The Annals of Statistics
1998, Vol. 26, No. 4, 1411–1433

MAXIMIN CLUSTERS FOR NEAR-REPLICATE REGRESSION LACK OF FIT TESTS

BY FORREST R. MILLER, JAMES W. NEILL AND BRIAN W. SHERFEY

Kansas State University

To assess the adequacy of a nonreplicated linear regression model, Christensen introduced the concepts of orthogonal between- and within-cluster lack of fit with corresponding optimal tests. However, the properties of these tests depend on the choice of near-replicate clusters. In this paper, a graph theoretic framework is presented to represent candidate clusterings. A clustering is then selected according to a proposed maximin power criterion from among the clusterings consistent with a specified graph on the predictor settings. Examples are given to illustrate the methodology.

1. Introduction. Christensen (1989, 1991) derived uniformly most powerful invariant tests for detecting orthogonal between- and within-cluster lack of fit in linear regression models. These tests can be useful for assessing model adequacy for the common circumstance in which replicate measurements are not available. Green (1971), Breiman and Meisel (1976), Atwood and Ryan (1977), Lyons and Proctor (1977), Shillington (1979), Daniel and Wood (1980), Utts (1982), Neill and Johnson (1985) and Joglekar, Schuenemeyer and LaRiccia (1989) have also proposed lack of fit tests for the case of nonreplication. However, Christensen's approach is of particular interest since the lack of fit space was characterized as a sum of orthogonal subspaces with corresponding optimal tests.

All of the preceding tests require that the data be grouped into clusters and, depending on the test statistic, one of two approaches can be used as a basis for cluster selection. The near-replicate approach groups observations according to measures of nearness in the predictor space. The rationale behind this approach parallels the reasoning which motivates the classical lack of fit test with replication as introduced by Fisher (1922). Alternatively, a second approach determines clusters so that within each group the response is well approximated by the hypothesized model or a polynomial model in the specified predictors. An estimate of error variance is determined by pooling the residual sums of squares obtained from the local least squares model fittings within each cluster.

Joglekar, Schuenemeyer and LaRiccia (1989) summarized the various grouping methodologies, which can be classified as one of the two previous approaches. They noted that data-directed clusterings, which are generally

Received December 1995; revised September 1997.

AMS 1991 *subject classifications*. Primary 62J05; secondary 62F03.

Key words and phrases. Regression, lack of fit, nonreplication, between clusters, within clusters, maximin power, graph theory.

inherent to the methods based on the local fittings approach, lead to intractable distribution theory problems. In addition, current distance measures used for determining near-replicate clusters do not always admit an explicit, much less optimal, relation to the power of the corresponding lack of fit tests. Nonparametric regression techniques have also been developed to test the adequacy of parametric linear models [Hart (1997)]. The choice of smoothing parameter in this context parallels the problem of cluster selection as discussed above. A lack of fit test for generalized linear models which avoids the need to cluster observations was suggested by Su and Wei (1991). The test is based on the supremum of a partial sum process determined by the residuals. However, in the absence of relatively long sequences of residuals with like sign, this test may be expected to have diminished power. In addition, a rational way of ordering the cases before computing partial sums is needed.

The tests derived by Christensen involve models based on the near-replicate approach to clustering. These tests are in fact uniformly most powerful invariant for detecting orthogonal between- and within-cluster lack of fit, given a specific grouping of the data into near replicates. Thus, properties of the tests depend heavily on the choice of such clusters. In this paper we address the question of how to select near-replicate clusters which provide maximin power properties associated with the optimal tests.

In Section 2, the models and tests for orthogonal between- and within-cluster lack of fit are reviewed. A graph theoretic framework is presented in Section 3 to represent a collection of candidate groupings. A maximin power criterion is proposed in Section 4 in order to then select an optimal clustering from among the candidate groupings. The methodology is illustrated in Section 5.

2. Orthogonal between- and within-cluster lack of fit tests. The normal theory linear regression model is given by

$$Y = X\beta + \varepsilon, \tag{1}$$

where Y is an n-dimensional random response vector, X is a known, nonrandom $n \times p$ matrix of predictor variables, β is an unknown parameter in R^p and ε is an n-dimensional random error vector distributed as $N(\mathbf{0}, \sigma^2 I_n)$ with unknown $\sigma^2 > 0$. Lack of fit is said to exist when the linear structure $X\beta$ in model (1) does not adequately describe the mean of Y, that is, $E(Y) \neq X\beta$. In such case we may suppose that a true model which accounts for the inadequacy of model (1) is of the form

$$Y = X\theta + Q\gamma + \varepsilon, \tag{2}$$

where $C(X) \perp C(Q)$ and $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. The notation $C(A)$ denotes the column space of a matrix A.

In practice, a true model, and thus Q, are of course not known. Also, in the usual paradigm for testing lack of fit, there are no additional predictor variables available for inclusion in model (1). Hence, a lack of fit vector $Q\gamma$ in

model (2) is known only to be contained in the lack of fit space $C(X)^{\perp}$. A decomposition of $C(X)^{\perp}$ into orthogonal subspaces was obtained by Christensen (1989, 1991). The characterization provides the basis for the construction of optimal tests for assessing the existence of various types of lack of fit. In particular, the lack of fit space can be written as

$$C(X)^{\perp} = \left(C(X)^{\perp} \cap C(Z)\right) \oplus \left(C(X)^{\perp} \cap C(Z)^{\perp}\right) \oplus S,$$

where S denotes the orthogonal complement of the sum of the first two subspaces with respect to $C(X)^{\perp}$. The $n \times c$ matrix Z contains indicator variables for the near-replicate clusters, and thus contains only zeroes and ones. The column dimension represents the number of groups for a specified near-replicate clustering of the observations. Thus, the nonzero values in the ith column of Z correspond to the observations in the ith cluster of near replicates, $i = 1, 2, \ldots, c$. A clustering determined by such a Z will also be called a grouping or partition of the observations. In addition, the first two subspaces in the preceding decomposition of $C(X)^{\perp}$ are called the orthogonal between- and within-cluster lack of fit subspaces, respectively. This terminology corresponds to one-way analysis of variance in which clusters of replicates are identified with different treatment groups. In the special case that replication exists in the row structure of X, that is, $C(X) \subseteq C(Z)$, the tests for between and within cluster lack of fit compare model (1) with model (2), with $Q\gamma$ replaced by lack of fit vectors in $C(Z)$ and $C(Z)^{\perp}$, respectively. Thus, the test for between-cluster lack of fit reduces to the classical lack of fit test in the case of replication. Also note that the test for within-cluster lack of fit would, for example, allow for detection of a trend in time within each group of replicates whenever the replicates are observed in a time sequence. The interpretation for the case of near replication generalizes the preceding concepts.

For computational purposes, projections onto the lack of fit subspaces can be determined by the following lemma. The proof of the lemma is straightforward and thus omitted. The notation P_A denotes the orthogonal projection operator onto a subspace A of R^n.

LEMMA 1. *Let U and V be subspaces of R^n and suppose W is the subspace of R^n given by $W = P_U(V^{\perp})$. Then $P_{U \cap V} = P_U - P_W$.*

By letting $U = C(Z)$ and $V = C(X)^{\perp}$ in the preceding lemma,

$$P_{C(X)^{\perp} \cap C(Z)} = P_{C(Z)} - P_{C(X_0^Z)},$$

where $X_0^Z = P_{C(Z)}X$. Thus, the likelihood ratio test statistic for testing

$$H_o^B\colon E(Y) \in C(X)$$

versus

$$H_a^B\colon E(Y) \in C(X) \oplus \left(C(X)^{\perp} \cap C(Z)\right) \text{ and } E(Y) \notin C(X)$$

is given by

$$F_B = \frac{\left\|\left(P_{C(Z)} - P_{C(X_0^Z)}\right)Y\right\|^2 / r_1^B}{\left\|\left(P_{C(X)^\perp} - \left(P_{C(Z)} - P_{C(X_0^Z)}\right)\right)Y\right\|^2 / r_2^B}$$

where $r_1^B = c -$ dimension $C(X_0^Z)$ and $r_2^B = n - p - r_1^B$ with $c =$ dimension $C(Z)$ and $p =$ dimension $C(X)$. Also, for $x \in R^n$, $\|x\|^2$ denotes the squared Euclidean length $x^{\mathrm{T}}x$ of x. Since $F_B \sim F_{r_1^B, r_2^B}(\delta_B)$ where

$$\delta_B = \left\|P_{C(X)^\perp \cap C(Z)} E(Y)\right\|^2 / \sigma^2,$$

a uniformly most powerful invariant size α test of H_o^B versus H_a^B rejects H_o^B provided

$$F_B > F^{\alpha}_{r_1^B, r_2^B}.$$

The notation $F^{\alpha}_{d_1, d_2}$ represents the upper α point of a central F distribution with d_1 and d_2 degrees of freedom.

By symmetry, the likelihood ratio test statistic F_W and its distribution $F_{r_1^W, r_2^W}(\delta_W)$ for testing

$$H_o^W\colon E(Y) \in C(X)$$

versus

$$H_a^W\colon E(Y) \in C(X) \oplus \left(C(X)^\perp \cap C(Z)^\perp\right) \quad \text{and} \quad E(Y) \notin C(X)$$

can be obtained by replacing $C(Z)$ with $C(Z)^\perp$ in F_B and $F_{r_1^B, r_2^B}(\delta_B)$, respectively.

The following sections specifically address the problem of choosing a grouping matrix Z to effectively test H_o^B against H_a^B. Although symmetry considerations can be used in part to derive optimal clusters for testing H_o^W against H_a^W, a separate paper will be written to discuss this problem more fully.

3. Graph theoretic representation of candidate groupings. For the maximin power criterion, as defined in the following section, to be computationally feasible one must generally restrict the number of potential groupings under consideration. For example, with $n = 16$, the number of all possible groupings is 10,480,142,147, as calculated by the recurrence relations satisfied by Stirling numbers of the second kind [Constantine (1987)]. The following graph theoretic framework may be viewed in part as a device to eliminate absurd groupings directly and thus reduce the number of partitions under consideration. In addition, the framework based on graph theory provides a useful representation of the collection of candidate groupings to which the maximin criterion is applied.

Suppose the rows of the $n \times p$ matrix of predictor variables X are denoted by $r_i \in \mathbf{R}^P$, $i = 1, \ldots, n$. A graph G may be defined with n vertices $v_1, \ldots, v_n$ identified with $r_1, \ldots, r_n$, respectively. Furthermore, an edge is said to exist between two predictor settings r_i and r_j if and only if the corresponding

observations are candidates for near-replicate status. Several approaches may be considered for determining the edge set of G. For example, a nearness parameter $\varepsilon > 0$ can be specified along with a distance function d on $R^P \times R^P$. If $d(r_i, r_j) \leq \varepsilon$ then the ith and jth observations are candidates for near-replicate status, and thus an edge exists between vertices v_i and v_j. More generally, different nearness parameters ε_k, $k = 1, \ldots, p$, can be used for predictors with different scales. In such a case an edge exists between v_i and v_j provided $d_k(r_i^k, r_j^k) \leq \varepsilon_k$, $k = 1, \ldots, p$, where r_i^k denotes the kth component of r_i, and d_k is a specified distance function appropriate for the scaling of the kth predictor variable. A distance function d on $R^P \times R^P$ can also be generated as $d = \Sigma w_k d_k$ with the d_k as described above and for selected weights w_k, $k = 1, \ldots, p$. Alternatively, a family of overlapping subsets $\{S_1, \ldots, S_m\}$ may be specified in $R^P (R^{P-1}$ in case X has a column of ones or its equivalent). An edge exists between vertices v_i and v_j if r_i and r_j lie in S_k for some $k = 1, \ldots, m$. For example, the subsets may be overlapping grids in R^P which are chosen in a manner that excludes extreme pairings of the predictor settings from being clustered together. Each of the preceding approaches for determining the edge set is illustrated below with a simple example in R^2. In addition, the overlapping grid approach is used in the more complex setting of Example 2 in Section 5. In either case, a graph is determined with a corresponding collection of groupings. Specifically, the collection of groupings associated with a graph are naturally defined as in Definition 1 below. In this definition, let G be a graph with vertex set V and recall that a subgraph of G induced by $V_o \subseteq V$ is the graph obtained by deleting all vertices not in V_o from the graph on V, together with all edges that do not join two vertices of V_o. In addition, a subgraph is complete if there is an edge between every two of its vertices. Note that singletons are considered to be complete subgraphs.

DEFINITION 1. The collection of groupings consistent with a graph G consists of all partitions $P = \{A_1, \ldots, A_k\}$ of $V = \{v_1, \ldots, v_n\}$ such that each A_i determines a complete subgraph of G.

To illustrate the concepts given in the preceding paragraph and definition, consider a simple example with

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.3 & 1.8 & 2.6 & 2.7 & 3.4 \end{bmatrix}$$

and suppose the graph on the five vertices (rows of X) has edge set determined by Euclidean distance in R^2 and a nearness parameter $\varepsilon = 1$, as described above. The resulting graph is given by

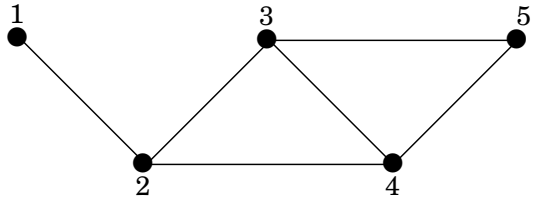

TABLE 1
Consistent groupings

	Dimension $C(Z)$														
Vertex	**2**	**3**							**4**						**5**
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	2	1	1	1	2	2	2	2	2	1	2
3	2	3	3	2	2	2	2	2	3	3	3	3	2	2	3
4	2	2	3	3	2	3	3	2	4	4	2	3	3	3	4
5	2	3	3	3	3	2	3	3	3	4	4	4	4	4	5

Note that the same graph is obtained by specifying overlapping subsets in R^1 given by the intervals $S_1 = (1, 2)$, $S_2 = (1.5, 3)$ and $S_3 = (2.5, 4)$. For example, an edge exists between vertices v_2 and v_3 since 1.8 and 2.6 lie in S_2. However, there is no edge between v_1 and v_3 since 1.3 and 2.6 do not lie in a common S_i for some $i = 1, 2, 3$. For $n = 5$ there are 52 possible groupings, 15 of which are consistent with this graph. The consistent groupings are listed in Table 1 and classified according to the dimension of $C(Z)$. The notation used in the table represents a particular grouping matrix Z by a corresponding n-tuple $\mathbf{z} = (z_1, \ldots, z_n)^T$ where z_i is the cluster number for the ith row of X.

It should be emphasized that specification of a graph only indicates which pairs of observations are potential near replicates. The specification of a nearness parameter, for example, does not determine a unique grouping but rather a collection of groupings consistent with the associated graph. The maximin power criterion is then applied to this collection. One is still generally choosing a maximin clustering from a large number of candidate groupings represented by the collection of groupings consistent with the specified graph. Of course the complete graph may be selected, in which case the collection of consistent groupings consists of all possible groupings. However, for most practical problems, the number of all possible groupings is enormous, as indicated above.

4. Maximin power criterion for clustering near replicates.

4.1. *Definition of the criterion.* The tests discussed in Section 2 are optimal for orthogonal between- and within- cluster lack of fit, which depend on the choice of the grouping matrix Z. For the purpose of clustering in observational experiments, the matrix X may be regarded as fixed while the matrix Z remains to be chosen. To identify clusterings that allow the corresponding optimal lack of fit test to discriminate effectively between model (1) and a type of lack of fit represented by model H_a^B, a maximin strategy is proposed. In particular, since the power of the F test for lack of fit is an increasing function of the noncentrality parameter, a maximin criterion will be applied to δ_B to provide a cluster selection associated with the optimal test. It will be shown below that the degrees of freedom parameters of the F

distribution for the optimal test, corresponding to a maximin clustering, are inherently in concordance with the objective of maximal power. Jones and Mitchell (1978), Atkinson and Fedorov (1975) and Atkinson (1972) discuss design criteria which also use a maximin power approach for detecting lack of fit. However, these criteria are concerned with the selection of design points which determine the rows of X. In addition, the alternative models upon which these criteria are based are of the form of model (2) where Q is a matrix of known functions of the settings of the predictor variables in X. For this form of model (2), Shelton, Khuri and Cornell (1983) also suggested a maximin power criterion to select check points for use in assessing lack of fit. Model discrimination designs for polynomial regression were discussed by Dette (1994, 1995). Specifically, optimal designs which maximize the minimum power of a given set of alternatives were discussed, in addition to generalized c-optimal designs for polynomial regression.

In order to present the maximin criterion for cluster selection, a set of candidate groupings is required. Assume a graph G has been specified along with the collection of groupings consistent with the graph as defined in Section 3. This collection must be restricted according to the dimension considerations given in the following definition.

DEFINITION 2. Let Ξ_G denote the set of grouping matrices Z which give partitions consistent with G and satisfy $C(X)^{\perp} \cap C(Z) \neq \{0\}$, excluding the trivial partitions which cluster all or none of the observations together.

For example, returning to the X matrix and associated graph introduced in Section 3, Ξ_G consists of those partitions consistent with the graph and having dimension $C(Z)$ equal to three or four. This observation follows from the fact that dimension $C(X)^{\perp} \cap C(Z) = c - \text{rank } X_0^Z$ where $c =$ dimension $C(Z)$, and noting that rank $X_0^Z = 2$ for all the matrices Z indicated in Table 1. Thus, excluding the trivial partition corresponding to $c = 5$, only clusterings with $c = 3, 4$ provide orthogonal between-cluster lack of fit subspaces with positive dimension. Note that the trivial partition which clusters all of the observations together, that is, the case $c = 1$, is automatically excluded for this example since X has a column of ones and thus $C(X)^{\perp} \cap C(Z) = \{0\}$.

Suppose the alternative model is assumed to be of the form given by H_a^B for some $Z \in \Xi_G$, and one wishes to test the adequacy of the H_o^B model. A maximin criterion will next be discussed in the present context to determine an appropriate $Z \in \Xi_G$, and hence a clustering. To explain the basic concept underlying the maximin criterion for cluster selection, let τ denote a quadratic form defined on the lack of fit space $C(X)^{\perp}$. As seen in the following discussion, τ must be positive definite on each of the lack of fit subspaces $C(X)^{\perp} \cap C(Z)$ for $Z \in \Xi_G$. A specific class of such forms which allow for nearness considerations is given in Section 4.2. For fixed $Z \in \Xi_G$, determine the vector in the corresponding orthogonal between-cluster lack of fit space

which minimizes the noncentrality parameter δ_B, subject to the restriction that the vector lies in a set of the form $\{\tau \geq \Delta\}$ for some $\Delta > 0$. We want to determine a $Z \in \Xi_G$ which maximizes such minimum noncentrality parameter values. Specifically, a $Z \in \Xi_G$ is sought which maximizes

$$\Lambda_Z = \inf\{\|v\|^2 : v \in C(X)^\perp \cap C(Z), \tau(v) \geq \Delta\}$$

for some fixed $\Delta > 0$. Next, letting

$$l_Z = \inf\left\{ \frac{\|v\|^2}{\tau(v)} : v \in C(X)^\perp \cap C(Z), v \neq 0 \right\} \tag{3}$$

and noting that $\|sv\|^2/\tau(sv) = \|v\|^2/\tau(v)$ for nonzero $v \in C(X)^\perp \cap C(Z)$ and any nonzero scalar s, it follows that

$$l_Z = \inf\left\{ \frac{\|v\|^2}{\tau(v)} : v \in C(X)^\perp \cap C(Z), \tau(v) = \Delta \right\} = \frac{1}{\Delta}\Lambda_Z.$$

Consequently, a clustering matrix $Z \in \Xi_G$ which maximizes Λ_Z does not depend on Δ. Accordingly, a maximin clustering matrix is formally defined as follows.

DEFINITION 3. A clustering matrix $Z \in \Xi_G$ which maximizes l_Z as given by (3) is defined to be a maximin clustering matrix from among the candidate groupings in Ξ_G.

Note that a maximin clustering is invariant under reparametrizations and does not depend on the data vector Y. Also, implicit in the above discussion is the fact that the quadratic form τ allows comparisons of vectors in the lack of fit subspaces $C(X)^\perp \cap C(Z)$ for different $Z \in \Xi_G$. In particular, suppose $\xi \in C(X)^\perp \cap C(Z_0)$ and $\eta \in C(X)^\perp \cap C(Z)$ where $Z_0, Z \in \Xi_G$ and $\tau(\xi) = \tau(\eta)$. Furthermore, if Z_0 is a maximin clustering matrix and $\|\xi\|^2 = \Lambda_{Z_0}$ and $\|\eta\|^2 = \Lambda_Z$, then $\|\xi\|^2 \geq \|\eta\|^2$. Thus, for lack of fit vectors possessing the preceding properties, the corresponding noncentrality parameter values based on a maximin clustering are maximal as compared to corresponding values based on nonmaximin clusterings. In fact, by the definition of Λ_{Z_0} it follows that $\|v\|^2 \geq \|\eta\|^2$ for any $v \in C(X)^\perp \cap C(Z_0)$ with $\tau(v) = \tau(\eta)$ and $\|\eta\|^2 = \Lambda_Z$. A calculational form for Λ_Z (equivalently l_Z) is derived in the Appendix.

4.2. *Contour quadratic forms based on power and nearness.* A specific class of quadratic forms τ for use in Λ_Z will next be defined. First note that permuting the columns of a specified grouping matrix Z results in different Z matrices, but each such Z corresponds to the same clustering. Such Z matrices will be considered as being equivalent. Thus, when a set of grouping matrices is specified, a set of equivalence classes is inherently specified. In particular, the notation $\sum_{Z \in \zeta}\{\ \}$, where ζ is a collection of grouping matrices, will indicate that the sum is over all distinct groupings determined by ζ. That is, each equivalence class of ζ contributes just one term to the sum.

Now note that the possible alternative models are of the form

$$Y = X\theta + P_{C(X)^{\perp} \cap C(Z)}u + \varepsilon,$$

where $Z \in \Xi_G$ and $u \in C(X)^{\perp}$. To motivate a choice for τ, let $u \in C(X)^{\perp}$ be fixed. If $\|P_{C(X)^{\perp} \cap C(Z)}u\|^2 < \Delta$ for all $Z \in \Xi_G$ where $\Delta > 0$ is a small preassigned number, then one may suppose that u will not contribute to any detectable between-cluster lack of fit. Next let $\Xi_E \subseteq \Xi_G$ consist of those groupings determined by a single edge in the edge set E of the specified graph G. That is, each edge of G determines a clustering in Ξ_G, which clusters only the two connected vertices, with all other vertices being singleton clusters. For example, note that Ξ_E for the example introduced in Section 3 consists of those partitions with dimension $C(Z)$ equal to four. Now observe that $\|P_{C(X)^{\perp} \cap C(Z)}u\|^2 < \Delta$ for all $Z \in \Xi_G$ if and only if $\|P_{C(X)^{\perp} \cap C(Z)}u\|^2 < \Delta$ for all $Z \in \Xi_E$. Note that the "if" part of the preceding claim follows since for any $Z \in \Xi_G$ there exists a $Z_1 \in \Xi_E$ such that $C(Z) \subseteq C(Z_1)$, and thus $\|P_{C(X)^{\perp} \cap C(Z)}u\|^2 \leq \|P_{C(X)^{\perp} \cap C(Z_1)}u\|^2$. The point of introducing Ξ_E and the preceding equivalence is that the cardinality of Ξ_E is generally much smaller than that of Ξ_G, from which computational advantages are obtained. Thus, based on the preceding, consider τ defined by

$$\tau(u) = \sum_{Z \in \Xi_E} w_Z \|P_{C(X)^{\perp} \cap C(Z)}u\|^2 \tag{4}$$

for $u \in C(X)^{\perp}$ with $w_Z \geq 0$ and $\sum_{Z \in \Xi_E} w_Z = 1$. Since $\tau(u)$ is a convex combination, $\|P_{C(X)^{\perp} \cap C(Z)}u\|^2 < \Delta$ for all $Z \in \Xi_E$ implies that $\tau(u) < \Delta$. Hence, $\tau(u) \geq \Delta$ implies that $\|P_{C(X)^{\perp} \cap C(Z)}u\|^2 \geq \Delta$ for at least one $Z \in \Xi_E$, and u may contribute to detectable between-cluster lack of fit for some $Z \in \Xi_E$.

To incorporate nearness as measured by $\|X - X_0^Z\|^2$ into the weight w_Z, let

$$w_Z = \|X - X_0^Z\|^2 \Big/ \sum_{Z^* \in \Xi_E} \|X - X_0^{Z^*}\|^2.$$

The notation $\|A\|^2$ denotes the squared matrix norm defined by $\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2$ where a_{ij} is the (ij)th element of an $n \times p$ matrix A. The motivation for the above measure of nearness is derived from the fact that the jth row of the matrix $X_0^Z = P_{C(Z)}X$ is obtained by averaging over the rows of X corresponding to the cluster, as determined by Z, which contains the jth observation, $j = 1, \ldots, n$. Thus, as we illustrate below, the effect of using the preceding weights w_Z in τ is to penalize those grouping matrices Z which cluster observations corresponding to rows of X which are far apart. Consider the case with

$$X^T = \begin{bmatrix} 1, & 1, \ldots, 1 \\ x_1, & x_2, \ldots, x_n \end{bmatrix},$$

and note that τ reduces to

$$\tau(u) = \sum_{\substack{[x_i, x_j] \in E \\ i<j}} w_{Z_{ij}} \|P_{C(X)^{\perp} \cap C(Z_{ij})}u\|^2,$$

where Z_{ij} is the grouping matrix that clusters the ith and jth vertices only and

$$w_{Z_{ij}} = (x_i - x_j)^2 \Big/ \sum_{\substack{[x_k, x_m] \in E \\ k<m}} (x_k - x_m)^2.$$

In the following discussion, suppose there are only two values x_{i_o} and x_{j_o} in the second column of X which are relatively far part. If there are more than two such values, then an analogous argument holds in order to obtain the following conclusions. Now note that if $Z \in \Xi_G$ groups x_{i_o} and x_{j_o} where $(x_{i_o} - x_{j_o})^2$ is relatively large, then $w_{Z_{i_o j_o}} \simeq 1$. As a result, we claim that $l_Z \simeq 1$ for such Z. To see this, let $u \in C(X)^\perp \cap C(Z)$ and without loss of generality suppose $\|u\|^2 = 1$. Then, since $C(X)^\perp \cap C(Z) \subseteq C(X)^\perp \cap C(Z_{i_o j_o})$, $\|P_{C(X)^\perp \cap C(Z_{i_o j_o})} u\|^2 = \|u\|^2 = 1$ and thus $w_{Z_{i_o j_o}} \leq \tau(u) \leq 1$. Consequently, $\tau(u) \simeq 1$ for $u \in C(X)^\perp \cap C(Z)$ with $\|u\|^2 = 1$, and hence $l_Z \simeq 1$. This result indicates that Z matrices which cluster observations corresponding to rows of X which are far apart are not favored according to the maximin clustering criterion. The preceding conclusion follows, since $l_Z \geq 1$ for every $Z \in \Xi_G$ when τ has the general form given by (4). Of course τ may be defined with alternative choices for the weights w_Z. For example, a uniform weight may be taken with $w_Z = (\text{cardinality } \Xi_E)^{-1}$.

For the maximin approach to clustering to be operative, the quadratic form τ must be positive definite on each subspace $C(X)^\perp \cap C(Z)$ for $Z \in \Xi_G$. The following proposition provides a condition to ensure that any τ of the form given by (4) is in fact positive definite on $C(X)^\perp$. The proof of the proposition is given in the Appendix.

PROPOSITION 1. *τ is positive definite on $C(X)^\perp$ provided*

$$\sum_{Z \in \Xi_{E^+}} C(X)^\perp \cap C(Z) = C(X)^\perp, \tag{5}$$

where $\Xi_{E^+} \subseteq \Xi_E$ consists of those $Z \in \Xi_E$ such that $w_Z > 0$.

4.3. *Refinements and atoms.* As will be shown, the maximin approach to clustering necessarily results in a cluster selection from among those partitions which group as many observations together as possible. To facilitate the following discussion, the concepts of refinement and atoms are introduced in the next definition. These concepts will be illustrated later in this subsection in the context of the example introduced in Section 3.

DEFINITION 4. For $Z_0, Z_1 \in \Xi_G$, Z_1 is said to be a refinement of Z_0 provided $C(Z_0) \subseteq C(Z_1)$. In addition, Z_0 is an atom of Ξ_G provided Z_0 is not a refinement of any other member of Ξ_G. Let Ξ_0 denote the set of atoms in Ξ_G.

Note that if Z_1 is a refinement of Z_0 in Ξ_G then

$$\left\{ \frac{\|v\|^2}{\tau(v)} : v \in C(X)^{\perp} \cap C(Z_0), v \neq 0 \right\} \subseteq \left\{ \frac{\|v\|^2}{\tau(v)} : v \in C(X)^{\perp} \cap C(Z_1), v \neq 0 \right\}.$$

Hence, $l_{Z_0} \geq l_{Z_1}$, so that maximization of l_Z can be made with respect to the atoms $Z \in \Xi_o$. Since atoms represent those partitions consistent with the graph which group as many observations together as possible, the claim made at the beginning of this subsection has been verified. It should also be noted that this discussion concerning refinements holds for *any quadratic form* τ which is positive definite on each subspace $C(X)^{\perp} \cap C(Z)$ for $Z \in \Xi_G$.

Based on the preceding, in order to determine a maximin clustering for a given X, the set of atoms Ξ_o in Ξ_G must be determined. The following theorem characterizes the atoms associated with a graph. The proof of Theorem 1 is given in the Appendix. In Theorem 1 and in subsequent discussions, a clique of a graph G is defined as a maximal complete subgraph of G. In addition, the definition of a minimal cover for a set is recalled for use in determining the atoms of a graph.

DEFINITION 5. A cover of a set S is a collection of sets $\{A_\alpha\}$ whose union is S. A cover $\{A_\alpha\}$ of S is minimal provided $A_\alpha \setminus \bigcup_{\beta \neq \alpha} A_\beta \neq \phi$ for each α. Note that this is equivalent to the case that for every α there is an element of S which is in A_α but not A_β for $\beta \neq \alpha$.

THEOREM 1. *Let G be a graph with vertex set V identified with the rows of X as indicated at the beginning of Section 3. Let C denote the set of cliques of G. Also assume that dimension $C(X) = p < k$ where k is the cardinality of C.*

(i) *If $Z \in \Xi_G$ determines an atom of Ξ_G and $c =$ dimension $C(Z)$, then $c \leq k$.*

(ii) *If G has the property that C is a minimal cover for V, then the atoms Ξ_o of Ξ_G consist of exactly those $Z \in \Xi_G$ with dimension $C(Z) = k$.*

To illustrate the use of Theorem 1 for determining atoms, recall the example introduced in Section 3 and first note by inspection that the set of cliques for the graph of this example is given by $C = \{\{1, 2\}, \{2, 3, 4\}, \{3, 4, 5\}\}$. In addition, note that C is not a minimal cover for the vertices of the graph since the set $\{2, 3, 4\}$ does not contain an element of V which is not contained in $\{1, 2\}$ or $\{3, 4, 5\}$. Thus, Theorem 1(ii) is not applicable. However, Theorem 1(i) ensures that the atoms of Ξ_G have dimension $C(Z) \leq 3$, where 3 is the cardinality of C. Since as shown previously there are no $Z \in \Xi_G$ with dimension $C(Z) < 3$, the set of atoms Ξ_0 consists of those $Z \in \Xi_G$ with dimension $C(Z) = 3$. As argued above, a maximin clustering would be chosen

from among those partitions in Ξ_0. Example 2 in Section 5 will be used to illustrate the case in which the set of cliques is a minimal cover for the vertices of the graph, and thus allows application of Theorem 1(ii) to determine the set of atoms.

Next suppose that a specified graph G breaks up into $k > p$ (k and p as defined in Theorem 1) disjoint complete subgraphs $G_1, \ldots, G_k$ with no edges between the vertices of G_i and G_j for $i, j = 1, \ldots, k$ and $i \neq j$. Let Z_0 be a clustering matrix which groups only the vertices of G_i together for $i = 1, \ldots, k$. Since dimension $C(X_0^Z) \leq p$ for every clustering matrix Z, dimension $C(X)^\perp \cap C(Z_0)$ = dimension $C(Z_0)$ − dimension $C(X_0^Z) \geq k - p$. Then, since $k > p$, Z_0 is in Ξ_G and every clustering in Ξ_G is a refinement of Z_0. Thus, by Definition 4, Z_0 is the only atom in Ξ_G. Hence, when there is a clear grouping, as in this case, the maximin power criterion chooses it. However, the utility of the criterion is more fully realized when there is not one clear grouping. In such cases, we want to allow sufficiently many potential near-replicate pairs so that the graph is connected, as in Examples 1 and 2 of Section 5. Recall that a graph G is connected if for every two distinct vertices $v_i, v_j \in V$ there is a continuous path consisting of a finite sequence of distinct edges joining v_i to v_j. For the case of disconnected graphs, it is possible, depending on τ, that the maximin clustering criterion provides no discrimination between the candidate groupings. This point is further discussed in the Appendix. However, it should be emphasized that connected graphs provide the appropriate clustering possibilities in case there is not one clear grouping, and thus connected graphs are of most interest.

4.4. *Power considerations and implementation.* The power function of the optimal size α test for testing orthogonal between-cluster lack of fit is given by

$$q\left(r_1^B, r_2^B, \delta_B, \alpha\right) = P\left(F_B > F^{\alpha}_{r_1^B, r_2^B} | \delta_B\right)$$

for a specified $Z \in \Xi_G$. Note that q is an increasing function of r_2^B for fixed values of r_1^B and δ_B and a decreasing function of r_1^B for fixed values of r_2^B and δ_B [Ghosh (1973)]. Now if $Z \in \Xi_G$ and Z_0 is an atom in Ξ_G with Z a refinement of Z_0, then $l_{Z_o} \geq l_Z$ as shown in Section 4.3. Consequently, maximization of l_Z can be restricted to the atoms of Ξ_G. Furthermore, since $C(Z_o)$ is a subspace of $C(Z)$, it follows that $r_1^B(Z_o)$ = dimension $C(X)^\perp \cap C(Z_o) \leq$ dimension $C(X)^\perp \cap C(Z) = r_1^B(Z)$ and $r_2^B(Z_o) = n - p - r_1^B(Z_o) \geq n - p - r_1^B(Z) = r_2^B(Z)$. Thus, the degrees of freedom parameters for the distribution of F_B, based on clusterings corresponding to the atoms, are inherently in concordance with the objective of maximal power.

If the degrees of freedom parameters $r_1^B(Z)$ and $r_2^B(Z)$ are constant on the atoms $Z \in \Xi_o$, then the atoms can be compared for power on the basis of the l_Z values alone for $Z \in \Xi_o$. An atom which maximizes l_Z with respect to $Z \in \Xi_o$ thus corresponds to a maximin clustering as defined in Section 4.1.

For implementation purposes, the constancy of $r_1^B(Z)$ and $r_2^B(Z)$ is determined by directly evaluating dimension $C(X)^\perp \cap C(Z)$ for each atom $Z \in \Xi_o$. Furthermore, since dimension $C(X)^\perp \cap C(Z)$ = dimension $C(Z)$ − dimension $C(X_o^Z)$, there are two ways this constancy may not obtain. In particular, if dimension $C(Z)$ is not constant on the atoms $Z \in \Xi_o$ then constancy of the degrees of freedom parameters may not hold. On the other hand, if dimension $C(Z)$ is constant but dimension $C(X_o^Z)$ is not constant on the atoms then constancy of the degrees of freedom may not follow. However, according to Theorem 1, if the number of cliques for the specified graph G is greater than dimension $C(X)$ and the cliques form a minimal cover for the vertices of G, then dimension $C(Z)$ is equal to the cardinality of the set of cliques for every atom $Z \in \Xi_o$. In addition, we claim that for *most* predictor matrices X, if dimension $C(Z)$ is constant on the atoms then dimension $C(X)^\perp \cap C(Z)$ is constant on the atoms as well. This claim, which is justified by Theorem 2 below, leads to the concept of a generic predictor matrix as defined next. The proof of Theorem 2 is given in the Appendix.

DEFINITION 6. Let X be an $n \times p$ matrix of predictor variables with dimension $C(X) = p$ and let c = dimension $C(Z)$ for Z a grouping matrix. If

$$\text{dimension } C\left(X_o^Z\right) = \begin{cases} p, & \text{if } p \le c, \\ c, & \text{if } p > c, \end{cases}$$

for all grouping matrices Z then X is generic.

THEOREM 2. *Consider the $n \times p$ matrix $X = (x_1, \ldots, x_p)$, where x_1 is a column of ones or its equivalent, to be determined by an element of $R^{n(p-1)}$. With this identification, the set of generic X matrices is open and dense in $R^{n(p-1)}$. If X does not provide for an intercept in the model, then the preceding holds with R^{np} in place of $R^{n(p-1)}$.*

Theorem 2 justifies and makes more precise the claim made above. In particular, except for nongeneric $n \times p$ X matrices which constitute a set of Lebesgue measure zero in R^{np}($R^{n(p-1)}$ in case of an intercept), if dimension $C(Z)$ is constant on the atoms then dimension $C(X)^\perp \cap C(Z)$ is also constant on the atoms. However, it should be emphasized that for implementation of the maximin clustering criterion, we do not need to check whether X is generic. Rather, as indicated previously, we simply check whether dimension $C(X)^\perp \cap C(Z)$ is constant on the atoms. The significance of Theorem 2 is based on the assurance that the constancy of dimension $C(X)^\perp \cap C(Z)$ on the atoms will in fact ordinarily be realized whenever dimension $C(Z)$ is constant on the atoms. The latter will hold, for example, when the cliques form a minimal cover for the vertices of the specified graph. In this case, dimension $C(X)^\perp \cap C(Z) = k - p$ for all $Z \in \Xi_o$ whenever X is generic with $p < k$ and k denotes the number of cliques of G. For completeness we remark that in case the specified graph produces atoms such that dimension $C(X)^\perp \cap C(Z)$,

$Z \in \Xi_o$, is not constant then one can proceed as follows. First group the atoms into classes according to dimension $C(X)^{\perp} \cap C(Z)$. Next choose an atom from each class with maximum l_Z value, and then select from among such atoms the clustering which provides a dominant power curve.

4.5. *Consistency of the test based on maximin clustering*. In addition to atom determination, the cliques of a graph can also be related to the consistency of the test based on F_B. Suppose that as the total number of observations $n \to \infty$, the number of clusters c_n = dimension $C(Z_n) \to \infty$ such that $c_n/n \to \rho$, $0 < \rho < 1$. In addition, assume the true model is of the form given by model (2) with dimension $C(X_n, Q_n)$ fixed and dimension $C(X_n) = p$ for all n. Under the preceding asymptotic scheme and assuming independent zero mean random errors with common second, third and fourth moments, Christensen (1989) proved consistency of the test based on F_W when the alternative only holds in the limit. By symmetry, analogous conditions for consistency of the test based on F_B can be given. In particular, using the same asymptotic scheme and method of proof, the following proposition holds.

PROPOSITION 2. *If* $\|Q_n\gamma\|^2/c_n \to \eta > 0$ *and* $\|P_{C(X_n)+C(Z_n)^{\perp}} Q_n\gamma\|^2/(n - c_n) \to 0$ *then* $F_B \to_p 1 + \eta/\sigma^2$ *as* $n \to \infty$.

Note that the conditions of Proposition 2 ensure that the true model contains lack of fit $Q_n\gamma$, which remains substantial as $n \to \infty$ and is asymptotically between clusters. Next suppose that G_n and V_n denote a graph and associated vertex set corresponding to the generic predictor matrix X_n. Also let C_n represent the set of cliques for G_n, and suppose C_n forms a minimal cover for V_n. Assume the predictor space is unbounded in R^P and let c_n = cardinality $C_n \to \infty$ as $n \to \infty$ with $c_n/n \to \rho$, $0 < \rho < 1$. Finally, suppose the lack of fit vector $Q_n\gamma \in C(X_n)^{\perp} \cap C(Z_n)$ for all n and $\|Q_n\gamma\|^2/c_n \to \eta > 0$ as $n \to \infty$. Then, by Proposition 2, the lack of fit test based on F_B is consistent. For example, suppose that $|\,\|Q_n\gamma\|^2 - nb| < B$ for some positive constants b and B for all n. Then $\|Q_n\gamma\|^2/c_n \to b/\rho > 0$ as $n \to \infty$ so that the lack of fit remains substantial for consistency purposes.

5. Examples.

EXAMPLE 1. Some fundamental concepts of the maximin clustering criterion have been illustrated in Sections 3 and 4 based on the simple matrix X and associated graph given in Section 3. In particular, the set of atoms $\Xi_o \subseteq \Xi_G$ was found to be the collection of grouping matrices consistent with the graph with dimension $C(Z) = 3$, as listed in Table 1. Accordingly, a maximin clustering is selected from this collection and the atoms can be compared for power by using the maximin noncentrality parameter alone. Utilizing the calculational form of l_Z given in the Appendix, $\max_{Z \in \Xi_o} l_Z =$ 1.789945 corresponding to the maximin grouping matrix Z represented by

$\mathbf{z} = (1, 1, 2, 2, 3)^T$. As in Table 1, like coordinates indicate that the corresponding observations are to be grouped together.

EXAMPLE 2. The data for this example are taken from Draper and Smith [(1981), page 215], which reports percentage of properly sealed bars of soap (y), sealer plate clearance (x_1) and sealer plate temperature (x_2) with

$$\mathbf{x}_1^T = (130, 174, 134, 191, 165, 194, 143, 186, 139, 188, 175, 156, 190, 178, 132, 148),$$
$$\mathbf{x}_2^T = (190, 176, 205, 210, 230, 192, 220, 235, 240, 230, 200, 218, 220, 210, 208, 225).$$

Let $X = (\mathbf{1}_{16}, \mathbf{x}_1, \mathbf{x}_2)$. Figure 1 gives a plot of the predictor settings in R^2 with the ith row of X corresponding to i in the figure, $i = 1, \ldots, 16$. To illustrate the overlapping grid approach for determining a graph, let $I_1 = [130, 171]$ and $I_2 = [153, 194]$ be two overlapping intervals whose union includes the range of values in $\mathbf{x}_1$. Similarly, let $J_1 = [176, 217]$ and $J_2 = [199, 240]$ cover the values in $\mathbf{x}_2$. The corresponding overlapping grid elements in R^2 are then given by $S_{ij} = I_i \times J_j$ for $i, j = 1, 2$. That is, using the cell designations given in Figure 1, $S_{11} = A_1 \cup A_2 \cup A_4 \cup A_5$, $S_{21} = A_2 \cup A_3 \cup A_5 \cup A_6$, $S_{12} = A_4 \cup A_5 \cup A_7 \cup A_8$ and $S_{22} = A_5 \cup A_6 \cup A_8 \cup A_9$. The S_{ij} for $i, j = 1, 2$ determine a graph G with vertices V corresponding to the

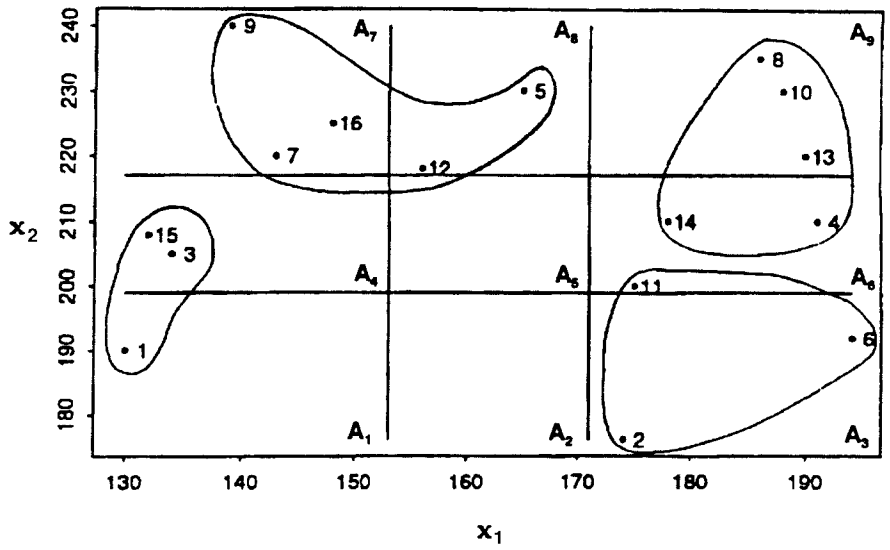

FIG. 1. *Overlapping grid elements* $S_{ij} = \cup_{I_{ij}} A_k$ *for* $i, j = 1, 2$, *where* $I_{11} = \{1, 2, 4, 5\}$, $I_{21} = \{2, 3, 5, 6\}$, $I_{12} = \{4, 5, 7, 8\}$ *and* $I_{22} = \{5, 6, 8, 9\}$, *of the predictor space in* R^2 *and maximin grouping for Example* 2.

16 rows of X and edges as discussed in Section 3. For example, an edge exists between vertices v_1 and v_3 since v_1 and v_3 lie in S_{11}. However, there is no edge between v_1 and v_7 since v_1 and v_7 do not lie in a common S_{ij} for some $i,j = 1,2$. In addition, the cliques of G are the sets $S_{ij} \cap V$ for $i, j = 1, 2$. A method is given in the Appendix to show this is the case. The method is also applicable to Example 1, although the cliques in the simple setting of that example can be determined by inspection. Thus, the set of cliques for the graph G is given by

$$C = \{\{1,3,15\},\{2,4,6,11,14\},\{3,5,7,9,12,15,16\},\{4,5,8,10,11,12,13,14\}\}.$$

Furthermore, C is a minimal cover for V since v_1 is in the first set but no other, v_2 is in the second set but no other, v_9 is in the third set but no other and v_8 is in the fourth set but no other. Thus, by Theorem 1(ii), the set of atoms $\Xi_o \subseteq \Xi_G$ consists of those groupings consistent with G and having dimension $C(Z) = 4$, where 4 is the cardinality of C. Specifically, the atoms for this example are the groupings that have observation 1 confined to cluster 1, observations 2 and 6 confined to cluster 2, observations 7, 9 and 16 confined to cluster 3 and observations 8, 10 and 13 confined to cluster 4. Note that the cluster designations are arbitrary. Also note that observation 1 is confined to cluster 1 since this observation lies in a portion of S_{11} (in particular, the A_1 cell) which does not overlap with any of the remaining grid elements S_{21}, S_{12} or S_{22}. Thus, there can be no edge between v_1 and any vertex which lies in the A_3, A_7 or A_9 cells by construction of the graph according to the overlapping grid approach. Analogous reasoning can be used to justify the claims made for the other three clusters. The remaining observations lie in one of the four clusters described above. As determined by computer enumeration, there are 128 atoms for this example while the cardinality of the set of all possible groupings is 10,480,142,147 as indicated in Section 3. A Fortran program has been written and implemented for general use to calculate $\max_{Z \in \Xi_o} l_Z$, which for this example is 1.10291 corresponding the maximin grouping matrix Z represented by $\mathbf{z} = (1,2,1,4,3,2,3,4,3,4,2,3,4,4,1,3)^T$ (notation as in Table 1) and shown in Figure 1.

APPENDIX

A calculational form for l_Z is derived in Part A of the Appendix. Proofs of Proposition 1 and Theorems 1 and 2 are given in Parts B, C and D, respectively. Disconnected graphs are discussed in Part E and a method for determining the cliques of a graph derived by the overlapping grid approach is developed and applied to Example 2 in Part F.

Part A. To calculate l_Z or equivalently Λ_Z for $Z \in \Xi_G$, let $\{e_1, \ldots, e_d\}$ be an orthonormal basis for $C(X)^\perp \cap C(Z)$ where d denotes the dimension of $C(X)^\perp \cap C(Z)$. Thus, letting $(v^1, \ldots, v^d)$ be the d-tuple in R^d representing

$v \in C(X)^{\perp} \cap C(Z)$ with respect to this basis,

$$\|v\|^2 = \sum_{i=1}^{d} (v^i)^2$$

and

$$\tau(v) = \sum_{i=1}^{d} \sum_{j=1}^{d} v^i v^j b_{ij},$$

where

$$b_{ij} = \tau(e_i, e_j) \quad \text{for } i, j = 1, \ldots, d.$$

Next let γ_i and Γ_i, $i = 1, \ldots, d$, represent the eigenvalues and corresponding orthonormal eigenvectors, respectively, of the matrix $B = (b_{ij})$. Hence, letting $x \cdot y$ denote the usual Euclidean inner product in R^n,

$$\|v\|^2 = \sum_{i=1}^{d} (\xi^i)^2$$

and

$$\tau(v) = \sum_{i=1}^{d} \gamma_i (\xi^i)^2,$$

where

$$\xi^i = v \cdot \Gamma_i \quad \text{for } i = 1, \ldots, d.$$

Thus, Λ_Z may be computed by considering the constrained extrema problem:

$$\text{minimize } f(x) = \sum_{i=1}^{d} x_i^2 \text{ subject to } g(x) = \Delta$$

for $x = (x_1, \ldots, x_d) \in R^d$, $x \neq 0$, where $g(x) = \sum_{i=1}^{d} \gamma_i x_j^2$ and Δ is a positive constant. Critical points of the function $F(x) = f(x) + \lambda(\Delta - g(x))$ are determined by $\nabla f = \lambda \nabla g$ and given by

$$x_i = \sqrt{\frac{\Delta}{\gamma_i}}, \qquad x_j = 0 \quad \text{for } j \neq i, i = 1, \ldots, d,$$

where λ denotes the Lagrange multiplier. Thus, the minimum value of f subject to $g = \Delta$ is $\Delta/\gamma_{\max}$ where $\gamma_{\max}$ is the largest eigenvalue of the matrix B. Hence,

$$\Lambda_Z = \Delta/\gamma_{\max} \quad \text{and} \quad l_Z = 1/\gamma_{\max}.$$

Part B. The proof of Proposition 1 follows Lemma 2. The proof of Lemma 2 is straightforward and thus omitted.

LEMMA 2. *Let U and V be subspaces in R^n and suppose W_1 and W_2 are the subspaces of R^n given by $W_1 = P_U(V^{\perp})$ and $W_2 = P_U V$, respectively. Then*:

(i) $U \cap V = U \cap W_1^{\perp}$,
(ii) $U \cap (U^{\perp} + V) = W_2$.

PROOF OF PROPOSITION 1. First note that by Lemma 2(i),

$$C(X)^{\perp} \cap \left(P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right)\right)^{\perp} = C(X)^{\perp} \cap C(Z)$$

for $Z \in \Xi_{E^+}$. Thus, by condition (5),

$$C(X)^{\perp} \subseteq \sum_{Z \in \Xi_{E^+}} \left(P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right)\right)^{\perp}. \tag{6}$$

Next note that (6) is equivalent to

$$\bigcap_{Z \in \Xi_{E^+}} P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right) = \{0\}. \tag{7}$$

This equivalence will be established below. Now note that (7) implies that τ is positive definite on $C(X)^{\perp}$. To see this, let $v \in C(X)^{\perp}$ and suppose that

$$P_{C(X)^{\perp} \cap C(Z)} v = 0$$

for all $Z \in \Xi_{E^+}$. Thus, $v \in C(X) + C(Z)^{\perp}$ for all $Z \in \Xi_{E^+}$. By Lemma 2(ii),

$$v \in C(X)^{\perp} \cap \left(C(X) + C(Z)^{\perp}\right) = P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right)$$

for all $Z \in \Xi_{E^+}$. Hence,

$$v \in \bigcap_{Z \in \Xi_{E^+}} P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right)$$

so that $v = 0$ by (7). Thus, for $v \in C(X)^{\perp}$, $\tau(v) = 0$ implies that $v = 0$.

The equivalence of (6) and (7) follows by first observing that (7) holds if and only if

$$R^n = \sum_{Z \in \Xi_{E^+}} \left(P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right)\right)^{\perp}. \tag{8}$$

Thus, (8) implies (6) immediately. In addition, since

$$P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right) \subseteq C(X)^{\perp}$$

for all $Z \in \Xi_{E^+}$, we have that

$$C(X) \subseteq \sum_{Z \in \Xi_{E^+}} \left(P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right)\right)^{\perp}.$$

Thus, assuming that (6) holds,

$$R^n = C(X) \oplus C(X)^{\perp} \subseteq \sum_{Z \in \Xi_{E^+}} \left(P_{C(X)^{\perp}}\left(C(Z)^{\perp}\right)\right)^{\perp}.$$

Hence, (6) implies (8) and the equivalence of (6) and (7) follows. □

Part C. The proof of Theorem 1 follows directly from Lemmas 3 and 4 below. The proof of Lemma 3 is straightforward and thus omitted.

LEMMA 3. *Let G be a graph with vertex set V. Let $C = \{C_1, \ldots, C_k\}$ be the collection of cliques of G. Then we have the following.*

(i) *C is a cover for V.*

(ii) *If* $P = \{A_1, \ldots, A_m\}$ *is a partition of V which is consistent with G, then for each* $i = 1, \ldots, m$, $A_i \subseteq C_j$ *for some* $j \in \{1, \ldots, k\}$.

LEMMA 4. *Let* $C = \{C_1, \ldots, C_k\}$ *be a cover of a set T. Suppose L is the collection of all partitions P of T satisfying: if* $P = \{A_1, \ldots, A_m\}$ *then for each* $i = 1, \ldots, m$ *there exists* $j \in \{1, \ldots, k\}$ *such that* $A_i \subseteq C_j$.

(i) *Let* $P \in L$ *with cardinality of P equal to* $m \geq k + 1$. *Then there exists* $P^* \in L$ *such that P is a refinement of* P^* *and the cardinality of* P^* *is equal to* $m - 1$.

(ii) *Suppose the cover C is minimal. If* $P = \{A_1, \ldots, A_m\} \in L$, *then the cardinality of P is greater than or equal to k.*

PROOF. (i) Since $m > k$, there exists $1 \leq a, b \leq m$ with $a \neq b$ and $1 \leq j \leq k$ such that $A_a \subseteq C_j$ and $A_b \subseteq C_j$. Thus, $P^* = \{A_i | A_i \in P,\ 1 \leq i \leq m,\ i \neq a,\ i \neq b\} \cup \{A_a \cup A_b\} \in L$ with cardinality $m - 1$.

(ii) For each $i = 1, \ldots, m$, choose $j_i \in \{1, \ldots, k\}$ such that $A_i \subseteq C_{j_i}$. Let ϕ: $\{i | 1 \leq i \leq m\} \to \{j | 1 \leq j \leq k\}$ where $\phi(i) = j_i$. For each $j = 1, \ldots, k$, choose $t_j \in C_j \setminus \bigcup_{a \neq j} C_a$. Define ψ: $\{j | 1 \leq j \leq k\} \to \{i | 1 \leq i \leq m\}$ by $t_j \in A_{\psi(j)}$. Since $t_j \notin C_a$ for $1 \leq a \leq k$ and $a \neq j$, it follows that $A_{\psi(j)} \not\subseteq C_a$ for $a \neq j$. Thus $\phi(\psi(j)) = j$. Hence, ϕ is onto and $m \geq k$. □

Part D. The proof of Theorem 2 follows directly from Theorems 3 and 4 below. In the following, let $X = (x_1, \ldots, x_p)$ where each x_i is a vector in R^n with $p \leq n$, and identify X with a point in R^{np}. As in Section 2, let P_A: $R^n \to R^n$ denote the orthogonal projection onto a subspace $A \subseteq R^n$. The proof of Lemma 5 is straightforward and thus omitted.

LEMMA 5. *Suppose* $P(z_1, \ldots, z_N)$ *is a polynomial function on* R^N *which is not identically zero. Then the set* $\{z \in R^N: P(z) \neq 0\}$ *is open and dense in* R^N.

LEMMA 6. *There exists an open dense subset* $O \subseteq R^{np}$ *such that* $X \in O$ *implies dimension* $C(X) = p$.

PROOF. Consider X as a variable point in R^{np} and let $P(X)$ be the determinant of the upper p rows of X. Since there are X for which P is not zero, the set of X such that dimension $C(X) = p$ is open and dense in R^{np} by Lemma 5. □

LEMMA 7. *Suppose B is a subspace of* R^n *with dimension* $B + p \leq n$. *Then the set of X for which* $B \cap C(X) = \{0\}$ *contains an open dense subset in* R^{np}.

PROOF. Let $\{v_1, \ldots, v_k\}$ be a basis for B and extend this collection to a linearly independent set denoted by $\{v_1, \ldots, v_k, v_{k+1}, \ldots, v_{n-p}\}$. Next let V be the $n \times (n - p)$ matrix with columns given by v_j, $j = 1, \ldots, n - p$, and

define $P(X)$ as the determinant of the $n \times n$ matrix (V, X). Since there are X for which P is not zero, the set of X such that $B \cap C(X) = \{0\}$ contains an open dense subset in R^{np} by Lemma 5. □

In the following proofs, the fact that in a complete metric space the intersection of finitely many open dense subsets is open and dense will be used.

LEMMA 8. *Suppose A is a subspace of R^n with dimension $A = c \geq p$. Then there exists an open dense subset $O_A \subseteq R^{np}$ such that $X \in O_A$ implies dimension $P_A C(X) =$ dimension $C(X) = p$.*

PROOF. First note that dimension $P_A C(X) =$ dimension $C(X)$ if and only if $A^{\perp} \cap C(X) = \{0\}$, with dimension $C(X) \leq p \leq c$ and dimension $A^{\perp} = n - c$. This follows by considering P_A: $C(X) \to A$ so that dimension $C(X) =$ dimension(kernel $P_A \cap C(X)$) + dimension $P_A C(X) =$ dimension $(A^{\perp} \cap C(X)) +$ dimension $P_A C(X)$. Next note by Lemma 7 there exists an open dense subset $V_A \subseteq R^{np}$ such that $X \in V_A$ implies $A^{\perp} \cap C(X) = \{0\}$. Hence, dimension $P_A C(X) =$ dimension $C(X)$. Now take $O_A = V_A \cap O$ where, by Lemma 6, $O \subseteq R^{np}$ is open and dense such that $X \in O$ implies dimension $C(X) = p$. □

LEMMA 9. *Suppose A is a subspace of R^n with dimension $A = c < p$. Then there exists an open dense subset $O_A \subseteq R^{np}$ such that $X \in O_A$ implies dimension $P_A C(X) = c$ and dimension $C(X) = p$.*

PROOF. Let $\hat{X} = (x_1, \ldots, x_c) \in R^{nc}$. By Lemma 8, there exists an open dense subset $\hat{O}_A \subseteq R^{nc}$ such that $\hat{X} \in \hat{O}_A$ implies dimension $P_A C(\hat{X}) =$ dimension $C(\hat{X}) = c$. Now take $O_A = \hat{O}_A \times R^{n(p-c)} \cap O$ where, by Lemma 6, $O \subseteq R^{np}$ is open and dense such that $X \in O$ implies dimension $C(X) = p$. Note also that O_A is open and dense in R^{np}, and since $P_A C(\hat{X}) \subseteq P_A C(X) \subseteq A$, dimension $P_A C(X) = c$. □

THEOREM 3. *Suppose $A_1, \ldots, A_s$ are subspaces of R^n. Then there exists an open dense subset $O \subseteq R^{np}$ such that $X \in O$ implies dimension $C(X) = p$ and dimension $P_{A_i} C(X) = p$ if $p \leq$ dimension A_i, while dimension $P_{A_i} C(X) =$ dimension A_i if $p >$ dimension A_i, $i = 1, \ldots, s$.*

PROOF. Let O_{A_i}, $i = 1, \ldots, s$, be given by Lemma 8 or Lemma 9, and take $O = \bigcap_{i=1}^{s} O_{A_i}$. □

The following generalization of Theorem 3 may be used to cover, for example, the case in which the predictor matrix provides for an intercept in the model.

THEOREM 4. *Suppose $A_1, \ldots, A_s$ are subspaces of R^n and also let Y be a subspace of R^n where $Y \subseteq A_i$ for each $i = 1, \ldots, s$. Suppose dimension $Y = k$*

and consider $X = (x_1, \ldots, x_p) = (y_1, \ldots, y_k, z_1, \ldots, z_{p-k})$ *where* $\{y_1, \ldots, y_k\}$ *is a fixed basis for* $Y, k < p \leq n$. *Then there exists an open dense subset* $O \subseteq R^{n(p-k)}$ *such that* $Z = (z_1, \ldots, z_{p-k}) \in O$ *implies dimension* $C(X) = p$ *and dimension* $P_{A_i}C(X) = p$ *if* $p \leq$ *dimension* A_i, *while dimension* $P_{A_i}C(X) =$ *dimension* A_i *if* $p >$ *dimension* A_i, $i = 1, \ldots, s$.

PROOF. First note by Lemma 7 there exists an open dense subset $O_1 \subseteq R^{n(p-k)}$ such that $Z \in O_1$ implies $Y \cap C(Z) = \{0\}$. Then $C(X) = Y \oplus C(Z)$ and $A_i^{\perp} \cap C(X) = \{0\}$ if and only if $(A_i^{\perp} \oplus Y) \cap C(Z) = \{0\}$, $i = 1, \ldots, s$. These observations allow the analysis for Theorem 3 to be used to conclude the proof of Theorem 4.

Part E. Consider a disconnected graph $G = (V, E)$ which breaks up into q connected components with

$$(V, E) = (V_1, E_1) \cup (V_2, E_2) \cup \cdots \cup (V_q, E_q).$$

In this case there is a subspace R which is common to $C(Z)$ for every Z consistent with G. In fact, $R = \bigcap_{Z \in \Xi_G} C(Z) = \bigcap_{Z \in \Xi_o} C(Z)$ and dimension $R = q$. If $q > p =$ dimension $C(X)$ then $\bigcap_{Z \in \Xi_o} C(X)^{\perp} \cap C(Z) = C(X)^{\perp} \cap R \neq \{0\}$. Furthermore, $R = C(Z_R)$ where Z_R is the grouping matrix corresponding to the partition $\{V_1, V_2, \ldots, V_q\}$. Note that Z_R need not be an element of Ξ_G. Thus, $C(X)^{\perp} \cap R$ is common to every alternative model of the type H_a^B under consideration and $C(X) \oplus (C(X)^{\perp} \cap C(Z)) = C(X) \oplus (R \cap C(X)^{\perp}) \oplus (R^{\perp} \cap C(X)^{\perp} \cap C(Z))$ for all $Z \in \Xi_G$. Consequently, $l_Z = 1$ for all $Z \in \Xi_G$ whenever τ has the form given by (4) and G is disconnected as specified above, and hence provides no discrimination between the candidate groupings. Thus, in comparing the atoms consistent with the graph, l_Z may be modified as

$$l_Z^* = \inf\left\{ \frac{\|v\|^2}{\tau(v)} : v \in R^{\perp} \cap C(X)^{\perp} \cap C(Z), v \neq 0 \right\}.$$

Note that τ is not affected by R.

Part F. Let G be a graph determined by a family of overlapping subsets $\{S_1, \ldots, S_m\}$ in R^p as discussed in Section 3. Let $V = \{v_1, \ldots, v_n\}$ denote the set of vertices of G and let $V_i = S_i \cap V$, $i = 1, \ldots, m$. Conditions will be given under which $\{V_1, \ldots, V_m\}$ is the collection of cliques of G. We assume that (1) $\cup V_i = V$ and (2) $V_i \neq V_j$ with $V_i \setminus V_j \neq \phi$ and $V_j \setminus V_i \neq \phi$ for $i \neq j$. Next recall that for $v_i, v_j \in V$ there is an edge joining v_i to v_j if and only if $v_i, v_j \in V_k$ for some $k = 1, \ldots, m$. Thus, the sets V_i, $i = 1, \ldots, m$, themselves form complete subgraphs. The question that remains is whether every complete subgraph is contained in some V_i. That is, $\{V_1, \ldots, V_m\}$ is not the collection of cliques of G if and only if there exists a subset $O \subseteq V$ which induces a complete subgraph of G and O is not contained in any of the sets V_i. Suppose such subsets O exist. From all such subsets choose O_0 to be of minimal cardinality. The cardinality of O_0 must be at least three since any

set of cardinality two which has its points joined by an edge must lie in one of the sets V_i by construction of the graph according to the overlapping grid approach. Next choose distinct vertices v_{01}, v_{02} and v_{03} from O_0, and note by the minimal cardinality of O_0 that there exist sets V_{01}, V_{02} and V_{03} from the collection $\{V_1, \ldots, V_m\}$ such that

$$O_0 \setminus \{v_{01}\} \subseteq V_{01}, v_{01} \notin V_{01},$$
$$O_0 \setminus \{v_{02}\} \subseteq V_{02}, v_{02} \notin V_{02},$$
$$O_0 \setminus \{v_{03}\} \subseteq V_{03}, v_{03} \notin V_{03}.$$

The preceding discussion establishes the following lemma.

LEMMA 10. *If the set of cliques of G is not the collection* $C = \{V_1, \ldots, V_m\}$ *then there exists three distinct vertices* v_{01}, v_{02} *and* v_{03} *and three distinct sets* V_{01}, V_{02} *and* V_{03} *from C such that*

$$\{v_{02}, v_{03}\} \subseteq V_{01}, v_{01} \notin V_{01},$$
$$\{v_{01}, v_{03}\} \subseteq V_{02}, v_{02} \notin V_{02},$$
$$\{v_{01}, v_{02}\} \subseteq V_{03}, v_{03} \notin V_{03}.$$

The following corollary follows directly from Lemma 10 and provides conditions under which $\{V_1, \ldots, V_m\}$ is the collection of cliques of G.

COROLLARY 1. *If there does not exist such a configuration* $(v_{01}, v_{02}, v_{03}, V_{01}, V_{02}, V_{03})$ *as described in Lemma* 10, *then the collection of cliques of G is given by* $\{V_1, \ldots, V_m\}$.

The preceding methodology for determining the cliques of a graph derived by the overlapping grid approach is now applied to Example 2. In this example there are four overlapping grid elements in R^2 given by S_{11}, S_{21}, S_{12} and S_{22}. The corresponding sets V_i, $i = 1, \ldots, 4$, may be identified as

$$V_1 = S_{11} \cap V = \{1, 3, 15\},$$
$$V_2 = S_{21} \cap V = \{2, 4, 6, 11, 14\},$$
$$V_3 = S_{12} \cap V = \{3, 5, 7, 9, 12, 15, 16\},$$
$$V_4 = S_{22} \cap V = \{4, 5, 8, 10, 11, 12, 13, 14\}.$$

Note for example that $V_1 \cap V_4 = \phi$ and $V_2 \cap V_3 = \phi$ so that by choosing any three distinct sets V_{01}, V_{02}, V_{03} from the collection $\{V_1, \ldots, V_4\}$ there exists at least one pair with empty intersection. Thus, there cannot exist a configuration of the type described in Lemma 10. Hence, by Corollary 1, the collection of cliques of G in Example 2 is given by $\{V_1, \ldots, V_4\}$.

REFERENCES

ATKINSON, A. C. (1972). Planning experiments to detect inadequate regression models. *Biometrika* **59** 275–293.

ATKINSON, A. C. and FEDOROV, V. V. (1975). The design of experiments for discriminating between two rival models. *Biometrika* **62** 57–70.

ATWOOD, C. L. and RYAN, T. A., JR. (1977). A class of tests for lack of fit to a regression model. Unpublished manuscript.

BREIMAN, L. and MEISEL, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71** 301–307.

CHRISTENSEN, R. R. (1989). Lack of fit based on near or exact replicates. *Ann. Statist.* **17** 673–683.

CHRISTENSEN, R. R. (1991). Small sample characterizations of near replicate lack of fit tests. *J. Amer. Statist. Assoc.* **86** 752–756.

CONSTANTINE, G. M. (1987). *Combinatorial Theory and Statistical Design*. Wiley, New York.

DANIEL, C. and WOOD, F. S. (1980). *Fitting Equations to Data*, 2nd ed. Wiley, New York.

DETTE, H. (1994). Discrimination designs for polynomial regression on compact intervals. *Ann. Statist.* **22** 890–903.

DETTE, H. (1995). Optimal designs for identifying the degree of a polynomial regression. *Ann. Statist.* **23** 1248–1266.

DRAPER, N. R. and SMITH, H. (1981). *Applied Regression Analysis*, 2nd. ed. Wiley, New York.

FISHER, R. A. (1922). The goodness of fit of regression formulae and the distribution of regression coefficients. *J. Roy. Statist. Soc.* **85** 597–612.

GHOSH, B. K. (1973). Some monotonicity theorems for χ^2, F and t distributions with applications. *J. Roy. Statist. Soc. Ser. B* **35** 480–492.

GREEN, J. R. (1971). Testing departure from a regression without using replication. *Technometrics* **13** 609–615.

HART, J. D. (1997). *Nonparametric Smoothing and Lack of Fit Tests*. Springer, New York.

JOGLEKAR, G., SCHUENEMEYER, J. H. and LARICCIA, V. (1989). Lack of fit testing when replicates are not available. *Amer. Statist.* **43** 135–143.

JONES, E. R. and MITCHELL, T. J. (1978). Design criteria for detecting model inadequacy. *Biometrika* **65** 541–551.

LYONS, N. I. and PROCTOR, C. H. (1977). A test for regression function adequacy. *Comm. Statist. Theory Methods* **A6** 81–86.

NEILL, J. W. and JOHNSON, D. E. (1985). Testing linear regression function adequacy without replication. *Ann. Statist.* **13** 1482–1489.

SHELTON, J. H., KHURI, A. I. and CORNELL, J. A. (1983). Selecting check points for testing lack of fit in response surface models. *Technometrics* **25** 357–365.

SHILLINGTON, E. R. (1979). Testing for lack of fit in regression without replication. *Canad. J. Statist.* **7** 137–146.

SU, J. Q. and WEI, L. J. (1991). A lack of fit test for the mean function in a generalized linear model. *J. Amer. Statist. Assoc.* **86** 420–426.

UTTS, J. M. (1982). The rainbow test for lack of fit in regression. *Comm. Statist. Theory Methods* **A11** 2801–2815.

F. R. MILLER
DEPARTMENT OF MATHEMATICS
KANSAS STATE UNIVERSITY
MANHATTAN, KANSAS 66506-0802

J. W. NEILL
B. W. SHERFEY
DEPARTMENT OF STATISTICS
KANSAS STATE UNIVERSITY
MANHATTAN, KANSAS 66506-0802
E-MAIL: jwneill@stat.ksu.edu