Rephrasing the Lengthy and Involved Proof of Kristof’s Theorem: A Tutorial with Some New Findings

Haruhiko Ogasawara

Otaru University of Commerce, Otaru, Japan
emt-hogasa@emt.otaru-uc.ac.jp

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 Jr (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 (1951)’s 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 the 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 $\mathbf{A}\mathbf{D}$ and $\mathbf{D}\mathbf{A}$ 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, and 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 \( a^*_1, \ldots, a^*_m \) and \( b^*_1, \ldots, b^*_m \) 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 \( b^*_k \) 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 \( a_i b_j + a_j b_i - (a_i b_j + a_j b_i) \geq 0 \) \((1 \leq i \leq j \leq m)\). Using this possibly exchanged permutation, redefine \( b^*_1, \ldots, b^*_m \). Then, when \( b_2^* \neq b_2 \), exchange \( b_2^* \) and \( b^*_k = b_2 \). Repeat this process until the possible exchange of \( b^*_m-1 \) and \( b_m^* \) when \( b_{m-1}^* \neq b_m^* \). 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 \( b_1^*, \ldots, b_m^* \) using pairwise exchanges after the final one increases \( \sum_{i=1}^{m} a_i b_i^* \). \( \square \)

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 the proof of Simon (2005, Lemma 1.8., p. 4), 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 \( a^{(j)}_1 \geq \cdots \geq a^{(j)}_m \geq 0 \) \((j = 1, \ldots, n; n \geq 2)\), let \( a^{(j)*}_1, \ldots, a^{(j)*}_m \) be an arbitrary case in the \( j \)-th set of \( m! \) permutations including possibly the same ones. Then, the maximum of \( \sum_{i=1}^{m} a^{(1)*}_i \cdots a^{(n)*}_i \) over the permutations is given by \( \sum_{i=1}^{m} a^{(1)}_i \cdots a^{(n)}_i \).

**Proof.** Consider the case of \( n = 3 \). For two sets \( a^{(j)*}_1, \ldots, a^{(j)*}_m \) \((j = 1, 2)\) in the three sets, Lemma 1 gives the maximum of the sum of the products as \( \sum_{i=1}^{m} a^{(1)}_i a^{(2)*}_i \). Then, for the two sets of \( m \) products \( a^{(1)}_1 a^{(2)*}_1, \ldots, a^{(1)}_m a^{(2)*}_m \) and \( m \) numbers \( a^{(3)*}_1, \ldots, a^{(3)*}_m \), the maximum of \( \sum_{i=1}^{m} a^{(1)*}_i a^{(2)*}_i a^{(3)*}_i \) over the \( m! \) permutation in the third set is similarly obtained by \( \sum_{i=1}^{m} a^{(1)}_i a^{(2)*}_i a^{(3)*}_i \). Since any permutation for the maximized one including the first two sets decreases the product sum or remains unchanged as seen in Lemma 1, \( \sum_{i=1}^{m} a^{(1)*}_i a^{(2)*}_i a^{(3)*}_i \) is the global maximum. The cases with \( n \geq 4 \) is similarly obtained. \( \square \)
Lemma 2: The same ranges of the traces irrespective of their absolute values of the diagonal elements with permutation (Kristof, 1970, Lemma 2). Let \( \Gamma^*_1 \) and \( \Gamma^*_2 \) be \( m \times m \) diagonal matrices; and \( \Gamma_1 \) and \( \Gamma_2 \) be those with the corresponding diagonal elements replaced by their absolute values located in the weakly descending (non-increasing) orders, respectively. Suppose that \( X_1 \) and \( X_2 \) independently vary over all the \( m \times m \) orthonormal matrices. Then, \( \text{tr}(X_1 \Gamma^*_1 X_2 \Gamma^*_2) \) has the same range as that of \( \text{tr}(X_1 \Gamma_1 X_2 \Gamma_2) \).

Proof. Kristof’s derivation is didactically repeated. Note that \( \Gamma_i \) is obtained by \( \Gamma_i = P_i S_i \Gamma_i^* P_i^T \), where \( S_i \) is the signed identity matrix replacing the diagonal elements of \( \Gamma_i^* \) by their corresponding absolute values, and \( P_i \) is the permutation matrix to have the weakly descending order mentioned earlier \((i = 1, 2)\). Noting that \( \Gamma_i^* = S_i P_i^T \Gamma_i P_i \), we have

\[
\text{tr}(X_1 \Gamma^*_1 X_2 \Gamma^*_2) = \text{tr}\{X_1 (S_1 P_1^T \Gamma_1 P_1) X_2 (S_2 P_2^T \Gamma_2 P_2)\}
= \text{tr}\{(P_2 X_1 S_1 P_1^T) \Gamma_1 (P_1 X_2 S_2 P_2^T) \Gamma_2\}.
\]

In the last result, \( P_2 X_1 S_1 P_1^T \) and \( P_1 X_2 S_2 P_2^T \) are products of orthonormal matrices and consequently orthonormal with the same variations of \( X_1 \) and \( X_2 \), which shows the required same ranges of \( \text{tr}(X_1 \Gamma^*_1 X_2 \Gamma^*_2) \) and \( \text{tr}(X_1 \Gamma_1 X_2 \Gamma_2) \).

The following lemma gives a derivation of Kristof’s theorem shown later when \( n = 1 \), which will also be used in the cases when \( n = 2, 3, \ldots \) in the derivation of the theorem by induction.

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

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

In the above proof, the identity matrix maximizing the trace is didactically obtained in two ways as follows. (i) Note that \( \text{tr}(V^T U \Lambda) = \sum_{i=1}^m v_i^T u_i \lambda_i \), where \( v_i \) and \( u_i \) are the \( i \)-th columns of \( V \) and \( U \), respectively with \( v_i^T v_i = u_i^T u_i = 1 \) \((i = 1, \ldots, m)\) by assumption. Using \( v_i^T u_i = v_i^T u_i / \{(v_i^T v_i)^{1/2}(u_i^T u_i)^{1/2}\} \leq 1 \) \((i = 1, \ldots, m)\) by the Cauchy-Schwarz (C-S) inequality, we have \( \text{tr}(V^T U \Lambda) = \sum_{i=1}^m v_i^T u_i \lambda_i \leq \sum_{i=1}^m \lambda_i \). Since the maximum in the C-S inequality is obtained when \( v_i = u_i \) \((i = 1, \ldots, m)\), we have \( V^T U = V^T V = U^T U = I_m \) (the \( m \times m \) identity matrix) for maximizing \( \text{tr}(V^T U \Lambda) \).
(ii) The second derivation is given without using the C-S inequality. Since $W = \{w_{ij}\} \equiv V^T U$ is orthonormal as seen from $(V^T U)^T (V^T U) = U^T V V^T U = U^T U = I_m$, we have $w_{ij} \leq 1$ ($i, j = 1, \ldots, m$). Consequently, $\text{tr}(V^T U A) = \sum_{i=1}^m w_{ii} \lambda_i \leq \sum_{i=1}^m \lambda_i = \text{tr}(\Lambda)$, where the maximum is attained when $w_{ii} = 1$ ($i = 1, \ldots, m$), which is the case of $W = V^T U = I_m$.

(iii) It is of interest to find that (ii) gives the C-S inequality in (i) via Kristof’s theorem, which will be addressed later as an application of Kristof’s theorem.

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

**Proof.** By assumption, $AD = (AD)^T = DA^T$ and $DA = (DA)^T = A^T D$. From the last equation, we have $D A D^{-1} = A^T$. Using $A^T = D A D^{-1}$ on the right-hand-side of the first equation $A D = D A^T$ and post-multiplying $D^{-1}$ on both sides of the equation, we obtain $A = D^2 A D^{-2}$, which indicates that

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

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

When $A$ is diagonal, we have $AD = DA$, where $A$ and $D$ 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 $X_i$, $D_i$ and $\Delta_i$ be $m \times m$ matrices, where $X_i$ is orthonormal while $D_i$ and $\Delta_i$ are diagonal and of full rank ($i = 1, 2$). Suppose that

$$X_1 D_1 X_2 = \Delta_1 \quad \text{and} \quad X_2 D_2 X_1 = \Delta_2.$$

Then, $X_1$ and $X_2$ 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, $X_1$ and $X_2$ are diagonal matrices with their diagonal elements being $\pm 1$, and $\Delta_1 D_2 = \Delta_2 D_1$.

**Proof.** Left-multiplying $X_1 D_1$ on both sides of $X_2 D_2 X_1 = \Delta_1$ using $X_1 D_1 X_2 = \Delta_1$, we have $\Delta_1 D_2 X_1 = X_1 D_1 \Delta_2$ giving $\Delta_1 D_2 = X_1 D_1 \Delta_2 X_1^T$. The last result shows the spectral decomposition of the diagonal matrix $\Delta_1 D_2$, which indicates that the orthonormal matrix $X_1$ becomes a signed and/or permuted identity matrix and when without permutation $\Delta_1 D_2 = \Delta_2 D_1$. For $X_2$, exchanging the subscripts “1” and “2” due to symmetry, we obtain $\Delta_2 D_1 = X_2 D_2 \Delta_1 X_1^T$ indicating the same results as for $X_1$. \qed
Theorem 1: Two-fold or bilinear case \((n = 2)\) (von Neumann (1937, Theorem 1); Kristof (1970, Theorem (first version))). Let \(\Gamma_1^*\) and \(\Gamma_2^*\) 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 \(\text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*)\) which are attained, where \(X_1\) and \(X_2\) independently vary over all \(m \times m\) orthonormal matrices. Then,

\[-\text{tr}(\Gamma_1\Gamma_2) \leq \text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*) \leq \text{tr}(\Gamma_1\Gamma_2),\]

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

Proof. By Lemma 2, since the range of \(\text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*) = \text{tr}(\Gamma_2^*X_1\Gamma_1X_2)\) is the same as that of \(\text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*)\), we consider the former range. Due to Lemma 3, when the former trace is maximized, \(X_1^*\Gamma_1^*X_2\Gamma_2^*\) and similarly \(\Gamma_2^*X_1\Gamma_1X_2\) become symmetric. Then, using Lemma 4 we find that \(X_1^*\Gamma_1^*X_2\Gamma_2^*\) is diagonal. Due to symmetry with \(\text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*) = \text{tr}(\Gamma_2^*X_1\Gamma_1X_2)\), \(X_2\Gamma_2^*X_1\) is also found to be diagonal. By Lemma 5, these two diagonal conditions give the maximum \(\text{tr}(\Gamma_1\Gamma_2)\) of \(\text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*)\) when \(X_1\) and \(X_2\) are the same signed identity matrices. The global maximum \(\text{tr}(\Gamma_1\Gamma_2)\) among the permuted diagonal elements of \(\Gamma_1\) and \(\Gamma_2\) is shown by Lemma 1. The minimum is given by replacing e.g., \(X_1\) by \(-X_1\).

\(\square\)

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 \(\text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*) = \text{tr}(P\Delta Q^T) = \text{tr}(PQ^T\Delta)\), whose optima are attained when \(P = Q\) or \(P = -Q\) as \(\text{tr}(\Delta) \leq \text{tr}(X_1^*\Gamma_1^*X_2\Gamma_2^*) \leq \text{tr}(\Delta)\). This result is correct. Then, Waller (2018, Equation (22)) gave inequalities \(\text{tr}(\Gamma_1\Gamma_2) \leq \text{tr}(\Delta) \leq \text{tr}(\Gamma_1\Gamma_2)\) 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 \(\text{tr}(\Delta) = \text{tr}(\Gamma_1\Gamma_2)\), 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 trilinear 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 $A, X, D_i$ and $\Delta_i$ be $m \times m$ matrices of full rank, where $X$ is orthonormal while $D_i$ and $\Delta_i$ are diagonal ($i = 1, 2$). Suppose that

$$AD_1X = \Delta_1 \text{ and } XD_2A = \Delta_2.$$  

Then, $X$ 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, $X$ is diagonal with its diagonal elements being $\pm 1$, and $\Delta_1D_2 = \Delta_2D_1$.

**Proof.** Right-multiplying $D_1X$ on both sides of $XD_2A = \Delta_2$ using $AD_1X = \Delta_1$, we have $XD_2\Delta_1 = \Delta_2D_1X$ giving $\Delta_2D_1 = X\Delta_1D_2X^T$. The last result shows the spectral decomposition of the diagonal matrix $\Delta_2D_1$, which indicates that the orthonormal matrix $X$ becomes a signed and/or permuted identity matrix and when without permutation $\Delta_1D_2 = \Delta_2D_1$. \hfill \Box

**Remark 2** Lemma 5 is seen as a special case of Lemma 6 when $A = X_1$ an orthonormal matrix and $X$ is denoted by $X_2$. However, in Lemma 5, both $X_1$ and $X_2$ were found to be signed and/or permuted identity matrices. Note also that Kristof (1970) dealt with the case when $A = G_1\Gamma_1G_2\Gamma_2 \cdots G_{n-1}\Gamma_{n-1}G_n$, $D_1 = \Gamma_n$, $D_2 = \Gamma_{n+1}$ and $X = G_{n+1}$, where $\Gamma_j$ and $G_i$ are diagonal and orthonormal matrices, respectively. This specification was necessary for his derivation by induction though the involved expression $G_1\Gamma_1G_2\Gamma_2 \cdots G_{n-1}\Gamma_{n-1}G_n$ may hide the basic structure in Lemma 6.

**Theorem 2**: Three-fold or trilinear case ($n = 3$) (Kristof, 1970, Theorem (first version)). Let $\Gamma_i^*$ ($i = 1, 2, 3$) 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 $\text{tr}(X_1\Gamma_1^*X_2\Gamma_2^*X_3\Gamma_3^*)$ which are attained, where $X_i$ ($i = 1, 2, 3$) independently vary over all $m \times m$ orthonormal matrices. Then,

$$-\text{tr}(\Gamma_1^*\Gamma_2^*\Gamma_3^*) \leq \text{tr}(X_1\Gamma_1^*X_2\Gamma_2^*X_3\Gamma_3^*) \leq \text{tr}(\Gamma_1^*\Gamma_2^*\Gamma_3^*),$$

where $\Gamma_i = \text{diag}(\gamma_{i1}, ..., \gamma_{im})$ is given by $\Gamma_i^*$ 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 the proof for Theorem 1, using Lemma 2 in a similar manner with $\Gamma_i = P_iS_i\Gamma_i^*P_i^*$ and unconstrained orthonormal $X_i$ ($i = 1, 2, 3$), the range of $\text{tr}(X_1\Gamma_1^*X_2\Gamma_2^*X_3\Gamma_3^*)$ ($= \text{tr}(\Gamma_3^*X_1\Gamma_1^*X_2\Gamma_2^*X_3^*)$) is found to be the same as that of $\text{tr}(X_1\Gamma_1^*X_2\Gamma_2^*X_3\Gamma_3^*)$. Due to Lemma 3, when the former trace is maximized, $X_1\Gamma_1^*X_2\Gamma_2^*X_3\Gamma_3^* \equiv BF_3$ and similarly $\Gamma_3^*X_1\Gamma_1^*X_2\Gamma_2^*X_3^* = \Gamma_3^*B$ become symmetric. Then, using Lemma 4 we find that $B = X_1\Gamma_1^*X_2\Gamma_2^*X_3^* \equiv A\Gamma_2^*X_3^*$ is diagonal. Similarly, since $\text{tr}(X_1\Gamma_1^*X_2\Gamma_2^*X_3\Gamma_3^*) = \text{tr}(A\Gamma_2^*X_3^*) = \text{tr}(A\Gamma_2^*X_3^*) = \text{tr}(\Gamma_1^*\Gamma_2^*\Gamma_3^*) \leq \text{tr}(\Gamma_1^*\Gamma_2^*\Gamma_3^*)$.
tr(X₃Γ₃AΓ₂), X₃Γ₃A is also diagonal. From Lemma 6, these two diagonal conditions can make X₃ an identity matrix. Then, using Theorem 1, the maximum tr(X₁Γ₁X₂Γ₂X₃Γ₃) is obtained when X₁ and X₂ are identity matrices as tr(Γ₁Γ₂Γ₃). The minimum −tr(Γ₁Γ₂Γ₃) is obtained as in Theorem 1.

Remark 2A In the proof of Theorem 2, the key result is tr(X₁Γ₁X₂Γ₂X₃Γ₃) = tr(X₁Γ₁X₂Γ₂Γ₃). Since Γ₂Γ₃ ≡ Γ₂₃ is diagonal, the maximum of tr(X₁Γ₁X₂Γ₂Γ₃) = tr(X₁Γ₁X₂Γ₂₃) 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: The n-fold case (n = 2, 3,...) (Kristof, 1970, Theorem (first version)). Let Γᵢ (i = 1,...,n) be fixed diagonal matrices. Consider the maximum and minimum of tr(X₁Γ₁ᵢXₙΓₙ) which are attained, where Xᵢ(i = 1,...,n) independently vary over all orthonormal matrices. Then,

\[ -\text{tr}(Γ₁ \cdots Γₙ) \leq \text{tr}(X₁Γ₁ᵢ \cdots XₙΓₙ) \leq \text{tr}(Γ₁ \cdots Γₙ), \]

where Γᵢ = diag(γᵢ₁, ..., γᵢₘ) is given by Γᵢ with their diagonal elements replaced by the corresponding absolute values with possible permutation to have the weakly descending order i.e., γᵢ₁ ≥ ... ≥ γᵢₘ ≥ 0 (i = 1,...,n).

Proof. As in Kristof (1970), first suppose that γᵢ₁ > ... > γᵢₘ > 0 (i = 1,...,n). Using the result of Theorem 2, increase n one by one as n = 4, 5,... When n = 4, redefine B ≡ X₂Γ₂X₃Γ₃X₄ ≡ AΓ₃X₄. Then, from Lemma 4, B = AΓ₃X₄ becomes diagonal. Similarly, X₄Γ₄A is also diagonal. Then, as before X₄ can become an identity matrix. Using Theorem 2, the maximum of tr(X₁Γ₁ᵢX₄Γ₄) = tr(X₁Γ₁ᵢX₄Γ₄) is given by tr(Γ₁ᵢX₄). The minimum is similarly obtained as −tr(Γ₁ᵢX₄). Increasing n successively one by one, we obtain the required results.

Further, consider the weakly ordered case γᵢ₁ ≥ ... ≥ γᵢₘ ≥ 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 W = diag(w₁,...,wₘ) with w₁ > ... > wₘ > 0. Redefine Γᵢ as Γᵢ + εW(i = 1,...,n) with ε > 0. Then, we have the same above result since γᵢ₁ > ... > γᵢₘ > 0. When ε approaches zero with fixed Xᵢ(i = 1,...,n), the same required result is given by substituting ε = 0 for Γᵢ + εW under the limiting condition of γᵢ₁ ≥ ... ≥ γᵢₘ ≥ 0 (i = 1,...,n).

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 γᵢ₁ ≥ ... ≥ γᵢₘ ≥ 0. Though when γᵢ₁ ≥ ... ≥ γᵢₘ = 0, which is given in the limiting case of ε = 0, Γᵢ becomes singular while Γᵢ + εW with ε > 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 $X_i$’s with $\text{rank}(X_i) = r_i^* \leq r_i$ are $m_{i-1} \times m_i$ possibly non-square suborthonormal matrices i.e., submatrices of orthonormal ones ($i = 1, ..., n$; $m_0 = m_n$). Let a semiorthonormal matrix be a non-square submatrix of its parent orthonormal one with the same number of the rows or columns (not both) as that of the parent. Note that if $X_i$ is suborthonormal rather than semi- or fully orthonormal, $r_i^*$ can be 0. For this result, we consider the restricted parent orthonormal matrix taking a block-diagonal form $X^*_i = \begin{pmatrix} X_{i11} & X_{i12} \\ X_{i21} & X_{i22} \end{pmatrix} = \begin{pmatrix} X_{i11} & 0 \\ 0 & X_{i22} \end{pmatrix}$.

where $X_i^T X_i^* = X_i^* X_i^T = I_{m_{i-1} + m_i}$, $X_{i11}^T X_{i11} = X_{i11}^T X_{i11}^T = I_{m_{i-1}}$ and $X_{i22}^T X_{i22} = X_{i22}^T X_{i22}^T = I_{m_i}$. Then, when $X_i^*$ is one of the two off-block-diagonal zero submatrices $X_{i12} = X_{i21} = 0$, the rank of $X_i^*$ is zero i.e., $r_i^* = 0$. It is assumed that $X_i^*$ varies unrestrictedly under the block-diagonal form. 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 $X_i$ varies unrestrictedly. In other words, when $r_i < \min\{m_{i-1}, m_i\}$, $X_i$ does not vary unrestrictedly. If $X_i$ is semi- or fully orthonormal, we have $r_i^* = r_i = \min\{m_{i-1}, m_i\}$.

Let $\Gamma_i^*$ and $\Gamma_i$ 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 $m = \max(m_1, ..., m_n)$. Define $\Delta_i$ and $E_r$ as the $m \times m$ diagonal matrices containing $\Gamma_i$ and $I_e$ in their upper left corners with zeroes elsewhere, respectively. Then, Ten Berge gave the following result.

**Theorem 4:** The generalized Kristof theorem (Ten Berge (1983, Theorem 1); see also Kiers and Ten Berge (1989); and Ten Berge (1993, Sections 3.2 and 3.3)). Under the definitions and assumptions given above, when $X_i$ varies with the condition $\text{rank}(X_i) \leq r_i$ ($i = 1, ..., n$), we have

$$-\text{tr}(\Delta_1 \cdots \Delta_n E_r) \leq \text{tr}(X_1 \Gamma_1^* \cdots X_n \Gamma_n^*) \leq \text{tr}(\Delta_1 \cdots \Delta_n E_r).$$

For the proof of Theorem 4, Ten Berge (1983) defined $Y_i$ as the $m \times m$ matrix containing $X_i$ in its upper left corner with zeroes elsewhere. Then, he defined its SVD as $Y_i = P_i D_i Q_i^T$, where “$P_i$ and $Q_i$ are orthonormal and $D_i$ is diagonal” (loc.cit., p. 521). Note that he employed the SVD using the non-negative singular values rather than the positive ones, where $P_i$, $Q_i$, and $D_i$ are $m \times m$ square matrices. This is seen in his inequalities (loc.cit., Equation (10))

$$-\text{tr}(D_1 \Delta_1 \cdots D_n \Delta_n) \leq \text{tr}(X_1 \Gamma_1^* \cdots X_n \Gamma_n^*) \leq \text{tr}(D_1 \Delta_1 \cdots D_n \Delta_n).$$

For this derivation, he used $Y_i = P_i D_i Q_i^T$ and $\Delta_i^*$ defined similarly to $\Delta_i$ using $\Gamma_i^*$ ($i = 1, ..., n$), which gives

$$\text{tr}(X_1 \Gamma_1^* \cdots X_n \Gamma_n^*) = \text{tr}(Y_1 \Delta_1^* \cdots Y_n \Delta_n^*) = \text{tr}(P_1 D_1 Q_1^T \Delta_1^* \cdots P_n D_n Q_n^T \Delta_n^*)$$
He applied Kristof’s theorem to this result supposing that all the $2^n m \times m$ matrices $P_i$ and $Q_i$ ($i = 1, \ldots, n$) 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 $D_i = \text{diag}(d_1, \ldots, d_m)$ with $d_1 \geq \cdots \geq d_r^* > 0$ and $d_{r+1} = \cdots = d_m = 0$, orthonormal matrices $P_i$ and $Q_i$ in $Y_i = P_i D_i Q_i^T$ should be of the form

$$P_i = \begin{pmatrix} P_{i1} & 0 \\ O & P_{i2} \end{pmatrix} \quad \text{and} \quad Q_i = \begin{pmatrix} Q_{i1} & 0 \\ O & Q_{i2} \end{pmatrix}$$

where $P_{i1}$ and $Q_{i1}$ are $r_i^* \times r_i^*$ orthonormal submatrices while $P_{i2}$ and $Q_{i2}$ are $(m-r_i^*) \times (m-r_i^*)$ similar ones unless vanishing when $r_i^* = m$. This formulation is due to Ten Berge’s special definition of $Y_i$ whose elements are zero except the upper left submatrix.

Although $P_i$ and $Q_i$ 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 $D_i \Delta_i$ 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 $E_r$ in the statement of Theorem 4 is due to the unrestricted rank condition of the diagonal matrix $\Gamma_i$ in the upper left corner of $\Delta_i$ employed by Ten Berge. Since the rank of $\Delta_i$ may be greater than that of $D_i$, the upper bound in his Equation (10) becomes

$$\text{tr}(X_1 \Gamma_1^* \cdots X_n \Gamma_n^*) \leq \text{tr}(D_1 \Delta_1 \cdots D_n \Delta_n)$$

$$= \text{tr}(D_1 \cdots D_n \Delta_1 \cdots \Delta_n) \leq \text{tr}(\Delta_1 \cdots \Delta_n E_r),$$

where the last inequality is due to the range $[0,1]$ of the singular values of suborthonormal matrices (loc.cit., Lemma 2) yielding $D_1 \cdots D_n \leq E_r$ 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 $X_i$ 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 $X_i$ (recall the condition $\text{rank}(X_i) = r_i^* \leq r_i$) is less than $\min\{m_{i-1}, m_i\}$, $X_i$ does not vary unrestrictedly, but varies over a subset of the suborthonormal matrices satisfying $r_i^* \leq r_i < \min\{m_{i-1}, m_i\}$. That is, in the subset, $X_i$ cannot be semi- or fully orthonormal. In other words, in this subset the sum of the squared elements in
each row or column of $X_i$ is smaller than 1. Under this restriction, the optima may not be obtained. Note also that Ten Berge 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 $X_i$ be the $m_{i-1} \times m_i$ (\( i = 1, ..., n; \ m_0 \equiv m_n \)) possibly non-square matrix with \( \text{rank}(X_i) = r^*_i \leq r_i = \min\{ m_{i-1}, m_i \} \), 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 $X_i^*$, which includes $X_i$ as a submatrix.

Let $\Gamma_i^*$ and $\Gamma_i$ be $m_i \times m_i$ fixed diagonal matrices defined as in Theorems 3 and 4. In the modification, however, $\Gamma_i^*$ and $\Gamma_i$ are assumed to be non-singular without loss of generality. This is seen from the form \( \text{tr}(X_i \Gamma_i^* X_{i+1} \Gamma_{i+1}^*) \) to be optimized later, since when $\Gamma_i^*$ is singular, $\Gamma_i^*$ can be redefined by deleting the row(s) and column(s) corresponding to the zero diagonal elements of $\Gamma_i^*$. Then, in the similar manner, the corresponding column(s) of $X_i$ and row(s) of $X_{i+1}(X_{n+1} \equiv X_1)$ can be deleted without changing the value of \( \text{tr}(X_i \Gamma_i^* X_{i+1} \Gamma_{i+1}^*) \), where $r^*_i(r^*_{i+1})$ and $r_i(r_{i+1})$ may be adjusted for the reduced $X_i(X_{i+1})$ when necessary.

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

\[
-\text{tr}(\Delta_1 \cdots \Delta_n) \leq \text{tr}(X_1 \Gamma_1^* \cdots X_n \Gamma_n^*) \leq \text{tr}(\Delta_1 \cdots \Delta_n),
\]

where the optima are attained.

\textbf{Proof}. Define $\Delta_i^*$ using $\Gamma_i^*$ similarly to $\Delta_i$. Then, we obtain

\[
\text{tr}(X_1 \Gamma_1^* \cdots X_n \Gamma_n^*) = \text{tr}(X_1^* \Delta_1^* \cdots X_n^* \Delta_n^*).
\]

Noting that the assumption of the independent and unrestricted variations of $X_i^* (i = 1, ..., n)$ satisfies that of Kristof’s theorem, the required results with the attained optima follow. \( \square \)

\textbf{Remark 4} In Theorem 5, the assumption for the variations of $X_i^*$ automatically gives \( \text{rank}(X_i) = r^*_i \leq r_i = \min\{ m_{i-1}, m_i \} \). The optima are obtained when $r^*_i = r_i$, which indicates that $X_i$ is semi- or fully orthonormal with non-zero singular value(s) being unity when the optima are attained. This makes the matrix $E_r$ 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 \( X_i \) may be less than unity and since the SVD form of \( Y_i \) restricts the permutation of the diagonal elements of \( \Gamma_i^* \) to have \( \Gamma_i \). In other words, when \( \Gamma_i^* = \Gamma_i \), the latter restriction vanishes. Note that the former restriction corresponds to \( r_i^* < r_i \). That is, under this assumption \( X_i \) 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 \( X_i \) 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 “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 \( i_B = 1 \), \( X_i \) becomes a usual orthonormal matrix as in Kristof’s theorem, and when \( i_B = m \), \( X_i \) is the signed identity matrix whose diagonal elements are ±1, where \( X_i \) and its diagonal elements i.e., ±1 are \( m \times m \) and \( 1 \times 1 \) orthonormal matrices with unit singular value(s), respectively.

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 \( X_i(i = 1, \ldots, n) \) as

\[
X_i = Bdiag(X_{i1}, \ldots, X_{in}),
\]

where \( X_{ij} \) is the \( i_j \times i_j \) diagonal block \( (j = 1, \ldots, i_B) \) with \( i_1 + \ldots + i_{i_B} = m \). Suppose that \( X_i(i = 1, \ldots, n) \) independently and unrestrictedly vary with \( \text{rank}(X_{ij}) = i_j \) \( (j = 1, \ldots, i_B) \). Use \( \Gamma_i(i = 1, \ldots, n) \) as defined in Kristof’s theorem. Then, we have

\[
-\text{tr}(\Gamma_1 \cdots \Gamma_n) \leq \text{tr}(X_1 \Gamma_1 \cdots X_n \Gamma_n) \leq \text{tr}(\Gamma_1 \cdots \Gamma_n),
\]

where the optima are attained.

**Proof.** The case when \( \Gamma_i^* = \Gamma_i(i = 1, \ldots, n) \) can be used in Kristof’s theorem. Although \( X_i(i = 1, \ldots, n) \) take the block-diagonal forms, the optima with \( X_i = I_m \) \( (i = 1, \ldots, n) \) are attained in the subset of varying \( X_i(i = 1, \ldots, n) \) under the block diagonal restriction. □

The usage of \( \Gamma_i \) in place of \( \Gamma_i^*(i = 1, \ldots, n) \) is due to the lack of permutation across the different diagonal blocks when using the block diagonal \( X_i \). Note that Kiers and Ten Berge (1989) used the assumptions that \( \Gamma_i \)'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 \( i_B = 1 \), \( X_i \) becomes a usual orthonormal matrix as in Kristof’s theorem, and when \( i_B = m \), \( X_i \) is the signed identity matrix whose diagonal elements are ±1, where \( X_i \) and its diagonal elements i.e., ±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 $\Gamma_i$ to be an unconstrained square matrix of the same rank as that of $\Gamma_i$ ($i = 1, ..., n$), which is mathematically immaterial. Since under the relaxed condition the SVD of $\Gamma_i \equiv U_i A_i V_i^T$ with $U_i^T U_i$ and $V_i^T V_i$ being an identity matrix of the same order as rank($\Gamma_i$) ($i = 1, ..., n$), redefine $X_1$ and $X_i$ as $V_n X_1 U_1^T$ and $V_{i-1} X_i U_i^T$ ($i = 2, ..., n$). Then, the problem becomes the same as that using the diagonal matrix $\Lambda_i$ ($i = 1, ..., n$). 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 $\Gamma_i$.

5 Applications of Kristof’s theorem and its generalizations

While as mentioned in the introductory section, “underutilization” of Kristof’s theorem and its generalization seem 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 the Cauchy-Schwarz inequality using his theorem are the cases with attained optima.

Maximization of $\text{tr}(MAL)$ or $\text{tr}(A^T A)$; Green Jr (1969, Appendix B) ($n = 1$). In this problem, $M, L$ and $A$ are fixed $m \times r, s \times m$ and $r \times s$ ($r \geq s$) matrices, respectively while $\Lambda$ 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 $A = (LM)^T$. 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 $\text{tr}\{(A - BT)(A - BT)^T\}$, where $A$ and $B$ are fixed $r \times s$ ($r \geq s$) matrices while $T$ is an orthonormal matrix to be derived, which reduces to maximizing $\text{tr}(A^T BT)$, a case with $n = 1$. Note that $T$ gives permutations with sign changes (reflections) of the columns of $B$ as well as the rotation of $B$. 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 that $T$ can give permutations and reflections of the columns of $B$.

On the other hand, we have a problem of “procrustes transformation” without rotation. This happens in e.g., simulations when $B$ is one of sample loading matrices rotated by a fixed method e.g., the geomin, which is to be matched to the population geomin-rotated $A$ to see the sampling variation of $B$, where
only permutations and reflections of the columns of $B$ are possible while unconstrained rotation is not allowed since otherwise $BT$ no longer becomes a geomin-rotated loading matrix. This problem is seen as a subproblem of Kristof’s theorem when $n = 1$ using a constrained orthonormal matrix $T$ considering only permutations and reflections of the columns of $B$. Problems using constrained orthonormal matrices are included in Ten Berge’s (1983) generalized Kristof theorem, as addressed earlier.

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

\[
\begin{align*}
\max \{ \text{tr}(A^T B T) \} &= \max \{ \text{tr}\{ U \text{diag}(\gamma_1, \ldots, \gamma_{s^*}) V^T T \} \} \\
&= \max \{ \text{tr}\{ \text{diag}(\gamma_1, \ldots, \gamma_{s^*}) V^T T U \} \} = \text{tr}\{ \text{diag}(\gamma_1, \ldots, \gamma_{s^*}) \} \\
&= \sum_{i=1}^{s^*} \gamma_i,
\end{align*}
\]

where $V^T U$ is a suborthonormal matrix since the product of suborthonormal matrices is suborthonormal (Ten Berge, 1983, Lemma 4). Actually, $U$ and $V$ of full column rank by definition are semiorthonormal (Ten Berge, 1983, Definition 2) while $T$ 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 the 1960s in psychometrics.

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

**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 $\text{tr}\{ (Y - XB)(Y - XB)^T \}$, where $Y$ is an $s \times u$ matrix for $u$ dependent variables and $X$ of full column rank is an matrix for $t$ independent variables. The well-known solution $\hat{B} = (X^T X)^{-1} X^T Y$ 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 $R$ is a $k \times k$ correlation matrix of rank $r \leq k$. The problem is to maximize the sum of squared loadings $\text{tr}(B^T R^2 B)$, where $B$ is a $k \times p$ loading matrix with $p \leq r$ subject to the uncorrelated components with unit variances $B^T RB = I_p$. Note that an atypical assumption of the possibly singular $R$ is used.

Let $R = KAK^T$ with $K^T K = I_r$ be the SVD using the positive diagonal elements of the $r \times r$ diagonal matrix $\Lambda$. The solution can be given by maximizing

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

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

$$I_p = B^T RB = B^T KAK^T B = (B^T K \Lambda^{1/2})(\Lambda^{1/2} K^T B).$$

Noting that $B^T K \Lambda^{1/2}$ is a semiorthonormal matrix, Ten Berge’s generalized Kristof theorem gives $\text{max}\{\text{tr}(B^T R^2 B)\} = \sum_{i=1}^{p} \lambda_i$, where $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and the unrotated loading matrix $B$ is a submatrix of $K \Lambda^{-1/2}$ taking its first $p$ columns to give $B^T K \Lambda^{1/2}$ a submatrix of $I_r$ consisting its first $p$ rows.

The formulation using suborthonormal matrices $\text{tr}(B^T R^2 B)$ is of interest since the restriction $B^T RB = I_p$ 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 $R$, the maximum of $\text{tr}(B^T R^2 B)$ is obtained by Kristof’s theorem as $\text{tr}(\Lambda)$, where $\Lambda$ is the $k \times k$ diagonal matrix with positive diagonals. In this case, $\Lambda^{1/2} K^T B$ becomes $I_k$ yielding the well-known unrotated loading matrix $B = K \Lambda^{-1/2}$.

**Canonical correlations:** Kristof (1970, General comment (a)) ($n = 1$), Ten Berge (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 Cauchy-Schwarz inequality:** Ten Berge (1983, p. 522) ($n = 2$); Paragraph (iii) after Lemma 3 ($n = 1$) 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 \(x\) and \(y\) be non-zero vectors of order \(k\). Suppose that \(x\) and \(y\) vary independently and unrestrictedly. Then,

\[
x^T(x^Tx)^{-1/2}y(y^Ty)^{-1/2} = x^T(x^Tx)^{-1/2}I_ky(y^Ty)^{-1/2}1
\]

\[
= x^T(x^Tx)^{-1/2}G_2^*y(y^Ty)^{-1/2}2,
\]

where \(G_1^* = G_1 = I_k\) and \(G_2^* = G_2 = 1\). Define \(k \times k\) diagonal matrices \(\Delta_1 = \Delta_1 = I_k\) and \(\Delta_2 = E_{11}\), where the first diagonal element of \(E_{11}\) is unity elsewhere zero. Since the vectors \(x^T(x^Tx)^{-1/2}\) and \(y^T(y^Ty)^{-1/2}\) are semiorthonormal, their non-zero singular values are unity. Consequently, when applying Ten Berge’s theorem, the maximum is attained as \(tr(\Delta_1 \Delta_2) = tr(I_k E_{11}) = 1\) with minimum obtained similarly. Then, we have the Cauchy-Schwarz inequality

\[-1 \leq x^T(x^Tx)^{-1/2}y(y^Ty)^{-1/2} = \frac{x^Ty}{\sqrt{(x^Tx)\sqrt{(y^Ty)}}} \leq 1.\]

As addressed in Paragraph (iii) after Lemma 3, the inequality is given via Kristof’s theorem \((n = 1)\). The same result is obtained by using Theorem 5 of this article. Define the \(k \times k\) parent orthonormal matrices \(X\) and \(Y\), whose first rows are \(x^T(x^Tx)^{-1/2}\) and \(y^T(y^Ty)^{-1/2}\), respectively, where \(X\) and \(Y\) independently and unrestrictedly vary over the sets of orthonormal matrices. Then, using Theorem 5, the maximum of \(x^T(x^Tx)^{-1/2}y(y^Ty)^{-1/2}\) is attained as \(tr(X I_k Y E_{11}) \leq tr(I_k E_{11}) = 1\) with the minimum \(tr(X I_k Y E_{11}) \geq -tr(I_k E_{11}) = -1\) obtained similarly.

Generalized linear form: Ten Berge (1993, Equation (48)), Yanai and Takane (2007, Property 11) and Adachi (2020, Theorem A.4.2) \((n = 1)\). This is a problem maximizing \(tr(X^T A)\), where \(X\) and fixed \(A\) are \(p \times q\) \((p \geq q)\) matrices with the constraint \(X^T X = I_q\) and \(rank(A) \leq q\). To the author’s knowledge, this problem was first solved by Green Jr (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 \(A = UAV^T\) with \(U^T U = V^T V = I_q\) and \(A\) 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

\[
max\{tr(X^T A)\} = max\{tr(X^T U A V^T)\} = max\{tr(V^T X^T U A)\} = tr(\Lambda)
\]

since \(V^T X^T U\) is suborthonormal. The maximum is attained when \(V^T X^T U = I_q\) or \(X = U V^T\).

Note that the definition of the SVD using the non-negative rather than positive singular values is important considering the case \(rank(A) \equiv q^* < q\). That is, in the last case when using only positive singular values, \(U\) and \(V\) become \(p \times q^*\) and \(q \times q^*\) semiorthonormal matrices, respectively, yielding \(X^T X \neq I_q\), which does not satisfy the assumption. Note that Green Jr (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 \(U\), \(V\) and \(U V^T\). 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 \(X_i\) are no longer required to be orthonormal. Second, the \(X_i\) 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 \(\text{rank}(X_i) = r^*_i < r_i = \min\{m_{i-1}, m_i\}\) 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., $X_1 (n \times r_1)$ of rank $r^*_1$ and $X_2 (n \times r_2)$ 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 $B_1 (r_1 \times r^*)$ and $B_2 (r_2 \times r^*)$ with other ones, the maximum of $\text{tr}(B_1^T X_1^T X_2 B_2)$ was attained.

References


Proof of Kristof’s Theorem: A Tutorial


Appendix A Technical results

**Lemma A:** Fan’s (1951, Lemma 1A) inequality for the doubly substochastic matrix. Let \( \mathbf{a} = (a_1, \ldots, a_n)^T \) and \( \mathbf{b} = (b_1, \ldots, b_n)^T \) 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 \( \sum_{j=1}^{n} p_{ij} \leq 1 \) \( (i = 1, \ldots, n) \) and \( \sum_{i=1}^{n} p_{ij} \leq 1 \) \( (j = 1, \ldots, n) \) (see e.g., Marshall et al. (2011, Section 2.C)). Then, \( \mathbf{a}^T \mathbf{P} \mathbf{b} \leq \mathbf{a}^T \mathbf{b} \).

**Proof** (a slight extension of Uchida (2023)). Let \( c_1, \ldots, c_n \) and \( d_1, \ldots, d_n \) be non-negative numbers. Then, we can write \( a_i = \sum_{k=1}^{n} c_k \) and \( b_i = \sum_{k=1}^{n} d_k \). Using these expressions,

\[
\mathbf{a}^T \mathbf{b} - \mathbf{a}^T \mathbf{P} \mathbf{b} = \sum_{i,j=1}^{n} (\delta_{ij} - p_{ij}) a_i b_j = \sum_{i,j=1}^{n} (\delta_{ij} - p_{ij}) \sum_{k=1}^{n} c_k \sum_{l=1}^{n} d_l = \sum_{k,j=1}^{n} c_k d_l \sum_{i=1}^{k} \sum_{j=1}^{l} (\delta_{ij} - p_{ij})
\]
follows, where $\delta_{ij}$ is the Kronecker delta. In the above expression, define the doubly stochastic (rather than substochastic) matrix $P^* = \{p^*_{ij}\}$ satisfying $p^*_{ij} \geq p_{ij}$ ($i, j = 1, ..., n$). Then, consider the case $k \geq l$ in the above expression. We obtain

$$\sum_{i=1}^{k} \sum_{j=1}^{l} (\delta_{ij} - p_{ij}) = \sum_{i=1}^{k} \sum_{j=1}^{l} \delta_{ij} - \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij}$$

$$= \sum_{i=1}^{l} \delta_{ii} - \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} = l - \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij}$$

$$\geq l - \sum_{j=1}^{n} \sum_{j=1}^{l} p^*_{ij} \geq 0.$$ 

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 $a^Tb - a^TPb \geq 0$. □

**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 $P^*$, which is a special case of the doubly substochastic matrix $P$.

The second proof of Theorem 1 (von Neumann’s trace inequality; Mirsky (1975, Section 3, p. 305)). \(|\text{tr}(X_1\Gamma_1X_2\Gamma_2)| \leq \text{tr}(\Gamma_1\Gamma_2)|\) is derived. Let $X_i = \{x_{i(jk)}\}$ ($j, k = 1, ..., m$) and $\Gamma_i = \text{diag}(\gamma(i1), ..., \gamma(im))$ ($i = 1, 2$). Then, using Fan’s (1951) inequality for doubly (sub)stochastic matrices, we have

$$|\text{tr}(X_1\Gamma_1X_2\Gamma_2)| = |\sum_{j,k=1}^{m} x_{i(jk)}\gamma(i1)k_{j}\gamma(i2)j|$$

$$\leq \sum_{j,j=1}^{m} |x_{i(jk)}k_{j}\gamma(i1)k_{j}\gamma(i2)j|$$

$$\leq \frac{1}{2} \sum_{j,j=1}^{m} |x_{i(jk)}|\gamma(i1)k_{j}\gamma(i2)j + \frac{1}{2} \sum_{j,j=1}^{m} |x_{i(jk)}|\gamma(i1)k_{j}\gamma(i2)j$$

$$\leq \frac{1}{2} \sum_{j=1}^{m} \gamma(i1)\gamma(i2)j + \frac{1}{2} \sum_{j=1}^{m} \gamma(i1)\gamma(i2)j = \text{tr}(\Gamma_1\Gamma_2).$$

□

The 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 and Kurata (2013)). He did not mention or use Ten Berge’s theorem, but employed calculus and Lagrange multipliers. Neudecker also used the SVD $A = U_0A_0V^T_0$, where he employed only positive singular values i.e., $A_0 > O$ in Lönner’s sense.

An advantage of Neudecker’s derivation is to give the set of explicit expressions of $X$ maximizing $\text{tr}(X^TA)$ as $X = U_0V^T_0 = UV^T$ when $\text{rank}(A) = q^* = q$ and $X = U_0V^T_0 + Q(I_q - V_0V^T_0)$ with a $p \times q$ matrix $Q$ when $q^* < q$ as “the
general solution” (Neudecker, 2004, Equation (2.5)). In these expressions,

\[ U_0 V_0^T = X ((X^T X)^{1/2})^+ = X ((X^T X)^+)^{1/2} \equiv X (X^T X)^+ \]

and

\[ V_0 V_0^T = (X^T X)^+(X^T X)^{1/2} = (X^T X)^{1/2} (X^T X)^+ \]

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 \(Y = PQ^T\) employing only the positive singular values, we have \(Y^+ = Q \Gamma^{-1} P^T\), which satisfies the conditions of the MP \(g\)-inverse: \(YY^+ = Y, Y^+ YY^+ = Y^+, YY^+ = (YY^+)^T\) and \(Y^+ Y = (Y^+ Y)^T\).

Neudecker (2004, Equation (2.5)) stated that “\(Q\) arbitrary”. This is misleading since when \(Q\) is a zero matrix, the rank of \(X = U_0 V_0^T\) becomes \(q^* < q\) and does not satisfy \(X^T X = I_q\) though this will give the same maximum. Instead, \(Q\) should be defined as a \(p \times q\) arbitrary suborthonormal matrix \(Q = U_1 V_1^T\) of rank \(q - q^*\), where \(U_1\) and \(V_1\) are \(p \times (q - q^*)\) and \(q \times (q - q^*)\) semiothornormal matrices, respectively such that \(U_1^T U_1 = V_1^T V_1 = I_{q - q^*}\), \(U_0^T U_1 = O\) and \(V_0^T V_1 = O\), which shows the arbitrary property of \(Q\) stated earlier in that when \(U_1\) and \(V_1\) are replaced by \(U_1 U_1^T\) and \(V_1 V_1^T\) with \(U_1^T\) and \(V_1^T\) being arbitrary \((q - q^*) \times (q - q^*)\) orthonormal matrices, these can be used with a different \(Q^* \equiv U_1^T [V_1^T V_1^T]^T \neq Q\). Note that these arbitrary \(U_1\) and \(V_1\) give \(V_0^T V_0 + V_1^T V_1 = (V_0 : V_1)^T (V_0 : V_1) = (V_0 : V_1 V_1^T)^T (V_0 : V_1^T) = I_q\).

Then, it is found that

\[ X = U_0 V_0^T + Q(I_q - V_0 V_0^T) = U_0 V_0^T + U_1 V_1^T (I_q - V_0 V_0^T) \]

\[ = U_0 V_0^T + U_1 V_1^T = (U_0 : U_1)(V_0 : V_1)^T \]

satisfying

\[ X^T X = (V_0 : V_1)(U_0 : U_1)^T (U_0 : U_1)(V_0 : V_1)^T \]

\[ = (V_0 : V_1)(V_0 : V_1)^T = I_q. \]

Using the above \(X\), the maximum is given by

\[ \text{tr}(A^T X) = \text{tr}\{(U_0 A_0 V_0^T)^T (U_0 : U_1)(V_0 : V_1)^T\} \]

\[ = \text{tr}\{V_0 A_0 U_0^T (U_0 V_0^T + U_1 V_1^T)\}\]

\[ = \text{tr}(V_0 A_0 V_0^T) = \text{tr}(V_0^T V_0 A_0) \]

\[ = \text{tr}(A_0), \]

where \(\text{tr}(A_0) = \text{tr}(A)\) and the singular diagonal matrix \(A\) of rank \(q^*\) was used earlier. Define the semiothornormal matrix \(U \equiv (U_0 : U_1)\) and orthonormal \(V \equiv (V_0 : V_1)\). Then, \(X = U V^T\) is equal to that obtained by Ten Berge as shown earlier.
Recall the expression $X = U_0 V_0^T + Q(I_q - V_0 V_0^T)$. The matrix $Q$ can be an arbitrary $p \times q$ semiorthonormal one denoted by $U^*$ with $U_0^T U^* = 0$ and $U^* U^* = I_q$. This is seen by

$$X^T X = (U_0 V_0^T + U^*(I_q - V_0 V_0^T))^T (U_0 V_0^T + U^*(I_q - V_0 V_0^T))$$

$$= V_0 U_0^T U_0 V_0^T + (I_q - V_0 V_0^T) U^* U^*(I_q - V_0 V_0^T)$$

$$= V_0 V_0^T + I_q - V_0 V_0^T = I_q,$$

satisfying the assumption, where the idempotent property $(I_q - V_0 V_0^T)^2 = I_q - V_0 V_0^T$ is used. The matrix $X$ gives the same maximum

$$\text{tr}(A^T X) = \text{tr} [V_0 A_0 U_0^T (U_0 V_0^T + U^*(I_q - V_0 V_0^T))]$$

$$= \text{tr}(V_0 A_0 U_0^T U_0 V_0^T) = \text{tr}(V_0 A_0 V_0^T)$$

$$= \text{tr}(A_0).$$

In the expression $X = U_0 V_0^T + Q(I_q - V_0 V_0^T)$, the term $V_0 V_0^T$ is the projection matrix onto the space spanned by the $q^*$ orthonormal columns of $V_0$ given by $V_0 (V_0^T V_0) = V_0 V_0^T$. Recall that $I_q = V_0 V_0^T + V_1 V_1^T$. Then, $I_q - V_0 V_0^T = V_1 V_1^T$ is also the projection matrix onto the space spanned by the $q - q^*$ orthonormal columns of $V_1$. In the term $Q(I_q - V_0 V_0^T)$ is projected onto the space spanned by $V_1$.

An arbitrary property of $U_1$ and $V_1$ in $X = UV^T = U_0 V_0^T + U_1 V_1^T$ of Ten Berge's (1983) solution similar to that with $Q$ is that $U_1$ and $V_1$ can be replaced by $U_1 W_{U1}$ and $V_1 W_{V1}$, respectively, where $W_{U1}$ and $W_{V1}$ are arbitrary $(q - q^*) \times (q - q^*)$ orthonormal matrices yielding $X^* \equiv U_0 V_0^T + U_1 W_{U1} V_1^T \neq X = U_0 V_0^T + U_1 V_1^T$.

For completeness, arbitrary aspects in $A = U_0 A_0 V_0^T$ of rank $q^* \leq q$ are noted. Under the standard definition of $A_0 = \text{diag}(\lambda_1, \ldots, \lambda_{q^*})$ with $\lambda_1 \geq \cdots \geq \lambda_{q^*} \geq 0$, $A_0$ is identified while $U_0$ and $V_0$ 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., $\lambda_j = \lambda_{j+1} = \cdots = \lambda_{j+k-1} \equiv \lambda_{jk}$ with multiplicity $k(>1)$, we have

$$A = U_{0(-k)} A_{0(-k)} V_{0(-k)}^T + U_{0(k)} A_{0(k)} V_{0(k)}^T$$

$$= U_{0(-k)} A_{0(-k)} V_{0(-k)}^T + U_{0(k)} \lambda_j I_k V_{0(k)}^T$$

$$= U_{0(-k)} A_{0(-k)} V_{0(-k)}^T + \lambda_j U_{0(k)} W_{(k)} (V_{0(k)} W_{(k)})^T,$$

where $U_{0(k)}$ and $V_{0(k)}$ are semiorthonormal submatrices of $U_0$ and $V_0$, respectively corresponding to the multiple $\lambda_{jk}$ with $A_{0(k)}$ defined similarly; and $U_{0(-k)}$, $V_{0(-k)}$ and $A_{0(-k)}$ are matrices given by using the singular values except $\lambda_j = \lambda_{j+1} = \cdots = \lambda_{j+k-1}$; and $W_{(k)}$ 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.