MULTIVARIATE RANKING AND SELECTION WITHOUT REDUCTION TO A UNIVARIATE PROBLEM

Edward J. Dudewicz
Vidya S. Taneja

ABSTRACT

"Ranking and selection" procedures are statistical procedures appropriate for use in situations where the experimenter's goal is to "select the best" (selection) or to "rank competing alternatives" (ranking). These goals are often present in simulation studies, which are often performed in order to select that one of several procedures (for running a real-world system) which is "best." (For a discussion of ranking in simulation, see (2).)

Most previous work in this area has dealt with situations where either one has a univariate response, or where a simple univariate function of the multivariate response characterizes the "goodness" of a procedure. (For example, see (3) for an introduction and some procedures especially useful in univariate-response simulation settings, (8) for a comprehensive review of the area, and (4) for a discussion of selection in simulation and related statistical problems and procedures.) In a recent excellent expository book on the area (7), Gibbons, Olkin, and Sobel noted (Chapter 15, p. 390) that "The whole field [of multivariate-response ranking and selection] is as yet undeveloped and the reader is encouraged to regard this chapter as an introduction to a wide area that will see considerable development in the future as more meaningful models are formulated."

In this paper we outline a selection model recently developed for this multiple-response problem (6) and develop an example of its use and recommendations for its implementation.

I. INTRODUCTION

In many settings it is reasonable to assume that the response (output) from a simulation of a procedure has a multivariate normal distribution. (For example, this is usually done in optimization, as in (1).) We will assume that when alternative i (denoted π_i) is simulated the response (output of a run) follows a multivariate normal distribution with $p(\geq 1)$ component variates, mean vector $\underset{\sim}{\mu}_i$, and variance-covariance matrix Σ_i (where $i = 1,\ldots,k$ if k alternatives are being comparatively evaluated). This is usually abbreviated by saying π_i is $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ $(i = 1,\ldots,k)$. We desire to study the problem of selecting the "best" of $\pi_1, \pi_2,\ldots,\pi_k$.

In contrast to the optimization approach (see (1)), we do not assume the p responses are independent, and we specifically acknowledge that they are random (whereas in the optimization approach one usually disregards the randomness) and seek to run our simulation so as to be (e.g.) 90% sure that we do in fact select the best alternative.

II. MODEL

Let π_i be $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ for $i = 1,\ldots,k$, and assume $p \geq 1$. Here the k $p{\times}p$ variance-covariance matrices $\Sigma_1,\ldots,\Sigma_k$ are assumed unknown, and need not be equal. Let $g(\underset{\sim}{\mu}_1,\ldots,\underset{\sim}{\mu}_k)$ be an experimenter-specified function with possible values $1,2,\ldots,k$ and such that

$$g(\underset{\sim}{\mu}_1,\ldots,\underset{\sim}{\mu}_k) = j \tag{II.1}$$

if and only if, given a choice among $\underset{\sim}{\mu}_1,\ldots,\underset{\sim}{\mu}_k$, the experimenter would prefer $\underset{\sim}{\mu}_j$. Let $\underset{\approx}{\mu} = (\underset{\sim}{\mu}_1,\ldots,\underset{\sim}{\mu}_k)$ denote the set of true mean vectors and

$$P_j = \{\underset{\approx}{\mu}:\ g(\underset{\approx}{\mu}) = j\},\ j = 1,\ldots,k, \tag{II.2}$$

and note that $P_1,\ldots,P_k$ are disjoint preference sets whose union is kp dimensional Euclidean space, $\mathbb{R}^{kp}$. Define the distance between any two points $\underset{\approx}{a}$ and $\underset{\approx}{b}$ of $\mathbb{R}^{kp}$ as the usual Euclidean distance

$$d(\underset{\approx}{a}, \underset{\approx}{b}) = \left(\sum_1^{kp} (a_i - b_i)^2\right)^{1/2}, \tag{II.3}$$

and denote the distance from $\underset{\approx}{\mu}$ to the boundary of $P_{g(\underset{\approx}{\mu})}$ by

$$d_B(\underset{\approx}{\mu}) = \inf\{d(\underset{\approx}{\mu}, \underset{\approx}{b}):\ \underset{\approx}{b} \notin P_{g(\underset{\approx}{\mu})}\}. \tag{II.4}$$

We now set our <u>probability requirement</u> as

$$P(CS|\mathcal{P}) \geq P^* \text{ whenever } d_B(\underset{\sim}{\mu}) \geq d^*. \qquad \text{(II.5)}$$

That is, we desire a selection procedure $\mathcal{P}$ which has probability at least P^* of choosing the true best population (event "CS") whenever the mean vector $\underset{\sim}{\mu}$ is at least Euclidean distance d^* from mean vectors where other populations are best.

Consider the following procedure $\mathcal{P}_{HM}$ specified by its sampling rule (telling how many observations, or simulation runs, are needed on each alternative) and terminal decision rule (telling which alternative to select once all the runs have been made).

Sampling Rule for $\mathcal{P}_{HM}$. Select $z > 0$, an integer $n_0 > p$, and a $p \times p$ positive-definite matrix (α_{rs}). Take observations from each and every population π_c $(c = 1,\dots,k)$ as follows. Take n_0 initial observations $X_{c1},\dots,X_{cn_0}$ where $X_{ci} = (X_{c1i}, X_{c2i},\dots,X_{cpi})'$ $(i = 1,2,\dots,n_0)$ and compute

$$\bar{X}_{ci} = \frac{1}{n_0} \sum_{\ell=1}^{n_0} X_{ci\ell},$$

$$S_{cij} = \sum_{\ell=1}^{n_0} (X_{ci\ell} - \bar{X}_{ci})(X_{cj\ell} - \bar{X}_{cj}), \qquad \text{(II.6)}$$

$$s_{cij} = \frac{1}{(n_0-1)} S_{cij}, \quad i, j = 1,2,\dots,p .$$

Define the positive integer N_c by

$$N_c = \max\{n_0 + p^2, \; [z^{-1} \sum_{i,j=1}^{p} \alpha_{ij}\, s_{cij}] + 1\}, \qquad \text{(II.7)}$$

where $[q]$ denotes the largest integer less than q, and select p $(p \times N_c)$ matrices

$$A_{cr} = \begin{bmatrix} a_{cr_{11}} & \cdots & a_{cr_{1N_c}} \\ \vdots & & \\ a_{cr_{p1}} & \cdots & a_{cr_{pN_c}} \end{bmatrix} \qquad (r = 1,2,\dots,p)$$

in such a way that:

1) $a_{cr_{i1}} = \dots = a_{cr_{in_0}}$;

2) $A_{cr}\,\eta_c = \epsilon_r$ where η_c is the $N_c \times 1$ vector $(1,1,\dots,1)'$ and ϵ_r is the $p \times 1$ vector whose r^{th} element is 1 and all other elements are zero;

and

3) $A_c A_c' = z(\alpha^{rs}) \otimes (s_c^{ij})$, where $A_c' = (A_{c1}', A_{c2}',\dots,A_{cp}')$, $\otimes$ denotes the direct product, and (b^{ij}) denotes the inverse of the matrix (b_{ij}), $r,i = 1,2,\dots,p$.

Next take $N_c - n_0$ additional observations $X_{c,n_0+1},\dots,X_{cN_c}$ and compute

$$\tilde{\bar{X}}_{cr} = \sum_{i=1}^{p} \sum_{\ell=1}^{N_c} a_{cr_{i\ell}} X_{ci\ell} \quad (r=1,2,\dots,p). \qquad \text{(II.8)}$$

For π_c construct the p-dimensional vector $\tilde{\bar{X}}_c = (\tilde{\bar{X}}_{c1},\dots,\tilde{\bar{X}}_{cp})$, $c=1,2,\dots,k$.

Terminal Decision Rule for $\mathcal{P}_{HM}$. Select

$$\pi_{g(\tilde{\bar{X}}_1,\dots,\tilde{\bar{X}}_k)} . \qquad \text{(II.9)}$$

It is shown in (6) that selection procedure $\mathcal{P}_{HM}$ satisfies probability requirement (II.5), i.e. if we use $\mathcal{P}_{HM}$ we have probability at least P^* of choosing the best alternative whenever $d_B(\underset{\sim}{\mu}) \geq d^*$.

The constant $z > 0$ in $\mathcal{P}_{HM}$ is to be chosen so that

$$P[\sum_{i=1}^{k} (\tilde{\bar{X}}_i - \mu_i)'(\tilde{\bar{X}}_i - \mu_i) \leq (d^*)^2] = P^* . \qquad \text{(II.10)}$$

While the distribution of $(\tilde{\bar{X}}_1,\dots,\tilde{\bar{X}}_k)$ is independent of $(\mu_1,\dots,\mu_k)$, it is very complicated (see equations (2.12), (2.13) of (5)), hence calculation of $z > 0$ which satisfies (II.10) is not a simple matter. However, for large n_0 (a design constant under the experimenter's control) we may approximate the solution since (see (6)) as $n_0 \to \infty$ the $z > 0$ which solves (II.10) approaches the solution of (II.10) when $(\tilde{\bar{X}}_1,\dots,\tilde{\bar{X}}_k)$ is replaced by $(Y_1,\dots,Y_k)$ where $Y_1,\dots,Y_k$ are independent random variables and Y_i is $N_p(\mu_i, zp(\alpha^{rs}))$. This solution may be calculated from equations (4.3) and and (4.5) in (6) or, more simply, by Monte Carlo. (The authors are currently calculating tables of z.)

III. EXAMPLE

In practice it may sometimes be reasonable to specify the function $g(\mu_1,\dots,\mu_k)$ of equation (II.1) as follows. Let $V_1,\dots,V_k$ be

$$V_j = V_j(\mu_1,\dots,\mu_k) = \sum_{i=1}^{p} c_i \left(\sum_{\ell=1}^{k} H_i(\mu_{ji}, \mu_{\ell i})\right) \tag{III.1}$$

where $c_1 + \dots + c_p = 1$ (the c_i's are weights depending on the relative importance of the p factors in the multivariate output or response) and $H_i(\mu_{ji}, \mu_{\ell i})$ is a function expressing the goodness of the $i^{\underline{th}}$ component of μ_j relative to the $i^{\underline{th}}$ component of μ_ℓ. (The functions $H_i(\cdot,\cdot)$, $i = 1,2,\dots,p$, will in general not be linear. One may often expect to have H_i be monotone increasing in $\mu_{ji} - \mu_{\ell i}$ if large means of the $i^{\underline{th}}$ component are desirable.) We then set

$$g(\mu_1,\dots,\mu_k) = j \text{ iff } V_j > \max(V_1,\dots,V_{j-1}, V_{j+1},\dots,V_k) \tag{III.2}$$

$(j = 1,\dots,k)$. Thus V_j can be interpreted as the "value" of population π_j relative to $\pi_1,\dots,\pi_{j-1}, \pi_{j+1},\dots,\pi_k$ and we desire the alternative with the largest relative value.

Suppose, for example, that $p = 3$ and we take

$$H_i(\mu_{ji}, \mu_{\ell i}) = \mu_{ji} + e^{\mu_{ji} - \mu_{\ell i}} \tag{III.3}$$

for all $i = 1,2,3$. If $k = 3$ (we have three alternatives to consider), one possible set of μ_1, μ_2, μ_3 is (as an example)

$$\mu_1 = \begin{pmatrix}\mu_{11}\\ \mu_{12}\\ \mu_{13}\end{pmatrix} = \begin{pmatrix}2.0\\ 2.0\\ 1.0\end{pmatrix}, \mu_2 = \begin{pmatrix}\mu_{21}\\ \mu_{22}\\ \mu_{23}\end{pmatrix} = \begin{pmatrix}1.5\\ 1.5\\ 3.0\end{pmatrix}, \quad \mu_3 = \begin{pmatrix}\mu_{31}\\ \mu_{32}\\ \mu_{33}\end{pmatrix} = \begin{pmatrix}2.0\\ 1.5\\ 2.0\end{pmatrix}. \tag{III.4}$$

For convenience of discussion, assume we are considering three $(k = 3)$ alternative transportation systems and each has $p = 3$ attributes (e.g., convenience, pollution, and cost). If equation (III.3) applies, then

$$\begin{aligned} V_1 &= c_1(H_1(2.0,2.0) + H_1(2.0, 1.5) + H_1(2.0,2.0)) \\ &\quad + c_2(H_2(2.0,2.0) + H_2(2.0, 1.5) + H_2(2.0,1.5)) \\ &\quad + c_3(H_3(1.0,1.0) + H_3(1.0, 3.0) + H_3(1.0,2.0)) \end{aligned} \tag{III.5}$$

$$= 9.6487c_1 + 10.2974c_2 + 4.5032c_3,$$

$$\begin{aligned} V_2 &= c_1(H_1(1.5,2.0) + H_1(1.5, 1.5) + H_1(1.5,2.0)) \\ &\quad + c_2(H_2(1.5,2.0) + H_2(1.5, 1.5) + H_1(1.5,1.5)) \\ &\quad + c_3(H_3(3.0,1.0) + H_3(3.0, 3.0) + H_3(3.0,2.0)) \end{aligned} \tag{III.6}$$

$$= 6.7131c_1 + 7.1065c_2 + 20.1073c_3,$$

$$\begin{aligned} V_3 &= c_1(H_1(2.0,2.0) + H_1(2.0, 1.5) + H_1(2.0,2.0)) \\ &\quad + c_2(H_2(1.5,2.0) + H_2(1.5, 1.5) + H_2(1.5,1.5)) \\ &\quad + c_3(H_3(2.0,1.0) + H_3(2.0, 3.0) + H_3(2.0,2.0)) \end{aligned} \tag{III.7}$$

$$= 9.6487c_1 + 7.1065c_2 + 10.0862c_3,$$

and (if the three components are weighted equally, i.e. $c_1 = c_2 = c_3 = 1/3$) then

$$V_1 = 8.1498, \quad V_2 = 11.3090, \quad V_3 = 8.9471 \tag{III.8}$$

and we find that of the mean vectors (III.4) alternative two is preferred:

$$g(\mu_1, \mu_2, \mu_3) = 2. \tag{III.9}$$

(With other weightings (c_1, c_2, c_3), other alternatives will be preferred. For example, if convenience is most important and cost least we may take $c_1 = .6$, $c_2 = .3$, $c_3 = .1$. Then $V_1 = 9.3288$, $V_2 = 8.1705$, $V_3 = 8.9298$ and alternative 1 will be preferred.) Note that if only alternatives 1 and 2 were present, we would have

$$\begin{cases} V_1 = 6.6487c_1 + 6.6487c_2 + 3.1353c_3 \\ V_2 = 4.6065c_1 + 4.6065c_2 + 14.3891c_3 \end{cases} \tag{III.10}$$

and could (for some c_1, c_2, c_3) prefer alternative 1 while (with the same c_1, c_2, c_3) we would prefer alt. 2 of 1,2,3. (For example, if $c_1 = c_2 = .418$ while $c_3 = .164$, then in the case of three alternatives $V_1 = 9.0760$, $V_2 = 9.0742$, $V_3 = 8.6578$ and alternative 1 is preferred, while if only alternatives 1 and 2 are present then $V_1 = 6.0725$, $V_2 = 6.2108$ and alternative 2 is preferred.) This non-transitivity is, as discussed in (6), both desirable and present in many situations of true multivariate nature, even though it has been disparaged in most of the literature to date (just as, at one time, non-normality was believed widely to be ab-normality). In cases where it is difficult to specify g(.) in advance, a method of judging of the results by experts (generals, managers, etc.) as in (9) may be used.

BIBLIOGRAPHY

1. Biles, William E. and Swain, James J. "Strategies for Optimization of Multiple-Response Simulation Models," in Proceedings of the 1977 Winter Simulation Conference, edited by H. J. Highland, R. G. Sargent, and J. W. Schmidt, 1977, pp. 135-142.

2. Bishop, Thomas A. "Designing Simulation Experiments to Completely Rank Alternatives," in Proceedings of the 1978 Winter Simulation Conference, 1978.

3. Dudewicz, Edward J. "Statistics in Simulation: How to Design for Selecting the Best Alternative," in Proceedings of the 1976 Bicentennial Winter Simulation Conference, edited by H. J. Highland, T. J. Schriber, and R. G. Sargent, 1976, pp. 67-71.

4. Dudewicz, Edward J. "New Procedures for Selection Among (Simulated) Alternatives," in Proceedings of the 1977 Winter Simulation Conference, edited by H. J. Highland, R. G. Sargent, and J. W. Schmidt, 1977, pp. 59-62.

5. Dudewicz, Edward J. and Bishop, Thomas A. "The Heteroscedastic Method," Technical Report No. 153, Department of Statistics, The Ohio State University, Columbus, Ohio 43210, December 1977. To appear in Optimizing Methods in Statistics - II to be published by Academic Press, New York.

6. Dudewicz, Edward J. and Taneja, Vidya S. "A Multivariate Solution of the Multivariate Ranking and Selection Problem," Technical Report No. 167, Department of Statistics, The Ohio State University, Columbus, Ohio 43210, August 1978.

7. Gibbons, Jean Dickinson, Olkin, Ingram, and Sobel, Milton Selecting and Ordering Populations: A New Statistical Methodology. John Wiley & Sons, Inc., New York, 1977.

8. Kleijnen, Jack P. C. Statistical Techniques in Simulation, Part II. Marcel Dekker, Inc., New York, 1975.

9. Lee, Young Jack and Dudewicz, Edward J. "How to Select the Best Contender," in Annual Technical Conference Transactions of the American Society for Quality Control, Vol. 32 (1978), pp. 546-552.