SELECTION OF NONREPEATABLE OBSERVATIONS FOR ESTIMATION

G. ELFVING
UNIVERSITY OF HELSINGFORS AND COLUMBIA UNIVERSITY

1. Introduction and summary

This paper deals with the following particular type of design problem. Let there be given a set of possible observations, of the form

$$x_i = u_{i1}a_1 + \cdots + u_{ik}a_k + \xi_i = u_i'a + \xi_i\,, \qquad i = 1, \cdots, N\,, \tag{1.1}$$

where the coefficient vectors u_i are known, the parameter vector a is unknown, and the error terms ξ_i are uncorrelated random variables with mean 0 and variance 1. The last requirement can obviously be met by a change of scale if the original error variances are known. Let the aim of the investigator be to estimate a particular parametric form $\theta = c'a$. If it is required to do this on the basis of a subset comprising, say, $n < N$ observations, and if N and n are too large to permit trying out all possible combinations, one has to find some feasible selection procedure leading to a least-squares estimator $\hat{\theta}$ with as small variance as possible.

A practical situation where this problem is encountered is the following one, arising in psychology. Let the x_i's be the scores associated with various possible test items, and assume that a factor analysis has been performed, yielding a more or less approximate representation of the scores in terms of certain common factors $a_1, \cdots, a_k$ and mutually uncorrelated specific factors ξ_i. If the scores are normalized to specific variance one, and if the common factors are considered as parameters characteristic of the individual, we are concerned with the model (1.1). Further, let z be a "criterion score" measuring, for example, some ability of particular interest, for which, by the same factor analysis, a representation $z = c'a + \zeta$ has been found. For practical reasons, a planned routine prediction of z often has to be based on a moderate size subset of the original large set of items. The question then arises how to select this subset.

Our problem is closely connected with the allocation problem which, in its simplest form, can be stated as follows. Given a set (1.1) of possible observations, each of which can be independently *repeated* as many times as we please, which of them should we select for estimating $\theta = c'a$, and how many times should we repeat the selected ones when a fixed total n of actual observations is allowed? This problem may be considered as a special case of the previous one, namely, the case that all different coefficient vectors u_i occur in the given set with at least multiplicity n. This is approximately the situation when the "item points" u_i, in k-space, appear in clusters. In such a case it is possible to make use of certain geometric allocation methods developed by the writer [1], [2].

In the present paper, we shall be concerned with the opposite situation where the u_i's are more or less smoothly distributed, so as to permit an idealized description by means

This work was sponsored by the School of Aviation Medicine, Randolph Field, Texas, under Contract AF 18(600)-941.

of a density function. With this idealization, we shall show (i) that the question of selection reduces to a variational problem, the varied element being a region S in k-space; (ii) that this problem leads to the conditions (3.2) which essentially constitute a system of $k(k+1)/2+1$ equations with equally many unknowns; (iii) that, if the above-mentioned density function is spherically symmetric, or can be reduced to such symmetry by a linear transformation of the argument, the equations (3.2) admit a unique solution, obtained by taking for S a twin half-space $\{u: |g'u| > \kappa\}$.

It seems very likely that these results will prove modifiable so as to cover also the original discrete problem. We hope to be able to return to this question in another paper.

2. The variational problem

Let ω denote any subset of the set $(1, \cdots, N)$ of integers. Every ω determines a set of observations (1.1), and a corresponding system of normal equations in $\alpha_1, \cdots, \alpha_k$, with "information" matrix $M = \sum_\omega u_i u_i'$. The least-squares estimator $\hat{\theta}$ for $\theta = c'\alpha$ has variance $D^2(\hat{\theta}) = c'M^{-1}c$. Thus, in the discrete case, our problem is a restricted minimum problem in ω,

$$c'\left(\sum_\omega u_i u_i'\right)^{-1} c = \min_\omega; \qquad \sum_\omega 1 = n. \tag{2.1}$$

Before proceeding to the idealization referred to in the introduction, it is useful to introduce what might be called a symmetry convention. The contribution of any observation (1.1) to the information matrix M depends solely on the vector ("item point") u_i in k-space. Items with opposite u_i's obviously yield the same contribution. For reasons of symmetry we shall henceforth describe each item by means of the *pair* of opposite points $\pm u_i$. Denoting by S the centrally symmetric set of all points u_i and $-u_i$ corresponding to a particular selection ω, we note that the sums in (2.1) may be written as sums over S, divided by 2. Introducing for convenience a constant factor $1/N$, which obviously does not affect the minimization, we may rewrite the problem (2.1) as

$$c'\left(\frac{1}{2N}\sum_S uu'\right)^{-1} c = \min_S; \qquad \frac{1}{2N}\sum_S 1 = \epsilon, \tag{2.2}$$

where $\epsilon = n/N$ is the selection rate.

Now, let $f(u)$ denote a centrally symmetric probability density function in k-space, with finite second-order moments, and assume that $f(u)$ provides a reasonably accurate description of the distribution of the item points $\pm u_i$; that is, the number of such points in a region A is approximately $2N\int_A f(u)du$ where du denotes the volume element. With this idealization, our problem (2.2) turns into the variational problem

$$c'M^{-1}c = \min_S; \tag{2.3}$$

$$\int_S f(u)\, du = \epsilon, \tag{2.4}$$

where S is the "selection region" sought for, and M denotes the $k \times k$ information matrix

$$M = \int_S uu' f(u)\, du, \tag{2.5}$$

with elements

$$\mu_{ij} = \int_S u_i u_j f(u)\, du. \tag{2.6}$$

We shall in the sequel confine ourselves to this continuous problem, returning to the original discrete one only to establish the practical procedure presented at the end of the paper.

3. Necessary conditions for extremal regions

By an argument somewhat similar to that of the Neyman-Pearson lemma, we are led to

THEOREM 3.1. *Any extremal region of the problem presented in equations* (2.3), (2.4), *and* (2.5) *is, aside from a set of f-measure* 0, *a twin half-space of form*

$$S = \{u: |\, c'M^{-1}u\,| > h\} \tag{3.1}$$

where the matrix M in turn depends on S according to (2.5).

We note that (2.5), (2.4), and (3.1) constitute a system

$$M = \int_S uu' f(u)\, du; \qquad \int_S f(u)\, du = \epsilon; \qquad S = \{u: |\, c'M^{-1}u\,| > h\} \tag{3.2}$$

of equations in M, h, S; upon insertion of S from the last equation, the two first ones make $k(k+1)/2 + 1$ equations in the equally many unknowns $\mu_{11}, \mu_{12}, \cdots, \mu_{kk}, h$.

PROOF. Let S be an extremal set of equations (2.3), (2.4), and (2.5), and S^* its complement. It is no restriction to assume that any neighborhood of a point $u \in S$ contains a subset of S with positive f-measure, and similarly for S^*. Take $u \in S$ and $u^* \in S^*$ such that $f(u) \neq 0$, $f(u^*) \neq 0$, but otherwise arbitrary. At each of these points, take a differential set, $du \in S$ and $du^* \in S^*$, respectively, of nonvanishing Euclidean measure (we use, for simplicity, the same notation for the sets and their measures), and such that

$$f(u)\, du = f(u^*)\, du^*. \tag{3.3}$$

The variation $S - du + du^*$ then obviously is admissible with respect to the size condition (2.4). Since S yields a minimum of $c'M^{-1}c$, with M depending on S according to (2.5), the corresponding variation of the first-mentioned quantity must be nonnegative. Using the differential formula $dM^{-1} = -M^{-1} \cdot dM \cdot M^{-1}$ and noting that the differential of (2.5) is the integration element, we find the effect on $c'M^{-1}c$ of subtracting du,

$$\begin{aligned} -\delta(c'M^{-1}c) &= c'M^{-1} \cdot dM \cdot M^{-1}c \\ &= c'M^{-1} \cdot uu' f(u)\, du \cdot M^{-1}c \\ &= (c'M^{-1}u)^2 f(u)\, du\,, \end{aligned} \tag{3.4}$$

and a similar expression for the effect of adding du^*. By the minimum property of S, we thus have

$$(c'M^{-1}u)^2 f(u)\, du \geqq (c'M^{-1}u^*)^2 f(u^*)\, du^*. \tag{3.5}$$

Dividing (3.5) by (3.3), we realize that $|c'M^{-1}u| \geqq |c'M^{-1}u^*|$ as soon as $u \in S$ ∩

($f > 0$) and $u^* \in S^* \cap (f > 0)$. It follows that there exists a constant $h > 0$ (in case $f = 0$ in some parts of u-space, h might not be uniquely determined) such that

$$(3.6) \qquad |\, c'M^{-1}u \,| \begin{cases} \geqq h & \text{in} \quad S \cap (f > 0) \\ \leqq h & \text{in} \quad S^* \cap (f > 0)\,. \end{cases}$$

This completes the proof of theorem 3.1.

We do not as yet know whether, in the general case, the system (3.2) possesses a solution, or whether this, if existing, is unique and constitutes a solution of the variational problem presented in equations (2.3), (2.4) and (2.5). It seems likely that some iterative procedure might be designed for solving the system mentioned. We shall here only discuss a special case in which an explicit solution is easily obtainable. The resulting procedure will probably be useful as an approximate solution also in more general cases.

4. A transformation lemma

Before proceeding to the special case referred to above, we shall prove the following simple

LEMMA 4.1. *Let $v = Lu$ be a linear transformation of k-space onto itself, with $|L| = 1$. If M, S, c satisfy* (3.2), *then*

$$(4.1) \qquad \bar{M} = LML', \qquad \bar{S} = LS\,, \qquad \bar{c} = Lc$$

satisfy the same system, with u replaced by v and $f(u)$ by $\bar{f}(v) = f(L^{-1}v)$; and conversely.

PROOF. Applying the transformation $v = Lu$ to the second integral in (3.2), we find that the second part of (3.2) remains true when S and f are replaced by $\bar{S}$ and $\bar{f}$, respectively. Applying the same transformation to the first part of (3.2) we find

$$(4.2) \qquad M = L^{-1} \cdot \int_{\bar{S}} v\, v'\, \bar{f}\,(v)\; dv \cdot L'^{-1};$$

it follows that the matrix $\bar{M} = LML'$, together with $\bar{S}$ and $\bar{f}$, satisfies the first part of (3.2). As to the third equation of (3.2), we have

$$(4.3) \qquad \bar{S} = LS = \{\, v: |\, c'M^{-1}L^{-1}v \,| > h \,\};$$

replacing c by $L^{-1}\bar{c}$ and M by $L^{-1}\bar{M}L'^{-1}$, we find that $\bar{S}$, $\bar{M}$, $\bar{c}$ satisfy the third part of (3.2). The converse is proved in the same way, replacing L by L^{-1}.

5. The spherically symmetric case

When the density function $f(u)$ is spherically symmetric with respect to the origin, it is intuitively almost obvious that the twin half-space of theorem 3.1 has to be symmetric with respect to the "relevant direction" determined by the vector c. This is the content of the following proposition.

THEOREM 5.1. *If $f(u)$ is constant on every sphere $u'u = C$, then the system* (3.2) *has a unique solution, determined by the region*

$$(5.1) \qquad S = \{\, u: |\, c'u \,| > \kappa \,\}$$

where κ has to be chosen so as to satisfy the second part of (3.2).

PROOF. (i) We shall first prove that the solution (5.1) is sufficient in the standardized case $c = \gamma e$ where e is the first coordinate vector $(1, 0, \cdots, 0)'$ and γ any positive constant.

Take S according to (5.1), determine κ so as to satisfy the second part of (3.2), and M from the first part of (3.2). It remains to show that, for an appropriate h, the third part of (3.2) is satisfied. For this purpose, we first note that in the present case, M is of form

$$M = \begin{bmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & m \end{bmatrix}; \tag{5.2}$$

this follows from (2.6), noting that (a) S is now of form $|u_1| > \text{constant}$, (b) $f(u)$ is an even function of each variable u_j separately, and hence, (c) the integral from $-\infty$ to $+\infty$ of $u_j f(u) du_j$, $j = 2, \cdots, k$, vanishes. It follows from (5.2) that the coefficient row vector appearing in the third part of (3.2) may be written as

$$c'M^{-1} = \gamma\, e'M^{-1} = \frac{\gamma\, e'}{m_1} = \frac{c'}{m_1}. \tag{5.3}$$

Hence, the region S may be written

$$S = \left\{u: |\, c'M^{-1}u\,| > \frac{\kappa}{m_1}\right\} \tag{5.4}$$

and the third part of (3.2) is satisfied if we choose $h = \kappa/m_1$.

(ii) Next, we shall prove that the solution (5.1) is sufficient for a general c. For this purpose, take an *orthogonal* transformation $v = Lu$ mapping c into a vector along the first coordinate axis, that is, such that $c = \gamma L'e$, $\gamma > 0$. From (i) we know that the entities

$$\bar{c} = \gamma\, e\,, \qquad \bar{S} = \left\{ v: |\, e'v\,| > \frac{\kappa}{\gamma}\right\}, \qquad \bar{M} = \int_{\bar{S}} v\, v'\, \bar{f}\,(v)\, dv\,, \tag{5.5}$$

with appropriate choice of κ, satisfy (3.2). Applying lemma 4.1 and noting that in the spherically symmetric case $\bar{f}(v) = f(v)$, we conclude that the entities $c = L'\bar{c}$, $S = L'\bar{S}$, and $M = L'\bar{M}L$, satisfy the same system (3.2). Moreover,

$$S = L'\bar{S} = \left\{u: |\, e'Lu\,| > \frac{\kappa}{\gamma}\right\} = \{u: |\, c'u\,| > \kappa\} \tag{5.6}$$

is actually the region (5.1).

(iii) It remains to show that a region S which, together with the corresponding M and h, satisfies (3.2) is necessarily of form (5.1); that is, that the coefficient vector $c'M^{-1}$ in the third part of (3.2) is proportional to c'.

For this purpose, take an orthogonal transformation $v = Lu$ that takes the vector $M^{-1}c$ onto γe, $\gamma > 0$, that is, such that

$$M^{-1}c = \gamma L'\, e\,. \tag{5.7}$$

This transformation takes the region $S = \{u: |c'M^{-1}u| > h\}$ into

$$\bar{S} = LS = \left\{ v: |\, v_1\,| > \frac{h}{\gamma}\right\}. \tag{5.8}$$

By lemma 4.1 the transformed matrix $\bar{M} = LML'$ satisfies

$$\bar{M} = \int_{\bar{S}} v\, v' f\,(v)\, dv\,, \tag{5.9}$$

and hence, by the argument of (i), is of form (5.2) with first diagonal element, say, $\bar{m}_1$. From (5.7) then follows

$$c = \gamma ML'e = \gamma \cdot L'\bar{M}L \cdot L'e = \gamma L'\bar{M}e = \gamma \bar{m}_1 L'e\ , \tag{5.10}$$

and hence, again by (5.7), $c = \bar{m}_1 M^{-1}c$. But this is the proportionality that we set out to establish, and so the proof of theorem 5.1 is complete.

6. The quadrically symmetric case

Theorem 5.1 is easily generalized to the case that $f(u)$ can be made spherically symmetric by a linear transformation of the argument.

Assume that there is a nonsingular linear transformation $v = Lu$ such that $\bar{f}(v) = f(L^{-1}v)$ is spherically symmetric. Since a constant factor in the argument does not affect this property, we may without restriction assume $|L| = 1$. Our assumption says that f remains constant whenever the squared distance

$$(L^{-1}v)'(L^{-1}v) = v'(LL')^{-1}v \tag{6.1}$$

remains constant, that is, on each member of a certain family of homothetic ellipsoids. We shall refer to this situation as the quadrically symmetric case.

Consider any set of entities c, S, M in u-space and their counterparts $\bar{c} = Lc$, $\bar{S} = LS$, $\bar{M} = LML'$ in v-space. According to lemma 4.1, the former set satisfies (3.2) if and only if the latter set satisfies the same equations, with u replaced by v and $f(u)$ by $\bar{f}(v) = f(L^{-1}v)$. Since $\bar{f}(v)$ is spherically symmetric, we know from theorem 5.1 that the latter system has a unique solution, generated by the region

$$\bar{S} = \{v: |\bar{c}'v| > \kappa\} \tag{6.2}$$

where κ has to be determined so as to meet the size condition. It then follows that the original system (3.2) has a unique solution generated by the transformed region

$$S = L^{-1}\bar{S} = \{u: |\bar{c}'Lu| > \kappa\} = \{u: |c'(L'L)u| > \kappa\}. \tag{6.3}$$

The matrix $L'L$ can be expressed in terms of the covariance matrix Λ of the distribution $f(u)$. Integrating over the whole u-space, we have

$$\Lambda = \int uu' f(u)\, du = L^{-1} \cdot \int vv' \bar{f}(v)\, dv \cdot L'^{-1}\ . \tag{6.4}$$

Since $\bar{f}(v)$ is spherically symmetric, the last integral is of form θI where θ is a positive scalar. It follows that $\Lambda = \theta(L'L)^{-1}$, $L'L = \theta\Lambda^{-1}$. Inserting this result in (6.3) and denoting $\kappa/\theta = \lambda$, we have the following theorem.

THEOREM 6.1. *If $f(u)$ is quadrically symmetric, then the system* (3.2) *has a unique solution generated by the region*

$$S = \{u: |c'\Lambda^{-1}u| > \lambda\}, \tag{6.5}$$

where Λ is the covariance matrix of the f-distribution, and where λ has to be determined so as to satisfy the size condition of the second part of (3.2).

7. Practical procedure

We now finally turn back to our original discrete problem. If the distribution of the item points in u-space is regular enough to justify a description by means of a quadrically symmetric density function, then we may apply theorem 6.1 and use for Λ the

empirical covariance matrix of the item points. The size requirement may be met simply by counting off from the "outer end," that is, in order of decreasing $|c'\Lambda^{-1}u|$, as many items as desired. We thus end up with the following practical procedure:

(i) Compute the moment matrix Λ with elements

$$\lambda_{jh} = \frac{1}{N}\sum_{i=1}^{N} u_{ij}u_{ih}, \qquad j, h = 1, \cdots, k. \tag{7.1}$$

(ii) Compute the vector $g = \Lambda^{-1}c$, that is, solve the equations

$$\begin{aligned} &\lambda_{11}g_1 + \cdots + \lambda_{1k}g_k = c_1, \\ &\cdots\cdots\cdots\cdots \\ &\lambda_{k1}g_1 + \cdots + \lambda_{kk}g_k = c_k. \end{aligned} \tag{7.2}$$

(iii) Compute, for each item i, the quantity

$$w_i = g'u_i = \sum_{j=1}^{k} g_j u_{ij} \tag{7.3}$$

and select the items with largest $|w_i|$.

◊ ◊ ◊ ◊ ◊

Note added in proof: A direct treatment of the discrete selection problem will be published in *Soc. Sci. Fenn. Comment. Phys.-Math.* (1956).

REFERENCES

[1] G. ELFVING, "Optimum allocation in linear regression theory," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 255–262.

[2] ———, "Geometric allocation theory," *Skand. Aktuar.*, Vol. 37 (1954), pp. 170–190.