Multiple Subspace Selection in Estimation of Multivariate Normal Means

by

Mary Ellen Bock
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #80-31

September 1980

Section 0. Summary

This paper considers the problem of estimating the multivariate normal mean vector restricted to a certain subspace of the original parameter space after selecting the subspace from several possible subspaces (not necessarily nested or overlapping). Results show that the usual estimator in the selected subspace may be improved upon for the types of selection procedures normally considered in practice provided that the chosen subspace satisfies certain conditions.

Section 1. Introduction

The problem under consideration is that of estimating the mean vector θ of a p-dimensional multivariate normal distribution with identity covariance matrix while restricting the estimator to belong to one of several subspaces of the parameter space $\mathbb{R}^p$. Charles Stone (1) considered the problem in a recent paper and the model used here is similar. (In the case that the positive definite covariance matrix A for the random vector Y is not the identity, define X to be $A^{-\frac{1}{2}}Y$ with the identity covariance matrix and label θ as $A^{-\frac{1}{2}}E[Y]$ in the discussion.) For instance, these subspaces might be generated by requiring that certain components of the vector $\underset{\sim}{\theta}$ be 0. When these results are generalized to the non-identity covariance matrix for X, this would correspond to deletion of certain independent variables from the linear regression model. The possible subspaces are denoted V_j, $j = 1,\dots,S$ and the index of the chosen subspace is m. It is not assumed that one of the subspaces is necessarily all of $\mathbb{R}^p$. Associated with each subspace V_j is a loss function $f_j(\theta,\hat{\theta}^j)$ with $\hat{\theta}^j \in V_j$ and

$$f_j(\theta,\hat{\theta}^j) = ||\hat{\theta}^j - \theta||^2 + C_j,$$

where C_j is a constant "cost" associated with choosing subspace V_j. Also associated with each subspace V_j is a function $h_j(X)$, which may be viewed as an estimate of the loss associated with the choice of the subspace V_j.

The choice of one subspace V_m among S subspaces V_j, $j = 1,\ldots,S$, of the parameter space is made by minimizing $h_j(X)$, $j = 1,\ldots,S$ at the value X. (Clearly $m = m\,(X)$.) Thus

(*) $$h_m(X) \le h_j(X), \quad 1 \le j \le S.$$

It is assumed that the above inequalities are all strict except on a set of X values of measure 0.

If the selection functions $h_j(X)$ of interest don't result in a unique minimum for "most" values of X (i.e. for all X values except those which have measure 0 for every value of θ), they can be redefined to be

$$h_j'(X) = \begin{cases} h_j(X), \text{ if } h_j(X) \neq h_m(X) \text{ for } m < j \\ (h_j(X) + 1), \text{ if } h_j(X) = h_m(X) \text{ for } m < j. \end{cases}$$

This has the effect of choosing the subspace V_m with the smallest index such that

$$h_m(X) = \min_{1\le j\le S} h_m(X).$$

<u>Example</u>. Akaike's Information Criterion (AIC) may be calculated for each subspace and then V_m would be the subspace with minimal AIC, i.e.

$$h_j(X) = (||P^jX||^2) + 2\{\text{dimension}(V_j)\}$$

is minimized over j, where $P^j = I_p - P_j$ and P_j is the orthogonal projection on V_j.

Remark. The loss function for this paper is actually

$$L(\theta,\hat{\theta}) = \sum_{m=1}^{S} \left\{ \prod_{\substack{1\le j\le s \\ j\ne m}} I_{(h_m(X),\infty)}(h_j(X)) \right\} f_m(\theta,\hat{\theta}^m(X))$$

and $\hat{\theta}^m$ is in V_m. Observe that an estimator $\hat{\theta}$ is restricted in form.

A natural choice of estimator for θ once V_m has been chosen is

$$\hat{\theta}_0^m(X) = P_m X,$$

where P_m is the orthogonal projection on V_m. This paper is an examination of conditions under which $\hat{\theta}_0^m$ may be improved.

Section 2. Improved Estimators

The theorem of this section will detail conditions under which $\hat{\theta}_0^m$ may be improved and exhibit improved estimators.

Define V_{m_0} to be a subspace of V_m. (It is not assumed that V_{m_0} is necessarily one of the original S subspaces V_j, $j = 1,\dots,S$.) Let P_{m_0} be the orthogonal projection on V_{m_0}. We consider V_{m_0} to satisfy the following:

Assumption I. Each of the inequalities

$$h_m(X) < h_j(X)$$

either has no effect on $P_{m_0} X$ (i.e. it involves only random variables independent of $P_{m_0} X$) or is equivalent to a lower bound for $||P_{m_0} X||^2$. The lower bound may be random but depends only on variables independent of $P_{m_0} X$.

Remark. Thus the S inequalities imply the constraint

$$||P_{m_0} X||^2 > \gamma(X)$$

where $\gamma(X)$ is the maximum of the lower bounds for $||P_{m_0} X||^2$ induced by the

S inequalities. (Note that the distribution of $\gamma(X)$ is independent of $P_{m_0}X$.)

Thus we may assume that $\prod_{\substack{1\le j\le S \\ j\ne m}} I_{(h_m(X),\infty)}(h_j(X))$

$= \prod_{\substack{j\ne m \\ 1\le j\le S}} I_{\{A_{j,m}(X)\}}(\delta_{j,m}(X)) \cdot I_{(\gamma(X),\infty)}(||P_{m_0}X||^2)$ where $\delta_{j,m}(X)$, $\gamma(X)$ and $A_{j,m}(X)$ are independent of $P_{m_0}X$.

The following theorem provides a large class of improved estimators if the dimension of V_{m_0} is three or more.

<u>Remark</u>. In Section 3, it will be shown that V_{m_0} satisfies Assumption I if $V_m \subseteq V_j$ or $V_{m_0} \subseteq V_j^*$ for $j = 1,\ldots,S$, where V_j^* is the orthogonal complement space of V_j.

<u>Theorem</u>. Suppose that V_{m_0} is a subspace of V_m with dimension ℓ and that

a.) V_m satisfies Assumption I

and b.) $\ell \ge 3$.

Define

$$\tilde{\theta}^m(X) = (P_m - P_{m_0})X + h(||P_{m_0}X||^2)P_{m_0}X$$

where h is a nonconstant function satisfying, for $\mu \ge 0$,

(1) $0 \le h(u) \le 1$

and (2) $g(u) = u(1 - h(u)) \le 2(\ell - 2)$ and $g(u)$ is nondecreasing.

Define $\tilde{\theta}^j(X) = \hat{\theta}_0^j(X)$ for $j \ne m$. If $\tilde{\theta}$ and $\hat{\theta}_0$ use the same selection function $h_j(X)$, then $\tilde{\theta}$ dominates $\hat{\theta}_0$ if the selection function chooses the subspace V_m with positive probability for some θ.

Remark. The proof does not require that $h(\cdot) \geq 0$, but $\tilde{\theta}^m$ may be improved by making $h(\cdot) \geq 0$.

Remark. The estimators in the theorem shrink towards the subspace $V_{m_0}^c \cap V_m$ where $V_{m_0}^c$ is the orthogonal complement of V_{m_0}.

Proof of Theorem. The difference in risk for $\tilde{\theta}$ and $\hat{\theta}_0$ is

$$R(\theta,\tilde{\theta}) - R(\theta,\hat{\theta}_0) = E_\theta\Big[\sum_{m=1}^{S} \prod_{\substack{1\leq j\leq S\\ j\neq m}} I_{(h_m(X),\infty)}(h_j(X))\Big(||\tilde{\theta}^m-\theta||^2-||\hat{\theta}_0^m-\theta||^2\Big)\Big]$$

$$= \sum_{\substack{m=1\\ V_{m_0}\subset V_m}}^{S} E_\theta\Big[\prod_{\substack{j\neq m\\ 1\leq j\leq S}} I_{\{A_{j,m}}(X)\}(\delta_{j,m}(X))E_\theta\Big[I_{(\gamma(X),\infty)}(||P_{m_0}X||^2)\ \cdot$$

$$\cdot\Big([h(||P_{m_0}X||^2) - 1]^2||P_{m_0}X||^2 - 2[h(||P_{m_0}X||^2) - 1]\ \cdot$$

$$\cdot\ [(P_{m_0}\theta)'P_{m_0}X - ||P_{m_0}X||^2]\Big)\,|\gamma(X)]\Big],$$

using the notation in the first Remark after Assumption I.

Thus it suffices to show that (**) is negative where

$$(**)\qquad = E_\theta\Big[I_{(\gamma,\infty)}(||P_{m_0}X||^2)\{(h(||P_{m_0}X||^2) - 1)^2||P_{m_0}X||^2$$

$$- 2\,(h(||P_{m_0}X||^2) - 1)[(P_{m_0}\theta)'P_{m_0}X - ||P_{m_0}X||^2]\}\Big]$$

since $\gamma = \gamma(X)$ is independent of $P_{m_0}X$.

$$\text{Now }(**) = E\Big[I_{(\gamma,\infty)}(\chi^2_{\ell,\lambda})[-(h(\chi^2_{\ell,\lambda}) - 1)]\{(-\chi^2_{\ell,\lambda})(h(\chi^2_{\ell,\lambda}) - 1) - 2\chi^2_{\ell,\lambda}\}\Big]$$

$$+\lambda E[I_{(\gamma,\infty)}(\chi^2_{\ell+2,\lambda})2(1 - h(\chi^2_{\ell+2,\lambda}))]$$

(where $\ell = \dim V_{m_0}$ and $\lambda = ||P_{m_0}\theta||^2$).

$$\text{Thus }(**) = E[I_{(\gamma,\infty)}(\chi^2_{\ell,\lambda})[1 - h(\chi^2_{\ell,\lambda})]\ \{\chi^2_{\ell,\lambda}(1 - h(\chi^2_{\ell,\lambda}))$$

$$- 2(\ell-2) - 2\chi^2_{\ell,\lambda}\}]$$

$$+ E[I_{(\gamma,\infty)}(\chi^2_{\ell-2,\lambda})2\chi^2_{\ell-2,\lambda}(1 - h(\chi^2_{\ell-2,\lambda}))]$$

because $2\lambda E[g(\chi^2_{\ell+2,\lambda})] = 2E[g(\chi^2_{\ell-2,\lambda}) \cdot \chi^2_{\ell-2,\lambda}] - 2(\ell-2)E[g(\chi^2_{\ell,\lambda})]$.

We have (**) $\le$

$$E\left[I_{(\gamma,\infty)}(\chi^2_{\ell-2,\lambda})2\chi^2_{\ell-2,\lambda}(1 - h(\chi^2_{\ell-2,\lambda}))\right] - E\left[I_{(\gamma,\infty)}(\chi^2_{\ell,\lambda})2\chi^2_{\ell,\lambda}(1 - h(\chi^2_{\ell,\lambda}))\right]$$

(if $u(1 - h(u)) \le 2(\ell-2)$) and $(1 - h(u)) \ge 0$).

This last upper bound for (**)is negative if $g(u) = u(1 - h(u))$ is non-decreasing since

$$g(\chi^2_{\ell-2,\lambda}) \le g(\chi^2_{\ell,\lambda}) = g(\chi^2_{\ell-2,\lambda} + \chi^2_2)$$

and

$$I_{(\gamma,\infty)}(\chi^2_{\ell-2,\lambda}) \le I_{(\gamma,\infty)}(\chi^2_{\ell,\lambda}) = I_{(\gamma,\infty)}(\chi^2_{\ell-2,\lambda} + \chi^2_2).$$

qed.

Remark: It is clear that one may improve on $\tilde{\theta}$ if there are further subspaces V_m with appropriate V_{m_0}.

Section 3. Selection functions

In this section conditions on V_{m_0} and conditions on the subspace selection functions h_j are imposed which guarantee the satisfaction of Assumption I.

Lemma. Assume $V_{m_0} \subseteq V_m$ and each function $h_j(x)$ is representable as a continuous strictly increasing function of $||P^jX||^2$. If $V_{m_0} \subseteq V_j$, then the inequality

$$h_m(X) < h_j(X)$$

is equivalent to the inequality

$$||P^jX||^2 > \gamma_{j,m}(X)$$

where $P^j X$ and $\gamma_{j,m}(X)$ are independent of P_{m_0}. If $V_{m_0} \subseteq$, the orthogonal complement space of V_j, then the inequality

$$h_m(X) < h_j(X)$$

is equivalent to the inequality

$$||P_{m_0} X||^2 > \gamma'_{j,m}(X)$$

where $\gamma'_{j,m}(X)$ is independent of $P_{m_0} X$.

<u>Proof</u>. Setting $g_j(||P^j X||^2) = h_j(X)$ where g_j is a continuous strictly increasing function implies that

$$h_m(X) < h_j(X)$$

is equivalent to

$$g_m(||P^m X||^2) < g_j(||P^j X||^2)$$

if and only if

$$g_j^{-1}(g_m(||P^m X||^2)) < ||P^j X||^2.$$

Note that $P_{m_0} X$ and $P^m X$ are independent. (To see this note that $V_{m_0} \subseteq V_m$ implies that $P^m P_{m_0} X = 0$. Thus

$$\begin{aligned} P^m X &= P^m P_{m_0} X + P^m P^{m_0} X \\ &= P^m P^{m_0} X. \end{aligned}$$

Furthermore, the covariance matrix of $P_{m_0} X$ and $P^m P^{m_0} X$ is $P^m P^{m_0} P_{m_0} = [0]$.) Suppose $V_{m_0} \subseteq V_j$. Then $P_{m_0} X$ is independent of $P^j X$ by the sort of reasoning used in the preceding parentheses. The equivalence of the first set of inequalities stated in the corollary follows if we set

$\gamma_{j,m}(x) = g_j^{-1}(g_m(||P^m X||^2))$.

Suppose $V_{m_0} \subseteq V_j^*$. Then

$$P^j X = P^j P_{m_0} X + P^j P^{m_0} X$$
$$= P_{m_0} X + P^j P^{m_0} X,$$

since P^j is actually the projection to V_j^*. Note that $P_{m_0} X$ is orthogonal to $P^j P^{m_0} X$ since

$$X^t P^{m_0} P^j P_{m_0} X = 0.$$

Also, $P_{m_0} X$ is independent of $P_j P^{m_0} X$ because their covariance is

$$P_{m_0} (P^j P^{m_0})^t = P_{m_0} P^{m_0} P^j = [0].$$

Setting $\gamma'_{j,m}(X) = g_j^{-1}(g_m(||P^m X||^2)) - ||P^j P^{m_0} X||^2$, we have that the last two inequalities stated in the corollary are equivalent when we substitute

$$||P_{m_0} X||^2 + ||P^j P^{m_0} X||^2 \quad \text{for} \quad ||P^j X||^2$$

in the inequality

$$||P^j X||^2 > g_j^{-1}(g_m(||P^m X||^2)).$$

qed.

<u>Example</u>. Setting

$$h_j(X) = ||P^j X||^2 + C_j$$

where C_j is a constant (possibly a cost proportional to the dimension of the space V_j), the assumptions of the lemma about h_j are satisfied. Observe that for $V_i \subset V_j$, we have $||P^j X||^2 \leq ||P^i X||^2$ for all X and V_i is effectively removed from the set of choices unless $C_j > C_i$.

Example. Various forms of Akaike's Information Criterion have the form

$$h_j(X) = r(||P^jX||^2) + \beta_j$$

where β_j is independent of X and r is a continuous strictly increasing function. These are continuous strictly increasing functions of $||P^jX||^2$, and the previous lemma applies for the h_j.

The next corollary examines the implications of Assumption I for the type of h_j's considered in the lemma.

Corollary: (1) Assume that each function h(X) is representable as a continuous strictly increasing function of $||P^jX||^2$, j = 1,...,S.

(2) Assume that V_{m_0} is a linear subspace of the subspace V_{m_0} (not necessarily equal (to V_{m_0}) such that for each j, either $V_{m_0} \subseteq V_j$ or $V_{m_0} \subseteq V_j^*$, j = 1,...,S.

Then if the dimension of V_{m_0} is three or more, the estimator given in the theorem dominates the estimator $\hat{\theta}_0$.

Proof: The last lemma shows that (1) and (2) imply Assumption I and the theorem may be applied.

qed.

Remark: These results imply the inadmissibility of the estimator $\hat{\theta}_0$ under special assumptions about the relationships of the V_j's to one another. Namely, there is a subspace V_{m_0} such that $V_{m_0} \subseteq V_j$ or $V_{m_0} \subseteq V_j^*$ for j = 1,...,S and V_{m_0} is contained in at least one of the subspaces $V_1,\ldots,V_S$.

<u>Example</u>: Set p = 7 and

$$\text{let } V_1 = \{\theta \in \mathbb{R}^p: \ \theta_1 = \theta_2 = \theta_3 = 0\}$$

and

$$V_2 = \{\theta \in \mathbb{R}^p: \ \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7 = 0\}$$

and

$$V_{m_0} = \{\theta \in \mathbb{R}^p: \ \theta_1 = \theta_2 = \theta_3 = \theta_4 = 0\}.$$

Then $\ell = \dim V_{m_0} = 3$ and $V_{m_0} \subseteq V_1$ and $V_{m_0} \subseteq V_2^*$.

<u>References</u>:

(1) Stone, Charles J., "Admissible Selection of an Accurate and Parsimonious Normal Linear Regression Model" (1980) to appear in <u>Annals of Statistics</u>.