Statistical Science
2003, Vol. 18, No. 3, 394–410

Inference Functions and Quadratic Score Tests

Bruce G. Lindsay and Annie Qu

Abstract. A general expository description is given of the use of quadratic score test statistics as inference functions. This methodology allows one to do efficient estimation and testing in a semiparametric model defined by a set of mean-zero estimating functions. The inference function is related to a quadratic minimum distance problem. The asymptotic chi-squared properties are shown to be the consequences of asymptotic projection properties. Shortcomings of the asymptotic theory are discussed and a bootstrap method is shown to correct for anticonservative testing behavior.

Key words and phrases: Bootstrapping, chi-squared test, Edgeworth expansion, generalized estimating equation, generalized method of moments, likelihood, quadratic inference function, quasi-likelihood, semiparametric model.

1. INTRODUCTION

In the years since C. R. Rao developed the score test (1948), there has been a wide diversity of research connected with this procedure. Although the original work was in the area of parametric statistical inference, we will instead focus here on the following line of research within semiparametric inference based on estimating equations.

A quadratic form test statistic, such as Rao's, has some surprising properties when it is treated as an *inference function* instead of a test statistic. That is, when it is treated as a function of the parameters instead of the data, one can use it as if it were a likelihood function, deriving chi-squared tests of goodness of fit, profile tests of parameter components and so forth.

This methodology has roots in the statistics literature dating back over 50 years, but its application since that time has been minimal except for the world of econometrics, where its use has blossomed under the name of the "generalized method of moments," or GMM (Hansen, 1982). However, it is a method whose applicability in the world of "estimating functions" (Godambe, 1960) surely exceeds the domain of economic examples, and so we here attempt to introduce the main ideas to a wider audience of statisticians.

We will review this work here, add some clarifying points and point to the large range of possible applications. This methodology is particularly fruitful when there are more equations than unknowns, as the inference function then provides a simple but optimal method for combining the equations. We can think of each estimating equation that is added to the quadratic form statistic as being an additional model assumption. The quadratic inference function then, just as a likelihood function, provides an optimal estimation and testing method for the corresponding semiparametric model.

Indeed, the method is so flexible that it can be used to suit many statistical purposes. In our review of this area, we will show that it can be used to minimize assumptions, increase information about parameters and evaluate goodness of fit. Along the way, we will see that the methodology can be analyzed from parametric, semiparametric and nonparametric functional points of view. Our goal will be to carefully distinguish among these three points of view. To our knowledge, this last perspective is new to the literature. Throughout the paper, our goal is to present the development of ideas, and

Bruce G. Lindsay is Distinguished Professor, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802 (e-mail: bgl@psu.edu). Annie Qu is Assistant Professor, Department of Statistics, Oregon State University, Corvallis, Oregon 97331 (e-mail: qu@stat.orst.edu).

so we suppress technical derivations when they do not provide additional insight.

Section 2 introduces the construction of basic scores under parametric, semiparametric and nonparametric settings. Section 3 presents an overview of the asymptotic properties of the quadratic inference function (QIF) and evaluates its performance based on efficiency and robustness. Section 4 provides a brief history of the QIF and relates it to minimum chi-squared, Rao's score test, generalized least squares and empirical likelihood. Section 5 examines the QIF more closely and illustrates the projection representations of the QIF. Section 6 presents bootstrap strategies to improve higher-order accuracy for QIF test size.

2. CONSTRUCTION

We start with a vector of *basic score functions* $\mathbf{b}(x,\theta)$ that are assumed to have mean 0 when θ is the true parameter. These basic scores are also called *estimating functions* or *moment conditions* in the statistics and econometrics literature, respectively. It is a key element of the QIF method that there are more distinct score components $b_j,\ j=1,\ldots,q$, than there are distinct parameters $\theta_k,\ k=1,\ldots,p$. Thus, we cannot simply set the scores to 0 to solve for the unknown parameter θ.

To motivate our method, we investigate a real data example found in Rotnitzky and Wypij (1994), Chen and Little (1999) and Qu and Song (2002). The data concern studies of pediatric asthma in Steubenville, Ohio. Dichotomous outcomes recording asthma status were recorded for children at ages 9 and 13. The marginal probability will be modeled as a logistic regression (Rotnitzky and Wypij, 1994) with gender and age as covariates:

$$\operatorname{logit}\{\operatorname{pr}(y_{it}=1)\}=\theta_0+\theta_1 I(\text{male})+\theta_2 I(\text{age}=13),$$

where $y_{it}=1$ if the ith child had asthma at time $t=1,2$ and $I(E)$ is the indicator function for event E. About 20% of the children had asthma status missing at age 13. That is, every child had his or her asthma status recorded at age 9, but for some the asthma status was missing at age 13. Note that there are three parameters θ_0,θ_1 and θ_2 in the model with complete observations, but only two identifiable parameters, θ_0 and θ_1, for the incomplete case.

If we are interested in estimating θ_0, θ_1 and θ_2 using all subjects, it is not obvious how to combine $\hat\theta_0$, $\hat\theta_1$ and $\hat\theta_2$ from complete and incomplete data optimally, especially where dimensions of parameters are different for different missing patterns. It is natural to create the basic score functions

$$\mathbf{b}_1(\theta_0,\theta_1,\theta_2)=\sum (X_i^c)'(y_i^c-\mu_i^c),$$

$$\mathbf{b}_2(\theta_0,\theta_1)=\sum (X_i^m)'(y_i^m-\mu_i^m),$$

where (X^c,y^c) and (X^m,y^m) are data from complete and incomplete observations, respectively, and $\mu_i^c=\operatorname{pr}(y_i^c=1)$, $\mu_i^m=\operatorname{pr}(y_i^m=1)$. Notice that the total dimension of $\mathbf{b}_1$ and $\mathbf{b}_2$ is 5, which is more than 3, the dimension of the parameter. We might not be able to find solutions of θ_0, θ_1 and θ_2 by setting both $\mathbf{b}_1$ and $\mathbf{b}_2$ to 0, but we can estimate them by setting a weighted combination of $\mathbf{b}_1$ and $\mathbf{b}_2$ to 0, and the question becomes, what are the optimal weights for $\mathbf{b}_1$ and $\mathbf{b}_2$?

As another motivating example, consider the two basic scores

$$(1)\qquad b_1(x,\theta)=(x-\theta),\quad b_2(x,\theta)=I\{x-\theta\ge 0\}.$$

If we solve $\sum b_1(X_i;\theta)=0$ for θ, for a given set of data $X_1,\ldots,X_n$, then the solution is the sample mean. If for the same data we solve $\sum b_2(X_i;\theta)=0$, then the solution is the sample median. We cannot set them both to 0 to solve for a single θ unless the median and mean are equal. However, we might ask how we can optimally combine these two score functions to obtain a compromise estimator.

The method we consider combines the basic scores in a quadratic form test statistic, which will then be treated as an inference function. With this method, the resulting estimator is asymptotically fully efficient under both the normal model and the double exponential model, corresponding to the fact that the two basic scores b_1 and b_2 in (1) are the respective likelihood scores for these models.

2.1 The Semiparametric Model

In the settings we are considering, there can be some ambiguity about the nature of the model under investigation. We start by clarifying this issue.

Given an arbitrary distribution F and solving the equations

$$E_F[b_i(X,\theta)]=0$$

for θ in the mean-and-median example above (1) yields the mean of F for $i=1$ and the median of F for $i=2$. If there is a value of θ that solves both equations, as in a symmetric distribution with finite mean, then we will say that the pair of scores are *compatible* under F and that θ is the parameter value corresponding to F. (In

our example, uniqueness of the mean implies that the solution set has at most a single element.) If there is no value of θ for which both scores are mean 0, then the scores are *incompatible* under F.

The above definitions of compatible and incompatible can be extended to any set of q estimating equations with p unknown parameters, where $q > p$. We say a distribution is *compatible* with a vector of basic scores if there is a p-dimensional solution θ to the q-equations $E_F[\mathbf{b}(X, \theta)] = 0$. We will say that $F \in \mathcal{M}_\theta$ if there is a solution.

This in turn defines a *semiparametric model*

$$\mathcal{M} = \bigcup \mathcal{M}_\theta.$$

It consists of all distributions F that are compatible with the vector of basic scores for some θ. It is semiparametric, not nonparametric, because there are implicit restrictions on $\mathcal{M}$ arising from the fact that there are more equations q than unknowns p. For example, in the exponential distribution, the mean and median do not coincide, so this distribution is *not* in the semiparametric model $\mathcal{M}$ determined by compatibility with the mean and median equations.

Since it is possible that the semiparametric model assumptions are false, we might well ask how any procedure we develop will behave when compatibility fails. We will call this the *nonparametric* setting. For the methods we describe, the parameters, estimators and tests will have a minimum quadratic distance interpretation which gives them valid nonparametric interpretation as well.

In the end, we could potentially have three levels of models. The basic scores $\mathbf{b}(\theta)$ could have arisen from a *parametric model* of interest, such as the normal, where the mean and median scores both generate valid inference. This model would then be nested within the *semiparametric model* $\mathcal{M}$ compatible with the scores. Valid inference on θ in the semiparametric model is then an automatically valid inference in the parametric model, and we can think of the semiparametric model as being a weakening of the parametric model assumptions. Finally, because we can extend the definition of the parameter θ to the nonparametric setting using minimum distance ideas, we can consistently estimate it there and perform valid tests of hypotheses concerning its value.

2.2 The Quadratic Distance

To illustrate the idea of the minimum distance interpretation, we use a second type of example. In fact, the following minimum chi-squared methods based on partitioning continuous data into "cells" have great historical significance for the development of this methodology. We start with a parametric model. Given an underlying parametric distribution M_θ, we could use for basic scores the indicator functions for a partition $A_1, \ldots, A_K$ of the x axis, with $A_k = (a_{k-1}, a_k]$, minus their expectation under the parametric model $\mathcal{M}_\theta$ at parameter value θ:

$$b_k(x, \theta) = I\{x \in A_k\} - P_\theta(A_k). \tag{2}$$

These scores in turn define a semiparametric model which consists of all distributions τ that satisfy, for some value of θ, the mean-zero score equations:

$$E_\tau[b_k(x, \theta)] = P_\tau(A_k) - P_\theta(A_k) = 0 \quad \text{for all } k.$$

That is, τ is in the semiparametric model if and only if it has the same cell probabilities as one of the parametric distributions M_θ, in which case we say that τ has parameter value θ, or that $\tau \in \mathcal{M}_\theta$.

Now suppose the semiparametric model is not true. Given this set of functions and for each θ, we can define the vector of discrepancies $\delta(\theta)$ between the model M_θ and true distribution τ by

$$E_\tau[\mathbf{b}(X, \theta)] = \delta(\theta). \tag{3}$$

If the semiparametric model is correct, then $\delta(\theta) = 0$ for some θ which we call $\theta(\tau)$. If not, then model failure implies that the vector δ is never zero for any θ. However, given a particular measure of distance, we can define $\theta(\tau)$ as a nonparametric function on the space of all distributions by letting it be that θ which makes the distance between $\delta(\theta)$ and 0 the smallest.

We will use a Mahalanobis-type distance. We start by defining the covariance matrix

$$C_\theta = \mathrm{Var}_G(\mathbf{b}(x, \theta)). \tag{4}$$

The subscript G on Var represents the distribution G that is used to evaluate the variance. The most widely desirable choice is $G = \tau$, the true distribution, but there may be practical reasons to prefer a model-based covariance estimator; this will be discussed in Section 3.2.

The *quadratic distance function* between the true distribution and the model distribution as determined through the basic scores is then

$$\rho^2(\tau, \mathcal{M}_\theta) = \delta(\theta)^t [C_\theta]^{-1} \delta(\theta). \tag{5}$$

(If the variance is not invertible, then we would replace it with the Moore–Penrose generalized inverse.) From the quadratic distance function, one can now define the

parameter θ outside the semiparametric model as the nonparametric function

$$\theta(\tau) = \arg\min_{\theta} \rho^2(\tau, \mathcal{M}_\theta). \tag{6}$$

This function gives the value of θ for which the basic scores are closest to mean 0.

2.3 From Distance to Quadratic Inference Function

The next step is to create an empirical quadratic distance function that can be used for inference. If our data $X_1, \ldots, X_n$ are independent and identically distributed under F, then a natural step is to replace $E[\mathbf{b}]$ in (3) with its empirical estimator $\bar{\mathbf{b}}(\theta) = n^{-1}\sum \mathbf{b}(X_i, \theta)$ and to choose a suitable estimator of $\mathrm{Var}_G(\bar{\mathbf{b}})$, say $\hat{C}_\theta$, ending up with a *quadratic inference function* of the form

$$Q^2(\theta) = \bar{\mathbf{b}}'_\theta \hat{C}_\theta^{-1} \bar{\mathbf{b}}_\theta. \tag{7}$$

The choice of $\hat{C}_\theta$ is an important issue to be discussed later. The quadratic inference function then becomes an estimator of the quadratic distance found in (5). Generally, we would construct the quadratic inference function so as to converge to the quadratic distance, and so that the *QIF estimator*, found as

$$\hat{\theta} = \arg\inf_{\theta}\{Q^2(\theta)\},$$

is consistent for the nonparametric θ functional in (6). In particular, if the semiparametric model is correct, then the QIF estimator is consistent for the true value of θ.

Suppose one applies this to the partitioning scores in (2), and one uses the covariance matrix C_θ generated by the model M_θ. In this case, (7) can be shown to be a version of the *Pearson chi-squared distance*:

$$Q^2(\theta) = \sum \frac{[n_i - n P_\theta(A_i)]^2}{n P_\theta(A_i)}. \tag{8}$$

Here n_i is the observed count in the ith cell A_i (we provide the proof in the Appendix). The quadratic inference estimator in this case is better known as the minimum chi-squared estimator (Neyman, 1949). For this reason, there is some justification in describing the methodology we are describing as a generalization of the *minimum chi-squared methods*, a point elaborated in Section 4.

For other scores, we have generalized the chi-squared distance function to allow for other types of statistical questions. For example, we could also use the moment functions, $b_k(x, \theta) = x^k - E_\theta[X^k]$. When we use the same number of moments as parameters, we have the method of moments. If we use more moments than parameters, we are led to a *generalized method of moments*.

One of the great strengths of the quadratic inference method is that the asymptotic theory requires very mild assumptions. Suppose that the semiparametric model is correct and that in (7) we have used a vector of basic scores $\bar{\mathbf{b}}$ and a covariance estimator $\hat{C}_\theta$. Much of the theory that follows holds if there is an asymptotic embedding of the problem that provides a central limit theorem of the form

$$\hat{C}_\theta^{-1/2} \bar{\mathbf{b}}(\theta) \to N(0, I),$$

where I is the identity matrix. Here we have suppressed the dependence of the statistics on some sample-size-like parameter n that goes to ∞. An obvious consequence of this is that $Q^2(\theta)$ in (7) is itself asymptotically chi-squared, with degrees of freedom equal to q, the number of basic scores. In particular, we can construct methods for independent but not identically distributed data, or for stochastic process data, based on the appropriate central limit theory. We will not dwell on these points here, but point to the extensive econometrics literature (White, 1980, 1982; Hansen, 1982; Lee, 1996; Newey and McFadden, 1994; Mátyás, 1999).

3. SOME FUNDAMENTAL ISSUES

Two important questions arise at this point.

- First, how does one choose the covariance estimator $\hat{C}_\theta$?
- Second, how does one select the basic score vector $\mathbf{b}$?

The answers to these questions are tied to the asymptotic properties of the quadratic inference function. We therefore start by summarizing some of the properties of the quadratic inference function found in Qu, Lindsay and Li (2000).

3.1 Overview of Asymptotic Results

The following results all require that the semiparametric model is true and that $\hat{C}_\theta$ is consistent in this model.

1. The quadratic inference function combines the score functions in an optimal way, and so can yield highly efficient procedures. In fact, it can be shown that the point estimators are equivalent to the estimators based on the best linear combinations of the basic scores $\mathbf{b}$, where "best" means "best possible asymptotic variance in the semiparametric model."

2. Suppose we have a particular parametric model in mind, such as the normal, and we include the likelihood scores of that model in our set $\mathbf{b}$. In this case, the QIF estimators will also be fully efficient in the parametric model when it is true. That is, they will be asymptotically equivalent to the maximum likelihood estimator. Thus, for example, if the original scores are the partitioning scores (2), one could add the parametric likelihood scores to the indicators and obtain full efficiency in the minimum chi-squared parameter estimators. If we use the partitioning scores alone, there is a loss of efficiency due to discretization.
3. As an inference function, $Q^2(\theta)$ mimics the properties of the log-likelihood function. In particular, if the semiparametric model is true, with parameter value θ_0, then:

   (a) $Q^2(\theta_0) - Q^2(\hat{\theta})$ is asymptotically chi-squared with degrees of freedom equal to the dimension of θ;

   (b) the profile test statistic $Q^2(\psi_0, \hat{\lambda}_0) - Q^2(\hat{\psi}, \hat{\lambda})$, where (ψ, λ) is a partitioning of the parameter θ into an interest parameter and a nuisance parameter, is asymptotically chi-squared as a test of $H : \psi = \psi_0$, with degrees of freedom equal to the dimension of ψ;

   (c) $Q^2(\hat{\theta})$ is asymptotically chi-squared as a test statistic for testing whether the semiparametric model is true.

These general properties have been known since Hansen (1982), although given in a more specialized case in Ferguson (1958). Noncentrality parameters for local alternatives are given in Newey and West (1987) and Qu, Lindsay and Li (2000). These latter are important for power and sample size calculations.

We note that the goodness-of-fit test given in item 3(c) above has a direct equivalent when we are doing a likelihood analysis in a multinomial model, as there we can conduct a goodness-of-fit test of the parametric model against the unspecified multinomial. However, in a continuous model, the likelihood does not generate such a direct goodness-of-fit procedure, whereas the quadratic inference function does.

The choice of covariance estimation can alter these asymptotic results as follows:

4. If the semiparametric model is correct, but $\hat{C}_\theta \to C^* \neq \mathrm{Var}(\mathbf{b}(\theta)) = C_\theta$, then all semiparametric efficiency properties are lost, and the limiting distributions in item 3 are replaced by linear combinations of chi-squares. This can occur if one uses a parametric model-based estimation of C_θ when the parametric model is false.
5. If the semiparametric model is false, but $\hat{C}_\theta$ is consistent for $\mathrm{Var}(\mathbf{b}(\theta))$, then:

   (a) $\hat{\theta}$ is consistent for $\theta(\tau)$, the minimum distance parameter;

   (b) the test results in items 3(a) and 3(b) still hold for inference on the parameter $\theta(\tau)$;

   (c) the statistic $Q^2(\hat{\theta})$ in item 3(c) converges to $\rho^2(\tau, \mathcal{M}_\theta)$ defined in (5).

3.2 Remarks on Estimating Covariance

The foregoing results indicate that the selection of $\hat{C}_\theta$ can play an important role in the asymptotic properties of the inference function. We next offer some further insights into this problem.

If one has a fully parametric model, then one might wish to use for $\hat{C}_\theta$ the model-based covariance matrix

$$C_\theta = \mathrm{Var}_\theta(\mathbf{b}). \tag{9}$$

Similar in spirit to creating a z-type test for the problem, this has the advantage of adding no additional source of variability. One might conjecture that this would result in improved efficiency of estimation in the model, but this is false as we show below. Moreover, it has the disadvantage of being the incorrect covariance if the parametric model is incorrect, with the consequences we saw in item 4 of Section 3.1.

A second natural approach is to estimate the covariance structure empirically, with some nonparametric estimator $\hat{C}_\theta$ which is globally consistent for the variance of $\mathbf{b}$. If we have the case where $\bar{\mathbf{b}}(\theta) = n^{-1}\sum \mathbf{b}(X_i, \theta)$ and the $\mathbf{b}(X_i, \theta)$ are independent or uncorrelated in (7), it is natural to use

$$\hat{C}_\theta = n^{-2}\sum \mathbf{b}(X_i, \theta)\mathbf{b}'(X_i, \theta). \tag{10}$$

We note that $n\hat{C}_\theta$ in (10) estimates the covariance of $\mathbf{b}$ consistently only if the mean of $\mathbf{b}$ is 0, so that one might consider replacing $\mathbf{b}(X_i, \theta)$ in (10) with $\mathbf{b}(X_i, \theta) - \bar{\mathbf{b}}(\theta)$:

$$\begin{aligned}\tilde{C}_\theta = n^{-2}\sum &\big[\mathbf{b}(X_i, \theta) - \bar{\mathbf{b}}(\theta)\big] \\ &\cdot \big[\mathbf{b}(X_i, \theta) - \bar{\mathbf{b}}(\theta)\big]'.\end{aligned} \tag{11}$$

The use of (11) turns $Q^2(\theta)$ given in (7) into a Hotelling T^2-type statistic. However, the two quadratic inference functions Q_{u}^2 and Q_{c}^2 that are generated by (10) and (11), respectively, have a simple numerical relationship:

$$Q_{\mathrm{u}}^2 = Q_{\mathrm{c}}^2/(1 + n^{-1}Q_{\mathrm{c}}^2). \tag{12}$$

The proof is provided in the Appendix.

As a consequence, there is no difference between the parameter estimators, and for large sample sizes only small differences in the test functions. (We will examine the difference for small sample sizes in a later section.) Hereafter, we will generally restrict our attention to the uncentered covariance function (10), as the resulting quadratic distance has a nice geometric interpretation to be described later, as well as more conservative testing properties.

To further illustrate the role of the choice of C_θ estimators, we return to the partitioned chi-square scores given in (2). If we use the model-based covariance matrix in (9), we have already noted that the quadratic inference function Q^2 equals the partitioned chi-squared distance, with the Pearson (1900) denominators:

$$Q^2_{\mathrm{P}}(\theta) = \sum \frac{[n_i - n P_\theta(A_i)]^2}{n P_\theta(A_i)}.$$

If we use the centered quadratic inference function generated by (11), then the corresponding quadratic inference function is the Neyman chi-squared function (Neyman, 1949)

$$Q^2_{\mathrm{N}}(\theta) = \sum \frac{[n_i - n P_\theta(A_i)]^2}{n_i}.$$

If one wished to compromise between Neyman and Pearson distances for reasons of balancing efficiency and robustness, a simple method is to use

$$\hat{C}_\theta = (1-\alpha) C_{1\theta} + \alpha \hat{C}_{2\theta},$$

where $C_{1\theta}$ is the model-based covariance and $\hat{C}_{2\theta}$ is the empirical variance estimator. This results in the blended chi-squared inference function

$$Q^2_\alpha(\theta) = \sum \frac{(n_i - n P_\theta(A_i))^2}{(1-\alpha) n P_\theta(A_i) + \alpha n_i},$$

and for $\alpha = 0$ and $\alpha = 1$ we recover Q^2_{P} and Q^2_{N}, respectively.

Lindsay (1994) carefully studied the chi-squared distances given above for their efficiency and robustness properties. We can summarize those findings as follows, where here by parametric model we mean the multinomial model for discretized data.

1. All three methods generate fully efficient estimators in the parametric model.
2. One can compare the estimators based on their second-order efficiency in the model. The highest efficiency comes for $\alpha = 2/3$, and Q^2_{N} is more second-order efficient than Q^2_{P}. This is quite surprising given the increased variability of the covariance estimator $\hat{C}_{2\theta}$ compared with $C_{1\theta}$.
3. The robustness of the estimators to outliers increases in the parameter α, with greatest robustness for Q^2_{N}. This is not surprising given the relationship between z and t statistics.

Although these results apply strictly to the partitioning scores, they do point to some important advantages to using empirical covariance matrices. A secondary issue regarding the choice of $\hat{C}_\theta$ is its role in determining the accuracy of the chi-squared approximations in item 3 of Section 3.1. This will be discussed later.

3.3 Selection of Basic Scores Based on Goals

The above asymptotic properties of the quadratic inference function suggest a number of ways to select a set of basic scores for use in the quadratic inference approach. We may point to three important inferential goals, each of which yields a method for selecting a set of scores.

1. To combine a set of *model-defining scores* in the most efficient way. For example, Qu, Lindsay and Li (2000) have shown that the method can be used to improve efficiency in the area of generalized estimating equations (GEE's; Liang and Zeger, 1986). Instead of using "working correlation matrices," they suggest using an extended set of regression scores and combining them optimally using QIF. If the scores are chosen correctly, the estimators are more widely efficient than those found using the same working correlation matrices.
2. To incorporate *goodness-of-fit questions* into parameter estimation via the construction of additional goodness-of-fit scores. For example, the goodness-of-fit test is applicable for testing whether the missingness of data is ignorable by constructing basic scores based on different missing patterns, as Qu and Song (2002) point out. They distinguish between ignorable and nonignorable missingness in estimating equation approaches by whether the mean-zero assumption of the estimating equations holds, since unbiased estimating equations lead to a consistent estimator.

   Chen and Little (1999) proposed a Wald-type test for detecting whether missing data is ignorable, although their Wald test requires the maximum identifiable parameter transformation when dimensions of parameters are different under different missing patterns, as illustrated in Section 2. However, the transformation is not unique, and it has not been investigated as to whether the estimator and the

test statistics are invariant under different transformations. The overall performance of the goodness-of-fit test applying the QIF is better than the Chen–Little Wald-type test.

3. To balance *robustness and efficiency* in point estimation by combining robust and nonrobust-but-efficient scores. For example, Park (2000) showed that if one combines robust and efficient scores, such as the mean and median scores, then one can be simultaneously efficient in both heavy- and light-tailed distributions, and can be consequently robust and efficient simultaneously. The breakdown point in the mean and median case is 25%, for example. In addition, the QIF method provides a goodness-of-fit statistic that assesses the symmetry of the data by testing whether the mean and median are equal.

Given a set of candidate scores, one might ask whether it is wiser to select a subset of them for use in QIF rather than the full set. This is a difficult question to answer, as it depends on the following tradeoff:

- If the true distribution is compatible with the full set of scores, then there can only be a gain in first-order efficiency (asymptotic variance) from using it instead of a subset.
- However, if the reduced subset of scores already generates a fully first-order efficient score, then, even though both reduced and full are ostensibly equivalent, there is a hidden "second-order" extra cost to using the full set that will show up in smaller samples.

To illustrate, consider the mean and median scores b_1 and b_2 and the use of reduced set $\{b_1\}$ versus the use of b_1 and b_2. If the double exponential model is true, then the full set has greater first-order efficiency. However, if the normal model is true, then using the full set has an extra adaptation cost over using b_1.

This additional variability is the price for adaptively estimating from the scores as opposed to simply combining them using prespecified weights. The GMM literature provides some guidance on this point. Imbens (1997) gives an example where including a second moment may lose precision of the estimator in small samples. Harris and Mátyás (1999) also provide examples where there is no improvement of the estimator in asymptotic efficiency when additional moment conditions are highly or perfectly collinear to the existing moment functions, and further, they make the computation more complicated.

To obtain highly informative moment conditions with a reasonable dimension, Gallant and Tauchen (1996) proposed an auxiliary model and generated auxiliary scores to substitute for true scores from the parametric model, and showed that the estimator is nearly efficient if the auxiliary model approximates the true distribution well. Small (2002) developed a criterion similar to the Bayesian information criterion for selection of moment conditions for panel data. Qu and Lindsay (2003) applied the conjugate gradient method to choose optimal linear combinations of moment conditions in quasi-likelihood equation settings.

4. A BRIEF HISTORY

We have seen that the quadratic inference function approach to inference could be fairly called a generalization of the minimum chi-squared method. However, many applications of the method now go under the heading of the generalized method of moments, the name given by Hansen (1982) and now employed extensively in the econometrics literature. In this section, we give a brief review of important literature that is closely related to this methodology. There are two important streams to identify. The first follows the line of Rao and considers the construction of quadratic test statistics. The second stream, largely separated from the first, considers the use of quadratic distances as inference functions.

4.1 Rao's Score Test

Rao (1948) introduced quadratic score test statistics with the form

$$R^2 = \mathbf{s}(\tilde{\theta})' I(\tilde{\theta})^{-1} \mathbf{s}(\tilde{\theta}). \tag{13}$$

Here $\mathbf{s}(\tilde{\theta})$ is the score function, that is, the partial derivative of the log-likelihood function with respect to θ; $\tilde{\theta}$ is the restricted MLE of θ under the null hypothesis; and $I(\tilde{\theta})$ is the Fisher information. It can be shown that *Rao's score test* is asymptotically equivalent to the likelihood ratio test (Neyman and Pearson, 1933) and the Wald (1943) test under both the null and the Pitman alternative hypotheses (Serfling, 1980, page 156), and all follow the chi-squared distribution asymptotically. Indeed, if the null hypothesis is fully specified, the equivalence to the likelihood ratio statistic

$$R^2 = 2\{\log L(\hat{\theta}) - \log L(\tilde{\theta})\} + o_p(1)$$

is suggestive of the fact that Rao's test statistic can also be used as an inference function, where $\hat{\theta}$ is the unrestricted MLE of θ. Rao's score test has some advantages over the Wald and likelihood ratio tests since the Wald test is not invariant to reparameterization and

the likelihood ratio test requires an additional unrestricted MLE of θ. In general, all have limiting distributions that are weighted chi-squared if the model is misspecified (Foutz and Srivastava, 1977; Rotnitzky and Jewell, 1990). However, in Rao's test statistic, as we have noted in Section 3.2, using an empirical information/covariance estimator does repair this problem (Boos, 1992).

There are a multitude of names for Rao's score test with some variations. In fact, the Pearson chi-squared test (1900) can be derived using Rao's score test for multinomial distributions (Rao, 1973, page 442; Cox and Hinkley, 1974, page 316). In the econometrics literature, a parallel version of Rao's score test is called the *Lagrangian multiplier test* (Aitchison and Silvey, 1958; Silvey, 1959) since the restricted likelihood equations can be solved using Lagrangian multipliers.

4.2 Generalized Score Tests

In recent years, there has been considerable interest in the reduction of assumptions by using a general estimating function approach to statistical inference. Although much of the interest in the areas of generalized linear models and generalized estimating functions has been focused on Wald-type tests and quasi-likelihood procedures, there have been some developments of the equivalent score tests. Boos (1992) provided a general discussion of the extension of Rao's score test for general estimating equations, calling them *generalized score tests*. The main ideas are as follows:

- The likelihood scores $\mathbf{s}(\theta)$ in Rao's score test (13) can be replaced by any estimating function arising from likelihood, quasi-likelihood, least squares and robust M-estimation.
- The parameter θ is estimated by solving $\mathbf{s}(\theta) = 0$, and the Fisher information $I(\theta)$ in (13) is replaced by the asymptotic covariance matrix of $\mathbf{s}(\theta)$.

Notice that Boos' approach is quite similar to the use of the quadratic inference function. The main difference between these two approaches is that in the quadratic inference function approach the dimension of $\mathbf{s}$ could be greater than that of the parameter and $\hat{\theta}$ is obtained by minimizing the quadratic inference function.

Another distinction arises in the treatment of tests for interest parameters, as in Section 3.1, item 3(b). Boos' work implicitly treats a subset of the scores as being identified with the nuisance parameters, as in the scores generated by differentiating a likelihood. In the present work, we need not make such special identification.

4.3 Generalized Least Squares

Turning to methods that use quadratic forms to create inference functions, one of the most important ancestors is generalized least squares (Sprent, 1966). In generalized least squares, we minimize a least squares criterion such as

$$\sum_{i=1}^{K} (y_i - \mu_i(\theta))' \Sigma_i^{-1} (y_i - \mu_i(\theta))$$

over θ, where y_i is the response for the ith outcome, $\mu_i = E(y_i)$ and Σ_i is the covariance matrix of y_i. This has the form of a quadratic inference function for the basic scores $b_i = (y_i - \mu_i(\theta))$, where the scores are assumed to be independent across i. If Σ_i does not depend on θ, then differentiation of the least squares criterion leads one to the quasi-score function and consistent estimation.

However, if the covariances Σ_i depend on the parameters, minimization can lead to inconsistent estimators as the number of responses increases (Singh and Mantel, 1998). This problem has been avoided in the quadratic inference function approach by making sure that the number of scores stays fixed. In particular, in the QIF approach we take the sums over the independent scores *before* we create the quadratic form, which reduces the asymptotic bias generated by θ in the covariance function.

Multiple-root problems often arise in practice for estimating equation approaches. Heyde and Morton (1998) apply generalized least squares as an objective function to choose the correct root of a general estimating equation in cases of multiple roots. However, Singh and Mantel (1998) argue that using the form similar to the generalized score test is better than the generalized least squares approach for selecting a consistent root.

The generalized least squares method has also been applied in GEE settings (Liang and Zeger, 1986). Chaganty (1997) and Shults and Chaganty (1998) combine the generalized least squares method with the GEE for serially correlated data. Their method allows for a wide range of working correlation structures. They refer to it as the quasi-least squares method.

4.4 Ferguson and Minimum Chi-squared

As noted earlier, another foundation for quadratic inference functions is found in the minimum chi-squared work of Neyman (1949). Ferguson (1958) expanded upon the minimum chi-squared method. The motivation for his development of a generalized minimum chi-squared method was to find a computationally

simpler method to estimate interest parameters in parametric models without losing the usual asymptotic properties of best asymptotically normal estimates (Barankin and Gurland, 1951). Ferguson's approach is to estimate the parameter by minimizing the quadratic form

$$n\big(Z_n - P(\theta)\big)'\Sigma^{-1}(\theta)\big(Z_n - P(\theta)\big),$$

where $Z_n = n^{-1}\sum_{i=1}^n X_i$, the X_i's are independent identically distributed q-dimensional random vectors, $P(\theta) = E_\theta[X]$ and $\Sigma(\theta)$ is the variance of X. As in our development, the number of "scores" q is possibly greater than or equal to the dimension of θ. However, the focus was entirely on parametric models and simplified estimation.

Ferguson obtained the general asymptotic results described earlier. Other than the fact that the scores were formed strictly as contrasts between random variables and their expectations, rather than being arbitrary, the asymptotic description was nearly completely developed at this point. However, these ideas entered a quiet phase until they were recreated in the econometrics literature under a different name.

4.5 Hansen and Generalized Method of Moments

Hansen (1982) introduced the generalized method of moments (GMM). This has become popular in the econometrics field, where conditional heteroscedasticity (White, 1980) and serially correlated data arise often, and therefore it is difficult to formulate the full likelihood function. Thus, there was greater interest in semiparametric models with reduced assumptions. For that reason, the GMM does not require the complete specification of the model, but requires the specification of zero-mean *moment conditions* which the model satisfies, that is, $E(g(X,\theta)) = 0$. In regression settings, a commonly used moment condition is $E[X'(y - X\theta)] = 0$, which is the one associated with ordinary least squares.

The dimension of the moment conditions (q) is usually greater than the dimension of the parameters (p) in econometrics. We might contrast this with the statistical mainstream, where g are often called *estimating functions* or *scores* and the dimension of g is set equal to p. The corresponding estimation procedures are also known as *M-estimation* in the robustness literature.

Since there are more equations than unknowns, the parameter vector is said to be *overidentified*. A GMM estimator of the parameter of interest is obtained by setting linear combinations of r moment conditions as close to 0 as possible, that is,

$$\hat\theta_{\mathrm{GMM}} = \arg\min_\theta g_N'\Sigma^{-1}g_N,$$

where the optimal weighting matrix Σ, based on minimizing the asymptotic variance of the estimator, is found to equal the covariance of the moment conditions.

Since Σ is often unknown for finite samples, the traditional two-step approach for the GMM estimator is to apply an initial consistent, but inefficient estimator $\hat\theta$ to obtain $\hat C$, the estimator of Σ, then update $\hat\theta$ by minimizing $g_N'\hat C^{-1}g_N$ in the second step. However, the two-step GMM estimator is not invariant to linear transformation of the moment conditions and may be severely biased in small samples. To improve GMM estimators in small samples, Hansen, Heaton and Yaron (1996) proposed the continuous updating estimator, Imbens (1997) proposed the one-step method and Smith (1997) proposed semiparametric quasi-likelihood approximations to the likelihood function. These alternative approaches are all invariant to linear transformation of the moments, and their estimators are asymptotically efficient without relying on the initial choice of the weighting matrix.

Hansen's GMM and its development have largely been in the econometrics literature. We point next to one area of cross-fertilization between mainstream statistics and econometrics.

4.6 Empirical Chi-Squared

The quadratic inference methodology is also closely linked to another method for combining estimating functions by the construction of an inference function, and through that linkage it might rightfully be called the "empirical chi-squared method."

The empirical likelihood method of Owen (1988), as applied to the problem of combining estimating functions, can be described as follows. First, suppose we have a sample $X_1,\ldots,X_n$, assumed to be i.i.d. for simplicity. We will treat the observed data points $x_1,\ldots,x_n$ as being fixed, much as one does in the bootstrap. However, instead of sampling from the observed data, we will build a model by allowing arbitrary discrete distributions F on this support set. Thus, we write $p_i = \mathrm{pr}_F[X_i = x_i]$, so that the distributions can be represented by a vector $\mathbf{p}$.

In this class of discrete distributions, we can generate a mirror of the semiparametric model by letting $\mathcal{M}_\theta = \{\mathbf{p} : \sum p_i\mathbf{b}(x_i,\theta) = \mathbf{0};\ \sum p_i = 1\}$ and setting $\mathcal{M} = \bigcup \mathcal{M}_\theta$. We also create a multinomial-type likelihood by letting $\mathcal{L}(\mathbf{p}) = \prod p_i$. This corresponds to having one observation from each of the observed x_i. The *empirical likelihood* of the parameter θ is then

$$L(\theta) = \sup\{\mathcal{L}(\mathbf{p}) : \mathbf{p} \in \mathcal{M}_\theta\}.$$

To find the empirical likelihood, we need to optimize $\mathcal{L}(\mathbf{p})$ over $\mathbf{p} \in \mathcal{M}_\theta$ for each fixed θ. Note that the set $\mathcal{M}_\theta$ is defined by $q+1$ linear constraints, so that the optimization can be carried out for each θ using the method of Lagrange multipliers. Also, corresponding to each θ, there is a $\hat{\mathbf{p}}_\theta$, corresponding to a discrete distribution $\hat{F}_\theta$, that is the maximizing argument, and so $\hat{F}_\theta$ is an estimator of the true distribution in the semiparametric model.

The next step is to maximize the empirical likelihood $L(\theta)$ over θ, yielding the maximum empirical likelihood estimator $\hat{\theta}$. If the dimensions of the score vector and the parameter vector are the same, then this estimator is just the solution to the equations $\sum \mathbf{b}(x_i, \theta) = \mathbf{0}$ because $\hat{p}_i = 1/n$ maximizes $\mathcal{L}(\mathbf{p})$ over all $\mathbf{p}$ in this case. If there are more scores than parameters, then we can think of the empirical likelihood as finding the discrete semiparametric distribution best fitting the observed empirical distribution.

Owen (1988) showed that the empirical likelihood behaves much like a true likelihood, for example, that $2[\log L(\hat{\theta}) - \log L(\theta_0)]$ is an asymptotic chi-squared test for $\theta = \theta_0$. Qin and Lawless (1994) extended this to the semiparametric model. They linked estimating equations and the empirical likelihood and proposed to optimally combine information when there are more estimating equations than unknown parameters. They applied Owen's (1988) empirical likelihood as an objective function, rather than the quadratic inference function we propose here.

The QIF approach is more direct than Qin and Lawless' approach since they have to estimate the optimal weights p_i through Lagrange multipliers, whereas in the QIF approach minimizing the QIF automatically provides the optimal weights, as we show next.

The parallel between the empirical likelihood and the quadratic inference function can be developed as follows. Rather than use a likelihood measure of discrepancy between the empirical weights $1/n$ and the semiparametric weights p_i, we form a chi-squared-type distance:

$$\chi^2(\mathbf{p}) = \sum \left(p_i - \frac{1}{n} \right)^2.$$

This would correspond to using the Neyman chi-squared distance where we treat the observed data as $1/n$ at each fixed x_i. For each fixed θ, we can then generate a measure of distance from the observed data to the best fitting semiparametric model via

$$R^2(\theta) = \inf\{\chi^2(\mathbf{p}) : \mathbf{p} \in \mathcal{M}_\theta\}.$$

It is a standard exercise in Lagrange multipliers to show that $R^2(\theta) = Q_u^2(\theta)$, that is, the quadratic inference function with the uncentered estimator of covariance.

5. PROJECTION REPRESENTATIONS OF QIF

In this section, we display two projection representations for Q^2 that are an aid to a deeper understanding of its basic properties. The results in this section were derived in Park (2000), where proofs may be found. In this section, we suppose that the basic scores are of the form $\bar{\mathbf{b}}(\theta) = n^{-1} \sum \mathbf{b}(X_i, \theta)$.

We start by creating an $n \times q$ matrix B with (i, j) element $b_j(x_i; \theta)$. That is, the jth column of B is an n vector giving the values of the jth score function at the n values of x_i. We can form a projection matrix that will project an arbitrary vector in $\mathbb{R}^n$ onto the column space of B by $P_B = B(B'B)^{-1}B'$.

The geometric interpretation of $Q^2(\theta)$, with the uncentered covariance estimator, is that it is the squared length of the projection of the $\mathbf{1}$ vector onto the column space of B. That is, it is easily verified that

$$Q_{\mathrm{u}}^2(\theta) = \mathbf{1}' P_B \mathbf{1} = \|\mathbf{P}_B \mathbf{1}\|^2.$$

We can interpret this as follows. Each column of B is orthogonal to $\mathbf{1}$ if and only if the corresponding sample score $n^{-1} \sum b_j(X_i, \theta)$ equals 0. Since we cannot make all the scores simultaneously equal to 0, we instead make their departure from orthogonality to $\mathbf{1}$ as small as possible by finding the overall length of the projection onto this space and minimizing it.

The projection representation above can be used to show a number of nonobvious properties of $Q_{\mathrm{u}}^2(\theta)$. For example, Q_{u}^2 must always have a value less than $\|\mathbf{1}\|^2 = n$. In addition, if we add a new score function to our basic set, we increase the column space and so necessarily increase Q_{u}^2, the length of the projection of $\mathbf{1}$ onto the column space.

The second projection representation of Q_{u}^2 is useful in asymptotics. We start with the basic assumption that by construction

$$\mathbf{Z} \stackrel{\mathrm{def}}{=} \hat{C}_\theta^{-1/2} \bar{\mathbf{b}}(\theta) \to N(0, I).$$

From this, we clearly have the relationship

$$Q_{\mathrm{u}}^2(\theta) \to \chi_q^2,$$

where q is the number of scores being used.

We gain an ANOVA-style decomposition of variance from the following asymptotic approximation (under the null hypothesis):

$$Q_{\mathrm{u}}^2(\hat{\theta}) = \|(I - \mathbf{P}_W)\mathbf{Z}\|^2 + o_p(1), \tag{14}$$

where the projection matrix $\mathbf{P}_W$ in (14) is

$$P_W = C^{-1/2}D(DC^{-1}D')^{-1}D'C^{-1/2},$$

with $C = C_\theta$ and $D = D_\theta = E[\nabla \mathbf{b}(X, \theta)]$. Note that P_W is the projection matrix yielding projection onto the columns of $W = C^{-1/2}D$, where W has p columns. Hence, if W is full rank, then $\mathrm{tr}(P_W) = p$, and from (14) the goodness-of-fit test statistic satisfies

$$Q_{\mathrm{u}}^2(\hat{\theta}) \to \chi^2_{q-p}.$$

From the projection representation, it is also clear that

$$\begin{aligned} Q_{\mathrm{u}}^2(\theta) - Q_{\mathrm{u}}^2(\hat{\theta}) &\approx \|\mathbf{Z}\|^2 - \|(I - \mathbf{P}_W)\mathbf{Z}\|^2 \\ &= \|\mathbf{P}_W\mathbf{Z}\|^2 \to \chi^2_p. \end{aligned}$$

Finally, one can show that the test statistic for a parameter of interest ψ in the composite hypothesis $H_0 : \psi = \psi_0$, namely $Q_{\mathrm{u}}^2(\psi_0, \tilde{\lambda}) - Q_{\mathrm{u}}^2(\hat{\psi}, \hat{\lambda})$, is chi-squared with degrees of freedom equal to the dimension of ψ, by showing that the relevant projection subspaces for the full model, say W, and the null hypothesis model, say V, are nested. Then the test statistic corresponds asymptotically to the length of a projection onto the part of W that is orthogonal to V.

6. SECOND-ORDER IMPROVEMENTS

We have noted that Owen (1988, 1990) introduced the concept of empirical likelihood when the parametric likelihood is unknown, and provided results of the asymptotic theory analogous to the parametric likelihood. In particular, the empirical likelihood ratio test can be approximated by a chi-squared distribution, a nonparametric version of Wilks' theorem, with error of order n^{-1}, and can be improved by Bartlett adjustment to an error of order n^{-2}.

Corcoran (1998) extended these results to show that quadratic statistics such as QIF are not Bartlett correctable. Despite the seeming potential distributional superiority of the empirical likelihood method, simulation studies (Corcoran and Davison, 1995) in small and moderate samples show that the empirical likelihood ratio test (adjusted or unadjusted) performs poorly in the right tail of the distribution. In particular, it is anticonservative (i.e., the actual test size is greater than the nominal size α) and often not very Bartlett correctable. In fact, the behavior of the empirical likelihood test statistic is similar to that found in simulations for the QIF method in Qu and Lindsay (1999) and Davidson and MacKinnon (1983, 1984) where Rao's score tests were calculated using sample variance instead of information.

We next present an Edgeworth expansion that shows the primary source of asymptotic difficulty for the QIF method, then indicate how bootstrapping can help. That is, a bootstrap resampling strategy (Efron, 1987; Hall and Horowitz, 1996; Hu and Kalbfleisch, 2000) can be a simple and effective way to achieve higher-order accuracy for test size, and it works effectively for relatively small sample sizes in our case.

6.1 Edgeworth Expansion for QIF

The following Edgeworth expansion (see Hall, 1992, page 39) of the quadratic inference function indicates the asymptotic source of inaccuracy for the method.

Suppose that $n^{1/2}(t_n - \mu)$ is asymptotically normally distributed with mean 0 and variance σ^2. If t_n is a smooth function of sample means, then the distribution of $n^{1/2}(t_n - \mu)$ can be approximated as an *Edgeworth expansion* using a series of the form

$$\begin{aligned} &P\{n^{1/2}(t_n - \mu)/\sigma \le x\} \\ &\quad = \Phi(x) + n^{-1/2}p_1(x)\phi(x) + \cdots \\ &\qquad + n^{-j/2}p_j(x)\phi(x) + \cdots, \end{aligned} \tag{15}$$

where $\Phi(x)$ and $\phi(x)$ are the standard normal cumulative distribution and probability density functions. The functions p_j are polynomials with coefficients depending on cumulants of the sample means, and are even or odd functions according to whether j is odd or even, respectively.

When the empirical covariance estimators are used to estimate C_θ, the quadratic function has the same structure as a Hotelling T^2 statistic. The quadratic inference function Q^2 can therefore be represented as a smooth function of the first and second moments of the basic scores. To simplify the analysis, we will illustrate the Edgeworth expansion of Q^2 for the simplest case when the dimension of the score statistic is $q = 1$. In this case, Q_{u}^2 is the square of an "uncentered" t-statistic, that is,

$$t_u = \frac{\sum b_i}{\sqrt{\sum b_i^2}} = \frac{\sqrt{n}\bar{b}}{\sqrt{\sum b_i^2/n}}.$$

If we were to center the covariance estimator, we would have the usual centered t as

$$t_c = \frac{\sqrt{n}\bar{b}}{\sqrt{\sum (b_i - \bar{b})^2/n}}.$$

Now we are squaring these statistics, so we are interested only in the error in the sum of the two tail probabilities. For these probabilities, if we apply the

formal Edgeworth expansion of t of the form (15), we get a simplification due to the even/odd nature of the polynomials involved:

$$G(x) = P[T^2 \le x^2] = P[-x \le T \le x]$$
$$= \Phi(x) - \Phi(-x) + 2n^{-1}p_2(x)\phi(x) + O(n^{-2}), \tag{16}$$

since $p_1(-x) = p_1(x)$, $p_2(-x) = -p_2(x)$ and $\phi(x) = \phi(-x)$. Note that the use of symmetric intervals for T, as implied by using T^2, means that the errors of orders $n^{-1/2}$ and $n^{-3/2}$ are 0, so we can focus on the order n^{-1} term to make second-order comparisons.

Hall (1992, page 73) proved that, for the centered t, the second Edgeworth polynomial is

$$p_2^{\mathrm{c}}(x) = x\{\tfrac{1}{12}\kappa(x^2-3) - \tfrac{1}{18}\gamma^2(x^2-1)(x^2+3) - \tfrac{1}{4}(x^2+3)\},$$

where γ and κ are standardized skewness and kurtosis, respectively. The formula for the uncentered t is given by Qu and Lindsay (1999) as

$$p_2^{\mathrm{u}}(x) = x\{\tfrac{1}{12}\kappa(x^2-3) - \tfrac{1}{18}\gamma^2(x^2-1)(x^2+3) + \tfrac{1}{4}(x^2-3)\}.$$

The relationship between the centered and uncentered versions of p_2 is very simple:

$$p_2^{\mathrm{u}}(x) = p_2^{\mathrm{c}}(x) + \tfrac{1}{2}x^3. \tag{17}$$

We can interpret this formula as follows. First, notice that a negative value of p_2 from (16) at any x is undesirable because it makes the normal approximation anticonservative at that x. Moreover, it is clear from (17) that, for positive x, the uncentered p_2 is larger than the centered p_2, so that using the uncentered t_{u} is always more conservative than t_{c}. If the skewness γ is 0, then both polynomials simplify greatly, and we find that the tail approximation based on the uncentered t_{u} is quite good. In particular, at $x = \sqrt{3}$, corresponding to roughly the 90th percentile, the order n^{-1} error is 0, and there is only order n^{-2} error. Our simulation results bear out that, for symmetric data, the chi-squared approximations work very well for Q^2 with the uncentered estimator of covariance.

Unfortunately, when skewness is not 0, γ plays an important role for large values of $|x|$. In both centered and uncentered cases, the polynomial p_2 becomes dominated by its largest order term, namely $-\gamma^2 x^5$, which, being negative, always points to nonconservative behavior. For this reason, one must be very cautious about using asymptotic approximations in skewed data. We can, however, offer the following suggestion.

6.2 Bootstrap Sampling for QIF Tests

Bootstrap simulation is a simple and effective way to estimate the distribution of a pivotal quantity whose limiting distribution does not depend on unknown quantities, and it provides more accurate critical values for test statistics than asymptotic results provide (Singh, 1981; Beran, 1988; Hall, 1986, 1992; Hall and Horowitz, 1996). In particular, Hall and Horowitz (1996) proposed bootstrap critical values for tests based on the GMM. Notice that in our case t^2 is an asymptotic pivotal quantity.

We can motivate the improvement in accuracy for the bootstrap as follows. The Edgeworth expansion of the bootstrap distribution, as in (16), is

$$\hat{G}(x) = P(|T^*| \le x|\chi) = 2\Phi(x) - 1 + 2n^{-1}\hat{p}_2(x)\phi(x) + O(n^{-2}),$$

where T^*, the bootstrap version of T, is obtained from a resample χ^* instead of the true sample χ, and $\hat{p}_2$ is the bootstrap distribution version of p_2. The difference between $\hat{p}_2$ and p_2 is of order $n^{-1/2}$ in probability, which leads to $\hat{G}(x) - G(x) = O_p(n^{-3/2})$. It is clear that using the bootstrap approximation of G is sharper than using a normal approximation, since it is in error by $n^{-3/2}$ instead of n^{-1}.

In the models we have considered here, one could perform three kinds of bootstrapping: parametric, semiparametric or nonparametric. Parametric and nonparametric bootstrapping would follow the well-established recipes given in Efron and Tibshirani (1993) as well as in other texts.

However, devising a bootstrapping recipe that uses the semiparametric model is more challenging because the distribution one samples from should be an estimated element of the semiparametric model. The problem arises because the original data do not satisfy the equation $n^{-1}\sum b_j(x_i, \hat{\theta}) = 0$. It follows that when one simulates from the nonparametric bootstrap distribution, with mass $1/n$ at each x_i, the scores do not have mean 0, and so this sampling distribution is not in the semiparametric model.

Hall and Horowitz (1996) solved this problem by recentering. That is, for each resampled x_i^*, we create the recentered moment condition $\mathbf{b}^*$ as follows:

$$\mathbf{b}^*(x_i^*, \theta) = \mathbf{b}(x_i^*, \theta) - n^{-1}\sum_j \mathbf{b}(x_j, \hat{\theta}), \tag{18}$$

where the x_j's are from the original sample and $\hat{\theta}$ is the QIF estimator of the original sample. The covariance C_θ^* of $\mathbf{b}^*$ can be calculated as $n^{-2}\sum(\mathbf{b}^*{}_i)(\mathbf{b}^*{}_i)'$.

Hall and Horowitz (1996) showed that this recentering procedure gives asymptotically valid bootstrap critical values.

We note that the bootstrapping distribution should be centered in some way if we wish to use the corresponding simulated critical value for testing goodness of fit with $Q^2(\hat{\theta})$. If one uses an uncentered bootstrap distribution, then the null hypothesis, which specifies the mean-zero property of all the scores, does not hold perfectly in the sample score distribution, and so one is sampling from an alternative hypothesis in which the means are not identically 0. This will inflate the critical value a small amount if the null is actually true (because the estimated means will be near 0), but a lot if the null is false. Thus, the size might be nearly correct, but the power will be quite poor.

On the other hand, the test $Q^2(\theta) - Q^2(\hat{\theta})$ is a valid test of a nonparametric hypothesis regarding the minimum distance parameter as well as a semiparametric hypothesis. Using semiparametric bootstrapping should provide greater accuracy when this model is true, but nonparametric bootstrap testing should be valid as well.

We will illustrate how bootstrap resampling techniques correct for the effect of skewness in a QIF test involving correlated Poisson–Gamma data which is inherently highly skewed. The setting is a GEE-type model, where vector responses y_i correspond to measurements taken on the ith cluster. In our simulation, the response variable within the ith cluster y_i, with cluster size 10, was generated from Poisson$(\lambda_i e^{X_i\theta})$, where the cluster-specific latent variable λ_i is generated from Gamma$(1, 1)$. The covariate vector X_i was equal to $(0.1, 0.2, \ldots, 1.0)$ in each cluster, with the number of clusters being $N = 20, 50$ or 100. For the simulation, the parameter θ was set to 1. Note that although this model is "subject specific" due to the presence of the latent variable λ, it still satisfies the marginal model

$$E[y_i|X_i] = e^{X_i\theta} \triangleq \mu_i, \tag{19}$$

and we can apply the generalized estimating equation method to the marginal means (19).

Qu, Lindsay and Li (2000) showed that one could create an efficient quadratic inference function method for models such as this by creating basic scores from the "working independence" model scores and adding additional scores designed to increase efficiency under other correlation structures. In the case of our simulation, we assumed the Poisson link with exchangeable correlation structure and used the basic scores

$$\mathbf{b}(\theta) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}(\dot{\mu}_i)' A_i^{-1}(y_i - \mu_i) \\ \sum_{i=1}^{N}(\dot{\mu}_i)' A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \end{pmatrix}, \tag{20}$$

where $A_i = \text{diag}(\mu_{ij})$ is the diagonal marginal variance of the Poisson model and M_1 is a 10×10 matrix with 0 on the diagonal and 1 elsewhere. Here choosing I and M_1 preserves full efficiency under the exchangeable correlation structure (see Example 1 of Qu, Lindsay and Li, 2000). The QIF is calculated using (7) with the uncentered covariance estimator.

For this example, we first simulated a single data set from the Poisson–Gamma distribution described above and then took $B = 5000$ bootstrap random samples of the N clusters (using $N = 20$, 50 and 100). We applied the block bootstrap strategy here (Künsch, 1989; Lahiri, 1996). That is, we sample N clusters randomly with replacement to form a bootstrap sample. For comparison, we bootstrapped the sample moment conditions with recentering as in (18) and without recentering. The QIF test statistics were calculated from each bootstrap resample data set.

We also simulated 5000 data sets from the true distribution. In Figure 1, for small sample size with $N = 20$, we plot quantiles of the QIF from the true distribution and the bootstrap samples after recentering the moment conditions against the quantiles of the asymptotic chi-squared distributions. The Q–Q plots of $Q^2(\theta)$, $Q^2(\hat{\theta})$ and $Q^2(\theta) - Q^2(\hat{\theta})$ shown in Figure 1 indicate that the true distribution is closer to the single bootstrap estimate of the distribution than to the chi-squared approximations in the tails, though tests based on bootstrap and asymptotic critical values were still liberal.

We also compared the level of tests using the asymptotic and bootstrap (with and without recentering of moment conditions) critical values for different cluster sizes. Table 1 lists nominal test size $\alpha = 0.05$ and cluster size $N = 20, 50$ and 100. Notice that, for $N = 20, 50$ and 100, the test levels based on the asymptotic chi-squared distribution are liberal and perform worst for the small sample size $N = 20$. Test levels based on bootstrap (*with* recentering moment conditions) critical values are less liberal than the asymptotic tests. Test levels based on bootstrap (*without* recentering moment conditions) critical values seem to

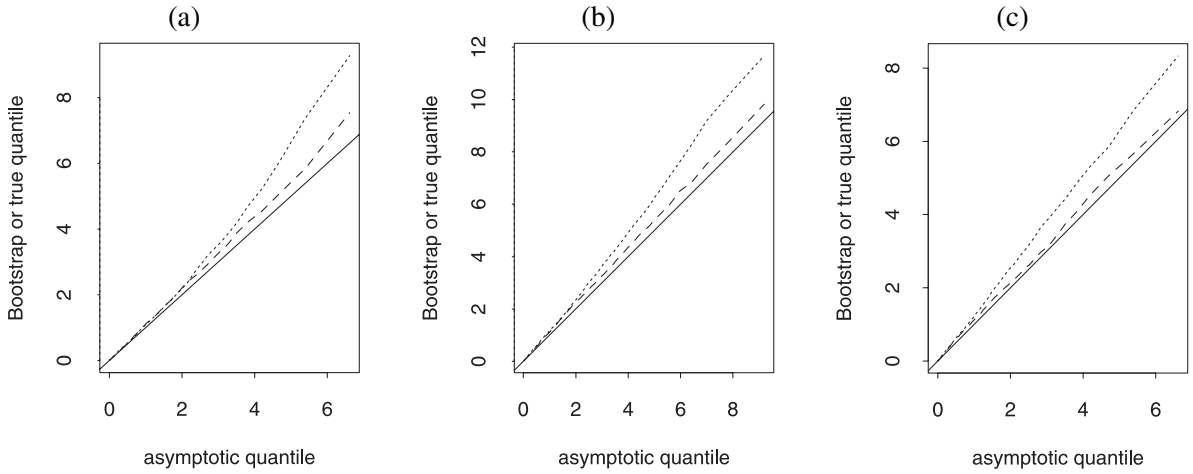

FIG. 1. *Quantile–quantile plot of QIF test for correlated Poisson–Gamma data with* $N = 20$, *where the solid line indicates the reference line*, *the dotted line is the distribution of the QIF test based on* 5000 *simulated data and the dashed line is the recentering bootstrap distribution of the QIF test*. (a) *Quantile–quantile plot of* $Q^2(\hat{\theta})$ *vs.* χ^2_1. (b) *Quantile–quantile plot of* $Q^2(\theta)$ *vs.* χ^2_2. (c) *Quantile–quantile plot of* $Q^2(\theta) - Q^2(\hat{\theta})$ *vs.* χ^2_1.

be very conservative for $N = 20$. As the sample size increases, these three test sizes are fairly similar and close to the nominal test level.

However, the difference between the bootstrap with recentering and without recentering becomes obvious when the mean-zero model assumption does not hold. We create the second score in (20) by using $y_i - \mu_i - 0.1$ instead of $y_i - \mu_i$, and keep the first one as before. Table 2 indicates that the power of rejecting the mean-zero model assumption is much lower for the bootstrap without recentering moment conditions than for the bootstrap with recentering. The power based on asymptotic chi-squared is slightly higher than the power using bootstrap with recentering, because the test size based on the asymptotic result is inflated.

Alternatives to using bootstrapping or empirical likelihood methods to improve small sample properties include the exponential tilting estimator of Kitamura and Stutzer (1997) and Imbens, Spady and Johnson (1998). These authors applied Cressie and Read's (1984) power divergence statistics to this problem and therefrom introduced the exponential tilting estimator, which is based on minimizing the Kullback–Leibler information criterion. Smith (1997) and Newey and Smith (2002) showed that the empirical likelihood and exponential tilting estimators are all in a class of

TABLE 1
Simulated size of nominal 0.05 *level tests based on the quadratic inference function with asymptotic and bootstrap critical values* (with *recentering and* without *recentering moment conditions*); *the simulation standard error is approximately* $\sqrt{0.05 \cdot 0.95/5000} = 0.003$

N		$Q^2(\hat{\theta})$	$Q^2(\theta) - Q^2(\hat{\theta})$
20	Asymptotic	0.072	0.084
20	Bootstrap (with)	0.060	0.073
20	Bootstrap (without)	0.015*	0.049
50	Asymptotic	0.065	0.069
50	Bootstrap (with)	0.058	0.041
50	Bootstrap (without)	0.057*	0.038
100	Asymptotic	0.054	0.061
100	Bootstrap (with)	0.039	0.050
100	Bootstrap (without)	0.036*	0.066

*Not recommended; see text.

TABLE 2
Simulated test power based on the quadratic inference function with asymptotic and bootstrap critical values (with *recentering and* without *recentering moment conditions*) *when the mean-zero model assumption does not hold*; *the simulation standard error is approximately* $\sqrt{0.05 \cdot 0.95/5000} = 0.003$

N		$Q^2(\hat{\theta})$
50	Asymptotic	0.370
50	Bootstrap (with)	0.328
50	Bootstrap (without)	0.065*
100	Asymptotic	0.533
100	Bootstrap (with)	0.503
100	Bootstrap (without)	0.279*

*Not recommended; see text.

generalized empirical likelihood estimators and have the same asymptotic distribution as GMM, but the asymptotic properties of higher orders are different among these estimators.

Another alternative approach for more accurate inference is the saddlepoint approximation for M-estimators (Field, 1982; Jing and Robinson, 1994; Robinson, Ronchetti and Young, 2003; Ronchetti and Trojani, 2002). The relationship between empirical likelihood and empirical saddlepoint approximation was investigated by Monti and Ronchetti (1993).

7. CONCLUSION

In this paper, we hope to have broadened the horizon of potential applications of the quadratic inference function. The QIF is built on a semiparametric framework defined by a set of mean-zero estimating functions, but is also applicable to parametric or nonparametric settings. The QIF has advantages of the estimating function approach such as not requiring the specification of the likelihood function, but also overcomes limitations of the estimating function approach such as a lack of objective functions for selecting a correct root in multiple-root problems (Small, Wang and Yang, 2000) and a lack of likelihood-type functions for testing.

The origin of the QIF can be traced back to Pearson's (1900) χ^2 test and Rao's score test (1948), but it is more closely related to Ferguson's (1958) minimum χ^2 method and Hansen's (1982) generalized method of moments, popular in econometrics. We compared the QIF to similar existing approaches such as generalized score tests (Rotnitzky and Jewell, 1990; Boos, 1992) and generalized weighted least squares, and tied special cases of the QIF to Pearson's (1900) and Neyman's (1949) χ^2 tests.

We have paid particular attention as to how to select basic scores, since this can provide robust and efficient estimation of regression parameters by combining robust and efficient scores optimally (Park, 2000). Selecting scores from different missing patterns also provides a simple tool to test for ignorable missingness in estimating equation approaches (Qu and Song, 2002). Notice that for these applications of the QIF it is not necessary to satisfy the mean-zero assumption of the moment conditions, as the QIF approach is also valid under a nonparametric interpretation. The goodness-of-fit test of the QIF plays an important role in testing the mean-zero assumption. This test is rather simple to use compared to other goodness-of-fit tests in the GEE literature (Barnhart and Williamson, 1998; Pan, 2002).

Finally, the QIF is also related to empirical likelihood which is popular for nonparametric models. We illustrated the Edgeworth expansion of the QIF and showed how a bootstrap strategy could improve testing accuracy for small samples. Overall, the improvements using bootstrapping were modest. It is worth pointing out, however, that an additional advantage to carrying out a bootstrap analysis is that the existence of large differences between the bootstrap and asymptotic p-values can be used as a diagnostic for the failure of the large-sample theory.

APPENDIX

PROOF OF (8). Let $\bar{\mathbf{b}} = 1/n \sum_{i=1}^{n} \mathbf{b}_i$, where $\mathbf{b}_i = (b_{i1}, \ldots, b_{iK})'$ and $b_{ij} = I\{x_i \in A_j\} - P_\theta(A_j)$ for $j = 1, \ldots, K$. Therefore,

$$\bar{b}_j = \frac{1}{n}\sum_{i=1}^{n}\big(I\{x_i \in A_j\} - P_\theta(A_j)\big) = \frac{n_j}{n} - P_\theta(A_j) = \hat{p}_j - p_j.$$

Let $\mathbf{p} = (p_1, \ldots, p_K)'$ and $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_K)'$. Since $\mathbf{b}_i$ follows a multinomial distribution,

$$\mathrm{Var}(\mathbf{b}_i) = \mathrm{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}' \stackrel{\triangle}{=} \Sigma_\theta,$$

where $\mathrm{Diag}(\mathbf{p})$ is a diagonal matrix with p_j as a diagonal component; we denote Diag as D hereafter. This matrix is not full rank, so we consider the following argument to find a generalized inverse. Notice that

$$\{D(\mathbf{p})\}^{-1/2}\Sigma_\theta\{D(\mathbf{p})\}^{-1/2} = I - \{D(\mathbf{p})\}^{-1/2}\mathbf{p}\mathbf{p}'\{D(\mathbf{p})\}^{-1/2} = A$$

is an idempotent matrix. The eigenvalues of an idempotent matrix are either 0 or 1; the generalized inverse $A^- = \sum_{j=1}^{K} \lambda_j^{-1}\mathbf{e}_j\mathbf{e}_j'$, where the λ_i's are all 1's. Therefore, $A^- = A$. This implies

$$\Sigma_\theta^- = \{D(\mathbf{p})\}^{-1/2}A\{D(\mathbf{p})\}^{-1/2} = \{D(\mathbf{p})\}^{-1}\Sigma_\theta\{D(\mathbf{p})\}^{-1} = D(1/\mathbf{p}) - \mathbf{1}\mathbf{1}'.$$

Now $\mathrm{Var}(\bar{\mathbf{b}}) = n^{-1}\Sigma_\theta$, so the QIF

$$Q_{\mathrm{u}}^2 = \bar{\mathbf{b}}'\{\mathrm{Var}(\bar{\mathbf{b}})\}^{-1}\bar{\mathbf{b}} = n(\hat{\mathbf{p}} - \mathbf{p})'\{D(1/\mathbf{p}) - \mathbf{1}\mathbf{1}'\}(\hat{\mathbf{p}} - \mathbf{p}) = n(\hat{\mathbf{p}} - \mathbf{p})'D(1/\mathbf{p})(\hat{\mathbf{p}} - \mathbf{p}),$$

which is Pearson's χ^2. □

PROOF OF (12). By the definition of the QIF,

$$\begin{aligned} Q_{\mathrm{u}}^2 &= \Big(\sum \mathbf{b}_i\Big)' \Big(\sum \mathbf{b}_i \mathbf{b}_i'\Big)^{-1} \Big(\sum \mathbf{b}_i\Big) \\ &= \Big(\sum \mathbf{b}_i\Big)' \Big(\sum (\mathbf{b}_i - \bar{\mathbf{b}})(\mathbf{b}_i - \bar{\mathbf{b}})' + n\bar{\mathbf{b}}\bar{\mathbf{b}}'\Big)^{-1} \\ &\quad \cdot \Big(\sum \mathbf{b}_i\Big) \\ &= \Big(\sum \mathbf{b}_i\Big)' (S^2 + n\bar{\mathbf{b}}\bar{\mathbf{b}}')^{-1} \Big(\sum \mathbf{b}_i\Big) \\ &= \Big(\sum \mathbf{b}_i\Big)' S^{-1} (I + nS^{-1}\bar{\mathbf{b}}\bar{\mathbf{b}}' S^{-1})^{-1} S^{-1} \Big(\sum \mathbf{b}_i\Big). \end{aligned}$$

It can be shown that $(I + \mathbf{c}\mathbf{c}')^{-1} = I - k\mathbf{c}\mathbf{c}'$, where $k = (1 + \mathbf{c}'\mathbf{c})^{-1}$. We can calculate $k = (1 + n^{-1}Q_{\mathrm{c}}^2)^{-1}$ in our case. Hence, to simplify Q_{u}^2, we have

$$\begin{aligned} Q_{\mathrm{u}}^2 &= Q_{\mathrm{c}}^2 - n^{-1}Q_{\mathrm{c}}^4 (1 + n^{-1}Q_{\mathrm{c}}^2)^{-1} \\ &= Q_{\mathrm{c}}^2/(1 + n^{-1}Q_{\mathrm{c}}^2). \end{aligned}$$

□

ACKNOWLEDGMENTS

We appreciate constructive suggestions from the referees, an Editor and the Executive Editor. The research was supported by National Science Foundation Grants DMS-01-04443 (Lindsay) and DMS-01-03513 (Qu).

REFERENCES

AITCHISON, J. and SILVEY, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29** 813–828.

BARANKIN, E. and GURLAND, J. (1951). On asymptotically normal efficient estimators. I. *Univ. California Publ. Statist.* **1** 89–129.

BARNHART, H. X. and WILLIAMSON, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics* **54** 720–729.

BERAN, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83** 687–697.

BOOS, D. D. (1992). On generalized score tests. *Amer. Statist.* **46** 327–333.

CHAGANTY, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *J. Statist. Plann. Inference* **63** 39–54.

CHEN, H. Y. and LITTLE, R. J. A. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika* **86** 1–13.

CORCORAN, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika* **85** 967–972.

CORCORAN, S. A. and DAVISON, A. C. (1995). Reliable inference from empirical likelihoods. Unpublished manuscript.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

CRESSIE, N. and READ, T. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464.

DAVIDSON, R. and MACKINNON, J. G. (1983). Small sample properties of alternative forms of the Lagrange multiplier test. *Econom. Lett.* **12** 269–275.

DAVIDSON, R. and MACKINNON, J. G. (1984). Convenient specification tests for logit and probit models. *J. Econometrics* **25** 241–262.

EFRON, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82** 171–200.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

FERGUSON, T. S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Ann. Math. Statist.* **29** 1046–1062.

FIELD, C. A. (1982). Small sample asymptotic expansions for multivariate M-estimates. *Ann. Statist.* **10** 672–689.

FOUTZ, R. V. and SRIVASTAVA, R. C. (1977). The performance of the likelihood ratio test when the model is incorrect. *Ann. Statist.* **5** 1183–1194.

GALLANT, A. R. and TAUCHEN, G. (1996). Which moments to match? *Econometric Theory* **12** 657–681.

GODAMBE, V. P. (1960). An optimum property of regular maximum-likelihood estimation. *Ann. Math. Statist.* **31** 1208–1211.

HALL, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14** 1431–1452.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

HALL, P. and HOROWITZ, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* **64** 891–916.

HANSEN, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.

HANSEN, L., HEATON, J. and YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14** 262–280.

HARRIS, D. and MÁTYÁS, L. (1999). Introduction to the generalized method of moments estimation. In *Generalized Method of Moments Estimation* (L. Mátyás, ed.) 1–30. Cambridge Univ. Press.

HEYDE, C. C. and MORTON, R. (1998). Multiple roots in general estimating equations. *Biometrika* **85** 954–959.

HU, F. and KALBFLEISCH, J. D. (2000). The estimating function bootstrap (with discussion). *Canad. J. Statist.* **28** 449–499.

IMBENS, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Rev. Econom. Stud.* **64** 359–383.

IMBENS, G. W., SPADY, R. H. and JOHNSON, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* **66** 333–357.

JING, B. and ROBINSON, J. (1994). Saddlepoint approximations for marginal and conditional probabilities of transformed variables. *Ann. Statist.* **22** 1115–1132.

KITAMURA, Y. and STUTZER, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65** 861–874.

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241.

LAHIRI, S. N. (1996). On Edgeworth expansion and moving block bootstrap for Studentized M-estimators in multiple linear regression models. *J. Multivariate Anal.* **56** 42–59.

LEE, M. J. (1996). *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*. Springer, New York.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

LINDSAY, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22** 1081–1114.

MÁTYÁS, L., ed. (1999). *Generalized Method of Moments Estimation*. Cambridge Univ. Press.

MONTI, A. C. and RONCHETTI, E. (1993). On the relationship between empirical likelihood and empirical saddlepoint approximations for multivariate M-estimators. *Biometrika* **80** 329–338.

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (R. Engle and D. McFadden, eds.) **4** 2111–2245. North–Holland, Amsterdam.

NEWEY, W. K. and SMITH, R. J. (2002). Higher order properties of GMM and generalized empirical likelihood estimators. Unpublished manuscript.

NEWEY, W. K. and WEST, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *Internat. Econom. Rev.* **28** 777–787.

NEYMAN, J. (1949). Contribution to the theory of the χ^2 test. *Proc. Berkeley Symp. Math. Statist. Probab.* 239–273. Univ. California Press.

NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London Ser. A* **231** 289–337.

OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120.

PAN, W. (2002). Goodness-of-fit tests for GEE with correlated binary data. *Scand. J. Statist.* **29** 101–110.

PARK, C. (2000). Robust estimation and testing based on quadratic inference functions. Ph.D. dissertation, Dept. Statistics, Pennsylvania State Univ.

PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser. 5* **50** 157–175.

QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325.

QU, A. and LINDSAY, B. G. (1999). Bootstrapping for hypothesis testing using quadratic inference functions for longitudinal data. Technical report, Likelihood Center, Dept. Statistics, Pennsylvania State Univ.

QU, A. and LINDSAY, B. G. (2003). Building adaptive estimating equations when inverse-of-covariance estimation is difficult. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 127–142.

QU, A., LINDSAY, B. G. and LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87** 823–836.

QU, A. and SONG, P. X. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika* **89** 841–850.

RAO, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Philos. Soc.* **44** 50–57.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

ROBINSON, J., RONCHETTI, E. and YOUNG, G. A. (2003). Saddlepoint approximations and tests based on multivariate M-estimates. *Ann. Statist.* **31** 1154–1169.

RONCHETTI, E. and TROJANI, F. (2002). Saddlepoint approximations and test statistics for accurate inference in overidentifying moment conditions models. Unpublished manuscript.

ROTNITZKY, A. and JEWELL, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77** 485–497.

ROTNITZKY, A. and WYPIJ, D. (1994). A note on the bias of estimators with missing data. *Biometrics* **50** 1163–1170.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SHULTS, J. and CHAGANTY, N. R. (1998). Analysis of serially correlated data using quasi-least squares. *Biometrics* **54** 1622–1630.

SILVEY, S. D. (1959). The Lagrangian multiplier test. *Ann. Math. Statist.* **30** 389–407.

SINGH, A. C. and MANTEL, H. J. (1998). Minimum chi-square estimating function and the problem of choosing among multiple roots. In *Proc. Biometrics Section* 102–107. Amer. Statist. Assoc., Alexandria, VA.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

SMALL, C. G., WANG, J. and YANG, Z. (2000). Eliminating multiple root problems in estimation (with discussion). *Statist. Sci.* **15** 313–341.

SMALL, D. (2002). Inference and model selection for instrumental variables regression. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

SMITH, R. J. (1997). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Econom. J.* **107** 503–519.

SPRENT, P. (1966). A generalized least-squares approach to linear functional relationships (with discussion). *J. Roy. Statist. Soc. Ser. B* **28** 278–297.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25.