Contributions to Multivariate Analysis due to C. R. Rao and Associated Developments

Yasunori Fujikoshi

Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi Hiroshima, Hiroshima 739-8626, Japan

Abstract

C. R. Rao has made various significant contributions to multivariate analysis. Among them, we consider the following topics: (i) Rao's U-statistic in discriminant analysis, (ii) MANOVA tests, (iii) Asymptotic expansion and Rao's F approximation for Λ statistic, (iv) Growth curve analysis, and (v) Information criteria for the selection of variables. Some of these were introduced at the dawn of multivariate analysis. Under topic (v), we also discuss recent developments on the selection of variables in discriminant analysis.

AMS 2000 Subject Classification: primary 62H15; secondary 62H10

Key Words and Phrases: Additional information, Growth curve analysis, MANOVA tests, Model selection. Rao's F approximation, Rao's U-statistic.

Abbreviated title: Contributions to Multivariate Analysis due to Rao

1. Introduction

In this paper we consider some important contributions to multivariate analysis due to C. R. Rao, and overview associated developments. In Section 2, we focus on Rao's U-statistic for additional information in two-group discriminant analysis. This research leads to a development of statistical methods for the selection of variables. In Section 3, multivariate analysis of variance (MANOVA) problems are discussed, based on Rao (1948). We note that Rao developed various types of tests based on real data, which are essentially LR tests. One of these is to test an additional information hypothesis for a set of response variables. Section 4 considers the distribution of a Lambda statistic, $\Lambda_p(q, n-q)$, which appears as the null distribution for various tests, including a MANOVA test. It is noted that an asymptotic expansion of $T = -\{n - (p+q+1)/2\}\Lambda_p(q, n-q)$ was first obtained by Rao (1948). Afterwards, Box (1952) gave an asymptotic expansion for a class of statistics including T. Rao (1952) proposed a highly accurate F approximation for a transformed version of $\Lambda_p(q, n-q)$. Section 5 is concerned with analysis of growth curve data. Rao (1958, 1965) introduced two types of models for such data and developed statistical inference of the growth curve models.

It is important to examine whether a set of variables has additional information in the presence of a given set of variables. Such notions were discussed by Rao and others in various models. Applying information criteria such as AIC and BIC to such models, variable selection methods have been proposed. After explaining these, in Section 6, we provide more detail on discriminant analysis.

We note that there have been many other important contributions to multivariate analysis due to Rao that are not covered in this paper, some of which are concerned with topics in the following areas: (a) Factor analysis (Rao (1955), etc.). (b) Principal component analysis (Rao (1964), etc.). (c) Correspondence analysis (Rao (1997), etc.). (d) Separation theorems and

reduction of dimensionality (Rao (1979), etc.).

2. Rao's U-Statistic in Discriminant Analysis

In two-group discriminant analysis, Rao (1946) investigated whether some variables can be dropped without losing discriminative information. One of his motivations was to reduce computational problems, in addition to enabling efficient discrimination. He proposed the following test of an additional information hypothesis, which determines whether augmenting a given set of p variables with another set of q variables provides additional discrimination between two populations. Suppose that there are n_i samples from $(p+q)$-variate populations, and let D^2_{p+q} and D^2_p be the squared Mahalanobis distances based on the $(p+q)$ variate and the p variate. He proposed a test statistic

$$U = \frac{n-(p+q)-1}{q}\frac{n_1 n_2(D^2_{p+q}-D^2_p)}{n(n-2)+n_1 n_2 D^2_p}. \tag{2.1}$$

whose null distribution is an F-distribution with degrees of freedom q and $n-(p+q)-1$, where $n=n_1+n_2$. The statistics U or $c(D^2_{p+q}-D^2_p)/\{n-2+cD^2_p\}$, where $c=n_1n_2/n$, were named Rao's U-statistic by Kshirsagar (1972). The test based on U-statistic is called U-test.

The above additional information hypothesis can be formulated as follows. In two-group discriminant analysis we have two populations $\Pi_i, i=1,2$, and n_i observations from Π_i of p-dimensional variate $\boldsymbol{Y}$. The mean vectors of $\boldsymbol{Y}$ when $\boldsymbol{Y}\in\Pi_i$ are

$$\mathrm{E}(\boldsymbol{Y}\mid\Pi_i)=\boldsymbol{\mu}^{(i)},\quad i=1,2,$$

where it is assumed that the covariance matrices are the same, i.e., $\mathrm{Var}(\boldsymbol{Y}\mid\Pi_i)=\boldsymbol{\Sigma}$. In discriminant analysis, we are interested in which set of variables are important, or which set of variables are redundant. Let $\boldsymbol{Y}$ be decomposed as $\boldsymbol{Y}=(\boldsymbol{Y}_1',\boldsymbol{Y}_2')'$ where $\boldsymbol{Y}_i; p_i\times 1$. Let us formulate the notion that $\boldsymbol{Y}_2$ provides no additional information for the discriminant analysis in the presence of $\boldsymbol{Y}_1$,

or simply that $\boldsymbol{Y}_1$ is sufficient or $\boldsymbol{Y}_2$ is redundant. We refer to such a notion as the sufficiency of $\boldsymbol{Y}_1$ or the redundancy of $\boldsymbol{Y}_2$.

When the parameters are known, it is natural to classify a new observation $\boldsymbol{Y}$ into Π_1 if

$$(\boldsymbol{Y}-\boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}^{(1)}) < (\boldsymbol{Y}-\boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}^{(2)}) \tag{2.2}$$

and otherwise classify $\boldsymbol{Y}$ into Π_2. This expression (2.2) is equivalent to $L(\boldsymbol{Y},\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)},\boldsymbol{\Sigma})>0$, where

$$L(\boldsymbol{Y},\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)},\boldsymbol{\Sigma}) = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2), \tag{2.3}$$

which is called the population discriminant function. The coefficients of the population discriminant function are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}) = (\boldsymbol{\beta}_1',\boldsymbol{\beta}_2')',$$

where $\boldsymbol{\beta}_1 : p_1\times 1$ and $\boldsymbol{\beta}_2 : p_2\times 1$. One way to define the redundancy of $\boldsymbol{Y}_2$ is to define it as $\boldsymbol{\beta}_2=\boldsymbol{0}$. Let δ and δ_1 be the population Mahalanobis distances between Π_1 and Π_2 based on $\boldsymbol{Y}$ and $\boldsymbol{Y}_1$, respectively. Then,

$$\begin{aligned}\delta^2 &= (\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}),\\ \delta_1^2 &= (\boldsymbol{\mu}_1^{(1)}-\boldsymbol{\mu}_1^{(2)})'\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\mu}_1^{(1)}-\boldsymbol{\mu}_1^{(2)}),\end{aligned}$$

where $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\Sigma}$ are partitioned as

$$\boldsymbol{\mu}^{(i)} = \begin{pmatrix}\boldsymbol{\mu}_1^{(i)}\\ \boldsymbol{\mu}_2^{(i)}\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}, \tag{2.4}$$

$\boldsymbol{\mu}_i^{(g)} : p_i\times 1, g=1,2$ and $\boldsymbol{\Sigma}_{ij} : p_i\times p_j$. It is also reasonable to define the redundancy of $\boldsymbol{Y}_2$ as $\delta^2=\delta_1^2$. Note that we have

$$\delta^2 = \delta_1^2 + \delta_{2\cdot 1}^2, \tag{2.5}$$

where

$$\begin{aligned}\delta_{2\cdot 1}^2 &= (\boldsymbol{\mu}_{2\cdot 1}^{(1)}-\boldsymbol{\mu}_{2\cdot 1}^{(2)})'\Sigma_{22\cdot 1}^{-1}(\boldsymbol{\mu}_{2\cdot 1}^{(1)}-\boldsymbol{\mu}_{2\cdot 1}^{(2)}),\\ \boldsymbol{\Sigma}_{22\cdot 1} &= \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12},\\ \boldsymbol{\mu}_{2\cdot 1}^{(i)} &= \boldsymbol{\mu}_2^{(i)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1^{(i)}, \quad i=1,\ 2.\end{aligned}$$

This relation is obtained by substituting a well-known inverse matrix formula,

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} + \begin{pmatrix} -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \mathbf{I}_{p-k} \end{pmatrix} \boldsymbol{\Sigma}_{22\cdot 1}^{-1} \begin{pmatrix} -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I}_{p-k} \end{pmatrix}$$

for $\boldsymbol{\Sigma}^{-1}$ in δ^2. On the other hand, the coefficient vector of the linear discriminant function is expressed as

$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_1^{(2)}) - \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_2, \quad \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{22\cdot 1}^{-1}(\boldsymbol{\mu}_{2\cdot 1}^{(1)} - \boldsymbol{\mu}_{2\cdot 1}^{(2)}).$$

From these results, we can see that, as proved by Rao (1970), the following three statements are equivalent:

$$\text{(i)} \quad \delta^2 = \delta_1^2, \quad \text{(ii)} \quad \boldsymbol{\mu}_{2\cdot 1}^{(1)} = \boldsymbol{\mu}_{2\cdot 1}^{(2)}, \quad \text{(iii)} \quad \boldsymbol{\beta}_2 = 0.$$

The second statement is related to the equality of conditional means. In fact

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_2^{(i)} \mid \boldsymbol{Y}_1^{(i)}) &= \boldsymbol{\mu}_2^{(i)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y}_1^{(i)} - \boldsymbol{\mu}_1^{(i)}) \\ &= \boldsymbol{\mu}_{2\cdot 1}^{(i)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{Y}_1^{(i)}, \quad i = 1, 2. \end{aligned} \tag{2.6}$$

Statements (i) and (iii) help in understanding that $\boldsymbol{Y}_2$ provides no additional information for the discriminant analysis in the presence of $\boldsymbol{Y}_1$. Statement (iii) is used for obtaining a likelihood ratio test for (i) or (iii), which is equivalent to a U-test. Statements (i), (ii) and (iii) and their equivalence were extended to the case of several groups by Fujikoshi (1982). Gupta et al. (2006) derive a large sample asymptotic expansion of Rao's U-statistic under nonnormality. Pynnönen (1987) extended the notion of redundancy to the case where the covariance matrices are different.

In general, it is important to formulate that a subset of response or explanatory variables is sufficient, or the set of remainder variables has no additional information or redundant, as in discriminant analysis. It is also important to extend statistical inferences for such formulations. For some of such results, see Fujikoshi (1989, 1992). In section 6, we see that such formulations are used in variable selection methods.

3. MANOVA Tests

Rao made many important contributions to MANOVA through the analysis of various real data. First we note that he gives an example in his book (1952) where Mahalanobis D^2 (or Hotelling T^2) based on two variables showed no significance between the two populations; whereas two sample t-tests based on each of the variables were highly significant. This is the first example of what is called the "curse of dimensionality" in multivariate analysis, which was named as Rao's paradox by Healy (1969) and Rencher (2002, p.116). Rencher (2002) explained this paradox in detail, and also showed the situation that, conversely, the multivariate test is more powerful in some situations, despite the univariate tests are not being significant. In general, the "curse of dimensionality" phrase was introduced by Bellman (1957) for describing the problem caused by the exponential increase in volume associated with adding extra dimensions to the Euclidean space. When we are concerned with the analysis of a p variate, we might be concerned with the analysis of various subsets of the p variate. Related to this problem, Rao (1966a) gave conditions under which additional variables are useful in tests of significance.

As in a typical MANOVA model, consider a multivariate one-way analysis of variance model, in which we measure p dependent variables on each experimental unit instead of just one variable. We consider q treatments and assign n_i subjects to the ith treatment. It is assumed that all of the n $(= n_1 + \cdots + n_q)$ observations are normally distributed with the common covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{in_i}$ be samples from the ith treatment group $\mathrm{N}_p(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$. For testing the equality of the mean vectors, i.e., $H_0 : \boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(q)}$, let $\mathbf{B}$ and $\mathbf{W}$ be the matrices of sums of squares and products due to treatments (between groups) and errors (within groups), respectively. These matrices are defined by

$$\mathbf{B} = \sum_{i=1}^{q} n_i(\bar{\boldsymbol{Y}}_{i\cdot} - \bar{\boldsymbol{Y}}_{\cdot\cdot})(\bar{\boldsymbol{Y}}_{i\cdot} - \bar{\boldsymbol{Y}}_{\cdot\cdot})', \quad \mathbf{W} = \sum_{i=1}^{q}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i\cdot})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i\cdot})'$$

where $\bar{\boldsymbol{Y}}_{i\cdot} = (1/n_i)\sum_{j=1}^{n_i} \boldsymbol{Y}_{ij}$ and $\bar{\boldsymbol{Y}}_{\cdot\cdot} = (1/n)\sum_{i=1}^{q}\sum_{j=1}^{n_i} \boldsymbol{Y}_{ij}$. Then, under H_0, B and W are independently distributed as Wishart distributions $\mathrm{W}_p(q-1, \boldsymbol{\Sigma})$ and $\mathrm{W}_p(n-q, \boldsymbol{\Sigma})$, respectively. Letting $\mathsf{T} = \mathsf{B} + \mathsf{W}$, an LR test for H_0 is based on $\Lambda = |\mathsf{W}|/|\mathsf{T}|$, whose null distribution does not depend on $\boldsymbol{\Sigma}$ and is denoted by $\Lambda_p(q-1, n-q)$. Such Λ was called Wilks Lambda in Rao (1948), based on the underlying theory of Λ due to Wilks (1932).

In MANOVA, there are two types of problems. One is the problem of comparing the mean vectors as in the above one-way MANOVA model. The other is the problem of comparing within the mean vectors. Rao (1948) gave various types of MANOVA methods through real data, most of which can also be formulated as a general testing problem in a multivariate linear model. A multivariate linear model is given by

$$\mathsf{Y} = \mathbf{A}\boldsymbol{\Theta} + \mathsf{E}, \tag{3.1}$$

where $\mathbf{A}$ is an $n \times k$ given matrix and $\boldsymbol{\Theta}$ is a $k \times p$ unknown parameter matrix. It is assumed that the rows of the error matrix E are independently distributed as a p-variate normal distribution with mean zero and unknown covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Various hypotheses are expressed as

$$H_g : \ \mathbf{C}\Theta\mathbf{D} = \mathbf{O}, \tag{3.2}$$

where $\mathbf{C}$ and $\mathbf{D}$ are given matrices of $c \times k$ and $p \times d$ with ranks c and d, respectively. In fact, a relation between the row vectors of $\boldsymbol{\Theta}$ and a relation within the row vectors of $\boldsymbol{\Theta}$ are expressed by respectively defining $\mathbf{C}$ and $\mathbf{D}$, as appropriate. The likelihood ratio test is based on

$$\Lambda = \frac{|\mathsf{S}_e|}{|\mathsf{S}_e + \mathsf{S}_h|} = \frac{|\mathsf{S}_e|}{|\mathsf{S}_t|}, \tag{3.3}$$

whose null distribution is $\Lambda_d(c, n-k)$, where

$$\begin{aligned}
\mathsf{S}_h &= \{\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathsf{Y}\mathbf{D}\}'\{\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\}^{-1}\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathsf{Y}\mathbf{D}, \\
\mathsf{S}_e &= \mathbf{D}'\mathsf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathsf{Y}\mathbf{D}, \\
\mathsf{S}_t &= \mathsf{S}_h + \mathsf{S}_w.
\end{aligned}$$

Rao (1948) proposed a test for whether $\boldsymbol{Y}_2 = (Y_{k+1}, \ldots, Y_p)'$ brings out further differences in q populations when the differences due to $\boldsymbol{Y}_1 = (Y_1, \ldots, Y_k)'$ are removed. Let us consider this problem in a multivariate one-way MANOVA or multi-group discriminant model. Such an additional information hypothesis may be defined as

$$\boldsymbol{\mu}_{2\cdot 1}^{(1)} = \cdots = \boldsymbol{\mu}_{2\cdot 1}^{(q)}, \tag{3.4}$$

where $\boldsymbol{\mu}_{2\cdot 1}^{(i)} = \boldsymbol{\mu}_2^{(i)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1^{(i)}, i = 1, \ldots, q$. Here, $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\Sigma}$ have been decomposed as in (2.4). Let us decompose B and W as

$$\mathsf{B} = \begin{pmatrix} \mathsf{B}_{11} & \mathsf{B}_{12} \\ \mathsf{B}_{21} & \mathsf{B}_{22} \end{pmatrix}, \quad \mathsf{W} = \begin{pmatrix} \mathsf{W}_{11} & \mathsf{W}_{12} \\ \mathsf{W}_{21} & \mathsf{W}_{22} \end{pmatrix},$$

and decompose T similarly. Then, the LR test is based on

$$\Lambda_{2\cdot 1} = \frac{|\mathsf{W}|/|\mathsf{W}_{11}|}{|\mathsf{T}|/|\mathsf{T}_{11}|} = \frac{|\mathsf{W}_{22\cdot 1}|}{|\mathsf{T}_{22\cdot 1}|}, \tag{3.5}$$

where $\mathsf{W}_{22\cdot 1} = \mathsf{W}_{21} - \mathsf{W}_{21}\mathsf{W}_{11}^{-1}\mathsf{W}_{21}$ and $\mathsf{T}_{22\cdot 1} = \mathsf{T}_{21} - \mathsf{T}_{21}\mathsf{T}_{11}^{-1}\mathsf{T}_{21}$. The null distribution is $\Lambda_{p-k}(q-1, n-q-k)$. For a proof of the result, see, for example, Fujikoshi et al. (2010, Theorem 3.3.2).

4. Asymptotic expansion and Rao's F approximation for Λ statistic

We consider the lambda distribution, defined as the distribution of

$$\Lambda = \frac{|\mathsf{W}|}{|\mathsf{W} + \mathsf{B}|} \sim \Lambda_p(q, n-q), \tag{4.1}$$

where B and W are independently distributed and follow the Wishart distributions $\mathrm{W}_p(q, \boldsymbol{\Sigma})$ and $\mathrm{W}_p(n-q, \boldsymbol{\Sigma})$, respectively. Such Λ appears, for example, as a likelihood ratio test for testing the equality of mean vectors $\boldsymbol{\mu}_i$, $i = 1, \ldots, q+1$, based on an N_i sample from $\mathrm{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. In this case, $N = N_1 + \cdots + N_{q+1}$ and $n = N - 1$. When we consider the distribution of Λ,

we may assume $\mathbf{\Sigma} = \mathbf{I}_p$. The likelihood ratio criterion is based on $\lambda = \Lambda^{n/2}$. The hth moment of Λ is given by

$$\mathrm{E}[\Lambda^h] = \prod_{j=1}^{p} \frac{\Gamma[\frac{1}{2}(n-q-j+1)+h]\Gamma[\frac{1}{2}(n-j+1)]}{\Gamma[\frac{1}{2}(n-q-j+1)]\Gamma[\frac{1}{2}(n-j+1)+h]}. \tag{4.2}$$

We consider an asymptotic expansion of the distribution of $-2\rho \log \Lambda^{n/2}$ with a Bartlett correction factor ρ under a large sample framework:

$$p, q; \text{ fixed}, \quad n \to \infty.$$

Here, ρ is chosen as $1 - (p+q+1)/(2n)$, and we set

$$m = n\rho = n - \frac{1}{2}(p+q+1). \tag{4.3}$$

Then, the characteristic function of $V = -m \log \Lambda$ is expressed as

$$\begin{aligned} C(t) &= \mathrm{E}\left[\Lambda^{-mit}\right] \\ &= \prod_{j=1}^{p} \frac{\Gamma[\frac{1}{2}m(1-2it) + \frac{1}{4}(p-q+1) - \frac{1}{2}(j-1)]}{\Gamma[\frac{1}{2}m + \frac{1}{4}(p-q+1) - \frac{1}{2}(j-1)]} \\ &\times \frac{\Gamma[\frac{1}{2}m + \frac{1}{4}(p+q+1) - \frac{1}{2}(j-1)]}{\Gamma[\frac{1}{2}m(1-2it) + \frac{1}{4}(p+q+1) - \frac{1}{2}(j-1)]}. \end{aligned} \tag{4.4}$$

We can derive an expansion for $C(t)$ by using the generalized version of Stirling's formula for the gamma function

$$\begin{aligned} \log \Gamma(z+h) = \log \sqrt{2\pi} + \left(z + h - \frac{1}{2}\right) \log z - z \\ - \sum_{r=1}^{m} (-1)^r \frac{B_{r+1}}{r(r+1)z^r} + R_{m+1}(z), \end{aligned}$$

where $R_{m+1}(z) = \mathrm{O}\big(z^{-(m+1)}\big)$ and $B_r(h)$ is the Bernoulli polynomial of degree r defined by

$$\frac{\tau \mathrm{e}^{h\tau}}{\mathrm{e}^{\tau} - 1} = \sum_{r=0}^{\infty} \frac{\tau^r}{r!} B_r(h).$$

The first three of these are $B_0(h) = 1,\ B_1(h) = h - \frac{1}{2},\ B_2(h) = h^2 - h + \frac{1}{6}$. The final result is given as follows:

$$C(t) = (1-2it)^{-f/2}\left[1 + \frac{\gamma_2}{m^2}\{(1-2it)^{-2} - 1\} + \frac{1}{m^4}\left\{\gamma_4((1-2it)^{-4} - 1) - \gamma_2^2((1-2it)^{-2} - 1)\right\}\right] + \mathrm{O}(m^{-5}), \quad (4.5)$$

where $f = pq$, $\gamma_2 = pq(p^2 + q^2 - 5)/48$, and

$$\gamma_4 = pq\{3p^4 + 3q^4 + 10p^2q^2 - 50(p^2 + q^2) + 159\}/1920.$$

Inverting the above characteristic function formally, we have an asymptotic expansion:

$$\mathrm{P}(-m\log\Lambda \le x) = G_f(x) + \frac{\gamma_2}{m^2}[G_{f+4}(x) - G_f(x)] + \frac{1}{m^4}[\gamma_4\{G_{f+8}(x) - G_f(x)\} - \gamma_2^2\{G_{f+4}(x) - G_f\}] + \mathrm{O}(m^{-5}), \quad (4.6)$$

where $G_f(x)$ is the distribution function of χ_f^2.

It may be noted that result (4.6) was first derived by Rao (1948), based on an expression due to Wald and Brookner (1941). On the other hand, Box (1949) obtained the result as a special case of a general asymptotic expansion of the distribution of a random variable whose moments belong to a class of Box-type moments.

Rao (1951) proposed a better F approximation of the distribution of another function of $\Lambda = \Lambda_p(q, n-q)$. The approximation is to consider

$$\frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \cdot \frac{ms + 2\lambda}{pq} \quad (4.7)$$

as an F approximation with pq and $ms + 2\lambda$ degrees of freedom, where

$$\lambda = -\frac{1}{4}pq + \frac{1}{2}, \quad s = \left(\frac{p^2q^2 - 4}{p^2 + q^2 - 5}\right)^{1/2}. \quad (4.8)$$

For $p = 1$ or 2 (or $q = 1$ or 2) the F-distribution is exactly as given. If $ms + 2\lambda$ is not an integer, interpolation between two integer values can be

used. The F approximation may be also written as a beta approximation $\beta(\frac{1}{2}(ms+2\lambda), \frac{1}{2}pq)$ for $Y = \Lambda^{1/s}$ which was obtained in Rao (1951) as follows. From (4.6) the density function of $V = -m \log \Lambda$ can be expressed as

$$f_V(v) = g_r(v) \left[1 + \frac{\gamma_2}{m^2} \left\{ \frac{v^2}{r(r+2)} - 1 \right\} + \mathrm{O}(m^{-3}) \right], \tag{4.9}$$

where $r = pq$ and $g_r(v)$ is the density function of χ^2_r, by using $g_{r+2}(v) = (v/r)g_r(v)$. Rao considered a better approximation for

$$Y = \Lambda^{1/s} = \mathrm{e}^{-V/(sm)}, \tag{4.10}$$

introducing a constant s. The density function of Y is expressed as

$$\begin{aligned} f_Y(y) &= f_V(sm(-\log y)) \frac{sm}{y} \\ &= \frac{1}{\Gamma(r/2)2^{r/2}} (ms)^{r/2} y^{(ms)/2+\lambda-1} y^{-\lambda} (-\log y)^{r/2-1} \\ &\times \left\{ 1 + \frac{\gamma_2 s^2}{r(r+2)} (-\log y)^2 + \mathrm{o}(m^{-2}) + \mathrm{o}((1-y)^2) \right\}, \end{aligned} \tag{4.11}$$

introducing a constant λ. Now we use

$$\begin{aligned} y^{-\lambda} &= \{1 - (1-y)\}^{-\lambda} \\ &= 1 + \lambda(1-y) + \frac{1}{2}\lambda(1+\lambda)(1-y)^2 + \cdots, \\ (-\log y)^{r/2-1} &= [-\log \{1 - (1-y)\}]^{r/2-1} \\ &= (1-y)^{r/2-1} \left[1 + \frac{1}{2} \left(\frac{1}{2}r - 1 \right) (1-y) \right. \\ &\left. + \left\{ \frac{1}{3} \left(\frac{1}{2}r - 1 \right) + \frac{1}{8} \left(\frac{1}{2}r - 1 \right) \left(\frac{1}{2}r - 2 \right) \right\} (1-y)^2 + \cdots \right]. \end{aligned}$$

Substituting the above expansions to the density of Y given by (4.11), we have

$$\begin{aligned} f_Y(y) &= \frac{1}{\Gamma(r/2)2^{r/2}} (ms)^{r/2} y^{(ms)/2+\lambda-1} (1-y)^{r/2-1} \\ &\times \left\{ 1 + a_1(1-y) + a_2(1-y)^2 + \mathrm{o}(m^{-2}) + \mathrm{o}((1-y)^2) \right\}, \end{aligned} \tag{4.12}$$

where

$$a_1 = \lambda + \frac{1}{2}\left(\frac{1}{2}r - 1\right),$$
$$a_2 = \frac{1}{2}\lambda(1+\lambda) + \left(\frac{1}{2}r - 1\right)\left\{\frac{1}{2}\lambda + \frac{1}{3} + \frac{1}{8}\left(\frac{1}{2}r - 2\right) + \frac{\gamma_2 s^2}{r(r+2)}\right\}.$$

Here, we note that the defining s and λ as the ones in (4.8) is equivalent to the defining a_1 and a_2 as zero. Further, using the generalized version of Stirling's formula for the gamma function, we can see that

$$\frac{\Gamma(\frac{1}{2}ms + \lambda + \frac{1}{2}r)}{\Gamma(\frac{1}{2}ms + \lambda)\Gamma(\frac{1}{2}r)} = \frac{1}{\Gamma(\frac{1}{2}r)2^{r/2}}(ms)^{r/2} + \mathrm{O}(m^{-1}).$$

This shows that the distribution of $Y = \Lambda^{1/s}$ has an expansion whose leading term is a Beta distribution $\beta(\frac{1}{2}(ms + 2\lambda), \frac{1}{2}pq)$ with a smaller error.

Now we note there have been developments related to asymptotic approximations of Λ. A computable error bound for large-sample approximations was derived based on an error bound in the L_1-norm for a multivariate scale mixture; see Fujikoshi and Ulyanov (2006). A high-dimensional approximation and its error bound have been studied under $p/n \to c \in (0, 1)$ by Fujikoshi et al. (2010) and Wakaki (2007). The distribution of Λ is called the nonnull distribution of Λ when $\mathbf{B}$ is distributed as a noncentral Whishart distribution $\mathrm{W}_p(q, \mathbf{\Sigma}; \mathbf{\Omega})$. An extension of (4.6) up-to the order m^{-2} to the nonnull case was given by Sugiura and Fujikoshi (1969). Kulp and Nagarsenker (1984) gave an asymptotic expansion of the nonnull distribution of $Y = \Lambda^{1/s}$.

5. Growth Curve Analysis

Research of growth curve analysis dates back to Wishart (1938), who compared the growth curves of animals under different treatments. In particular, the weight of each animal under each treatment was ascertained each week

for a number of weeks. For the original measurements of weekly weights $(y_1, \ldots, y_p)$, Wishart (1938) fitted orthogonal polynomials, for example,

$$a + b_1\phi_1(t) + b_2\phi_2(t)$$

to each growth curve dataset and replaced the original measurements of weekly weights $(y_1, \ldots, y_p)$ by (y_1, b_1, b_2). Then, a univariate analysis of variance on b_1 or b_2 was considered, using y_1 as a concomitant.

Rao (1959) proposed to analyze such growth curve data by considering a multivariate structure in addition to a growth curve structure. In general, suppose that a single variable Y is measured at p time points $t_1, \ldots, t_p$ (or different conditions) on n subjects, chosen at random from a group. One way to analyze such repeated measures is to specify a polynomial regression for Y on the time variable t, and to assume that the covariance matrix of $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ is unknown and positive definite.

Let the observations $Y_{i1}, \ldots, Y_{ip}$ of the ith subject be denoted by

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})', \quad i = 1, \ldots, n.$$

Then, in the growth curve model, it is assumed that for $i = 1, \ldots, n$,

$$\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\theta}, \tag{5.1}$$

and $\mathrm{Var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}$, where $\mathbf{X}$ is a given $p \times q$ matrix with rank q, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ is an unknown parameter vector, and $\boldsymbol{\Sigma}$ is unknown positive and definite. The matrix $\mathbf{X}$ is called a within-design matrix. In the growth curve model (5.1), Rao (1959) proposed and developed the following theory:

(1) Is the specification (5.1) adequate?

(2) How can estimators of $\theta_1, \ldots, \theta_q$ be obtained and the precision of the estimators be expressed?

(3) How can general linear hypotheses concerning $\theta_1, \ldots, \theta_q$ be tested?

(4) How can simultaneous confidence limits for a class of linear functions of $\theta_1, \ldots, \theta_q$ be obtained?

In addition, Rao (1987) proposed approaches to the following problem.

(5) Suppose that the measurements of growth at the time points $t_1, \ldots, t_p$, t_{p+1} are available for n individuals and only at $t_1, \ldots, t_p$ for an $(n+1)-$th individual. How do we predict the measurement at t_{p+1} for the $(n+1)-$th individual?

The growth curve model for above one-group data was extended by Potthoff and Roy (1964) as follows. Suppose that the rows of $\mathbf{Y}$ are independently distributed as p-dimensional normal distributions with a common covariance matrix $\mathbf{\Sigma}$, and

$$\mathrm{E}(\mathbf{Y}) = \mathbf{A\Theta X'}, \tag{5.2}$$

where $\mathbf{A}$ is the $n \times k$ between-group design matrix, $\mathbf{X}$ is the $p \times q$ within design matrix, and $\mathbf{\Theta}$ is the $k \times q$ unknown parameter matrix. A general testing problem is to test

$$H_g : \mathbf{C\Theta D} = \mathbf{O}, \text{against } K_g : \mathbf{C\Theta D} \neq \mathbf{O}. \tag{5.3}$$

Here, $\mathbf{C}$ is a given $c \times k$ matrix with rank c, and $\mathbf{D}$ is a given $q \times d$ matrix with rank d. The growth curve model (5.2) is reduced to a MANOVA model when the within-individual design matrix $\mathbf{X}$ is $\mathbf{I}_p$. In this sense, the growth curve model is a generalized MANOVA model.

In order to relate the growth curve model to a multivariate linear model, consider the transformation from $\mathbf{Y}$ to $(\mathbf{U}\ \mathbf{V})$:

$$(\mathbf{U}\ \mathbf{V}) = \mathbf{Y}(\mathbf{G}_1\ \mathbf{G}_2), \tag{5.4}$$

where $\mathbf{G}_1$ and $\mathbf{G}_2$ are the same matrices as in the group, $\mathbf{G}_1 = \mathbf{X}(\mathbf{X'X})^{-1}$, $\mathbf{G}_2 = \tilde{\mathbf{X}}$, and $\tilde{\mathbf{X}}$ is a $p \times (p-q)$ matrix satisfying $\tilde{\mathbf{X}}'\mathbf{X} = \mathbf{O}$ and $\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \mathbf{I}_{p-q}$. Then, the rows of $(\mathbf{U}\ \mathbf{V})$ are independently distributed as p-dimensional normal distributions with means

$$\mathrm{E}[(\mathbf{U}\ \mathbf{V})] = (\mathbf{A}\Theta\ \mathbf{O})$$

and the common covariance matrix

$$\mathbf{\Psi} = \mathbf{G}'\mathbf{\Sigma}\mathbf{G} = \begin{pmatrix} \mathbf{G}_1'\mathbf{\Sigma}\mathbf{G}_1 & \mathbf{G}_1'\mathbf{\Sigma}\mathbf{G}_2 \\ \mathbf{G}_2'\mathbf{\Sigma}\mathbf{G}_1 & \mathbf{G}_2'\mathbf{\Sigma}\mathbf{G}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{\Psi}_{11} & \mathbf{\Psi}_{12} \\ \mathbf{\Psi}_{21} & \mathbf{\Psi}_{22} \end{pmatrix},$$

where $\mathbf{G} = (\mathbf{G}_1 \ \mathbf{G}_2)$. This transformation can be regarded as one from $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ to a q-dimensional main variable $\boldsymbol{U} = (U_1, \ldots, U_q)'$ and a $(p-q)$-dimensional auxiliary variable $\boldsymbol{V} = (V_1, \ldots, V_{p-q})'$. The growth curve model is equivalent to the following two models:

(1) The conditional distribution of $\mathbf{U}$ given $\mathbf{V}$ is

$$\mathbf{U} \mid \mathbf{V} \sim N_{n\times q}(\mathbf{A}^*\mathbf{\Xi},\ \mathbf{\Psi}_{11\cdot 2}). \tag{5.5}$$

(2) The marginal distribution of $\mathbf{V}$ is

$$\mathbf{V} \sim \mathrm{N}_{n\times(p-q)}(\mathbf{O},\ \mathbf{\Psi}_{22}), \tag{5.6}$$

where

$$\mathbf{A}^* = (\mathbf{A}\ \mathbf{V}), \quad \mathbf{\Xi} = \begin{pmatrix} \mathbf{\Theta} \\ \mathbf{\Gamma} \end{pmatrix},$$
$$\mathbf{\Gamma} = \mathbf{\Psi}_{22}^{-1}\mathbf{\Psi}_{21},\ \mathbf{\Psi}_{11\cdot 2} = \mathbf{\Psi}_{11} - \mathbf{\Psi}_{12}\mathbf{\Psi}_{22}^{-1}\mathbf{\Psi}_{21}.$$

Rao (1965) also gave the above reduction, and called $\mathbf{V}$ the observation matrix of concomitant variables. Statistical methods based on likelihood were introduced by Rao (1959, 1965), Khatri (1966), Gleser and Olkin (1970), and others. The LR test was first given by Khatri (1966). Gleser and Olkin (1970) gave the LR test based on a canonical for the testing problem (5.3). The LR test is based on

$$\Lambda = |\mathbf{S}_e|/|\mathbf{S}_e + \mathbf{S}_h|,$$

where

$$\mathbf{S}_e = \mathbf{D}'(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{D}, \quad \mathbf{S}_h = (\mathbf{C}\hat{\Theta}\mathbf{D})(\mathbf{C}\mathbf{R}\mathbf{C}')^{-1}\mathbf{C}\hat{\Theta}\mathbf{D}$$

and

$$\mathbf{R} = (\mathbf{A}'\mathbf{A})^{-1} + (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}\mathbf{S}^{-1}\{\mathbf{S} - \mathbf{X}'(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}')^{-1}\mathbf{X}\}$$
$$\times\, \mathbf{S}^{-1}\mathbf{Y}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}.$$

Here, $\hat{\mathbf{\Theta}}$ and $\mathbf{S}$ are given by

$$\hat{\mathbf{\Theta}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}\mathbf{S}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1},$$
$$\mathbf{S} = \frac{1}{m}\mathbf{Y}'(\mathbf{I}_n - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{Y}, \quad m = n - k.$$

Further, the null distribution of Λ is $\Lambda_d(c, n-k-(p-q))$.

Rao (1965, 1966) and Grizzle and Allen (1969) discuss the possibility of using fewer than $p-q$ covariables. Fujikoshi and Rao (1991) proposed two types of formulation for the hypotheses of redundancy of a given set of covariables. The likelihood ratio criteria were obtained for testing these hypotheses. Further, using these results, they proposed information criteria such as for selection of the best subset of covariables.

In the growth curve models as in (5.1) and (5.2), it is necessary that the observations be observed at the same time points for each of the subjects, and that each of the groups have the same within-design matrix $\mathbf{X}$. In order to resolve the latter assumption, a general growth curve model was proposed by Rosen (1988), Verbyla and Venables (1988), etc., as follows:

$$\mathrm{E}(\mathbf{Y}) = \sum_{i=1}^{r} \mathbf{A}_i \mathbf{\Theta}_i \mathbf{X}_i', \tag{5.7}$$

which is called the sum-of-profiles model. On the other hand, in order to incorporate individual effects fully, the following random coefficients model or mixed effects model was considered:

$$\boldsymbol{Y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i, \quad i = 1, \ldots, n,$$
$$\boldsymbol{\beta}_i = \boldsymbol{\theta} + \boldsymbol{b}_i, \quad i = 1, \ldots, n,$$

where $\mathbf{X}_i$ is a $p_i \times k$ known matrix,

$$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n \sim \text{i.i.d. } \mathrm{N}_k(\mathbf{0}, \mathbf{\Delta}), \mathbf{\Delta} \geq \mathbf{O},$$
$$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n \text{ are independent}, \boldsymbol{e}_i \sim \mathrm{N}_{p_i}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}),$$
$$\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\} \text{ and } \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\} \text{ are independent.}$$

This model is a special case of mixed effects and random coefficients models (see Laird and Ware (1982), and Vonesh and carter (1987)). Rao (1965) considered the above model in the case $\mathbf{X}_1 = \cdots = \mathbf{X}_r$, and developed its statistical inference.

At the end of this section we consider some topics on discriminant analysis of growth curve data. Such problems were first discussed by Burnaby (1966). The paper pointed out a need for a general procedure of eliminating either a single growth factor or several nuisance factors from discriminant functions or generalized distances between a number of populations. Some results were given with the help of Rao's comments. Rao (1966) and, in his book, Rao (1973) treated this problem in a more general form which was called discrimination between composite hypotheses, as follows. Let $\boldsymbol{Y}$ be a p-variate random vector depending on a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let H_1 be the hypothesis that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ and H_2 that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_2$, where $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are two disjoint subsets of $\boldsymbol{\Theta}$. The problem involves choosing between H_1 and H_2 on the basis of an observed value of $\boldsymbol{Y}$. More concretely, let $\boldsymbol{Y}$ be a p-variate normal vector such that

$$\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{\theta}_i, H_i) = \boldsymbol{a}_i + \mathbf{X}\boldsymbol{\theta}_i, \quad \mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{\theta}_i, H_i) = \boldsymbol{\Sigma}, \quad i = 1, 2.$$

Here, $\mathbf{X}$ is a given $p \times k$ matrix of rank k. Let $\mathbf{Z}$ be a $p \times (p-k)$ matrix of rank $p-k$ such that $\mathbf{X}'\mathbf{Z} = \mathbf{O}$. Then,

$$\mathrm{E}(\mathbf{Z}'\boldsymbol{Y} \mid H_i) = \mathbf{Z}'\boldsymbol{a}_i, \quad \mathrm{Var}(\mathbf{Z}'\boldsymbol{Y} \mid H_i) = \mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}', \quad i = 1, 2.$$

From (2.3), the discriminant function based on $\mathbf{Z}'\boldsymbol{Y}$ is given by

$$(\mathbf{Z}'\boldsymbol{a}_1 - \mathbf{Z}'\boldsymbol{a}_2)'(\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z})^{-1}\mathbf{Z} = (\boldsymbol{a}_1 - \boldsymbol{a}_2)'\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{Y},$$

which is reduced as

$$(\boldsymbol{a}_1 - \boldsymbol{a}_2)'\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\}\boldsymbol{Y}. \tag{5.8}$$

Further, it was shown that

$$\sup_{\mathbf{X}'\boldsymbol{\ell}=\mathbf{0}} \frac{[\mathrm{E}(\boldsymbol{\ell}'\boldsymbol{Y} \mid H_1) - \mathrm{E}(\boldsymbol{\ell}'\boldsymbol{Y} \mid H_2)]^2}{2^{-1}[\mathrm{Var}(\boldsymbol{\ell}'\boldsymbol{Y} \mid H_1) + \mathrm{Var}(\boldsymbol{\ell}'\boldsymbol{Y} \mid H_2)]} \tag{5.9}$$

is attained at

$$\boldsymbol{\ell}_* = \{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1}\}(\boldsymbol{a}_1 - \boldsymbol{a}_2). \tag{5.10}$$

The result follows by using the fact that, under the condition $\mathbf{X}'\boldsymbol{\ell} = \mathbf{0}$, expression (5.9) is reduced to

$$\sup_{\mathbf{X}'\boldsymbol{\ell}=\mathbf{0}} \frac{[\mathrm{E}\{\boldsymbol{\ell}'(\boldsymbol{a}_1 - \boldsymbol{a}_2)\}^2]}{\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}}. \tag{5.11}$$

The discriminant function (5.8) is $\boldsymbol{\ell}_*'\boldsymbol{Y}$, where $\boldsymbol{\ell}_*$ is as defined in (5.10).

On the other hand, the usual discriminant method and its modifications have been studied for some growth curve models. For example, assume that $\boldsymbol{Y}$ is observed at two populations $\Pi_i, i = 1, 2$, and

$$\boldsymbol{Y} \mid \Pi_i \sim \mathrm{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),\ i = 1, 2,$$

where $\boldsymbol{\mu}^{(i)} = \mathbf{X}\boldsymbol{\theta}_i,\ i = 1, 2$. Further, let $\widehat{\boldsymbol{\theta}}_i, i = 1, 2$ and $\widehat{\boldsymbol{\Sigma}}$ be the MLEs of $\boldsymbol{\theta}_i, i = 1, 2$ and $\boldsymbol{\Sigma}$, based on n_i samples from $\Pi_i,\ i = 1, 2$. Then, there is a situation to decide which a new observation $\boldsymbol{Y}$ belongs to Π_1 and Π_2. A natural method is to discriminate $\boldsymbol{Y}$ based on the discriminant function $L(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}^{(1)}, \widehat{\boldsymbol{\mu}}^{(2)}, \hat{\boldsymbol{\Sigma}})$ in (2.3). As another example, Lee (1982) considered the classification of growth curves from a non-Bayesian and Baysian viewpoint, under the case where $\boldsymbol{\Sigma}$ is arbitrary positive definite and is of Rao's simple structure (Rao (1967)). In some cases, it will be necessary to evaluate the expected probabilities of misclassification. However, this subject has not been much well researched.

6. Information Criteria for Selection of Variables

Related to the selection of variables in multivariate analysis, Rao stated the following in the foreword of Multivariate Analysis IV (R. R. Krishnaiah, ed., 1977, North-Holland Publishing Company):

"While refinement of Fisherian methods continue to be made, relatively few new lines of investigations are started. New extensions of univariate methods to multiple measurement are being made, which are no doubt useful, but there has not been adequate discussion of the number or choice of variables. In spite of the enormous increase in the multivariate methods, they do not seem to be rich enough to meet all practical demands..."

As we have seen in Section 2, Rao proposed a U-statistic for testing a hypothesis that $\boldsymbol{Y}_2$ provides no additional information to a discriminant analysis in the presence of $\boldsymbol{Y}_1$, where $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2)'$ and $\boldsymbol{Y} : p_i \times 1$. However, if several other specifications are considered, we need to decide upon the best specification. One approach is to apply model selection criteria such as AIC, BIC and C_p. In order to represent these approaches, it is standard to formulate the notions of sufficiency or redundancy of a subset of variables such that its likelihood is obtained in a computable form. Before describing it in detail in the case of two-group discriminant analysis, here we note that such an approach has been extended for the selection of variables in various multivariate models. Corresponding results have been obtained for, for example, the selection of the response variables and the explanatory variables in multivariate linear models, the selection of the main variables and the covariables in growth curve models, the selection of variables in canonical correlation analyses, and the selection of dimensionality in principal component analyses.

In the following we state a more detailed two-group discriminant analysis, following Fujikoshi and Sakurai (2020). Suppose that $\boldsymbol{j}$ denotes a subset of $\boldsymbol{\omega} = \{1, \ldots, p\}$ containing p_j elements, and $\boldsymbol{Y}_j$ denotes the p_j vector consisting of the elements of $\boldsymbol{Y}$, indexed by the elements of $\boldsymbol{j}$. We use the notation $D_{\boldsymbol{j}}$ and $D_{\boldsymbol{\omega}}$ for D based on $\boldsymbol{Y}_{\boldsymbol{j}}$ and $\boldsymbol{Y}_{\boldsymbol{\omega}} (= \boldsymbol{Y})$, respectively. Let $M_{\boldsymbol{j}}$ be a variable selection model, defined by

$$M_{\boldsymbol{j}} : \ \beta_i \neq 0 \text{ if } i \in \boldsymbol{j}, \text{ and } \beta_i = 0 \text{ if } i \notin \boldsymbol{j}. \tag{6.1}$$

The model $M_{\boldsymbol{j}}$ is equivalent to $\Delta_{\boldsymbol{j}} = \Delta_{\boldsymbol{\omega}}$, i.e., the Mahalanobis distance based on $\boldsymbol{Y}_j$ is the same as the one based on the full set of variables, $\boldsymbol{Y}$. We identify the selection of $M_{\boldsymbol{j}}$ with the selection of $\boldsymbol{Y}_{\boldsymbol{j}}$. Let $\mathrm{AIC}_{\boldsymbol{j}}$ be the AIC

for $M_{\boldsymbol{j}}$. Then, it is known (see, e.g., Fujikoshi (1985)) that

$$\begin{aligned} \mathrm{A}_{\boldsymbol{j}} &= \mathrm{AIC}_{\boldsymbol{j}} - \mathrm{AIC}_{\boldsymbol{\omega}} \\ &= n \log \left\{ 1 + \frac{g^2 (D^2_{\boldsymbol{\omega}} - D^2_{\boldsymbol{j}})}{n - 2 + g^2 D^2_{\boldsymbol{j}}} \right\} - 2(p - p_j), \end{aligned} \tag{6.2}$$

where $g = \sqrt{(n_1 n_2)/n}$. Similarly, let BIC_j be the BIC for $M_{\boldsymbol{j}}$, and we have that $\mathrm{B}_{\boldsymbol{j}} = \mathrm{BIC}_{\boldsymbol{j}} - \mathrm{BIC}_{\boldsymbol{\omega}}$ is the one replaced 2 in $\mathrm{A}_{\boldsymbol{j}}$ by $\log n$.

The variable selection methods based on AIC and BIC are given as $\min_{\boldsymbol{j}} \mathrm{AIC}_{\boldsymbol{j}}$ and $\min_{\boldsymbol{j}} \mathrm{BIC}_{\boldsymbol{j}}$, respectively. Therefore, such criteria become computationally onerous when p is large. To circumvent this issue, we can use a test-based method (TM, see Fujikoshi and Sakurai (2020)) or KOO method (Zhao et al. (1986), Nishii et al. (1988), Bai et al. (2018)), drawing on the significance of each variable. A critical region for "$\beta_i = 0$" based on the likelihood ratio principle is expressed (see, e.g., Rao (1946, 1973)) as

$$\mathrm{T}_{d,i} = n \log \left\{ 1 + \frac{g^2 (D^2_{\boldsymbol{\omega}} - D^2_{(-i)})}{n - 2 + g^2 D^2_{(-i)}} \right\} - d > 0, \tag{6.3}$$

where $(-i), i = 1, \ldots, p$ is the subset of $\boldsymbol{\omega} = \{1, \ldots, p\}$ obtained by omitting the i from $\boldsymbol{\omega}$, and d is a positive constant that may depend on p and n. Note that

$$\mathrm{T}_{2,i} > 0 \iff \mathrm{AIC}_{(-i)} - \mathrm{AIC}_{\boldsymbol{\omega}} > 0.$$

A test-based method or KOO method is defined by selecting the set of suffixes or the set of variables given by

$$\mathrm{TM}_d = \{ i \in \boldsymbol{\omega} \mid \mathrm{T}_{d,i} > 0 \}, \tag{6.4}$$

or $\{ Y_i \in \{Y_1, \ldots, Y_p\} \mid \mathrm{T}_{d,i} > 0 \}$. The notation $\widehat{\boldsymbol{j}}_{\mathrm{TM}_d}$ is also used for TM_d.

In general, if d is large, a small number of variables are selected. On the other hand, if d is small, a large number of variables are selected. Ideally, we want to select only the true variables whose discriminant coefficients are not zero. Consistency properties of AIC, BIC, and TM_d have been studied under a large-sample framework ($n \to \infty$) and a high-dimensional framework

$(n/p \to c \in (0,1))$; see Fujikoshi (1985), Nishii et al.(1988), and Fujikoshi and Sakurai (2020). In general, we note that the conclusions of asymptotic consistencies of model selection criteria may be reversed. For example, in the selection of the explanatory variables in a multivariate regression model, it is known (Nishii et al. (1988)) that under a large-sample framework, BIC is consistent, but AIC is not consistent. On the other hand, it is known (Yanagihara et al. (2015), Bai et al. (2018)) that under a high-dimensional framework, AIC is consistent, but BIC is not consistent.

For high-dimensional data such that $p > n$, Lasso and other regularization methods have been extended. For such studies, see, e.g., Clemmensen et al. (2011), Wtten and Tibshirani (2011), and Hao and Dong (2015).

Acknowledgements

We would like to thank the two referees and Dr. K. P. Choi for their careful reading of our manuscript and many helpful comments which improved the presentation of this paper.

References

[1] Bai, Z., Fujikoshi, Y. and Hu, J. (2018). Strong consistency of the AIC, BIC, Cp and KOO methods in high-dimensional multivariate linear regression. *Hiroshima Statistical Research Group*, TR; 18-09.

[2] Bellman, R. E. (1957). *Dynamic Programing*. Princeton Univ. Press, Princeton.

[3] Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Ann. Math. Statist.*, **42**, 241–259.

[4] Burnaby, T. P. (1966). Growth invariant discriminant functions and generalized distance. *Biometrics*, **22**, 96–110.

[5] Clemmensen, L., Hastie, T., Witten, D. M. and Ersbell, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**, 406–413.

[6] Fujikoshi, Y. (1982). A test for additional information in canonical correlation analysis. *Ann. Inst. Statist. Math.*, **34**, 137–144.

[7] Fujikoshi, Y. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. In *Multivariate Analysis-VI* (P. R. Krishnaiah, ed.), 219–236, North-Holland.

[8] Fujikoshi, Y. (1989). Tests for redundancy of some variables in multivariate analysis. In *Recent Developments in Statistical Data Analysis and Inference* (Y. Dodge, ed.), 141–163, Elesevier Science Publishers B.V., Amsterdam.

[9] Fujikoshi, Y. (1992). Redundancy of some variables in multivariate analysis. *Japanese J. Behaviormetrics*, **19**, 18-28 (in Japanese).

[10] Fujikoshi, Y. and Rao, C. R. (1991). Selection of covariables in the growth curve model. *Biometrika*, **78**, 779–785.

[11] Fujikoshi, Y. and Ulyanov, V. V. (2006). On accuracy of asymptotic expansion for Wilks' lambda distribution. *J. Multivariate Anal.*, **97**, 1941–1957.

[12] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* Wiley, Hoboken, N.J.

[13] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-typ criteria in multivariate linear regression. *Journal of Multivariate Analysis*, **123**, 184–200.

[14] Fujikoshi, Y. and Sakurai, T. (2019). Consistency of test-based method for selection of variables in high-dimensional two group-

discriminant analysis. *Japanese Journal of Statistics and Data Science*, **2**, 155–171.

[15] Gleser, L. J. and Olkin, I. (1970). Linear models in multivariate analysis. In *Essays in Probability and Statistics*, (R.C. Bose et al., eds.), 267–292, University of North Carolina Press, Chapel Hill, NC.

[16] Gupta, A. K., Xu, J. and Fujikoshi, (2006). An asymptotic expansion of the distribution of Rao's U-statistic under a general condition. *J. Multivariate Anal.*, **97**, 492–513.

[17] Hao, N. and Bin, B. (2015). Sparcifying the Fisher linear discriminant by rotation. *Journal of the Royal Statistical Society: Series B*, **77**, 827–851.

[18] Healy, M. J. R. (1969). Rao's paradox concerning multivariate tests of significance. *Biometrics*, **25**, 411–413.

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

[20] Kshirsagar, A. M. (1972). *Multivariate Analysis*. Marcel Dekker, New York.

[21] Kulp, R. W. and Nagarsenker, B. N. (1984). An asymptotic expansion of the nonnull distribution of Wilks criterion for testing the multivariate linear hypothesis. *Ann. Statist.*, **12**, 1576–1583.

[22] Laird, N. M. and Ware, J. H. (1982). Random-effects model for longitudinal data. *Biometrics*, **38**, 963–974.

[23] Lee, J. C. (1982). Classification of growth curves. In *Handbook of Statistics-2*(P. R. Krishnaiah and K. L. Kanal, eds.), 121-132, North-Holland Publishing Company, Amsterdam.

[24] Nishii, R. , Bai, Z. D. and Krishnaia, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Mathematical Journal*, *18*, 451–462.

[25] Oda, R., Suzuki, Y., Yanagihara, H. and Fujikoshi, Y. (2020). A consistent variable selection method in high-dimensional canonical discriminant analysis. *J. Multivariate Anal.*, **175**, 1–13.

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

[27] Pynnönen, S. (1987). Selection of variables in nonlinear discriminant analysis by information criteria. In *Proceedings of the Second International Tampere Conference in Statistics* (T. Pukkila and S. Puntanen, eds.), 627–636, Univ. of Tampere, Tampare.

[28] Rao, C. R. (1946). Tests with discriminant functions in multivariate analysis. *Sankhyā*, **7**, 407-414.

[29] Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58-79.

[30] Rao, C. R. (1951). An asymptotic expansion of the distribution of Wilk's criterion. *Bulletin of the International Statistical Institute*, **33**, Part 2, 177–180.

[31] Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research.* John Wiley & Sons, New York.

[32] Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychchometrika*, **20**, 93–111.

[33] Rao, C. R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, **46**, 49–58.

[34] Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā A*, **26**, 329–358.

[35] Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447–458.

[36] Rao, C. R. (1966a). Covariance adjustment and related problems in multivariate analysis. In *Multivariate Analysis-I*(P. R. Krishnaiah, ed.), 87–103, Academic Press, Inc., New York.

[37] Rao, C. R. (1966b). Discriminant function between composite hypotheses and related problems. *Biometrika*, **53**, 315–321.

[38] Rao, C. R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proc. Fifth Berkerley Symp. Math. Statist. and Prob.*, **1**, 355–372.

[39] Rao, C. R. (1970). Inference on discriminant function coefficients. In *Essays in Prob. and Statist.* (R. C. Bose et al., eds.), 587–602.

[40] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*(2nd ed.). John Wiley & Sons, New York.

[41] Rao, C. R. (1977). Foreword in Multivariate Analysis IV (P. R. Krishnaia, ed.). North Holland.

[42] Rao, C. R. (1979). Separation theorems for singular values of matrices and their applications in multivariate analysis. *J. Multivariate Anal.*, **9**, 362–377.

[43] Rao, C. R. (1987). Prediction of future observations in growth curve type models. *J. Statist. Sci.*, **2**, 434–471.

[44] Rao, C. R. (1997). An alternative to correspondence analysis using Hellinger distance. *Proc. Int. Symp. on Contemporary Multivariate Analysis and Its Applications*, A11–A29, Hong Kong.

[45] Rencher, A. C. (2002). *Methods of Multivariate Analysis*, 2nd ed., John Wiley & Sons, New York.

[46] Verbyla, A. P. and Venables, W. N. (1988). An extension of the growth curve models. *Biometrika*, **75**, 129–138.

[47] von Rosen, D. (1991). Maximum likelihood estimators in multivariate linear normal models. *J. Multivariate Anal.*, **31**, 187–200.

[48] Vonesh, E. F. and Carter, R. L. (1987). Efficient inference for random-coefficient growth curve models with unbalanced data. *Biometrics*, **43**, 617–628.

[49] Sugiura, N. and Fujikoshi, Y. (1969). Asymptotic expansions of the non-null distributions of the likelihood ratio criteria for multivariate linear hypothesis and independence. *Ann. Math. Statist.*, **40**, 942–952.

[50] Wakaki, H. (2007). An error bound for high-dimensional Edgeworth expansion for Wilks' Lambda distribution. *Hiroshima Statistical Research Group*, TR; 07-03.

[51] Wald, A. and Brookner, R., J. (1941). On the distribution of Wilks' statistic for testing the independence of several groups of variates. *Ann. Math. Statist.*, **12**, 137–152.

[52] Wilks, S. S. (1932). Certain generalization in the analysis of variance. *Biometrika*, **27**, 471–494.

[53] Wishart, J. (1938). Growth rate determination in nutrition studies with the bacon pig, and their analysis. *Biometrika*, **30**, 16–28.

[54] Witten, D. W. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. Roy. Statist. Soc.: Series B*, **73**, 753–772.

[55] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electronic Journal of Statistics*, **9**, 869–897.

[56] Zhao, L. C. , Krishnaiah, P. R. and Bai, Z. D. (1986). On determination of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1-25.