CENTER FOR COMPUTATIONAL RESEARCH
IN ECONOMICS AND MANAGEMENT SCIENCE

# NEW GEOMETRIC THEORY FOR THE LINEAR MODEL

by

R. Wayne Oldford*

Technical Report No. 48 January 1985

ALFRED P. SLOAN SCHOOL OF MANAGEMENT
MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139

**ABSTRACT**

A mathematical theory is presented which extends the geometric theory of vector spaces to deal particularly with finite collections of vectors. This theory is then exploited in the case of the linear model to describe the geometry of certain practically relevant issues such as least-squares regression diagnostics.

**Keywords:** Geometry of linear model, Numerical ranks, Regression diagnostics, Ridge regression.

# 1. Introduction

Consider the geometry of the linear model

$$y = Xb + e \tag{1.1}$$

where $b \epsilon I\!R^m, X = (X_1, ..., X_m)$ and $y, e, X_1, ..., X_m \epsilon I\!R^n$. The response vector $y$, is to be fitted by some vector $\hat{y} = Xb$ which lies in the subspace of $I\!R^n$ generated by the vectors $X_1, ..., X_m$. The residual vector $e \equiv y - \hat{y}$ lies in a complementary subspace and gives the error with which $y$ is fitted by $\hat{y}$. This framework has powerful theoretical and pedagogical value, particularly when used to study or motivate such standard statistical techniques as least-squares regression or analysis of variance ( see e.g. Scheffé [1959], Seber [1966], Kruskal [1968], or Gurrman [1983]). In light of the geometry, such procedures become intuitively compelling.

In what follows, a more detailed geometric framework is proposed which supplements the usual vector space approach and makes it possible to bring geometric intuition to bear on a greater variety of statistical methods. In particular, the new frame work will be shown to yield geometric insight into matrix rank measures, collinearity problems, least-squares influential data diagnostics and the minimax properties of certain ridge regression estimators. The new frame work takes as a basic premise that in practice the observed vectors are often of as much interest as the subspaces they generate. Hence, instead of spaces and subspaces, finite collections of vectors will be the objects of geometric interest.

The mathematical theory involved is presented in the next section (proofs of certain results appear in the appendices). As will be shown, strong parallels exist between the new geometric framework and the more familiar geometry of finite dimensional vector spaces. The remaining sections apply the theory to different areas of practical statistical interest.

# 2. A finite Collection Extension

In this section, theory is presented to describe certain properties of finite collections of vectors. The vectors may come from any finite dimensional vector space $P$ with inner product $(u, v)$ for $u, v \epsilon P$. However, in the exposition it will be simpler to assume that $P = I\!R^m$ However, in the exposition it will be simpler to assume that $P = I\!R^m$ for some finite $m \geq 1$, and that $(u, v) = u^T v$, the usual inner product. What follows is an extension of vector space theory which focuses on particular finite collections of vectors. New objects, operators and attributes are defined and linked to their familiar counterparts in finite dimensional vector space theory. Table 1 summarizes some of the relationships to be discussed.

To motivate the theory, suppose that some practical problem under investigation yields $N$ data vectors $p_1, ..., p_N \epsilon I\!R^m$ to be studied. Typically, these vectors will have been selected in an arbitrary fashion but will have some meaning attached to them. For example, $p_1$ may represent $m$ measurements of the gross domestic product of a nation, $p_2 m$ measurements of the price of oranges and so on. Conceivably, each of these vectors will have some intrinsic meaning within the problem while an arbitrary linear combination of them will not. A geometric frame work which does not ignore the individuality of each vector is therefore proposed.

Table 1. Analogies between items of the proposed extension and those of the finite dimensional vector space theory.

| | Finite Dimensional<br>Vector Space Theory | Finite Collection<br>Extension |
|---|---|---|
| Objects | -vector space $P$<br>-subspace $S$ of $P$ | - Star P<br>- Substar S of P |
| Operators | -$S \subset P \equiv$ ‘$S$ is a subspace of $P$’<br>-$S \cap T \equiv$ vector space intersections<br>-$S \oplus T \equiv$ vector space addition<br>-isomorphic vector spaces $P$ and $Q$ | - S $\overset{\star}{\subset}$ P $\equiv$ ’S is a subset of P’<br>- S $\overset{\star}{\cap}$ T $\equiv$ ’substar intersection’<br>- S $\overset{\star}{\cup}$ T $\equiv$ ’substar union’<br>-‘star-equivalent’ stars P and Q (P $\overset{\star}{=}$ Q) |
| Attributes | -dim $(P) \equiv$ dimension of the vector space $P$<br>-dim$(S) \equiv$ dimension of the subspace $S$ | -d$_\alpha$ (P;P) $\equiv$ ‘effective dimension of the star P’<br>-d$_\alpha$ (S;P) $\equiv$ ‘effective dimension of the substar S of P’ |

Pictorially, $p_1, ..., p_N$ can be represented as rays emanating from the origin to the points $p_1, ..., p_N$ in $I\!R^m$ as in Figure 1, where $N = 4$ and $m = 2$.

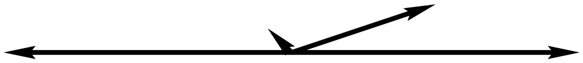


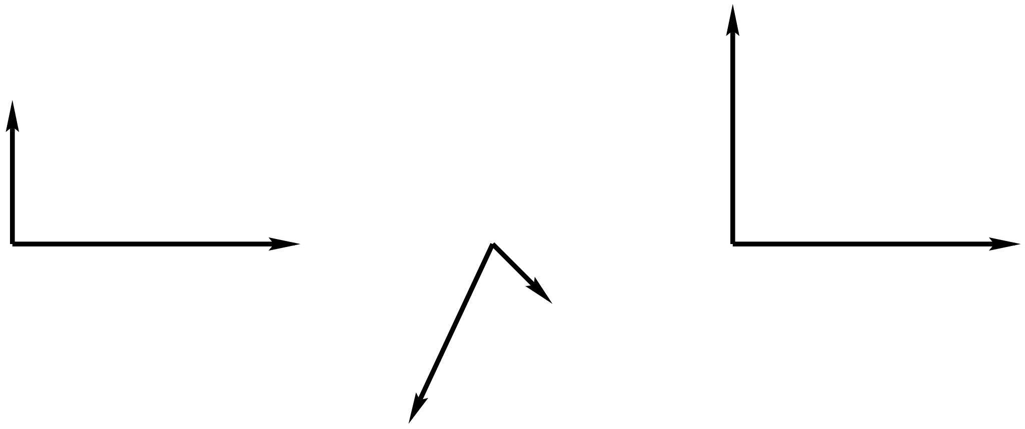
Figure 1: $< p_1, p_2, p_3, p_4 >$ in $I\!R$

Such a representation suggests the following definition.

Definition 2.1: A collection P, denoted P $=< p_1, ..., p_N >=< p_i : i \epsilon \{1, ..., N\} >$, of $Nm-$dimensional vectors $p_1, ..., p_N \epsilon I\!R^m$ such that $1 \leq m \leq \infty$ and $N \geq 0$ is called a *star* in $I\!R^m$

Note that the possibility that $p_i = p_j$ for some $i \neq j$ is not excluded. Also, if no vectors are in P$(N =))$ then P is the *null star* and denoted by the empty set symbol $\emptyset$.

Stars are basic objects of the extension and, as suggested in Table 1, may be compared to the set of all vectors in the space spanned by the vectors of the star. For example, consider again the vectors of Figure 1. The same vector space, $I\!R^2$, is produced by the span of two or more of the vectors in the figure. The vector spaces denoted by span$(p_1, p_2)$, span$(p_2, p_3)$, and span$(p_1, p_2, p_3)$ are identical. A more general notion of equivalence of vector spaces would be isomorphism. When it comes to stars, equivalence of the corresponding vector spaces will not do. The stars $< p_1, p_2 >, < p_2, p_3 >$ and $< p_1, p_2, p_3 >$ are clearly distinct and any definition of the equivalence of stars should distinguish these three as different. Roughly speaking, a star, P, will be said to be equivalent to another star S, if P can be scaled by a single scale factor c, and/or rotated in $I\!R^m$, so as to fit on top of S. For example, Figures 2(a), (b) and (c) depict three different stars in $I\!R^2$



(a) (b) (c)

**Figure 2:** Three stars in $I\!R^2$: (a) and (b) show equivalent stars; star in (c) is not equivalent to either of the others.

The star in Figure 2(b) may be rotated and scaled up to be identical to the star in Figure 2(a). The star in Figure 2(c) has the same orientation as that in Figure 2(a), but no single scale factor applied to both of its vectors will yield the star of Figure 2(a). HEnce the stars of Figures 2(a) and 2(c) are not equivalent.

More formally, let P = $< p_1, ..., p_N >$ and S $=< s_1, ..., s_K >$ be two stars in $I\!R^m$ and suppose the indices of the vectors are arranged so that $p_1 = 0$ and $s_j = 0$ whenever $i > n$ and $j > k$, for some $n \leq N$ and $k \leq K$. The following definition determines when P and S are judged to be equivalent.

Definition 2.2: P and S, as defined above, are said to be *star-equivalent*, written P$\overset{\star}{=}$S, if there exist a non-zero scalar c$\epsilon I\!R$, an $m \times m$ orthogonal matrix ), and a permutation function $\pi$ of the indices $i = 1, ..., n$, such that

(i) $n = k$

and (ii) $p_{\pi(i)} =$ c $\cdot 0 s_j$ for $i = 1, ..., n$.

The permutation, $\pi$, makes the equivalence independent of the order in which the vectors of P and S are indexed.

For every vector space $P$, a subspace $S$ can be constructed by selecting vectors, $p_1, ..., p_k$ from $P$ and forming $S =$ span$(p_1, ..., p_k)$. Similarly, selecting $k$ vectors from the star P = $< p_1, ..., p_N >$ will yield another star $< p_1, ..., p_k >$, say, which “fits inside” P. We say that, $< p_1, ..., p_k >$ is a *substar* of P. Before formalizing this concept, consider Figure 3. There an arbitrary star P, shown in Figure 3(a), is represented by dashed lines and other stars represented by solid lines are placed on top of P in Figures 3(b)-(e). Only Figures 3(b) and 3(c) show substars of P. Note that the solid stars of Figures 3(c) and (e) are star-equivalent to the dashed star but only that of Figure 3(c) is a substar of P.

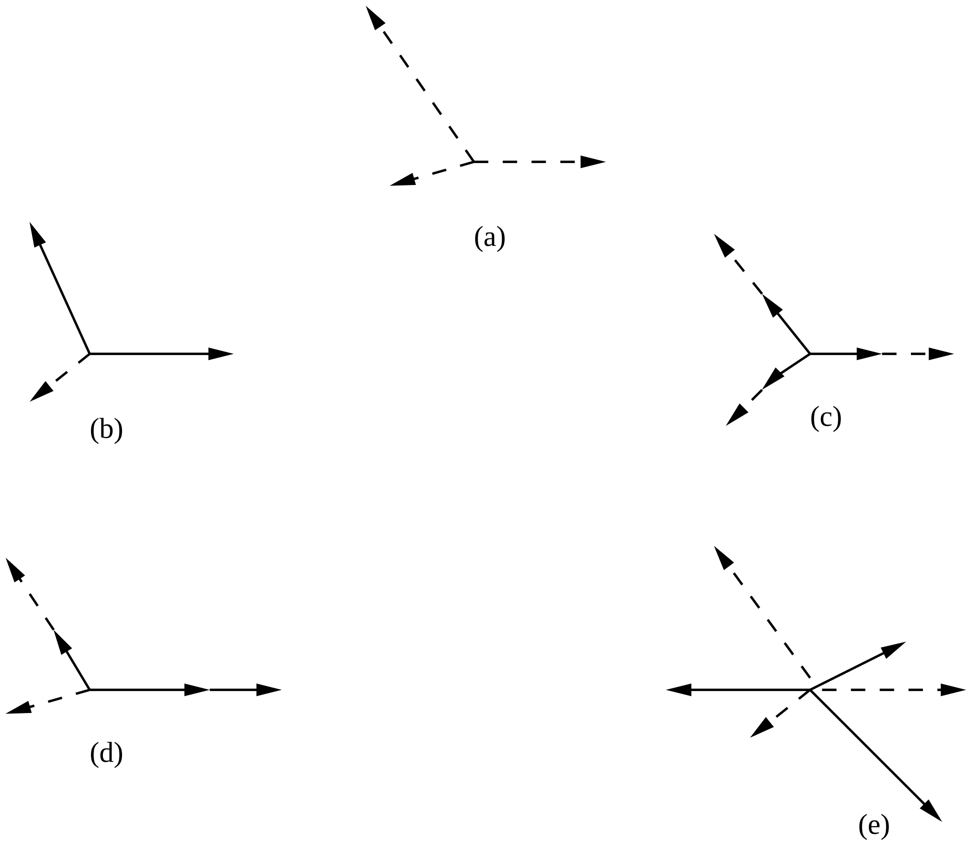


**Figure 3:** (a) P; (b) a substar of P; (c) a substar of P; (d) not a substar; (e) not a substar.

Definition 2.3: For P and S as above, S is said to be a *substar* of P, written S $\overset{\star}{\subseteq}$ P, if there exists scalars c$_i$(S)$\epsilon(0,1] i = 1, ..., k$ and a permutation function $\pi$ of the indices $1, ..., n$ such that

(i) $0 \leq k \leq n$
and (ii) $s_i = c_i(S)p_{\pi(i)}$ for $i = 1, ..., k$.

As is the case with subspaces, the substar operator "$\overset{\star}{\subseteq}$" is a transitive one.

Having defined substars of a given star P, we next consider operations on substars of P that lead to new substars of P. For these purposes, let $S = < s_1, ..., s_K >$ and $Q = < q_1, ..., q_L >$ be substars of $P = < p_1, ..., p_N >$ such that the following hold:

$$\begin{array}{rl} (i) & s_i = c_i(S)p_{\pi(i)} \text{ for } i = 1, ..., k \\ (ii) & s_i = 0 \text{ for} i = k+1, ...., K \end{array} \qquad (2.1)$$

and similarly

$$\begin{aligned} (iii) &\quad q_j = c_j(Q)p_{\gamma(j)} \text{ for } j = 1, ..., \ell \\ (iv) &\quad q_j = 0 \text{ for } j = \ell + 1, ..., L \end{aligned} \tag{2.2}$$

where $k, \ell \leq N, C_i(S)$ and $c_j(Q)\epsilon(0, 1]$ for all $i = 1, ..., k$ and $j = 1, ..., \ell,$ and $\pi(i)$ and $\gamma(i)$ are two specified permutation functions of the indices $i = 1, ..., n$.

Unless $c_i(S) = c_j(Q)$ whenever $\pi(i) = \gamma(j)$, the collection of vectors formed by the union of the vectors of S and Q will not necessarily be a substar of P. Alternatively, the collection formed by the vectors in the intersection of the set of vectors in S and the set of vectors in Q will always product a substar of P.[1] However, this substar will often be empty or contain only the zero vector. More intuitive versions of the union- and intersection- like operations which will always produce substars of P are now defined. For expository reasons we set $c_i(S) = 0$ for $k < i \leq n$ and $c_j(Q) = 0$ and $\ell < j \leq n$.

Definition 2.4: For $S, Q \overset{\star}{\subset} P$ as defined above, union $(\overset{\star}{\cup})$ and intersection $(\overset{\star}{\cap})$ operators on $S$ and $Q$ are defined to be

(a) $S \overset{\star}{\cup} Q =< d_h p_h : h = 1, ..., n >$

(b) $S \overset{\star}{\cap} Q =< e_h p_h : h = 1, ..., n >$

where for $h = \pi(i) = \gamma(j) = 1, ..., n$[2]

$d_h = max\{c_i(S), c_j(Q)\}$[3]

and

$e_h = min\{c_i(S), c_j(Q)\}.$

The intuitive motivation for these operators, $\overset{\star}{\cup}$ and $\overset{\star}{\cap}$, can be seen by examining the five stars of Figures 4 (a)-(e). In (a) some star P is shown. The remaining stars (b)-(e) are substars of P, (b) and (c) representing two arbitrary substars, $S$ and $Q$, and (d) and (e) representing their substar union $S \overset{\star}{\cup} Q$ and substar intersection $S \overset{\star}{\cap} Q$ respectively.

As might be expected, a number of properties of substar union and intersection are easily shown to hold. Among them are the following:

(i) $< 0 > \overset{\star}{\subseteq} S \overset{\star}{\cap} Q \overset{\star}{\subseteq} S \overset{\star}{\subseteq} S \overset{\star}{\cup} Q \overset{\star}{\subseteq} P$

(ii) $S \overset{\star}{\cap} P = S$

[1] If $S$ and $Q$ are subspaces of a vector space $P$, the analogy persists: a subspace of $P$ is not necessarily produced by $S \cup Q$, while one is always produced by $S \cap Q$. Introducing vector space addition $\oplus$, however, will yield a subspace $S \oplus Q$.

[2] Note that the permutation functions $\pi$ and $\gamma$ used to define S and Q determine the sequence of $i$'s and $j$'s, respectively.

[3] $max\{\cdot\}$ denotes the maximum element in the set $\{\cdot\}$ and min $\{\cdot\}$ the minimum element.

(iii) $S \overset{\star}{\cup} P = P$

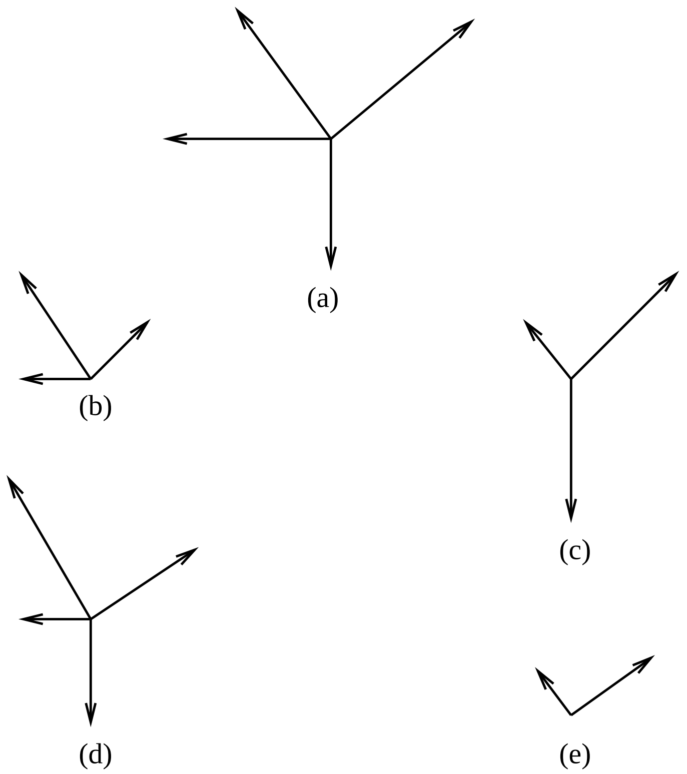


Figure 4: (a) P; (b) $S \overset{\star}{\subseteq} P$; (c) $Q \overset{\star}{\subseteq} P$; (d) $S \overset{\star}{\cup} Q$; (e) $S \overset{\star}{\cap} Q$

A potential ambiguity exists unless each non-zero vector of the substars $S$ and $Q$ above is given explicitly as a specified scalar $c_i \epsilon (0, 1]$ times $p_i$. For example, suppose that $p_1 = p_2, s_1 = 0.5p_i$ and $q_1 = 0.7p_j$ where it is only known that $i$ and $j$ are either 1 or 2. The ambiguity in $s_1$ and $q_1$ causes an unresolvable ambiguity in $S \overset{\star}{\cup} Q$. If $i = 1$ and $j = 2$, then $S \overset{\star}{\cup} Q$ includes both $0.5p_1$ and $0.7p_2$p; whereas if $i = j = 1$ then $S \overset{\star}{\cup} Q$ might only include $0.7p_1$. The same ambiguity arises for $S \overset{\star}{\cap} Q$. For this reason, application of the star-union and star-intersection operators requires a precise specification oft he vectors in each substar like that given in equations (2.1) and (2.2) for $S$ and $Q$.

From Table 1, the only item remaining to be introduced and examined in this extension is an attribute of stars and substars called the *effective dimension*[4]. This attribute takes non-negative real values which can be given an interpretation similar to that of the usual dimension. Given that attention is to be focussed on the particular collection of vectors in hand and also given the possible inexactness of these vectors in

[4] For a more general axiomatic treatment see Oldford [1983].

practice, the basic premise of an effective dimension is that certain orthogonal directions, or dimensions, are better determined by the collection of others. For example, consider again Figure 1. The premise suggests that the horizontal dimension is more clearly defined by this star than is the vertical dimension. In this sense, a value of 2 for the dimension is considered misleading. A number between 1 and 2 would more closely reflect the ambivalence with which the vertical dimension is regarded. For the star of Figure 2(c), on the other hand, each dimension is equally well defined and the value of 2 is appropriate.

Capturing this notion of relative merits of dimensions is further complicated when the object of interest is a substar, $S$, of a given star $P$. For example, let $P$ be the star of Figure 3(a) and $S$ be the substar in Figure 3(c). Given $P$, the effective dimensionality suggested by $S$ should be smaller than that suggested by $P$ alone. However, $S$ is star equivalent to $P$ and when taken alone it should have the same dimensionality as that of $P$. Which is appropriate for $S$ depends upon whether its dimensionality is to be taken with respect to $P$ or not. The function to be proposed as an effective dimension will therefore operate on a (substar, star) pair. Since a star is always a substar of itself, when the effective dimension is to be taken of $S$ alone, the appropriate pair will be $(S, S)$.

For any star $Q$ let $\lambda_i(Q)$ denote the $i$-th largest singular value [5] of the matrix having column (or row) vectors equivalent to the vectors of $Q$. We then have

Definition 2.5: The $d_\alpha$ - *effective dimension* of $S \overset{\star}{\subseteq} P$ where $\alpha > 0$ is given by $d_\alpha(S; P)$ where

$$d_\alpha(S;P) = \begin{cases} 0 \text{ if } S =< 0 >, \text{ or} S = \emptyset \\ [\lambda_1(P)]^{-\alpha} \sum_{i=1}^{k} [\lambda_i(S)]^\alpha \text{ otherwise} \end{cases} \tag{2.3}$$

As $\alpha \to 0, d_\alpha(S; P)$ will give the dimension [6], $dim(S)$, of the subspace of $I\!R^m$ spanned by the vectors of $S$. For $\alpha > 0, d_\alpha(S; P)$ will equal some non-negative real number less than or equal to this dimension.

In what follows, similarities and differences between the $d_\alpha$ - effective dimension of a subspace may be seen by simply identifying the elements of the first column of Table 1 with the corresponding ones of column 2 whenever these appear. That is, to check whether any results given below is analogous to, or fundamentally different from, a result concerning vector spaces, stars like $< \mathrm{a}, \mathrm{b}, \mathrm{c} >$ must be replace by the span (a, b, c), and so on. for example, if $u, v \epsilon I\!R^m$ are both non-zero, we have $1 = d_\alpha(< v >; < v >) = dim(\mathrm{span}(v)) > d_\alpha(< v >; < u, v >)$ for all $\alpha > 0$. while some similarity exists between $d_\alpha$ and $dim$ when operating on a single vector, this simple example underscores the importance of the substar-star relationship to the $d_\alpha$-effective dimension.

[5] For vector spaces other than $I\!R^m$ and arbitrary inner products, $\lambda_i(Q)$ may be defined as the square root of the $i$-th largest eigenvalue of the matrix of inner products of the vectors of $Q$.

[6] The number of basis vectors necessary to generate the subspace.

The effect this relationship has on $d_\alpha$ is investigated in the following series of propositions (proofs are detailed in Appendix A). With the exception of the first proposition, $S$ and $T$ will denote substars, while $P$ and $Q$ denote stars. Occasional reference to the pictures of Figure 4 should make most results transparent.

Proposition 2.1: If $\overset{\star}{\subseteq} T \overset{\star}{\subseteq} Q \overset{\star}{\subseteq} P$ then for all $\alpha > 0$

$$d_\alpha(S; P) \leq d_\alpha(T; P) \leq d_\alpha(T; Q) \tag{2.4}$$

This result describes the effect of increasing, or decreasing, in size either partner of the (substar, star) pair. Increasing the size of the substar or decreasing the size of the star will increase the effective dimension. Equivalently, a decrease in $d_\alpha$ results if the opposite changes in the sizes of the substar and star are made. Proposition 2.1 may be illustrated by letting $S, T$ and $Q,$ or $T, Q,$ and $P,$ be the stars of Figures 4(b), (d) and (a) respectively. Note also that when the vector space quantities are substituted in (2.4) only the left hand inequality holds and the right hand one becomes equality.

Proposition 2.2:[7] If $S \overset{\star}{\subseteq} P$ and $S \overset{\star}{\subseteq} Q$ then for all $\alpha > 0$

$$d_\alpha(S; P \cup Q) \leq \frac{1}{2}\left[d_\alpha(S; P) + d_\alpha(S; Q)\right]; \tag{2.5}$$

Proposition 2.3:[7] If $S \cup T \overset{\star}{\subseteq} P$ then for $\alpha = 1$ and $\alpha = 2,$

$$d_\alpha(S \cup T; P) \leq d_\alpha(S; P) + d_\alpha(T; P) \tag{2.6}$$

While (2.5) and (2.6) describe similar bounds for $d_\alpha$ applied to the set-union of stars and substars, note that they differ in one important respect, namely, (2.6) requires $\alpha$ to be 1 or 2. for other values of $\alpha > 0,$ it is an open problem as the whether (2.6) applies or not.

In terms of dimensions of vector spaces, strict inequality holds in Proposition 2.2, unless $dim(S)$ is zero. As it stands, Proposition 2.3 does not make sense in terms of $dim.$ The following result is more meaningful.

Proposition 2.4: If $S \overset{\star}{\subseteq} P$ and $T \overset{\star}{\subseteq} P$, then for $\alpha = 1$ and $\alpha = 2$

$$d_\alpha(S \overset{\star}{\cup} T; P) \leq d_\alpha(S; P) + d_\alpha(T; P). \tag{2.7}$$

If $S \cup T \overset{\star}{\subseteq} P$ then (2.7) follows from (2.4) and (2.6), otherwise the proof given in Appendix A is required.

[7] Here the set-union operation is used on stars. The standard interpretation holds: e.g., $< a, b > \cup < b, c > = < a, b, c >$ and $< a, b > \cup < \frac{1}{2}b, c > = < a, b, \frac{1}{2}b, c >$

Substitution of the corresponding finite vector space quantities into (2.7) yields the familiar property

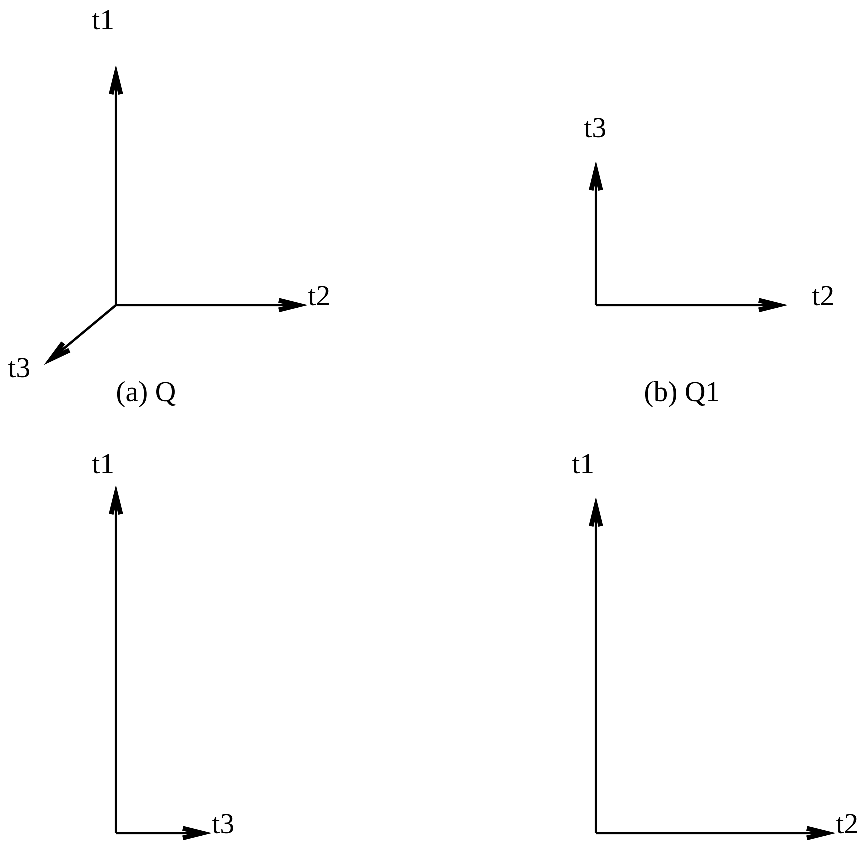
$$dim(S \oplus T) \leq dim(S) + dim(T)$$

which follows from the fact that

$$dim(S \oplus T) = dim(S) + dim(T) - dim(S \cap T) \tag{2.8}$$

Considering the definition of $d_\alpha$, it may come as no surprise that results analogous to (2.8) hold when the vectors of the collection of interest (substar or star) are mutually orthogonal. Additionally, the essential differences between $d_\alpha$ and $dim$ are easily examined under these conditions. These differences and the effect on $d_\alpha$ of the conditions themselves are explored in the remaining propositions of this section (see Appendix B for these proofs).

Some new notation will simplify the presentation. As before, $P$ will denote an arbitrary star. However, for the rest of this section $Q$ and $T$ will denote a star and substar each containing only mutually orthogonal vectors. That is, $t =< t_1, ..., t_k >$ and $Q =< q_1, ..., q_\ell >$ are such that $t_i^T t_j = 0$ and $q_i^T q_j = 0$ whenever $i \neq j$. Further, if the length of a vector $v$ is denoted by $||v|| = (v^T v)^{\frac{1}{2}}$, then $T$ and $Q$ are also such that $||t_1|| \geq ||t_2|| \geq \cdots \geq ||t_k|| > 0$ and $||q_1|| \geq ||q_2|| \geq \cdots \geq ||q_\ell|| > 0$. These substars are $Q_j =< q_i : i = 1, ..., j-1, j+1, ..., \ell >$ for $j = 1, ..., \ell, T_j =< t_i : i = 1, .; .., j-1, j+1..., k >>$ for $j = 1, ..., k$, and $Q_I =< q_i : i \epsilon I >, Q_J =< Lq_i : i \epsilon J >$ where $I$ and $J$ are arbitrary subsets of $\{1, ..., \ell\}$. A picture may help clarify the notation. Figure 5(a) is a



(a) Q

(b) Q1

(c) Q2

(d) Q3

representation in $I\!\!R^3$ of $Q$, when $\ell = 3$. Figure 5(b), (c) and (d) represent $Q_1, Q_2$ and $Q_3$ respectively. These have been reoriented to lie in the plane of the page. These same three pictures also represent $Q_I(\text{or } \mathrm{Q_J})$ when $I$ (or $J$) equals {2, 3}, {1, 3} and {1, 2}, respectively. Similar figures could be drawn for $T$ and its substars.

With this notation, the essential differences between $d_\alpha$ and $dim$ are illustrated by two propositions. The first is directed at assessing the $d_\alpha$-effective dimension of stars with respect to themselves. The second is an assessment of the $d_\alpha$ of substars with respect to a fixed star.

Proposition 2.5: For $i > j > 1$ and for all $\alpha > 0$

$$d_\alpha(Q_i; Q_i) \geq d_\alpha(Q_j; Q_j) \tag{2.9}$$

and equality holds if, and only if, $||q_i|| = ||q_j||$.

There are two things to note here. First, since $\lambda_1$ appears as a standardizing factor in the denominator of $d_\alpha, i \text{ and } j$ are restricted to be greater than 1. If $j = 1$, the inequality could go either way. Second, (2.9) states that the star of Figure 5(d) has larger effective dimension than that of Figure 5(c). This corresponds to the intuitive notion that when the largest orthogonal directions, here the vertical $t_1$, of Figures 5(c) and (d) is fixed, the horizontal dimension is better determined by $Q_3$ than it is by $Q_2$. Of course with $dim$ no such distinction would be made. Both Figures 5(c) and 5(d) would yield $dim = 2$.

As a corollary to Proposition 2.5, it may easily be seen that by replacing $q_i$ in $Q$ by $c \cdot q_i$, where $i > 1$ and $c$ is some positive scalar such that $c \cdot ||q_i|| \leq ||q_1||$, $d_\alpha(Q; Q)$ will be decreased or increased as $c$ is less than or greater than 1. Again, $dim$ will be completely unresponsive to such changes provided $c > 0$.

Similar results hold for substars of $T$. However, since the standardization $[\lambda_1(P)]^{-\alpha}$ of $d_\alpha$ is based on a larger common star $P$, there need be no concern about inclusion or exclusion of $t_1$. Corresponding to Proposition 2.5, we have the following.

Proposition 2.6: For a star $P$, such that $T \overset{\star}{\subseteq} P$, for $i > j > 0$ and for all $\alpha > 0$

$$d_\alpha(T_i; P) \geq d_\alpha(T_j; P) \tag{2.10}$$

and equality holds if, and only if, $||t_i|| = ||t_j||$.

The conditions under which an equality similar to (2.8) holds for $d_\alpha$ are now described. As before, propositions for two cases are given. The first has $d_\alpha$ operating on particular stars with respect to themselves, or (star, star) pairs, and the second shows $d_\alpha$'s operation on certain substars of a star $P$, or (substar, $P$) pairs. For the first result, recall the sets $I, J \subset \{1, ..., \ell\}$ and the corresponding substars $Q_I, Q_J \overset{\star}{\subseteq} Q$. Further, let $M_1 = \max_{i \epsilon I} ||q_i||$ and $M_2 = \max_{i \epsilon J} ||q_i||$ denote the maximum attained by $||q_i||$ for all $i \epsilon I$, and for all $i \epsilon J$ respectively. In Appendix B, the following result is shown to hold.

Proposition 2.7: If $M_1 = M_2$ and either (i) $I \cap J = \emptyset$, (ii) $\max_{i \epsilon I \cap J} ||q_i|| = M_1$, then for all $\alpha > 0$

$$\begin{aligned} & d_\alpha(Q_I \overset{\star}{\cup} Q_J; Q_I \overset{\star}{\cup} Q_J) \\ = {} & d_\alpha(Q_I; Q_I) + d_\alpha(Q_J; Q_J) - d_\alpha(Q_I \overset{\star}{\cap} Q_J; Q_I \overset{\star}{\cap} Q_J). \end{aligned} \tag{2.11}$$

The corresponding result for substars with respect to a common star is more simply written.

Proposition 2.8: If $T \overset{\star}{\subseteq} P$, and $S_1$ and $S_2$ are arbitrary substars of $T$, then for all $\alpha > 0$

$$d_\alpha(s_1 \overset{\star}{\cup} S_2; P) = d_\alpha(S_1; P) + d_\alpha(S_2; P) - d_\alpha(S_1 \overset{\star}{\cap} S_2; P). \tag{2.12}$$

As with the earlier result (2.9), care must be taken when the star part of the (substar, star) pair is changing. This is entirely due to the normalizing factor $\lambda_1^{-\alpha}$ in $d_\alpha$. The conditions of Proposition 2.7 are required to keep this factor constant.

Finally, note that while $d_\alpha(S:P) \leq dim(\mathcal{S})$, in general no stronger relationship exists. If $S_1$ and $S_2$ are substars of $P$ and $\mathcal{S}_1$ and $\mathcal{S}_2$ are the corresponding vector spaces, then $d_1(S_1;P) \leq d_1(S_2;P)$ does not imply anything about the relation ship between $dim(\mathcal{S}_1)$ and $dim(\mathcal{S}_2)$. Similarly, $d_1(S_1;P) \leq d_1(S_2;P)$ in general implies nothing about the relationship between $d_2(S_1;P)$ and $d_2(S_2;P)$. If additionally $S_1 \overset{\star}{\subseteq} S_2$, then Proposition 2.1 applies and $d_2(S_1;P) \leq d_2(S_2;P)$, otherwise the inequality does not necessarily hold.
Because of such considerations, for those pairs $(S;P)$ where no value of $\alpha$ is preferred over any other, the *dimension indices* $\eta_i = \lambda_i(S)/\lambda_1(P)$ will be used to summarize the dimensionality information $(d_\alpha(S;P) = \sum \eta_i^\alpha)$.

# 3. Some Applications

The theory of the last section will now be applied to a number of problem areas which deal with the linear model of (1.1). For these applications stars will be formed by taking the column (or row) vectors of some given matrix. For example, if $A = (A_1, ..., A_n)$ is an $m \times n$ matrix, then a star $P$ might be formed by the columns of $A$, that is $P =< A_1, ..., A_n >$. With some abuse of notation, $\lambda_i(A)$ will also be used to denote $\lambda_1(P)$.

To begin, the close relation between $d_\alpha(S;P)$ and the numerical determination of the rank of a matrix is briefly illustrated. Other problems to be discussed are minimax ridge-A estimates, collinearity and least squares influential data diagnostics.

## 3.1 Numerical Ranks

Mathematically, the rank of a matrix A is easily determined. It equals the number of non-zero singular values of A, or equivalently, the dimension of the row or column-space of A. [8] Computationally, however, the determination of the rank must take into account the precision of the machine.

The condition number of a matrix A, $\lambda_{max}(A)/\lambda_{min}(A)$, has long been used to numerically determine whether or not A is of full rank (Wilkinson [1965]). More recently, Chambers [1977] has defined the numerical rank of A, for a given $\epsilon > 0$, to be $r_\epsilon = r$ if $\lambda_r(A) \geq \epsilon \cdot \lambda_1(A) > \lambda_{r+1}(A)$. the similarity to $d_\alpha(S;P)$ is clear. If $\eta_i$ is the $i$-th dimension index, then $r_\epsilon = r$ if

[8] Hence it is always greater than the corresponding $d_\alpha(S;P)$ where $S = P =< A_1, ..., A_n >$ if A $= (A_1, ..., A_n)$

$$\eta_r \geq \epsilon > \eta_{r+1}$$

Equivalently,

$$r_\epsilon(P;P) = \sum I_{[\epsilon,1]}(\eta_i)$$

where $I_{\{\cdot\}}(\cdot)$ is the indicator function for the set $\{\cdot\}$. If the $\eta_i$'s were plotted against their indices, $i$, the numerical rank $r_\epsilon$ would count the number of points above some cutoff $\epsilon > 0$, whereas the $d_\alpha$-effective dimension would sum the $\alpha^{\text{th}}$ power of the heights of the points. Note also that the numerical rank operates only on a $(P;P)$ pair (or $A$).

## 3.2 A Minimax Result

Given the model (1.1), assume that the errors, $e_i$, are independent and identically distributed as $N(0, \sigma^2)$ with known $\sigma^2 > 0$. An alternative to the least squares estimator, $\hat{b}$, is the adaptive Ridge-A estimator (Thisted [1982]) given by (when shrinking $\hat{b}$ to 0)

$$\hat{b}_a = (X^T X + k^2 I_m)^{-1} X^T y \tag{3.2.1}$$

where $k^2 = a\sigma^2/(\hat{b}^T V D_W V^T \hat{b})$. Here $a > 0, D_W$ equals some diagonal matrix of weights, and $V$ is the matrix having the eigenvectors of $X^T X$ as columns. Further, suppose that the expected loss of an estimator $\delta$ of $b$ can be given by $E\left[(\delta\ b)^T L(\delta - b)\right]$ for some positive semi-definite matrix $L$. Thisted [1982] has shown that the $d_\alpha$-effective dimension plays much the same role in this setting as does the dimension in the well-known James-Stein result.

The relevant star for this result is $P- < p_1, ..., p_m >$ where $p_i$ is the $i$-th column vector of the matrix

$$D_W^{\frac{1}{2}} V^T Var(\hat{b}) L Var(\hat{b}) V D_W^{\frac{1}{2}} \tag{3.2.2}$$

The following result is proved by Thisted and Morris [1980] and may be found in Thisted [1982].

Proposition 3.1: For suitable choices of $a \geq 0$, Ridge-A estimators given by (3.2.1), are minimax with respect to the above loss function, if and only if $d_1(P;P) > 2$.

Note that whereas the James-Stein result required $dim > 2$, the above result requires a $s_\alpha > 2$. For this reason, Thisted [1982] has called $d_1(P;P)$ *the* effective dimension and denoted it as ED. Since $d_\alpha$ shares many geometric properties with $dim$ for values of $\alpha$ other than $\alpha = 1$, the term "$d_\alpha$-effective dimension" is preferred here.

## 3.3 Collinearity

In a collinearity analysis [9] it helps to distinguish between those procedures used to detect the presence of collinearity and those used to ascertain its effect on the problem of interest. In this subsection both kinds of procedures are examined.

For detection, consider the star given by $P_0 =< X_1, ..., X_m >$ in $I\!R^n$. Given that the $X_i$'s are in a structurally interpretable form (see Belsley [1984] and Belsley and Oldford [1984] for discussion), collinearity is judged to be present if at least one of the dimension indices, $\eta_i$, of $d_1(P_0; P_0)$ is small, [10] and inestimability occurs if at least one is zero. Thus, collinearity is present if at least one orthogonal direction is not well determined. Further, the $\eta_i$'s, whose inverses are called "condition" indices by Belsley, Kuh and Welsch [1980], are used to assess the extent of the collinearity. The greater the number of poorly determined orthogonal directions of $P_0$, the more extensive is the collinearity.

Consider now the effect on the Ridge-A estimator. Thisted [1980, 1982] has suggested that $d_1$-effective dimension o Proposition 3.1 be used to assess the effect of collinearity o the minimax property of the Ridge-A estimators. As will be demonstrated, the statistics itself is not at all related to the presence or absence of collinearity. However, since Ridge-A estimators are often suggested in place of the least-squares estimator when collinearity is present, it is of interest in this case to see when minimaxity obtains.

In particular, ;et $P_1$ and $P_2$ be the stars having as vectors the columns of the matrix of (3.2.2) with $D_W = I$ when $L = I$ and when $L = X^T X$, respectively. it can be shown that

$$\begin{aligned} d_1(P_1; P_1) &= \lambda_1(P_1)^{-1} \sum \lambda_i(P_1) \\ &= \lambda_m(X)^4 \sum \lambda_i(X)^4 \end{aligned} \tag{3.3.1}$$

and

$$\begin{aligned} d_1(P_2; P_2) &= \lambda_1(P_2)^{-1} \sum \lambda_i(P_2) \\ &= \lambda_m(X)^2 \sum \lambda_i(X)^{-2} \end{aligned} \tag{3.3.2}$$

where $\lambda_1(X) \geq ... \geq \lambda_m(X) > 0$, are the singular values of the X-matrix. Thisted [1980, 1982] has called $d_1(P_1; P_1)$ and $d_2(P_2; P_2)$ the multicollinearity index (mci) and the predictive multicollinearity index (pmci), respectively. From Proposition 3.1, each quantity is relevant to the minimaxity of a particular $(D_W = I)$ Ridge-A estimator, first when the expected loss is that of the mean-square-error of the estimator $(l = I)$ and second when the expected loss is that of the mean-square-error of the predicted response at the observed $X(L = X^T X)$. Values of mci or pmci less than two indicate that the corresponding minimax

[9] Recently, formal definitions have been proposed by Gunst [1984], and by Belsley and Oldford [1984].

[10] Based on experimental evidence, Belsley, Kuh and Welsch [1980] suggest that those $\eta_i$'s less than 0.033 be regarded as small

property is lost.

That mci and pmci bear no relationship to the presence or absence of collinearity, as assessed by the dimension indices $\eta_i$ of $d_1(P_0; P_0)$, is easily demonstrated by an example. Let $X$ be of full mathematical rank $m = 4$ and denote by $\Lambda$ the vector of ordered singular values of $X$, written as $\Lambda = (\lambda_1(X), ..., \lambda_4(X))$. Now consider the following three possibilities for $\Lambda$

(i) $\Lambda_1 = (1, 1, 1, \epsilon)$

(ii) $\Lambda_2 = (1, 1, \epsilon, \epsilon)$

(iii) $\Lambda_1 3 = (1, \epsilon, \epsilon, \epsilon)$

where $0 < \epsilon < 1$. Corresponding to each case are the values of $d_1(P_1; P_1)$ and $d_1(P_2; P_2)$,

(i) $(1 + 3\epsilon^4)$ and $(1 + 3\epsilon^2)$

(ii) $(2 + 2\epsilon^4)$ and $(2 + 2\epsilon^2)$

(iii) $(3 + \epsilon^4)$ and $(3 + \epsilon^2)$

Suppose first that $\epsilon = 1/5$. In all cases, the condition number of the X-matrix which results is five and collinearity is not likely to be judged present. However, the values of $d_1(P_1; P_1)$ are (i) 1.0048, (ii) 2.0032, and (iii) 3.0018 giving minimaxity of the Ridge-A estimator in the last two cases but not in the first. Now suppose that $\epsilon = 10^{-5}$ yielding 100, 000 as the condition number of X. Most likely, collinearity will be judged to be present. But the minimaxity or not of the Ridge-A estimator remains the same in each case as when $\epsilon = 1/5$. Indeed, when there are three out of four mutually orthogonal linear combinations of the parameters which are very nearly inestimable, as in case (iii), the minimax property of the estimator is assured, whereas in the case (i) of least extensive collinearity the minimaxity is lost.

Although mci and pmci and their dimension indices are of little uses for the general diagnosis of collinearity, they do provide interesting geometric information about the minimaxity of the Ridge-A estimator. The Ridge-A estimator (3.2.1) can be thought of as an estimator which shrinks the least squares estimates toward zero (or some other specified point). This shrinkage is done selectively, as Smith and Campbell [1980] point out, shrinking most those parameter estimates having the greatest variance. Letting $\hat{\gamma} = V^T \hat{b}$ with $\hat{b}$ and $V$ in (3.2.1), then the components of $\hat{\gamma}$ have variances equal to $\sigma^2 \lambda^{-2}(X)$. In the above examples, these correspond to the following vectors of variances,

(i) $(\sigma^2, \sigma^2, \sigma^2, \sigma^2/\epsilon^2)$

(ii) $(\sigma^2, \sigma^2, \sigma^2/\epsilon^2, \sigma^2/\epsilon^2)$

(iii) $(\sigma^2, \sigma^2/\epsilon^2, \sigma^2/\epsilon^2, \sigma^2/\epsilon^2)$

For small $\epsilon > 0$, the variance $\sigma^2$ is negligible when compared to $\sigma^2/\epsilon^2$. The first case, (i), this means that there is essentially only one least squares estimate, $\hat{\gamma}_4$ say, with non-negligible variance, or equivalently, there is effectively only one random quantity to shrink. Not until $\epsilon^2$ (or $\epsilon^4$) is greater that 1/3 does $d_1(P_2; P_2)$(or $d_1(P_1; P_1))$ produce a value larger than two. In terms of variance, as long as $Var(\hat{\gamma}_4) \geq 9Var(\hat{\gamma}_i)$ for $i \neq 4$, by comparison the $\hat{\gamma}_i$ for $i \neq 4$ act as fixed quantities. This interpretation makes sense of the fact that in case (i) large values of $\epsilon$, which might properly be ignored by a collinearity detection diagnosis, cannot be tolerated by the minimaxity property. Similar remarks and interpretations apply to the cases (ii) and (iii).

It has been demonstrated that there is a role to be played by the $d_1$-effective dimension in the detection of collinearity and in the determination of the minimaxity of the ridge-A estimator. In the latter its role is critical as demonstrated by Proposition 3.1, whereas in the former the dimension indices of $d_1(P_0; P_0)$ are the important quantities. Further, in each instance a geometric interpretation is available by considering the dimension indices of certain stars.

Beyond the geometrical interpretation, it is tempting to interpret the $d_1$-effective dimension as the effective number of explanatory variables in the regression model (1.1), or in the case of the ridge-A estimator as the effective number of least-squares parameter estimates which might reasonably be regarded as random quantities for the purposes of minimaxity. Certainly when collinearity is present one often feels, as Thisted[1980] has pointed out, that there are in fact fewer variables available than are given by the rank of the X-matrix. The cases (i) to (ii) above, examined with respect to variances, also lend support to this thesis for the ridge-A estimator. While this interpretation is a tantalizing one, it is not clear that is should be adopted. For instance, except for Proposition 3.1, there does not seem to be any particular reason for preferring $d_1$ over any other $d_\alpha$ [11]. This being the case, it must be noted that in general for two stars $P$ and $Q$ it is possible to have $d_1(P; P) > d_1(Q; Q)$ and also $d_2(P; P) < d_2(Q; Q)$. It may be the case that only in special circumstances, such as those given in Proposition 3.1, will such an interpretation be permitted. Nevertheless, as will be seen in the remaining subsection this kind of interpretation presents itself again.

### 3.4 Influential Observations in Least Squares Regression

Let $\hat{e}$ be the residual from the least squares fir $\hat{y}$ of (1.1). If it is felt that $k$ observations, without loss of generality the last $k$, may be suspect in this fit, then $k$ parameters could be added to the model yielding

$$y = Xb + \delta + e \tag{3.4.1}$$

where $\delta = \binom{0}{b_2}$ is an $n \times 1$ vector and $b_2$ is $k \times 1$. Many peculiarities of the data and fit can be incorporated in this manner (e.g., see Andrews [1971]). Let $e^*$ denote the least squares residual for the model of (3.4.1). The influence these $k$ observations have on the original fit can be measured by the difference in

[11] Proposition 2.3 admits at least $\alpha = 1$ and $\alpha = 2$

fits $y^* - \hat{y} = \hat{e} - e^* = \Delta e$, say.

Assuming that X is of full rank let

$$R = (I - X(X^TX)^{-1}X^T) = \left[ \begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array} \right] \qquad (3.4.2)$$

where $R_{11}$ is $(n-k) \times (n-k),\ R_{22}$ is $k \times k$ and $R_{12} = r_{21}^T$ is $(n-k) \times k$. Now assume also that the $k$ observations are such that $R_{22}$ is of rank $k$. The following result is proved in Appendix C.

Proposition 3.2: Given $e$ and $e^*$ as above and $R$ as in (3.4.2), then

$$\Delta e = \begin{pmatrix} R_{12} \\ R_{22} \end{pmatrix} R_{22}^{-1}(R_{12}^T, R_{22})\hat{e}. \qquad (3.4.3)$$

Denote the $i^{\text{th}}$ column of $R$ by the vector $a_i$ so that $R = (a_1, ..., a_n)$, and let $R_2$ denote $(R_{21}, R_{22})^T = (a_{n-k+1}, ..., a_n)$. Associated with each matrix is a star, respectively P $\equiv < a_1, ..., a_n >$ and S $\equiv < a_{n-k+1}, ..., a_n >$ with S $\overset{\star}{\subseteq}$ P, and a subspace of $I\!R^n$, respectively the error space $\mathcal{P} \equiv$ span $(a_1, ..., a_n)$ and $\mathcal{S} \equiv$ span $(a_{n-k+1}, ..., a_n)$ with $\mathcal{S} \subseteq \mathcal{P}$.

Since $R_{22} = R_2^T R_2$, Proposition 3.2 shows that the change in fit, $\Delta e$, equals the orthogonal projection of $\hat{e}$ onto the subspace $\mathcal{S}$. Thus the influence of the last $k$ observations is expressible as a perturbation in the direction of the vectors of $S$. The size of the perturbation will depend upon the actual vectors of $S$ and the orientation of $S$ relative to $\hat{e}$. These two components, $S$ and its orientation to $\hat{e}$, provide simple starting places to generate summaries of the information in $\Delta e$. As will be shown, summaries of one or both of these components are quite common.

The structure of $S$ may be summarized by the dimension indices of $d_\alpha(S;P)$ and the principal direction vectors of $S$ given the eigenvectors of matrix $R_2R_2^T$. These two sets of summary statistics are sufficient to describe the "leverage" [12] or "potential influence" any group of $k$ observations may have on the determination of the least-squares fit of the remaining $n-k$ observations. To show this denote the dimension indices by $eta_1 \geq ...geq\eta_k > 0$ and, since $\eta_i = \lambda_i(S)$, the singular value decomposition (svd) of $R_2$ by $R_2 = UHV^T$ where $H \equiv$ diag$(\eta_1...\eta_k)$, and $U^TU = V^TV = VV6t = I_k$ so that $U$ contains the principal direction vectors of $S$. Letting $U^T = (U_1^T, U_2^T)$ where $U_2$ has $k$ rows, it is easily shown that

$$U_2^TU_2 = H^2$$

and

$$U_1^TU_1 = I_k - H^2. \qquad (3.4.4)$$

[12] This differs slightly from that of Hoaglin and Welsch [1978] where the leverage of a single observation is described in terms of the potential influence it has on its own fit.

Further, (3.4.3) may be rewritten as

$$\Delta e = U(U^T \hat{e}). \tag{3.4.5}$$

Together (3.4.4) and (3.4.5) indicate that the fit will be perturbed only at the $k$ suspect observations if, and only if, $\eta_1 = ... = \eta_k = 1$ regardless of $\hat{e}$. If some of these dimension indices are small, then the fit of the other observations may be perturbed. The extent of the perturbation will also depend upon $\hat{e}$. If some of these dimension indices are small, then the fit of the other observations may be perturbed. The extent of the perturbation will also depend upon $\hat{e}$. Therefore examination of $U$ and the dimension indices alone can determine only the potential influence of the group. Hence the word "leverage". [13] Note that, as was the case with collinearity, to determine the presence of high leverage it is best to examine the entire set of dimension indices rather than any single unidimensional summary like $d_\alpha(S; P)$.

Now consider summarizing the orientation of $S$ with respect to $\hat{e}$. The simplest, and most common (see e.g., Andrews [1971], Dempster and Gasko-Green [1981]), way to proceed is to ignore the particulars of $S$ and summarize instead the orientation of $\hat{e}$ to the subspace $\mathcal{S}$. A summary of this orientation is given by the angle, $\theta$, that $\hat{e}$ makes with $\mathcal{S}$. The closer $\hat{e}$ lies to $\mathcal{S}S$, the smaller $\theta \epsilon [0, \pi/2]$ is, and the greater the influence of the $k$ observations. This is equivalent to the extra-sum-of-squares principle and has been advocated in the literature to test for outliers by Gentleman [1980] and Draper and John [1981]. However, examination of the second factor of (3.4.5) indicates that the orientation of the principal directions of $S$ to $\hat{e}$ contain important information about the change in fit. for example, suppose that $\eta_k < 1$ and all other $\eta_i$'s equal 1. Only the last column of $U_1$ is non-zero and the fit of the first $n-k$ observations will be perturbed if, and only if, $\hat{e}$ is orthogonal to the $k^{\text{th}}$ column vector of $U$. The amount of the perturbation will increase as the acute angle between this vector and $\hat{e}$ decreases, regardless of the value taken by $\theta$. The value of $\theta$ serves only as a lower bound for this angle. The $k$ angles $\theta_1, ..., \theta_k$, say, between $\hat{e}$ and the principal directions of $S$, or equivalently the scaled vector of cosines $U^T \hat{e}$, provide more detailed information on the orientation of $S$ to $\hat{e}$ and hence on the actual influence.

Separate examination of the structure of $S$ in $P$ and the orientation of $S$ to $\hat{e}$, suggests that the $n$-dimensional information contained in $\Delta e$ can be reasonably summarized by the two $k$-dimensional statistics $(\eta_1, ..., \eta_k)$ and $(\theta_1, ..., \theta_k)$. These statistics correspond roughly to the "leverage" and "outlyingness" components of the influence of a group of $k$ observations. respectively. However, if many groups of $k$ observations are to be examined and $k$ is large enough, then practicality will require still lower dimensional summaries. Two obvious unidimensional statistics which maintain the above distinction are $d_\alpha(S; P)$ and $\theta$. Any further reduction would sacrifice this distinction.

A number of single unidimensional statistics have been suggested in the literature. Two of them, Cook's [1977] statistic and the Andrews-Pregibon [1978] statistic are now described geometrically. These

[13] When $k = 1$ $\eta_1^2 = ||a_n||^2 = ||a_n||^2 = 1 - h_n$ where $h_n$ is the $n^{\text{th}}$ diagonal element of the "hat matrix" $H = I - R$. Thus for the one-at-a-time case this notation of leverage corresponds directly to the self-influence one given by Hoaglin and Welsch[1978].

have also been discussed recently by Dempster and Gasko-Green [1981], and by Draper and John [1981]. As will be shown, each captures the structure of $S$ (leverage) and its orientation to $\hat{e}$ (outlyingness) in different ways.

Let $Q$ be the extra sum of squares, $\Delta e^T \Delta e$, due to fitting $b_2$. The angle, $\theta$, between $\hat{e}$ and the span $a_{n-k+1}, ..., a_n)$ may be measured by,

$$\cot^2\theta = \frac{Q}{\hat{e}^T \hat{e} - Q},$$

or equivalently,

$$\sin^2\theta = \frac{\hat{e}^T \hat{e} - Q}{\hat{e}^T \hat{e}},$$

One-at-a-time statistics combine this outlier information with the leverage information, $||a_i||^2$, in simple ways (see Dempster and Gasko-Green [1981]).

For $k$-at-a-time diagnostics, the Andrews-Pregibon (AP) and Cook (C) statistics may be expressed, up to a multiplicative constant, as follows,

$$\text{AP} = \text{sim}^2\theta \cdot \det(R_{22}) = \sin^2\theta \prod_{i=1}^{k} \eta_i^2$$

and

$$\text{C} = \cos^2\theta \left( \frac{\hat{e}^T R_2 R_{22}^{-2} R_2^T \hat{e}}{Q} - 1 \right) \tag{3.4.6}$$

Small values of AP and large values of C indicate influential groups.

Since a single unidimensional statistic is used in either case, care must be taken when combining the two sources of information. For example, having the "outlier" part, $\theta$, enter each statistic multiplicatively through $\sin^2\theta$ and $\cos^2\theta$ has certain drawbacks. In particular, if the fit including the $i^{\text{th}}$ observation is exactly the same as that excluding it, then $\cos^2\theta = 0$ and AP may be small if $||a_i||^2 = (1 - h_i)$ is small. Draper and John [1981] recommend C over AP for this reason alone. If $\cos^2\theta = 1$, then removal of the $i^{\text{th}}$ observation gives a perfect fit to the remaining observations. C does not detect this. This has been pointed out by Dempster and Gasko-Green [1981].

While the AP statistic factors simply into an "outlier" part, $\sin^2\theta$, representing the orientation of $\hat{e}$ to $\mathcal{S}$, and a "leverage" part, $\prod_{i=1}^{k} \eta_i^2$, summarizing the structure of $S$, this is not the case for the C statistic. Its second factor in (3.4.6) is not a simple "leverage" component. Rather, the second factor summarizes both the structure of $S$ and the orientation of $S$ to $\hat{e}$. This may be seen by reexpressing this factor (+1) as

$$\frac{1}{\cos^2\theta} \cdot \sum_{i=1}^{k} \left( \frac{\cos\theta_i}{\eta_i} \right)^2 \tag{3.4.7}$$

where $\theta_i$ is the angle between $\hat{e}$ and the $i$'th principal direction of $S$. Clearly, both the structure of $S$ and its orientation to $\hat{e}$ are captured by (3.4.7). Indeed (3.4.6) can now be rewritten as

$$\mathrm{C} = \sum_{i=1}^{k} \left( \frac{\cos\theta_i}{\eta_i} \right)^2 - \cos^2\theta. \tag{3.4.8}$$

The closer $\hat{e}$ lies to principal direction of $S$ which has a small dimension index the large is C. This is a fundamental difference between C and AP. The Andrews-Pregibon statistic uses the orientation of $\mathcal{S}$ to $\hat{e}$ whereas the Cook statistic uses the orientation of $S$ to $\hat{e}$ with that of $\mathcal{S}$ to $\hat{e}$ appearing more as a correction factor in (3.4.8).

Analogous to the case of collinearity, it is tempting here to interpret some $d_\alpha$-effective dimension as a measure of the effective number of observations determining the least-squares fit. Certainly when few observations are overly influential it will be found that the error vector $\hat{e}$ lies close to some smaller dimensional space, $\mathcal{S}$, and/or the vectors of $R$ which generate $\mathcal{S}$ form a star $S$ having one or more small dimension indices. Further, if $\hat{e}$ lies close to a principal direction of $S$ associated with a small dimension index then, as in the Cook statistic, the influence of these points will be greater still. Unfortunately there are many such substars, $S$, to consider so that until there exists a result similar to Proposition 3.1 which would specify a particular $S \overset{\star}{\subseteq} P$ and an $\alpha > 0$ such an interpretation is unavailable.

# 4. Concluding Remarks

As always, a number of issues are left open. Among those that can be stated precisely are the following strictly mathematical ones. Does (2.6) of Proposition 2.3 hold for values of $\alpha$ other than 1 or 2? Further, what, if any, functions of $(S; P)$ satisfy it many properties? It can be shown for example (Oldford [1983]) that any such function operating on $(P; P)$ must be a function of the singular values $\lambda_i(P)$. Statistically, one can speculate on other possible applications of $d_\alpha$ and the dimension indices. Huber [1981, p. 160] has called $h_i^{-1}$ "the equivalent number of observations entering into the determination of $\hat{y}_i$" and Mosteller and Tukey [1977, p. 348] have referred to the sum of weights from a robust regression as indicative of the "equivalent number of equivariable observations". Can these vague notions of "equivalent" or "effective" number of observations being used be formalized through some $d_\alpha$-effective dimension as was done for the number of least-squares parameter estimates by Proposition 3.1? Section 3.4 lends support to this possibility for least-squares regression.

In conclusions, then, a mathematical theory has been presented which supplements the theory of vector spaces by focusing on finite collections of vectors. When applied to one area of statistical interest, namely, the linear model (1.1), it has been found to describe a variety of practical statistical concerns, some of which previously were not described within an $n$-dimensional framework. It remains to be seen what other areas of statistical interest might benefit from this theory.

## Appendix A: Propositions 2.1 - 2.4

Some notation and preliminary results are necessary before proceeding with the proofs of the propositions.

For any star $P =< p_1, ..., p_N >$ and $m \times N$ matrix $A = (p_1, ..., p_N)$ may be formed having ordered singular values denoted by $\lambda_i(A)$, or equivalently by $\lambda_i(P)$ for $i = 1, ..., min(m, N)$. Now, if S $\overset{\star}{\subseteq}$ P, it will be assumed without loss of generality that the vectors of S have been indexed and sufficient zero vectors added to either S or P so that the matrix formed from S may be written as A $\cdot \mathrm{D_c}$ (S) where $\mathrm{D_c}$ (S) is an $N \times N$ diagonal matrix with diagonal elements $\mathrm{c_1(S), ..., c_N(S)}$. Here the scalars $\mathrm{c_1(S), ..., c_N(S)}$ are constants in [0,1] peculiar to the pari (S;P).

Three lemmas are now given on singular values of certain matrices. The first two may be found in the literature and hence are presented without proof.

Lemma A.1: let A me an $m \times N$ matrix and B the $m \times (N-1)$ matrix resulting from deleting any column of A. Then for

$$m \geq N \ \lambda_1(\mathrm{A}) \geq \lambda_1(\mathrm{B}) \geq \lambda_2(\mathrm{A}) \geq, ..., \geq \lambda_{N-1}(\mathrm{B}) \geq \lambda_N(\mathrm{A}) \geq 0$$

and for

$$m < N \ \lambda_1(\mathrm{A}) \geq \lambda_1(\mathrm{B}) \geq, ..., \geq \lambda_m(\mathrm{A}) \geq \lambda_m(\mathrm{B}) \geq 0$$

Proof: See e.g., Lawson and Hanson [1974].

Lemma A.2: Let A, B and C = (A + B) be $m \times N$ matrices. Then, for $k = 1, ..., min(m, N)$

$$\sum_{i=1}^{k} \lambda_i(\mathrm{C}) \leq \sum_{i=1}^{k} \lambda_i(\mathrm{A}) + \sum_{i=1}^{k} \lambda_i(\mathrm{B}).$$

Proof: Ky Fan [1951].

Lemma A.3: Let A = $(p_1, ..., p_N)$, B = $(c_1 p_1, ..., c_N p_N)$ where $c_i \, \epsilon$ [0,1] for $i = 1, ..., N$, then

$$\lambda_i(\mathrm{A}) \geq \lambda_i(\mathrm{B}) \geq 0 \quad i = 1, ..., min(m, N).$$

Proof: For $i = 1, ..., min(m, N)$ and any $2N \times 2N$ orthogonal matrix $\Gamma$

$$\lambda_i n(\mathrm{A}) = \lambda_i((p_1, 0, p_2, 0, ..., p_N, 0) \cdot \Gamma).$$

Now take $\Gamma$ to be the block diagonal matrix diag$(\Gamma_1, ..., \Gamma_N)$ where

$$\Gamma_i = \begin{pmatrix} c_i & (1-c_i^2)^{\frac{1}{2}} \\ (1-c_i^2)^{\frac{1}{2}} & c_i \end{pmatrix} \qquad i = 1, ..., N.$$

Then by repeated application of Lemma A.1,

$$\begin{aligned} \lambda_i(\mathrm{A}) &= \lambda_i((c_1 p_1, (1-c_1^2)^{\frac{1}{2}} p_1, ..., C_N p_N, (1-c_1^2)^{\frac{1}{2}} p_n)) \\ &\geq \lambda_i(\mathrm{B}) \qquad i = 1, ..., min(m, N) \end{aligned}$$

which completes the proof.

We are now in a position to prove the propositions stated in Section 2.

Proposition 2.1 (proof): S $\overset{\star}{\subseteq}$ T $\overset{\star}{\subseteq}$ Q $\overset{\star}{\subseteq}$ P

$$\begin{aligned} d_\alpha(\mathrm{S};\mathrm{P}) &= \lambda^{-\alpha}(\mathrm{P}) \sum_i \lambda_i^\alpha(\mathrm{S}) \\ &\leq \lambda_i^\alpha(\mathrm{P}) \sum_i \lambda_i^\alpha(\mathrm{T}) \qquad \text{by Lemma A.3} \\ & d_\alpha(\mathrm{T};\mathrm{P}) \end{aligned}$$

Similarly for all $\alpha > 0$, $d_\alpha(\mathrm{T};\mathrm{P}) \leq d_\alpha(\mathrm{T};\mathrm{Q})$.

Proposition 2.2 (proof): S $\overset{\star}{\subseteq}$ P and S $\overset{\star}{\subseteq}$ Q. Let A = ($p_1, ..., p_N$) and B = ($q_1, ..., q_L$) be the matrices which correspond to P and Q, respectively. Then by Lemma A.3, for all $\alpha > 0$

$$\lambda_1^{-\alpha}(\mathrm{P} \cup \mathrm{Q}) \leq \lambda_1^{-\alpha}(\mathrm{P})$$

$$\text{and} \quad \lambda_1^{-\alpha}(\mathrm{P} \cup \mathrm{Q}) \leq \lambda_1^{-\alpha}(\mathrm{Q})$$

$$\Longrightarrow \quad \lambda_1^{-\alpha}(\mathrm{P} \cup \mathrm{Q}) \leq \tfrac{1}{2}[\lambda_1^{-\alpha}(\mathrm{P}) + \lambda_1^{-\alpha}(\mathrm{Q})]$$

$$\Longrightarrow \quad d_\alpha(\mathrm{S}; \mathrm{P} \cup \mathrm{Q}) \leq \tfrac{1}{2}[d_\alpha(\mathrm{S};\mathrm{P}) + d_\alpha(\mathrm{S};\mathrm{Q})].$$

Proposition 2.3 (proof): S $\cup$ T $\overset{\star}{\subseteq}$ P. Let C denote the matrix containing the vectors S and T have in common and A and B be the matrices of the remaining vectors in S and T respectively. Then,

$$\begin{aligned} d_\alpha(\mathrm{S} \cup \mathrm{T}; \mathrm{P}) &= \lambda_i^{-\alpha}(\mathrm{P}) \sum \lambda_i^\alpha(\mathrm{S} \cup \mathrm{T}) \\ &= \lambda_i^{-\alpha}(\mathrm{P}) \sum \lambda_i^\alpha((\mathrm{A}, \mathrm{C}, \mathrm{B})) \\ &\leq \lambda_i^{-\alpha}(\mathrm{P}) \sum \lambda_i^\alpha((\mathrm{A}, \mathrm{C}, \mathrm{B}, \mathrm{C})) \end{aligned}$$

by Lemma A.3. For $\alpha = 1$, application of Lemma A.2 yields the required result. The same lemma may be applied for $\alpha = 2$ after recognizing that

$$\lambda_i^2((\mathrm{A}, \mathrm{C}, \mathrm{B}, \mathrm{C})) = \lambda_i(\mathrm{AA}^\mathrm{T} = \mathrm{CC}^\mathrm{T} + \mathrm{BB}^\mathrm{T} + \mathrm{CC}^\mathrm{T}).$$

Proposition 2.4 (proof): This proof is entirely similar to that of Proposition 2.3, the only difference being that two possible different matrices $\mathrm{C}_1$ and $\mathrm{C}_2$ will be required from the single matrix C owing to the difference between $\overset{\star}{\cup}$ and $\cup$.

## Appendix B: Propositions 2.5 - 2.8

Proposition 2.5: For $i > j > 1$ and for all $\alpha > 0$

$$d_\alpha(\mathrm{Q}_i; \mathrm{Q}_i) \geq d_\alpha(\mathrm{Q}_j; \mathrm{Q}_j)$$

and equality holds if, and only if, $||q_i|| = ||q_j||$

Proof: Recall the definitions of $\mathrm{Q}_i$ and $\mathrm{Q}_j$. since $i > 1$ and $j > 1$,

$$d_\alpha(\mathrm{Q}_i; \mathrm{Q}_i) = ||q_1||^{-\alpha} \sum_{\substack{k \neq i \\ k \neq j}} ||q_k||^\alpha + ||q_1||^{-\alpha} ||q_j||^\alpha$$

and $$d_\alpha(\mathrm{Q}_j; \mathrm{Q}_j) = ||q_1||^{-\alpha} \sum_{\substack{k \neq i \\ k \neq j}} ||q_k||^\alpha + ||q_1||^{-\alpha} ||q_i||^\alpha$$

Since $||q_i|| \leq ||q_j||$ whenever $i > j$, the inequality is true fro all $\alpha = 0$ and further it is clear that equality holds if $||q_i|| = ||q_j||$.

Proposition 2.6: For a star p, such that $\mathrm{T} \overset{\star}{\subseteq} \mathrm{P}$ and T is defined as in Section 2, for $i > j > 0$ and for all $\alpha > 0$

$$d_\alpha(\mathrm{T}_i; \mathrm{P}) \geq d_\alpha(\mathrm{T}_j; \mathrm{P})$$

and equality holds if, and only if, $||t_i|| = ||t_j||$.

Proof: The proof is virtually identical to that presented above with $||q_1||$ replaced by $\lambda_1(\mathrm{P})$ and all other $q$'s by $t$'s. The major difference here is that $j$ may now equal 1 since the same star P is used as a reference.

Proposition 2.7: If $\mathrm{M}_1 = \mathrm{M}_2$ and

either (i) $\mathrm{I} \cap \mathrm{J} = \emptyset$

or (ii) $\max\limits_{i \epsilon \mathrm{I} \cap \mathrm{J}} ||q_i|| = \mathrm{M}_1$,

for all $\alpha > 0$

$$d_\alpha(\mathrm{Q_I} \overset{\star}{\cup} \mathrm{Q_J}; \mathrm{Q_I} \overset{\star}{\cup} \mathrm{Q_J})$$

$$= d_\alpha(\mathrm{Q_I}; \mathrm{Q_I}) + d_\alpha(\mathrm{Q_J}; \mathrm{Q_J}) - d_\alpha(\mathrm{Q_I} \overset{\star}{\cap} \mathrm{Q_J}; \mathrm{Q_I} \overset{\star}{\cap} \mathrm{Q_J}).$$

Proof: Recall the definitions of $\mathrm{Q_I}, \mathrm{Q_J}, \mathrm{M_1},$ and $\mathrm{M_2}$. Suppose $\mathrm{M_1} = \mathrm{M_2}$, then $\lambda_1(\mathrm{Q_I}) = \lambda_1(\mathrm{Q_J}) = \lambda_1(\mathrm{Q_I} \overset{\star}{\cup} \mathrm{Q_J}) = \mathrm{M_1}$ and hence

$$\begin{aligned} d_\alpha(\mathrm{Q_I}; \mathrm{Q_I}) &= \mathrm{M}_1^{-\alpha} \sum_{i\epsilon \mathrm{I}} ||q_i||^\alpha \\ d_\alpha(\mathrm{Q_J}; \mathrm{Q_J}) &= \mathrm{M}_1^{-\alpha} \sum_{i\epsilon \mathrm{J}} ||q_i||^\alpha \end{aligned}$$

and

$$\begin{aligned} d_\alpha(\mathrm{Q_I} \overset{\star}{\cup} \mathrm{Q_J}; \mathrm{Q_I} \overset{\star}{\cup} \mathrm{Q_J}) &= \mathrm{M}_1^{-\alpha} \sum_{i\epsilon \mathrm{I}\cup\mathrm{J}} ||q_i||^\alpha \\ &= \mathrm{M}_1^{-\alpha}(\sum_{i\epsilon \mathrm{I}} ||q_i||^\alpha + \sum_{i\epsilon \mathrm{J}} ||q_i||^\alpha - \sum_{i\epsilon \mathrm{I}\cap\mathrm{J}} ||q_i||^\alpha). \end{aligned}$$

Now if (i) $\mathrm{I} \cap \mathrm{J} = \emptyset$ then $d_\alpha(\mathrm{Q_I} \overset{\star}{\cap} \mathrm{Q_J}; \mathrm{Q_I} \overset{\star}{\cap} \mathrm{Q_J}) = 0$ and for all $\alpha > 0$

$$d_\alpha(\mathrm{Q_I} \overset{\star}{\cup} \mathrm{Q_J}; \mathrm{Q_I} \overset{\star}{\cup} \mathrm{Q_J}) = d_\alpha(\mathrm{Q_I}; \mathrm{Q_I}) + d_\alpha(\mathrm{Q_J}; \mathrm{Q_J})$$

as required. If instead (ii) $\max\limits_{i\epsilon \mathrm{I}\cap\mathrm{J}} ||q_i|| = \mathrm{M_1}$, then

$$d_\alpha(\mathrm{Q_I} \overset{\star}{\cap} \mathrm{Q_J}; \mathrm{Q_I} \overset{\star}{\cap} \mathrm{Q_J}) = \mathrm{M}_1^{-\alpha}{}_{i\epsilon \mathrm{I}\cap\mathrm{J}} ||q_i||^\alpha$$

and the required equality will again be satisfied for all $\alpha < 0$.

Proposition 2.8: If $\mathrm{T} \overset{\star}{\subseteq} \mathrm{P}$, and $\mathrm{S_1}$ and $\mathrm{S_2}$ are arbitrary substars of T, then for all $\alpha > 0$

$$d_\alpha(\mathrm{S_1} \overset{\star}{\cup} \mathrm{S_2}; \mathrm{P}) = d_\alpha(\mathrm{S_1}; \mathrm{P}) + d_\alpha(\mathrm{S_2}; \mathrm{P}) - d_\alpha(\mathrm{S_1} \overset{\star}{\cap} \mathrm{S_2}; \mathrm{P}).$$

Proof: Again recall the definition of T and further without loss of generality suppose that $\mathrm{S_1}$ and $\mathrm{S_2}$ are representable as $\mathrm{S_1} =< \mathrm{c_1(S_1)t_1}, ..., \mathrm{c_k(S_1)t_k} >$ and $\mathrm{S_2} =< \mathrm{c_1(S_2)t_1}, ..., \mathrm{c_k(S_2)t_k} >$ for some scalar constants $\mathrm{c}_i(\mathrm{S_1})$ and $\mathrm{c}_i(\mathrm{S_2})$ in [0,1] for $i = 1, ..., k$. Then $\mathrm{S_1} \overset{\star}{\cup} \mathrm{S_2} =< \mathrm{d_1t_1}, ..., \mathrm{d_kt_k} >$ and $\mathrm{S_1} \overset{\star}{\cap} \mathrm{S_2} =< \mathrm{e_1t_1}, ..., \mathrm{e_kt_k} >$ where $d_i = \max(\mathrm{c}_i(\mathrm{S_1}, \mathrm{c}_i(\mathrm{S_2}))$ and $e_i = \min(\mathrm{c}_i(\mathrm{S_1}, \mathrm{c}_i(\mathrm{S_2}))$ for $i = 1, ..., k$.

Thus, $d_i = (\mathrm{c}_i(\mathrm{S_1} + \mathrm{c}_i(\mathrm{S_2})) - e_i$ for all $i$ and it is easily seen that

$$\begin{aligned} d_\alpha(\mathrm{S}_1 \overset{\star}{\cup} \mathrm{S}_2; \mathrm{P}) &= \lambda_1^{-\alpha}(\mathrm{P}) \sum_i d_i^\alpha ||t_i||^\alpha \\ &= \lambda_1^{-\alpha}(\mathrm{P})[\sum_i \mathrm{c}_i(\mathrm{S}_1)^\alpha ||t_i||^\alpha + \sum_i \mathrm{c}_i(\mathrm{S}_2)^\alpha ||t_i||^\alpha \\ &\quad - \sum_i e_i^\alpha ||t_i||^\alpha] \\ &= d_\alpha(\mathrm{S}_1; \mathrm{P}) + d_\alpha(\mathrm{S}_2; \mathrm{P}) - d_\alpha(\mathrm{S}_1 \overset{\star}{\cap} \mathrm{S}_2; \mathrm{P}) \end{aligned}$$

as required.

## Appendix C. Proof of Proposition 3.2

Given $X$ and $R_{22}$ are of full rank, Draper and John [1981] show that

$$e^* = \begin{pmatrix} R_{11} - R_{12}R_{22}^{-1}R_{21} & 0 \\ 0 & 0 \end{pmatrix} y.$$

Thus $\Delta e = \hat{e} - e^*$

$$= \begin{pmatrix} R_{12}R_{22}^{-1}R_{21} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} y. \tag{C.1}$$

Since $R$ is idempotent, the following results hold

$$R_{11}^2 + R_{12}R_{21} = R_{11} \tag{C.2}$$

$$R_{21}R_{11} + R_{22}R_{21} = R_{21} \tag{C.3}$$

and $$R_{21}R_{12} + R_{22}^2 = R_{22}. \tag{C.4}$$

Multiplying (C.3) and (C.4) on the left by $R_{12}R_{22}^{-1}$ gives

$$R_{12}R_{22}^{-1}R_{21} = R_{12}R_{21}^{-1}R_{21}R_{11} + R_{12}R_{21} \tag{C.5}$$

and

$$R_{12} = R_{12}R_{22}^{-1}R_{21}R_{22} + R_{21}R_{22} \tag{C.6}$$

Now equations (C.3) to (C.6) are substituted into (C.1) to give

$$\begin{aligned}
\Delta e &= \begin{pmatrix} R_{12}R_{22}^{-1}R_{21}R_{11} + R_{12}R_{21} & R_{12}R_{22}^{-1}R_{21}R_{12} + R_{12}R_{21} \\ R_{21}R_{11} + R_{22}R_{21} & R_{21}R_{12} + R_{22}^2 \end{pmatrix} \; y. \\
&= \begin{pmatrix} R_{12}R_{22}^{-1}R_{21} & R_{12} \\ R_{12} & R_{22} \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \; y \\
&= \tbinom{R_{12}}{R_{22}} \; R_{22}^{-1} \; (R_{21}R_{22}) \; \hat{e}
\end{aligned}$$

which completes the proof of Proposition 3.2.

# References

Andrews, D.F. [1971], "Significance Tests Based on Residuals," *Biometrika*, 58, pp. 139-148.

Andrews, D.F. and Pregibon, D. [1978], "Finding the Outliers that Matter," *J.R.S.S.* B, 40, pp. 85-93.

Belsley, D.A. [1984], "Demeaning Conditioning Diagnostics Through Centering (with discussion)," *The American Statistician*, 38, p. 73-94.

Belsley, D.A., Kuh, E., and Welsch, R.E. [1980], *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons Inc., New York.

Belsley, D.A. and Oldford, R.W. [1984], "The general problem of ill conditioning in statistical analysis," *Proceedings of the American Statistical Association: Business and Economic Statistics Section*.

Chambers, John M. [1977], *Computational Methods for Data Analysis*," John Wiley and Sons Inc., New York.

Cook, R.D. [1977], "Detection of Influential Observations in Linear Regression," *Technometrics*, 19. pp. 15-18.

Dempster, A.P. and Gasko-Green, M. [1981], "New Tools for Residual Analysis," *Annals of Statistics*, 9, pp. 945-959.

Draper, N.R. and John, J.A. [1981], "Influential Observations and Outliers in Regression," *Technometrics*, 23, pp. 21-26.

Fan, Ky [1951], "Maximum Properties and Inequities for the Eigenvalues of Completely Continuous Operators," *Proceedings of the National Academy of Sciences (U.S.A.)*, 37, pp. 760-766.

Gentleman, J.F. [1980], "Finding the K Most Likely Outliers in Two-way Tables," *Technometrics*, 22, pp. 591-600.

Gunst, R. [1984], "Comment: Toward a Balanced Assessment of Collinearity Diagnostics," *The American Statistician*, 32, pp. 17-22.

Guttman, I. [1983], *Linear Models: An Introduction*, John Wiley and Sons Inc., New York.

Hoaglin, D.C. and Welsch, R.E [1978], "The Hat Matrix in regression and Anova," *American Statistician*, 32, pp. 17-22.

Huber, P.J. [1981] *Robust Statistics*, John Wiley and Sons Inc., New York.

Kruskal, W. [1968], "When are Gauss-Markov and least squares estimates identical? A coordinate-free approach." *Annals of Mathematical Statistics*, 39, pp. 70-75.

Lawson, C.L. and Hanson, R.J. [1974], *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, New Jersey.

Mosteller, F. and Tukey, J.W. [1977], *Data Analysis and Regression*, Addison-Wesley Publishing Co., Reading, Massachusetts.

Oldford, R.W. [1983], "Collinearity and Influential Observations as Dimensionality Problems," *Proceedings of the American Statistical Association: Statistical Computing Section*, pp. 153-158.

Scheffé, H. [1959], *The Analysis of Variance*, John Wiley and Sons Inc., New York.

Seber, G.A.F. [1966], *The Linear Hypothesis: A General Theory*, Griffin's Statistical Monographs No. 19. Griffin: London.

Smith, G. and F. Campbell [1980], "A Critique of Some Ridge Regression Methods (Invited Paper)," *JASA*, 75, pp. 74-104.

Thisted R.A. [1980], Discussion of 'A Critique of Some Ridge Regression Methods (Invited Paper)," by G. Smith and F. Campbell, *JASA*, 75, pp. 81-86.

Thisted, R.A. [1982], "Decision-Theoretic Regression Diagnostics," *Statistical Decision Theory and Related Topics III*, Vol. 2, pp. 363-382, Academic Press, Inc., New York.

Thisted, R.A. and Morris, C.N. [1980], "Theoretical Results for Adaptive Ordinary Ridge Regression Estimators," Technical Report No., 94 (revised), University of Chicago, Department of Statistics, Chicago.

Wilkenson, J.H. [1965], *The Algebraic Eigenvalue Problem*, Oxford University Press.