TESTING FOR LACK OF FIT IN NONLINEAR MODELS
BASED ON FUZZY CLUSTERS

Forrest R. Miller, James W. Neill, Kansas State University
James W. Neill, Department of Statistics, Kansas State University, Manhattan, KS 66506

Key Words: Lack of Fit, Maximin Power, Nonlinear Regression, Fuzzy Clusters, Nonreplication

1. Introduction

We consider lack of fit tests for parametric regression models

\[ y_i = f(x_i, \beta) + \varepsilon_i, \quad 1 \leq i \leq n \]

where \( f(x_i, \beta) \) is a specified regression function. The \( x_i \) are assumed to be deterministic input variables (dial settings) in \( R^q \) and, for asymptotic purposes, the observable random variables \( \{y_{ni}: 1 \leq i \leq n, n \geq 1\} \) form a triangular array of rowwise independent variables. The random errors \( \varepsilon_i \) are assumed to be identically distributed with zero mean and unknown positive finite variance \( \sigma^2 \) and \( \beta \) ranges over a subset \( B \) of \( R^p \). Let \( y_n = (y_{n1}, \ldots, y_{nn})^T, \varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{nn})^T \) and \( f_n(\beta) = (f(x_{n1}, \beta), \ldots, f(x_{nn}, \beta))^T \). For general functions \( f(x, \beta) \), the expectation surface

\[ M_c = \{m \in R^p: m = f_c(\beta), \beta \in B\}, \]

generated by \( \beta \), is a curved surface in \( R^p \). The special case of linear regression occurs when \( f(x, \beta) = \beta^T x \) (where \( g \) is a continuous vector-valued function of the \( q \) input dial settings) and gives a flat affine variety as its expectation surface. Lack of fit is said to exist when \( f_n(\beta) \) does not adequately describe the mean of \( y_n \) in the sense that \( E(yn) \) is not an element of the expectation surface. To check whether model (1) is useful for prediction and parameter inferences, it is important to examine the model for lack of fit.

Several approaches for assessing the existence of lack of fit have been proposed. Although overlapping in some aspects, these include the development of nonparametric regression techniques to test the adequacy of parametric regression models, tests associated with marked empirical processes based on residuals and graphical methods to determine the validity of regression models (see, e.g., Hart (1997), Cook and Weisburg (1997), Stute, Thies and Zhu (1998), Dette (1999), Fan and Huang (2001) and the references therein). In addition, lack of fit tests based on clusters of the data have been given. When replicate measurements (clusters) exist at the input dial settings then the classical lack of fit test introduced by Fisher (1922) may be applied to test for the existence of lack of fit in model (1). Neill (1988) presented a generalization of the classical test based on near replicate clusters, which was shown to be consistent whenever model (1) is the orthogonal projection of the true model. For normal theory linear regression models, Christensen (1989, 1991) derived uniformly most powerful invariant tests for detecting orthogonal between- and within-cluster lack of fit, given a choice of near replicate clusters of the input dial settings. In the following, we present likelihood based models and associated likelihood ratio tests for detecting general types of nonlinear lack of fit, derived from large classes of possible clusters.

The clustering based tests of Christensen (1989, 1991) and Neill (1988) require the selection of near replicate clusters, the choice of which was left open. In a series of papers, Miller, Neill and co-workers (1998, 1999, 2001) presented a statistically principled method for choosing a clustering of the input variables for linear models. Their method is based on using Christensen’s tests and involves developing an alternative model using a large class of nonparametric step functions.

The purpose of this paper is to extend these methods to the nonlinear case. In particular, we will construct statistical experiments which allow us to extend Christensen’s orthogonal between-cluster lack of fit test to the nonlinear case, permitting nonnormal errors as well. Each experiment will determine a decision, based on \( y_n \) and a likelihood ratio test (LRT), as to whether or not lack of fit exists in (1). Then we extend the methods of Miller, Neill and Sherfey (1998, 1999) to choose an optimal such experiment from a large class of such. Here we use the term statistical experiment in the sense developed by Le Cam (1972), and described by Shiryaev and Spokoiny (2000) and Le Cam and Yang (2001). The choice of an optimal alternative model involves comparing experiments and taking limits of experiments using the local asymptotic normality (LAN) of Le Cam.

2. Lack of Fit Experiments and Decisions

The regression function \( f(x, \beta) \) which gives model (1) remains fixed throughout this paper. In case lack of fit exists in (1) then we may suppose that \( E(y_n) = f_c(\beta) + \xi_n(\beta) \) where \( \xi_n(\beta) \) represents a lack of fit vector known only to be contained in some subspace of the lack of fit space \( T_pM_n^\perp \). The notation \( T_pM_n \) denotes the tangent vector space to the expectation surface \( M_n \) at \( \beta \), which is spanned by the vectors \( \partial f_c(\beta)/\partial \beta_j, 1 \leq j \leq p \). The superscript \( \perp \) denotes orthogonal complementation. We next parallel Christensen (1989, 1991) in that we construct general nonlinear alternative models by decomposing \( T_pM_n^\perp \) into orthogonal subspaces according to clusterings of the observations.
To specify a clustering of the $n$ observations into $c$ clusters we use an $n \times c$ matrix $Z_n = [z_{ij}]$. The $ij$th element of $Z_n$ assigns a measure of membership for the $i$th observation to the $j$th cluster. The weights $z_{ij}$ are allowed to take on any value in the interval $[0, 1]$, but must satisfy the constraint $\sum_{j=1}^{c} z_{ij} = 1$ for $1 \leq i \leq n$. Such a clustering is called a fuzzy clustering (see, e.g., Rousseeuw (1995), Seaver, Triantis and Reeves (1999)). As a special case, a so-called crisp clustering is obtained when each observation belongs to exactly one cluster. In this case, $Z_n$ contains only zeroes and ones, and the nonzero values in the $j$th column of $Z_n$ correspond to the observations in the $j$th cluster, $1 \leq j \leq c$. As discussed below, fuzzy clusters are used in the nonlinear case to obtain continuous and smooth alternative models for general types of lack of fit. Throughout the remainder of this paper, clusterings of the observations are considered to be fuzzy with crisp clusters specifically indicated.

In analogy with Christensen (1989, 1991) we now specifically discuss the detection of nonlinear between-cluster lack of fit.

We next discuss the class of clusterings from which $Z_n(\beta)$, $\beta \in B$, is chosen. In the linear case, as discussed by Miller, Neill and Sherfey (1998, 1999), we can take $Z_n(\beta)$ to be independent of $\beta$. In addition, the possible (crisp) clusterings for the maximin power criterion of cluster selection is determined from the dial settings $V_n = \{x_{ni}, 1 \leq i \leq n\}$, and it is shown that restricting to the atoms of this large class of possible clusterings give better power. To briefly elaborate (see Miller, Neill and Sherfey (1998, 1999) for complete discussion), the set of possible clusterings is given by the collection of clusterings consistent with a specified cover $C_n = \{C_{n1}, ..., C_{nm}\}$ of the dial settings $V_n$. That is (identifying the elements of each $C_{nj}$ with the indices of the corresponding observations), the collection of all partitions $P$ of the set of observation indices $\{1, ..., n\}$ which satisfy the condition: if $P = \{A_{i1}, ..., A_{im}\}$ then for each $1 \leq i \leq m$ there exists $1 \leq j \leq m$ such that $A_{ij} \subseteq C_{nj}$. The atoms represent those partitions consistent with the cover that group as many observations together as possible. Alternative models for linear between-cluster lack of fit constructed from the atoms lead to $F$-distributed test statistics with degrees of freedom parameters that are in concordance with the objective of maximal power. A method to determine a cover can be based on a family $F_n = \{F_{n1}, ..., F_{nm}\}$ of overlapping subsets in $R^n$ whose union includes the dial settings $V_n$. The cover elements are then given by $C_{nj} = F_{nj} \cap V_n$, $1 \leq j \leq m$. Furthermore, a method of selecting the overlapping subsets $F_n$ consists of dividing the input variable space into cells, where any cell indexed by a $q$-tuple with all odd components is referred to as an odd cell. The overlapping subsets in $F_n$ are then determined by taking the union of each nonempty odd cell with all contiguous cells. Finally, let $K_n, 1 \leq i \leq n$, be the set of indices of the odd cells that contain observations with which the $i$th observation can or must be clustered. As shown in Miller, Neill and Sherfey (1999), the set of atoms $K_{n,n}$, say, can be usefully represented as the product space $K_{n,n} = K_{n1} \times ... \times K_{nn}$ (3) where like coordinates within each particular $n$-tuple indicate that the corresponding observations are to be clustered together.

In the nonlinear case, in order to obtain continuous and smooth alternative models for between-cluster lack of fit, we choose $Z_n(\beta)$ for $\beta \in B$ from the class of fuzzy clusters given by the convex hull of the set of atoms $K_{n,n}$ as described below. We will denote this class by $K_n$ and refer to the elements as fuzzy atoms. Similar to the linear case, alternative models for nonlinear between-cluster lack of fit constructed from the fuzzy atoms lead to $(\text{asymptotically}) \chi^2$-distributed test statistics with degree of freedom parameters that are in concordance with the objective of better power. Under very general
circumstances (see Section 4), each such $Z_n(\beta)$ is $n \times c$ with $c$ constant, corresponding to the number of nonempty odd cells. In the following discussion we arbitrarily order and label the nonempty odd cells $1$ through $c$. In particular, suppose $K_i = \{ \omega_{ij} \}$ where $n_i$ is the cardinality of $K_i$ and $1 \leq \omega_{ij} \leq c$ for $1 \leq i \leq n$. Let $\text{co}(E)$ denote the convex hull of a set $E \subseteq \mathbb{R}^d$. Then $\text{co}(K_i)$ is formed by identifying $\omega_{ij}$ with the standard basis elements $e_1, \ldots, e_c$ of $\mathbb{R}^c$ and taking convex combinations of the $e_1, \ldots, e_c$ of $\mathbb{R}^c$, and taking convex combinations of the $e_{ij}, 1 \leq j \leq n_i$ in $\mathbb{R}^c$. We claim that $K_n = \text{co}(K_1) \times \ldots \times \text{co}(K_n)$. (4) This follows by first noting that the extreme points of the set on the right hand side of (4) are the atoms $K_{n\beta}$ as represented in (3). The equality in (4) then follows by the Krein-Milman theorem (see, e.g., Rudin (1991)). Next let $\sum_{j=1}^c \lambda_{ij} \omega_{ij} \in \text{co}(K_i), 1 \leq i \leq n$. By (4) and the identifications made above, the $i$th row of the corresponding $Z_n(\beta) \in K_n$ consists of $\lambda_{ij}$ as the column entries corresponding to the odd cells $\omega_{ij}, 1 \leq j \leq n_i$, and zeroes elsewhere. Since $\sum_{j=1}^c \lambda_{ij} = 1$ for $1 \leq i \leq n$, such a $Z_n(\beta)$ does in fact represent a fuzzy clustering as defined above. Thus, for the reasons given above, we consider experiments that assign a fuzzy atom $Z_n(\beta) \in K_n$ to each $\beta \in B$. As indicated above, each such $Z_n(\beta)$ is $n \times c$ with $c$ constant so that the $(\beta, \gamma)$ parameter space is the same for all experiments
\[
E_n(x_{n1}, \ldots, x_{nm}; Z_n(\beta) \in K_n, \beta \in B) = \mathcal{R}^n, A^n, \mathcal{P}^\infty_{\gamma^{\infty}}, (\beta, \gamma) \in B \times \mathcal{R}^c \times \mathcal{G}_n.
\]
We refer to such an experiment as a between-cluster lack of fit experiment (BCLFE).

3. Comparing Between-Cluster Lack of Fit Experiments
The strategy that we follow is the same as that used in the linear case by Miller, Neill and Sherfey (1998, 1999). That is, we ask if $E_1$ and $E_2$ were to give a good approximation for a ‘true model’, which experiment would give better power in detecting lack of fit? To make this comparison a positive definite quadratic form $\tau(\beta)$ of $z_n$ was introduced in the linear case and we maximized
\[
\tau(\beta) = \min \left\{ \frac{\| F_2(x) \|_{\gamma(\beta)}}{\tau(\beta)} : v \in C(Z_n) \cap C(G_n) \right\}
\]
over the atoms $Z_n \in K_n$. Here $G_n$ is the $n \times p$ matrix with $i$th row given by $g(x_{ni})^T$ as defined in Section 1, and $C(Z_n) \cap C(G_n)$ is the linear (orthogonal) between-cluster lack of fit subspace corresponding to $Z_n$ as introduced by Christensen (1991). An atom $Z_n$ is called a maximin power clustering, and determines an optimum alternative model for testing linear between-cluster lack of fit.

In the nonlinear case we consider
\[
\tau(\beta) = \min \left\{ \frac{\| F_2(x) \|_{\gamma(\beta)}}{\tau(\beta)} : v \in C(B, Z_n, \gamma, \theta) \cap T_{\gamma} \right\}
\]
for each BCLFE where $\tau(\beta)$ represents a positive definite quadratic form on $T_{\gamma}$. Recall that the columns of $B_{\gamma}$ give a basis for $C(Z_n(\beta)) \cap T_{\gamma}$. We will consider $\tau(\beta)$ to be of the same form as constructed in the linear case. In particular,
\[
\tau(\beta) = \sum_{\omega_{ij} \in \mathcal{E}_i} w_{\omega_{ij}} \left( P_{\omega_{ij}} x \right)^T
\]
for $v \in T_{\gamma} \cap \mathcal{E}_i$ with $w_{\omega_{ij}} \geq 0$ and $\sum_{\omega_{ij} \in \mathcal{E}_i} w_{\omega_{ij}} = 1$. $\mathcal{E}_i$ denotes the clusterings consistent with a given cover $C_n$ that cluster only two observations, with all other observations being singleton clusters. Also, $P_{\omega_{ij}}$ denotes the orthogonal projection operator onto the subspace $C(B, Z_n, \gamma, \theta)$. Let $X_n$ denote the $n \times q$ matrix of input dial settings with $i$th row given by $x_{ni}^T$. To incorporate nearness as measured by $\| X_n - X_n(\beta) \|_2^2$ into the weight $w_{\omega_{ij}}$, we let
\[
w_{\omega_{ij}} = \frac{\| X_n - X_n(\beta) \|_2^2}{\| X_n - X_n(\beta) \|_2^2}
\]
where $X_n(\beta) = P_{\omega_{ij}} X_n$ and $X_n(\beta) = P_{\omega_{ij}} X_n$ for $Z_n \in \mathbb{Z}_n$. The notation $\| A \|_2^2$ denotes the squared matrix norm defined by $\sum_{i=1}^n \sum_{j=1}^q a_{ij}^2$ where $a_{ij}$ is the $ij$th element of an $n \times q$ matrix $A$. Note that the weights in $\tau(\beta)$ are determined by $x_{ni}^T, \ldots, x_{nm}^T$ and do not depend upon $\beta$, nor do the $Z_n \in \mathbb{Z}_n$. However, the projections involve $C(B, Z_n, \gamma, \theta)$ and thus $\tau(\beta)$ does depend upon $\beta$.

For the nonlinear case let $Z_n(\beta)$, say, maximize $\tau(\beta)$ over the fuzzy atoms $K_n$ for each $\beta \in B$. Note that $Z_n(\beta)$ can be considered to be an optimal power selection clustering for testing the approximating linear model around $\beta = \beta_0$. We claim (see Section 4) that using $Z_n(\beta)$ for each $\beta$ provides (asymptotically) the optimum BCLFE for testing lack of fit in model (1). An experiment $E_n(x_{n1}, \ldots, x_{nm}; Z_n(\beta), \beta \in B)$ will be called a maximin between-cluster lack of fit experiment (MMBCLFE).
4. Asymptotics and the Parameter Space

In order to see more precisely the role of the $\mathbf{B}_{\mathbf{Z}_{\mathbf{B}}}$ in the asymptotic limit, we need to describe them in terms of the $(\beta, \gamma)$ parameter space instead of the observed data space $\mathbb{R}^n$. A choice of the basis vectors $\mathbf{B}_{\mathbf{Z}_{\mathbf{B}}}$ for $C_{\mathbf{Z}_{\mathbf{B}}} \cap \mathbb{T}_{\mathbf{p}} \mathbb{M}_{\mathbf{Z}_{\mathbf{B}}}$ determines a mapping $\Phi_{\mathbf{Z}_{\mathbf{B}}}: \mathbb{B} \times \mathbb{R}^{n-p} \to \mathbb{R}^n$ by

$$\Phi_{\mathbf{Z}_{\mathbf{B}}}(\beta, \gamma) = \mathbf{f}(\beta) + \mathbf{B}_{\mathbf{Z}_{\mathbf{B}}}(\gamma),$$

The Fisher information metric for the expectation surface parameters is

$$I_{\mathbf{Z}_{\mathbf{B}}}(\beta, \gamma) = \pi \left( \sum_{i=1}^{n} \frac{\partial \Phi_{\mathbf{Z}_{\mathbf{B}}}}{\partial \beta_i} \frac{\partial \Phi_{\mathbf{Z}_{\mathbf{B}}}}{\partial \gamma_i} \right),$$

where $\theta = (\beta, \gamma)$ and $\kappa = \pi$, \( d\mu = \kappa \mathbb{E}_{\mathbf{Z}_{\mathbf{B}}}(\gamma) \) (Euclidean metric). Here $\Phi_{\mathbf{Z}_{\mathbf{B}}}$ denotes the pull-back operation under $\Phi_{\mathbf{Z}_{\mathbf{B}}}$. Note that $\kappa$ is independent of $n$, the $x_{ni}$ for $1 \leq i \leq n$, and $Z_{\mathbf{B}}(\beta)$.

We now choose the columns of $\mathbf{B}_{\mathbf{Z}_{\mathbf{B}}}$ to be an orthonormal basis of $C_{\mathbf{Z}_{\mathbf{B}}} \cap \mathbb{T}_{\mathbf{p}} \mathbb{M}_{\mathbf{Z}_{\mathbf{B}}}$ with respect to $\tau(\beta)$. It then follows that along the space $\gamma = 0$ in the $\gamma$-direction, $\Phi_{\mathbf{Z}_{\mathbf{B}}}(\theta)$ is just the Euclidean metric with respect to the parameter $\beta$. But this is just the direction in which we calculate $I_{\mathbf{Z}_{\mathbf{B}}}(\beta, 0)$.

The $\frac{1}{n}$ is included because the LAN analysis shows that the convergence $\frac{1}{n} I_{\mathbf{Z}_{\mathbf{B}}}(\beta, 0) \to I(\beta, 0)$ is required (and is in fact obtainable) for the local type of LAN convergence of experiments.

Thus, suppose we have an array of values of the input variables $x_{ni}$ for $1 \leq i \leq n$ and $n \geq 1$. A simplifying assumption is that $x_{ni} = x_{ni}$ for $1 \leq i \leq m \leq n$ but this is not necessary. The determination of the atoms $\mathcal{K}_{\mathbf{m}}$ at stage $n$, as discussed in Section 2 and described in detail by Miller, Neill and Sherfey (1999), is made by dividing the input space into cells and considering the occupancies of the odd cells and their neighbors. This determines $c = \dim(C_{\mathbf{Z}_{\mathbf{B}}}(\beta))$ for each fuzzy atom $\mathbf{Z}_{\mathbf{B}}(\beta) \in \mathcal{K}_{\mathbf{m}}$. We assume that the occupancy pattern (as far as empty or not) is the same for each $n$. In this case we say that the input data has the same shape for each $n$. By choosing, for each $n$, any function $\beta \to Z_{\mathbf{B}}(\beta)$ we obtain a sequence of BCLFEs $\mathbf{E}_{\mathbf{Z}_{\mathbf{B}}}(x_{ni}, ..., x_{ni}; Z_{\mathbf{B}}(\beta), \beta \in \mathcal{K}_{\mathbf{m}} \beta \in \mathbf{B})$. Of course we have a special sequence of interest, the MMBCCLFEs $\mathbf{E}_{\mathbf{Z}_{\mathbf{B}}}(x_{ni}, ..., x_{ni}; Z_{\mathbf{B}}(\beta), \beta \in \mathbf{B})$.

Theorem 1 below provides the asymptotic distributional properties of the LRT for testing $H_0: \gamma = 0$ versus $H_1: \gamma \neq 0$ in a sequence of BCLFEs. The proof is given in the appendix and is based in part on the requirement that the sequence of BCLFEs be LAN. A key condition used to ensure LAN is that the sequence of BCLFEs satisfy the following definition.

Definition 1. A sequence of BCLFEs $\mathbf{E}_{\mathbf{Z}_{\mathbf{B}}}(x_{ni}, ..., x_{ni}; Z_{\mathbf{B}}(\beta) \in \mathcal{K}_{\mathbf{m}} \beta \in \mathbf{B})$ is uniformly differentiable in quadratic mean at $\theta = (\beta, \gamma)$ if for each $1 \leq i \leq n$ and $n \geq 1$ there exists a measurable function $\mathbf{g}_{\mathbf{Z}_{\mathbf{B}}}(\theta)$ such that

$$\int_{\mathbb{R}^c} \left( \sqrt{\mathbf{g}_{\mathbf{Z}_{\mathbf{B}}}(\theta)} - \sqrt{\mathbf{g}_{\mathbf{Z}_{\mathbf{B}}}(\theta)} - \frac{1}{2} \frac{d\mathbf{u}_{\mathbf{Z}_{\mathbf{B}}}}{d\mathbf{u}_{\mathbf{Z}_{\mathbf{B}}}} \right)^2 d\mathbf{u}_{\mathbf{Z}_{\mathbf{B}}} = \mathbb{E}(\mathbf{g}_{\mathbf{Z}_{\mathbf{B}}})$$

independent of $n$ and $i$ as $\mathbf{u}_{\mathbf{Z}_{\mathbf{B}}} \to \theta$ in $\mathbb{R}^c$, where $p_{\mathbf{Z}_{\mathbf{B}}} = \mathbb{E}(\mathbf{Z}_{\mathbf{B}})$ denotes the density of $\mathbf{Z}_{\mathbf{B}}$.

The following lemma gives mild conditions under which uniform differentiability in quadratic mean obtains. The proof is given in the appendix.

Lemma 1. Let $s_{ni}(\mathbf{y}, 2) = s(\mathbf{y} - \mathbf{M}_{ni}(2))$ for $1 \leq i \leq n$ and $n \geq 1$ where $s(\epsilon) = \sqrt{n(\epsilon)}$. Fix $\theta$ and $\delta > 0$ and suppose that there exists an $M > 0$ such that

$$|\phi_{\mathbf{Z}_{\mathbf{B}}}(\epsilon)| \leq M, \left| \frac{\partial \phi_{\mathbf{Z}_{\mathbf{B}}}(\epsilon)}{\partial \theta} \right| \leq M,$$

for all $n$ and $i$ where $\|\phi - \theta\| < \delta$. In addition, let $I(\mathbf{y}) = \sup_{\mathbf{E}_{\mathbf{Z}_{\mathbf{B}}}} |s(\mathbf{y} - \mathbf{E}_{\mathbf{Z}_{\mathbf{B}}})| + |s''(\mathbf{y} - \mathbf{E}_{\mathbf{Z}_{\mathbf{B}}})|$ and suppose

$$\mathbb{E}(I(\mathbf{y})) < \infty.$$ Then the corresponding sequence of BCLFEs is uniformly differentiable in quadratic mean at $\theta$.

Theorem 1. Suppose the conditions of Lemma 1 hold at $\theta = (\beta, 0)$ for a sequence of BCLFEs $\mathbf{E}_{\mathbf{Z}_{\mathbf{B}}}(x_{ni}, ..., x_{ni}; Z_{\mathbf{B}}(\beta) \in \mathcal{K}_{\mathbf{m}} \beta \in \mathbf{B})$ and that the $\frac{1}{n} I_{\mathbf{Z}_{\mathbf{B}}}(\beta, 0)$ converge to $I(\beta, 0)$.

Also let

$$\int \frac{P}{p} d\mathbb{P} < \infty, j = 2, 4, \text{ where } d\mathbb{P} = p.$$ If the unrestricted and restricted (by $H_o$) maximum likelihood estimators for $\theta$ are $\sqrt{n}$-consistent under $(\beta, 0)$, then the sequence of LRT statistics for the decisions converge under $(\beta, \gamma/\sqrt{n})$ in distribution to a noncentral $F$ with noncentrality parameter $I(\beta, 0)(\gamma, \gamma)$.

Since the power is an increasing function of the noncentrality parameter, Theorem 1 indicates the
importance of the $i_{z,0}$. The convergence $\frac{1}{n} I_n(\beta, 0) \to I(\beta, 0)$ implies the convergence of $i_{z,0} \to \hat{i}(\beta)$, say, since $i_{z,0}$ is the minimum eigenvalue of $\frac{1}{n} I_n(\beta, 0)$ in the $\gamma$-direction. Now consider the MMBCLFE sequence $B_n(x_{1i}, \ldots, x_{ni}, Z_n^*(\beta), \beta \in B)$ with its information metric $\frac{1}{n} I_n(\beta, \gamma)$ and $i_{z,0}$. Let $i^{M_n^{*}}(\beta) = \liminf i_{z,0}$. Theorem 2. If $E_n(x_{1i}, \ldots, x_{ni}; Z_n^*(\beta) \in K_n, \beta \in B)$ is any sequence of BCLFEs as described above satisfying the hypotheses of Theorem 1, then $i^{M_n^{*}}(\beta) \geq \hat{i}(\beta) = \lim i_{z,0}$.

Theorem 2, together with Theorem 1, justifies the conclusion that for any $n$, the MMBCLFE is a good choice to use in testing for lack of fit in terms of giving optimal power.

5. Implementation

For a given data size $n$ we approximate the MMBCLFE as follows:

a) choose a finite set $\beta_1, \ldots, \beta_N$ from $B$ and at these points take $Z_n^*(\beta)$ to be the crisp atom in $K_n$ which gives the largest $i_{z,0}$;

b) use the fact that the set of fuzzy atoms $K_n$ is convex to give an algorithm for calculating $Z_n^*(\beta)$, say, for the remaining $\beta \in B$.

We thus obtain an approximating sequence of BCLFEs $E_n(x_{1i}, \ldots, x_{ni}, Z_n^*(\beta), \beta \in B)$ which we will use in place of the corresponding $E_n(x_{1i}, \ldots, x_{ni}; Z_n^*(\beta), \beta \in B)$ for implementation purposes. Notice that this method will give the crisp atom maximin power clustering of Miller, Neill and Sherfey (1998, 1999) in the linear case since the choices in step a will always give the same crisp atom.

To use the data $y_n$ to calculate the LRT statistic we assume (for illustration) that the random errors $\epsilon_{ni}$ are normally distributed. Then it is a matter of finding the nearest point to $y_n$ on the expectation surface $M_n$ and also the nearest point to $y_n$ on the expectation surface $M_n^*$ for the alternative (full) model provided by $E_n(x_{1i}, \ldots, x_{ni}; Z_n^*(\beta), \beta \in B)$. Clearly, $M_n \subset M_n^*$. To find the nearest point to $y_n$ on $M_n$ we use the modified Gauss-Newton least squares estimation as described by Seber and Wild (1989). The nearest point calculation for $M_n^*$ can be reduced to a $\beta$-space calculation due to the fact that $M_n^*$ is fibered across $M_n$ by affine varieties. For each $\beta$ let $V_\beta = m \in \mathbb{R}^n; m = f_\beta(\gamma) + B_{z,0}\gamma; \gamma \in \mathbb{R}^{n-\gamma}$. Now suppose $(\beta, \gamma)$ gives the nearest point $y_n$ to $y_n$ on $M_n^*$. Then $y_n \in V_\beta$ and $y_n$ is the nearest point to $y_n$ on $V_\beta$. In fact, for each $\beta$ let $Y(\beta)$ give the nearest point to $y_n$ on $V_\beta$ and let $M_n^{*} = \{ m \in \mathbb{R}^n; m = f_\beta(\gamma) + B_{z,0}\gamma; \beta \in B_0 \}$ consisting of the nearest points to $y_n$ on the fibers $V_\beta$. Note that since each $V_\beta$ is an affine variety, standard linear model projection algorithms determine the parameter values $\gamma(\beta)$ for each $\beta$. Thus we have a calculational procedure for obtaining any point on $M_n^{*}$ coming from a particular value of $\beta$. This is all we need to apply the modified Gauss-Newton algorithm mentioned above to find the nearest point to $y_n$ on $M_n^{*}$, and hence on $M_n^*$.

References


