Review Article   **Open Access**

Yoshio Takane*

Professor Haruo Yanai and multivariate analysis

Abstract: The late Professor Yanai has contributed to many fields ranging from aptitude diagnostics, epidemiology, and nursing to psychometrics and statistics. This paper reviews some of his accomplishments in multivariate analysis through his collaborative work with the present author, along with some untold episodes for the inception of key ideas underlying the work. The various topics covered include constrained principal component analysis, extensions of Khatri's lemma, the Wedderburn-Guttman theorem, ridge operators, generalized constrained canonical correlation analysis, and causal inference. A common thread running through all of them is projectors and singular value decomposition, which are the main subject matters of a recent monograph by Yanai, Takeuchi, and Takane [60].

Keywords: Projectors, Singular value decomposition, Khatri's lemma, The Wedderburn-Guttman theorem, Ridge operators

1 Prolog

I met Professor Yanai for the first time almost fifty years ago. I was an undergraduate student in Psychology at the University of Tokyo, where every May they hold a big open house called May Festival for general public. We, students in Psychology Department, planned to host an event, and did some experiments on attitude change. After we collected our data, someone suggested that we might apply factor analysis, a statistical technique similar to principal component analysis (PCA), to our data. But no one really knew what to do. So I got in touch with a professor in Educational Psychology to seek some advice, and I was introduced to Professor Yanai, who was then a graduate student developing a diagnostic system that guided us to choose the best profession based on our ability, interests, personality, and so on using some "esoteric" multivariate analysis (MVA) techniques (e.g., canonical discriminant analysis). This was my first encounter to Quantitative Psychology (QP) and MVA. I was more interested in Social Psychology at that time, but through many subsequent interactions with Professor Yanai, I gradually became interested in QP until in the end I chose to become a quantitative psychologist. Professor Yanai was indeed instrumental to my entire career.

Professor Yanai passed away due to prostate cancer in December, 2013 at the age of 73. A quick glance at his homepage reveals that his contributions extend over seven broad categories, including aptitude diagnostics, test theories, educational psychology, epidemiology, nursing, linear algebra, statistics, and MVA. I am not capable of covering all these areas in this paper. Rather, I would like to focus on his contributions in the last category, namely MVA, primarily through his collaborative works with me. We have 15 joint publications including two books, a majority of which are on MVA. The specific topics we cover are:

(i) Constrained principal component analysis (CPCA)
(ii) Khatri's lemma

---

*Corresponding Author: Yoshio Takane:** Department of Psychology, University of Victoria, Victoria, BC, Canada, E-mail: yoshio.takane@mcgill.ca

(iii) The Wedderburn-Guttman (WG) theorem
(iv) Ridge operators
(v) Generalized constrained canonical correlation analysis (GCCANO)
(vi) Causal inference

Professor Yanai thought of MVA as partitioning a space into meaningful subspaces according to some external and internal criteria (Takeuchi, Yanai, and Mukherjee [50]). Two major tools for partitioning are:

(i) Projections
(ii) Singular value decomposition (SVD)

which are the main subject matters of a recent monograph by Yanai, Takeuchi, and Takane [60]. As is well known, projectors are used to partition the space of observation vectors into subspaces that can and cannot be explained by external information, while SVD finds the subspace of minimal dimensionality that captures the largest variability in the original space.

Before we start, let us introduce some basic notations we use throughout this paper. Let Sp($\mathbf{G}$) represent the space spanned by column vectors of $\mathbf{G}$, and let Ker($\mathbf{G}'$) represent the orthogonal complement subspace to Sp($\mathbf{G}$). Let

$$\mathbf{P}_G = \mathbf{G}(\mathbf{G}'\mathbf{G})^-\mathbf{G}' \tag{1}$$

denote the orthogonal projector onto Sp($\mathbf{G}$), and let

$$\mathbf{Q}_G = \mathbf{I} - \mathbf{P}_G \tag{2}$$

denote the orthogonal projector onto Ker($\mathbf{G}'$). As is well known, these projectors have the following properties:

$$\mathbf{P}'_G = \mathbf{P}_G, \mathbf{Q}'_G = \mathbf{Q}_G \text{ (symmetric).}$$
$$\mathbf{P}^2_G = \mathbf{P}_G, \mathbf{Q}^2_G = \mathbf{Q}_G \text{ (idempotent).}$$
$$\mathbf{P}_G\mathbf{Q}_G = \mathbf{Q}_G\mathbf{P}_G = \mathbf{O} \text{ (orthogonal).}$$

These projectors are called (I-orthogonal) projectors and are useful in partitioning $\mathbf{y}$, the vector of observations on the dependent variable in regression analysis, into $\mathbf{P}_G\mathbf{y}$, the portion of $\mathbf{y}$ that can be accounted for by the predictor variables $\mathbf{G}$, and $\mathbf{Q}_G\mathbf{y}$, the portion of $\mathbf{y}$ that is left unaccounted for by $\mathbf{G}$.

Slight generalizations of the I-orthogonal projectors above lead to K-orthogonal projectors, which are useful in weighted least squares (LS) estimation in regression analysis. Let $\mathbf{K}$ be a nonnegative definite (*nnd*) matrix such that rank($\mathbf{KG}$) = rank($\mathbf{G}$). Then,

$$\mathbf{P}_{G/K} = \mathbf{G}(\mathbf{G}'\mathbf{KG})^-\mathbf{G}'\mathbf{K}, \tag{3}$$

and

$$\mathbf{Q}_{G/K} = \mathbf{I} - \mathbf{P}_{G/K} \tag{4}$$

are called K-orthogonal projectors onto Sp($\mathbf{G}$) and Ker($\mathbf{G}'$), respectively, with respect to the metric matrix $\mathbf{K}$. These projectors have properties similar to those of the I-orthogonal projectors, namely

$$(\mathbf{KP}_{G/K})' = \mathbf{KP}_{G/K}, (\mathbf{KQ}_{G/K})' = \mathbf{KQ}_{G/K} \text{ (K-symmetric).}$$
$$\mathbf{P}^2_{G/K} = \mathbf{P}_{G/K}, \mathbf{Q}^2_{G/K} = \mathbf{Q}_{G/K} \text{ (idempotent).}$$
$$\mathbf{P}'_{G/K}\mathbf{KQ}_{G/K} = \mathbf{Q}'_{G/K}\mathbf{KP}_{G/K} = \mathbf{O} \text{ (K-orthogonal).}$$

When $\mathbf{K}$ is set equal to $\mathbf{P}_Z$, where $\mathbf{Z}$ represents the matrix of instrumental variables (IV), the K-orthogonal projectors

$$\mathbf{P}_{G/P_Z} = \mathbf{G}(\mathbf{G}'\mathbf{P}_Z\mathbf{G})^-\mathbf{G}'\mathbf{P}_Z$$

become those that follow from IV estimation. This type of projectors will be useful in the section entitled "Epilog." See Yanai [58] for other types of projectors.

2 Constrained Principal Component Analysis

In 1970, Professor Yanai [57] proposed a PCA (called partial PCA; PPCA) that eliminated certain prescribed effects from extracted components. The effects to be eliminated may be gender, age, levels of education, and so on. PPCA amounts to SVD of $\mathbf{Q}_G\mathbf{Y}$, where $\mathbf{Y}$ indicates the main data matrix (to which PCA is to be applied), and $\mathbf{G}$ the matrix of predictor variables whose effects are to be eliminated. If we look at this process more closely, we notice that PPCA consists of two distinct phases: (1) Decomposing $\mathbf{Y}$ into $\mathbf{P}_G\mathbf{Y}$ and $\mathbf{Q}_G\mathbf{Y}$ (a decomposition of $\mathbf{Y}$ according to the external information), and (2) applying SVD to the latter (a decomposition of $\mathbf{Q}_G\mathbf{Y}$ according to the internal criterion). While Professor Yanai did not explicitly suggest SVD of $\mathbf{P}_G\mathbf{Y}$, there was no reason why not. Indeed, there exists a MVA technique (called redundancy analysis; RA), which applies SVD to $\mathbf{P}_G\mathbf{Y}$. Constrained PCA (CPCA) has been developed with these two phases of PPCA (and RA) in mind. CPCA represents a direct application of the two most important ingredients of MVA, projections and SVD, according to Professor Yanai.

CPCA ([31–33, 39]) closely follows the above idea. It consists of two major phases, one called External Analysis and the other called Internal Analysis. External Analysis decomposes a main data matrix according to the external information about the rows and columns of the data matrix, which amounts to projections of the data matrix onto the space spanned by the external information. Internal Analysis further decomposes the matrices decomposed in the External Analysis into several components according to their importance by SVD.

In CPCA, we consider not only the row-side constraints $\mathbf{G}$, but also the column-side constraints $\mathbf{H}$, analogously to growth curve models (Potthoff and Roy [20]). This leads to a four-way decomposition of the main data matrix $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{P}_G\mathbf{Y}\mathbf{P}_H + \mathbf{Q}_G\mathbf{Y}\mathbf{P}_H + \mathbf{P}_G\mathbf{Y}\mathbf{Q}_H + \mathbf{Q}_G\mathbf{Y}\mathbf{Q}_H. \tag{5}$$

A similar decomposition is also possible with K-orthogonal projectors.

The decomposition above is a very basic one. When $\mathbf{G}$ and/or $\mathbf{H}$ consist of more than one set of variables, finer decompositions of $\mathbf{Y}$ are possible, corresponding to analogous decompositions of $\mathbf{P}_G$ and/or $\mathbf{P}_H$ (e.g., Takane and Yanai [40]). Let $\mathbf{G} = [\mathbf{M}, \mathbf{N}]$, for example. Then,

(i) $\mathbf{P}_G = \mathbf{P}_M + \mathbf{P}_N \Leftrightarrow \mathbf{M}'\mathbf{N} = \mathbf{O}$. ($\mathbf{M}$ and $\mathbf{N}$ are mutually orthogonal.)

(ii) $\mathbf{P}_G = \mathbf{P}_M + \mathbf{P}_N - \mathbf{P}_M\mathbf{P}_N \Leftrightarrow \mathbf{P}_M\mathbf{P}_N = \mathbf{P}_N\mathbf{P}_M$. ($\mathbf{P}_M$ and P_N commute, or equivalently $\mathbf{M}$ and $\mathbf{N}$ are mutually orthogonal except their common space if it exists; this decomposition is useful in ANOVA without interactions).

(iii) $\mathbf{P}_G = \mathbf{P}_M + \mathbf{P}_{Q_MN} = \mathbf{P}_N + \mathbf{P}_{Q_NM}$. (This decomposition always holds; it arises when we fit one of $\mathbf{M}$ and $\mathbf{N}$ first and the other to the residuals from the first).

(iv) $\mathbf{P}_G = \mathbf{P}_{M/Q_N} + \mathbf{P}_{N/Q_M} \Leftrightarrow \text{rank}(\mathbf{G}) = \text{rank}(\mathbf{M}) + \text{rank}(\mathbf{N})$. (This decomposition holds when $\mathbf{M}$ and $\mathbf{N}$ are disjoint; it arises when we fit both $\mathbf{M}$ and $\mathbf{N}$ simultaneously).

(v) $\mathbf{P}_G = \mathbf{P}_{GA} + \mathbf{P}_{G(G'G)^-B} \Leftrightarrow \mathbf{A}'\mathbf{B} = \mathbf{O}$, and $\text{Sp}(\mathbf{A}) \oplus \text{Sp}(\mathbf{B}) = \text{Sp}(\mathbf{G}')$. (A matrix of regression coefficients $\mathbf{C}$ is constrained either by $\mathbf{C} = \mathbf{A}\mathbf{C}^\star$ for some $\mathbf{C}^\star$ or by $\mathbf{B}'\mathbf{C} = \mathbf{O}$), where $\mathbf{A}$ and $\mathbf{B}$ are know matrices.

The first four decompositions were noted in Rao and Yanai [25], while (v) is due to Yanai and Takane [59]. Analogous decompositions are possible for $\mathbf{P}_H$, $\mathbf{P}_{G/K}$, and $\mathbf{P}_{H/L}$.

Note 1. The two terms in Decomposition (iv) above are not mutually orthogonal. Takane and Yanai ([40]) suggested a metric under which they were rendered orthogonal. This metric stipulates

$$\mathbf{K}_1 = \mathbf{Q}_M + \mathbf{Q}_N + \mathbf{T}\mathbf{D}\mathbf{T}', \tag{6}$$

where $\mathbf{T}$ is such that $\text{Sp}(\mathbf{T}) = \text{Ker}(\mathbf{G}')$, and $\mathbf{D}$ is an arbitrary positive definite (pd) matrix. It can be easily verified that $\mathbf{P}_G = \mathbf{P}_{G/K_1}$, $\mathbf{P}_{M/Q_N} = \mathbf{P}_{M/K_1}$, $\mathbf{P}_{N/Q_N} = \mathbf{P}_{N/K_1}$, and $\mathbf{P}'_{M/K_1}\mathbf{K}_1\mathbf{P}_{N/K_1} = \mathbf{O}$. Note that such an orthogonalizing metric is not unique (apart from the arbitrariness in $\mathbf{T}$ and $\mathbf{D}$). For example, the following $\mathbf{K}_2$ also has the effect of orthogonalizing the two terms in (iv):

$$\mathbf{K}_2 = (\mathbf{G}\mathbf{G}')^+ + \mathbf{T}\mathbf{D}\mathbf{T}', \tag{7}$$

where $^+$ indicates the Moore-Penrose inverse. This metric matrix has similar properties as $\mathbf{K}_1$. In fact, it not only orthogonalizes the two terms in (iv) ($\mathbf{M}'\mathbf{K}_2\mathbf{N} = \mathbf{O}$), but also normalizes them ($\mathbf{M}'\mathbf{K}_2\mathbf{M} = \mathbf{P}_{M'}$ and $\mathbf{N}'\mathbf{K}_2\mathbf{N} = \mathbf{P}_{N'}$, where $\mathbf{P}_{M'}$ and $\mathbf{P}_{N'}$ reduce to identity matrices if $\mathbf{M}$ and $\mathbf{N}$ are columnwise nonsingular. The metric $\mathbf{K}_2^{\star} = (\mathbf{G}\mathbf{G}')^+ = \mathbf{G}(\mathbf{G}'\mathbf{G})^{+2}\mathbf{G}'$ (obtained from $\mathbf{K}_2$ by setting $\mathbf{D} = \mathbf{O}$) is particularly interesting because it maps $\mathbf{G}$ into $\mathbf{G}(\mathbf{G}'\mathbf{G})^+$ (i.e., $(\mathbf{G}')^+$), which implies that it is a matrix analogue of reproducing kernel in the reproducing kernel Hilbert space (RKHS); see e.g., Ramsay and Silverman [22]. (A finite dimensional Euclidean space is always an RKHS.) Both $\mathbf{K}_1$ and $\mathbf{K}_2$ are special instances of the orthogonalizing metric discussed by Rao and Mitra ([24], Lemma 5.3.1).■ (which marks the end of the note.)

In Internal Analysis, on the other hand, we apply PCA to terms obtained by the external analysis, e.g., $\mathbf{P}_G\mathbf{Y}\mathbf{P}_H$, which amounts to SVD($\mathbf{P}_G\mathbf{Y}\mathbf{P}_H$), whose computation time can be shortened substantially by the following procedure:

A theorem on SVD($\mathbf{P}_G\mathbf{Y}\mathbf{P}_H$) (Takane and Hunter [32]). Let $\mathbf{F}_G$ and $\mathbf{F}_H$ be columnwise orthogonal matrices such that Sp($\mathbf{G}$) = Sp($\mathbf{F}_G$) and Sp($\mathbf{H}$) = Sp($\mathbf{F}_H$). Then, $\mathbf{P}_G\mathbf{Y}\mathbf{P}_H = \mathbf{F}_G(\mathbf{F}_G'\mathbf{Y}\mathbf{F}_H)\mathbf{F}_H'$. Let SVD($\mathbf{F}_G'\mathbf{Y}\mathbf{F}_H$) be denoted as $\mathbf{U}\mathbf{D}\mathbf{V}'$, and let SVD($\mathbf{F}_G\mathbf{F}_G'\mathbf{Y}\mathbf{F}_H\mathbf{F}_H'$) be denoted as $\mathbf{U}^{\star}\mathbf{D}^{\star}\mathbf{V}^{\star'}$. Then, $\mathbf{U}^{\star} = \mathbf{F}_G\mathbf{U}$, $\mathbf{V}^{\star} = \mathbf{F}_H\mathbf{V}$, and $\mathbf{D}^{\star} = \mathbf{D}$. Takane and Hunter ([32]) also extends this theorem to the case of non-identity metric matrices.

Takane ([31]) provides a more comprehensive account of CPCA including many applications and extensions.

3 Khatri's Lemma

Toward the end of 1980's, I was interested in the relationships among various methods of constrained correspondence analysis (CCA). Correspondence analysis (CA) is a PCA-like technique for the analysis of two-way contingency tables, allowing spatial representations of rows and columns of contingency tables. When I looked through the literature on CCA, I found that there were two ways of incorporating the constraints. Let $\mathbf{U}$ denote the row representation matrix. (For explanation, we consider only the row-side constraints.) Two equivalent ways of constraining $\mathbf{U}$ are: (i) $\mathbf{U} = \mathbf{A}\mathbf{U}^{\star}$ (e.g., ter Braak [51]), and (ii) $\mathbf{B}'\mathbf{U} = \mathbf{O}$ (e.g., Böckenholt and Böckenholt [1]), where $\mathbf{A}$ and $\mathbf{B}$ are known matrices. ($\mathbf{A}$ reparameterizes $\mathbf{U}$ by $\mathbf{U}^{\star}$, while $\mathbf{B}$ specifies the null space of $\mathbf{U}$.) When they are such that Sp($\mathbf{A}$) = Ker($\mathbf{B}'$), (i) and (ii) are equivalent. This is a rather trivial relationship, i.e.,

$$\mathbf{P}_A = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}' = \mathbf{I} - \mathbf{B}(\mathbf{B}'\mathbf{B})^-\mathbf{B}' = \mathbf{Q}_B \tag{8}$$

under the identity metric. However, twenty five years ago I was not sure what would happen if non-identity metric $\mathbf{K}$ was used, which was usually the case in CA. Khatri's lemma gives the exact relationship we needed (Takane, Yanai, and Mayekawa [49]):

Let $\mathbf{A}$ ($p \times r$) and $\mathbf{B}$ ($p \times (p - r)$) be matrices such that rank($\mathbf{A}$) = r, rank($\mathbf{B}$) = $p - r$, and $\mathbf{A}'\mathbf{B} = \mathbf{O}$. Then (Khatri [13]),

$$\mathbf{I} = \mathbf{A}(\mathbf{A}'\mathbf{K}\mathbf{A})^{-1}\mathbf{A}\mathbf{K} + \mathbf{K}^{-1}\mathbf{B}(\mathbf{B}'\mathbf{K}^{-1}\mathbf{B})^{-1}\mathbf{B}', \tag{9}$$

where $\mathbf{K}$ is a symmetric pd matrix. This lemma was used by Khatri [13] for rewriting a Q-type projector (written as $\mathbf{I}$ minus a P-type projector) into a single P-type projector in growth curve models.

Note 2. According to Schaffrin [26], "Khatri's" lemma was established well before 1966. In as early as 1907, Helmert [8] showed what we call Khatri's lemma in this paper. Let $\mathbf{y} = \mathbf{B}\mathbf{a} + \mathbf{e}$ be the regression equation, where $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$. The best linear unbiased predictor (BLUP) of $\mathbf{e}$ in this set-up is given by

$$\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{B}(\mathbf{B}'\mathbf{K}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{K}^{-1})\mathbf{y}. \tag{10}$$

Next, let $\mathbf{A}$ be such that $\text{Ker}(\mathbf{A}') = \text{Sp}(\mathbf{B})$. By premultiplying both sides of the above regression equation by $\mathbf{A}'$, we obtain

$$\mathbf{A}'\mathbf{y} = \mathbf{A}'\mathbf{e}, \tag{11}$$

eliminating the term related to $\mathbf{a}$. The BLUP of $\mathbf{e}$ for (11) is obtained by minimizing $\mathbf{e}'\mathbf{K}^{-1}\mathbf{e}$ subject to the constraint (11), and is given by

$$\hat{\mathbf{e}} = \mathbf{K}\mathbf{A}(\mathbf{A}'\mathbf{K}\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}. \tag{12}$$

The two $\hat{\mathbf{e}}$'s given above must be identical for any $\mathbf{y}$, leading to an equivalent formula to Khatri's lemma. ∎

Several remarks are in order on Khatri's original lemma given above. Khatri's lemma may sometimes be expressed in an alternative form:

$$\mathbf{K} = \mathbf{K}\mathbf{A}(\mathbf{A}'\mathbf{K}\mathbf{A})^{-1}\mathbf{A}\mathbf{K} + \mathbf{B}(\mathbf{B}'\mathbf{K}^{-1}\mathbf{B})^{-1}\mathbf{B}'. \tag{13}$$

Note also that $\mathbf{K}$ and $\mathbf{K}^{-1}$ are interchangeable. Khatri's lemma is extremely useful for rewriting P-type projectors into Q-type projectors, which occurs quite frequently (e.g., LaMotte [15]; Seber [27]; Shapiro [28]; Takane and Zhou [43]; Verbyla [55]). More recently, Khatri's lemma was used to derive additive decompositions of Pearson's chi-square ([19]) and Light and Margolin's ([16] C statistics for contingency tables ([17]), [38], [44]).

Khatri's lemma has been generalized in various ways, e.g., let $\mathbf{K}$ be square, but not necessarily symmetric or nonsingular, but such that $\text{Sp}(\mathbf{B}) \subset \text{Sp}(\mathbf{K})$ and $\text{Sp}(\mathbf{B}) \subset \text{Sp}(\mathbf{K}')$. Then (Khatri [14]),

$$\mathbf{K} = \mathbf{K}\mathbf{A}(\mathbf{A}'\mathbf{K}\mathbf{A})^{-}\mathbf{A}'\mathbf{K} + \mathbf{B}(\mathbf{B}'\mathbf{K}^{-}\mathbf{B})^{-}\mathbf{B}'. \tag{14}$$

Professor Yanai (Yanai and Takane [59]) further extended Khatri's lemma as follows. Let $\mathbf{A}$ ($p \times r$) and $\mathbf{B}$ ($p \times (p - r)$) be matrices such that $\text{rank}(\mathbf{A}) = r$ and $\text{rank}(\mathbf{B}) = p - r$, and let $\mathbf{M}$ and $\mathbf{N}$ be *nnd* matrices such that

        (i) $\mathbf{A}'\mathbf{M}\mathbf{N}\mathbf{B} = \mathbf{O}$,
        (ii) $\text{rank}(\mathbf{M}\mathbf{A}) = \text{rank}(\mathbf{A})$,
        (iii) $\text{rank}(\mathbf{N}\mathbf{B}) = \text{rank}(\mathbf{B})$.

Then,

$$\mathbf{I} = \mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^{-}\mathbf{A}'\mathbf{M} + \mathbf{N}\mathbf{B}(\mathbf{B}'\mathbf{N}\mathbf{B})^{-}\mathbf{B}'. \tag{15}$$

This reduces to the original lemma when $\mathbf{M} = \mathbf{K}$ and $\mathbf{N} = \mathbf{K}^{-1}$. Takane [31] further extends it to a rectangular $\mathbf{K}$. (Also, see the next section.)

4 The Wedderburn-Guttman Theorem

The Wedderburn-Guttman (WG) theorem is stated as follows. Let $\mathbf{Y}$ ($n \times p$) be of rank r, and let $\mathbf{A}$ ($n \times s$) and $\mathbf{B}$ ($p \times s$) be such that $\mathbf{A}'\mathbf{Y}\mathbf{B}$ is invertible. Then,

$$\text{rank}(\mathbf{Y}_1) = \text{rank}(\mathbf{Y}) - \text{rank}(\mathbf{Y}\mathbf{B}(\mathbf{A}'\mathbf{Y}\mathbf{B})^{-1}\mathbf{A}'\mathbf{Y}) \tag{16}$$

$$= \text{rank}(\mathbf{Y}) - \text{rank}(\mathbf{A}'\mathbf{Y}\mathbf{B}) = r - s, \tag{17}$$

where

$$\mathbf{Y}_1 = \mathbf{Y} - \mathbf{Y}\mathbf{B}(\mathbf{A}'\mathbf{Y}\mathbf{B})^{-1}\mathbf{A}'\mathbf{Y}. \tag{18}$$

Guttman [6] called the above theorem Lagrange's theorem, referring to Wedderburn [56], but there was no reference to Lagrange in [56]. We call it the WG theorem following Hubert, Meulman and Heiser [10]. Wedderburn [56] first proved the theorem for $s = 1$. Guttman [6] extended it to $s > 1$. Guttman [7] further proved the reverse, i.e., $\mathbf{Y}_1$ must be of the form (18) to satisfy the rank condition stated in (16) and (17).

It is interesting to note that Guttman [6] used the "matrix rank method" for a proof of the above theorem. In this method, we apply a series of elementary block matrix operations to a matrix to derive a rank formula. We apply another series of elementary block matrix operations to the same matrix to derive another rank

formula. Neither operations change the rank of the original matrix, so the two must be equal. Tian (e.g., [52], [53]) derived many interesting rank formula based on this method. It is intriguing to find that Guttman [6] already used the method in 1944 (cf. Khatri [12]). For your interest, Guttman's proof is given below.

Let

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_s & (\mathbf{A}'\mathbf{YB})^{-1}\mathbf{A}'\mathbf{Y} \\ \mathbf{YB} & \mathbf{Y} \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{YB} & \mathbf{I} \end{bmatrix},$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{I} & -(\mathbf{A}'\mathbf{YB})^{-1}\mathbf{A}'\mathbf{Y} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}.$$

Then,

$$\mathbf{ECF} = \begin{bmatrix} \mathbf{I}_s & \mathbf{O} \\ \mathbf{O} & \mathbf{Y}_1 \end{bmatrix},$$

so that

$$\mathrm{rank}(\mathbf{C}) = s + \mathrm{rank}(\mathbf{Y}_1). \tag{19}$$

On the other hand, let

$$\mathbf{G} = \begin{bmatrix} \mathbf{I} & -(\mathbf{A}'\mathbf{YB})^{-1}\mathbf{A}' \\ \mathbf{O} & \mathbf{I} \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{B} & \mathbf{I} \end{bmatrix}.$$

Then,

$$\mathbf{GCH} = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{Y} \end{bmatrix},$$

so that

$$\mathrm{rank}(\mathbf{C}) = \mathrm{rank}(\mathbf{Y}). \tag{20}$$

We obtain the WG theorem by combining (19) and (20).

My initial interest in this theorem stemmed from Hubert's talk (Hubert, Meulman, and Heiser [10]) at the 1998 Meeting of the Psychometric Society at the University of Illinois at Urbana-Champaign. This talk criticised computational linear algebraists (e.g., Chu, Funderlic, and Golub [2]) for failing to acknowledge Guttman's contributions (Guttman [6, 7]) to the WG theorem. When the talk was over, I asked Hubert a question: When $\mathbf{A}'\mathbf{YB}$ is not invertible, can we replace its regular inverse by a generalized inverse? I said yes, while Hubert said no. It has turned out that both of us are only half correct. The answer is yes, but it requires a condition.

I initially thought this was purely a rank additivity (subtractivity) problem. That is, we are to find a condition under which

$$\mathrm{rank}(\mathbf{Y} - \mathbf{YB}(\mathbf{A}'\mathbf{YB})^{-}\mathbf{A}'\mathbf{Y}) = \mathrm{rank}(\mathbf{Y}) - \mathrm{rank}(\mathbf{YB}(\mathbf{A}'\mathbf{YB})^{-}\mathbf{A}'\mathbf{Y}) \tag{21}$$

holds. My supposition also included that

$$\mathrm{rank}(\mathbf{YB}(\mathbf{A}'\mathbf{YB})^{-}\mathbf{A}'\mathbf{Y}) = \mathrm{rank}(\mathbf{A}'\mathbf{YB}) \tag{22}$$

always holds. It turned out that this was false since Tian and Styan [54] showed that the following held unconditionally:

$$\mathrm{rank}(\mathbf{Y} - \mathbf{YB}(\mathbf{A}'\mathbf{YB})^{-}\mathbf{A}'\mathbf{Y}) = \mathrm{rank}(\mathbf{Y}) - \mathrm{rank}(\mathbf{A}'\mathbf{YB}). \tag{23}$$

This implies that (22) requires a condition, so does (21), and that the two conditions must be equal.

Note 3. The result by Tian and Styan [54] can be more directly shown by a minor extension of the matrix rank method used by Guttman [6]. When $\mathbf{A}'\mathbf{YB}$ is not invertible, we may use its generalized inverse in $\mathbf{C}$, $\mathbf{F}$, and $\mathbf{G}$ in the proof of the original WG theorem. We then have

$$\mathbf{GCH} = \begin{bmatrix} \mathbf{I}_s - (\mathbf{A}'\mathbf{YB})^{-}\mathbf{A}'\mathbf{YB} & \mathbf{O} \\ \mathbf{O} & \mathbf{Y} \end{bmatrix}, \tag{24}$$

while **ECF** remains unchanged. Eq. (24) implies that

$$\text{rank}(\mathbf{C}) = \text{rank}(\mathbf{Y}) + (s - \text{rank}(\mathbf{A}'\mathbf{YB})), \tag{25}$$

where s is the number of columns of **B**. By combining (19) and (25), we obtain

$$\text{rank}(\mathbf{Y}_1) = \text{rank}(\mathbf{Y}) - \text{rank}(\mathbf{A}'\mathbf{YB}), \tag{26}$$

which may be called a generalized WG theorem. ∎

The necessary and sufficient condition for (21) and (22) to hold is stated as follows (Takane and Yanai [41]): Let $\mathbf{C} = \mathbf{B}(\mathbf{A}'\mathbf{YB})^-\mathbf{A}'$. Then, the necessary and sufficient condition is:

$$\mathbf{YCYCY} = \mathbf{YCY}. \tag{27}$$

There are a number of equivalent conditions, e.g., $(\mathbf{YCYY}^-)^2 = \mathbf{YCYY}^-$ or $(\mathbf{Y}^-\mathbf{YCY})^2 = \mathbf{Y}^-\mathbf{YCY}$. There are also a number of interesting sufficient (but not necessary) conditions, e.g., $(\mathbf{YC})^2 = \mathbf{YC}$ or $(\mathbf{CY})^2 = \mathbf{CY}$, and $\mathbf{CYC} = \mathbf{C}$ (Cline, Funderlic, and Golub [3]; Galantai [5]). The latter is even stronger than the idempotency of **YC** or **CY**.

Note 4. Although $(\mathbf{YC})^2$ is not necessarily equal to **YC** under (27), **YC** can be considered a projector in a broader sense (Rao and Yanai [25], Lemma 1). It is a projector in the restricted subspace of $\text{Sp}(\mathbf{Y}) = \text{Sp}(\mathbf{YC})$, satisfying

$$(\mathbf{YC})\mathbf{YCY} = \mathbf{YCY},$$

and

$$(\mathbf{YC})(\mathbf{Y} - \mathbf{YCY}) = \mathbf{O}.$$

We can turn **YC** into a full projector by postmultiplying it by $\mathbf{YY}^-$, a projector onto $\text{Sp}(\mathbf{Y})$ ([25], Corollary 1 to Lemma 2), which obtains $\mathbf{YCYY}^-$. This matrix is idempotent under (27), is the projector onto $\text{Sp}(\mathbf{Y})$ along $\text{Ker}(\mathbf{YCYY}^-) = \text{Sp}(\mathbf{I} - \mathbf{YCYY}^-)$, and has exactly the same effects as **YC** on matrices in $\text{Sp}(\mathbf{Y})$, e.g., **YCY** and $\mathbf{Y} - \mathbf{YCY}$. The matrices **CY** and $\mathbf{Y}^-\mathbf{YCY}$ are similar. ∎

Note 5. The condition (27) is similar in form to Cochran-Ogasawara-Takahashi's (Cochran [4]; Ogasawara and Takahashi [18]) necessary and sufficient condition for $\mathbf{x}'\mathbf{Hx}$ to follow a chi-square distribution, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. This condition is stated as

$$\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma}.$$

This is the reason why the condition stated in (27) is sometimes called the rectangular version of Cochran's condition. ∎

The WG theorem states the rank condition for the residual matrix. However, the decomposition of the data matrix **Y** the theorem implies, is even more interesting from a data analytic viewpoint:

$$\mathbf{Y} = \mathbf{YB}(\mathbf{A}'\mathbf{YB})^-\mathbf{A}'\mathbf{Y} + (\mathbf{Y} - \mathbf{YB}(\mathbf{A}'\mathbf{YB})^-\mathbf{A}'\mathbf{Y}). \tag{28}$$

Takane and Hunter [33] developed a new family of CPCA almost exclusively based on this decomposition. The second term of the above decomposition involves a Q-type projector, but it can be rewritten as a P-type projector (Takane [31]): Let $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$ be matrices such that

    (i) $\text{Sp}(\tilde{\mathbf{A}}) \subset \text{Sp}(\mathbf{Y})$,
    (ii) $\text{Sp}(\tilde{\mathbf{B}}) \subset \text{Sp}(\mathbf{Y}')$,
    (iii) $\text{rank}(\mathbf{A}'\mathbf{YB}) + \text{rank}(\tilde{\mathbf{B}}'\mathbf{Y}^-\tilde{\mathbf{A}}) = \text{rank}(\mathbf{Y})$,
    (iv) $\mathbf{A}'\mathbf{YY}^-\tilde{\mathbf{A}} = \mathbf{A}'\tilde{\mathbf{A}} = \mathbf{O}$,
    (v) $\tilde{\mathbf{B}}'\mathbf{Y}^-\mathbf{YB} = \tilde{\mathbf{B}}'\mathbf{B} = \mathbf{O}$.

Then,

$$\mathbf{Y} = \mathbf{YB}(\mathbf{A}'\mathbf{YB})^-\mathbf{A}'\mathbf{Y} + \tilde{\mathbf{A}}(\tilde{\mathbf{B}}'\mathbf{Y}^-\tilde{\mathbf{A}})^-\tilde{\mathbf{B}}'. \tag{29}$$

5 Ridge Operators

In the mid 2000's, I was interested in extending the ridge-type of regularized least squares (RLS) estimation to various MVA techniques. These extensions were rather straightforward, and I wrote most of the papers on them with my graduate students ([35–37, 45, 46]). I did not have to bother Professor Yanai. However, as I applied the RLS estimation to many MVA procedures, I thought it would be beneficial to write a paper on ridge operators, which were the common thread running through all of them ([30, 42]).

The simplest form of ridge operators is written as:

$$\mathbf{R}_G(\lambda) = \mathbf{G}(\mathbf{G}'\mathbf{G} + \lambda\mathbf{P}_{G'})^-\mathbf{G}', \tag{30}$$

where $\mathbf{P}_{G'} = \mathbf{G}'(\mathbf{GG}')^-\mathbf{G}$ is the orthogonal projector onto $\mathrm{Sp}(\mathbf{G}')$. ($\mathbf{P}_{G'} = \mathbf{I}$ if $\mathbf{G}$ is columnwise nonsingular.) This operator arises in the RLS estimation $\min_{\mathbf{c}} = \phi_\lambda(\mathbf{c})$ in regression analysis, where $\phi_\lambda(\mathbf{c}) = \mathrm{SS}(\mathbf{e}) + \lambda\mathrm{SS}(\mathbf{c})_{P_{G'}}$ and $\mathbf{e} = \mathbf{y} - \mathbf{Gc}$. (We assume, without loss of generality, that $\mathrm{Sp}(\mathbf{c}) \subset \mathrm{Sp}(\mathbf{G}')$.) The RLS estimation often obtains estimates of regression coefficients with smaller mean square errors (MSE) that are on overage closer to population parameters than the ordinary least squares (OLS) estimators [9]. It is particularly useful when the number of predictor variables is large relative to the number of cases, and/or when the predictor variables are nearly collinear (e.g., [22]). Reduced rank RLS estimates of the matrix of regression coefficients in RA can be obtained in a manner similar to reduced rank LS estimates, i.e., we first obtain non-reduced rank RLS estimates of regression coefficients, to which we apply generalized SVD for rank reduction ([36, 37]). (The generalized SVD is a special kind of SVD obtained under nonidentity metric matrices [29, 32]).

The $\mathbf{R}_G(\lambda)$ and $\mathbf{S}_G(\lambda)$ have properties similar to those of $\mathbf{P}_G$ and $\mathbf{Q}_G$, where $\mathbf{S}_G(\lambda) = \mathbf{I} - \mathbf{R}_G(\lambda)$. For example:
(i) $\mathbf{R}_G(\lambda)$ and $\mathbf{S}_G(\lambda)$ are symmetric and invariant over the choice of a g-inverse of $\mathbf{G}'\mathbf{G} + \lambda\mathbf{P}_{G'}$.
(ii) $\mathbf{R}_G(\lambda)\mathbf{K}_G(\lambda)\mathbf{R}_G(\lambda) = \mathbf{R}_G(\lambda)$ (i.e., $\mathbf{K}_G(\lambda) \in \{\mathbf{R}_G(\lambda)^-\}$, and further that $\mathbf{K}_G(\lambda) = \mathbf{R}_G(\lambda)^+$).
(iii) $\mathbf{R}_G(\lambda) - \mathbf{R}_G(\lambda)^2 = \mathbf{R}_G(\lambda)\mathbf{S}_G(\lambda) = \mathbf{S}_G(\lambda)\mathbf{R}_G(\lambda) \geq \mathbf{O}$. (This is why $\mathbf{R}_G(\lambda)$ is called a contraction matrix.)
(iv) $\mathbf{R}_G(\lambda)\mathbf{K}_G(\lambda) = \mathbf{P}_G$, etc.

Similar decompositions of $\mathbf{R}_G(\lambda)$ to those of $\mathbf{P}_G$ are also possible [42].

The ridge operators defined above can be rewritten as follows using a ridge metric matrix defined below: Let

$$\mathbf{K}_G(\lambda) = \mathbf{P}_G + \lambda(\mathbf{GG}')^+. \tag{31}$$

Then, $\mathbf{R}_G(\lambda)$ can be rewritten as:

$$\mathbf{R}_G(\lambda) = \mathbf{G}(\mathbf{G}'\mathbf{K}_G(\lambda)\mathbf{G})^-\mathbf{G}'. \tag{32}$$

The simple ridge operators introduced above can be generalized into generalized ridge operators:

$$\mathbf{R}_G^{(W,L)}(\lambda) = \mathbf{G}(\mathbf{G}'\mathbf{WG} + \lambda\mathbf{L})^-\mathbf{G}'\mathbf{W}, \tag{33}$$

where $\mathbf{L}$ is an *nnd* matrix such that $\mathrm{Sp}(\mathbf{L}) \subset \mathrm{Sp}(\mathbf{G}')$, and $\mathbf{W}$ is an *nnd* matrix such that $\mathrm{rank}(\mathbf{WG}) = \mathrm{rank}(\mathbf{G})$. As before, the generalized ridge operators can be rewritten as follows using a generalized ridge metric matrix defined below: Let

$$\mathbf{K}_G^{(W,L)}(\lambda) = \mathbf{P}_G + \lambda\mathbf{G}(\mathbf{G}'\mathbf{WG})^-\mathbf{L}(\mathbf{G}'\mathbf{WG})^-\mathbf{G}'\mathbf{W}. \tag{34}$$

Then,

$$\mathbf{R}_G^{(W,L)}(\lambda) = \mathbf{G}(\mathbf{G}'\mathbf{WK}_G^{(W,L)}(\lambda)\mathbf{G})^-\mathbf{G}'\mathbf{W}. \tag{35}$$

Ramsay and Silverman [22] use a nonidentity $\mathbf{L}$ to regulate the degree of smoothness in approximating continuous functions in functional regression analysis.

Note 6. The generalized ridge operator can also be characterized as follows. Let

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{O} \end{bmatrix}, \quad \tilde{\mathbf{G}} = \begin{bmatrix} \mathbf{G} \\ (\lambda\mathbf{L})^{1/2} \end{bmatrix}, \quad \text{and} \quad \tilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}.$$

The weighted LS estimation leads to a partitioned $\tilde{W}$-orthogonal projector,

$$\mathbf{P}_{\tilde{G}/\tilde{W}} = \left[\begin{array}{cc} \mathbf{R}_G^{(W,L)}(\lambda) & \mathbf{A} \\ \mathbf{A}'\mathbf{W} & \mathbf{C} \end{array} \right], \tag{36}$$

where $\mathbf{A} = (\lambda)^{1/2}\mathbf{G}(\mathbf{G}'\mathbf{WG} + \lambda\mathbf{L})^{-}\mathbf{L}^{1/2}$, and $\mathbf{C} = \lambda\mathbf{L}^{1/2}(\mathbf{G}'\mathbf{WG} + \lambda\mathbf{L})^{-}\mathbf{L}^{1/2}$. From the idempotency of $\mathbf{P}_{\tilde{G}/\tilde{W}}$, it follows that

$$\mathbf{R}_G^{(W,L)}(\lambda) - (\mathbf{R}_G^{(W,L)}(\lambda))^2 = \mathbf{A}\mathbf{A}'\mathbf{W} \geq \mathbf{O}, \tag{37}$$

$$\mathbf{C} - \mathbf{C}^2 = \mathbf{A}'\mathbf{WA} \geq \mathbf{O} \quad (\mathbf{C} \text{ is also a contraction matrix}), \tag{38}$$

and

$$\mathbf{R}_G^{(W,L)}(\lambda)\mathbf{A} + \mathbf{A}\mathbf{C} = \mathbf{A}. \tag{39}$$

(37) generalizes Property (iii) above. ∎

6 Generalized Constrained Canonical Correlation Analysis

In the external analysis of CPCA, a data matrix is decomposed into several components by external information. I initially thought ([34]) we could do the same in generalized constrained canonical correlation analysis (GCCANO). We decompose $\mathbf{X}$ and $\mathbf{Y}$ (the matrices of observations on the two sets of variables) separately into several orthogonal components, and then choose one term from each decomposition, and apply CANO to the pair, which amounts to SVD of the product of the orthogonal projectors. It has turned out that this strategy will not work.

CANO analyzes total association between $\mathbf{X}$ and $\mathbf{Y}$, i.e., $\text{tr}(\mathbf{P}_X\mathbf{P}_Y)$. However, $\mathbf{X} = \mathbf{M} + \mathbf{N}$, where $\mathbf{M}'\mathbf{N} = \mathbf{O}$ does not guarantee $\mathbf{P}_X = \mathbf{P}_M + \mathbf{P}_N$. This may be contrasted with a similar situation in which $\mathbf{X} = [\mathbf{M}, \mathbf{N}]$, where $\mathbf{M}'\mathbf{N} = \mathbf{O}$, in which case we indeed have $\mathbf{P}_X = \mathbf{P}_M + \mathbf{P}_N$. This suggests that we need orthogonal decompositions of projectors to derive additive decompositions of the total association. In CPCA, orthogonal decompositions of projectors were used to obtain orthogonal decompositions of data matrices, while in GCCANO, they were needed to be directly inserted into the trace operation (to secure additivity in the decompositions of the association).

Takane, Yanai, and Hwang [48] derived the following two orthogonal decompositions of $\mathbf{P}_{[X,G]}$ by combining two orthogonal decompositions ((iii) and (v)) of the orthogonal projector given in the CPCA section:

(1) Let $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{W}$ be matrices such that $\text{Sp}(\mathbf{A}) = \text{Ker}(\mathbf{H}'\mathbf{X}'\mathbf{P}_G\mathbf{X})$, $\text{Sp}(\mathbf{B}) = \text{Ker}(\mathbf{H}'\mathbf{X}'\mathbf{Q}_G\mathbf{X})$, and $\text{Sp}(\mathbf{W}) = \text{Ker}(\mathbf{X}'\mathbf{G})$. Then,

$$\mathbf{P}_{[X,G]} = \mathbf{P}_{P_GXH} + \mathbf{P}_{P_GXA} + \mathbf{P}_{Q_GXH} + \mathbf{P}_{Q_GXB} + \mathbf{P}_{GW}. \tag{40}$$

(2) Let $\mathbf{K}$, $\mathbf{U}$, and $\mathbf{V}$ be matrices such that $\text{Sp}(\mathbf{K}) = \text{Ker}(\mathbf{H}'\mathbf{X}'\mathbf{X})$, $\text{Sp}(\mathbf{U}) = \text{Ker}(\mathbf{G}'\mathbf{XH})$, and $\text{Sp}(\mathbf{V}) = \text{Ker}(\mathbf{G}'\mathbf{XK})$. Then,

$$\mathbf{P}_{[X,G]} = \mathbf{P}_{P_{XH}G} + \mathbf{P}_{XHU} + \mathbf{P}_{P_{XK}G} + \mathbf{P}_{XKV} + \mathbf{P}_{Q_XG}. \tag{41}$$

The meaning of the terms in the above decompositions are given in [48]. We can derive similar decompositions of $\mathbf{P}_{[Y,G_Y]}$ (The subscript Y is put on G to indicate that this is a $\mathbf{G}$ matrix for $\mathbf{Y}$.) We take one term each from a decomposition of $\mathbf{P}_{[X,G_X]}$ and that of $\mathbf{P}_{[Y,G_Y]}$, and apply SVD to the product of the two, e.g.,

$$\text{SVD}(\mathbf{P}_{Q_{G_X}XH_X}\mathbf{P}_{YH_YU_Y}). \tag{42}$$

A great variety of part CANO's are realized in this way ([48]).

7 Epilog

This paper overviewed Professor Yanai's contributions to MVA. He adamantly emphasized linear algebraic aspects of MVA in his approach to MVA ([60]). I feel extremely lucky that I was exposed to his idea on MVA when I was young ([50]). After almost half a century since then, I am still working strictly within the limit of his framework. I have little idea when I can break through the boundary.

Yet we have to move forward because there still is a long way to go. In this section, I would like to offer to introduce my last conversation with Professor Yanai. In many cases, our collaboration started this way, and if Professor Yanai had lived longer, this may have evolved into another joint work. It was on September 22, 2013, immediately after he got out of hospital due to his head injury incurred by falling off from stairs in his own house. Although he looked somewhat weak physically, his mood was high. He even told me that he intended to resume teaching at his institution (St. Luke's College of Nursing) on a wheel chair the following week. I was greatly relieved by his remark, and forgot all about his prostate cancer until the moment I received an email notifying me of his death in December of the same year. The topic of our conversation was causal inference in statistics. Just prior to my visit at Professor Yanai, I was attending a symposium in Osaka on missing data and causal inference in statistics, and I probably raised the issue during our conversation as part of a report on my recent activities. It turned out that it was one of the most memorable conversations with him.

When randomization is unavailable, there are a lot of pitfalls in establishing causal relationships based on correlational relationships alone. One crucial aspect of the problem is how to eliminate the effects of confounding variables. The "easiest" way to deal with the problem is to include the effects of the confounding variables in regression analysis along with the predictor variable of interest, although this is easier said than done. Identifying the set of confounding variables is not so easy, although here we assume that they are known. Let $\mathbf{y}$ denote the criterion variable, let $\mathbf{g}$ denote the predictor variable of interest, and let $\mathbf{U}$ denote the matrix of confounding variables. The suggested regression model can be written as:

$$\mathbf{y} = \mathbf{g}a_1 + \mathbf{U}\mathbf{c} + \mathbf{e}_1. \tag{43}$$

The ordinary least squares (OLS) estimate of $\mathbf{g}a_1$ is given by

$$\mathbf{g}\hat{a}_1 = \mathbf{P}_{g/Q_U}\mathbf{y} \tag{44}$$

.

Consider next the regression of $\mathbf{g}$ onto $\mathbf{U}$, i.e.,

$$\mathbf{g} = \mathbf{U}\mathbf{d} + \mathbf{e}_2. \tag{45}$$

The OLS estimate of $\mathbf{U}\mathbf{d}$ is given by

$$\mathbf{U}\hat{\mathbf{d}} = \mathbf{P}_U\mathbf{g}. \tag{46}$$

We call $\mathbf{P}_U\mathbf{g}$ linear propensity scores. Residuals from the above regression, $\mathbf{Q}_U\mathbf{g}$, represent the portions of $\mathbf{g}$ left unaccounted for by $\mathbf{U}$.

We next consider using $\mathbf{P}_U\mathbf{g}$ instead of $\mathbf{U}$ in the first regression, i.e.,

$$\mathbf{y} = \mathbf{g}a_2 + \mathbf{P}_U\mathbf{g}b + \mathbf{e}_3. \tag{47}$$

The OLS estimate of $\mathbf{g}a_2$ is given by

$$\mathbf{g}\hat{a}_2 = \mathbf{P}_{g/Q_{P_U g}}\mathbf{y}, \tag{48}$$

where $\mathbf{Q}_{P_U g} = \mathbf{I} - \mathbf{P}_U\mathbf{g}(\mathbf{g}'\mathbf{P}_U\mathbf{g})^{-1}\mathbf{g}'\mathbf{P}_U$.

Since

$$\mathbf{Q}_{P_U g}\mathbf{g} = \mathbf{g} - \mathbf{P}_U\mathbf{g}(\mathbf{g}'\mathbf{P}_U\mathbf{g})^{-1}\mathbf{g}'\mathbf{P}_U\mathbf{g} = \mathbf{Q}_U\mathbf{g}, \tag{49}$$

we obtain

$$\mathbf{P}_{g/Q_{P_U g}}\mathbf{y} = \mathbf{P}_{g/Q_U}\mathbf{y}. \tag{50}$$

This means (44) and (48) are equivalent. This gives the rationale for replacing $\mathbf{U}$ by $\mathbf{P}_U\mathbf{g}$. The latter is more convenient because it is a single variable, and matching on a single variable is much easier than matching on multiple variables.

More recently, methods of causal inference based on instrumental variables are getting popular. An instrumental variable $\mathbf{z}$ has the following properties:

(i) $\mathbf{z}'\mathbf{U} = \mathbf{0}$ ($\mathbf{z}$ and $\mathbf{U}$ are uncorrelated),
(ii) $\mathbf{z}'\mathbf{g} \neq 0$ ($\mathbf{z}$ and $\mathbf{g}$ are correlated),
(iii) $\mathbf{z}'\mathbf{Q}_{[U,g]}\mathbf{y} = 0$ (*i.e.*, $\mathbf{z}$ has a predictive power on $\mathbf{y}$ only through $\mathbf{g}$).

How is $\mathbf{z}$ related to $\mathbf{P}_U\mathbf{x}$ or $\mathbf{Q}_U\mathbf{g}$?

Assume $\mathbf{z} = c\mathbf{Q}_U\mathbf{g}$, where c is a normalization factor. This $\mathbf{z}$ satisfies (i) and (ii) above. That it also satisfies (iii) can be seen from:

$$(1/c)\mathbf{z}'\mathbf{Q}_{[U,g]}\mathbf{y} = \mathbf{g}'\mathbf{Q}_U\mathbf{Q}_{[U,g]}\mathbf{y} = \mathbf{g}'\mathbf{Q}_{[U,g]}\mathbf{y} = 0. \tag{51}$$

Now consider the simple regression model:

$$\mathbf{y} = \mathbf{g}a_3 + \mathbf{e}_4. \tag{52}$$

But here we use instrumental variable (IV) estimation instead of LS estimation. The IV estimate of $\mathbf{g}a_3$ is given by

$$\mathbf{g}\hat{a}_3 = \mathbf{P}_{g/P_z}\mathbf{y} = \mathbf{P}_{g/Q_U}\mathbf{y}. \tag{53}$$

Since $\mathbf{P}_z = \mathbf{Q}_U\mathbf{g}(\mathbf{g}'\mathbf{Q}_U\mathbf{g})^{-1}\mathbf{g}'\mathbf{Q}_U$ and $\mathbf{g}'\mathbf{P}_z = \mathbf{g}'\mathbf{Q}_U$, this is identical to (44) and (48). This implies that the $\mathbf{z}$ defined above is an ideal IV.

8 Concluding Remarks

In my view, MVA has been one of the most exciting areas of research in statistics. Recent trends in this area may be summarized as 1) nonlinearization, 2) regularization, and 3) nonparameterization. Linear algebra and theories of linear models have provided ground work for many of these extensions (e.g., Rao [23]; Puntanen, Styan, and Isotalo [21]). Kernel methods for nonlinear MVA, for example, linearize the problem using a so-called kernel trick [35]. The RLS estimation discussed in Section 5 is an effective method of regularization, but it basically consists of linear operations combined with a nonparametric resampling procedure (e.g., the bootstrap method) to select an "optimal" value of a tuning parameter. Professor Yanai has indeed been correct in his emphasis on linear algebra and models. There are as well limitations in his approach. One of them relates to the fact that while he did not actively oppose the use of iterative optimization procedures, he himself never ventured into developments of techniques involving such procedures ([11, 47]).

References

[1] U. Böckenholt, I. Böckenholt, Canonical analysis of contingency tables with linear constraints, Psychometrika, 55 (1990), 633–639.

[2] M. T. Chu, R. E. Funderlic, G. H. Golub, A rank-one reduction formula and its applications to matrix factorizations, SIAM Review, 37 (1995), 512–530.

[3] R. E. Cline, R. E. Funderlic, The rank of a difference of matrices and associated generalized inverses, Linear Algebra Appl., 24 (1979), 185–215.

[4]   W. G. Cochran, The distribution of quadratic forms in a normal system with applications to analysis of covariance, Proc. Camb. Phil. Soc., 30 (1934), 178–191.

[5]   A. Galantai, A note on generalized rank reduction, Act. Math. Hung., 116 (2007), 239–246.

[6]   L. Guttman, General theory and methods for matric factoring, Psychometrika, 9 (1944), 1–16.

[7]   L. Guttman, A necessary and sufficient formula for matric factoring, Psychometrika, 22 (1957), 79–81.

[8]   F. R. Helmert, Adjustment Computations by the Method of Least Squares (in German), 2nd Edition (Teubnei, Leipzig, 1907).

[9]   A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12 (1970), 55–67.

[10] L. Hubert, J. Meulman, W. J. Heiser, Two purposes of matrix factorization: A historical appraisal, SIAM Review, 42 (2000), 68–82.

[11] H. Hwang, Y. Takane, Generalized Structured Component Analysis: A Component-Based Approach to Structural Equation Modeling (Chapman and Hall/CRC Press, Boca Raton, FL., 2014).

[12] C. G. Khatri, A simplified approach to the derivation of the theorems on the rank of a matrix, J. of the Maharaja Sayajirao University of Baroda 10 (1961), 1–5.

[13] C. G. Khatri, A note on a MANOVA model applied to problems in growth curve, Ann. I. Stat. Math., 18 (1966), 75–86.

[14] C. G. Khatri, Some properties of BLUE in a linear model and canonical correlations associated with linear transformations, J. Multivariate Anal., 34 (1990), 211–226.

[15] L. R. LaMotte, A direct derivation of the REML likelihood function, Stat. Pap., 48 (2007), 321–327.

[16] R. J. Light, B. H. Margolin, An analysis of variance of categorical data, J. Am. Stat. Assoc., 66 (1971), 534–544.

[17] S. Loisel, Y. Takane, Partitions of Pearson's chi-square statistic for frequency tables: A comprehensive account, Computation. Stat., in press.

[18] T. Ogasawara, M. Takahashi, Indepedence of quadratic forms in normal system, J. Sci. Hiroshima University, 15 (1951), 1–9.

[19] K. Pearson, On the criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag., 50 (1900), 157–172.

[20] R. F. Potthoff, S. N. Roy, A generalized multivariate analysis of variance model useful for growth curve problem, Biometrika, 51 (1964), 313–326.

[21] S. Puntanen, G. P. H. Styan, J. Isotalo, Matrix Tricks for Linear Statistical Models (Springer, Berlin, 2011).

[22] J. O. Ramsay, B. M. Silverman, Functional Data Analysis, Second Edition (Springer, New York, 2005).

[23] C. R. Rao, Linear Statistical Inference and Its Applications, Second Edition (Wiley, New York, 1973).

[24] C. R. Rao, S. K. Mitra, Generalized Inverse of Matrices amd Its Applications (Wiley, New York, 1971).

[25] C. R. Rao, H. Yanai, General definition and decomposition of projectors and some applications to statistical problems, J. Stat. Plan. Infer., 3 (1979), 1–17.

[26] B. Schaffrin, Personal communication (2015).

[27] G. F. A. Seber, Multivariate Observations (Wiley, New York, 1984.)

[28] A. Shapiro, Asymptotic theory of overparameterized structural models, J. Am. Stat. Assoc., 81 (1986), 142–149.

[29] Y. Takane, Relationships among various kinds of eigenvalue and singular value decompositions, In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. Meulman (Eds.), New Developments in Psychometrics (Springer, Tokyo, 2003), 45–56.

[30] Y. Takane, More on regularization and (generalized) ridge operators, In: K. Shigemasu, A. Okada, T. Imaizumi, T. Hoshino (Eds.), New Trends in Psychometrics (University Academic Press, Tokyo, 2008) 443–452.

[31] Y. Takane, Constrained Principal Component Analysis and Related Techniques (Chapman and Hall/CRC Press, Boca Raton, FL, 2013).

[32] Y. Takane, M. A. Hunter, Constrained principal component analysis: A comprehensive theory, Appl. Algebr. Eng. Comm., 12 (2001), 391–419.

[33] Y. Takane, M. A. Hunter, New family of constrained principal component analysis (CPCA), Linear Algebra Appl., 434 (2011), 2539–2555.

[34] Y. Takane, H. Hwang, Generalized constrained canonical correlation analysis, Multivar. Behav. Res., 37 (2002), 163–195.

[35] Y. Takane, H. Hwang, Regularized multiple correspondence analysis. In: J. Blasius, M. J. Greenacre (Eds.), Multiple correspondence analysis and related methods (Chapman and Hall, London, 2006) 259–279.

[36] Y. Takane, S. Jung, Regularized partial and/or constrained redundancy analysis, Psychometrika, 73 (2008), 671–690.

[37] Y. Takane, S. Jung, Regularized nonsymmetric correspondence analysis, Comput. Stat. Data An., 53 (2009), 3159–3170.

[38] Y. Takane, S. Jung, Tests of ignoring and eliminating in nonsymmetric correspondence analysis, Adv. Data Anal. Classif., 3 (2009), 315–340.

[39] Y. Takane, T. Shibayama, Principal component analysis with external information on both subjects and variables, Psyhometrika, 56 (1991), 97–120.

[40] Y. Takane, H. Yanai, On oblique projectors, Linear Algebra Appl., 289 (1999), 297–310.

[41] Y. Takane, H. Yanai, On the Wedderburn-Guttman theorem, Linear Algebra Appl. 410 (2005), 267–278.

[42] Y. Takane, H. Yanai, On ridge operators, Linear Algebra Appl., 428 (2008), 1778–1790.

[43] Y. Takane, L. Zhou, On two expressions of the MLE for a special case of the extended growth curve models, Linear Algebra Appl., 436 (2012), 2567–2577.

[44] Y. Takane, L. Zhou, Anatomy of Pearson's chi-square statistic in three-way contingency tables, In: R. E. Millsap, L. A. van der Ark, D. M. Bolt, C. M. Woods (Eds.), New Developments in Quantitative Psychology (Springer, New York, 2013), 41–57.

[45] Y. Takane, H. Hwang, H. Abdi, Regularized multiple-set canonical correlation analysis, Psychometrika, 73 (2008), 753–775.

[46] Y. Takane, K. Jung, H. Hwang, Regularized growth curve models, Comput. Stat. Data An., 55 (2011), 1041–1052.

[47] Y. Takane, H. A. L. Kiers, J. de Leeuw, Component analysis with different constraints on different dimensions, Psychometrika, 60 (1995), 259–280.

[48] Y. Takane, H. Yanai, H. Hwang, An improved method for generalized constrained canonical correlation analysis, Comp. Stat. Data An., 50 (2006), 221–241.

[49] Y. Takane, H. Yanai, S. Mayekawa, Relationships among several methods of linearly constrained correspondence analysis, Psychometrika, 56 (1991), 667–684.

[50] K. Takeuchi, H. Yanai, B. N. Mukherjee, The Foundation of Multivariate Analysis (Wiley Eastern, New Delhi, and Halsted Press, New York, 1982).

[51] C. J. F. ter Braak, Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis, Ecology, 67 (1986), 1167–1179.

[52] Y. Tian, The Moore-Penrose inverses of $m \times n$ block matrices and their applications, Linear Algebra Appl., 283 (1998), 35–60.

[53] Y. Tian, Upper and lower bounds for ranks of matrix expressions using generalized inverses, Linear Algebra Appl., 355 (2002), 187–214.

[54] Y. Tian, G. P. H. Styan, On some matrix equalities for generalized inverses with applications, Linear Algebra Appl., 430 (2009), 2716–2733.

[55] A. P. Verbyla, A conditional derivation of residual maximum likelihood, Aust. J. Stat., 32 (1990), 227–230.

[56] J. H. M. Wedderburn, Lectures on Matrices, Colloquium Publication, Vol. 17 (American Mathematical Society, Providence, 1934).

[57] H. Yanai, Factor analysis with external criteria, Jpn. Psychol. Res., 12 (1970), 143–153.

[58] H. Yanai, Some generalized forms of least squares g-inverse, minimum norm g-inverse and Moore-Penrose inverse matrices, Comput. Stat. Data An., 10 (1990), 251–260.

[59] H. Yanai, Y. Takane, Canonical correlation analysis with linear constraints, Linear Algebra Appl., 176 (1992), 75–82.

[60] H. Yanai, K. Takeuchi, Y. Takane, Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition (Springer, New York, 2011).