The above proof is a “heuristic” one finding the maximum successively. Waller (2018, Topic III) also used a similar “constructive proof” for the above lemma based on Simon’s (2005, Lemma 1.8., p.4) proof, which is elementary though of interest. However, the above heuristic proof seems to be simpler than Waller’s didactic one. Note that “heuristic” is synonymous with “constructive” in this case.

**Lemma 1A: The maximum of the sum of products with arbitrary number of factors (Kristof, 1970, Lemma 1)**. *For n sets of m numbers with* $a_1^{(j)} \geq \cdots \geq a_m^{(j)} \geq 0$ $(j = 1,...,n;\ n \geq 2)$, *let* $a_1^{(j)*},...,a_m^{(j)*}$ *be an arbitrary case in the j-th set of m! permutations including possibly the same ones. Then, the maximum of* $\sum_{i=1}^m a_i^{(1)*} \cdots a_i^{(n)*}$ *over the permutations is given by* $\sum_{i=1}^m a_i^{(1)} \cdots a_i^{(n)}$ .

Proof. Consider the case of $n = 3$ . For two sets $a_1^{(j)*},...,a_m^{(j)*}$ $(j = 1, 2)$ in the three sets, Lemma 1 gives the maximum of the sum of the products as $\sum_{i=1}^m a_i^{(1)} a_i^{(2)}$ . Then, for the two sets of *m* products $a_1^{(1)}a_1^{(2)},...,a_m^{(1)}a_m^{(2)}$ and *m* numbers $a_1^{(3)*},...,a_m^{(3)*}$, the maximum of $\sum_{i=1}^{m} a_i^{(1)}\, a_i^{(2)} a_i^{(3)*}$ over the *m*! permutation in the third set is similarly obtained by $\sum_{i=1}^{m} a_i^{(1)}\, a_i^{(2)} a_i^{(3)}$. Since any permutation for the maximized one including the first two sets decreases the product sum or remains unchanged as seen in Lemma 1, $\sum_{i=1}^{m} a_i^{(1)}\, a_i^{(2)} a_i^{(3)}$ is the global maximum. The cases with $n \geq 4$ is similarly obtained. Q.E.D.

**Lemma 2: The same ranges of the traces irrespective of their absolute values of the diagonal elements with permutation (Kristof, 1970, Lemma 2)**. *Let* $\mathbf{\Gamma}_1^*$ *and* $\mathbf{\Gamma}_2^*$ *be* $m \times m$ *diagonal matrices; and* $\mathbf{\Gamma}_1$ *and* $\mathbf{\Gamma}_2$ *be those with the corresponding diagonal elements replaced by their corresponding absolute values located in the weakly descending (non-increasing) orders, respectively. Suppose that* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *independently vary over all the* $m \times m$ *orthonormal matrices. Then,* $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*)$ *has the same range as that of* $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2)$.

Proof. Kristof's derivation is didactically repeated. Note that $\mathbf{\Gamma}_i$ is obtained by $\mathbf{\Gamma}_i = \mathbf{P}_i\mathbf{S}_i\mathbf{\Gamma}_i^*\mathbf{P}_i^{\mathrm{T}}$, where $\mathbf{S}_i$ is the signed identity matrix replacing the diagonal elements of $\mathbf{\Gamma}_i^*$ by their corresponding absolute values, and $\mathbf{P}_i$ is the permutation matrix to have the weakly descending order mentioned earlier $(i = 1,2)$. Noting that $\mathbf{\Gamma}_i^* = \mathbf{S}_i\mathbf{P}_i^{\mathrm{T}}\mathbf{\Gamma}_i\mathbf{P}_i$, we have

$$\begin{aligned}\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*) &= \mathrm{tr}\{\mathbf{X}_1(\mathbf{S}_1\mathbf{P}_1^{\mathrm{T}}\mathbf{\Gamma}_1\mathbf{P}_1)\mathbf{X}_2(\mathbf{S}_2\mathbf{P}_2^{\mathrm{T}}\mathbf{\Gamma}_2\mathbf{P}_2)\} \\ &= \mathrm{tr}\{(\mathbf{P}_2\mathbf{X}_1\mathbf{S}_1\mathbf{P}_1^{\mathrm{T}})\mathbf{\Gamma}_1(\mathbf{P}_1\mathbf{X}_2\mathbf{S}_2\mathbf{P}_2^{\mathrm{T}})\mathbf{\Gamma}_2)\}.\end{aligned}$$

In the last result, $\mathbf{P}_2\mathbf{X}_1\mathbf{S}_1\mathbf{P}_1^{\mathrm{T}}$ and $\mathbf{P}_1\mathbf{X}_2\mathbf{S}_2\mathbf{P}_2^{\mathrm{T}}$ are products of orthonormal matrices and consequently orthonormal with the same variations of $\mathbf{X}_1$ and $\mathbf{X}_2$, which shows the required same ranges of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*)$ and $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2)$. Q.E.D.

**Lemma 3: A symmetric condition for a maximized trace (Kristof, 1970, (iv) of the proof of Theorem (first version))**. *Let* $\mathbf{G}$ *be a square matrix of full rank whose singular* *value decomposition (SVD) is* $\mathbf{G} = \mathbf{U\Lambda V}^{\mathrm{T}}$*, where* $\mathbf{U}$ *and* $\mathbf{V}$ *are orthonormal, and* $\mathbf{\Lambda}$ *is a diagonal matrix with positive diagonal elements in a prescribed order. When* $\mathrm{tr}(\mathbf{G})$ *is maximized with a given* $\mathbf{\Lambda}$*,* $\mathbf{G}$ *becomes symmetric.*

Proof. Since $\mathrm{tr}(\mathbf{G}) = \mathrm{tr}(\mathbf{U\Lambda V}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{V}^{\mathrm{T}}\mathbf{U\Lambda})$ with $\mathbf{V}^{\mathrm{T}}\mathbf{U}$ being orthonormal, $\mathrm{tr}(\mathbf{G})$ is maximized when $\mathbf{V}^{\mathrm{T}}\mathbf{U}$ is an identity matrix, which indicates that $\mathbf{U} = \mathbf{V}$ and consequently $\mathbf{G} = \mathbf{U\Lambda U}^{\mathrm{T}}$ is symmetric. Q.E.D.

**Lemma 4: Symmetric $\mathbf{AD}$ and $\mathbf{DA}$ with diagonal $\mathbf{D}$ make $\mathbf{A}$ diagonal (Kristof, 1970, (iv) of the proof of Theorem (first version)).** *Let* $\mathbf{A} = \{a_{ij}\}$ *and* $\mathbf{D}$ *be* $m \times m$ *matrices with* $\mathbf{D}$ *being diagonal. Suppose that the diagonal elements* $d_i (i = 1,...,m)$ *of* $\mathbf{D}$ *are nonzero and* $|d_i|$*'s are mutually different. Suppose further that* $\mathbf{AD}$ *and* $\mathbf{DA}$ *are both symmetric. Then,* $\mathbf{A}$ *is diagonal.*

Proof. By assumption, $\mathbf{AD} = (\mathbf{AD})^{\mathrm{T}} = \mathbf{DA}^{\mathrm{T}}$ and $\mathbf{DA} = (\mathbf{DA})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{D}$. From the last equation, we have $\mathbf{DAD}^{-1} = \mathbf{A}^{\mathrm{T}}$. Substituting $\mathbf{A}^{\mathrm{T}}$ for the right hand-side of the first equation $\mathbf{AD} = \mathbf{DA}^{\mathrm{T}}$, we obtain $\mathbf{A} = \mathbf{D}^2\mathbf{AD}^{-2}$, which indicates that

$$a_{ij} = a_{ij} d_i^2 / d_j^2 (i, j = 1,...,m).$$

Since $d_i^2 / d_j^2 \neq 1$ when $i \neq j$ by assumption, $a_{ij} = 0 (i \neq j)$ follow. Q.E.D.

When $\mathbf{A}$ is diagonal, we have $\mathbf{AD} = \mathbf{DA}$, where $\mathbf{A}$ and $\mathbf{D}$ are said to commute. Using this formulation, a lemma equivalent to Lemma 4 was stated by Kiers and ten Berge (1989, Lemma 1).

**Lemma 5: Two products of orthonormal and diagonal matrices (a special case of Kristof, 1970, (iv) of the proof of Theorem (first version)).** *Let* $\mathbf{X}_i$, $\mathbf{D}_i$ *and* $\mathbf{\Delta}_i$ *be* $m \times m$ *matrices, where* $\mathbf{X}_i$ *is orthonormal while* $\mathbf{D}_i$ *and* $\mathbf{\Delta}_i$ *are diagonal and of full rank* $(i = 1, 2)$*. Suppose that*

$$\mathbf{X}_1\mathbf{D}_1\mathbf{X}_2 = \mathbf{\Delta}_1 \textit{ and } \mathbf{X}_2\mathbf{D}_2\mathbf{X}_1 = \mathbf{\Delta}_2.$$

*Then,* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are the* $m \times m$ *signed and/or permuted identity matrices with m nonzero* *elements being* $\pm 1$*, where permutation indicates row- or column-wise one. When without permutation,* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are diagonal matrices with their diagonal elements being* $\pm 1$*, and* $\mathbf{\Delta}_1\mathbf{D}_2 = \mathbf{\Delta}_2\mathbf{D}_1$.

Proof. Left-multiplying $\mathbf{X}_1\mathbf{D}_1$ on both sides of $\mathbf{X}_2\mathbf{D}_2\mathbf{X}_1 = \mathbf{\Delta}_2$ using $\mathbf{X}_1\mathbf{D}_1\mathbf{X}_2 = \mathbf{\Delta}_1$, we have $\mathbf{\Delta}_1\mathbf{D}_2\mathbf{X}_1 = \mathbf{X}_1\mathbf{D}_1\mathbf{\Delta}_2$ giving $\mathbf{\Delta}_1\mathbf{D}_2 = \mathbf{X}_1\mathbf{D}_1\mathbf{\Delta}_2\mathbf{X}_1^{\mathrm{T}}$. The last result shows the spectral decomposition of the diagonal matrix $\mathbf{\Delta}_1\mathbf{D}_2$, which indicates that the orthonormal matrix $\mathbf{X}_1$ becomes a signed and/or permuted identity matrix and when without permutation $\mathbf{\Delta}_1\mathbf{D}_2 = \mathbf{\Delta}_2\mathbf{D}_1$. For $\mathbf{X}_2$, exchanging the subscripts “1” and “2” due to symmetry, we obtain $\mathbf{\Delta}_2\mathbf{D}_1 = \mathbf{X}_2\mathbf{D}_2\mathbf{\Delta}_1\mathbf{X}_2^{\mathrm{T}}$ indicating the same results as for $\mathbf{X}_1$. Q.E.D.

**Theorem 1: Two-fold or bilinear case (*n* = 2) (von Neumann, 1937, Theorem 1; Kristof, 1970, Theorem (first version)).** *Let* $\mathbf{\Gamma}_1^*$ *and* $\mathbf{\Gamma}_2^*$ *be fixed diagonal matrices of full rank with the absolute values of the diagonal elements being mutually different in each matrix. Consider the maximum and minimum of* $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*)$ *which are attained, where* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *independently vary over all* $m \times m$ *orthonormal matrices. Then,*

$$-\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2) \le \mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*) \le \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2),$$

*where* $\mathbf{\Gamma}_i = \mathrm{diag}(\gamma_{i1},...,\gamma_{im})$ *is given by* $\mathbf{\Gamma}_i^*$ *with their diagonal elements replaced by the corresponding absolute values with possible permutation to have the descending order i.e.,* $\gamma_{i1} > \cdots > \gamma_{im} > 0 \quad (i = 1,2)$.

Proof. By Lemma 2, since the range of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2)(= \mathrm{tr}(\mathbf{\Gamma}_2\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2))$ is the same as that of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*)$, we consider the former range. Due to Lemma 3, when the former trace is maximized, $\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2$ and similarly $\mathbf{\Gamma}_2\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2$ become symmetric. Then, using Lemma 4 we find that $\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2$ is diagonal. Due to symmetry with $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2) = \mathrm{tr}(\mathbf{\Gamma}_2\mathbf{X}_2\mathbf{\Gamma}_1\mathbf{X}_1)$, $\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_1$ is also found to be diagonal. By Lemma 5, these two diagonal conditions give the maximum $\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2)$ of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2)$ when $\mathbf{X}_1$ and $\mathbf{X}_2$ are the same signed identity matrices. The global maximum $\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2)$ among the permuted diagonal elements of $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ is shown by Lemma 1. The minimum is given by replacing e.g., $\mathbf{X}_1$ by $-\mathbf{X}_1$. Q.E.D.

**Remark 1**. The second simple proof of Theorem 1 (Mirsky, 1975) using an associated property of the doubly stochastic matrix obtained by Fan (1951, Lemma 1A) will be shown with Fan's lemma in the appendix. Note that a doubly stochastic matrix is a square one, where the sum of each row and that of each column are unities. An example is the matrix consisting of the squared elements of an orthonormal matrix. Mirsky's proof has been known in the mathematical community as a short and simple derivation of von Neumann's trace inequality.

**Remark 1A**. In his tutorial, Waller (2018, Equation (21)) explained the result of Theorem 1 using the symmetric condition as given in Lemma 3 with the SVD $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2) = \mathrm{tr}(\mathbf{P\Delta Q}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{PQ}^{\mathrm{T}}\mathbf{\Delta})$, whose optima are attained when $\mathbf{P} = \mathbf{Q}$ or $\mathbf{P} = -\mathbf{Q}$ as $-\mathrm{tr}(\mathbf{\Delta}) \le \mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2) \le \mathrm{tr}(\mathbf{\Delta})$. This result is correct. Then, Waller (2018, Equation (22)) gave inequalities $-\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2) \le \mathrm{tr}(\mathbf{\Delta}) \le \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2)$ using our notation followed by the statement "the bounds by Kristof's theorem can be achieved". These are also correct. However, the most important result in Theorem 1 is $\mathrm{tr}(\mathbf{\Delta}) = \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2)$, whose proof has been shown by using Lemma 5 as well as the second one in the appendix.

## 3. Didactic proofs of Kristof's theorem

In this section, an independent lemma in linear algebra is provided, which is an extension corresponding to the result in the proof of Kristof (1970). The tri-linear case is given as Theorem 3 for didactic purposes, followed by a short proof of Kristof's general theorem using several lemmas.

**Lemma 6: Two products of square, diagonal and orthonormal matrices (an extension of Kristof, 1970, (iv) of the proof of Theorem (first version))**. *Let* $\mathbf{A}, \mathbf{X}, \mathbf{D}_i$ *and* $\mathbf{\Delta}_i$ *be* $m \times m$ *matrices of full rank, where* $\mathbf{X}$ *is orthonormal while* $\mathbf{D}_i$ *and* $\mathbf{\Delta}_i$ *are* *diagonal* $(i=1,2)$. *Suppose that*

$$\mathbf{A}\mathbf{D}_1\mathbf{X} = \mathbf{\Delta}_1 \text{ and } \mathbf{X}\mathbf{D}_2\mathbf{A} = \mathbf{\Delta}_2 .$$

*Then,* $\mathbf{X}$ *is the* $m \times m$ *signed and/or permuted identity matrices with m nonzero elements being* $\pm 1$, *where permutation indicates row- or column-wise one. When without permutation,* $\mathbf{X}$ *is diagonal with its diagonal elements being* $\pm 1$, *and* $\mathbf{\Delta}_1\mathbf{D}_2 = \mathbf{\Delta}_2\mathbf{D}_1$.

Proof. Right-multiplying $\mathbf{D}_1\mathbf{X}$ on both sides of $\mathbf{X}\mathbf{D}_2\mathbf{A} = \mathbf{\Delta}_2$ using $\mathbf{A}\mathbf{D}_1\mathbf{X} = \mathbf{\Delta}_1$, we have $\mathbf{X}\mathbf{D}_2\mathbf{\Delta}_1 = \mathbf{\Delta}_2\mathbf{D}_1\mathbf{X}$ giving $\mathbf{\Delta}_2\mathbf{D}_1 = \mathbf{X}\mathbf{\Delta}_1\mathbf{D}_2\mathbf{X}^{\mathrm{T}}$. The last result shows the spectral decomposition of the diagonal matrix $\mathbf{\Delta}_2\mathbf{D}_1$, which indicates that the orthonormal matrix $\mathbf{X}$ becomes a signed and/or permuted identity matrix and when without permutation $\mathbf{\Delta}_1\mathbf{D}_2 = \mathbf{\Delta}_2\mathbf{D}_1$. Q.E.D.

**Remark 2**. Lemma 5 is seen as a special case of Lemma 6 when $\mathbf{A} = \mathbf{X}_1$ an orthonormal matrix and $\mathbf{X}$ is denoted by $\mathbf{X}_2$. However, in Lemma 5, both $\mathbf{X}_1$ and $\mathbf{X}_2$ were found to be signed and/or permutated identity matrices. Note also that Kristof (1970) dealt with the case when $\mathbf{A} = \mathbf{G}_1\mathbf{\Gamma}_1\mathbf{G}_2\mathbf{\Gamma}_2 \cdots \mathbf{G}_{n-1}\mathbf{\Gamma}_{n-1}\mathbf{G}_n$, $\mathbf{D}_1 = \mathbf{\Gamma}_n$, $\mathbf{D}_2 = \mathbf{\Gamma}_{n+1}$ and $\mathbf{X} = \mathbf{G}_{n+1}$, where $\mathbf{\Gamma}_i$ and $\mathbf{G}_i$ are diagonal and orthonormal matrices, respectively. This specification was necessary for his derivation by induction though the involved expression $\mathbf{G}_1\mathbf{\Gamma}_1\mathbf{G}_2\mathbf{\Gamma}_2 \cdots \mathbf{G}_{n-1}\mathbf{\Gamma}_{n-1}\mathbf{G}_n$ may hide the basic structure in Lemma 6.

**Theorem 2: Three-fold or trilinear case (*n* = 3) (Kristof, 1970, Theorem (first version))**. *Let* $\mathbf{\Gamma}_i^*$ $(i=1,2,3)$ *be fixed diagonal matrices of full rank with the absolute values of the diagonal elements being mutually different in each matrix. Consider the maximum and minimum of* $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*\mathbf{X}_3\mathbf{\Gamma}_3^*)$ *which are attained, where* $\mathbf{X}_i (i=1,2,3)$ *independently vary over all* $m \times m$ *orthonormal matrices. Then,*

$$-\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2\mathbf{\Gamma}_3) \le \mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*\mathbf{X}_3\mathbf{\Gamma}_3^*) \le \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2\mathbf{\Gamma}_3) ,$$

*where* $\mathbf{\Gamma}_i = \mathrm{diag}(\gamma_{i1},...,\gamma_{im})$ *is given by* $\mathbf{\Gamma}_i^*$ *with their diagonal elements replaced by the corresponding absolute values with possible permutation to have the descending order i.e.,*

$\gamma_{i1} > \cdots > \gamma_{im} > 0 \ \ (i = 1, 2, 3)$.

Proof. As in Theorem 1, the range of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3\mathbf{\Gamma}_3)(= \mathrm{tr}(\mathbf{\Gamma}_3\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3))$ is the same as that of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^*\mathbf{X}_3\mathbf{\Gamma}_3^*)$. Due to Lemma 3, when the former trace is maximized, $\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3\mathbf{\Gamma}_3 \equiv \mathbf{B}\mathbf{\Gamma}_3$ and similarly $\mathbf{\Gamma}_3\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3 = \mathbf{\Gamma}_3\mathbf{B}$ become symmetric. Then, using Lemma 4 we find that $\mathbf{B} = \mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3 \equiv \mathbf{A}\mathbf{\Gamma}_2\mathbf{X}_3$ is diagonal. Similarly, since $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3\mathbf{\Gamma}_3) = \mathrm{tr}(\mathbf{A}\mathbf{\Gamma}_2\mathbf{X}_3\mathbf{\Gamma}_3) = \mathrm{tr}(\mathbf{X}_3\mathbf{\Gamma}_3\mathbf{A}\mathbf{\Gamma}_2)$, $\mathbf{X}_3\mathbf{\Gamma}_3\mathbf{A}$ is also diagonal. From Lemma 6, these two diagonal conditions can make $\mathbf{X}_3$ an identity matrix. Then, using Theorem 1, the maximum $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3\mathbf{\Gamma}_3) = \mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{\Gamma}_3)$ is obtained when $\mathbf{X}_1$ and $\mathbf{X}_2$ are identity matrices as $\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2\mathbf{\Gamma}_3)$. The minimum $-\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2\mathbf{\Gamma}_3)$ is obtained as in Theorem 1. Q.E.D.

**Remark 2A**. In the proof of Theorem 2, the key result is $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{X}_3\mathbf{\Gamma}_3)$ $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{\Gamma}_3)$. Since $\mathbf{\Gamma}_2\mathbf{\Gamma}_3 \equiv \mathbf{\Gamma}_{2*3}$ is diagonal, the maximum of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2\mathbf{\Gamma}_3)$ $= \mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_{2*3})$ is obtained by Theorem 1 for the bilinear case. This suggests a heuristic proof for the general case, which is employed in the following result.

**Theorem 3: *n*-fold case (*n* = 2, 3,...) (Kristof, 1970, Theorem (first version))**. *Let* $\mathbf{\Gamma}_i^* \ (i = 1,\ldots,n)$ *be fixed diagonal matrices. Consider the maximum and minimum of* $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^* \cdots \mathbf{X}_n\mathbf{\Gamma}_n^*)$ *which are attained, where* $\mathbf{X}_i (i = 1,\ldots,n)$ *independently vary over all* $m \times m$ *orthonormal matrices. Then,*

$$-\mathrm{tr}(\mathbf{\Gamma}_1 \cdots \mathbf{\Gamma}_n) \le \mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^* \cdots \mathbf{X}_n\mathbf{\Gamma}_n^*) \le \mathrm{tr}(\mathbf{\Gamma}_1 \cdots \mathbf{\Gamma}_n),$$

*where* $\mathbf{\Gamma}_i = \mathrm{diag}(\gamma_{i1},\ldots,\gamma_{im})$ *is given by* $\mathbf{\Gamma}_i^*$ *with their diagonal elements replaced by the corresponding absolute values with possible permutation to have the weakly descending order i.e.,* $\gamma_{i1} \ge \cdots \ge \gamma_{im} \ge 0 \ \ (i = 1,\ldots,n)$.

Proof. As in Kristof (1970), first suppose that $\gamma_{i1} > \cdots > \gamma_{im} > 0 \ \ (i = 1,\ldots,n)$. Using the result of Theorem 2, increase *n* one by one as *n* = 4, 5,... When *n* = 4, redefine $\mathbf{B} \equiv \mathbf{X}_1\mathbf{\Gamma}_1 \cdots \mathbf{X}_3\mathbf{\Gamma}_3\mathbf{X}_4 \equiv \mathbf{A}\mathbf{\Gamma}_3\mathbf{X}_4$. Then, from Lemma 4, $\mathbf{B} = \mathbf{A}\mathbf{\Gamma}_3\mathbf{X}_4$ becomes diagonal.

Similarly, $\mathbf{X}_4\mathbf{\Gamma}_4\mathbf{A}$ is also diagonal. Then, as before $\mathbf{X}_4$ can become an identity matrix. Using Theorem 2, the maximum of $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\cdots\mathbf{X}_3\mathbf{\Gamma}_3\mathbf{X}_4\mathbf{\Gamma}_4)=\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\cdots\mathbf{X}_3\mathbf{\Gamma}_3\mathbf{\Gamma}_4)$ is given by $\mathrm{tr}(\mathbf{\Gamma}_1\cdots\mathbf{\Gamma}_4)$. The minimum is similarly obtained as $-\mathrm{tr}(\mathbf{\Gamma}_1\cdots\mathbf{\Gamma}_4)$. Increasing *n* successively one by one, we obtain the required results.

Further, consider the weakly ordered case $\gamma_{i1}\geq\cdots\geq\gamma_{im}\geq 0\ \ (i=1,...,n)$. As in Kristof (1970, (v) of the proof of Theorem (first version) based on the suggestion by Bary G. Wingersky), let $\mathbf{W}=\mathrm{diag}(w_1,...,w_m)$ with $w_1>\cdots>w_m>0$. Redefine $\mathbf{\Gamma}_i$ as $\mathbf{\Gamma}_i+\varepsilon\mathbf{W}(i=1,...,n)$ with $\varepsilon>0$. Then, we have the same above result since $\gamma_{i1}>\cdots>\gamma_{im}>0$. When $\varepsilon$ approaches zero with fixed $\mathbf{X}_i(i=1,...n)$, the same required result is given by substituting $\varepsilon=0$ for $\mathbf{\Gamma}_i+\varepsilon\mathbf{W}$ under the limiting condition of $\gamma_{i1}\geq\cdots\geq\gamma_{im}\geq 0\ \ (i=1,...,n)$. Q.E.D.

**Remark 3**. The heuristic derivation of Theorem 3 is essentially equal to that by induction, where the latter was employed by Kristof (1970). The method of successively finding maxima was shown for didactic purposes as well as a direct derivation in Theorem 2 when *n* = 2. Since Theorems 1 and 2 are special cases of Theorem 3, the former results also hold under $\gamma_{i1}\geq\cdots\geq\gamma_{im}\geq 0\ \ (i=1,...,n)$. Note that when $\gamma_{i1}\geq\cdots\geq\gamma_{im}=0$, which is given in the limiting case of $\varepsilon=0$, $\mathbf{\Gamma}_i$ becomes singular while $\mathbf{\Gamma}_i+\varepsilon\mathbf{W}$ with $\varepsilon>0$ is non-singular, this rank difference does not affect the maximum attained.

## 4. Generalizations of Kristof's theorem

ten Berge (1983) gave a generalized version of Kristof's theorem when $\mathbf{X}_i$'s with $\mathrm{rank}(\mathbf{X}_i)\equiv r_i^*\leq r_i$ are $m_{i-1}\times m_i$ possibly non-square suborthonormal matrices i.e., submatrices of orthonormal ones $(i=1,...n;\ m_0\equiv m_n)$. Let a semiorthonormal matrix be a non-square submatrix of its parent orthonormal one with the same number of the tows or columns (not both) as that of the parent. Note that if $\mathbf{X}_i$ is suborthonormal rather than semi- or fully orthonormal, $r_i^*$ can be 0 especially when the dimensionality of the parent orthonormal matrix is more than or equal to $m_{i-1} + m_i$, and when the latter matrix varies unrestrictedly. Though ten Berge did not fully explain the cases with $r_i^* < r_i$, $r_i$ is seen as an upper bound of $r_i^*$, which is given by $r_i = \min\{m_{i-1}, m_i\}$. Note that $r_i = \min\{m_{i-1}, m_i\}$ is the smallest upper bound when $\mathbf{X}_i$ varies unrestrictedly. In other words, when $r_i < \min\{m_{i-1}, m_i\}$, $\mathbf{X}_i$ does not vary unrestrictedly. If $\mathbf{X}_i$ is semi- or fully orthonormal, we have $r_i^* = r_i = \min\{m_{i-1}, m_i\}$.

Let $\mathbf{\Gamma}_i^*$ and $\mathbf{\Gamma}_i$ be $m_i \times m_i$ fixed diagonal matrices defined as in Theorem 3 with possible different $m_i$'s $(i = 1, ...n)$. Let $r = \min(r_1, ..., r_n)$ and $m = \max(m_1, ..., m_n)$. Denote the $r \times r$ identity matrix by $\mathbf{I}_r$. Define $\mathbf{\Delta}_i$ and $\mathbf{E}_r$ as the $m \times m$ diagonal matrices containing $\mathbf{\Gamma}_i$ and $\mathbf{I}_r$ in their upper left corners with zeros elsewhere, respectively. Then, ten Berge gave the following result.

**Theorem 4: The generalized Kristof theorem (ten Berge, 1983, Theorem 1; see also Kiers & ten Berge, 1989; ten Berge, 1993/2005, Sections 3.2 and 3.3).** *Under the definitions and assumptions given above, when* $\mathbf{X}_i$ *varies with the condition* $\mathrm{rank}(\mathbf{X}_i) \le r_i$ $(i = 1, ..., n)$, *we have*

$$-\mathrm{tr}(\mathbf{\Delta}_1 \cdots \mathbf{\Delta}_n \mathbf{E}_r) \le \mathrm{tr}(\mathbf{X}_1 \mathbf{\Gamma}_1^* \cdots \mathbf{X}_n \mathbf{\Gamma}_n^*) \le \mathrm{tr}(\mathbf{\Delta}_1 \cdots \mathbf{\Delta}_n \mathbf{E}_r).$$

For the proof of Theorem 4, ten Berge (1983) defined $\mathbf{Y}_i$ as the $m \times m$ matrix containing $\mathbf{X}_i$ in its upper left corner with zeroes elsewhere. Then, he defined its SVD as $\mathbf{Y}_i = \mathbf{P}_i \mathbf{D}_i \mathbf{Q}_i^{\mathrm{T}}$, where "$\mathbf{P}_i$ and $\mathbf{Q}_i$ are orthonormal and $\mathbf{D}_i$ is diagonal" (loc.cit., p. 521). Note that he employed the SVD using the non-negative singular values rather than the positive ones, where $\mathbf{P}_i$, $\mathbf{Q}_i$ and $\mathbf{D}_i$ are $m \times m$ square matrices. This is seen in his inequalities (loc.cit., Equation (10))

$$-\mathrm{tr}(\mathbf{D}_1 \mathbf{\Delta}_1 \cdots \mathbf{D}_n \mathbf{\Delta}_n) \le \mathrm{tr}(\mathbf{X}_1 \mathbf{\Gamma}_1^* \cdots \mathbf{X}_n \mathbf{\Gamma}_n^*) \le \mathrm{tr}(\mathbf{D}_1 \mathbf{\Delta}_1 \cdots \mathbf{D}_n \mathbf{\Delta}_n).$$

For this derivation, he used $\mathbf{Y}_i = \mathbf{P}_i \mathbf{D}_i \mathbf{Q}_i^{\mathrm{T}}$ and $\mathbf{\Delta}_i^*$ defined similarly to $\mathbf{\Delta}_i$ using $\mathbf{\Gamma}_i^*$

$(i=1,\ldots,n)$, which gives

$$\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\cdots\mathbf{X}_n\mathbf{\Gamma}_n^*)=\mathrm{tr}(\mathbf{Y}_1\mathbf{\Delta}_1^*\cdots\mathbf{Y}_n\mathbf{\Delta}_n^*)=\mathrm{tr}(\mathbf{P}_1\mathbf{D}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{\Delta}_1^*\cdots\mathbf{P}_n\mathbf{D}_n\mathbf{Q}_n^{\mathrm{T}}\mathbf{\Delta}_n^*)$$

(loc.cit., Equation (9)). He applied Kristof's theorem to this result supposing that all the $2n$ $m\times m$ matrices $\mathbf{P}_i$ and $\mathbf{Q}_i$ $(i=1,\ldots,n)$ vary over all the orthonormal matrices, which is required in Kristof's theorem. Then, he showed his Equation (1) shown earlier.

However, it is found that when $\mathbf{D}_i=\mathrm{diag}(d_1,\ldots,d_m)$ with $d_1\ge\cdots\ge d_{r_i^*}>0$ and $d_{r_i^*+1}=\cdots=d_m=0$, orthonormal matrices $\mathbf{P}_i$ and $\mathbf{Q}_i$ in $\mathbf{Y}_i=\mathbf{P}_i\mathbf{D}_i\mathbf{Q}_i^{\mathrm{T}}$ should be of the form

$$\mathbf{P}_i=\begin{pmatrix}\mathbf{P}_{i1} & \mathbf{O}\\ \mathbf{O} & \mathbf{P}_{i2}\end{pmatrix}\text{ and }\mathbf{Q}_i=\begin{pmatrix}\mathbf{Q}_{i1} & \mathbf{O}\\ \mathbf{O} & \mathbf{Q}_{i2}\end{pmatrix}$$

where $\mathbf{P}_{i1}$ and $\mathbf{Q}_{i1}$ are $r_i^*\times r_i^*$ orthonormal submatrices while $\mathbf{P}_{i2}$ and $\mathbf{Q}_{i2}$ are $(m-r_i^*)\times(m-r_i^*)$ similar ones unless vanishing when $r_i^*=m$. This formulation is due to ten Berge's definition of $\mathbf{Y}_i$ whose elements are zero except the upper left submatrix.

Although $\mathbf{P}_i$ and $\mathbf{Q}_i$ are orthonormal rather than semi- or suborthonormal, their block diagonal forms have substantial restrictions in the variations of orthonormal matrices required by Kristof's theorem. One of the severe restrictions is the lack of giving permutations across two sets of variables. Consequently, the upper and lower bounds using Kristof's theorem may not be attained when the diagonal elements of $\mathbf{D}_i\mathbf{\Delta}_i$ are not located in the weakly descending order. This restriction was not mentioned by ten Berge though he did not state that the bounds are attained in his theorem.

The necessity of $\mathbf{E}_r$ in the statement of Theorem 4 is due to the unrestricted rank condition of the diagonal matrix $\mathbf{\Gamma}_i$ in the upper left corner of $\mathbf{\Delta}_i$ employed by ten Berge. Since the rank of $\mathbf{\Delta}_i$ may be greater than that of $\mathbf{D}_i$, the upper bound in his Equation (10) becomes

$$\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\cdots\mathbf{X}_n\mathbf{\Gamma}_n^*)\le\mathrm{tr}(\mathbf{D}_1\mathbf{\Delta}_1\cdots\mathbf{D}_n\mathbf{\Delta}_n)=\mathrm{tr}(\mathbf{D}_1\cdots\mathbf{D}_n\mathbf{\Delta}_1\cdots\mathbf{\Delta}_n)\le\mathrm{tr}(\mathbf{\Delta}_1\cdots\mathbf{\Delta}_n\mathbf{E}_r),$$

where the last inequality is due to the range [0, 1] of the singular values of suborthonormal matrices (loc.cit., Lemma 2) yielding $\mathbf{D}_1 \cdots \mathbf{D}_n \le \mathbf{E}_r$ in Löwner's (1934, p. 177) sense. ten Berge explicitly wrote that "the statement that "the limits can be attained" has to be omitted" (loc.cit., p. 521). It is to be noted that he added that "the limits ... can be attained if the $\mathbf{X}_i$ are varying independently and (except for the rank) unrestrictedly over the set of suborthonormal matrices" (loc.cit., p. 521).

The meaning of the parenthetical phrase "(except for the rank)" is not clear since when the upper bound $r_i$ of the rank of $\mathbf{X}_i$ (recall the condition $\mathrm{rank}(\mathbf{X}_i) = r_i^* \le r_i$) is less than $\min\{m_{i-1}, m_i\}$ , $\mathbf{X}_i$ does not vary unrestrictedly, but varies over a subset of the suborthonormal matrices satisfying $r_i^* \le r_i < \min\{m_{i-1}, m_i\}$. That is, in the subset, $\mathbf{X}_i$ cannot be semi- or fully orthonormal. In other words, in this subset the sum of the squared elements in each row or column of $\mathbf{X}_i$ is smaller than 1. Under this restriction, the optima may not be obtained. Note also that ten Berge also mentioned the typical cases with i.e., $r_i^* = r_i$ as "this modification does not affect the validity" (loc.cit., p. 521) of his generalized theorem though the optima may not be attained due to the difficulty of applying Kristof's theorem using constrained parent orthonormal matrices.

In the following modification with attained optima, fully unconstrained suborthonormal matrices are considered. Let $\mathbf{X}_i$ be the $m_{i-1} \times m_i$ $(i = 1, ..., n;\ m_0 \equiv m_n)$ possibly non-square matrix with $\mathrm{rank}(\mathbf{X}_i) = r_i^* \le r_i = \min\{m_{i-1}, m_i\}$, which is supposed to vary unrestrictedly and independently over the set of $m_{i-1} \times m_i$ suborthonormal matrices in the corresponding $m \times m$ parent orthonormal matrix with $m = \max(m_1, ..., m_n)$ as given earlier. The parent orthonormal matrix is denoted by $\mathbf{X}_i^*$, which includes $\mathbf{X}_i$ as a submatrix.

Let $\mathbf{\Gamma}_i^*$ and $\mathbf{\Gamma}_i$ be $m_i \times m_i$ fixed diagonal matrices defined as in Theorems 3 and 4. In the modification, however, $\mathbf{\Gamma}_i^*$ and $\mathbf{\Gamma}_i$ are assumed to be non-singular without loss of generality. This is seen from the form $\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1^*\mathbf{X}_2\mathbf{\Gamma}_2^* \cdots \mathbf{X}_n\mathbf{\Gamma}_n^*)$ to be optimized later, since when $\mathbf{\Gamma}_i^*$ is singular, $\mathbf{\Gamma}_i^*$ can be redefined by deleting the row(s) and column(s) corresponding to the zero diagonal elements of $\mathbf{\Gamma}_i^*$. Then, in the similar manner, the corresponding column(s) of $\mathbf{X}_i$ and row(s) of $\mathbf{X}_{i+1} (\mathbf{X}_{n+1} \equiv \mathbf{X}_1)$ can be deleted without changing the value of $\mathrm{tr}(\mathbf{X}_1 \mathbf{\Gamma}_1^* \mathbf{X}_2 \mathbf{\Gamma}_2^* \cdots \mathbf{X}_n \mathbf{\Gamma}_n^*)$, where $r_i^* (r_{i+1}^*)$ and $r_i (r_{i+1})$ may be adjusted for the reduced $\mathbf{X}_i (\mathbf{X}_{i+1})$ when necessary.

**Theorem 5: A modified generalized Kristof theorem (a modification of ten Berge, 1983, Theorem 1)**. *Let* $\mathbf{X}_i$, $\mathbf{X}_i^*$, $\mathbf{\Gamma}_i^*$ *and* $\mathbf{\Gamma}_i$ $(i = 1, ..., n)$ *be as defined above. Define* $\mathbf{\Delta}_i$ *as the* $m \times m$ *diagonal matrix, whose upper left submatrix is* $\mathbf{\Gamma}_i$ *elsewhere zero, as defined earlier. Then, when the parent orthonormal matrices* $\mathbf{X}_i^*$ $(i = 1, ..., n)$ *vary independently and unrestrictedly over the set of orthonormal matrices, we have*

$$-\mathrm{tr}(\mathbf{\Delta}_1 \cdots \mathbf{\Delta}_n) \le \mathrm{tr}(\mathbf{X}_1 \mathbf{\Gamma}_1^* \cdots \mathbf{X}_n \mathbf{\Gamma}_n^*) \le \mathrm{tr}(\mathbf{\Delta}_1 \cdots \mathbf{\Delta}_n),$$

*where the optima are attained.*

Proof. Define $\mathbf{\Delta}_i^*$ using $\mathbf{\Gamma}_i^*$ similarly to $\mathbf{\Delta}_i$. Then, we obtain

$$\mathrm{tr}(\mathbf{X}_1 \mathbf{\Gamma}_1^* \cdots \mathbf{X}_n \mathbf{\Gamma}_n^*) = \mathrm{tr}(\mathbf{X}_1^* \mathbf{\Delta}_1^* \cdots \mathbf{X}_n^* \mathbf{\Delta}_n^*).$$

Noting that the assumption of the independent and unrestricted variations of $\mathbf{X}_i^*$ $(i = 1, ..., n)$ satisfies that of Kristof's theorem, the required results with the attained optima follow. Q.E.D.

**Remark 4**. In Theorem 5, the assumption for the variations of $\mathbf{X}_i^*$ automatically gives $\mathrm{rank}(\mathbf{X}_i) = r_i^* \le r_i = \min\{m_{i-1}, m_i\}$. The optima are obtained when $r_i^* = r_i$, which indicates that $\mathbf{X}_i$ is semi- or fully orthonormal with non-zero singular value(s) being unity when the optima are attained. This makes the matrix $\mathbf{E}_r$ used in ten Berge's theorem unnecessary. Recall that the optima may not be obtained in his theorem since the non-zero singular value(s) of $\mathbf{X}_i$ may be less than unity and since the SVD form of $\mathbf{Y}_i$ restricts the permutation of the diagonal elements of $\mathbf{\Gamma}_i^*$ to have $\mathbf{\Gamma}_i$. In other words, when $\mathbf{\Gamma}_i^* = \mathbf{\Gamma}_i$, the latter restriction vanishes. Note that the former restriction corresponds to $r_i^* < r_i$. That is, under this assumption $\mathbf{X}_i$ is suborthonormal (not semi- or fully orthonormal). On the other hand, the case $r_i^* = r_i$, addressed earlier with ten Berge's statement, indicates that $\mathbf{X}_i$ is semi- or fully orthonormal. Although generally this case does not satisfy the assumption of the unconstrained variation of the parent orthonormal matrix, which is the assumption in Kristof's theorem, the restricted variation also gives the same optima as for the unrestricted case since the non-zero singular value(s) are unities as long as $r_i^* = r_i$. For ten Berge's generalized theorem, Kiers and ten Berge (1989, p. 132) stated that "*r* is the minimum of the ranks of $\mathbf{\Gamma}_1,...,\mathbf{\Gamma}_k$ and $\mathbf{X}_1,...,\mathbf{X}_k$", where *k* = *n*. This is misleading and should be corrected as "*r* is the minimum of the ranks of $\mathbf{\Gamma}_1,...,\mathbf{\Gamma}_k$ and $r_1,...,r_k$" since when $\mathrm{rank}(\mathbf{X}_i) = r_i^* \le r_i$, $r_i^*$ can be smaller than $r_i = \min\{m_{i-1}, m_i\}$ in the subset of the variation of $\mathbf{X}_i$ unless the restricted case of $r_i^* = r_i$ is used.

Remark 4 indicates the following modification of Kristof's theorem.

**Theorem 6: A modification of Kristof's theorem**. *In Theorem 3 of Kristof's theorem (first version), redefine the orthonormal matrices* $\mathbf{X}_i (i = 1,...,n)$ *as*

$$\mathbf{X}_i = \mathrm{Bdiag}(\mathbf{X}_{i1},...,\mathbf{X}_{ii_\mathrm{B}}),$$

*where* $\mathbf{X}_{ij}$ *is the* $i_j \times i_j$ *diagonal block* $(j = 1,...,i_\mathrm{B})$ *with* $i_1 + ... + i_{i_\mathrm{B}} = m$ *. Suppose that* $\mathbf{X}_i (i = 1,...,n)$ *independently and unrestrictedly vary with* $\mathrm{rank}(\mathbf{X}_{ij}) = i_j \quad (j = 1,...,i_\mathrm{B})$. *Use* $\mathbf{\Gamma}_i (i = 1,...,n)$ *as defined in Kristof's theorem. Then, we have*

$$-\mathrm{tr}(\mathbf{\Gamma}_1 \cdots \mathbf{\Gamma}_n) \le \mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1 \cdots \mathbf{X}_n\mathbf{\Gamma}_n) \le \mathrm{tr}(\mathbf{\Gamma}_1 \cdots \mathbf{\Gamma}_n),$$

*where the optima are attained.*

Proof. The case when $\mathbf{\Gamma}_i^* = \mathbf{\Gamma}_i (i = 1,...,n)$ can be used in Kristof's theorem. Although $\mathbf{X}_i (i = 1,...,n)$ take the block-diagonal forms, the optima with $\mathbf{X}_i = \mathbf{I}_m \ (i = 1,...,n)$ are attained in the subset of varying $\mathbf{X}_i (i = 1,...,n)$ under the block diagonal restriction. Q.E.D.

The usage of $\mathbf{\Gamma}_i$ in place of $\mathbf{\Gamma}_i^* (i = 1,...,n)$ is due to the lack of permutation across the different diagonal blocks when using the block diagonal $\mathbf{X}_i \quad (i = 1,...,n)$. Note that Kiers and ten Berge (1989) used the assumptions that $\mathbf{\Gamma}_i$'s are available when *n* = 2 stating that "in most applications the assumptions are satisfied, if they are not satisfied only minor modifications will typically be involved" (pp. 126-127). Note also that when $i_{\mathrm{B}} = 1$, $\mathbf{X}_i$ becomes a usual orthonormal matrix as in Kistof's theorem, and when $i_{\mathrm{B}} = m$, $\mathbf{X}_i$ is the signed identity matrix whose diagonal elements are $\pm 1$, where $\mathbf{X}_i$ and its diagonal elements i.e., $\pm 1$ are $m \times m$ and $1 \times 1$ orthonormal matrices with unit singular value(s), respectively.

An extension in Theorems 4 and 5, is to relax $\mathbf{\Gamma}_i$ to be an unconstrained square matrix of the same rank as that of $\mathbf{\Gamma}_i$ $(i = 1,\ldots,n)$, which is mathematically immaterial. Since under the relaxed condition the SVD of $\mathbf{\Gamma}_i \equiv \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{V}_i^{\mathrm{T}}$ with $\mathbf{U}_i^{\mathrm{T}} \mathbf{U}_i$ and $\mathbf{V}_i^{\mathrm{T}} \mathbf{V}_i$ being an identity matrix of the same order as $\mathrm{rank}(\mathbf{\Gamma}_i)$ $(i = 1,\ldots,n)$, redefine $\mathbf{X}_1$ and $\mathbf{X}_i$ as $\mathbf{V}_n \mathbf{X}_1 \mathbf{U}_1^{\mathrm{T}}$ and $\mathbf{V}_{i-1} \mathbf{X}_i \mathbf{U}_i^{\mathrm{T}}$ $(i = 2,\ldots,n)$. Then, the problem becomes the same as that using the diagonal matrix $\mathbf{\Lambda}_i$ $(i = 1,\ldots n)$. Note that this unconstrained condition was used in von Neumann's (1937) trace inequality and Kristof's (1970) Theorem (second version). Kristof (1970, p. 523) stated that "A distinction of the two versions will not be emphasized", where the second version uses the unconstrained square matrix $\mathbf{\Gamma}_i$.

## 5. Applications of Kristof's theorem and its generalizations

While as mentioned in the introductory section, "underutilization" of Kristof's theorem and its generalization seems to be still true considering its simplicity, generality and tractability yielding solutions in various applications. In this section, basic or important cases in multivariate analysis showing advantages of Kristof's theorem and its generalizations are provided. Examples or applications are shown below mostly in the chronological order. Although in ten Berge's generalized Kristof's theorem, the optima may not be attained, all his three examples in multivariate analysis and an illustration of Caucy-Schwarz inequality using his theorem are the cases with attained optima.

**Maximization of $\mathrm{tr}(\mathbf{M\Lambda L})$ or $\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda})$ : Green (1969, Appendix B) (*n* = 1)**. In this problem, $\mathbf{M}, \mathbf{L}$ and $\mathbf{A}$ are fixed $m \times r, s \times m$ and $r \times s$ $(r \geq s)$ matrices, respectively while $\mathbf{\Lambda}$ is the $r \times s$ matrix varying over all the semiorthonormal matrices when $r > s$ or orthonormal when $r = s$. This is seen as an extended application of von Neumann's trace inequality when $n = 1$ with unconstrained and possibly rectangular $\mathbf{A} = (\mathbf{LM})^{\mathrm{T}}$. An application of this problem to have an optimal linear combination with a specified correlation matrix was investigated. This problem will also be addressed later.

**Orthogonal procrustes transformation: Kristof's (1970) Example 1 (*n* = 1).** This problem is to minimize $\mathrm{tr}\{(\mathbf{A} - \mathbf{BT})(\mathbf{A} - \mathbf{BT})^{\mathrm{T}}\}$, where $\mathbf{A}$ and $\mathbf{B}$ are fixed $r \times s$ $(r \geq s)$ matrices while $\mathbf{T}$ is an orthonormal matrix to be derived, which reduces to maximizing $\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{BT})$, a case with $n = 1$. Note that $\mathbf{T}$ gives permutations with sign changes (reflections) of the columns of $\mathbf{B}$ as well as the rotation of $\mathbf{B}$. Although the term "procrustes rotation" is usually used as stated by Kristof (1970, p. 523) especially in factor analysis using factor rotation, it is important $\mathbf{T}$ can give permutations and reflections of the columns of $\mathbf{B}$.

The method of "procrustes rotation" is seen as that when the permutation with possible reflection is unnecessary or has been appropriately chosen. This is important in applications since the permutation and reflection are not always unnecessary as in simulations when $\mathbf{B}$ is one of sample loading matrices randomly obtained and is to be matched to $\mathbf{A}$. An advantage of von Neumann's inequality or Kristof's theorem is to consider permutation and reflection as well as rotation while the method of differentiation with the Lagrange multipliers generally requires checking all the possible permutations and reflections.

At the end of Example 1, Kristof (1970, p. 524) stated that "The generalization of the present problem to allow $\mathbf{A}^{\mathrm{T}}\mathbf{B}$ to be singular is immediate and does not require special discussion." It is found that the singular case with $\mathrm{rank}(\mathbf{A}^{\mathrm{T}}\mathbf{B}) \equiv s^* < s \leq r$ gives the $s^*$ positive singular values, say, $\gamma_1 \geq \cdots \geq \gamma_{s^*} > 0$. Then, using the SVD $\mathbf{A}^{\mathrm{T}}\mathbf{B} = \mathbf{U}\,\mathrm{diag}(\gamma_1,\ldots,\gamma_{s^*})\mathbf{V}^{\mathrm{T}}$, we obtain

$$\begin{aligned}&\max\{\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{BT})\} = \max[\mathrm{tr}\{\mathbf{U}\,\mathrm{diag}(\gamma_1,\ldots,\gamma_{s^*})\mathbf{V}^{\mathrm{T}}\mathbf{T}\}]\\ &= \max[\mathrm{tr}\{\mathrm{diag}(\gamma_1,\ldots,\gamma_{s^*})\mathbf{V}^{\mathrm{T}}\mathbf{TU}\}] = \mathrm{tr}\{\mathrm{diag}(\gamma_1,\ldots,\gamma_{s^*})\}\\ &= \sum\nolimits_{i=1}^{s^*}\gamma_i,\end{aligned}$$

where $\mathbf{V}^{\mathrm{T}}\mathbf{TU}$ is a suborthonormal matrix since the product of suborthonormal matrices is suborthonormal (ten Berge, 1983, Lemma 4). Actually, $\mathbf{U}$ and $\mathbf{V}$ of full column rank by definition are semiorthonormal (ten Berge, 1983, Definition 2) while $\mathbf{T}$ is orthonormal as well as suborthonormal.

Kristof (1970, p. 524) added three examples with $n = 1$ ("the trivial case of the theorem" in his terminology) for determinations of e.g., orthogonal matrices with specific properties developed in the1960s in psychometrics.

**The two-sided orthogonal procrustes problem: Kristof's (1970) Example 2 ($n = 2$).** This problem based on Schönemann (1968) minimizes $\mathrm{tr}\{(\mathbf{B}-\mathbf{T}^{\mathrm{T}}\mathbf{AS})(\mathbf{B}-\mathbf{T}^{\mathrm{T}}\mathbf{AS})^{\mathrm{T}}\}$, where $\mathbf{A}$ and $\mathbf{B}$ are fixed square matrices while $\mathbf{T}$ and $\mathbf{S}$ are orthonormal matrices such that $\mathbf{A}$ is to be matched to $\mathbf{B}$. This problem reduces to maximizing $\mathrm{tr}(\mathbf{T}^{\mathrm{T}}\mathbf{ASB}^{\mathrm{T}})$, a case with unconstrained $\mathbf{A}$ and $\mathbf{B}$ when $n = 2$. The maximum of $\mathrm{tr}(\mathbf{T}^{\mathrm{T}}\mathbf{ASB}^{\mathrm{T}})$ is obtained as the product of the positive singular values of $\mathbf{A}$ and $\mathbf{B}$.

**Multivariate multiple regression: Kristof's (1970) Example 4 and ten Berge's (1983) Application 1 ($n = 2$).** Kristof's example included an unnatural restriction of the same numbers of the dependent and independent variables, which was removed by ten Berge's application. The problem is to minimize $\mathrm{tr}\{(\mathbf{Y}-\mathbf{XB})(\mathbf{Y}-\mathbf{XB})^{\mathrm{T}}\}$, where $\mathbf{Y}$ is an $s \times u$ matrix for $u$ dependent variables and $\mathbf{X}$ of full column rank is an $s \times t$ matrix for $t$ independent variables. The well-known solution $\hat{\mathbf{B}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ was obtained without calculus when $n = 2$ though the method is somewhat tedious.

**Principal components analysis: ten Berge's (1983) Application 2 ($n = 1$).** This application deals with the jointly determining $p$ principal components when $\mathbf{R}$ is a $k \times k$ correlation matrix of rank $r \le k$. The problem is to maximize the sum of squared loadings $\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{R}^2\mathbf{B})$, where $\mathbf{B}$ is a $k \times p$ loading matrix with $p \le r$ subject to the uncorrelated components with unit variances $\mathbf{B}^{\mathrm{T}}\mathbf{RB} = \mathbf{I}_p$. Note that an atypical assumption of the possibly singular $\mathbf{R}$ is used.

Let $\mathbf{R} = \mathbf{K}\mathbf{\Lambda}\mathbf{K}^{\mathrm{T}}$ with $\mathbf{K}^{\mathrm{T}}\mathbf{K} = \mathbf{I}_r$ be the SVD using the positive diagonal elements of the $r \times r$ diagonal matrix $\mathbf{\Lambda}$ The solution can be given by maximizing

$$\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{R}^2\mathbf{B}) = \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^2\mathbf{K}^{\mathrm{T}}\mathbf{B}) = \mathrm{tr}\{(\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^{1/2})\mathbf{\Lambda}(\mathbf{\Lambda}^{1/2}\mathbf{K}^{\mathrm{T}}\mathbf{B})\}$$
$$= \mathrm{tr}\{(\mathbf{\Lambda}^{1/2}\mathbf{K}^{\mathrm{T}}\mathbf{B})(\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^{1/2})\mathbf{\Lambda}\}$$

with $n = 1$, where $(\mathbf{\Lambda}^{1/2}\mathbf{K}^{\mathrm{T}}\mathbf{B})(\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^{1/2})$ is suborthonormal since

$$\mathbf{I}_p = \mathbf{B}^{\mathrm{T}}\mathbf{R}\mathbf{B} = \mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}\mathbf{K}^{\mathrm{T}}\mathbf{B} = (\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^{1/2})(\mathbf{\Lambda}^{1/2}\mathbf{K}^{\mathrm{T}}\mathbf{B}).$$

Noting that $\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^{1/2}$ is a $p \times r\,(p \le r)$ semiorthonormal matrix, ten Berge's generalized Kristof theorem gives $\max\{\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{R}^2\mathbf{B})\} = \sum_{i=1}^{p} \lambda_i$, where $\lambda_1 \ge \cdots \ge \lambda_r > 0$ and the unrotated loading matrix $\mathbf{B}$ is a submatrix of $\mathbf{K}\mathbf{\Lambda}^{-1/2}$ taking its first $p$ columns to give $\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^{1/2}$ a submatrix of $\mathbf{I}_r$ consisting its first $p$ rows.

The formulation using suborthonormal matrices $\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{R}^2\mathbf{B}) = \mathrm{tr}\{(\mathbf{B}^{\mathrm{T}}\mathbf{K}\mathbf{\Lambda}^{1/2})\mathbf{\Lambda}(\mathbf{\Lambda}^{1/2}\mathbf{K}^{\mathrm{T}}\mathbf{B})\}$ is of interest since the restriction $\mathbf{B}^{\mathrm{T}}\mathbf{R}\mathbf{B} = \mathbf{I}_p$ is cleverly used in the maximizing function without calculus or Lagrange multipliers. While the above example was employed by ten Berge to show an application of his generalized Kristof theorem, when all the $k$ components including minor ones are obtained in the usual non-singular case of $\mathbf{R}$, the maximum of $\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{R}^2\mathbf{B})$ is obtained by Kristof's theorem as $\mathrm{tr}(\mathbf{\Lambda})$, where $\mathbf{\Lambda}$ is the $k \times k$ diagonal matrix with positive diagonals. In this case, $\mathbf{\Lambda}^{1/2}\mathbf{K}^{\mathrm{T}}\mathbf{B}$ becomes $\mathbf{I}_k$ yielding the well-known unrotated loading matrix $\mathbf{B} = \mathbf{K}\mathbf{\Lambda}^{-1/2}$.

**Canonical correlations: Kristof (1970, General comment (a)) (*n* = 1), ten Berges (1983, Application 3) (*n* = 1), Ogasawara (2000 with errata, Theorem 1) (*n* = 2), and Waller (2018, pp. 195-196) (*n* = 1).** Kristof (1970) suggested a formulation of canonical correlations applying his theorem when $n = 1$, which was fully described by ten Berge (1983) using an associated SVD as in principal components. Ogasawara (2000) used Kristof's theorem with $n = 2$ to have a lower bound of the mean squared canonical correlations between factors in factor analysis and the corresponding principal components. Waller also showed an application of Kristof's theorem for canonical correlations using two sets of principal components in the two sets of original variables. However, his results are those when the numbers of original variables in each set are the same. Note that use of the principal components needs justification when the numbers of the original variables are not equal or the number of the canonical correlations is less than the minimum of the numbers of the original variables.

**The Caucy-Schwarz inequality: ten Berge (1983, p. 522) and Theorem 5 of the current article (*n* = 2).** This example is simple but impressive in that the example well shows the simplicity and generality of Kristof's theorem and its extensions without calculus. Let $\mathbf{x}$ and $\mathbf{y}$ be non-zero vectors of order *k*. Suppose that $\mathbf{x}$ and $\mathbf{y}$ vary independently and unrestrictedly. Then,

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}\mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1/2} &= \mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}\mathbf{I}_k\mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1/2}1 \\ &= \mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}\mathbf{\Gamma}_1^*\mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1/2}\mathbf{\Gamma}_2^*,\end{aligned}$$

where $\mathbf{\Gamma}_1^* = \mathbf{\Gamma}_1 = \mathbf{I}_k$ and $\mathbf{\Gamma}_2^* = \mathbf{\Gamma}_2 = 1$. Define $k \times k$ diagonal matrices $\mathbf{\Delta}_1 = \mathbf{\Gamma}_1 = \mathbf{I}_k$ and $\mathbf{\Delta}_2 = \mathbf{E}_{11}$, where the first diagonal element of $\mathbf{E}_{11}$ is unity elsewhere zero. Since the vectors $\mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}$ and $\mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}$ are semiorthonormal, their non-zero singular values are unity. Consequently, when applying ten Berge's theorem, the maximum is attained as $\mathrm{tr}(\mathbf{\Delta}_1\mathbf{\Delta}_2) = \mathrm{tr}(\mathbf{I}_k\mathbf{E}_{11}) = 1$ with minimum $-\mathrm{tr}(\mathbf{\Delta}_1\mathbf{\Delta}_2) = -1$ obtained similarly. Then, we have the Caucy-Schwarz inequality $-1 \le \mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}\mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1/2} = \dfrac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{1/2}} \le 1$.

The same result is obtained by using Theorem 5 of this article. Define the $k \times k$ parent orthonormal matrices $\mathbf{X}$ and $\mathbf{Y}$, whose first rows are $\mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}$ and $\mathbf{y}^{\mathrm{T}}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1/2}$, respectively, where $\mathbf{X}$ and $\mathbf{Y}$ independently and unrestrictedly vary over the sets of orthonormal matrices. Then, using Theorem 5, the maximum of $\mathbf{x}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1/2}\mathbf{y}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1/2}$ is attained as $\mathrm{tr}(\mathbf{X}\mathbf{I}_k\mathbf{Y}\mathbf{E}_{11}) \le \mathrm{tr}(\mathbf{I}_k\mathbf{E}_{11}) = 1$ with the minimum $\mathrm{tr}(\mathbf{X}\mathbf{I}_k\mathbf{Y}\mathbf{E}_{11}) \ge -\mathrm{tr}(\mathbf{I}_k\mathbf{E}_{11}) = -1$ obtained similarly.

**Generalized linear form: ten Berge (1993/2005, Equation (48)), Yanai and Takane (2007, Property 11) and Adachi (2016, Theorem A.4.2) (*n* = 1).** This is a problem maximizing $\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A})$, where $\mathbf{X}$ and fixed $\mathbf{A}$ are $p \times q$ $(p \ge q)$ matrices with the constraint $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}_q$ and $\mathrm{rank}(\mathbf{A}) \le q$. To the author's knowledge, this problem was first solved by Green (1969) as mentioned in the first example. After Kristof (1970), ten Berge (1993) reformulated the problem as the generalized linear form. He defined the SVD

$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$ with $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_q$ and $\boldsymbol{\Lambda}$ being the diagonal matrix with the non-negative diagonals using an application of his generalized Kristof theorem with $n$ = 1. Then, the maximum is given by

$$\max\{\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A})\} = \max\{\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}})\} = \max\{\mathrm{tr}(\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda})\} = \mathrm{tr}(\boldsymbol{\Lambda})$$

since $\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{U}$ is suborthonormal. The maximum is attained when $\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_q$ or $\mathbf{X} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$.

Note that the definition of the SVD using the non-negative rather than positive singular values is important considering the case $\mathrm{rank}(\mathbf{A}) \equiv q^* < q$. That is, in the last case when using only positive singular values, $\mathbf{U}$ and $\mathbf{V}$ become $p \times q^*$ and $q \times q^*$ semiorthonormal matrices, respectively, yielding $\mathbf{X}^{\mathrm{T}}\mathbf{X} \neq \mathbf{I}_q$, which does not satisfy the assumption. Note that Green (1969, p. 317) correctly considered the two cases $q^* < q$ and $q^* = q$ as well as the cases of multiple or equal positive singular values in terms of the uniqueness of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{U}\mathbf{V}^{\mathrm{T}}$. For this example, Neudecker's (2004, Section 2) derivation as "a Kristof-type theorem" with a correction and added explanation will be shown in the appendix

## 6. Discussion

**(a) The trivial case (*n* = 1) and the bilinear case i.e., von Neumann's trace inequality (*n* = 2)**. In the previous section, only the examples of *n* = 1 or 2 are shown. Though Kristof (1970) used the term "trivial case" when *n* = 1, its applications are meaningful ones as shown earlier. Note that only the derivation of Kristof's theorem is trivial or self-evident when *n* = 1. A case of *n* = 4 was provided by Kristof (1970, Example 6) as a generalization of Meredith's (1964) problem for a multivariate selection of subpopulations from a common parent. However, most of the applications of Kristof's theorem and ten Berge's generalized one seem to be those of *n* = 1 or 2. Kiers and ten Berge (1989, p. 126) stated that "All practical applications we have encountered so far apply to the cases *k* = 1 or *k* = 2", where *k* is used for *n*. That said, it is to be noted that ten Berge (1983, p. 509) stated that "Theorems should be derived in the greatest possible generality".

**(b) Alternative proofs**. Proofs in seminal papers tend to be complicated. After the discoveries, alternative simple or short proofs follow. For von Neumann's (1937) trace inequality, the elementary alternative proof by Mirsky (1975) with Fan's (1951) lemma may be the simplest one as shown in the appendix. In this tutorial, rephrasing or breaking down the proof by Kristof for his theorem has been shown. However, the logic is essentially the same as Kristof's one using induction. Marshall, Olkin and Arnold (2011, Chapter 20, Theorem B.2) also showed a similar proof by induction though they stated that "We give an inductive proof that is elementary, though still somewhat lengthy" (p. 791). Finding alternative simple, self-contained and hopefully short proofs of Kristof's theorem when $n \geq 3$ is an open problem. Mirsky used the doubly stochastic matrix. Applications or generalizations of Mirsky's proof to the inequalities when $n \geq 3$ seem to be difficult as far as the author conjectures.

**(c) Equivalent and inequivalent cases of the Kristof and ten Berge theorems**. As mentioned earlier, ten Berge (1983, Theorem 2) extended Kristof's theorem. For the differences of the theorems, he stated that "The most striking difference is that the $\mathbf{X}_i$ are no longer requited to be orthonormal. Second, the $\mathbf{X}_i$ need no longer to be square" (p. 521), which are advantages of ten Berge's theorem over Kristof's one claimed by ten Berge. The third difference is the lack of the attained optima, which is not an advantage. The two claimed advantages may be handled by Kristof's theorem by considering the parent orthonormal matrices as used by ten Berge and Theorem 5 in this article. So, the two theorems may be seen as equivalent when the optima are attained. Note that all the four examples in ten Berge (1983) are the cases of attained optima. Probably, the cases of unattained optima due to $\mathrm{rank}(\mathbf{X}_i) = r_i^* < r_i = \min\{m_{i-1}, m_i\}$ may be theoretical or special, if any, in practice.

It is conjectured that even in this special case, some adjustment giving $r_i^* \leq r_i$ may be obtained. For instance, consider canonical correlation analysis for two sets of standardized data matrices i.e., $\mathbf{X}_1(n \times r_1)$ of rank $r_1^*$ and $\mathbf{X}_2(n \times r_2)$ of rank $r_2^*$. Then, when $r^* < \min\{r_1^*, r_2^*\}$ canonical correlations are optimally derived in a least squares sense among $\min\{r_1^*, r_2^*\}$ possible ones, this seems to yield a similar problem. Actually, as ten

Berge (1983, p. 523) formulated the situation using the coefficient matrices $\mathbf{B}_1$ $(r_1 \times r^*)$ and $\mathbf{B}_2$ $(r_2 \times r^*)$ with other ones, the maximum of $\mathrm{tr}(\mathbf{B}_1^{\mathrm{T}}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\mathbf{B}_2)$ was attained.

## Appendix

**Lemma A: Fan's (1951, Lemma 1A) inequality for the doubly substochastic matrix**. *Let* $\mathbf{a} = (a_1,...,a_n)^{\mathrm{T}}$ *and* $\mathbf{b} = (b_1,...,b_n)^{\mathrm{T}}$ *be fixed vectors with* $a_1 \geq \cdots \geq a_n \geq 0$ *and* $b_1 \geq \cdots \geq b_n \geq 0$. *Define an* $n \times n$ *doubly substochastic matrix* $\mathbf{P} = \{p_{ij}\}$ *with non-negative elements satisfying* $\sum_{j=1}^{n} p_{ij} \leq 1$ $(i = 1,...,n)$ *and* $\sum_{i=1}^{n} p_{ij} \leq 1$ $(j = 1,...,n)$ *(see e.g., Marshall, Olkin & Arnold, 2011, Section 2.C). Then,* $\mathbf{a}^{\mathrm{T}}\mathbf{P}\mathbf{b} \leq \mathbf{a}^{\mathrm{T}}\mathbf{b}$.

Proof (a slight extension of Uchida, 2023). Let $c_1,...,c_n$ and $d_1,...,d_n$ be non-negative numbers. Then, we can write $a_i = \sum_{k=i}^{n} c_k$ and $b_i = \sum_{k=i}^{n} d_k$. Using these expressions,

$$
\begin{aligned}
\mathbf{a}^{\mathrm{T}}\mathbf{b} - \mathbf{a}^{\mathrm{T}}\mathbf{P}\mathbf{b} &= \sum_{i,j=1}^{n} (\delta_{ij} - p_{ij}) a_i b_j = \sum_{i,j=1}^{n} (\delta_{ij} - p_{ij}) \sum_{k=i}^{n} c_k \sum_{l=j}^{n} d_l \\
&= \sum_{k,l=1}^{n} c_k d_l \sum_{i=1}^{k} \sum_{j=1}^{l} (\delta_{ij} - p_{ij})
\end{aligned}
$$

follows, where $\delta_{ij}$ is the Kronecker delta. In the above expression, define the doubly stochastic matrix $\mathbf{P}^* = \{p_{ij}^*\}$ satisfying $p_{ij}^* \geq p_{ij}$ $(i, j = 1,...n)$. Then, consider the case $k \geq l$ in the above expression. We obtain

$$
\begin{aligned}
\sum_{i=1}^{k} \sum_{j=1}^{l} (\delta_{ij} - p_{ij}) &= \sum_{i=1}^{k} \sum_{j=1}^{l} \delta_{ij} - \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} \\
&= \sum_{i=1}^{l} \delta_{ii} - \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} = l - \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} \\
&\geq l - \sum_{i=1}^{n} \sum_{j=1}^{l} p_{ij}^* \geq l - l = 0.
\end{aligned}
$$

When, $k \leq l$, in a similar manner we obtain $\sum_{i=1}^{k} \sum_{j=1}^{l} (\delta_{ij} - p_{ij}) \geq 0$. These two inequalities give the required result $\mathbf{a}^{\mathrm{T}}\mathbf{b} - \mathbf{a}^{\mathrm{T}}\mathbf{P}\mathbf{b} \geq 0$. Q.E.D.

**Remark A**. The original proof by Fan (1951) is a short one though it is not self-contained in that "Abel's lemma" is used. The author could not identify the Abel lemma with an associated reference among Abel's formulas. The above proof is a slight extension of the result by Uchida (2023) who dealt with only the doubly stochastic matrix $\mathbf{P}^*$, which is a special case of the doubly substochastic matrix $\mathbf{P}$.

**The second proof of Theorem 1** (von Neuman's trace inequality; Mirsky, 1975, Section 3, p. 305). $|\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2)| \le \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2)$ is derived. Let $\mathbf{X}_i = \{x_{(i)jk}\}\,(j,k = 1,...,m)$ and $\mathbf{\Gamma}_i = \mathrm{diag}(\gamma_{(i)1},...,\gamma_{(i)m})\ (i = 1,\ 2)$. Then, using Fan's (1951) inequality for doubly (sub)stochastic matrices, we have

$$\begin{aligned}
&|\mathrm{tr}(\mathbf{X}_1\mathbf{\Gamma}_1\mathbf{X}_2\mathbf{\Gamma}_2)| = \sum\nolimits_{j,k=1}^{m} x_{(1)jk}\gamma_{(1)k}x_{(2)kj}\gamma_{(2)j} \\
&\le \sum\nolimits_{j,k=1}^{m} |\,x_{(1)jk}x_{(2)kj}\,|\,\gamma_{(1)k}\gamma_{(2)j} \\
&\le \frac{1}{2}\sum\nolimits_{j,k=1}^{m} x_{(1)jk}^{2}\gamma_{(1)k}\gamma_{(2)j} + \frac{1}{2}\sum\nolimits_{j,k=1}^{m} x_{(2)kj}^{2}\gamma_{(1)k}\gamma_{(2)j} \\
&\le \frac{1}{2}\sum\nolimits_{j=1}^{m} \gamma_{(1)j}\gamma_{(2)j} + \frac{1}{2}\sum\nolimits_{j=1}^{m} \gamma_{(1)j}\gamma_{(2)j} = \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2).
\end{aligned}$$

Q.E.D.

**A Kristof-type theorem for correlation preserving predictors of factor scores: Neudecker (2004, Section 2) (*n* = 1).** Neudecker obtained the same solution of the first example by Green and the last reformulated one by ten Berge in Section 5 as "A Kristof-type theorem" in the context of the derivations of correlation preserving predictors of factor scores (for these predictors see the references in Neudecker, 2004; and Mori & Kurata, 2013). He did not mention or use ten Berge's theorem, but employed calculus and Lagrange multipliers. Neudecker also used the SVD $\mathbf{A} = \mathbf{U}_0\mathbf{\Lambda}_0\mathbf{V}_0^{\mathrm{T}}$ , where he employed only positive singular values i.e., $\mathbf{\Lambda}_0 > \mathbf{O}$ in Löwner's sense.

An advantage of Neudecker's derivation is to give the set of explicit expressions of $\mathbf{X}$ maximizing $\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A})$ as $\mathbf{X} = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ when $\mathrm{rank}(\mathbf{A}) = q^* = q$ and $\mathbf{X} = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{Q}(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})$ with a $p \times q$ matrix $\mathbf{Q}$ when $q^* < q$ as "the general solution" (Neudecker, 2004, Equation (2.5)). In these expressions,

$$\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} = \mathbf{X}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2}\}^{+} = \mathbf{X}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}\}^{1/2} \equiv \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(+)1/2} \text{ and}$$

$$\mathbf{V}_0\mathbf{V}_0^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(+)1/2}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{(+)1/2},$$

where $(\cdot)^{1/2}$ is the matrix square root of a matrix; and $(\cdot)^{+}$ is the Moore-Penrose generalized (MP *g*-) inverse of a possibly rectangular matrix, which is obtained by using the SVD. That

is, when the SVD of a matrix is $\mathbf{Y} = \mathbf{P\Gamma Q}^{\mathrm{T}}$ employing only the positive singular values, we have $\mathbf{Y}^{+} = \mathbf{Q\Gamma}^{-1}\mathbf{P}^{\mathrm{T}}$, which satisfies the conditions of the MP *g*-inverse:
$\mathbf{YY}^{+}\mathbf{Y} = \mathbf{Y},\ \mathbf{Y}^{+}\mathbf{YY}^{+} = \mathbf{Y}^{+},\ \mathbf{YY}^{+} = (\mathbf{YY}^{+})^{\mathrm{T}}$ and $\mathbf{Y}^{+}\mathbf{Y} = (\mathbf{Y}^{+}\mathbf{Y})^{\mathrm{T}}$.

Neudecker (2004, Equation (2.5)) stated that "$\mathbf{Q}$ arbitrary". This is misleading since when $\mathbf{Q}$ is a zero matrix, the rank of $\mathbf{X} = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}}$ becomes $q^{*}(< q)$ and does not satisfy $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}_q$ though this will give the same maximum. Instead, $\mathbf{Q}$ should be defined as a $p \times q$ arbitrary suborthonormal matrix $\mathbf{Q} = \mathbf{U}_1\mathbf{V}_1^{\mathrm{T}}$ of rank $q - q^{*}$, where $\mathbf{U}_1$ and $\mathbf{V}_1$ are $p \times (q - q^{*})$ and $q \times (q - q^{*})$ semiorthonormal matrices, respectively such that $\mathbf{U}_1^{\mathrm{T}}\mathbf{U}_1 = \mathbf{V}_1^{\mathrm{T}}\mathbf{V}_1 = \mathbf{I}_{q-q^{*}}$, $\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_1 = \mathbf{O}$ and $\mathbf{V}_0^{\mathrm{T}}\mathbf{V}_1 = \mathbf{O}$, which shows the arbitrary property of $\mathbf{Q}$ stated earlier in that when $\mathbf{U}_1$ and $\mathbf{V}_1$ are replaced by $\mathbf{U}_1\mathbf{U}_1^{*}$ and $\mathbf{V}_1\mathbf{V}_1^{*}$ with $\mathbf{U}_1^{*}$ and $\mathbf{V}_1^{*}$ being arbitrary $(q - q^{*}) \times (q - q^{*})$ orthonormal matrices, these can be used with a different $\mathbf{Q}^{*} \equiv \mathbf{U}_1\mathbf{U}_1^{*}(\mathbf{V}_1\mathbf{V}_1^{*})^{\mathrm{T}} \neq \mathbf{Q}$. Note that these arbitrary $\mathbf{U}_1$ and $\mathbf{V}_1$ give
$\mathbf{V}_0^{\mathrm{T}}\mathbf{V}_0 + \mathbf{V}_1^{\mathrm{T}}\mathbf{V}_1 = (\mathbf{V}_0 : \mathbf{V}_1)^{\mathrm{T}}(\mathbf{V}_0 : \mathbf{V}_1) = (\mathbf{V}_0 : \mathbf{V}_1\mathbf{V}_1^{*})^{\mathrm{T}}(\mathbf{V}_0 : \mathbf{V}_1\mathbf{V}_1^{*}) = \mathbf{I}_q$.

Then, it is found that

$$\begin{aligned}\mathbf{X} &= \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{Q}(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}}) = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}_1\mathbf{V}_1^{\mathrm{T}}(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}}) \\ &= \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}_1\mathbf{V}_1^{\mathrm{T}} = (\mathbf{U}_0 : \mathbf{U}_1)(\mathbf{V}_0 : \mathbf{V}_1)^{\mathrm{T}}\end{aligned}$$

satisfying

$$\begin{aligned}\mathbf{X}^{\mathrm{T}}\mathbf{X} &= (\mathbf{V}_0 : \mathbf{V}_1)(\mathbf{U}_0 : \mathbf{U}_1)^{\mathrm{T}}(\mathbf{U}_0 : \mathbf{U}_1)(\mathbf{V}_0 : \mathbf{V}_1)^{\mathrm{T}} \\ &= (\mathbf{V}_0 : \mathbf{V}_1)(\mathbf{V}_0 : \mathbf{V}_1)^{\mathrm{T}} = \mathbf{I}_q.\end{aligned}$$

Using the above $\mathbf{X}$, the maximum is given by

$$\begin{aligned}&\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{X}) = \mathrm{tr}\{(\mathbf{U}_0\mathbf{\Lambda}_0\mathbf{V}_0^{\mathrm{T}})^{\mathrm{T}}(\mathbf{U}_0 : \mathbf{U}_1)(\mathbf{V}_0 : \mathbf{V}_1)^{\mathrm{T}}\} \\ &= \mathrm{tr}\{\mathbf{V}_0\mathbf{\Lambda}_0\mathbf{U}_0^{\mathrm{T}}(\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}_1\mathbf{V}_1^{\mathrm{T}})\} \\ &= \mathrm{tr}(\mathbf{V}_0\mathbf{\Lambda}_0\mathbf{V}_0^{\mathrm{T}}) = \mathrm{tr}(\mathbf{V}_0^{\mathrm{T}}\mathbf{V}_0\mathbf{\Lambda}_0) \\ &= \mathrm{tr}(\mathbf{\Lambda}_0),\end{aligned}$$

where $\mathrm{tr}(\mathbf{\Lambda}_0) = \mathrm{tr}(\mathbf{\Lambda})$ and the singular diagonal matrix $\mathbf{\Lambda}$ of rank $q^{*}$ was used earlier. Define the semiorthonormal matrix and $\mathbf{U} \equiv (\mathbf{U}_0 : \mathbf{U}_1)$ and orthonormal $\mathbf{V} \equiv (\mathbf{V}_0 : \mathbf{V}_1)$.

Then, $\mathbf{X} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ is equal to that obtained by ten Berge as shown earlier.

Recall the expression $\mathbf{X} = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{Q}(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})$. The matrix $\mathbf{Q}$ can be an arbitrary $p \times q$ semiorthonormal one denoted by $\mathbf{U}^*$ with $\mathbf{U}_0^{\mathrm{T}}\mathbf{U}^* = \mathbf{O}$ and $\mathbf{U}^{*\mathrm{T}}\mathbf{U}^* = \mathbf{I}_q$. This is seen by

$$\begin{aligned}\mathbf{X}^{\mathrm{T}}\mathbf{X} &= \{\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}^*(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})\}^{\mathrm{T}}\{\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}^*(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})\}\\ &= \mathbf{V}_0\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + (\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})\mathbf{U}^{*\mathrm{T}}\mathbf{U}^*(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})\\ &= \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}} = \mathbf{I}_q,\end{aligned}$$

satisfying the assumption, where the idempotent property $(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})^2 = \mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}}$ is used. The matrix $\mathbf{X}$ gives the same maximum

$$\begin{aligned}\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{X}) &= \mathrm{tr}\left[\mathbf{V}_0\mathbf{\Lambda}_0\mathbf{U}_0^{\mathrm{T}}\{\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}^*(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})\}\right]\\ &= \mathrm{tr}(\mathbf{V}_0\mathbf{\Lambda}_0\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}}) = \mathrm{tr}(\mathbf{V}_0\mathbf{\Lambda}_0\mathbf{V}_0^{\mathrm{T}})\\ &= \mathrm{tr}(\mathbf{\Lambda}_0).\end{aligned}$$

In the expression $\mathbf{X} = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{Q}(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}})$, the term $\mathbf{V}_0\mathbf{V}_0^{\mathrm{T}}$ is the projection matrix onto the space spanned by the $q^*$ orthonormal columns of $\mathbf{V}_0$ given by $\mathbf{V}_0(\mathbf{V}_0^{\mathrm{T}}\mathbf{V}_0)^{-1}\mathbf{V}_0^{\mathrm{T}} = \mathbf{V}_0\mathbf{I}_{q*}^{-1}\mathbf{V}_0^{\mathrm{T}} = \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}}$ . Recall that $\mathbf{I}_q = \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{V}_1\mathbf{V}_1^{\mathrm{T}}$. Then, $\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}} = \mathbf{V}_1\mathbf{V}_1^{\mathrm{T}}$ is also the projection matrix onto the space spanned by the $q - q^*$ orthonormal columns of $\mathbf{V}_1$. In the term $\mathbf{Q}(\mathbf{I}_q - \mathbf{V}_0\mathbf{V}_0^{\mathrm{T}}) = \mathbf{Q}\mathbf{V}_1\mathbf{V}_1^{\mathrm{T}}$, each row of $\mathbf{Q}$ is projected onto the space spanned by $\mathbf{V}_1$.

An arbitrary property of $\mathbf{U}_1$ and $\mathbf{V}_1$ in $\mathbf{X} = \mathbf{U}\mathbf{V}^{\mathrm{T}} = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}_1\mathbf{V}_1^{\mathrm{T}}$ of ten Berge's (1983) solution similar to that with $\mathbf{Q}$ is that $\mathbf{U}_1$ and $\mathbf{V}_1$ can be replaced by $\mathbf{U}_1\mathbf{W}_{\mathrm{U}1}$ and $\mathbf{V}_1\mathbf{W}_{\mathrm{V}1}$, respectively, where $\mathbf{W}_{\mathrm{U}1}$ and $\mathbf{W}_{\mathrm{V}1}$ are arbitrary $(q - q^*) \times (q - q^*)$ orthonormal matrices yielding $\mathbf{X}^* \equiv \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}_1\mathbf{W}_{\mathrm{U}1}(\mathbf{V}_1\mathbf{W}_{\mathrm{V}1})^{\mathrm{T}} \neq \mathbf{X} = \mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} + \mathbf{U}_1\mathbf{V}_1^{\mathrm{T}}$.

For completeness, arbitrary aspects in $\mathbf{A} = \mathbf{U}_0\mathbf{\Lambda}_0\mathbf{V}_0^{\mathrm{T}}$ of rank $q^* \le q$ are noted. Under the standard definition of $\mathbf{\Lambda}_0 = \mathrm{diag}(\lambda_1, ..., \lambda_{q^*})$ with $\lambda_1 \ge \cdots \ge \lambda_{q^*} \ge 0$, $\mathbf{\Lambda}_0$ is identified while $\mathbf{U}_0$ and $\mathbf{V}_0$ are identified up to the sign changes (orientations or reflections) of the pairs of their corresponding columns. Further, suppose that some positive singular values are multiple e.g., $\lambda_j = \lambda_{j+1} = \cdots = \lambda_{j+k-1} \equiv \lambda_{j(k)}$ with multiplicity $k(>1)$, we have

$$\begin{aligned}\mathbf{A} &= \mathbf{U}_{0(-k)}\mathbf{\Lambda}_{0(-k)}\mathbf{V}_{0(-k)}^{\mathrm{T}} + \mathbf{U}_{0(k)}\mathbf{\Lambda}_{0(k)}\mathbf{V}_{0(k)}^{\mathrm{T}} \\ &= \mathbf{U}_{0(-k)}\mathbf{\Lambda}_{0(-k)}\mathbf{V}_{0(-k)}^{\mathrm{T}} + \mathbf{U}_{0(k)}\lambda_{j(k)}\mathbf{I}_k\mathbf{V}_{0(k)}^{\mathrm{T}} \\ &= \mathbf{U}_{0(-k)}\mathbf{\Lambda}_{0(-k)}\mathbf{V}_{0(-k)}^{\mathrm{T}} + \lambda_{j(k)}\mathbf{U}_{0(k)}\mathbf{W}_{(k)}(\mathbf{V}_{0(k)}\mathbf{W}_{(k)})^{\mathrm{T}},\end{aligned}$$

where $\mathbf{U}_{0(k)}$ and $\mathbf{V}_{0(k)}$ are semiorthonormal submatrices of $\mathbf{U}_{0(k)}$ and $\mathbf{V}_{0(k)}$, respectively corresponding to the multiple $\lambda_{j(k)}$ with $\mathbf{\Lambda}_{0(k)}$ defined similarly; and $\mathbf{U}_{0(-k)}$, $\mathbf{V}_{0(-k)}$ and $\mathbf{\Lambda}_{0(-k)}$ are matrices given by using the singular values except $\lambda_j = \lambda_{j+1} = \cdots = \lambda_{j+k-1}$; and $\mathbf{W}_{(k)}$ is an arbitrary $k \times k$ orthonormal matrix. The last arbitrary property is similar to that in the so-called "rotational indeterminacy" in factor analysis.

## References

Adachi, K. (2016). *Matrix-based introduction to multivariate data analysis*. Singapore: Springer Singapore.

Fan, K. (1951). Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences*, *37* (11), 760-766.

Green Jr, B. F. (1969). Best linear composites with a specified structure. *Psychometrika*, *34* (3), 301-318.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1934). *Inequalities*. New York: Cambridge University Press.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1952). *Inequalities* (2nd ed.). New York: Cambridge University Press. Reprinted 1994.

Kiers, H. A., & ten Berge, J. M. (1989). Optimality conditions for the trace of certain matrix products. *Linear Algebra and its Applications*, *126*, 125-134.

Kristof, W. (1964). Die beste orthogonale Transformation zur gegenseitigen Überführung zweier Faktorenmatrizen. *Diagnostica*, *10*, 87-90.

Kristof, W. (1969, March). A theorem on the trace of certain matrix products and some applications. *ETS Research Bulletin, RB-69-21*, Educational Testing Service, Princeton, NJ. https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1969.tb00400.x, https://doi.org/10.1002/j.2333-8504.1969.tb00400.x

Kristof, W. (1970). A theorem on the trace of certain matrix products and some applications. *Journal of Mathematical Psychology*, *7* (3), 515-530.

Levin, J. (1979). Applications of a theorem on the traces of certain matrix products. *Multivariate Behavioral Research*, *14* (1), 109-113.

Löwner, C. (1934). Über monotone Matrixfunctionen. *Mathematische Zeitschrift*, *38*, 177-216.

Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of majorization and its applications* (2nd ed.). New York: Springer.

Meredith, W. (1964). Rotation to achieve factorial invariance. *Psychometrika*, *29* (2), 187-206.

Miranda, H., & Thompson, R. C. (1993). A trace inequality with a subtracted term. *Linear Algebra and its Applications*, *185*, 165-172.

Mirsky, L. (1975). A trace inequality of John von Neumann. *Monatshefte für mathematik*, *79* (4), 303-306.

Mori, K., & Kurata, H. (2013). Optimal correlation preserving linear predictors of factor scores in factor analysis. *Journal of the Japan Statistical Society*, *43*(1), 79-89.

Neudecker, H. (2004). On best affine unbiased covariance-preserving prediction of factor scores. *SORT, 28* (1), 27-36.

Ogasawara, H. (2000). Some relationships between factors and components. *Psychometrika, 65,* 167-185 (errata, *65*, 551).

Schönemann, P. H. (1968). On two-sided orthogonal Procrustes problems. *Psychometrika*, *33* (1), 19-33.

Simon, B. (2005). *Trace ideals and their applications* (2nd ed., No. 120). Providence, RI: American Mathematical Society.

ten Berge, J. M. (1983). A generalization of Kristof's theorem on the trace of certain matrix products. *Psychometrika*, *48* (4), 519-523.

ten Berge, J. M. (1993). *Least squares optimization in multivariate analysis*. Leiden, The Netherlands: DSWO Press, Leiden University. (pdf version with minor corrections, 2005, https://kiers.webhosting.rug.nl/leastsquaresbook.pdf)

Uchida, T. (2023, December 11). *von Neumann's trace inequality*. Unpublished document available at https://note.com/utaka233/n/n511cbf78f89d. (in Japanese).

von Neumann, J. (1937). Some matrix-inequalities and metrization of matric-space. *Tomsk University Review*, *1*, 286-300. Reprinted in A. H. Taub (Ed.), *John von Neumann collected works*, Vol. IV (pp. 205-219). New York: Pergamon (1962).

Waller, N. (2018). An introduction to Kristof's theorem for solving least-square optimization problems without calculus. *Multivariate Behavioral Research*, *53* (2), 190-198.

Yanai, H., & Takane, Y. (2007). Matrix methods and their applications to factor analysis. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 345-366). Amsterdam: North-Holland.