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 \( \theta \) 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 \( A^{-1/2}Y \) with the identity covariance matrix and label \( \theta \) as \( A^{-1/2}\mathbb{E}[Y] \) in the discussion.) For instance, these subspaces might be generated by requiring that certain components of the vector \( \theta \) 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 \( V_j \), \( j = 1, \ldots, S \) 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
\[ f_j(\hat{e}, \hat{e}_j) = ||\hat{e}_j - e||^2 + C_j, \]

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

\[
(*): \quad h_m(X) \leq h_j(X), \quad 1 \leq j \leq S.
\]

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 \( e \)), they can be redefined to be

\[
h'_j(X) = \begin{cases} 
  h_j(X), & \text{if } h_j(X) \neq h_m(X) \text{ for } m \neq j \\
  (h_j(X) + 1), & \text{if } h_j(X) = h_m(X) \text{ for } m \neq j.
\end{cases}
\]

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

\[
h_m(X) = \min_{1 \leq j \leq S} h_m(X).
\]

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

\[
h_j(X) = (||P^j X||^2) + 2(\text{dimension}(V_j))
\]

is minimized over \( j \), where \( P^j = I - P_j \) and \( P_j \) is the orthogonal projection on \( V_j \).
Remark. The loss function for this paper is actually

$$L(\theta, \hat{\theta}) = \sum_{m=1}^{S} \left\{ \prod_{1 \leq j < S} I(h_m(x), \infty) (h_j(x)) \right\} f_m(\theta, \hat{\theta}^m(x))$$

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

A natural choice of estimator for $\theta$ 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 $V_j, j = 1, \ldots, S$.) 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

\[
\prod_{1 \leq i \leq S} I(h_m(X), \delta_i(X)) (h_j(X))
\]

\[
= \prod_{j \neq m} I(A_{j,m}(X))(\delta_{j,m}(X)) \cdot I(\gamma(X), \delta)(||P_{m_0} X||^2)
\]

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.

Remark. In Section 3, it will be shown that \( V_{m_0} \) satisfies Assumption I if \( V_m \subseteq V_{j_0} \) or \( V_{m_0} \subseteq V_{j_0}^\perp \) for \( j = 1, \ldots, S \), where \( V_{j_0}^\perp \) is the orthogonal complement space of \( V_{j_0} \).

Theorem. Suppose that \( V_{m_0} \) is a subspace of \( V_m \) with dimension \( \ell \) and that

a.) \( V_m \) satisfies Assumption I

and b.) \( \ell \geq 3 \).

Define

\[
\hat{\theta}^m(X) = (P_m - P_{m_0})X + h(||P_{m_0} X||^2)P_{m_0} X
\]

where \( h \) is a nonconstant function satisfying, for \( u \geq 0 \),

1. \( 0 \leq h(u) \leq 1 \)

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

Define \( \hat{\theta}_0^j(X) = \hat{\theta}_0^j(X) \) for \( j \neq m \). If \( \hat{\theta} \) and \( \hat{\theta}_0 \) use the same selection function \( h_j(X) \), then \( \hat{\theta} \) dominates \( \hat{\theta}_0 \) if the selection function chooses the subspace \( V_m \) with positive probability for some \( \theta \).
Remark. The proof does not require that $h(\cdot) \geq 0$, but $\dot{\theta}^m$ may be improved by making $h(\cdot) \geq 0$.

Remark. The estimators in the theorem shrink towards the subspace $V_{\theta_0}^c \cap V_{\gamma}^c$, where $V_{\theta_0}^c$ is the orthogonal complement of $V_{\theta_0}^c$.

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

$$R(\theta, \hat{\theta}) - R(\theta, \hat{\theta}_0) = E_{\theta} \left[ \sum_{m=1}^{S} \prod_{1 \leq j \leq S} \frac{I(h_m(X), \omega)}{(\omega_j^m) (|\dot{\theta}^m - \theta|| \dot{\theta}_0^m - \theta||^2)} \right]$$

$$= \sum_{m=1}^{S} E_{\theta} \left[ \prod_{j \neq m} \frac{I(\theta_j, \omega)}{(\delta_j^m(X))} \right] \left( \frac{E_{\theta} \left[ I(\gamma(X), \omega) \right]}{|P_{\theta_0}^m X|} \right) \cdot$$

$$\cdot \left( \frac{|h(||P_{\theta_0}^m X||^2) - 1||P_{\theta_0}^m X||^2 - 2|h(||P_{\theta_0}^m X||^2) - 1||}{|P_{\theta_0}^m X||^2} \right) \cdot$$

$$\cdot \left[ (P_{\theta_0}^m \cdot P_{\theta_0}^m X - ||P_{\theta_0}^m X||^2) \right] \cdot \gamma(X) \right),$$

using the notation in the first Remark after Assumption I.

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

$$\text{(**) } = E_{\theta} \left[ I(\gamma, \omega) \left( ||P_{\theta_0}^m X||^2 \right) \left( h(||P_{\theta_0}^m X||^2) - 1 \right)^2 ||P_{\theta_0}^m X||^2 \right.$$  

$$- 2 \left( h(||P_{\theta_0}^m X||^2) - 1 \right) \left( (P_{\theta_0}^m \cdot P_{\theta_0}^m X - ||P_{\theta_0}^m X||^2) \right) \right]$$

since $\gamma = \gamma(X)$ is independent of $P_{\theta_0}^m X$.

Now $\text{(**) } = E \left[ I(\gamma, \omega) \left( x_{\lambda, \lambda}^2 \right) \left( h(x_{\lambda, \lambda}^2) - 1 \right) \right.$  

$$\left. \left( - x_{\lambda, \lambda}^2 \right) \left( h(x_{\lambda, \lambda}^2) - 1 \right) - 2 x_{\lambda, \lambda}^2 \right]$$

$$+ \lambda E \left[ I(\gamma, \omega) \left( x_{\lambda+2, \lambda}^2 \right) \left( 1 - h(x_{\lambda+2, \lambda}^2) \right) \right]$$

(where $\lambda = \dim V_{\theta_0}^m$ and $\lambda = ||P_{\theta_0}^m \omega||^2$).

Thus $\text{(**) } = E \left[ I(\gamma, \omega) \left( x_{\lambda, \lambda}^2 \right) \left[ 1 - h(x_{\lambda, \lambda}^2) \right] \left( x_{\lambda, \lambda}^2 \right) \left( 1 - h(x_{\lambda, \lambda}^2) \right) \right.$  

$$\left. - 2(\lambda - 2) - 2 x_{\lambda, \lambda}^2 \right]$$

$$+ E \left[ I(\gamma, \omega) \left( x_{\lambda-2, \lambda}^2 \right) x_{\lambda-2, \lambda}^2 \left( 1 - h(x_{\lambda-2, \lambda}^2) \right) \right]$$
because \(2\lambda E[g(x_{\lambda+2, \lambda}^2)] = 2E[g(x_{\lambda-2, \lambda}^2) \cdot x_{\lambda-2, \lambda}^2] - 2(\lambda-2)E[g(x_{\lambda, \lambda}^2)].\)

We have (**): 
\[
E\left[I_{(\gamma, \infty)}(x_{\lambda-2, \lambda}^2)2x_{\lambda-2, \lambda}^2, \lambda (1 - h(x_{\lambda-2, \lambda}^2))\right] - E\left[I_{(\gamma, \infty)}(x_{\lambda, \lambda}^2)2x_{\lambda, \lambda}^2, \lambda (1 - h(x_{\lambda, \lambda}^2))\right]
\]
(if \(u(1 - h(u)) \leq 2(\lambda-2)\) and \(1 - h(u)) \geq 0).\)

This last upper bound for (**is negative if \(g(u) = u(1 - h(u))\) is non-decreasing since
\[
g(x_{\lambda-2, \lambda}^2) \leq g(x_{\lambda, \lambda}^2) = g(x_{\lambda-2, \lambda}^2 + x_2^2)
\]
and
\[
I_{(\gamma, \infty)}(x_{\lambda-2, \lambda}^2) \leq I_{(\gamma, \infty)}(x_{\lambda, \lambda}^2) = I_{(\gamma, \infty)}(x_{\lambda-2, \lambda}^2 + x_2^2).
\]
\[\text{QED}\]

\textbf{Remark:} It is clear that one may improve on \(\delta\) if there are further subspaces \(V_m\) with appropriate \(V_m\).

\textbf{Section 3. Selection functions}

In this section conditions on \(V_m\) and conditions on the subspace selection functions \(h_j\) are imposed which guarantee the satisfaction of Assumption I.

\textbf{Lemma.} Assume \(V_m \subseteq V_j\) and each function \(h_j(x)\) is representable as a continuous strictly increasing function of \(||p_j^jx||^2\). If \(V_m \subseteq V_j\), then the inequality
\[
h_m(X) < h_j(X)
\]
is equivalent to the inequality
\[
||p_j^jx||^2 > y_{j,m}(X)
\]
where \( p^j X \) and \( \gamma_{j,m}(X) \) are independent of \( p_{m_0} X \). 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 \( p_{m_0} X \).

**Proof.** 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

\[
p^m X = p_{m_0} p^m X + p_{m_0} p_0^m X
\]

\[
= p^m p_0^m X.
\]

Furthermore, the covariance matrix of \( p_{m_0} X \) and \( p^m p_0^m X \) is \( p^m p_0^m p_{m_0} = [0] \).

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
\[
p^j x = p^j_{m_0} x + p^j_{m_0} p^0 x
\]
\[
= p_{m_0} x + p^j_{m_0} p^0 x,
\]
since \( p^j \) is actually the projection to \( V_j^* \). Note that \( p_{m_0} x \) is orthogonal to \( p^j_{m_0} p^0 x \) since
\[
x^t p_{m_0} p^j_{m_0} x = 0.
\]

Also, \( p_{m_0} x \) is independent of \( p^j_{m_0} p^0 x \) because their covariance is
\[
p_{m_0}(p^j_{m_0} p^0)^t = p_{m_0} p^0 p^j = [0].
\]

Setting \( \gamma_j,m'(x) = g_j^{-1}(g_m(||p^m x||^2)) - ||p^j_{m_0} p^0 x||^2 \), we have that the last two inequalities stated in the corollary are equivalent when we substitute
\[
||p_{m_0} x||^2 + ||p^j_{m_0} p^0 x||^2 \quad \text{for} \quad ||p^j x||^2
\]
in the inequality
\[
||p^j x||^2 > g_j^{-1}(g_m(||p^m x||^2)).
\]

\[ \text{qed.} \]

**Example.** 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 \( V_i \subseteq V_j \), 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

$$h_j(x) = r(||p^j x||^2) + \beta_j$$

where $\beta_j$ is independent of $x$ and $r$ is a continuous strictly increasing function. These are continuous strictly increasing functions of $||p^j x||^2$, 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 $||p^j x||^2$, $j = 1, \ldots, S$.

(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 $V_{m_0} \subseteq V_j^*$, $j = 1, \ldots, S$.

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 $j = 1, \ldots, S$ and $V_{m_0}$ is contained in at least one of the subspaces $V_1, \ldots, V_S$. 
Example: Set $p = 7$ and

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

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 $k = \dim V_{m_0} = 3$ and $V_{m_0} \subseteq V_1$ and $V_{m_0} \subseteq V_2$.

References: