Statistical Science
2004, Vol. 19, No. 2, 239–263
DOI 10.1214/088342304000000152

The Indirect Method: Inference Based on Intermediate Statistics—A Synthesis and Examples

Wenxin Jiang and Bruce Turnbull

Abstract. This article presents an exposition and synthesis of the theory and some applications of the so-called indirect method of inference. These ideas have been exploited in the field of econometrics, but less so in other fields such as biostatistics and epidemiology. In the indirect method, statistical inference is based on an intermediate statistic, which typically follows an asymptotic normal distribution, but is not necessarily a consistent estimator of the parameter of interest. This intermediate statistic can be a naive estimator based on a convenient but misspecified model, a sample moment or a solution to an estimating equation. We review a procedure of indirect inference based on the generalized method of moments, which involves adjusting the naive estimator to be consistent and asymptotically normal. The objective function of this procedure is shown to be interpretable as an "indirect likelihood" based on the intermediate statistic. Many properties of the ordinary likelihood function can be extended to this indirect likelihood. This method is often more convenient computationally than maximum likelihood estimation when handling such model complexities as random effects and measurement error, for example, and it can also serve as a basis for robust inference and model selection, with less stringent assumptions on the data generating mechanism. Many familiar estimation techniques can be viewed as examples of this approach. We describe applications to measurement error, omitted covariates and recurrent events. A dataset concerning prevention of mammary tumors in rats is analyzed using a Poisson regression model with overdispersion. A second dataset from an epidemiological study is analyzed using a logistic regression model with mismeasured covariates. A third dataset of exam scores is used to illustrate robust covariance selection in graphical models.

Key words and phrases: Asymptotic normality, bias correction, consistency, efficiency, estimating equations, generalized method of moments, graphical models, indirect inference, indirect likelihood, measurement error, missing data, model selection, naive estimators, omitted covariates, overdispersion, quasi-likelihood, random effects, robustness.

Wenxin Jiang is Associate Professor, Department of Statistics, Northwestern University, Evanston, Illinois 60208, USA (e-mail: wjiang@northwestern.edu). Bruce Turnbull is Professor, Departments of Statistical Science and of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853, USA (e-mail: bwt2@cornell.edu).

1. INTRODUCTION

Methods of "indirect inference" have been developed and used in the field of econometrics where they have proved valuable for parameter estimation in highly complex models. However, it is not widely recognized that similar ideas are extant generally in a number of other statistical methods and applications, and there they have not been exploited as such to the fullest extent.

This article was motivated by our experience in analyzing repeated events data for the Nutritional Prevention of Cancer (NPC) trial (Clark et al., 1996). The results reported there were quite controversial, suggesting substantial health benefits from long term daily supplementation with a nutritional dose of selenium, an antioxident. Early on, it was recognized that the subject population was heterogeneous and that there were sources of variability and biases not accounted for by standard statistical analyses—these included covariate measurement error, omitted covariates, missing data and overdispersion. However, the dataset, being large and complex, did not lend itself well to statistical methods that required complicated computations. Instead, convenient available statistical software was used that was based on fairly straightforward (nonlinear) regression models. The outputted results based on these naive models were then examined in the light of known and putative deviations from the model and inferences were adjusted accordingly. The details of this case study were described in Jiang, Turnbull and Clark (1999).

This is an example of a general approach, termed *indirect inference* (Gouriéroux, Monfort and Renault, 1993), which was motivated by complex dynamic financial models. Here maximum likelihood (ML) estimates are difficult to obtain despite modern algorithms and computing power, due to the presence of many latent variables and high-dimensional integrals. Another consideration in these applications is the desire to obtain estimates that are robust to misspecification of the underlying model.

1.1 Indirect Inference

Suppose we have a dataset consisting of n independent units. The essential ingredients of the indirect approach are as follows.

1. There is a hypothesized true model M for data generation, with distribution $P^{(\theta)}$ which depends on an unknown parameter of interest θ, which is of dimension p.
2. One first computes an *intermediate* or *auxiliary* statistic $\hat{s} = \Psi(P^{(n)})$ of dimension $q \geq p$ which is a functional of the empirical distribution function $P^{(n)}$, say.
3. A *bridge* (or *binding*) relationship $s = \Psi(P^{(\theta)})$ is defined. The unknown quantity s is called the *auxiliary parameter*.
4. With the auxiliary estimate $\hat{s}$ replacing s, the bridge relationship above is used to compute an *adjusted* estimate $\hat{\theta}(\hat{s})$ for θ.

The goals to be achieved in this approach include the following. We would like the estimator $\hat{\theta}(\hat{s})$ to be (1) *robust* to model M misspecification, in the sense that $\hat{\theta}(\hat{s})$ remains a consistent estimator of θ under a larger class of models $\mathcal{M}$ that includes M, and (2) relatively easy to compute. To attain these two goals, we will base our inference on the auxiliary statistic $\hat{s}$ which may not be sufficient under model M. Therefore, a third goal is that the estimator $\hat{\theta}(\hat{s})$ have high efficiency under M.

The starting point is the choice of an intermediate statistic $\hat{s}$. This can be chosen as some set of sample moments or the solution of some estimating equations or the ML estimator (MLE) based on some convenient model M′, say, termed the *auxiliary* (or *naive*) model. If the last, then the model M′ is a simpler but misspecified or partially misspecified model. The choice of an intermediate statistic $\hat{s}$ is not necessarily unique; however, in any given situation there is often a natural one to use. The theory of properties of estimators obtained from misspecified likelihoods goes back at least as far as Cox (1962), Berk (1966) and Huber (1967), and is summarized in the comprehensive monograph by White (1994). The use of $\hat{s}$ (based on an auxiliary model M′) in indirect inference about θ (under model M) appeared recently in the field of econometrics to treat complex time series and dynamic models (see, e.g., Gouriéroux, Monfort and Renault, 1993; Gallant and Tauchen, 1996, 1999), as well as in the field of biostatistics to treat regression models with random effects and measurement error (see, e.g., Kuk, 1995; Turnbull, Jiang and Clark, 1997; Jiang, Turnbull and Clark, 1999).

The econometric applications of the indirect approach have been primarily motivated by goal 2; for example, to perform inference for financial data based on stochastic differential equation or stochastic volatility models, where the usual maximum likelihood-based approach is intractable (see, e.g., Mátyás, 1999, Chapter 10; Carrasco and Florens, 2002, for reviews). In contrast, the goal of robustness as described in goal 1

has been an important consideration in recent biostatistical applications (e.g., see Lawless and Nadeau, 1995, and further references in Section 2.5). Recent work (Genton and Ronchetti, 2003) has shown how indirect inference procedures can also be made robust in the sense of stability in the presence of outliers. Both senses of robustness are discussed further in Section 2.5.

1.2 Method of Moments as Indirect Inference

The method of moments can be formulated as indirect inference. Consider an intermediate statistic $\hat{s} = \Psi(F_n) = (\overline{X}, S^2, \dots)^T$ with components that contain some sample moments such as the mean $\overline{X}$ and the variance S^2. Then the bridge equation is $s = s(\theta) = \Psi(F_\theta) = (\mu(\theta), \sigma^2(\theta), \dots)^T$ with components of population moments, that is, mean $\mu(\theta)$, variance $\sigma^2(\theta)$ and so on. The vector of q population moments is the auxiliary parameter s.

In the usual *method of moments* (MM), $\dim(s) = q = p = \dim(\theta)$, we solve $\hat{s} = s(\hat{\theta})$ for $\hat{\theta}$, the MM estimator. (We assume the solution is uniquely defined.) If $q > p$, then we can instead take $\hat{\theta}$ as

$$\hat{\theta} = \arg\min_{\theta} \{\hat{s} - s(\theta)\}^T v^{-1} \{\hat{s} - s(\theta)\},$$

where v is a positive definite matrix, such as a sample estimate of the asymptotic variance (avar) of $\hat{s}$. This is an example of the *generalized method of moments* (GMM; Hansen, 1982). In the *simulated method of moments* (SMM; McFadden, 1989; Pakes and Pollard, 1989), the moments $s(\theta)$ are too difficult to compute analytically. Instead, $s(\theta)$ is evaluated as a function of θ by Monte Carlo simulation.

Now, the full GMM method is a very broad approach to estimation which includes maximum likelihood, estimating equations, least squares, two-stage least squares and many other estimation procedures as special cases (see, e.g., Imbens, 2002). Since the indirect method is also a unifying framework for estimation procedures, it is not surprising that there is a strong connection between it and GMM. This connection is described further in Section 2.7.

1.3 Three Pedagogic Examples

The steps involved in the indirect method are illustrated in the following simple pedagogic examples. In fact, in all three of these examples, the adjusted estimators can be viewed as MM estimators; however, it is instructive to consider them in the indirect inference framework of Section 1.1.

EXAMPLE 1 (Exponential observations with censoring). Consider lifetimes $\{T_1, \dots, T_n\}$, which are independent and identically distributed (i.i.d.) according to an exponential distribution with mean θ. The data are subject to Type I single censoring after fixed time c. Thus the observed data are $\{Y_1, \dots, Y_n\}$, where $Y_i = \min(T_i, c)$ $(i = 1, \dots, n)$. We consider indirect inference based on the intermediate statistic $\hat{s} = \overline{Y}$. This choice can be considered either as the basis for an MM estimator or as the MLE for a misspecified model M′ in which the presence of censoring has been ignored. The naive estimator $\overline{Y}$ in fact consistently estimates not θ, but the naive or auxiliary parameter

$$s = \theta[1 - \exp(-c/\theta)], \tag{1}$$

the expectation of $\overline{Y}$. Equation (1) is an example of what we term a bridge relationship. We can see the obvious effect of the misspecification, namely that $\hat{s}$ underestimates θ. However, a consistent estimate $\hat{\theta}$ of θ as $n \to \infty$ can be obtained by solving (1) for θ with s replaced by $\hat{s} = \overline{Y}$. (Note that this is not the MLE of θ, which is $n\overline{Y}/[\sum_{i=1}^n I(Y_i < c)]$.) In the later sections we will see how to obtain the standard error for the adjusted estimate $\hat{\theta}$.

EXAMPLE 2 (Zero-truncated Poisson data). The zero-truncated Poisson distribution $\{(\exp(-\theta)\theta^y)/(1 - \exp(-\theta)y!);\ y = 1, 2, \dots\}$ is a model for positive count data—the number of articles by an author, for example. Suppose $Y_1, \dots, Y_n$ is an i.i.d. sample from this distribution. Suppose, however, that the zero truncation is overlooked and the standard Poisson likelihood $\prod_{i=1}^n \{\exp(-\theta)\theta^{y_i}/y_i!\}$ is used. The naive estimator $\hat{s} = \overline{Y}$ is consistently estimating $E(\hat{s}) = s = \theta/[1 - \exp(-\theta)]$. This is the bridge relationship and, with $\hat{s}$ in place of s, it can be inverted to obtain a consistent estimator $\hat{\theta}$ of θ. In this case, it coincides with the MLE based on the true likelihood and is asymptotically efficient.

EXAMPLE 3 (Multinomial genetic data). Dempster, Laird and Rubin (1977, Section 1) fitted some phenotype data given by Rao (1973, page 369) to a genetic linkage model described by Fisher (1946, page 303). The sample consists of $n = 197$ progeny which are distributed multinomially into four phenotypic categories according to probabilities from an intercross model M of the genotypes AB/ab × AB/ab : $(\frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1 - \theta), \frac{1}{4}(1 - \theta), \frac{1}{4}\theta)$ for some $\theta \in [0, 1]$. The corresponding observed counts are

$$\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34).$$

For the first step, we define an intermediate statistic as a naive estimate of θ from a "convenient" but misspecified model M′ in which it is wrongly assumed that **y** is drawn from a four-category multinomial distribution with probabilities $(\frac{1}{2}s, \frac{1}{2}(1-s), \frac{1}{2}(1-s), \frac{1}{2}s)$. This corresponds to a backcross of the genotypes AB/ab × ab/ab. The naive model is convenient because the naive MLE is simply calculated as $\hat{s} = (y_1 + y_4)/n = (125 + 34)/197 = 0.8071$. In the second step, we derive a bridge relationship which relates the "naive parameter" s (large sample limit of $\hat{s}$) to the true parameter θ. Here the bridge relationship is $s = (1+\theta)/2$, since, under the true model, this is the almost sure limit of $\hat{s}$ as $n \to \infty$. The third step is to invert the bridge relationship to obtain the adjusted estimate $\hat{\theta} = 2\hat{s} - 1 = (y_1 + y_4 - y_2 - y_3)/n = 0.6142$. Of course, in this case the maximum likelihood estimate based on the true model, $\hat{\theta}_{\mathrm{ML}}$ say, can be computed explicitly as

$$\begin{aligned}\hat{\theta}_{\mathrm{ML}} &= \big(y_1 - 2y_2 - 2y_3 - y_4 \\ &\qquad + \sqrt{(y_1 - 2y_2 - 2y_3 - y_4)^2 + 8ny_4}\big)/(2n) \\ &= 0.6268,\end{aligned}$$

which can be obtained directly from solving the score equation. Alternatively, the expectation-maximization algorithm can be used as in Dempster, Laird and Rubin (1977, Section 1). The MLE $\hat{\theta}_{\mathrm{ML}}$ is biased, unlike the adjusted estimator $\hat{\theta}$, but has smaller variance than $\hat{\theta}$. We have $\operatorname{Var}\hat{\theta} = 4\operatorname{Var}\hat{s} = 4s(1-s)/n$, which can be estimated as $4\hat{s}(1-\hat{s})/n = 0.0032$. This compares with $\operatorname{Var}\hat{\theta}_{\mathrm{ML}} = 0.0026$, obtained from the sample Fisher information. The asymptotic efficiency of $\hat{\theta}$ relative to $\hat{\theta}_{\mathrm{ML}}$ is therefore estimated to be $0.0026/0.0032 = 0.81$. The loss of efficiency is due to model misspecification; $\hat{s}$ is not sufficient under model M.

When $\hat{\theta}$ is not efficient, a general method for obtaining an asymptotically fully efficient estimator $\tilde{\theta}$ is via a one-step Newton–Raphson correction or "efficientization" (e.g., see Le Cam, 1956; White, 1994, page 137; Lehmann and Casella, 1998, page 454). Specifically, since $\hat{\theta}$ is consistent and asymptotically normal, the estimator

$$\tilde{\theta} = \hat{\theta} - \{\partial_{\hat{\theta}} S(\hat{\theta})\}^{-1} S(\hat{\theta}), \tag{2}$$

where $S(\cdot)$ is the true score function, is asymptotically the same as the ML estimate and hence achieves full efficiency. For complicated likelihoods, the one-step efficientization method, which requires the evaluation of $S(\hat{\theta})$ and $\partial_{\hat{\theta}} S(\hat{\theta})$ only once, can greatly reduce the computational effort compared to that for $\hat{\theta}_{\mathrm{ML}}$. In our genetic linkage example the true log-likelihood function is

$$\begin{aligned}L = Y_1 \log\left(\frac{1}{2} + \frac{\theta}{4}\right) + (Y_2 + Y_3)\log\left(\frac{1}{4} - \frac{\theta}{4}\right) \\ + Y_4 \log\left(\frac{\theta}{4}\right).\end{aligned}$$

First- and second-order derivatives of L can easily be evaluated, leading to the one-step correction estimator

$$\begin{aligned}\tilde{\theta} &= \hat{\theta} + \frac{Y_1(2+\hat{\theta})^{-1} - (Y_2 + Y_3)(1-\hat{\theta})^{-1} + Y_4\hat{\theta}^{-1}}{Y_1(2+\hat{\theta})^{-2} + (Y_2 + Y_3)(1-\hat{\theta})^{-2} + Y_4\hat{\theta}^{-2}} \\ &= 0.6271.\end{aligned}$$

This estimate is closer to the MLE $\hat{\theta}_{\mathrm{ML}} = 0.6268$ and has the same asymptotic variance of 0.0026. Thus we have obtained a consistent and asymptotically efficient estimate.

Another way to increase efficiency is to incorporate more information into the intermediate statistics. For example, all information of the data is incorporated if we instead define the intermediate statistic $\hat{s} = (y_1/n, y_2/n, y_3/n)^T$ [the last cell frequency is determined by $(1 - \hat{s}_1 - \hat{s}_2 - \hat{s}_3)$]. Here $q = \dim(\hat{s}) = 3 > 1 = p = \dim(\theta)$. The new bridge relationship is $s = s(\theta) = (\frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta))$. If we use the generalized method of moments and choose v to be an estimate of the asymptotic variance $\widehat{\operatorname{var}}(\hat{s})$ of $\hat{s}$ with the jkth element being $(\hat{s}_j\delta_{jk} - \hat{s}_j\hat{s}_k)/n$ (δ_{jk} is the Kronecker delta), then the adjusted estimate is $\hat{\theta} = \arg\min_\theta \{\hat{s} - s(\theta)\}^T v^{-1} \{\hat{s} - s(\theta)\}$. This expression yields

$$\begin{aligned}\hat{\theta} &= \big(Y_1^{-1} + Y_2^{-1} + Y_3^{-1} + Y_4^{-1}\big)^{-1} \\ &\qquad \cdot \big(-2Y_1^{-1} + Y_2^{-1} + Y_3^{-1}\big) \\ &= 0.6264,\end{aligned}$$

which is closer to the ML estimator. Later, in Proposition 1(ii), we will show that the asymptotic variance of $\hat{\theta}$ can be estimated by $\widehat{\operatorname{var}}(\hat{\theta}) = 2(\partial_\theta^2 H)^{-1}|_{\theta=\hat{\theta}}$, where $\partial_\theta^2 H$ is the Hessian of the objective function $H = \{\hat{s} - s(\theta)\}^T v^{-1} \{\hat{s} - s(\theta)\}$. In this example, upon evaluation, we obtain

$$\begin{aligned}\widehat{\operatorname{var}}(\hat{\theta}) &= \frac{16}{n^2}\big(Y_1^{-1} + Y_2^{-1} + Y_3^{-1} + Y_4^{-1}\big)^{-1} \\ &= 0.0029.\end{aligned}$$

The avar estimate now is very close to that of the ML estimator. In fact, here $\hat{\theta}$ is fully efficient because now

it is based on an intermediate statistic $\hat{s}$ that is sufficient under model M. The difference of the avar estimates arises because of the finite sample size. One should note that the method here is the minimum chi-square approach of Ferguson (1958) recast in terms of the indirect method.

1.4 Outline of the Article

The approach described has been used in a variety of statistical problems, but has not really been exploited on a systematic basis, with the exception of the considerable work in the field of econometrics. The present article is intended to provide a synthesis of a number of different ideas from different fields, illustrating them with examples from various applications (in fields other than econometrics).

Our unifying concept is inference using the framework of an approximate likelihood based on the intermediate statistic (the *indirect* likelihood), instead of one based on the full data. The current article may be viewed as an attempt to extend an analysis based on "complete data plus a complete probability model" to an asymptotic analysis based on "some compressed data $\hat{s}$ plus a model for its asymptotic mean." This extension allows flexibility for a spectrum of tradeoffs between robustness and efficiency. Often, a more compressed intermediate statistic leads to a lower efficiency under model M, but produces a consistent indirect likelihood estimator that relies on less assumptions about M. This indirect approach offers the following advantages:

1. *Ease of computation.* The indirect method is often computationally simpler or more convenient (e.g., $\hat{s}$ often can be computed with standard software if it is based on a standard auxiliary model M′).
2. *Informativeness on the effect of model misspecification*. When $\hat{s}$ is a naive estimate obtained from a naive model M′ by neglecting certain model complexity, the current approach is very informative on the effect of model misspecification—the bridge relationship $s = s(\theta)$ provides a dynamic correspondence between M′ and M. In fact, such a relationship is of central importance in, for example, errors-in-variables regression, where such a relationship is sometimes termed an attenuation relationship (see, e.g., Carroll, Ruppert and Stefanski, 1995, Chapter 2), which tells how the regression slope can be underestimated when neglecting the measurement error in a predictor.
3. *Robustness.* We will see that the validity of the inference based on an intermediate statistic essentially relies on the correct specification of its asymptotic mean. This is often a less demanding assumption than the correct specification of a full probability model, which would be generally needed for a direct likelihood inference to be valid. Therefore, the inferential result based on the adjusted estimate $\hat{\theta}$ often remains valid despite some departure of the data generation mechanism from the hypothesized true model M. Another, perhaps more traditional, sense of robustness is that of protection against outliers. It is possible to make indirect inference procedures resistant to outliers. Both senses of robustness are further discussed in Section 2.5.

In Section 2 we summarize the theory, integrating literature from different fields. In Section 3, we present some applications of the bridge relationship in assessing the robustness and sensitivity of an unadjusted naive estimator regarding model misspecification (when M is misspecified as M′). Examples include Poisson estimation, omitted covariates, measurement error and missing data. Section 4 includes three analyses: a carcinogenicity dataset is modelled by a Poisson regression model with random effects (overdispersion); an epidemiological dataset concerns a mismeasured covariate; a well-known multivariate dataset of mathematics exam scores illustrates robust model selection. In the Conclusion, we list some more statistical procedures that can be recast as examples of indirect inference, including importance sampling and applications to gene mapping.

2. THEORY

2.1 Auxiliary Statistic

Under the hypothesized true model M, we suppose that the observed data $\mathbf{W}$ come from n subjects or units, independently generated by a probability distribution $P^{(\theta)}$, which depends on an unknown p-dimensional parameter θ. It is desired to make inferences concerning θ.

The indirect method starts with an *auxiliary* or *intermediate statistic* $\hat{s} = \hat{s}(\mathbf{W})$, which can be generated by the method of moments, least squares (LS) or a likelihood analysis based on a convenient misspecified model M′, for example. Most such intermediate statistics can be defined implicitly as a solution, $s = \hat{s}$, of a (q-dimensional) estimating equation of the form

$G(\mathbf{W}, s) = 0$, say. [Clearly this includes any statistic $\hat{s} = \hat{s}(\mathbf{W})$ that has an explicit expression as a special case, by taking $G = s - \hat{s}(\mathbf{W})$.] The estimating equation could be the normal equation from an LS analysis, the score equation based on some likelihood function or the zero-gradient condition for a GMM analysis.

Note that $\hat{s}$ is typically asymptotically normal (AN) and $\sqrt{n}$ consistent for estimating some $s = s(\theta)$, the auxiliary parameter (see, e.g., White, 1994, Theorem 6.4, page 92, for the case when G is a score function based on a naive/misspecified likelihood). In our exposition, the theory of indirect inference methods will be based on this AN property for the intermediate statistic $\hat{s}$ alone, noting that this property can hold even if the complete original model $P^{(\theta)}$ for the data $\mathbf{W}$ is invalid. Our intermediate model is now

$$n^{1/2}\{\hat{s} - s(\theta)\} \xrightarrow{D} N(0, v). \tag{3}$$

Here $\hat{s}$ and $s(\theta)$ are of dimension q, where the *auxiliary parameter* $s = s(\theta)$ is the asymptotic mean of $\hat{s}$. (When $\hat{s}$ is based on a naive model M′, we sometimes alternatively term s a *naive parameter*.) Also, $n^{-1}v = \mathrm{var}(\hat{s})$ is the $q \times q$ asymptotic variance (avar) of $\hat{s}$. In general, the avar of $\hat{s}$ has a sandwich form:

$$\begin{aligned} \mathrm{var}(\hat{s}) &= n^{-1}v \\ &= (E\,\partial_s G)^{-1}\,\mathrm{var}(G)(E\,\partial_s G)^{-T}|_{s=s(\theta)}. \end{aligned} \tag{4}$$

Here we use superscripts T for transpose, and $-T$ for inverse *and* transpose. The derivative matrix is defined by $[\partial_s G]_{jk} = \partial_{s_k} G_j$, $j, k = 1, \ldots, q$, $G = (G_1, \ldots, G_q)^T$ and $s = (s_1, \ldots, s_q)^T$.

2.2 The Bridge Equation

Note that, as an asymptotic mean of $\hat{s}$, $s(\theta)$ is not unique: $s(\theta) + o(n^{-1/2})$ would do as well. We usually choose a version of $s(\theta)$ which does not depend on n, if available. Alternatively, we may use the actual expectation, $s(\theta) = E_{\mathbf{W}|\theta}\hat{s}$. Now $s(\theta)$, the consistent limit of $\hat{s}$, is not equal to the true parameter θ in general and not even necessarily equal in dimension. For problems with model misspecification, the naive parameter $s(\theta)$ establishes a mapping which plays a central role in bias correction and is referred to as the *binding function* (Gouriéroux, Monfort and Renault, 1993) or *bridge relationship* (Turnbull, Jiang and Clark, 1997; Jiang, Turnbull and Clark, 1999), because it relates what the naive model really estimates to the true parameter.

Now we turn to the problem of deriving $s(\theta)$ in two cases:

CASE A. When the naive estimator $\hat{s} = \hat{s}(\mathbf{W})$ has an explicit expression, it is sometimes possible to use the law of large numbers to find its limit directly, as in the examples of Section 1.

CASE B. More commonly, $\hat{s}$ does not have an explicit expression. When $\hat{s}$ maximizes an objective function, its large sample limit may be obtained by maximizing the limit of the objective function. When $\hat{s}$ is implicitly defined as a solution of an estimating equation $G(\mathbf{W}, s) = 0$, and $G(\mathbf{W}, s)$ converges in probability to $E_{\mathbf{W}|\theta}G(\mathbf{W}, s) = F(\theta, s)$, say, as $n \to \infty$, we can find the naive parameter $s(\theta)$ by looking for the solution $s = s_0(\theta)$, say, of the equation $F(\theta, s) = 0$, and take $s(\theta) = s_0(\theta)$.

Note that Case A is a special case of Case B with $G(\mathbf{W}, s) = s - \hat{s}(\mathbf{W})$.

More generally, $\hat{s} = \hat{s}(\mathbf{W})$ is defined as a procedure which maps the data vector to $\Re^q$, and $\hat{s}$ is asymptotically normal. Then $s(\theta)$, being an asymptotic mean of $\hat{s}$, can be computed by $E_{\mathbf{W}|\theta}\hat{s}(\mathbf{W})$. If necessary, this expectation, as a function of θ, can be estimated by a Monte Carlo method: Simulate $\mathbf{W}^{(k)}$, $k = 1, \ldots, m$, i.i.d. $\mathbf{W}|\theta$, and use $s(\theta) \approx m^{-1}\sum_{k=1}^{m}\hat{s}(\mathbf{W}^{(k)})$. For examples, see McFadden (1989), Pakes and Pollard (1989) and Kuk (1995).

2.3 The Adjusted Estimator and the Indirect Likelihood

We now consider inference for the parameter θ under model M based on the intermediate statistic $\hat{s}$. From the assumed AN approximation (3) of $\hat{s}$, we define an *indirect likelihood* $L = L(\theta|\hat{s}) \equiv |2\pi v|^{-1/2}\exp(-H/2)$, where $H = H(\theta, \hat{s}) = \{\hat{s} - s(\theta)\}^T v^{-1}\{\hat{s} - s(\theta)\}$, v is (a sample estimate of) the avar of $\hat{s}$ and $|\cdot|$ denotes determinant. More generally, when $\hat{s}$ is defined implicitly as the solution to an equation of the form $G(\mathbf{W}, s) = 0$, in the definition of the indirect likelihood L, H is defined by $H(\theta, \hat{s}) = F(\theta, \hat{s})^T v^{-1}F(\theta, \hat{s})$, with $F(\theta, s) \equiv E_{\mathbf{W}|\theta}G(\mathbf{W}, s)$. Here v is (a sample estimate of) the avar of $F(\theta, \hat{s})$, which can be evaluated by the delta method (e.g., Bickel and Doksum, 2001, Section 5.3) and found to be the same as $\mathrm{var}(G)$ evaluated at $s = s(\theta)$ (the auxiliary parameter).

We then define the *adjusted estimator* (or the *indirect* MLE) $\hat{\theta}$ to be the maximizer of L or the minimizer of H. This maximizer of L bears properties that are analogous to the usual MLE under mild regularity conditions. The most important condition is the correct specification of the bridge relationship $s = s(\theta)$, or implicitly of $F(\theta, s) = 0$, for the asymptotic mean s of

the intermediate statistic. These results are summarized in the following proposition. We will outline the proof in the explicit form. The proof in the implicit form is similar and is actually asymptotically equivalent after applying the implicit function theorem to the partial derivatives on F.

PROPOSITION 1. *Analogy of the adjusted estimator to the MLE. Suppose*:

(a) $\sqrt{n}\{\hat{s} - s(\theta)\} \xrightarrow{D} N(0, \nu)$;

(b) ν *is positive definite and symmetric and* $n\nu \xrightarrow{p} \nu$;

(c) $s(\cdot)$ *is second-order continuously differentiable in a neighborhood of* θ *and the derivative matrix* s' *is full rank at* θ. [*In the implicit form this condition involves the following*: F *is bivariate continuously differentiable to the second order in a neighborhood of* $(\theta, s(\theta))$, $\partial_s F$ *and* $\partial_\theta F$ *are full rank at* $(\theta, s(\theta))$ *and* F *takes value zero at* $(\theta, s(\theta))$.]

Then we have the following:

(i) *Indirect score function*: *The asymptotic mean and variance of the indirect likelihood score function satisfy the usual relationships* $E(\partial_\theta \log L) = 0$ *and* $\operatorname{var}(\partial_\theta \log L) + E(\partial_\theta^2 \log L) = 0$.

(ii) *Asymptotic normality*: *There exists a closed ball* Θ *centered at the true parameter* θ, *in which there is a measurable adjusted estimator* $\hat{\theta}$ *such that* $\hat{\theta} = \arg\max_{\theta\in\Theta} \log L$ *and* $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N\{0, (s'(\theta)^T \cdot \nu^{-1} s'(\theta))^{-1}\}$. *Alternatively,* $\hat{\theta}$ *is AN with mean* θ, *and with avar estimated by* $-(\partial_\theta^2 \log L)^{-1}$ *or* $2(\partial_\theta^2 H)^{-1}$, *where consistent estimates are substituted for parameter values*.

(iii) *Tests*: *Likelihood-ratio statistics based on the indirect likelihood for testing simple and composite null hypotheses have the usual asymptotic* χ^2 *distributions* (*e.g., under* $H_0 : \theta = \theta_0$, $2\log L(\hat{\theta}) - 2\log L(\theta_0) \xrightarrow{D} \chi^2_{\dim\theta}$).

(iv) *Efficiency I*: *The adjusted estimator has smallest avar among all consistent asymptotically normal* (*CAN*) *estimators* $f(\hat{s})$ *of* θ, *which are constructed from the naive estimator* $\hat{s}$ *by continuously differentiable mappings* f.

PROOF. (i) From Assumption (a), we note that

$$n^{-1}\partial_\theta \log L = -0.5n^{-1}\partial_\theta H = s'(\theta)^T \nu^{-1}\{\hat{s} - s(\theta)\} + o_p(n^{-1/2})$$

and

$$-n^{-1}\partial_\theta^2 \log L = s'(\theta)^T \nu^{-1} s'(\theta) + O_p(n^{-1/2}).$$

Then

$$n^{-1/2}\partial_\theta \log L \xrightarrow{D} N\{0, s'(\theta)^T \nu^{-1} s'(\theta)\}$$

and

$$-n^{-1}\partial_\theta^2 \log L \xrightarrow{p} s'(\theta)^T \nu^{-1} s'(\theta).$$

In this sense, the asymptotic mean of $n^{-1/2}\partial_\theta \log L$ is zero, and the asymptotic variance $n^{-1}\operatorname{var}(\partial_\theta \log L)$ and the asymptotic mean of $-n^{-1}\partial_\theta^2 \log L$ are both equal to $s'(\theta)^T \nu^{-1} s'(\theta)$.

(ii) The AN result is proved by using a usual linear approximation and using the results in (i). The validity of the linear approximation depends on the consistency of $\hat{\theta}$ and a zero-gradient condition, which are justified below.

By conditions (a), (b) and (c) we can choose a closed ball Θ centered at the true parameter θ, such that $\sup_{t\in\Theta} |n^{-1}H(t, \hat{s}) - h(t)| \xrightarrow{p} 0$ and the limiting criterion function $h(t) = \{s(\theta) - s(t)\}^T \nu^{-1}\{s(\theta) - s(t)\}$ has a unique minimum $t = \theta$ located in the interior of Θ. Therefore, the minimizer $\hat{\theta} = \arg\min_{t\in\Theta} n^{-1} \cdot H(t, \hat{s}) \xrightarrow{p} \theta$ and satisfies a zero-gradient condition $\partial_t H(t, \hat{s})|_{t=\hat{\theta}} = 0 = \partial \log L(\hat{\theta})$ with probability tending to 1. Now we expand this zero-gradient condition around $\hat{\theta} \approx \theta$ and use the just-established consistency of $\hat{\theta}$ to characterize the remainder. We obtain $\hat{\theta} - \theta = -\{\partial_\theta^2 \log L(\theta)\}^{-1}\partial_\theta \log L(\theta) + o_p(n^{-1/2})$. Applying the results obtained in the proof of (i) and Slutsky's theorem, we obtain $\hat{\theta} - \theta = \{s'(\theta)^T \nu^{-1} s'(\theta)\}^{-1} s'(\theta)^T \cdot \nu^{-1}\{\hat{s} - s(\theta)\} + o_p(n^{-1/2})$, from which the AN result of (ii) follows.

(iii) Since the AN result (ii) for the parameter estimates has been established, the standard treatment in likelihood-based inference (e.g., Sen and Singer, 1993, Section 5.6) can be applied, based on a second-order Taylor expansion. This results in the the limiting χ^2 distribution of the likelihood-ratio statistics.

(iv) The delta method can be applied to derive $n\operatorname{var}(f(\hat{s})) = f'(s)\nu f'(s)^T$, while result (ii) gives $n\operatorname{var}(\hat{\theta}) = (s'(\theta)^T \nu^{-1} s'(\theta))^{-1}$. The consistency of $f(\hat{s})$ as an estimator of θ implies that $f(s(\theta)) = \theta$ for all θ, implying the constraint $f'(s)s'(\theta) = I$, which in turn implies that a positive semidefinite matrix

$$\begin{aligned}&\left(f' - (s'^T \nu^{-1} s')^{-1} s'^T \nu^{-1}\right) \\ &\quad \cdot \nu\left(f' - (s'^T \nu^{-1} s')^{-1} s'^T \nu^{-1}\right)^T \\ &\quad = f'(s)\nu f'(s)^T - \left(s'(\theta)^T \nu^{-1} s'(\theta)\right)^{-1}.\end{aligned}$$

This last equation shows that $n\operatorname{var}(f(\hat{s}))$ is never less than $n\operatorname{var}(\hat{\theta})$ in the matrix sense. □

This proposition represents a summary of results that have appeared in varying forms and generality and tailored for various applications. For example, (iv) is a stronger version and synthesis of various optimality results in the existing literature such as the optimal quadratic criterion function in indirect inference (Gouriéroux, Monfort and Renault, 1993, Proposition 4), the optimal linear combination of moment conditions in GMM (Hansen, 1982, Theorem 3.2; McCullagh and Nelder, 1989, page 341), the method of linear forms (Ferguson, 1958, Theorem 2) and the regular best AN estimates that are functions of sample averages (Chiang, 1956, Theorem 3).

Recognizing that the maximization of L is the same as minimizing H, we can often view the method of minimum χ^2 or GMM as likelihood inference based on an intermediate statistic. For example, in the simulated method of moments and indirect inference, either the explicit (McFadden, 1989; Pakes and Pollard, 1989; Gouriéroux, Monfort and Renault, 1993; Newey and McFadden, 1994) or the implicit form (Gallant and Tauchen, 1996, 1999; Gallant and Long, 1997) of the GMM criterion function H is used, and applied to econometric and financial problems. Applications of GMM in the settings of generalized estimating equations from biostatistics were discussed by Qu, Lindsay and Li (2000).

In a special case when the dimension of the intermediate statistic (q) equals that (p) of the parameter θ, and $s(\cdot)$ is a diffeomorphism on the parameter space Θ of θ, maximization of L is equivalent to the bias correction $\hat{\theta} = s^{-1}(\hat{s})$ [from solving $F(\theta, \hat{s}) = 0$], which is AN and consistent for θ (see, e.g., Kuk, 1995; Turnbull, Jiang and Clark, 1997; Jiang, Turnbull and Clark, 1999, for biostatistical applications). In fact, when $\hat{s}$ is itself already asymptotically unbiased, the above adjustment procedure can still be used to remove small-sample bias of order $O(1/n)$ by solving for $\hat{\theta}$ from $\hat{s} - E_{\mathbf{W}|\theta}\hat{s}(\mathbf{W}) = 0$ (MacKinnon and Smith, 1998).

When $q < p$, there are more unknown true parameters than naive parameters. In this case, the bridge relationship is many-to-one and does not, in general, permit the construction of adjusted estimates. It is mainly of interest for investigating the effects of misspecification when the naive estimators are constructed under misspecified models; see Section 3.3, for example. However, in such situations it may be possible to construct consistent estimates for a subset of true parameters, which may be of interest. In other situations, some components of the higher-dimensional true parameter are known or can be estimated from other outside data sources. This enables the other components to be consistently estimated by inverting the bridge relationship. Examples of this kind arising from errors-in-variables regression models are given in Sections 3.2 and 4.2.

2.4 Efficiency of the Adjusted Estimator

In general, the intermediate statistic $\hat{s}$ is not a sufficient statistic of θ under the true model M and the indirect MLE $\hat{\theta}$ based on the intermediate data $\hat{s}$ is not as efficient as the MLE $\hat{\theta}_{\mathrm{ML}}$ based on the complete data $\mathbf{W}$. However, Cox (1983) and Jiang, Turnbull and Clark (1999) provided examples of situations when the efficiencies of $\hat{\theta}$ are quite high for some parameter components; see also the example of Section 4.1.

Proposition 1(iv) has already given our first result concerning the efficiency of $\hat{\theta}$. Further results on the efficiency of $\hat{\theta}$ under model M are summarized in the following two propositions. Proposition 2 provides necessary and sufficient conditions for the entire vector of $\hat{\theta}$ [parts (i) or (ii)] or some of its components [part (ii)] to be as efficient as the MLE. Proposition 3 provides a geometric view of the relative efficiency and avars for the three CAN estimators considered in this article, with their avars decreasingly ordered: $f(\hat{s})$ (any CAN estimator of θ smoothly constructed from the intermediate data $\hat{s}$), $\hat{\theta}$ (indirect MLE based on $\hat{s}$) and $\hat{\theta}_{\mathrm{ML}}$ (MLE based on the complete data $\mathbf{W}$). The results in Propositions 2 and 3 have appeared in different forms in the literature. For example, part of the geometry was given by Hausman (1978, Lemma 2.1); result (ii) of Proposition 2 can be recognized as a consequence of the Hájek–Le Cam convolution theorem (Hájek, 1970); result (i) is used in the efficient method of moments (e.g., Gallant and Tauchen 1996, 1999; Gallant and Long, 1997) for choice of auxiliary models to achieve full or approximate efficiency in indirect inference.

Some notation and background knowledge for the propositions are the following. Let the intermediate statistic $\hat{s}$ be defined in a general implicit form $G(\mathbf{W}, \hat{s}) = 0$. Denote the indirect likelihood based on the intermediate data $\hat{s}$ as $L(\theta|\hat{s})$ and denote the likelihood based on the complete data as $L(\theta|\mathbf{W})$, which are maximized by the indirect MLE $\hat{\theta}$ and the MLE $\hat{\theta}_{\mathrm{ML}}$, respectively. We adopt the following notation. Two order $n^{-1/2}$ quantities are said to be asymptotically equal $(\approx)$ when their difference is of a lower order and are said to be orthogonal $(\perp)$ to each other if their covariance elements have a lower order than n^{-1}. All function or derivative values are evaluated at the asymptotic limits θ and/or $s(\theta)$ (for s). For

a generic column vector $\mathbf{v}$, $\mathbf{v}^{\otimes 2}$ denotes $\mathbf{v}\mathbf{v}^T$. Subscripts on F denote partial derivatives, for example, $F_\theta = \{\partial_\theta F(\theta, s)\}|_{s=s(\theta)}$.

PROPOSITION 2. *Efficiency II. Assume that the usual regularity conditions hold so that* $\hat{\theta}$ *and* $\hat{\theta}_{\mathrm{ML}}$ *are both AN.* (*Assume, e.g., conditions in Proposition* 1 *for the AN of* $\hat{\theta}$, *and the conditions in Sen and Singer,* 1993, *Section* 5.2, *for the AN of the MLE* $\hat{\theta}_{\mathrm{ML}}$.) *Denote the score function as* $S = \partial_\theta \log L(\theta|\mathbf{W})$ *and denote the indirect score function as* $T = \partial_\theta \log L(\theta|\hat{s})$. *Then we have the following results*:

(i) *The difference of the "information" matrices satisfies*

$$\begin{aligned}&\mathrm{var}(S) - \mathrm{var}(T)\\ &\quad = \mathrm{var}(\hat{\theta}_{\mathrm{ML}})^{-1} - \mathrm{var}(\hat{\theta})^{-1}\\ &\quad = \inf_{p\times q \text{ matrix } C} \mathrm{var}(S - CG) = \mathrm{var}(S - T).\end{aligned}$$

(ii) *The difference of* avar *matrices satisfies*

$$\begin{aligned}&\mathrm{var}(\hat{\theta}) - \mathrm{var}(\hat{\theta}_{\mathrm{ML}})\\ &\quad = E\{(ETT^T)^{-1}T - (ESS^T)^{-1}S\}^{\otimes 2}.\end{aligned}$$

Therefore, for any direction vector $\mathbf{a}$, $\mathbf{a}^T\hat{\theta}$ *is efficient for estimating* $\mathbf{a}^T\theta$ *iff the standardized score functions for the true likelihood and the indirect likelihood are asymptotically equal at* θ *when projected onto* $\mathbf{a}$.

PROOF. Note that

$$\begin{aligned}\hat{\theta}_{\mathrm{ML}} - \theta &\approx -\{E\,\partial_\theta^2 \log L(\theta|\mathbf{W})\}^{-1}\partial_\theta \log L(\theta|\mathbf{W})\\ &\approx (ESS^T)^{-1}S.\end{aligned}$$

On the other hand, from the linear approximation and the results about the indirect score function in Proposition 1, we have

$$\begin{aligned}\hat{\theta} - \theta &\approx -\{E\,\partial_\theta^2 \log L(\theta|\hat{s})\}^{-1}\partial_\theta \log L(\theta|\hat{s})\\ &\approx E(TT^T)^{-1}T.\end{aligned}$$

These relationships imply that $\mathrm{var}(\hat{\theta}) = \mathrm{var}(T)^{-1}$ and $\mathrm{var}(\hat{\theta}_M) = \mathrm{var}(S)^{-1}$ as used in (i).

(i) We first derive a relationship between the indirect score function T and the estimating function G. By taking the derivative

$$\partial_\theta \log L(\theta|\hat{s}) = \partial_\theta\{-F(\theta, \hat{s})^T(\mathrm{var}\ G)^{-1}F(\theta, \hat{s})/2\}$$

and a linear approximation in $(\hat{s} - s)$, we obtain

$$\begin{aligned}T &\approx -F_\theta^T E(GG^T)^{-1}F_s(\hat{s} - s)\\ &\approx -E(GS^T)^T E(GG^T)^{-1}\{G(\mathbf{W}, \hat{s}) - G(\mathbf{W}, s)\}\\ &= E(GS^T)^T E(GG^T)^{-1}G(\mathbf{W}, s),\end{aligned}$$

noting that $G(\mathbf{W}, \hat{s}) = 0$ and that $E(GS^T) = F_\theta$ (an identity derivable assuming the interchangeability of the derivative and the integration). Then the indirect score T is asymptotically equivalent to the projection of the direct score function (S) onto the span of the estimating function G. Then $(S - T) \perp T$ and it follows that (i) is a direct consequence of Pythagoras' theorem.

(ii) Note that $\hat{\theta}_{\mathrm{ML}} - \theta \approx (ESS^T)^{-1}S$ and $\hat{\theta} - \theta \approx E(TT^T)^{-1}T$. Also note that

$$\{E(TT^T)^{-1}T - (ESS^T)^{-1}S\} \perp (ESS^T)^{-1}S$$

is a consequence of $(S - T) \perp T$. Now (ii) follows from Pythagoras' theorem. □

Result (i) was used by Gallant and Tauchen (1996, 1999) and Gallant and Long (1997) for the choice of the auxiliary model M′ (a "score generator") that generates a naive score function $G(\mathbf{W}, s)$ to which the intermediate statistic $\hat{s}$ is a root, to guarantee full or approximate efficiency in indirect inference. Gallant and Tauchen (1996) showed that $\hat{\theta}$ is fully efficient if the auxiliary model M′ includes the true model M as a submodel by a smooth reparameterization. They claimed high efficiency can be achieved if the auxiliary model can well approximate the true model. They proposed the use of flexible families of auxiliary models such as semi-nonparametric models and neural network models to generate G and $\hat{s}$.

Some geometric relationships are established from the proof of the above proposition. The orthogonality argument in the proof of (ii) essentially says $(\hat{\theta} - \hat{\theta}_{\mathrm{ML}}) \perp (\hat{\theta}_{\mathrm{ML}} - \theta)$. When similar arguments are applied to the situation of comparing $\hat{\theta}$ with any CAN estimate $f(\hat{s})$ smoothly constructed from $\hat{s}$ in Proposition 1(iv), we arrive at the following results that summarize the geometric relationships among $f(\hat{s})$, $\hat{\theta}$ and $\hat{\theta}_{\mathrm{ML}}$, where we assume standard regularity conditions as in Proposition 2.

PROPOSITION 3. *Geometry.* $\hat{\theta}_{\mathrm{ML}} - \theta$, $\hat{\theta} - \hat{\theta}_{\mathrm{ML}}$ *and* $f(\hat{s}) - \hat{\theta}$ *are mutually orthogonal* (*see Figure* 1). *The following Pythagoras-type result holds and summarizes the efficiency results geometrically*:

$$\begin{aligned}&E\{f(\hat{s}) - \theta\}^{\otimes 2}\\ &\quad \approx E\{f(\hat{s}) - \hat{\theta}\}^{\otimes 2} + E(\hat{\theta} - \theta)^{\otimes 2}\\ &\quad \approx E\{f(\hat{s}) - \hat{\theta}\}^{\otimes 2} + E(\hat{\theta} - \hat{\theta}_{\mathrm{ML}})^{\otimes 2}\\ &\qquad + E(\hat{\theta}_{\mathrm{ML}} - \theta)^{\otimes 2}.\end{aligned}$$

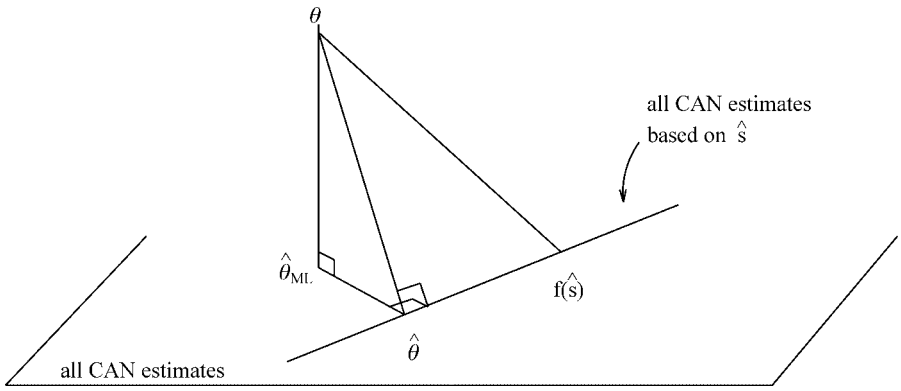


FIG. 1. *Geometry of efficiency results. Note that θ is the true parameter, $\hat{\theta}_{\mathrm{ML}}$ is the MLE, $\hat{\theta}$ is the optimal adjusted estimator based on $\hat{s}$ and $f(\hat{s})$ is any CAN estimator smoothly constructed from the intermediate statistic. The plane represents all CAN estimators constructed from the full dataset; the line across $\hat{\theta}$ and $f(\hat{s})$ represents all CAN estimators constructed from the intermediate statistic $\hat{s}$. The geometry uses the covariance as the matrix of inner products and uses the variance as the matrix of norms, and is accurate up to order $n^{-1/2}$. The closer a point is to θ, the less is the asymptotic variation. The distance from θ to the plane goes to zero as the size of the data increases.*

2.5 Robustness of the Adjusted Estimator

In the indirect approach, with the freedom of choosing what aspect of data information to be incorporated via the intermediate statistic, the inferential results can sometimes be made robust against certain departures from the hypothesized true model M, possibly at the cost of losing some efficiency when the true model is indeed M. The asymptotic properties of inferential procedures based on the indirect likelihood remain valid as long as the asymptotic mean of the intermediate statistic is correctly specified. In comparison, properties of the MLE usually depend on the correct specification of a full probability model. Thus inferences based on indirect likelihood are typically more robust to model misspecification. [This type of robustness has been considered by many authors, e.g., Box and Tiao (1973, Section 3.2), Foutz and Srivastava (1977) and Kent (1982).] It is typical to take robustness into consideration when choosing an intermediate statistic. For example, when one is only willing to assume a mean model for a response, then an intermediate statistic that is linear in the response variable is often used. Further such examples are illustrated in Sections 3 and 4.

The robustness discussed above refers to the consistency of estimators under violations of certain assumptions on the distribution of data. This sense of robustness has been the focus of much recent work in biostatistics. For example, the Poisson process estimation is termed robust by Lawless and Nadeau (1995) because the consistency holds regardless of the assumptions on higher order moments and correlations of the recurrent events. The generalized estimating equations (GEE; Liang and Zeger, 1986) allows consistency regardless of the assumptions on higher order moments or correlation structures of longitudinal data. The marginal method of Wei, Lin and Weissfeld (1989) is a popular method for achieving consistent estimation without modelling the dependence structure for multiple events in survival analysis.

Another sense of robustness refers to estimators that are resistant to outliers or gross errors (e.g., Huber, 1964; Hampel, 1968). Indirect inference procedures can also be made robust against outliers. A sequence of recent articles (Genton and de Luna, 2000; de Luna and Genton, 2001, 2002; Genton and Ronchetti, 2003) investigated the robustness of indirect inference in this sense of protecting against outliers and described many applications.

The key to robustness in the sense of resistance to outliers lies in the influence function (IF) of the estimator. Let $\hat{b}$ be a $\sqrt{n}$-consistent estimator of the parameter b based on n i.i.d. copies of data $\mathbf{W} = (W_1, \ldots, W_n)$. Then the IF is defined such that $\hat{b} - b = n^{-1}\sum_{i=1}^{n} \mathrm{IF}(W_i) + o_p(n^{-1/2})$. One can often compute IF via the Gateaux differential (Hampel, Ronchetti, Rousseeuw and Stahel, 1986, page 84). Note that $\sup_w |\mathrm{IF}(w)|$ shows how much one outlying observation can influence the value of $\hat{b}$. Therefore, the robustness of a consistent $\hat{b}$ against outliers can be characterized by a bounded $\mathrm{IF}(\cdot)$. Note that a bounded IF prevents a large loss of asymptotic efficiency under perturbations of the distributions assumed for W_i's (e.g., gross error), since the asymptotic variance $\mathrm{var}(\hat{b}) = n^{-1}\,\mathrm{var}\{\mathrm{IF}(W_i)\}$ will be bounded if $\mathrm{IF}(\cdot)$ is, whatever distribution W_i actually follows. For more

discussion on the general notion of influence function and robust estimation, see Bickel (1988) and Reid (1988).

Genton and de Luna (2000, Theorem 1) presented the key fact that relates the influence function IF_θ of the indirect estimator $\hat{\theta}$ to the influence function IF_s of the auxiliary estimator $\hat{s}$:

$$\mathrm{IF}_\theta(w) = \{s'(\theta)^T \nu^{-1} s'(\theta)\}^{-1} s'(\theta)^T \nu^{-1} \mathrm{IF}_s(w). \tag{5}$$

This result follows from the relationship $\hat{\theta} - \theta = \{s'(\theta)^T \nu^{-1} s'(\theta)\}^{-1} s'(\theta)^T \nu^{-1} \{\hat{s} - s(\theta)\} + o_p(n^{-1/2})$ derived in the proof of Proposition 1(ii). Therefore, $\hat{\theta}$ will have bounded influence and be resistant to outliers if a robust auxiliary statistic $\hat{s}$, having bounded influence, was used in the first place. (For the generalized method of moments procedure, there are parallel results that relate the influence function and the moment conditions; e.g., see Ronchetti and Trojani, 2001.)

Relationships between various norms of $\mathrm{IF}_\theta(\cdot)$ and $\mathrm{IF}_s(\cdot)$ are then derived from (5). Additional variation due to simulated approximation of $s(\theta)$ are accounted for in Genton and Ronchetti (2003). These ideas were applied in Genton and de Luna (2000), de Luna and Genton (2001, 2002) and Genton and Ronchetti (2003) to a variety of problems including stochastic differential equations models, time series and spatial data.

For one example in Genton and Ronchetti (2003), the assumed model M is the stochastic differential equation (geometric Brownian motion with drift). The auxiliary model M′ is based on a crude Euler discretization. The auxiliary estimators computed as $\hat{s}_{\mathrm{ml}}$, the maximum likelihood estimators under M′, or $\hat{s}_{\mathrm{r}}$, the robust estimators under M′ after using the "Huberized" estimating functions that have bounded influence. Indirect inference based on adjusting these auxiliary estimators then generates (respective) estimators $\hat{\theta}_{\mathrm{ml}}$ and $\hat{\theta}_{\mathrm{r}}$ that are both consistent under M. However, as might be expected, simulation experiments reported by Genton and Ronchetti (2003) showed that, generally in their applications, when there is gross error contamination on the assumed model M, $\hat{\theta}_{\mathrm{ml}}$, obtained from adjusting the naive MLE, behaves poorly, but the estimator $\hat{\theta}_{\mathrm{r}}$, obtained from adjusting a robustified auxiliary estimator, still behaves very well in terms of bias and variability.

2.6 Model Selection

Since the leading order properties of the criterion function $H(\cdot)$ are completely determined by its quadratic approximation around $\hat{\theta}$, which is analytically simpler, in this section we will denote by $H(\theta, \hat{s})$ the quadratic function $H(\hat{\theta}, \hat{s}) + 2^{-1}(\theta - \hat{\theta})^T \partial_\theta^2 H(\hat{\theta}, \hat{s})(\theta - \hat{\theta})$. For model selection, we can continue the process of the analogy and construct a Bayesian information criterion (BIC; Schwarz, 1978) based on the indirect likelihood $L(\theta|\hat{s}) \propto \exp(-H/2)$. Suppose that a submodel M of the original saturated model claims that θ lies in a d_M- ($\leq p$) dimensional submanifold Θ_M of the original parameter space (Θ, say). (Note that $\hat{\theta}$ is the minimizer of H in the original parameter space.) The BIC criterion function $-2\sup_{\theta\in\Theta_M} \log L(\theta|\hat{s}) + d_M \log n$ is, up to a constant of M, equal to the *Bayesian cost*

$$C(M) \equiv \inf_{\theta\in\Theta_M} H(\theta, \hat{s}) + d_M \log n.$$

For a set Φ (called the *scope*) of candidate model M's, the BIC (based on the intermediate statistic $\hat{s}$) chooses $\hat{M} = \arg\min_{M\in\Phi} C(M)$. This choice $\hat{M}$ enjoys the desirable frequentist property of consistency, when a single parameter (θ_0, say) is the true parameter based on which the data are generated. A true model in this case is a model which proposes a parameter space Θ_M that contains the true parameter.

PROPOSITION 4. *Consistency of BIC. Assume the conditions hold for the AN result in Proposition* 1(ii). *Then, with probability tending to* 1 *as the sample size* n *increases,* $\hat{M}$ *chooses a simplest true model* (*with lowest* d_M) *in the search scope* Φ. *If there is no true model in* Φ, *then* $\hat{M}$ *converges in probability to a model in* Φ *that is closest to the true parameter* θ_0, *that is, with smallest distance* $d(\theta_0, \Theta_M) \equiv \inf_{\theta\in\Theta_M}(\theta - \theta_0)^T \nu_\theta^{-1}(\theta - \theta_0)$, *where* $\nu_\theta = p\lim_{n\to\infty}\{n \operatorname{var}(\hat{\theta})\}$.

PROOF. This consistency result is easily proved by noting that $\inf_{\theta\in\Theta_M}[H(\theta, \hat{s}) - H(\hat{\theta}, \hat{s})]$ is positive and of order n when θ_0 is outside Θ_M, and is of order 1 when $\theta_0 \in \Theta_M$. These observations imply that, asymptotically, a true model is favored against a false model; when true models (M's for which $\theta_0 \in \Theta_M$) are compared, the complexity penalty dominates and a simplest model will be chosen. When all models in Φ are false, the behavior of the leading term of $C(M)$ is essentially $n\, d(\theta_0, \Theta_M)$ and the closest false model will be chosen. □

Continuing the Bayesian approach, conditional on the intermediate statistic $\hat{s}$, we define the posterior probability of a model and the Bayes factor (BF) for comparing two models. Suppose under model M θ can be parameterized as $\theta = \theta(\phi_M)$, where ϕ_M

lies in a d_M-dimensional manifold Φ_M. Then we can write $P(\hat{s}|M) = \int_{\Phi_M} P(\hat{s}|\theta(\phi_M))P(\phi_M|M)\,d\phi_M$, where $P(\phi_M|M)$ is a prior for the parameter ϕ_M.

The posterior conditional on $\hat{s}$ is defined as

$$P(M|\hat{s}) = P(\hat{s}|M)P(M)/P(\hat{s}),$$

and the Bayes factor BF_{12} for two models M_1 and M_2 is defined by $\mathrm{BF}_{12} = P(\hat{s}|M_1)/P(\hat{s}|M_2)$.

The following proposition is a straightforward application of the Laplace approximation:

$$-2\log P(\hat{s}|M) = d_M \log(n) - 2\sup_{t_M} \log P(\hat{s}|\theta(t_M)) + O(1)$$

(see, e.g., Draper, 1995, equation 11), and of the normal approximation $-2\log P(\hat{s}|\theta) = H(\theta,\hat{s}) + \log|2\pi\cdot\widehat{\mathrm{var}}(\hat{s})|$ coming from (3).

PROPOSITION 5. *Indirect posterior for a model and the Bayes factor.*

(i) $-2\log P(\hat{s}|M) = C(M) + \log|2\pi\,\hat{\mathrm{var}}(\hat{s})| + O(1)$ *and* $-2\log P(M|\hat{s}) = -2\log P(\hat{s}|M) - 2\log P(M) + 2\log P(\hat{s})$.

(ii) $-2\log \mathrm{BF}_{12} = C(M_1) - C(M_2) + O(1)$ *and if* $-2\log\{P(M_1)/P(M_2)\} = O(1)$, *then* $-2\log\{P(M_1|\hat{s})/P(M_2|\hat{s})\} = -2\log \mathrm{BF}_{12} + O(1)$.

(iii) *Let* $\hat{M} = \arg\min_{M\in\Phi} C(M)$. *Suppose* $-2\log\{P(M_1)/P(M_2)\} = O(1)$ *for all* M_1, M_2 *in* Φ. *Then*

$$\hat{M} = \arg\max_{M\in\Phi} \log Q(M|\hat{s}),$$

where $\log Q(M|\hat{s}) = \log P(M|\hat{s}) + O(1)$.

Roughly speaking, Proposition 5 implies that models with small Bayesian costs tend to have high leading order posterior probability. Together with the previous proposition, this implies that it may be desirable to report the models in the searching scope that have the smallest costs. We propose to report $\hat{M}$, as well as models that have $C(M) \le C(\hat{M}) + 6$, which corresponds roughly to reporting models with leading order posterior probability at least 0.05 times that of $\hat{M}$. We give an application of graphical model selection in Section 4.3.

2.7 Generalized Method of Moments and Indirect Inference

The generalized method of moments is an extremely general method of estimation that encompasses most well-known procedures, such as maximum likelihood, least squares, M estimation, instrumental variables and two-stage least squares (e.g., see Imbens, 2002). It is defined as follows (e.g., Mátyás, 1999, Chapter 1). Suppose the observed data $\mathbf{W}$ consist of n i.i.d. copies $(W_1,\ldots,W_n)$ of W from n units. Suppose also that under our model M, $E_\theta[h(W,\theta)] = 0$ for all θ. Here $h \in \Re^q$ and the q equations $E_\theta[h(W,\theta)] = 0$ are called the *moment conditions*. Define the sample analog $h_n(\theta) = n^{-1}\sum_{i=1}^n h(W_i,\theta)$. The GMM estimator of θ is then defined as

$$\hat{\theta}_{\mathrm{GMM}} = \arg\min_\theta h_n(\theta)^T A_n h_n(\theta), \tag{6}$$

where A_n is a positive definite weight matrix.

In Sections 2.1 and 2.3 we saw that indirect inference (II) was essentially a two-step procedure. In the first auxiliary step we obtained an intermediate statistic $\hat{s}_n$, which can often be defined implicitly from a set of q estimating equations $G(\mathbf{W},s) = 0$. The indirect estimator $\hat{\theta}_{\mathrm{II}}$ is then obtained in the second adjustment step as

$$\hat{\theta}_{\mathrm{II}} = \arg\min_\theta F(\theta,\hat{s}_n)^T v^{-1} F(\theta,\hat{s}_n),$$

where $F(\theta,s) = E_{\mathbf{W}|\theta} G(\mathbf{W},s)$ and v is a sample estimate of the avar of $F(\theta,\hat{s})$. This includes the explicit case when $F(\theta,\hat{s}_n) = \hat{s}_n - s(\theta)$.

In the definition of $\hat{\theta}_{\mathrm{GMM}}$, we may identify $A_n = v^{-1}$ and $h_n(\theta) = F(\theta,\hat{s}_n)$. The moment conditions for this choice are satisfied approximately because $E\{F(\theta,\hat{s}_n)|\theta\} \approx F\{\theta, E(\hat{s}_n|\theta)\} \approx F\{\theta,s(\theta)\} = 0$. These approximate equalities become exact if we interpret the E operator to denote the *asymptotic* mean. Thus the adjustment step of indirect inference can be considered as a GMM procedure where the moment conditions are asymptotically satisfied.

Conversely, it can be argued that GMM is a special example of the complete two-step procedure of indirect inference. Suppose we take the intermediate statistic $\hat{s}_n$ as a GMM estimator $\hat{s}_{\mathrm{GMM}}$ based on some auxiliary model M′. We can then go on to obtain an adjusted estimator $\hat{\theta}_{\mathrm{II}}$ under a true model M as described in Section 2.3. This possibility was suggested by Carrasco and Florens (2002) above their equation (15). The GMM becomes the same as indirect inference when the bridge relationship is trivial, so that $\hat{\theta}_{\mathrm{II}} = \hat{s}_{\mathrm{GMM}}$ even after the adjustment. This will happen if $\hat{s}_{\mathrm{GMM}}$ was obtained from a moment condition $h_n(\theta)$ that is correctly specified even under the true model M, that is, $E\{h_n(\theta)|\theta\} = 0$ under (both M′ and) M.

Although closely connected, the indirect inference approach, with its emphasis on an auxiliary (or intermediate) statistic and an indirect likelihood function,

gives a viewpoint that is somewhat different from the GMM approach. This viewpoint has been productive, leading to contributions in various application areas, especially econometrics.

3. APPLICATIONS OF BRIDGE RELATIONSHIPS

Often the auxiliary statistic $\hat{s}$ is constructed as a naive estimator of θ based on a simplified or naive model M′. The bridge relationship of Section 1.1 can be viewed as an expression for the large-sample limit of this naive estimator in terms of the true parameter. The relationship is then useful for assessing how sensitive or robust a naive analysis is against potential model misspecification. If the bridge relationship is trivial (i.e., $s = \theta$), the naive estimator obtained from M′ remains consistent for θ, even when the true model is M instead of M′. This demonstrates certain robustness (of the *naive* estimator). See examples in Sections 3.1 and 3.3. A number of estimating procedures can be considered in this perspective, which are also classifiable as the pseudo-maximum-likelihood methods in econometrics (Gouriéroux and Monfort, 1993; Broze and Gouriéroux, 1998). Nontrivial bridge relationships (biased naive estimates) reveal the effect of misspecification and are useful for sensitivity analysis and bias correction. See examples in Sections 3.2 and 3.4.

3.1 Poisson Process Estimation for Recurrent Events

For $i = 1, \ldots, n$, suppose $\{W_i(t),\ t \geq 0\}$ are n independent realizations of a point process (not necessarily Poisson) with respective multiplicative intensity functions $f_i(\beta)\lambda(t)$, where $f_i(\beta) = e^{x_i^T \beta}$, say, and x_i denotes a vector of covariates for the ith process. Here the true parameter is $\theta = (\beta, \{\lambda(t)\})$, with $\lambda(t)$ representing the nonparametric baseline intensity. It was shown by Lawless and Nadeau (1995) that naively assuming a model M′ in which the $W_i(t)$ follows a Poisson process but with a correct specification of the intensity function leads to a consistent naive estimator $\hat{s} = (\hat{\beta}, \{\hat{\lambda}(t)\})$ for the true parameter $(\beta, \{\lambda(t)\})$. (The consistency of the naive estimator is characterized by a trivial bridge relationship $s \equiv p\lim_{n\to\infty} \hat{s} = \theta$.) Here $\hat{\beta}$ is the Cox (1972) partial likelihood estimate and $\hat{\lambda}(t)$ is a discrete intensity estimate for $\{\hat{\lambda}(t)\}$ that corresponds to the Nelson–Aalen estimate of the cumulative intensity (see Andersen et al., 1993, Section VII.2.1). Jiang, Turnbull and Clark (1999) gave an application based on an overdispersed Poisson process model, where the overdispersion is caused by frailties (or random effects) that follow a gamma distribution. They showed that the naive estimator $\hat{\beta}$ from $\hat{s}$ not only remains consistent, but can also retain high efficiency relative to the MLE.

3.2 Measurement Error Problems

The main goal is to study the relationship between the response Y and the (true) covariate X, when only an error-contaminated version Z of X is observed. The regression model of interest is the one that relates Y and the true covariate X, which may be described by a conditional distribution $p_{Y|X}(y|x;\theta)$ that involves some unknown parameter(s) θ. It is desired to make inferences concerning θ. A common simplification assumes that Z is a "surrogate" of X in the sense that Y is independent of the surrogate Z when conditioning on the true X.

Let $(Y_i, X_i, Z_i), i = 1, \ldots, n$, be i.i.d. copies of (Y, X, Z), where X_i's are unobserved. The observed data consist of pairs $W_i = (Y_i, Z_i), i = 1, \ldots, n$.

If we denote $p_{Y|X,Z}$, $p_{X|Z}$ and p_Z as the probability density functions (pdfs) of $(Y_i|X_i, Z_i)$, $(X_i|Z_i)$ and Z_i, respectively, we have that the true likelihood based on the observed data $\{(Y_i, Z_i)\}$ is $\prod_{i=1}^n (\int p_{Y_i|x,Z_i} p_{x|Z_i} \times p_{Z_i}\, dx)$, which involves integration over unobserved X_i values. The maximization of the likelihood can be difficult computationally and there is unlikely to be any standard software available to be of aid. On the other hand, if we adopt a model M′ that simply ignores the covariate measurement error and treats Z_i as X_i for each i, we are led to a naive regression analysis for which standard software will very likely be available. A naive estimator $\hat{s}$ then is simply constructed by neglecting the measurement errors in Z_i, and maximizing the naive likelihood $\prod_{i=1}^n p_{Y|X}(Y_i|Z_i;s)$. The general method of Section 2 is to try to find a large sample limit $\hat{s} \to s(\theta)$ and then obtain the adjusted estimator $\hat{\theta}$ by solving $\hat{s} = s(\theta)$ for θ.

For a simple example, consider the case when the conditional distribution of Y_i given X_i is $N(\theta X_i, \sigma_\varepsilon^2)$, that is, simple linear regression through the origin with homoscedastic normal errors. A structural model of normal additive measurement error structure is assumed, that is, $Z_i = X_i + U_i$, where X_i and U_i are independent normal with variances σ_X^2 and σ_U^2, respectively. Then the naive MLE or naive LS estimator is $\hat{s} = \sum Y_i Z_i / \sum Z_i^2$, and $\hat{s} \to s(\theta)$ almost surely, where

$$s(\theta) = \frac{EY_iZ_i}{EZ_i^2} = \frac{EX_iZ_i}{EZ_i^2}\theta = \frac{EX_i^2}{EZ_i^2}\theta = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\theta.$$

Note that $|s| < |\theta|$, which is called the attenuation phenomenon: the magnitude of the naive slope estimate $|\hat{s}|$ underestimates $|\theta|$. This is a common feature when measurement error is ignored in analyzing regression models (Fuller, 1987, page 3). By solving $\hat{s} = s(\theta)$, a consistent adjusted estimator is easily found to be $\hat{\theta} = ((\sigma_X^2 + \sigma_U^2)/\sigma_X^2)\hat{s}$. Of course, this adjustment assumes that the measurement error parameters σ_X^2 and σ_U^2 are known. In practice, they will not be, and σ_X and σ_U should be considered as part of the parameter vector θ. We are in the situation discussed at the end of Section 2.3, where $(\dim\theta) > \dim(s)$. However, σ_X and σ_U can sometimes be estimated from a second or "validation" dataset in which pairs $(X_k, Z_k), k = 1, \ldots, m$, can be observed directly (Carroll, Ruppert and Stefanski, 1995, page 12). These estimates can then be plugged into the formula for $\hat{\theta}$. The uncertainty resulting from the fact that the measurement error parameters are not known but estimated can be incorporated into an estimate of $\operatorname{var}\hat{\theta}$ by the method of propagation of errors [see Taylor, 1997, (3.4), and Jiang, Turnbull and Clark, 1999, Appendix B]. Alternatively, instead of using a validation study, $\sigma_Z^2 = \sigma_X^2 + \sigma_U^2$ can be estimated from the sample variance of the observed Z values and σ_U^2 can be treated as a tuning parameter for a sensitivity analysis.

In the presence of covariate measurement error, similar explicit formulae that relate naive regression parameters and the true parameters were established by Jiang (1996) for Poisson, exponential and logistic regression models, by Turnbull, Jiang and Clark (1997) for negative binomial regression models and by Jiang, Turnbull and Clark (1999) for semiparametric Poisson process regression models. In these articles it was assumed that the distribution of X_i conditional on Z_i follows a normal linear model. In the following discussion, we introduce a method which does not require parametric assumptions on the distribution of (X_i, Z_i). In addition, only the first moment is specified for the parametric model of Y_i given X_i. This provides an example where the bias correction is robust in the sense that the consistency of the adjusted estimator depends on the correct specification of the mean function $E(Y_i|X_i)$ instead of a complete probability model. We also generalize the notion of a naive covariate Z_i to be a general surrogate of X_i. The dimensions of Z_i and X_i can differ. It is only assumed that $E(Y_i|X_i, Z_i) = E(Y_i|X_i)$, which corresponds to the assumption of nondifferential measurement error (Carroll, Ruppert and Stefanski, 1995, page 16).

Let Y, X and Z be three random vectors of dimensions d_y, d_x and d_z, respectively. Assume a nondifferential mean model $E(Y|X, Z) = \mu(X, \theta)$, where θ is a $p \times 1$ parameter. Suppose we observe a main dataset $\mathbf{W} = (Y_i, Z_i)_{i=1}^n$, that is, an i.i.d. realization of (Y, Z), as well as an independent validation dataset $\mathbf{V} = (X_j, Z_j)_{j=1}^m$, that is, an i.i.d. realization of (X, Z). The problem is to perform valid inference on θ based on the observed datasets.

Suppose we start with a naive $q \times 1$ estimator $\hat{s}$ $(q \geq p)$, which solves a $q \times 1$ linear estimating equation of the form $G(\mathbf{W}, s) = n^{-1}\sum_1^n h(Z_i, s)\{Y_i - m(Z_i, s)\} = 0$, where $h_{(q\times d_y)}$ and $m_{(d_y\times 1)}$ are fixed smooth functions. Typically, $\hat{s}$ is AN but not consistent for θ. We could then use the methods from Section 2 to adjust $\hat{s}$ to obtain a consistent estimator $\hat{\theta}$, for example, by maximizing the indirect likelihood $L(\theta|\hat{s})$ in the implicit form, or when $\dim(\hat{s}) = \dim(\theta)$, by solving $F(\theta, \hat{s}) = 0$, where $F(\theta, s)$ is the expectation of the estimating function G.

Here, the function $F(\theta, s) = E_{\mathbf{W}|\theta}G(\mathbf{W}, s)$ can be computed by noting that $E_{\mathbf{W}|\theta}G(\mathbf{W}, s) = E_{X,Z}$ $h(Z, s)\{\mu(X, \theta) - m(Z, s)\}$ by first taking the conditional mean given X, Z and using the nondifferential assumption. Then the expectation $E_{X,Z}$ can be approximated by the sample average based on the validation data $\mathbf{V}$. Consequently, F is estimated by $F^*(\theta, s) = m^{-1}\sum_1^m f(V_j; \theta, s)$, where $f(V_j; \theta, s) = h(Z_j, s)\{\mu(X_j, \theta) - m(Z_j, s)\}$ and $V_j = (X_j^T, Z_j^T)^T$. Using F^* to approximate F inflates the avar of the final estimator $\theta^*(\hat{s})$. Jiang and Turnbull (2003) showed that the avar can be estimated, in the limit of proportionally large n and m, based on a sample estimate of

$$\operatorname{var}\theta^*(\hat{s}) = (F_\theta)^{-1}\big(m^{-1}Eff^T + n^{-1}Egg^T\big)(F_\theta)^{-T}|_{s=s(\theta)}, \tag{7}$$

where $f = f(V_k; \theta, s)$ and $g = g(W_i, s) = h(Z_i, s) \times \{Y_i - m(Z_i, s)\}$. In Section 4.2 we will use an epidemiological dataset to illustrate the methodology described here.

3.3 Omitted Covariates

Gail, Wieand and Piantadosi (1984) considered the effect of omitting covariates in randomized clinical trials. Their method can be put into the formalism of establishing bridge relationships. Consider a special example where $\mathbf{W} = (W_1, \ldots, W_n)$ are i.i.d., $W_i = (Y_i, Z_i, O_i)$, and Y_i is the response following $Y_i|Z_i$, $O_i \sim \text{Poisson}(e^{\alpha + Z_i\beta + O_i\gamma})$ under model M. Here Z_i is a treatment assignment variable that takes value 0 or 1

with equal probability and is assumed to be independent of O_i, another covariate. The true parameter is $\theta = (\alpha, \beta, \gamma)^T$, and β is the regression coefficient for the treatment effect, which is of primary interest. Now consider a naive or simplified regression model M′, where the presence of the covariate O_i is ignored, that is, it is assumed that $Y_i|Z_i \sim \text{Poisson}(e^{a+Z_i b})$. The (naive) parameter in this model is $s = (a, b)^T$. Note that this is again a situation where there are fewer naive parameters than true parameters. The naive estimator $\hat{s} = (\hat{a}, \hat{b})^T$ maximizes the naive likelihood $\prod_{i=1}^n (e^{a+Z_i b})^{Y_i} \exp\{-e^{a+Z_i b}\}/Y_i!$, which neglects the covariate O_i. Therefore, $\hat{s}$ satisfies the naive score equation

$$G(\mathbf{W}, s) = n^{-1} \sum_{i=1}^{n} (1, Z_i)^T (Y_i - e^{a+Z_i b}) = 0$$

and its large sample limit $s = s(\theta)$ satisfies $EG(W, s) = 0$ or $E(1, Z_i)^T (Y_i - e^{a+Z_i b}) = 0$. Using

$$E(Y_i|Z_i) = E(e^{\alpha+Z_i\beta+O_i\gamma}|Z_i) = e^{\alpha+Z_i\beta}(Ee^{O_i\gamma}),$$

we obtain

$$E(1, Z_i)^T \big(\exp\big((\alpha + \log Ee^{O_i\gamma}) + Z_i\beta\big) - e^{a+Z_i b}\big) = 0.$$

Hence

$$a = \alpha + \log Ee^{O_i\gamma} \quad \text{and} \quad b = \beta,$$

establishing the bridge relationship $s = s(\theta)$ between $\theta = (\alpha, \beta, \gamma)^T$ and $s = (a, b)^T$. In this situation, neglecting the covariate O_i still leaves the treatment effect estimator $\hat{b}$ from $\hat{s} = (\hat{a}, \hat{b})^T$ consistent, since $b = \beta$. In a similar manner, Gail, Wieand and Piantadosi (1984) considered other regression models, for example, logistic and exponential regression models with various link functions, and presented a list of results on how the treatment effect estimator behaves in randomized clinical trials when covariates are omitted.

3.4 Missing Data

Rotnitzky and Wypij (1994) considered the bias of estimating equation methods (MLE and GEE) with missing data, when all available cases are used and the missing data mechanism is ignored, in the situation when the data may not be missing at random (Heckman, 1976; Little, 1994). The bias is obtained from examining the limit of the estimating equation and its solution—similar to finding the bridge relationship $s = s(\theta)$ from $F(\theta, s) = E_{\mathbf{W}|\theta} G(\mathbf{W}, s) = 0$ in Section 2.2.

Jiang (1996) considered the bridge relationship for finding the effect of neglecting incomplete cases in analysis of multivariate normal data. Assume that the complete data consist of $r \times 1$ random vectors Y_i, $i = 1, \ldots, n$, which are i.i.d. Associated with each subject there is a binary indicator M_i which takes value 1 if and only if all components of Y_i are observed. Denote Y_j^c, $j = 1, \ldots, n^c$, as the subsample where the M_j's are 1. A naive likelihood analysis is based on the complete cases and the multivariate normal assumption $Y_j^c \overset{\text{i.i.d.}}{\sim} N(m, S)$, where the naive parameter s contains all components of m and S. Therefore, we take as our intermediate statistic

$$\hat{s} = \arg\max_s \prod_{j=1}^{n^c} \left\{ \frac{1}{\sqrt{\det(2\pi S)}} \cdot \exp\left(-\frac{1}{2} (Y_j^c - m)^T S^{-1} (Y_j^c - m) \right) \right\}.$$

In fact, $\hat{s}$ estimates $s = (m, S)$, where $m = EY_j^c = E(Y_i|M_i = 1)$ and $S = \text{var}\, Y_j^c = \text{var}(Y_i|M_i = 1)$, which may be calculated according to different models of the missing mechanism. In a normal selection model (see Little, 1994), for example, $Y_i \sim N(\mu, \Sigma)$ and $M_i|Y_i$ follows a probit regression model $P(M_i = 1|Y_i) = \Phi(\alpha + \beta^T Y_i)$, where Φ is the cumulative distribution function (cdf) of the standard normal distribution. For this model, the pdf of $Y_i|M_i = 1$ is

$$P_{Y|M=1}(x) = \frac{\Phi(\alpha + \beta^T x)\phi_{\mu,\Sigma}(x)}{\int \Phi(\alpha + \beta^T y)\phi_{\mu,\Sigma}(y)\, dy} = \frac{\Phi(\alpha_0 + \beta^T (x - \mu))\phi_{0,\Sigma}(x - \mu)}{\int \Phi(\alpha_0 + \beta^T \eta)\phi_{0,\Sigma}(\eta)\, d\eta},$$

where $\phi_{\mu,\Sigma}$ is the probability density for the multivariate normal random variable with mean μ and variance Σ, and $\alpha_0 = \alpha + \beta^T \mu$. Note that when $\beta = 0$, $P_{Y|M=1} = \phi_{\mu,\Sigma} = P_Y$, which leads to the missing completely at random (MCAR) model. In that case, the bridge relationships are trivial, $m = \mu$ and $S = \Sigma$, implying that ignoring incomplete cases leads to consistent naive MLEs. This suggests that, for small β, we can perform a Taylor expansion when evaluating $E(Y_i|M_i = 1)$ and $\text{var}(Y_i|M_i = 1)$. Upon neglecting terms of $o(\beta^2)$ [or $o(\beta^T \Sigma \beta)$], this leads to approximate bridge relationships

$$\begin{aligned} m &= E(Y_i|M_i = 1) = \mu + \Phi(\alpha_0)^{-1} \phi(\alpha_0) \Sigma \beta, \\ S &= \text{var}(Y_i|M_i = 1) \\ &= \Sigma - \alpha_0 \Phi(\alpha_0)^{-1} \phi(\alpha_0) (\Sigma\beta)(\Sigma\beta)^T, \end{aligned} \tag{8}$$

where ϕ is the standard normal pdf. In fact, an exact formula for m is available, namely

$$m = \mu + (\Phi(\xi))^{-1}\phi(\xi)(1 + \beta^T\Sigma\beta)^{-1/2}\Sigma\beta,$$

where $\xi \equiv (1 + \beta^T\Sigma\beta)^{-1/2}(\alpha + \beta^T\mu)$; see Jiang (1996, equation 4.59).

We note that, in general, the bias of the naive mean estimator is determined by the sign of $\Sigma\beta$, and the naive variance estimator is typically biased downward, that is, $[S]_{kk} < [\Sigma]_{kk}$ for each k, $1 \le k \le r$, provided $\alpha_0 = \alpha + \beta^T\mu > 0$ (meaning that a majority of the cases are complete). The bridge relationships in (8) can be used to reduce the bias caused by the naive analysis that neglects incomplete cases, provided that the missing data parameter (α, β^T) can be estimated, perhaps from other studies, where missing data are tracked down with additional effort.

Alternatively, if such a dataset does not exist, but the missing at random (MAR) assumption (see Little, 1994, equation 9, page 473) is reasonable, we could estimate (α, β) from the original dataset. There we assume that the missingness M_i is only dependent on the complete components of Y_i which are observed for all subjects. For example, in the bivariate normal incomplete data situation, suppose $Y_i = (Y_{i(1)}, Y_{i(2)})^T$ and the first component, $Y_{i(1)}$, say, is always observed, but the second component $Y_{i(2)}$ is sometimes missing, when $M_i = 0$. In the MAR model we write $\beta = (\beta_{(1)}, \beta_{(2)})^T$ and may assume $\beta_{(2)} = 0$. Hence $(\alpha, \beta_{(1)}^T)$ can be obtained by performing a probit regression of the M_i's on the $Y_{i(1)}$'s, $i = 1, \ldots, n$, which are all available in the original dataset. Of course the uncertainty in estimating $(\alpha, \beta_{(1)}^T)$ must be incorporated in the asymptotic variance of the adjusted estimates for (μ, Σ). This can be done by a sensitivity analysis or, alternatively, by use of the propagation of errors method [Taylor, 1997, (3.4); Jiang, Turnbull and Clark, 1999]. Here we are more interested in assessing the effect of dropping incomplete cases in the complete case naive analysis. Notice that the MAR assumption does not ensure that the complete case analysis will give a consistent answer for estimating μ, since $\Sigma\beta$ is not necessarily zero even if $\beta_{(2)}$ is assumed to be zero.

4. THREE DATASETS

We illustrate the ideas of indirect inference procedures with analyses of three datasets. The first two use estimates from a naive model M′ as intermediate statistics as in the examples of Section 3. The third concerns model selection and uses sample moments.

4.1 Poisson Regression with Overdispersion: Animal Carcinogenicity Data

We use carcinogenicity data presented by Gail, Santner and Brown (1980) from an experiment conducted by Thompson, Grubbs, Moon and Sporn (1978) to illustrate our method for treating a Poisson regression model with random effects (overdispersion). Forty eight female rats that remained tumor-free after 60 days of pretreatment of a prevention drug (retinyl acetate) were randomized with equal probability into two groups. In Group 1 they continued to receive treatment ($Z = 1$); in Group 2 they received a placebo ($Z = 0$). All rats were followed for an additional 122 days and palpated for mammary tumors twice a week. The objective of the study was to estimate the effect of the preventive treatment (Z) on number of tumors (Y) diagnosed.

In the model, given Z and ε, Y is assumed to be Poisson with mean $e^{\alpha+Z\beta+\varepsilon}$. Here Z is observed but ε represents an unobserved random effect assumed normal with zero mean and constant variance σ^2, independent of Z. This unobserved random effect or unexplained heterogeneity could be caused by omitted covariates. We observe n i.i.d. pairs of $W_i = (Y_i, Z_i)$, $i = 1, \ldots, n$. The likelihood for the observed data involves integration over ε and is difficult to compute. We start with an auxiliary statistic $\hat{s} = (\hat{a}, \hat{b}, \hat{t}^2)^T$, where $(\hat{a}, \hat{b})$ are the regression coefficient estimates that maximize a naive log-likelihood $R = \sum_1^n \{Y_i(a + Z_i b) - \exp(a + Z_i b)\}$, and $\hat{t}^2 = n^{-1}\sum_{i=1}^n Y_i^2$ is the sample second moment. Here the naive parameter is $s = (a, b, t^2)$ and the true parameter is $\theta = (\alpha, \beta, \sigma^2)$. The use of the naive log-likelihood R corresponds to estimating the regression coefficients by neglecting the random effect ε. The second sample moment is included in the intermediate statistic to provide information for estimation of the variance parameter. Therefore, $\hat{s}$ is solved from the estimating equation $G(\mathbf{W}, s) = 0$, where (formally) $G = (n^{-1}\partial_a R, n^{-1}\partial_b R, \hat{t}^2 - t^2)^T$. The solution can be computed easily. For the rat carcinogenicity data, we obtain the naive estimates $\hat{a} = 1.7984$, $\hat{b} = -0.8230$ and $\hat{t}^2 = 31.875$. To obtain the adjusted estimates $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$, we must derive the bridge equation which comes from the large sample limit of $\hat{s} = (\hat{a}, \hat{b}, \hat{t}^2)$. Here, this limit is the solution of $F(\theta, s) = E_{\mathbf{W}|\theta}G(\mathbf{W}, s) = 0$, which can be explicitly solved to obtain $s = s(\theta)$. This yields bridge equations $a = \alpha + \sigma^2/2$, $b = \beta$ and $t^2 = \frac{1}{2}(1 + e^{\beta})e^{\alpha+\sigma^2/2} + \frac{1}{2}(1+e^{2\beta})e^{2(\alpha+\sigma^2)}$. These equations are inverted to obtain the adjusted estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$. Thus $\hat{\beta} = \hat{b}$

and $\hat{\alpha} = \hat{a} - \hat{\sigma}^2/2$, where $\hat{\sigma}^2 = \log\{(2\hat{t}^2 - e^{\hat{a}}(1 + e^{\hat{b}}))/(e^{2\hat{a}}(1 + e^{2\hat{b}}))\}$. For the rat data, this leads to adjusted estimates $\hat{\alpha} = 1.6808$ (0.1589), $\hat{\beta} = -0.8230$ (0.1968) and $\hat{\sigma} = 0.4850$ (0.1274). The estimated standard errors shown in parentheses are obtained from the sandwich formula (4) and the delta method.

Alternatively, the MLE of $\theta = (\alpha, \beta, \sigma^2)$ can be found by a somewhat tedious iterative numerical maximization of the true likelihood which involves numerical integration over the distribution of ε. These estimates are $\hat{\alpha}_{\mathrm{ML}} = 1.6717$ (0.1560), $\hat{\beta}_{\mathrm{ML}} = -0.8125$ (0.2078) and $\hat{\sigma}_{\mathrm{ML}} = 0.5034$ (0.0859). For the MLEs, the estimated standard errors are based on the inverse of the Fisher information matrix, evaluated at the corresponding estimate values.

The estimated standard errors suggest that the efficiency of the estimation of the treatment effect parameter β is high here in this example. Related results (Cox, 1983; Jiang, Turnbull and Clark, 1999) show that such high efficiency is achievable if the overdispersion is small or if the followup times are about the same across different subjects. Also it should be noted that the adjusted estimator $\hat{\beta}$ is robust in the sense that it remains consistent essentially as long as the mean function $E(Y|Z, \varepsilon)$ is correctly specified and ε and Z are independent. (Its standard error estimate from the sandwich formula is also model-independent and robust.) In particular, the consistency property does not depend on the specification of a complete probability model, namely that Y is Poisson and ε is normal.

Our approach, although formulated from the different perspective of using the naive model plus the method of moments, is intimately related to the work of Breslow (1990) based on quasi-likelihood and the method of moments. Breslow used a different linear combination of Y_i's based on quasi-likelihood (Wedderburn, 1974; McCullagh and Nelder, 1989) that enjoys general efficiency properties among linear estimating equations. However, (i) our approach can be interpreted as basing inference on the simple moments $\sum Y_i$, $\sum z_i Y_i$ and $\sum Y_i^2$ (which can be easily seen from writing out the naive score equations) and (ii) our approach shows clearly, by the use of bridge relationships, the sensitivity and robustness of parameter estimates to the omission of overdispersion in modelling. Also note that here we used a log-normal distribution to model the random effects and the variance parameter also enters the mean model (unconditional on ε), whereas Breslow (1990) focused on examples such as those with gamma multiplicative random effects in which the mean model does not change. For the only comparable parameter β (the treatment effect), the Breslow method [from his equations (1), (2) and (7)] gives exactly the same answer as our adjusted analysis: $\hat{\beta}_{\mathrm{Breslow}} = -0.8230$ (0.1968). This is because, for this special two-group design, both methods essentially use the log(frequency ratio) to estimate the treatment effect.

4.2 Logistic Regression with Measurement Error: Indoor Air Pollution Data

We consider data from Florey et al. (1979) on the prevalence of respiratory illness in relation to nitrogen dioxide (NO_2) exposure among primary school children in Middlesborough, England. Whittemore and Keller (1988) analyzed this dataset using a logistic regression where the NO_2 exposure variable is considered to be a covariate that is subject to measurement error. They used estimates based on modifying the estimates that result from a naive logistic regression model. Our method differs from theirs in that (i) it does not involve a small measurement error approximation, (ii) no parametric assumption is made concerning the measurement error distribution and (iii) adjustment is made for the effect of measurement errors both from the imperfection of the measurement method and from the incomplete knowledge of (grouped) measured data.

The study population consists of 103 primary school children and each child was classified into one of three exposure categories of the nitrogen dioxide (NO_2) concentration in the child's bedroom, which is a surrogate for personal exposure to NO_2. The response variable Y is 1 if a child has prevalent respiratory disease and 0 otherwise. A logistic regression model is assumed in which $\log\{EY/(1 - EY)\} = \alpha + \beta X$, where X is the personal exposure to NO_2. An imperfect measurement method for X is to use $\tilde{Z}$, the bedroom level of NO_2, as a surrogate of the personal exposure. However, the values of $\tilde{Z}$ reported by Florey et al. (1979) were only in three categories, namely less than 20 parts per billion (ppb), between 20 and 39 ppb, and exceeding 40 ppb. Since the individual levels are not published, Whittemore and Keller (1988, Section 6) used a further surrogate Z of $\tilde{Z}$ to perform the logistic regression analysis, where they coded $Z = 10$ if $\tilde{Z} <$ 20 ppb, $Z = 30$ if $\tilde{Z} \in [20, 40)$ ppb, and $Z = 60$ if $\tilde{Z} \geq$ 40 ppb. Table 1 is a recasting of Table 1 of Whittemore and Keller (1988) which summarizes the data.

Estimates and standard errors for the parameters α and β based on naive logistic regression analysis of Y on Z are displayed in the first row of Table 2 and agree with those of line 1 in Table 3 in Whittemore and Keller

TABLE 1
Number of children with or without respiratory disease by bedroom NO_2 *levels*

	$Z = 10$	$Z = 30$	$Z = 60$	Total
Cases ($Y = 1$)	21	20	15	56
Controls ($Y = 0$)	27	14	6	47
Total	48	34	21	103

NOTES. From Whittemore and Keller (1988). $Y = 1$ indicates the existence of respiratory illness and $Y = 0$ otherwise; $Z = 10$ if bedroom NO_2 exposure is under 20 ppb; $Z = 30$ if NO_2 exposure is between 20 and 39 ppb; $Z = 60$ if NO_2 exposure is 40 ppb or more.

(1988). However, two problems exist. First, bedroom level ($\tilde{Z}$) of NO_2 is only a surrogate for personal exposure (X), due to limitation of the measurement method. Second, the variable Z used in the analysis is only a coded version of bedroom exposure $\tilde{Z}$ caused by the grouping of this variable.

We proceed in a manner analogous to that outlined in Section 3.2. The dataset $\mathbf{W}$ consists of $n = 103$ i.i.d. pairs $\{(Y_i, Z_i)\}$, $1 \le i \le n$. The naive estimator $\hat{s} = (\hat{a}, \hat{b})$ is obtained from the logistic regression of the Y_i's on the Z_i's, maximizing the naive likelihood $\prod_{i=1}^n p_i^{Y_i}(1 - p_i)^{1-Y_i}$ in which the true covariate X_i is replaced by the surrogate Z_i. Thus the naive estimator $\hat{s} = (\hat{a}, \hat{b})^T$ satisfies the naive score equation $G \equiv n^{-1}\sum_1^n (1, Z_i)^T(Y_i - p_i) = 0$, where $p_i = \mathcal{H}(a + bZ_i)$, and $\mathcal{H}(u) = \exp(u)/[1 + \exp(u)]$. Its large-sample limit $s = (a, b)$ satisfies the limit of the naive score equation $F(\theta, s) = E_{\mathbf{W}|\theta}G = 0$ or $E[(1, Z)^T\{Y - \mathcal{H}(a + bZ)\}] = 0$. Note that Y is assumed to satisfy a logistic regression model on X (personal NO_2 exposure) instead of on Z, that is, $E(Y|X) = \mathcal{H}(\alpha + \beta X)$. We also assume that Z is a nondifferential surrogate of X (see Section 3.2), so that $E(Y|X, Z) = E(Y|X)$. Then we obtain

$$F(\theta, s) = E[(1, Z)^T\{\mathcal{H}(\alpha + \beta X) - \mathcal{H}(a + bZ)\}] = 0. \tag{9}$$

This obviously is a special example of the situation discussed at the end of Section 3.2, with the mean functions $\mu(\cdot)$ and $m(\cdot)$ both being logit-linear, and the naive estimator $\hat{s}$ having the same dimension as that of the true parameter $\theta = (\alpha, \beta)$. [Alternatively, we may regard the true parameter as also including the joint distribution (X, Z), which will be approximated in some sense by use of a validation dataset.]

The development in Section 3.2 suggests we approximate F in (9) by F^*, where the expectation on X and Z is approximated by a sample average based on a validation dataset. We will consider a validation study (Leaderer, Zagraniski, Berwick and Stolwijk, 1986) also considered by Whittemore and Keller (1988). Leaderer et al. (1986) discussed a dataset relating personal NO_2 exposure (X) to bedroom NO_2 concentration ($\tilde{Z}$) for 23 adults in New Haven, Connecticut. As in Whittemore and Keller (1988), we assumed the validation data are applicable to the English school children. In Leaderer et al. (1986), the data of X versus $\tilde{Z}$ were not published at the individual level, but their Figure 7 displays a scatter plot of X versus house average NO_2 level for the 23 subjects. To illustrate our method, we simulated two validation datasets of sizes $m = 23$ and 230 as follows. First, we simulated a dataset of 23 independent $(X, \tilde{Z})$'s to have the same distribution shape as Figure 7 in Leaderer et al. (1986). [We rescaled their data in Figure 7 to satisfy

TABLE 2
Logistic regression coefficients for respiratory illness versus personal NO_2 *exposure*

	α (standard error)	Z value	β (standard error)	Z value
Naive	−0.4536 (0.3490)	−1.299	0.0240 (0.0112)	2.138
$WK_{m=23}$	−0.5563 (0.3691)	−1.507	0.0296 (0.0125)	2.368
$RSW_{m=23}$	NA[a]	NA[a]	0.0264 (0.0133)	1.983
$Adjusted_{m=23}$	−0.5659 (0.4472)	−1.265	0.0304 (0.0188)	1.617
$RSW_{m=230}$	NA[a]	NA[a]	0.0270 (0.0127)	2.124
$Adjusted_{m=230}$	−0.6383 (0.4758)	−1.342	0.0314 (0.0186)	1.688

NOTE. The row labeled Naive gives the results obtained in a logistic regression using Z as the predictor and neglecting the presence of measurement error. The row labeled WK contains the results obtained by the modified method of Whittemore and Keller (1988). The rows labeled RSW contain the results obtained by the method of Rosner, Spiegelman and Willett (1990). The rows labeled Adjusted were obtained using the method described here.

[a]RSW did not provide a method for adjusting the intercept estimate. However, in case-control studies, as here, the intercept parameter is not of particular relevance.

the published regression fit $X = 4.48 + 0.76\tilde{Z}$ and $\mathrm{Var}(X|\tilde{Z}) = 81.14$.] From this simulated dataset, we grouped and coded the $\tilde{Z}$ values to obtain $m = 23$ pairs $(X_k, Z_k), k = 1, \ldots, 23$, which form the first validation dataset. Then a second (larger) validation dataset ($m = 230$) was obtained by sampling the first validation dataset with replacement.

Following Section 3.2, we approximate F in (9) by F^* constructed from the validation sample (X_k, Z_k), $k = 1, \ldots, m$, with $m = 23$ or 230, that is,

$$F^*(\theta, s) = m^{-1} \sum_{k=1}^{m} (1, Z_k)^T \{\mathcal{H}(\alpha + \beta X_k) - \mathcal{H}(a + bZ_k)\}.$$

Using the naive MLE $\hat{s} = (\hat{a}, \hat{b})$ (from line 1 of Table 2), consistent adjusted estimates $\theta^*(\hat{s}) = (\alpha^*, \beta^*)$ are obtained by solving $F^*(\theta, \hat{s}) = 0$; their values are listed in the fourth and sixth rows of Table 2. The standard errors (in parentheses) incorporate the sampling error from the validation data through use of (7), where $V_k = (X_k, Z_k)$ and $f(V_k; \theta, s) = (1, Z_k)^T \{\mathcal{H}(\alpha + \beta X_k) - \mathcal{H}(a + bZ_k)\}$ for $k = 1, \ldots, m$.

For comparison, we have included results from some alternative methods for treating covariate measurement error in logistic regression. In the second row of Table 2, we have included the parameter estimates that result from the approximation method of Whittemore and Keller (1988) (WK), which were listed in Table 3 of their article. In the third and fifth rows, we list the results from applying the method of Rosner, Spiegelman and Willett (1990) (RSW) based on regression calibration (Carroll, Ruppert and Stefanski, 1995, Chapter 3). Here a standard analysis was performed, but regression of X on Z was used in place of Z, the regression being based on estimates from the validation datasets. The method of RSW (1990) also provides a first-order correction to the bias, which is valid if the disease probability is small (RSW, 1990, Appendix 1) or if the effect of measurement error is small (see Carroll, Ruppert and Stefanski, 1995, page 65), which requires $\beta^T \mathrm{var}(X|Z)\beta$ to be small.

Our approach gives point estimates similar to those from Whittemore and Keller (1988), but our standard errors (s.e.s) are larger. Note, however, that the Whittemore and Keller (1988) results in the second row were obtained by treating Z (the coded values) as the true bedroom NO_2 level $\tilde{Z}$, and the s.e.'s were obtained by neglecting the sampling variation from the validation data. Our results are more comparable to those in the RSW rows, where variation from the validation data was incorporated using the delta method (RSW, 1990) and the coded values of Z were used both in the naive logistic analysis of the main data and in the linear regression of the validation data.

Our estimates of the slope β are larger than those obtained from the RSW method, showing that a correction based on regression calibration is not enough, probably due to a nonlinear bridge relationship between b and β implied by (9). In the special case when the distribution of X given Z is modelled as a normal linear regression in Z, this nonlinearity feature can be seen in the approximation formula (3.24) of Carroll, Ruppert and Stefanski (1995); see also Figure 4.1 in Jiang (1996). However, our Z values are lower than those obtained from the RSW method, due to an inflation of variance which more than compensates for the inflation of the parameter estimate. This is probably not related to the extra variation from the validation data, since in our approach as well as in the RSW approach, the s.e.s change little (less than 10%) when the variation from the validation data is neglected, for example, by removing the first summand in our (7) or the second summand of (A4) in RSW (1990). The nonproportional increase in s.e. is more likely due to the nonlinearity of the bridge relationship between b and β (see Jiang, 1996, equation 4.35).

Comparing the results derived from the two validation datasets, we see that the results are very similar despite the tenfold increase in m. This is not surprising, since (i) from the previous paragraph, we see that the changes in s.e. can be small even if we take m to be ∞, and (ii) this insensitivity probably is due to the small size of $\hat{b}_n$ (0.024). Point (ii) is easiest to understand by looking at $\mathrm{avar}(\beta^*_{nm}) = \mathrm{avar}(\hat{\lambda}_m^{-1}\hat{b}_n)$, the avar of the adjusted estimator using the antiattenuation formula (i.e., the RSW approach). It is apparent from the delta method that if $\hat{b}_n$ is very small, the precision of $\hat{\lambda}_m$ (or the validation sample size m) is not very relevant to $\mathrm{avar}(\beta^*_{nm})$.

In summary, our proposed adjustment method does not require modelling the validation data, in contrast to the WK and RSW methods, which both make use of a linear regression of X given Z. Second, the validity of our procedure is not restricted to the special cases of small measurement error (WK) or small disease probability (RSW).

4.3 Robust Covariance Selection: Mathematics Examination Marks Data

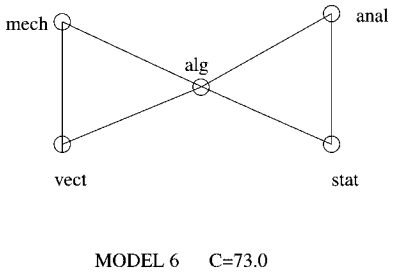
For continuous multivariate data, graphical models are attractive tools for summarizing visually the conditional irrelevance relationship among the variables.

However, most existing techniques for model selection depend on a complete probability model of the data such as joint normality. In the following example, an approach based on joint normality may be questionable due to the skewness and multimodality of some of the variables. On the other hand, the proposed indirect method can be used to produce inferential results that are robust against nonnormality.

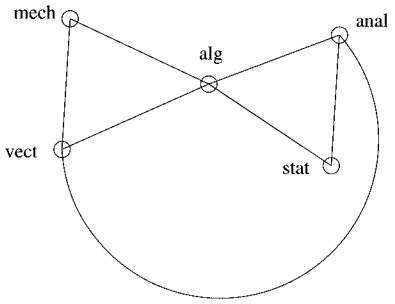
Whittaker (1990, Example 6.7.1) illustrated the graphical Gaussian model (or the covariance selection model) using a dataset of the examination marks of $n = 88$ students in the five mathematics subjects mechanics, vectors, algebra, analysis and statistics, representable as $n = 88$ i.i.d. copies of a five-dimensional random vector $X = (X_j),\ j = 1, \ldots, 5$. The dataset comes from Mardia, Kent and Bibby (1979) and is displayed in full in Table 1.1.1 of Whittaker (1990). Based on the matrix of partial correlations (Table 3), a butterfly graph (see Whittaker, 1990, page 181 or Model 6 in Figure 2 herein) that represents the conditional independence relationships among the five variables was shown to be an "excellent fit to the data," using a goodness-of-fit deviance test based on a multivariate normal model for the responses. By examining the histograms of the five variables (see Figure 3), it can be seen that some of the variables can exhibit left-skewness (analysis) and bimodality (mechanics). Because it is unclear how much the nonnormality affects the inferential results, it is desirable to investigate a robust method for selecting the graphical models for the structure of partial correlations. Note that the essential Markov properties of the graphs are preserved when we consider the weaker property of conditional irrelevance (see Dawid, 1998, page 149), that is, zero partial correlation, rather than the stronger property of conditional independence of the random variables. In such a graphical representation, a pair of vertices that represent two random variables is disconnected if and only if the partial correlation of these two variables is zero given the rest of the random variables. The concept of zero partial correlation is distribution-free (e.g., not dependent on a normal assumption on the vector X), and a corresponding distribution-free test is desirable, as is

TABLE 3
Mathematics marks data: *The sample partial correlation matrix*

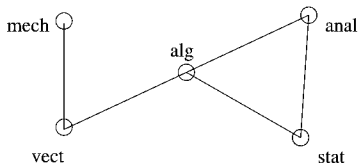
	mech	**vect**	**alg**	**anal**	**stat**
mech	1.0				
vect	0.33	1.0			
alg	0.23	0.28	1.0		
anal	0.00	0.08	0.43	1.0	
stat	0.02	0.02	0.36	0.25	1.0



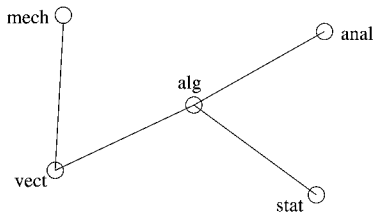
MODEL 6 C=73.0



MODEL 7 C=76.3



MODEL 5 C=74.3



MODEL 4 C=77.7

FIG. 2. *Mathematics marks data*: *Some robust graphical models with small Bayesian costs* ($C = \text{BIC} + \text{const}$).

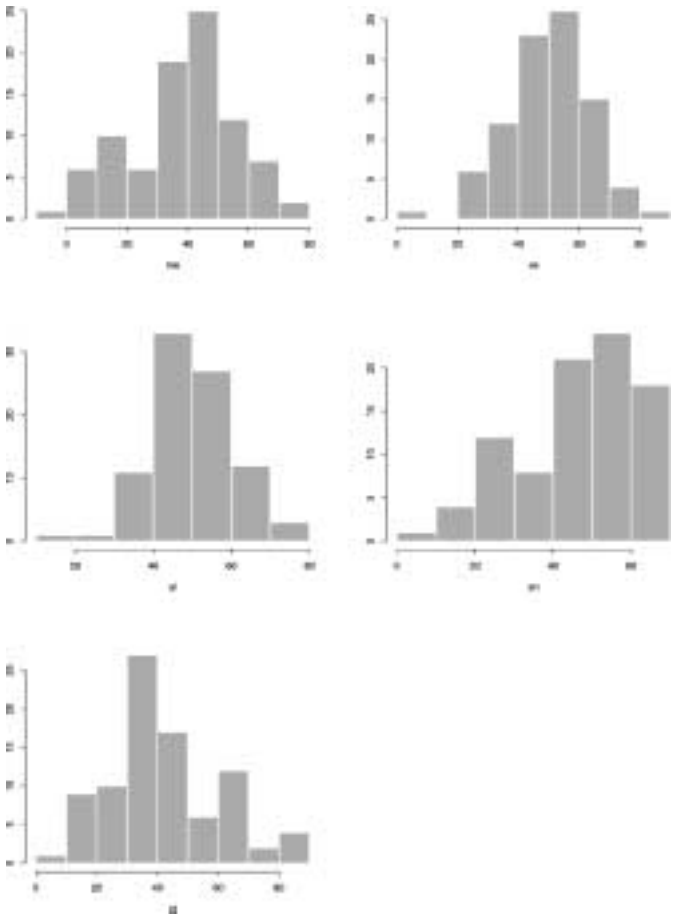


FIG. 3. *Histograms of mathematics examination marks.*

a robust method for selecting graphical models for the structure of partial correlations.

For such a distribution robust treatment, we consider inference based on the intermediate statistics $\hat{s}$ composed of the $(5 + 15)$ first- and second-order sample moments $n^{-1}\sum_{i=1}^{n} X_{ij}$ and $n^{-1}\sum_{i=1}^{n} X_{ij}X_{ij'}$ $(1 \le j \le j' \le 5)$, and using the objective function $H = \{\hat{s} - E(\hat{s}|\theta)\}^T v^{-1}\{\hat{s} - E(\hat{s}|\theta)\}$; see Section 2.6. Here the true parameter includes the five mean parameters $\mu = (\mu_1, \ldots, \mu_5)$, as well as the elements of the symmetric concentration matrix $\Gamma = \mathrm{var}(X)^{-1}$. The weight v is chosen as a sample estimate of the variance matrix of $\hat{s}$, that is, $(v)_{lk} = n^{-1}\sum_{i=1}^{n}(W_{il} - \bar{W}_{\cdot l})(W_{ik} - \bar{W}_{\cdot k})$ and $\bar{W}_{\cdot l} = n^{-1}\sum_{i=1}^{n} W_{il}$, where W_i is the 20-dimensional concatenated vector of X_{ij}'s and $X_{ij}X_{ij'}$'s for $1 \le j \le j' \le 5$ for each i $(1 \le i \le 88)$. This function H is minimized at zero by the saturated (or unrestricted) model with the same estimated means and concentration parameters as the MLEs derived using a multivariate normal specification for the distribution of X. When a subset of partial correlation parameters in θ is constrained to be zero, the minimum value for H of zero can no longer be achieved. For example, for the butterfly graph chosen by Whittaker for this data (Figure 2, Model 6), the concentration matrix has a block structure where elements that correspond to the index pairs {mech–anal, mech–stat, vect–anal, vect–stat} are constrained to be zero. The minimized H under this model equals 1.38 on 4 degrees of freedom and the goodness of fit is excellent (a similar deviance statistic of 0.895 was reported by Whittaker, 1990, page 182, but is based on the normal model of X).

Rather than using subjective judgment based on the observed concentration matrix, we may select a graphical model by considering a BIC analysis using the methods of Section 2.6 based on the intermediate statistic, namely the first- and second-order sample moments. The selection process involves computing the Bayesian cost $C(M)$ for all the models M in the entire searching scope Φ represented by the $2^{10} = 1024$ different graphs. For ease of illustration, we consider a reduced random scope Φ_r with just 10 models,

$M_1, \ldots, M_{10}$ say, where model M_k allows only those partial correlations with the k largest observed absolute values to be nonzero and restricts the remaining $10 - k$ off-diagonal entries in the concentration matrix to be zero. Thus M_{10} is the saturated model and, from Table 3, we see that the butterfly graphical model is M_6. [In general, it can be shown that such a (much) reduced scope, based on ordering the magnitudes of the partial correlations, will contain the simplest true model and the one with the lowest Bayesian cost almost surely in the large-sample limit.]

In Figure 4, the Bayesian cost $C(M)$ is plotted for each of the 10 models in Φ_r. The shape of the graph here appears to indicate that the Bayesian cost criterion penalizes overparametrization less than omission of true nonzero partial correlations. The best model, Model 6 ($k = 6$), corresponds to the butterfly model of Whittaker (1990, Figure 1.1.2), but is here chosen in a (somewhat) automated way. Model 6 suggests that {mechanics, vector} marks and the {analysis, statistics} marks are linearly related primarily through the algebra mark. The Bayesian costs also suggest some close competing models, M_4, M_5, M_7, which all have corresponding leading order a posteriori model probabilities at least 0.05 times that of Model M_6, as characterized by a Bayesian cost exceeding that of Model M_6 by no more than 6 (as represented by the dashed horizontal line in Figure 4). The corresponding graphical models of M_4, M_5, M_6, M_7, which represent the conditional linear irrelevance characterized by zero partial correlations, together with the Bayesian cost of each, are shown in Figure 2. Of course the Bayesian cost can be converted to the scale of posterior model probability. For example, with about 52% of the posterior probability of the favored butterfly model M_6, model M_5 additionally proposes linear irrelevance between students' marks in the mechanics subject and the algebra subject after controlling the vector subject mark. The models M_7 and M_4, on the other hand, are only about 19 and 10%, respectively, as likely as model M_6, based on the intermediate statistics of first- and second-order sample moments.

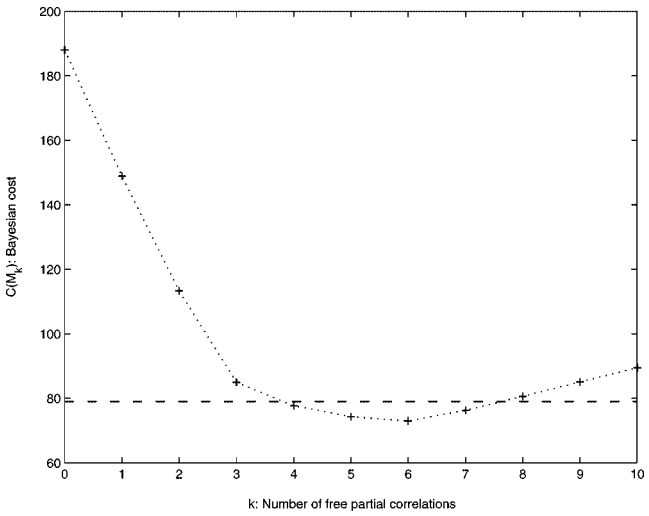


FIG. 4. *Mathematics marks data*: *Bayesian cost versus number of free partial correlation parameters in the model*.

5. CONCLUSION

A number of further applications of the indirect method are discussed in Jiang and Turnbull (2003). These include:

- The method of moment generating functions (mgf) (e.g., Quandt and Ramsey, 1978; Schmidt, 1982) can be regarded as indirect inference based on the intermediate statistic composed of some sample mgf values.
- Optimal linear combination of several consistent estimators (e.g., Serfling, 1980, page 127) can be regarded as the indirect inference based on an intermediate statistic with components including all those consistent estimators.
- The approximate relationship between the maximum likelihood estimates under the reduced model and the extended model [e.g., (5) and (6) of Cox and Wermuth (1990)] can be derived from indirect inference based on an intermediate statistic (the MLE from the extended model).
- The importance sampling estimator of a target distribution can be regarded as the indirect estimator based on an intermediate statistic that is the empirical cdf based on simulated data from the instrumental (naive) distribution.
- The method of least squares can be regarded as indirect inference based on the MLE from a naive regression model assuming independent normal errors with equal variances.
- The method of Gaussian estimation (e.g., Whittle, 1961; Crowder, 1985, 2001; Hand and Crowder, 1996, Chapter 7) can be regarded as indirect inference based on the MLE from a naive regression model assuming normal errors that may be correlated and have unequal variances.

There are other applications that are formally different but similar in spirit to the indirect approach that we discuss in this article. For example:

- Several articles concerning gene mapping (e.g., Wright and Kong, 1997; Sen, 1998) studied inference based on intermediate statistics generated from a naive single-gene normal quantitative trait locus model, when the "true model" can include nonnormality of phenotypic effect and polygenic traits.
- Some methods of nonparametric estimation of additive regression functions are built on marginal integration (e.g., Newey, 1994; Hengartner and Sperlich, 2002) or minimum L_2-distance treatment (e.g., Mammen, Linton and Nielsen, 1999) of an intermediate statistic, which is a full-dimensional local polynomial regression smoother.

ACKNOWLEDGMENTS

We thank the reviewers for very helpful comments and suggestions which led to substantial improvements to the original version. The authors were supported by a grant from the U.S. National Institutes of Health and a grant from the National Science Foundation.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

BERK, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** 51–58. [Correction **37** 745–746.]

BICKEL, P. (1988). Robust estimation. In *Encyclopedia of Statistical Sciences* (S. Kotz and N. L. Johnson, eds.) **8** 157–163. Wiley, New York.

BICKEL, P. J. and DOKSUM, K. A. (2001). *Mathematical Statistics* **1**, 2nd ed. Prentice Hall, Upper Saddle River, NJ.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison–Wesley, London.

BRESLOW, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Amer. Statist. Assoc.* **85** 565–571.

BROZE, L. and GOURIÉROUX, C. (1998). Pseudo-maximum likelihood method, adjusted pseudo-maximum likelihood method and covariance estimators. *J. Econometrics* **85** 75–98.

CARRASCO, M. and FLORENS, J.-P. (2002). Simulation-based method of moments and efficiency. *J. Bus. Econom. Statist.* **20** 482–492.

CARROLL, R. J., RUPPERT, D. and STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

CHIANG, C. L. (1956). On regular best asymptotically normal estimates. *Ann. Math. Statist.* **27** 336–351.

CLARK, L. C., COMBS, G. F., TURNBULL, B. W., SLATE, E. H., CHALKER, D. K., CHOW, J., DAVIS, L. S., GLOVER, R. A., GRAHAM, G. F., GROSS, E. G., KRONGRAD, A., LESHER, J. L., PARK, H. K., SANDERS, B. B., SMITH, C. L., TAYLOR, J. R. and THE NUTRITIONAL PREVENTION OF CANCER STUDY GROUP (1996). Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin: A randomized controlled trial. *J. American Medical Association* **276** 1957–1963; Editorial 1984–1985.

COX, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **24** 406–424.

COX, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

COX, D. R. (1983). Some remarks on overdispersion. *Biometrika* **70** 269–274.

COX, D. R. and WERMUTH, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77** 747–761.

CROWDER, M. (1985). Gaussian estimation for correlated binomial data. *J. Roy. Statist. Soc. Ser. B* **47** 229–237.

CROWDER, M. (2001). On repeated measures analysis with misspecified covariance structure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 55–62.

DAWID, A. P. (1998). Conditional independence. In *Encyclopedia of Statistical Sciences, Update Volume* (S. Kotz, C. B. Read and D. L. Banks, eds.) **2** 146–155. Wiley, New York.

DE LUNA, X. and GENTON, M. G. (2001). Robust simulation-based estimation of ARMA models. *J. Comput. Graph. Statist.* **10** 370–387.

DE LUNA, X. and GENTON, M. G. (2002). Simulation-based inference for simultaneous processes on regular lattices. *Stat. Comput.* **12** 125–134.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DRAPER, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 45–97.

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.

FISHER, R. A. (1946). *Statistical Methods for Research Workers*, 10th ed. Oliver and Boyd, Edinburgh.

FLOREY, C. DU V., MELIA, R. J. W., CHINN, S., GOLDSTEIN, B. D., BROOKS, A. G. F., JOHN, H. H., CRAIGHEAD, E. B. and WEBSTER, X. (1979). The relation between respiratory illness in primary schoolchildren and the use of gas for cooking, III—Nitrogen dioxide, respiratory illness and lung function. *International J. Epidemiology* **8** 347–353.

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.

FULLER, W. A. (1987). *Measurement Error Models*. Wiley, New York.

GAIL, M. H., SANTNER, T. and BROWN, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* **36** 255–266.

GAIL, M. H., WIEAND, S. and PIANTADOSI, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71** 431–444.

GALLANT, A. R. and LONG, J. R. (1997). Estimating stochastic differential equations efficiently by minimum chi-squared. *Biometrika* **84** 125–141.

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

GALLANT, A. R. and TAUCHEN, G. (1999). The relative efficiency of method of moments estimators. *J. Econometrics* **92** 149–172.

GENTON, M. G. and DE LUNA, X. (2000). Robust simulation-based estimation. *Statist. Probab. Lett.* **48** 253–259.

GENTON, M. G. and RONCHETTI, E. (2003). Robust indirect inference. *J. Amer. Statist. Assoc.* **98** 67–76.

GOURIÉROUX, C. and MONFORT, A. (1993). Simulation-based inference—A survey with special reference to panel-data models. *J. Econometrics* **59** 5–33.

GOURIÉROUX, C., MONFORT, A. and RENAULT, E. (1993). Indirect inference. *J. Applied Econometrics* **8S** 85–118.

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. dissertation, Univ. California, Berkeley.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions*. Wiley, New York.

HAND, D. and CROWDER, M. (1996). *Practical Longitudinal Data Analysis.* Chapman and Hall/CRC, London.

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

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271.

HECKMAN, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5** 475–492.

HENGARTNER, N. W. and SPERLICH, S. (2002). Rate optimal estimation with the integration method in the presence of many covariates. Working Paper 01-69, Carlos III de Madrid. Available at http://halweb.uc3m.es/esp/Personal/personas/stefan/papers/may2002.pdf.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press.

IMBENS, G. W. (2002). Generalized method of moments and empirical likelihood. *J. Bus. Econom. Statist.* **20** 493–506.

JIANG, W. (1996). Aspects of misspecification in statistical models: Applications to latent variables, measurement error, random effects, omitted covariates and incomplete data. Ph.D. dissertation, Cornell Univ.

JIANG, W. and TURNBULL, B. W. (2003). The indirect method—Robust inference based on intermediate statistics. Technical Report 1377, School of Operations Research and Industrial Engineering, Cornell Univ. Available at http://www.orie.cornell.edu/trlist/trlist.html.

JIANG, W., TURNBULL, B. W. and CLARK, L. C. (1999). Semiparametric regression models for repeated events with random effects and measurement error. *J. Amer. Statist. Assoc.* **94** 111–124.

KENT, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69** 19–27.

KUK, A. Y. C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *J. Roy. Statist. Soc. Ser. B* **57** 395–407.

LAWLESS, J. F. and NADEAU, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37** 158–168.

LEADERER, B. P., ZAGRANISKI, R. T., BERWICK, M. and STOLWIJK, J. A. J. (1986). Assessment of exposure to indoor air contaminants from combustion sources: Methodology and application. *American J. Epidemiology* **124** 275–289.

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 129–156. Univ. California Press.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.

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

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81** 471–483.

MACKINNON, J. G. and SMITH, A. A. (1998). Approximate bias correction in econometrics. *J. Econometrics* **85** 205–230.

MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490.

MARDIA, K. V., KENT, J. T. and BIBBY, J. (1979). *Multivariate Analysis*. Academic Press, New York.

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

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, New York.

MCFADDEN, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* **57** 995–1026.

NEWEY, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10** 233–253.

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

PAKES, A. and POLLARD, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057.

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

QUANDT, R. E. and RAMSEY, J. B. (1978). Estimating mixtures of normal distributions and switching regressions (with discussion). *J. Amer. Statist. Assoc.* **73** 730–752.

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

REID, N. (1988). Influence functions. In *Encyclopedia of Statistical Sciences* (S. Kotz and N. L. Johnson, eds.) **4** 117–119. Wiley, New York.

RONCHETTI, E. and TROJANI, F. (2001). Robust inference with GMM estimators. *J. Econometrics* **101** 37–69.

ROSNER, B., SPIEGELMAN, D. and WILLETT, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American J. Epidemiology* **132** 734–745.

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

SCHMIDT, P. (1982). An improved version of the Quandt–Ramsey MGF estimator for mixtures of normal distributions and switching regressions. *Econometrica* **50** 501–516.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SEN, P. K. and SINGER, J. M. (1993). *Large Sample Methods in Statistics*. Chapman and Hall, New York.

SEN, S. (1998). Confidence intervals for gene location: The effect of model misspecification and smoothing. Ph.D. dissertation, Dept. Statistics, Univ. Chicago.

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

TAYLOR, J. R. (1997). *An Introduction to Error Analysis,* 2nd ed. University Science Books, Sausalito, CA.

THOMPSON, H. F., GRUBBS, C. J., MOON, R. C. and SPORN, M. B. (1978). Continual requirement of retinoid for maintenance of mammary cancer inhibition. *Proc. Annual Meeting of the American Association for Cancer Research* **19** 74.

TURNBULL, B. W., JIANG, W. and CLARK, L. C. (1997). Regression models for recurrent event data: Parametric random effects models with measurement error. *Statistics in Medicine* **16** 853–864.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61** 439–447.

WEI, L. J., LIN, D. Y. and WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.* **84** 1065–1073.

WHITE, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge Univ. Press.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

WHITTEMORE, A. S. and KELLER, J. B. (1988). Approximations for regression with covariate measurement error. *J. Amer. Statist. Assoc.* **83** 1057–1066.

WHITTLE, P. (1961). Gaussian estimation in stationary time series. *Bull. Internat. Statist. Inst.* **39** 105–129.

WRIGHT, F. A. and KONG, A. (1997). Linkage mapping in experimental crosses: The robustness of single-gene models. *Genetics* **146** 417–425.