DEPARTMENT OF ECONOMICS
UNIVERSITY OF CYPRUS

THE MM, ME, ML, EL, EF AND GMM APPROACHES TO ESTIMATION: A SYNTHESIS

Anil K. Bera and Yannis Bilias

Discussion Paper 2001-09

P.O. Box 20537, 1678 Nicosia, CYPRUS Tel.: ++357-2-892430, Fax: ++357-2-892432
Web site: http://www.econ.ucy.ac.cy

The MM, ME, ML, EL, EF and GMM Approaches to Estimation: A Synthesis

Anil K. Bera*
Department of Economics
University of Illinois
1206 S. 6th Street
Champaign, IL 61820; USA

Yannis Bilias
Department of Economics
University of Cyprus
P.O. Box 20537
1678 Nicosia; CYPRUS

JEL Classification: C13; C52

Keywords: Karl Pearson's goodness-of-fit statistic; Entropy; Method of moment; Estimating function; Likelihood; Empirical likelihood; Generalized method of moments; Power divergence criterion; History of estimation.

*Corresponding author: Anil K. Bera, Department of Economics, University of Illinois, Urbana-Champaign, 1206 S. 6th Street, Champaign, IL 61820, USA. Phone: (217) 333-4596; Fax: (217) 244-6678; email: anil@fisher.econ.uiuc.edu

Abstract

The 20th century began on an auspicious statistical note with the publication of Karl Pearson's (1900) goodness-of-fit test, which is regarded as one of the most important scientific breakthroughs. The basic motivation behind this test was to see whether an assumed probability model adequately described the data at hand. Pearson (1894) also introduced a formal approach to statistical estimation through his method of moments (MM) estimation. Ronald A. Fisher, while he was a third year undergraduate at the Gonville and Caius College, Cambridge, suggested the maximum likelihood estimation (MLE) procedure as an alternative to Pearson's MM approach. In 1922 Fisher published a monumental paper that introduced such basic concepts as consistency, efficiency, sufficiency–and even the term "parameter" with its present meaning. Fisher (1922) provided the analytical foundation of MLE and studied its efficiency relative to the MM estimator. Fisher (1924a) established the asymptotic equivalence of minimum χ^2 and ML estimators and wrote in favor of using minimum χ^2 method rather than Pearson's MM approach. Recently, econometricians have found working under assumed likelihood functions restrictive, and have suggested using a generalized version of Pearson's MM approach, commonly known as the GMM estimation procedure as advocated in Hansen (1982). Earlier, Godambe (1960) and Durbin (1960) developed the estimating function (EF) approach to estimation that has been proven very useful for many statistical models. A fundamental result is that score is the optimum EF. Ferguson (1958) considered an approach very similar to GMM and showed that estimation based on the Pearson chi-squared statistic is equivalent to efficient GMM. Golan, Judge and Miller (1996) developed entropy-based formulation that allowed them to solve a wide range of estimation and inference problems in econometrics. More recently, Imbens, Spady and Johnson (1998), Kitamura and Stutzer (1997) and Mittelhammer, Judge and Miller (2000) put GMM within the framework of empirical likelihood (EL) and maximum entropy (ME) estimation. It can be shown that many of these estimation techniques can be obtained as special cases of minimizing Cressie and Read (1984) power divergence criterion that comes directly from the Pearson (1900) chi-squared statistic. In this way we are able to assimilate a number of seemingly unrelated estimation techniques into a unified framework.

1 Prologue: Karl Pearson's method of moment estimation and chi-squared test, and entropy

In this paper we are going to discuss various methods of estimation, especially those developed in the twentieth century, beginning with a review of some developments in statistics at the close of the nineteenth century. In 1892 W.F. Raphael Weldon, a zoologist turned statistician, requested Karl Pearson (1857-1936) to analyze a set of data on crabs. After some investigation Pearson realized that he could not fit the usual normal distribution to this data. By the early 1890's Pearson had developed a class of distributions that later came to be known as the Pearson system of curves, which is much broader than the normal distribution. However, for the crab data Pearson's own system of curves was not good enough. He dissected this "abnormal frequency curve" into two normal curves as follows:

$$f(y) = \alpha f_1(y) + (1-\alpha) f_2(y), \tag{1}$$

where

$$f_j(y) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp[-\frac{1}{2\sigma_j^2}(y-\mu_j)^2], \qquad j = 1, 2.$$

This model has five parameters[1] $(\alpha, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. Previously, there had been no method available to estimate such a model. Pearson quite unceremoniously suggested a method that simply equated the *first* five population moments to the respective sample counterparts. It was not easy to solve five highly nonlinear equations. Therefore, Pearson took an analytical approach of eliminating one parameter in each step. After considerable algebra he found a ninth-degree polynomial equation in one unknown. Then, after solving this equation and by repeated back-substitutions, he found solutions to the five parameters in terms of the first five sample moments. It was around the autumn of 1893 he completed this work and it appeared in 1894. And this was the beginning of the method of moment (MM) estimation. There is no general theory in Pearson (1894). The paper is basically a worked-out "example" (though a very difficult one as the first illustration of MM estimation) of a new estimation method.[2]

[1] The term "parameter" was introduced by Fisher (1922, p.311) [also see footnote 16]. Karl Pearson described the "parameters" as "constants" of the "curve." Fisher (1912) also used "frequency curve." However, in Fisher (1922) he used the term "distribution" throughout. "Probability density function" came much later, in Wilks (1943, p.8) [see, David (1995)]

[2] Shortly after Karl Pearson's death, his son Egon Pearson provided an account of life and work of the elder Pearson [see Pearson (1936)]. He summarized (pp.219-220) the contribution of Pearson (1894) stating, "The paper is particularly noteworthy for its introduction of the method of moments as a means of fitting a theoretical curve to observed data. This method is not claimed to be the best but is advocated from the utilitarian standpoint

After an experience of "some eight years" in applying the MM to a vast range of physical and social data, Pearson (1902) provided some "theoretical" justification of his methodology. Suppose we want to estimate the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ of the probability density function $f(y;\theta)$. By a Taylor series expansion of $f(y) \equiv f(y;\theta)$ around $y = 0$, we can write

$$f(y) = \phi_0 + \phi_1 y + \phi_2 \frac{y^2}{2!} + \phi_3 \frac{y^3}{3!} + \ldots + \phi_p \frac{y^p}{p!} + R, \tag{2}$$

where $\phi_0, \phi_1, \phi_2, \ldots, \phi_p$ depends on $\theta_1, \theta_2, \ldots, \theta_p$ and R is the remainder term. Let $\bar{f}(y)$ be the ordinate corresponding to y given by observations. Therefore, the problem is to fit a smooth curve $f(y;\theta)$ to p histogram ordinates given by $\bar{f}(y)$. Then $f(y) - \bar{f}(y)$ denotes the distance between the theoretical and observed curve at the point corresponding to y, and our objective would be to make this distance as *small as possible* by a proper choice of $\phi_0, \phi_1, \phi_2, \ldots, \phi_p$ [see Pearson (1902, p.268)].[3] Although Pearson discussed the fit of $f(y)$ to p histogram ordinates $\bar{f}(y)$, he proceeded to find a "theoretical" version of $f(y)$ that minimizes [see Mensch (1980)]

$$\int [f(y) - \bar{f}(y)]^2 dy. \tag{3}$$

Since $f(.)$ is the variable, the resulting equation is

$$\int [f(y) - \bar{f}(y)] \delta f dy = 0, \tag{4}$$

where, from (2), the differential δf can be written as

$$\delta f = \sum_{j=0}^{p} (\delta \phi_j \frac{y^j}{j!} + \frac{\partial R}{\partial \phi_j} \delta \phi_j). \tag{5}$$

Therefore, we can write equation (4) as

$$\int [f(y) - \bar{f}(y)] \sum_{j=0}^{p} (\delta \phi_j \frac{y^j}{j!} + \frac{\partial R}{\partial \phi_j} \delta \phi_j) dy = \sum_{j=0}^{p} \int [f(y) - \bar{f}(y)] (\frac{y^j}{j!} + \frac{\partial R}{\partial \phi_j}) dy \delta \phi_j = 0. \tag{6}$$

Since the quantities $\phi_0, \phi_1, \phi_2, \ldots, \phi_p$ are at our choice, for (6) to hold, each component should be independently zero, i.e., we should have

$$\int [f(y) - \bar{f}(y)] (\frac{y^j}{j!} + \frac{\partial R}{\partial \phi_j}) dy = 0, \qquad j = 0, 1, 2, \ldots, p, \tag{7}$$

on the grounds that it appears to give excellent fits and provides algebraic solutions for calculating the constants of the curve which are analytically possible."

[3] It is hard to trace the first use of smooth non-parametric density estimation in the statistics literature. Koenker (2000, p.349) mentioned Galton's (1885) illustration of "regression to the mean" where Galton averaged the counts from the four adjacent squares to achieve smoothness. Karl Pearson's minimization of the distance between $f(y)$ and $\bar{f}(y)$ looks remarkably modern in terms of ideas and could be viewed as a modern-equivalent of smooth non-parametric density estimation [see also Mensch (1980)].

which is same as

$$\mu_j = m_j - j! \int [f(y) - \bar{f}(y)](\frac{\partial R}{\partial \phi_j})dy, \qquad j = 0, 1, 2, \ldots, p. \tag{8}$$

Here μ_j and m_j are, respectively, the j-th moment corresponding to the theoretical curve $f(y)$ and the observed curve $\bar{f}(y)$.[4] Pearson (1902) then ignored the integral terms arguing that they involve the small factor $f(y) - \bar{f}(y)$, and the remainder term R, which by "hypothesis" is small for large enough sample size. After neglecting the integral terms in (8), Pearson obtained the equations

$$\mu_j = m_j, \qquad j = 0, 1, \ldots, p. \tag{9}$$

Then, he stated the principle of the MM as [see Pearson (1902, p.270)]: "To fit a good theoretical curve $f(y; \theta_1, \theta_2, \ldots, \theta_p)$ to an observed curve, express the area and moments of the curve for the given range of observations in terms of $\theta_1, \theta_2, \ldots, \theta_p$, and equate these to the like quantities for the observations." Arguing that, if the first p moments of two curves are identical, the higher moments of the curves becomes "*ipso facto* more and more nearly identical" for larger sample size, he concluded that the "equality of moments gives a good method of fitting curves to observations" [Pearson (1902, p.271)]. We should add that much of his theoretical argument is not very rigorous, but the 1902 paper did provide a reasonable theoretical basis for the MM and illustrated its usefulness.[5] For detailed discussion on the properties of the MM estimator see Shenton (1950, 1958, 1959).

After developing his system of curves [Pearson (1895)], Pearson and his associates were fitting this system to a large number of data sets. Therefore, there was a need to formulate a test to check whether an assumed probability model adequately explained the data at hand. He succeeded in doing that and the result was Pearson's celebrated (1900) χ^2 goodness-of-fit test. To describe the Pearson test let us consider a distribution with k classes with the

[4] It should be stressed that $m_j = \int y^j \bar{f}(y)dy = \sum_i^n y_i^j \pi_i$ with π_i denoting the area of the bin of the ith observation; this is not necessarily equal to the sample moment $n^{-1} \sum_i y_i^j$ that is used in today's MM. Rather, Pearson's formulation of empirical moments uses the efficient weighting π_i under a multinomial probability framework, an idea which is used in the literature of empirical likelihood and maximum entropy and to be described later in this paper.

[5] One of the first and possibly most important applications of MM idea is the derivation of t-distribution in Student (1908) which was major breakthrough in introducing the concept of finite sample (exact) distribution in statistics. Student (1908) obtained the first four moments of the sample variance S^2, matched them with those of the Pearson type III distribution, and concluded (p.4) "a curve of Professor Pearson's type III may be expected to fit the distribution of S^2." Student, however, was very cautious and quickly added (p.5), "it is probable that the curve found represents the theoretical distribution of S^2 so that although we have no actual proof we shall assume it to do so in what follows." And this was the basis of his derivation of the t-distribution. The name t-distribution was given by Fisher (1924b).

probability of j-th class being $q_j(\geq 0)$, $j = 1, 2, \ldots, k$ and $\sum_{j=1}^{k} q_j = 1$. Suppose that according to the assumed probability model, $q_j = q_{j0}$; therefore, one would be interested in testing the hypothesis, $H_0 : q_j = q_{j0}$, $j = 1, 2, \ldots, k$. Let n_j denote the observed frequency of the j-th class, with $\sum_{j=1}^{k} n_j = N$. Pearson (1900) suggested the goodness-of-fit statistic[6]

$$P = \sum_{j=1}^{k} \frac{(n_j - Nq_{j0})^2}{Nq_{j0}} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}, \tag{10}$$

where O_j and E_j denote, respectively, the observed and expected frequencies of the j-th class. This is the first constructive test in the statistics literature. Broadly speaking, P is essentially a *distance measure* between the observed and expected frequencies.

It is quite natural to question the relevance of this test statistic in the context of estimation. Let us note that P could be used to measure the *distance* between any two sets of probabilities, say, (p_j, q_j), $j = 1, 2, \ldots, k$ by simply writing $p_j = n_j/N$ and $q_j = q_{j0}$, i.e.,

$$P = N \sum_{j=1}^{k} \frac{(p_j - q_j)^2}{q_j}. \tag{11}$$

As we will see shortly a simple transformation of P could generate a broad class of distance measures. And later, in Section 5, we will demonstrate that many of the current estimation procedures in econometrics can be cast in terms of minimizing the distance between two sets of probabilities subject to certain constraints. In this way, we can tie and assimilate many estimation techniques together using Pearson's MM and χ^2-statistic as the unifying themes.

We can write P as

$$P = N \sum_{j=1}^{k} \frac{p_j(p_j - q_j)}{q_j} = N \sum_{j=1}^{k} p_j \left(\frac{p_j}{q_j} - 1 \right). \tag{12}$$

Therefore, the essential quantity in measuring the divergence between two probability distributions is the ratio (p_j/q_j). Using Steven's (1975) idea on "visual perception" Cressie and Read

[6] This test is regarded as one of the 20 most important scientific breakthroughs of this century along with advances and discoveries like the theory of relativity, the IQ test, hybrid corn, antibiotics, television, the transistor and the computer [see Hacking (1984)]. In his editorial in the inaugural issue of *Sankhyā, The Indian Journal of Statistics,* Mahalanobis (1933) wrote, "...the history of modern statistics may be said to have begun from Karl Pearson's work on the distribution of χ^2 in 1900. The Chi-square test supplied for the first time a tool by which the significance of the agreement or discrepancy between theoretical expectations and actual observations could be judged with precision." Even Pearson's lifelong arch-rival Ronald A. Fisher (1922, p.314) conceded, "Nor is the introduction of the Pearsonian system of frequency curves the only contribution which their author has made to the solution of problems of specification: of even greater importance is the introduction of an objective criterion of goodness of fit." For more on this see Bera (2000) and Bera and Bilias (2001).

(1984) suggested using the relative difference between the perceived probabilities as $(p_j/q_j)^\lambda - 1$ where λ "typically lies in the range from 0.6 to 0.9" but could theoretically be any real number [see also Read and Cressie (1988, p.17)]. By weighing this quantity proportional to p_j and summing over all the classes, leads to the following measure of divergence:

$$\sum_{j=1}^{k} p_j \left[\left(\frac{p_j}{q_j} \right)^\lambda - 1 \right]. \tag{13}$$

This is approximately proportional to the Cressie and Read (1984) power divergence family of statistics[7]

$$\begin{aligned} I_\lambda(p,q) &= \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} p_j \left[\left(\frac{p_j}{q_j} \right)^\lambda - 1 \right] \\ &= \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} q_j \left[\left\{ 1 + \left(\frac{p_j}{q_j} - 1 \right) \right\}^{\lambda+1} - 1 \right], \end{aligned} \tag{14}$$

where $p = (p_1, p_2, \ldots, p_n)'$ and $q = (q_1, q_2, \ldots, q_n)'$. Lindsay (1994, p.1085) calls $\delta_j = (p_j/q_j) - 1$ the "Pearson" residual since we can express the Pearson statistic in (11) as $P = N \sum_{j=1}^{k} q_j \delta_j^2$. From this, it is immediately seen that when $\lambda = 1$, $I_\lambda(p,q)$ reduces to P/N. In fact, a number of well-known test statistics can be obtained from $I_\lambda(p,q)$. When $\lambda \longrightarrow 0$, we have the likelihood (LR) test statistic, which, as an alternative to (10), can be written as

$$LR = 2 \sum_{j=1}^{k} n_j \ln \left(\frac{n_j}{N q_{j0}} \right) = 2 \sum_{j=1}^{k} O_j \ln \left(\frac{O_j}{E_j} \right). \tag{15}$$

Similarly, $\lambda = -1/2$ gives the Freeman and Tukey (FT) (1950) statistic, or Hellinger distance,

$$FT = 4 \sum_{j=1}^{k} (\sqrt{n_j} - \sqrt{n q_{j0}})^2 = 4 \sum_{j=1}^{k} (\sqrt{O_j} - \sqrt{E_j})^2. \tag{16}$$

All these test statistics are just different measures of distance between the observed and expected frequencies. Therefore, $I_\lambda(p,q)$ provides a very rich class of divergence measures.

Any probability distribution $p_i,\ i = 1, 2, \ldots, n$ (say) of a random variable taking n values provides a measure of *uncertainty* regarding that random variable. In the information theory literature, this measure of uncertainty is called *entropy*. The origin of the term "entropy" goes

[7] In the entropy literature this is known as Renyi's (1961) α-class generalized measures of entropy [see Maasoumi (1993, p.144), Ullah (1996, p.142) and Mittelhammer, Judge and Miller (2000, p.328)]. Golan, Judge and Miller (1996, p.36) referred to Schützenberger (1954) as well. This formulation has also been used extensively as a general class of decomposable income inequality measures, for example, see Cowell (1980) and Shorrocks (1980), and in time-series analysis to distinguish chaotic data from random data [Pompe (1994)].

back to thermodynamics. The second law of thermodynamics states that there is an inherent tendency for disorder to increase. A probability distribution gives us a measure of disorder. Entropy is generally taken as a measure of expected information, that is, how much information do we have in the probability distribution $p_i, \;\; i = 1, 2, \ldots, n$. Intuitively, information should be a decreasing function of p_i, i.e., the more unlikely an event, the more interesting it is to know that it can happen [see Shannon and Weaver (1949, p.105) and Sen (1975, pp.34-35)].

A simple choice for such a function is $-\ln p_i$. Entropy $H(p)$ is defined as a weighted sum of the information $-\ln p_i, \;\; i = 1, 2, \ldots, n$ with respective probabilities as weights, namely,

$$H(p) = -\sum p_i \ln p_i. \tag{17}$$

If $p_i = 0$ for some i, then $p_i \ln p_i$ is taken to be zero. When $p_i = 1/n$ for all i, $H(p) = \ln n$ and then we have the *maximum* value of the entropy and consequently the *least information* available from the probability distribution. The other extreme case occurs when $p_i = 1$ for one i, and $= 0$ for the rest; then $H(p) = 0$. If we do not weigh each $-\ln p_i$ by p_i and simply take the sum, another measure of entropy would be

$$H'(p) = -\sum_{i=1}^{n} \ln p_i. \tag{18}$$

Following (17), the cross-entropy of one probability distribution $p = (p_1, p_2, \ldots, p_n)'$ with respect to another distribution $q = (q_1, q_2, \ldots, q_n)'$ can be defined as

$$C(p, q) = \sum_{i=1}^{n} p_i \ln(p_i/q_i) = E[\ln p] - E[\ln q], \tag{19}$$

which is yet another measure of distance between two distributions. It is easy to see the link between $C(p, q)$ and the Cressie and Read (1984) power divergence family. If we choose $q = (1/n, 1/n, \ldots, 1/n)' = \mathbf{i}/n$ where $\mathbf{i}$ is a $n \times 1$ vector of ones, $C(p, q)$ reduces to

$$C(p, \mathbf{i}/n) = \sum_{i=1}^{n} p_i \ln p_i - \ln n. \tag{20}$$

Therefore, entropy maximization is a special case of cross-entropy minimization with respect to the uniform distribution. For more on entropy, cross-entropy and their uses in econometrics see Maasoumi (1993), Ullah (1996), Golan, Judge and Miller (1996, 1997 and 1998), Zellner and Highfield (1988), Zellner (1991) and other papers in Grandy and Schick (1991), Zellner (1997) and Mittelhammer, Judge and Miller (2000).

If we try to find a probability distribution that maximizes the entropy $H(p)$ in (17), the optimal solution is the uniform distribution, i.e., $p^* = \mathbf{i}/n$. In the Bayesian literature, it is common to maximize an entropy measure to find non-informative priors. Jaynes (1957) was the first to consider the problem of finding a prior distribution that maximizes $H(p)$ subject to certain side conditions, which could be given in the form of some moment restrictions. Jaynes' problem can be stated as follows. Suppose we want to find a *least informative* probability distribution $p_i = \Pr(Y = y_i), i = 1, 2, \ldots, n$ of a random variable Y satisfying, say, m moment restrictions $E[h_j(Y)] = \mu_j$ with known μ_j's, $j = 1, 2, \ldots, m$. Jaynes (1957, p.623) found an explicit solution to the problem of maximizing $H(p)$ subject to the above moment conditions and $\sum_{i=1}^{n} p_i = 1$ [for a treatment of this problem under very general conditions, see, Haberman (1984)]. We can always find some (in fact, many) solutions just by satisfying the constraints; however, maximization of (17) makes the resulting probabilities p_i $(i = 1, 2, \ldots, n)$ *as smooth as possible*. Jaynes (1957) formulation has been extensively used in the Bayesian literature to find priors that are as noninformative as possible given some prior partial information [see Berger (1985, pp.90-94)]. In recent years econometricians have tried to estimate parameter(s) of interest say, θ, utilizing only certain moment conditions satisfied by the underlying probability distribution, known as the generalized method of moments (GMM) estimation. The GMM procedure is an extension of Pearson's (1895, 1902) MM when we have more moment restrictions than the dimension of the unknown parameter vector. The GMM estimation technique can also be cast into the information theoretic approach of maximization of entropy following the empirical likelihood (EL) method of Owen (1988, 1990, 1991) and Qin and Lawless (1994). Back and Brown (1993), Kitamura and Stutzer (1997) and Imbens, Spady and Johnson (1998) developed information theoretic approaches of entropy maximization estimation procedures that include GMM as a special case. Therefore, we observe how seemingly distinct ideas of Pearson's χ^2 test statistic and GMM estimation are tied to the common principle of measuring distance between two probability distributions through the entropy measure. The modest aim of this review paper is essentially this idea of assimilating distinct estimation methods. In the following two sections we discuss Fisher's (1912, 1922) maximum likelihood estimation (MLE) approach and its relative efficiency to the MM estimation method. The MLE is the forerunner of the currently popular EL approach. We also discuss the minimum χ^2 method of estimation, which is based on the minimization of the Pearson χ^2 statistic. Section 4 proceeds with optimal estimation using an estimating function (EF) approach. In Section 5, we discuss the instrumental variable (IV) and GMM estimation procedure along with their recent variants. Both EF and GMM approaches were devised in order to handle problems of method of moments estimation where the number of moment restrictions is larger than the number of parameters. The last section provides some concluding remarks. While doing the survey, we also try to provide some personal perspectives on researchers who contributed to the amazing progress in statistical and econometrics estimation techniques that we have witnessed in the last 100 years. We do this since in many instances the original motivation and philosophy of various statistical techniques have become clouded over time. And to the best of our knowledge these materials have not found a place in econometric textbooks.

2 Fisher's (1912) maximum likelihood, and the minimum chi-squared methods of estimation

In 1912 when R. A. Fisher published his *first* mathematical paper, he was a third and final year undergraduate in mathematics and mathematical physics in Gonville and Caius College, Cambridge. It is now hard to envision exactly what prompted Fisher to write this paper. Possibly, his tutor the astronomer F. J. M. Stratton (1881-1960), who lectured on the theory of errors, was the instrumental factor. About Stratton's role, Edwards (1997a, p.36) wrote: "In the Easter Term 1911 he had lectured at the observatory on *Calculation of Orbits from Observations,* and during the next academic year on *Combination of Observations* in the Michaelmas Term (1911), the first term of Fisher's third and final undergraduate year. It is very likely that Fisher attended Stratton's lectures and subsequently discussed statistical questions with him during mathematics supervision in College, and he wrote the 1912 paper as a result."[8]

The paper started with a criticism of two known methods of curve fitting, least squares and Pearson's MM. In particular, regarding MM, Fisher (1912, p.156) stated "a choice has been made without theoretical justification in selecting r equations ..." Fisher was referring to the equations in (9), though Pearson (1902) defended his choice on the ground that these lower-order moments have smallest relative variance [see Hald (1998, p.708)].

After disposing of these two methods, Fisher stated "we may solve the real problem directly" and set out to discuss his absolute criterion for fitting frequency curves. He took the probability density function (p.d.f) $f(y;\theta)$ (using our notation) as an ordinate of the theoretical curve of unit area and, hence, interpreted $f(y;\theta)\delta_y$ as the chance of an observation falling within the

[8] Fisher (1912) ends with "In conclusion I should like to acknowledge the great kindness of Mr. J.F.M. Stratton, to whose criticism and encouragement the present form of this note is due." It may not be out of place to add that in 1912 Stratton also prodded his young pupil to write directly to Student (William S. Gosset, 1876-1937), and Fisher sent Gosset a rigorous proof of t-distribution. Gosset was sufficiently impressed to send the proof to Karl Pearson with a covering letter urging him to publish it in *Biometrika* as a note. Pearson, however, was not impressed and nothing more was heard of Fisher's proof [see Box (1978, pp.71-73) and Lehmann (1999, pp.419-420)]. This correspondence between Fisher and Gosset was the beginning of a lifelong mutual respect and friendship until the death of Gosset.

range δ_y. Then he defined (p.156)

$$\ln P' = \sum_{i=1}^{n} \ln f(y_i;\theta)\delta_{y_i} \tag{21}$$

and interpreted P' to be "proportional to the chance of a given set of observations occurring." Since the factors δ_{y_i} are independent of $f(y;\theta)$, he stated that the "probability of any particular set of θ's is proportional to P," where

$$\ln P = \sum_{i=1}^{n} \ln f(y_i;\theta) \tag{22}$$

and the most probable set of values for the θ's will make P a maximum" (p.157). This is in essence Fisher's idea regarding maximum likelihood estimation.[9] After outlining his method for fitting curves, Fisher applied his criterion to estimate parameters of a normal density of the following form

$$f(y;\mu,h) = \frac{h}{\sqrt{\pi}} \exp[-h^2(y-\mu)^2], \tag{23}$$

where $h = 1/\sigma\sqrt{2}$ in the standard notation of $N(\mu,\sigma^2)$. He obtained the "most probable values" as[10]

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{24}$$

and

$$\hat{h}^2 = \frac{n}{2\sum_{i=1}^{n}(y_i-\bar{y})^2}. \tag{25}$$

Fisher's probable value of h did not match the conventional value that used $(n-1)$ rather than n as in (25) [see Bennett (1907-1908)]. By integrating out μ from (23), Fisher obtained

[9] We should note that nowhere in Fisher (1912) he uses the word "likelihood." It came much later in Fisher (1921, p.24), and the phrase "method of maximum likelihood" was first used in Fisher (1922, p.323) [also see Edwards (1997a, p.36)]. Fisher (1912) did not refer to the Edgeworth (1908, 1909) inverse probability method which gives the same estimates, or for that matter to most of the early literature (the paper contained only two references). As Aldrich (1997, p.162) indicated "nobody" noticed Edgeworth work until "Fisher had redone it." Le Cam (1990, p.153) settled the debate on who first proposed the maximum likelihood method in the following way: "Opinions on who was the first to propose the method differ. However Fisher is usually credited with the invention of the name 'maximum likelihood', with a major effort intended to spread its use and with the derivation of the optimality properties of the resulting estimates." We can safely say that although the method of maximum likelihood pre-figured in earlier works, it was first presented in its own right, and with a full view of its significance by Fisher (1912) and later by Fisher (1922).

[10] In fact Fisher did not use notations $\hat{\mu}$ and $\hat{h}$. Like Karl Pearson he did not distinguish between the parameter and its estimator. That came much later in Fisher (1922, p.313) when he introduced the concept of "statistic."

the "variation of h," and then maximizing the resulting marginal density with respect to h, he found the conventional estimator

$$\hat{\hat{h}}^2 = \frac{n-1}{2\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{26}$$

Fisher (p.160) interpreted

$$P = \prod_{i=1}^{n} f(y_i; \theta) \tag{27}$$

as the "relative probability of the set of values" $\theta_1, \theta_2, \ldots, \theta_p$. Implicitly, he was basing his arguments on inverse probability (posterior distribution) with noninformative prior. But at the same time he criticized the process of obtaining (26) saying "integration" with respect to μ is "illegitimate and has no meaning with respect to inverse probability." Here Fisher's message is very confusing and hard to decipher.[11] In spite of these misgivings Fisher (1912) is a remarkable paper given that it was written when Fisher was still an undergraduate. In fact, his idea of the likelihood function (27) played a central role in introducing and crystallizing some of the fundamental concepts in statistics.

The history of the minimum chi-squared (χ^2) method of estimation is even more blurred. Karl Pearson and his associates routinely used the MM to estimate parameters and the χ^2 statistic (10) to test the adequacy of the fitted model. This state of affairs prompted Hald (1998, p.712) to comment: "One may wonder why he [Karl Pearson] did not take further step to minimizing χ^2 for estimating the parameters." In fact, for a while, nobody took a concrete step in that direction. As discussed in Edwards (1997a) several early papers that advocated this method of estimation could be mentioned: Harris (1912), Engledow and Yule (1914), Smith (1916) and Haldane (1919a, b). Ironically, it was the Fisher (1928) book and its subsequent editions that brought to prominence this estimation procedure. Smith (1916) was probably the first to state explicitly how to obtain parameter estimates using the minimum χ^2 method. She started with a mild criticism of Pearson's MM (p.11): "It is an undoubtedly utile and accurate method; but the question of whether it gives the 'best' values of the constant has not been very fully studied."[12] Then, without much fanfare she stated (p.12): "From another standpoint,

[11] In Fisher (1922, p.326) he went further and confessed: "I must indeed plead guilty in my original statement in the Method of Maximum Likelihood [Fisher (1912)] to having based my argument upon the principle of inverse probability; in the same paper, it is true, I emphasized the fact that such inverse probabilities were relative only." Aldrich (1997), Edwards (1997b) and Hald (1999) examined Fisher's paradoxical views to "inverse probability" in detail.

[12] Kirstine Smith was a graduate student in Karl Pearson's laboratory since 1915. In fact her paper ends with the following acknowledgement: "The present paper was worked out in the Biometric Laboratory and I have to thank Professor Pearson for his aid throughout the work." It is quite understandable that she could not be too critical of Pearson's MM.

however, the 'best values' of the frequency constants may be said to be those for which" the quantity in (10) "is a minimum." She argued that when χ^2 is a minimum, "the probability of occurrence of a result as divergent as or more divergent than the observed, will be maximum." In other words, using the minimum χ^2 method the "goodness-of-fit" might be better than that obtained from the MM. Using a slightly different notation let us express (10) as

$$\chi^2(\theta) = \sum_{j=1}^{k} \frac{[n_j - Nq_j(\theta)]^2}{Nq_j(\theta)}, \tag{28}$$

where $Nq_j(\theta)$ is the expected frequency of the j-th class with $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ as the unknown parameter vector. We can write

$$\chi^2(\theta) = \sum_{j=1}^{k} \frac{n_j^2}{Nq_j(\theta)} - N. \tag{29}$$

Therefore, the minimum χ^2 estimates will be obtained by solving $\partial\chi^2(\theta)/\partial\theta = 0$, i.e., from

$$\sum_{j=1}^{k} \frac{n_j^2}{[Nq_j(\theta)]^2} \frac{\partial q_j(\theta)}{\partial \theta_l} = 0, \qquad l = 1, 2, \ldots, p. \tag{30}$$

This is Smith's (1916, p.264) system of equations (1). Since "these equations will generally be far too involved to be directly solved" she approximated these around MM estimates. Without going in that direction let us connect these equations to those from Fisher's (1912) ML equations. Since $\sum_{j=1}^{k} q_j(\theta) = 1$, we have $\sum_{j=1}^{k} \partial q_j(\theta)/\partial\theta_l = 0$, and hence from (30), the minimum χ^2 estimating equations are

$$\sum_{j=1}^{k} \frac{n_j^2 - [Nq_j(\theta)]^2}{[Nq_j(\theta)]^2} \frac{\partial q_j(\theta)}{\partial \theta_l} = 0, \qquad l = 1, 2, \ldots, p. \tag{31}$$

Under the multinomial framework, Fisher's likelihood function (27), denoted as $L(\theta)$ is

$$L(\theta) = N! \prod_{j=1}^{k} [(n_j)^{-1}] \prod_{j=1}^{k} [q_j(\theta)]^{n_j}. \tag{32}$$

Therefore, the log-likelihood function (22), denoted by $\ell(\theta)$, can be written as

$$\ln L(\theta) = \ell(\theta) = \text{ constant } + \sum_{j=1}^{k} n_j \ln q_j(\theta). \tag{33}$$

The corresponding ML estimating equations are $\partial\ell(\theta)/\partial\theta = 0$, i.e.,

$$\sum_{j=1}^{k} \frac{n_j}{q_j(\theta)} \frac{\partial q_j(\theta)}{\partial \theta_l} = 0, \tag{34}$$

i.e.,

$$\sum_{j=1}^{k} \frac{[n_j - Nq_j(\theta)]}{Nq_j(\theta)} \cdot \frac{\partial q_j(\theta)}{\partial \theta_l} = 0, \qquad l = 1, 2, \ldots, p. \tag{35}$$

Fisher (1924a) argued that the difference between (31) and (35) is of the factor $[n_j + Nq_j(\theta)]/Nq_j(\theta)$, which tends to value 2 for large values of N, and therefore, these two methods are asymptotically equivalent.[13] Some of Smith's (1916) numerical illustration showed improvement over MM in terms of goodness-of-fit values (in her notation P) when minimum χ^2 method was used. However, in her conclusion to the paper Smith (1916) provided a very lukewarm support for the minimum χ^2 method.[14] It is, therefore, not surprising that this method remained dormant for a while even after Neyman and Pearson (1928, pp.265-267) provided further theoretical justification. Neyman (1949) provided a comprehensive treatment of χ^2 method of estimation and testing. Berkson (1980) revived the old debate, questioned the sovereignty of MLE and argued that minimum χ^2 is the primary principle of estimation. However, the MLE procedure still remains as one of the most important principles of estimation and Fisher's idea of the likelihood plays the fundamental role in it. It can be said that based on his 1912 paper, Ronald Fisher was able to contemplate much broader problems later in his research that eventually culminated in his monumental paper in 1922. Because of the enormous importance of Fisher (1922) in the history of estimation, in the next section we provide a critical and detail analysis of this paper.

[13] Note that to compare estimates from two different methods Fisher (1924a) used the "estimating equations" rather than the estimates. Using the estimating equations (31) Fisher (1924a) also showed that $\chi^2(\hat{\theta})$ has $k-p-1$ degrees of freedom instead of $k-1$ when the $p \times 1$ parameter vector θ is replaced by its estimator $\hat{\theta}$. In Section 4 we will discuss the important role estimating equations play.

[14] Part of her concluding remarks was "...the present numerical illustrations appear to indicate that but little practical advantage is gained by a great deal of additional labour, the values of P are only slightly raised–probably always within their range of probable error. In other words the investigation justifies the method of moments as giving excellent values of the constants with nearly the maximum value of P or it justifies the use of the method of moments, if the definition of 'best' by which that method is reached must at least be considered somewhat arbitrary." Given that the time when MM was at its highest of popularity and Smith's position under Pearson's laboratory, it was difficult for her to make a strong recommendation for minimum χ^2 method [see also footnote 12].

3 Fisher's (1922) mathematical foundations of theoretical statistics and further analysis on MM and ML estimation

If we had to name the single most important paper on the theoretical foundation of statistical estimation theory, we could safely mention Fisher (1922).[15] The ideas of this paper are simply revolutionary. It introduced many of the fundamental concepts in estimation, such as, consistency, efficiency, sufficiency, information, likelihood and even the term "parameter" with its present meaning [see Stephen Stigler's comment on Savage (1976)].[16] Hald (1998, p.713) succinctly summarized the paper by saying: "For the first time in the history of statistics a framework for a frequency-based general theory of parametric statistical inference was clearly formulated."

In this paper (p.313) Fisher divided the statistical problems into three clear types:

> "(1) Problems of Specification. These arise in the choice of the mathematical form of the population.
> (2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivates, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.
> (3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known."

Formulation of the general statistical problems into these *three* broad categories was not really entirely new. Pearson (1902, p.266) mentioned the problems of (a) "choice of a suitable curve", (b) "determination of the constants" of the curve, "when the form of the curve has

[15] The intervening period between Fisher's 1912 and 1922 papers represented years in wilderness for Ronald Fisher. He contemplated and tried many different things, including joining the army and farming, but failed. However, by 1922, Fisher has attained the position of Chief Statistician at the Rothamsted Experimental Station. For more on this see Box (1978, ch.2) and Bera and Bilias (2000).

[16] Stigler (1976, p.498) commented that, "The point is that it is to Fisher that we owe the introduction of parametric statistical inference (and thus nonparametric inference). While there are other interpretations under which this statement can be defended, I mean it literally–Fisher was principally responsible for the introduction of the word "parameter" into present statistical terminology!" Stigler (1976, p.499) concluded his comment by saying "...for a measure of Fisher's influence on our field we need look no further than the latest issue of any statistical journal, and notice the ubiquitous "parameter." Fisher's concepts so permeate modern statistics, that we tend to overlook one of the most fundamental!"

been selected" and finally, (c) measuring the goodness-of-fit. Pearson had ready solutions for all three problems, namely, Pearson's family of distributions, MM and the test statistic (10), for (a), (b) and (c), respectively. As Neyman (1967, p.1457) indicated, Émile Borel also mentioned these three categories in his book, *Eléments de la Théorie des Probabilitiés*, 1909. Fisher (1922) did not dwell on the problem of *specification* and rather concentrated on the second and third problems, as he declared (p.315): "The discussion of theoretical statistics may be regarded as alternating between problems of estimation and problems of distribution."

After some general remarks about the then-current state of theoretical statistics, Fisher moved on to discuss some concrete criteria of estimation, such as, consistency, efficiency and sufficiency. Of these three, Fisher (1922) found the concept of "sufficiency" the most powerful to advance his ideas on the ML estimation. He defined "sufficiency" as (p.310): "A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated." Let t_1 be sufficient for θ and t_2 be any other statistic, then according to Fisher's definition

$$f(t_1, t_2; \theta) = f(t_1; \theta) f(t_2|t_1), \tag{36}$$

where $f(t_2|t_1)$ does not depend on θ. Fisher further assumed that t_1 and t_2 asymptotically follow bivariate normal (BN) distribution as

$$\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \sim BN\left[\begin{pmatrix} \theta \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right], \tag{37}$$

where $-1 < \rho < 1$ is correlation coefficient. Therefore,

$$E(t_2|t_1) = \theta + \rho\frac{\sigma_2}{\sigma_1}(t_1 - \theta) \quad \text{and} \quad V(t_2|t_1) = \sigma_2^2(1 - \rho^2). \tag{38}$$

Since t_1 is sufficient for θ, the distribution of $t_2|t_1$ should be free of θ and we should have $\rho\frac{\sigma_2}{\sigma_1} = 1$ i.e., $\sigma_1^2 = \rho^2\sigma_2^2 \leq \sigma_2^2$. In other words, the sufficient statistic t_1 is "efficient" [also see Geisser (1980, p.61), and Hald (1998, p.715)]. One of Fisher's aim was to establish that his MLE has *minimum* variance in large samples.

To demonstrate that the MLE has minimum variance, Fisher relied on two main steps. The first, as stated earlier, is that a "sufficient statistic" has the smallest variance. And for the second, Fisher (1922, p.330) showed that "the criterion of sufficiency is generally satisfied by the solution obtained by method of maximum likelihood ..." Without resorting to the central limit theorem, Fisher (1922, pp.327-329) proved the asymptotic normality of the MLE. For details on Fisher's proofs see Bera and Bilias (2000). However, Fisher's proofs are not satisfactory. He himself realized that and confessed (p.323): "I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect. Readers of the ensuing pages are invited to form their own opinion as to the possibility of the method of the maximum likelihood leading in any case to an insufficient statistic. For my own part I should gladly have withheld publication until a rigorously complete proof could have been formulated; but the number and variety of the new results which the method discloses press for publication, and at the same time I am not insensible of the advantage which accrues to Applied Mathematics from the co-operation of the Pure Mathematician, and this co-operation is not infrequently called forth by the very imperfections of writers on Applied Mathematics."

A substantial part of Fisher (1922) is devoted to the comparison of ML and MM estimates and establishing the former's superiority (pp.321-322, 332-337, 342-356), which he did mainly through examples. One of his favorite examples is the Cauchy distribution with density function

$$f(h;\theta) = \frac{1}{\pi}\frac{1}{[1+(y-\theta)^2]}, \qquad -\infty < y < \infty. \tag{39}$$

The problem is to estimate θ given a sample $y = (y_1, y_2, \ldots, y_n)'$. Fisher (1922, p.322) stated: "By the method of moments, this should be given by the first moment, that is by the mean of the observations: such would seem to be at least a good estimate. It is, however, entirely valueless. The distribution of the mean of such samples is in fact the same, identically, as that of a single observations." However, this is an unfair comparison. Since no moments exist for the Cauchy distribution, Pearson's MM procedure is just not applicable here.

Fisher (1922) performed an extensive analysis of the efficiency of ML and MM estimators for fitting distributions belonging to the Pearson (1895) family. Fisher (1922, p.355) concluded that the MM has an efficiency exceeding 80 percent only in the restricted region for which the kurtosis coefficient lies between 2.65 and 3.42 and the skewness measure does not exceed 0.1. In other words, only in the immediate neighborhood of the normal distribution, the MM will have high efficiency.[17] Fisher (1922, p.356) characterized the class of distributions for which the MM and ML estimators will be approximately the same in a simple and elegant way. The two

[17] Karl Pearson responded to Fisher's criticism of MM and other related issues in one of his very last papers, Pearson (1936), that opened with the italicized and striking line: *"Wasting your time fitting curves by moments, eh?"* Fisher felt compelled to give a frank reply immediately but waited until Pearson died in 1936. Fisher (1937, p.303) wrote in the opening section of his paper "...The question he [Pearson] raised seems to me not at all premature, but rather overdue." After his step by step rebutal to Pearson's (1936) arguments, Fisher (1937, p.317) placed Pearson's MM approach in statistical teaching as: "So long as 'fitting curves by moments' stands in the way of students' obtaining proper experience of these other activities, all of which require time and practice, so long will it be judged with increasing confidence to be waste of time." For more on this see Box (1978, pp.329-331). Possibly this was the temporary death nail for the MM. But after half a century, econometricians are finding that Pearson's moment matching approach to estimation is more useful than Fisher's ML method of estimation.

estimators will be identical if the derivative of the log-likelihood function has the following form

$$\frac{\partial \ell(\theta)}{\partial \theta} = a_0 + a_1 \sum_{i=1}^{n} y_i + a_2 \sum_{i=1}^{n} y_i^2 + a_3 \sum_{i=1}^{n} y_i^3 + a_4 \sum_{i=1}^{n} y_i^4 + \ldots\ldots \tag{40}$$

Therefore, we should be able to write the density of Y as[18]

$$f(y) \equiv f(y;\theta) = \mathcal{C} \exp[b_0 + b_1 y + b_2 y^2 + b_3 y^3 + b_4 y^4] \tag{41}$$

(keeping terms up to order 4), where $b_j (j = 0, 1, \ldots, 4)$ depend on θ and the constant $\mathcal{C}$ is such that the total probability is 1. Without any loss of generality, let us take $b_1 = 0$, then we have

$$\begin{aligned} \frac{d \ln f(y)}{dy} &= 2b_2 y + 3b_3 y^2 + 4b_4 y^3 \\ &= 2b_2 y \left(1 + \frac{3b_3 y}{2b_2} + \frac{2b_4 y^2}{b_2}\right). \end{aligned} \tag{42}$$

If b_3 and b_4 are small, i.e., when the density is sufficiently near to the normal curve, we can write (42) approximately as

$$\frac{d \ln f(y)}{dy} = 2b_2 y \left(1 - \frac{3b_3 y}{2b_2} - \frac{2b_4 y^2}{b_2}\right)^{-1}. \tag{43}$$

The form (43) corresponds to the Pearson (1895) general family of distributions. Therefore, the MM will have high efficiency within the class of Pearson family of distributions only when the density is close to the normal curve.

However, the optimality properties of MLE depend on the correct specification of the density function $f(y;\theta)$. Huber (1967), Kent (1982) and White (1982) analyzed the effect of misspecification on MLE [see also Bera (2000)]. Suppose the true density is given by $f^*(y)$ satisfying certain regularity conditions. We can define a distance between the two densities $f^*(y)$ and $f(y;\theta)$ by the continuous counterpart of $C(p,q)$ in (19), namely

$$C(f^*, f) = E_{f^*}[\ln\left(f^*(y)/f(y;\theta)\right)] = \int \ln\left(f^*(y)/f(y;\theta)\right) f^*(y) dy, \tag{44}$$

where $E_{f^*}[\cdot]$ denotes expectation under $f^*(y)$. Let θ^* be the value of θ that minimizes the distance $C(f^*, f)$. It can be shown that the quasi-MLE $\hat{\theta}$ converges to θ^*. If the model is

[18] Fisher (1922, p.356) started with this density function without any fanfare. However, this density has far-reaching implications. Note that (41) can be viewed as the continuous counterpart of entropy maximization solution discussed at the end of Section 1. That is, $f(y;\theta)$ in (41) maximizes the entropy $H(f) = -\int f(y) \ln f(y) dy$ subject to the moment conditions $E(y^j) = c_j, j = 1, 2, 3, 4$ and $\int f(y) dy = 1$. Therefore, this is a continuous version of the maximum entropy theorem of Jaynes (1957). A proof of this result can be found in Kagan, Linnik and Rao (1973, p.409) [see Gokhale (1975) and Mardia (1975) for its multivariate extension]. Urzúa (1988, 1997) further characterized the maximum entropy multivariate distributions and based on those devised omnibus tests for multivariate normality. Neyman (1937) used a density similar to (41) to develop his smooth goodness-of-fit test, and for more on this see Bera and Ghosh (2001).

correctly specified, i.e., $f^*(y) = f(y;\theta_0)$, for some θ_0, then $\theta^* = \theta_0$, and $\hat{\theta}$ is consistent for the true parameter.

To take account of possible misspecification, Choi, Hall and Presnell (2000) suggested a method of tilting the likelihood function $\ell(\theta) = \sum_{i=1}^n \ln f(y_i;\theta)$ by say, $\ell(\theta|\pi) = \sum_{i=1}^n \pi_i \ln f(y_i;\theta)$ where the π_i's are such that $\sum_{i=1}^n \pi_i = 1$. Therefore, tilting amounts to choosing unequal weights for different observations; in standard ML procedure $\pi_i = n^{-1}, i = 1, 2, \ldots, n$. We can consider the Cressie-Read power divergence measure $I_\lambda(\mathbf{i}/n, \pi)$ in (14) which provides a distance measure between $\pi = (\pi_1, \pi_2, \ldots, \pi_n)'$ and $\mathbf{i}/n = (n^{-1}, n^{-1}, \ldots, n^{-1})'$ and can set it at a given level say, $\delta(> 0)$. Then, the estimation procedure would be to maximize $\ell(\theta|\pi)$ subject to $I_\lambda(\mathbf{i}/n, \pi) = \delta$ and $\sum_{i=1}^n \pi_i = 1$. The appropriate lagrangian function for this problem would be

$$\mathcal{L} = \sum_{i=1}^n \pi_i \ln f(y_i;\theta) + \tau_1 \left(I_\lambda(\mathbf{i}/n, \pi) - \delta\right) + \tau_2 \left(\sum_{i=1}^n \pi_i - 1\right), \tag{45}$$

where τ_1 and τ_2 are two Lagrange multipliers. Simulation results reported in Choi et al. (2000) indicates that such estimators have improved robustness properties.

The optimality of MLE rests on the assumption that *we know the true underlying density function.* Of course, in econometrics that is rarely the case. Therefore, it is not surprising that recently econometricians are finding the generalized MM (GMM) procedure, which is an extension of Pearson's MM approach, more attractive. A similar venue was initiated around 1960's in the statistics literature with the estimating functions (EF) approach to optimal estimation. In some sense, the developments in statistics and econometrics has come full circle – after discarding the moment approach in favor of Fisher's maximum likelihood for more than a half century, Karl Pearson's century-old techniques are now found to be more useful in a world with limited information. When we discuss the GMM approach we also present its relation to minimum χ^2-method.

4 Estimating Functions

Durbin (1960) appears to be the first instance of the modern use of "estimating equations" in econometrics. Durbin's treatment was amazingly complete. At that time nothing was known, not even asymptotically, about the sampling distributions of the least squares estimators in the presence of lagged dependent variables among the predictors; a case where the finite sample Gauss-Markov theorem is inapplicable. Durbin observed that the equations from which we obtain the estimators as their roots, evaluated at the true parameters, preserve their unbiasedness. Thus, it is natural to expect some sort of optimality properties to be associated with the least squares estimators in this case as well!

To illustrate, let us consider the $AR(1)$ model for y_t

$$y_t = \theta y_{t-1} + u_t; \qquad u_t \sim iid(0, \sigma^2), \qquad t = 1, \ldots, n.$$

The least squares estimator of θ, $\hat{\theta} = \sum y_t y_{t-1} / \sum y_{t-1}^2$ is the root of the equation

$$g(y, \theta) = \sum y_t y_{t-1} - \theta \sum y_{t-1}^2 = 0, \tag{46}$$

where y denotes the sample data. The function $g(y, \theta)$ in (46) is linear in the parameter θ and $E[g(y, \theta)] = 0$. Durbin (1960) termed $g(y, \hat{\theta}) = 0$ an *unbiased linear estimating equation.* Such a class of estimating functions can be denoted by

$$g(y, \theta) = T_1(y) + \theta T_2(y), \tag{47}$$

where $T_1(y)$ and $T_2(y)$ are functions of the data only. Then, Durbin proceeded to define a minimum variance requirement for the unbiased linear estimating function reminiscent of the Gauss-Markov theorem. As it is possible to change its variance by multiplying with an arbitrary constant without affecting the estimator, it seems proper to standardize the estimating function by dividing through by $E(T_2(y))$, which is not but $E(\partial g / \partial \theta)$.

Durbin (1960) turned into the study of estimating functions as a means of studying the least squares estimator itself. In Durbin's context, let $t_1 = T_1(y)/E(T_2(y))$, $t_2 = T_2(y)/E(T_2(y))$, and write

$$g_s(y, \theta) = t_1 + \theta t_2$$

for the standardized linear estimating function. For the root of the equation $t_1 + \hat{\theta} t_2 = 0$, we can write:

$$t_2(\hat{\theta} - \theta) = -(t_1 + \theta t_2) \tag{48}$$

which indicates that the sampling error of the estimator depends on the properties of the estimating function. Clearly, it is more convenient to study a linear function and then transfer its properties to its nonlinear root instead of studying directly the estimator itself. Durbin used representation (48) to study the asymptotic properties of the least squares estimator when lagged dependent variables are included among the regressors. More generally, a first-order Taylor series expansion of $g(\hat{\theta}) = 0$ around θ,

$$n^{1/2}(\hat{\theta} - \theta) \approx -n^{-1/2} g(\theta) \times \left(n^{-1} \frac{\partial g}{\partial \theta} \right)^{-1} \approx -n^{1/2} g(\theta) \times \left[E \left(\frac{\partial g}{\partial \theta} \right) \right]^{-1},$$

indicates that in order to obtain an estimator with minimum limiting variance, the estimating function g has to be chosen with minimum variance of its standardized form, $g(\theta) \times [E(\partial g(\theta)/\partial \theta)]^{-1}$.

Let $\mathcal{G}$ denote a class of unbiased estimating functions. Let

$$g_s = \frac{g}{E[\partial g/\partial\theta]} \tag{49}$$

be the standardized version of the estimating function $g \in \mathcal{G}$. We will say that g^* is the *best unbiased estimating function* in the class $\mathcal{G}$, if and only if, its standardized version has minimum variance $Var(g_s^*)$ in the class; that is

$$Var(g_s^*) \leq Var(g_s) \qquad \text{for all other } g \in \mathcal{G}. \tag{50}$$

This optimality criterion was suggested by Godambe (1960) for general classes of estimating functions and, independently, by Durbin (1960) for linear classes of estimating functions defined in (47). The motivation behind the *Godambe-Durbin criterion* (50) as a way of choosing from a class of unbiased estimating functions is intuitively sound: it is desirable that g is as close as possible to zero when it is evaluated at the true value θ which suggests that we want $Var(g)$ to be as small as possible. At the same time we want any deviation from the true parameter to lead g as far away from zero as possible, which suggests that $[E(\partial g/\partial\theta)]^2$ be large. These two goals can be accomplished simultaneously with the Godambe-Durbin optimality criterion that minimizes the variance of g_s in (49).

These developments can be seen as an analog of the Gauss-Markov theory for estimating equations. Moreover, let $\ell(\theta)$ denote the loglikelihood function and differentiate $E[g(y,\theta)] = 0$ to obtain $E[\partial g/\partial\theta] = -Cov[g, \partial\ell(\theta)/\partial\theta]$. Squaring both sides of this equality and applying the Cauchy-Schwarz inequality yields the Cramér-Rao type inequality

$$\frac{E(g^2)}{[E(\partial g/\partial\theta)]^2} \geq \frac{1}{E\left[(\partial\ell(\theta)/\partial\theta)^2\right]}, \tag{51}$$

for estimating functions. Thus, a lower bound, given by the inverse of the information matrix of the true density $f(y;\theta)$, can be obtained for $Var(g)$. Based on (51), Godambe (1960) concluded that *for every sample size, the score $\partial\ell(\theta)/\partial\theta$ is the optimum estimating function.* Concurrently, Durbin (1960) derived this result for the class of linear estimating equations. It is worth stressing that it provides an exact or *finite sample* justification for the use of maximum likelihood estimation. Recall that the usual justification of the likelihood-based methods, when the starting point is the properties of the estimators, is asymptotic, as discussed in Section 3. A related result is that the optimal estimating function within a class $\mathcal{G}$ has maximum correlation with the true score function; cf. Heyde (1997). Since the true loglikelihood is rarely known, this result is more useful from practical point of view.

Many other practical benefits can be achieved when working with the estimating functions instead of the estimators. Estimating functions are much easier to combine and they are invariant under one-to-one transformations of the parameters. Finally, the approach preserves the spirit of the method of moments estimation and it is well suited for semiparametric models by requiring only assumptions on a few moments.

Example 1: The estimating function (46) in the $AR(1)$ model can be obtained in an alternative way that sheds light to the distinctive nature of the theory of estimating functions. Since $E(u_t) = 0$, $u_t = y_t - \theta y_{t-1}$, $t = 1, 2, \ldots, n$ are n *elementary* estimating functions. The issue is how we should combine the n available functions to solve for the scalar parameter θ. To be more specific, let h_t be an elementary estimating function of $y_1, y_2, \ldots, y_t$ and θ, with $E_{t-1}(h_t) \equiv E(h_t|y_1, \ldots, y_{t-1}) = 0$. Then, by the law of iterated expectations $E(h_s h_t) = 0$ for $s \neq t$. Consider the class of estimating functions g

$$g = \sum_{t=1}^{n} a_{t-1} h_t$$

created by using different weights a_{t-1}, which are dependent only on the conditioning event (here, $y_1, \ldots, y_{t-1}$). It is clear that, due to the law of iterated expectations, $E(g) = 0$. In this class, Godambe (1985) showed that the quest for minimum variance of the standardized version of g yields the formula for the optimal weights as

$$a^*_{t-1} = \frac{E_{t-1}\left(\partial h_t / \partial \theta\right)}{E_{t-1}(h_t^2)}. \tag{52}$$

Application of (52) in the $AR(1)$ model gives $a^*_{t-1} = -y_{t-1}/\sigma^2$, and so the optimal estimating equation for θ is

$$g^* = \sum_{t=1}^{n} y_{t-1}(y_t - \theta y_{t-1}) = 0.$$

In summary, even if the errors are not normally distributed, the least squares estimating function has some optimum qualifications once we restrict our attention to a particular class of estimating functions.

Example 1 can be generalized further. Consider the scalar dependent variable y_i with expectation $E(y_i) = \mu_i(\theta)$ modeled as a function of a $p \times 1$ parameter vector θ; the vector of explanatory variables $x_i, i = 1, \ldots, n$ is omitted for convenience. Assume that the y_i's are independent with variances $Var(y_i)$ possibly dependent on θ. Then, the optimal choice within the class of $p \times 1$ unbiased estimating functions $\sum_i a_i(\theta)[y_i - \mu_i(\theta)]$ is given by

$$g^*(\theta) = \sum_{i=1}^{n} \frac{\partial \mu_i(\theta)}{\partial \theta} \frac{1}{Var(y_i)} [y_i - \mu_i(\theta)], \tag{53}$$

which involves only assumptions on the specification of the mean and variance (as it is the case with the Gauss-Markov theorem). Wedderburn (1974) noticed that (53) is very close to the true score of all the distributions that belong to the exponential family. In addition, $g^*(\theta)$ has properties similar to those of a score function in the sense that,

1. $E[g^*(\theta)] = 0$ and,

2. $E[g^*(\theta)g^*(\theta)'] = -E[\partial g^*(\theta)/\partial\theta']$.

Wedderburn called the integral of g^* in (53) "quasi-likelihood," the equation $g^*(\theta) = 0$ "quasi-likelihood equation" and the root $\hat{\theta}$ "maximum quasi-likelihood estimator." Godambe and Heyde (1987) obtained (53) as an optimal estimating function and assigned the name "quasi-score." This is a more general result in the sense that for its validity we do not need to assume that the true underlying distribution belongs to the exponential family of distributions. The maximum correlation between the optimal estimating function and the true unknown score justifies their terminology for $g^*(\theta)$ as a "quasi-score."

Example 2: Let y_i, $i = 1, \ldots, n$ be independent random variables with $E(y_i) = \mu_i(\theta)$ and $Var(y_i) = \sigma_i^2(\theta)$, where θ is a scalar parameter. The quasi-score approach suggests that in the class of linear estimating functions we should solve

$$g^*(\theta) = \sum_{i=1}^{n} \frac{[y_i - \mu_i(\theta)]}{\sigma_i^2(\theta)} \frac{\partial \mu_i(\theta)}{\partial \theta} = 0. \tag{54}$$

Under the assumption of normality of y_i the maximum likelihood equation

$$\frac{\partial \ell(\theta)}{\partial \theta} = g^*(\theta) + \frac{1}{2}\sum_{i=1}^{n} \frac{[y_i - \mu_i(\theta)]^2}{\sigma_i^4(\theta)} \frac{\partial \sigma_i^2(\theta)}{\partial \theta} - \frac{1}{2}\sum_{i=1}^{n} \frac{1}{\sigma_i^2(\theta)} \frac{\partial \sigma_i^2(\theta)}{\partial \theta} = 0, \tag{55}$$

is globally optimal and the estimation based on the quasi-score (54) is inferior. If one were unwilling to assume normality, one could claim that the weighted least squares approach that minimizes $\sum_i [y_i - \mu_i(\theta)]^2/\sigma_i^2(\theta)$ and yields the estimating equation

$$w(\theta) = g^*(\theta) + \frac{1}{2}\sum_{i=1}^{n} \frac{[y_i - \mu_i(\theta)]^2}{\sigma_i^4(\theta)} \frac{\partial \sigma_i^2(\theta)}{\partial \theta} = 0 \tag{56}$$

is preferable. However, because of the dependence of the variance on θ, (56) delivers an inconsistent root, in general; see Crowder (1986) and McLeish (1984). The application of a law of large numbers shows that $g^*(\theta)$ is stochastically closer to the score (55) than is $w(\theta)$. In a way, the second term in (56) creates a bias in $w(\theta)$, and the third term in (55) "corrects" for this bias in the score equation. Here, we have a case in which the extremum estimator (weighted least squares) is inconsistent, while the root of the quasi-score estimating function is consistent and optimal within a certain class of estimating functions.

The theory of estimating functions has been extended to numerous other directions. For extensions to dependent responses and optimal combination of higher-order moments, see Heyde (1997) and the citations therein. For various applications see the volume edited by Godambe (1991) and Vinod (1998). Li and Turtle (2000) offered an application of the EF approach in the ARCH models.

5 Modern Approaches to Estimation in Econometrics

Pressured by the complicity of econometric models, econometricians looked for methods of estimation that bypass the use of likelihood. This trend was also supported by the nature of economic theories that do provide characterization of the stochastic laws only in terms of moment restrictions; see Hall (2001) for specific examples. In the following we describe the instrumental variables (IV) estimator, the generalized method of moments (GMM) estimator and some recent extensions of GMM based on empirical likelihood. Often, the first instances of these methods appear in the statistical literature. However, econometricians, faced with the challenging task of estimating economic models, also offered new and interesting twists. We start with Sargan's (1958) IV estimation.

The use of IV was first proposed by Reiersøl (1941, 1945) as a method of *consistent* estimation of linear relationships between variables that are characterized by measurement error.[19] Subsequent developments were made by Geary (1948, 1949) and Durbin (1954). A definite treatment of the estimation of linear economic relationships using instrumental variables was presented by Sargan (1958, 1959).

In his seminal work, Sargan (1958) described the IV method as it is used currently in econometrics, derived the asymptotic distribution of the estimator, solved the problem of utilizing more instrumental variables than the regressors to be instrumented and discussed the similarities of the method with other estimation methods available in the statistical literature. This general theory was applied to time series data with autocorrelated residuals in Sargan (1959).

In the following we will use x to denote a regressor that is possibly subjected to measurement error and z to denote an instrumental variable. It is assumed that a linear relationship holds,

$$y_i = \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + u_i; \qquad i = 1, \ldots, n, \tag{57}$$

[19] The history of IV estimation goes further back. In a series of work Wright (1920, 1921, 1925, 1928) advocated a graphical approach to estimate simultaneous equation systems what he referred to as "path analysis." In his 1971 Schultz Lecture, Goldberger (1972, p.938) noted: "Wright drew up a flow chart, … and read off the chart a set of equations in which zero covariances are exploited to express moments among observable variables in terms of structural parameters. In effect, he read off what we might now call instrumental-variable estimating equations." For other related references see Manski (1988, p.xii)

where the residual u in (57) includes a linear combination of the measurement errors in x's. In this case, the method of least squares yields inconsistent estimates of the parameter vector. Reiersøl's (1941, 1945) method of obtaining consistent estimates is equivalent to positing a zero sample covariance between the residual and each instrumental variable. Then, we obtain p equations, $\sum_{i=1}^{n} z_{ki}u_i = 0, k = 1, \ldots, p$, on the p parameters; the number of available instrumental variables is assumed to be equal to the number of regressors.

It is convenient to express Reiersøl's methodology in matrix form. Let Z be a $(n \times q)$ matrix of instrumental variables with kth column $z_k' = (z_{k1}, z_{k2} \ldots, z_{kn})'$, and for the time being we take $q = p$. Likewise X is the $(n \times p)$ matrix of regressors, u the $n \times 1$ vector of residuals and y the $n \times 1$ vector of responses. Then, the p estimating equations can be written more compactly as

$$Z'u = 0 \qquad \text{or} \qquad Z'(y - X\beta) = 0, \tag{58}$$

from which the *simple* IV estimator, $\hat{\beta} = (Z'X)^{-1}Z'y$, follows. Under the assumptions that (i) $plim\, n^{-1}Z'Z = \Sigma_{ZZ}$, (ii) $plim\, n^{-1}Z'X = \Sigma_{ZX}$, and (iii) $n^{-1/2}Z'u$ tends in distribution to a normal vector with mean zero and covariance matrix $\sigma^2\Sigma_{ZZ}$, it can be concluded that the simple IV estimator $\hat{\beta}$ has a normal limiting distribution with covariance matrix $Var(n^{1/2}\hat{\beta}) = \sigma^2\Sigma_{ZX}^{-1}\Sigma_{ZZ}\Sigma_{XZ}^{-1}$ [see Hall (1993)].

Reiersøl's method of obtaining consistent estimates can be classified as an application of the Pearson's MM procedure.[20] Each instrumental variable z induces the moment restriction $E(zu) = 0$. In Reiersøl's work, which Sargan (1958) developed rigorously, the number of these moment restrictions that the data satisfy are equal to the number of unknown parameters. A major accomplishment of Sargan (1958) was the study of the case in which the number of available instrumental variables is greater than the number of regressors to be instrumented.

[20] There is also a connection between the IV and the limited information maximum likelihood (LIML) procedures which generally is not mentioned in the literature. Interviewed in 1985 for the first issue of the *Econometric Theory*, Dennis Sargan recorded this connection as his original motivation. Around 1950s Sargan was trying to put together a simple Klein-type model for the UK, and his attempt to estimate this macroeconomic model led him to propose the IV method of estimation as he stated [see Phillips (1985, p.123)]: "At that stage I also became interested in the elementary methods of estimation and stumbled upon the method of instrumental variables as a general approach. I did not become aware of the Cowles Foundation results, particularly the results of Anderson and Rubin on LIML estimation until their work was published in the late 1940s. I realized it was very close to instrumental variable estimation. The article which started me up was the article by Geary which was in the JRSS in 1948. That took me back to the earlier work by Reiersøl, and I pretty early realized that the Geary method was very close to LIML except he was using arbitrary functions of time as the instrumental variables, particularly polynomials in the time variable. One could easily generalize the idea to the case, for example, of using lagged endogenous variables to generate the instrumental variables. That is really where my instrumental variable estimation started from."

Any subset of p instrumental variables can be used to to form p equations for the consistent estimation of the p parameters. Sargan (1958, pp.398-399) proposed that, p linear combinations of the instrumental variables are constructed with the weights chosen so as to minimize the covariance matrix of asymptotic distribution. More generally, if the available number of moment restrictions exceeds that of the unknown parameters, p (optimal) linear combinations of the moments can be used for estimation. This idea pretty much paved the way for the more recent advances of the GMM approach in the history of econometric estimation and the testing of overidentifying restrictions. As we show earlier, the way that the EF approach to optimal estimation proceeds is very similar.

Specifically, suppose that there are $q(> p)$ instrumental variables from which we construct a reduced set of p variables z^*_{ki},

$$z^*_{ki} = \sum_{j=1}^{q} \alpha_{kj} z_{ji}, \qquad i = 1, \ldots, n, \qquad k = 1, \ldots, p.$$

Let us, for ease of notation, illustrate the solution to this problem using matrices. For any $(q \times p)$ weighting matrix $\alpha = [\alpha_{kj}]$ with $rank[\alpha] = p$, $Z^* = Z\alpha$ produces a new set of p instrumental variables. Using Z^* in the formula of simple IV estimator we obtain an estimator of β with asymptotic distribution

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2(\alpha'\Sigma_{ZX})^{-1}(\alpha'\Sigma_{ZZ}\alpha)(\Sigma_{XZ}\alpha)^{-1}). \tag{59}$$

The next question is, which matrix α yields the minimum asymptotic variance-covariance matrix. It can be shown that the optimal linear combination of the q instrumental variables is the one that maximizes the correlation of Z^* with the regressors X. Since the method of least squares maximizes the squared correlation between the dependent variable and its fitted value, the *optimal* $(q \times p)$ matrix $\alpha = [\alpha_{kj}]$ is found to be $\alpha_0 = \Sigma^{-1}_{ZZ}\Sigma_{ZX}$, for which it is checked that

$$\begin{aligned} 0 &\leq [\alpha(\Sigma_{XZ}\alpha)^{-1} - \alpha_0(\Sigma_{XZ}\alpha_0)^{-1}]'\Sigma_{ZZ}[\alpha(\Sigma_{XZ}\alpha)^{-1} - \alpha_0(\Sigma_{XZ}\alpha_0)^{-1}] \\ &= (\alpha'\Sigma_{ZX})^{-1}\alpha'\Sigma_{ZZ}\alpha(\Sigma_{XZ}\alpha)^{-1} - (\alpha_0'\Sigma_{ZX})^{-1}\alpha_0'\Sigma_{ZZ}\alpha_0(\Sigma_{XZ}\alpha_0)^{-1}. \end{aligned} \tag{60}$$

In practice, α_0 is consistently estimated by $\hat{\alpha}_0 = (Z'Z)^{-1}Z'X$ and the proposed optimal instrumental matrix is consistently estimated by the fitted value from regressing X on Z,

$$\hat{Z}^*_0 = Z\hat{\alpha}_0 = Z(Z'Z)^{-1}Z'X \equiv \hat{X}. \tag{61}$$

Then, the instrumental variables estimation methodology applies as before with $\hat{Z}^*_0 \equiv \hat{X}$ as the set of instrumental variables. The set of estimating equations (58) is replaced by $\hat{X}'u = 0$ which in turn yields

$$\hat{\beta} = (\hat{X}'X)^{-1}\hat{X}'y \tag{62}$$

with covariance matrix of the limiting distribution $Var(n^{1/2}\hat{\beta}) = \sigma^2(\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX})^{-1}$. This estimator is termed *generalized IV estimator* as it uses the $q(> p)$ available moments $E(z_j u) = 0$ $j = 1, 2, \ldots, q$ for estimating the p-vector parameter β. While Sargan (1958) combined the q instrumental variables to form a new composite instrumental variable, the same result can be obtained using the sample analogs of q moments themselves. This brings us back to the minimum χ^2 and the GMM approaches to estimation.

The minimum χ^2 method and different variants are reviewed in Ferguson (1958). His article also includes a new method of generating best asymptotically normal (BAN) estimates. These methods are extremely close to what is known today in econometrics as the GMM. A recent account is provided by Ferguson (1996). To illustrate, let us start with a q-vector sample statistics T_n which is asymptotically normally distributed with $E(T_n) = P(\theta)$ and covariance matrix V. The minimization of the quadratic form

$$Q_n(\theta) = n[T_n - P(\theta)]' M [T_n - P(\theta)], \tag{63}$$

with respect to θ, was termed as *minimum chi-square (χ^2) estimation.* In (63), the $q \times q$ matrix M has the features of a covariance matrix that may or may not depend on θ and it is not necessarily equal to V^{-1}.

Clearly, the asymptotic variance of the minimum χ^2 estimates $\hat{\theta}$ depends on the choice of the weighting matrix M. Let $\dot{P}$ be the $q \times p$ matrix of first partial derivatives of $P(\theta)$, where for convenience we have omitted the dependence on θ. It can be shown $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma)$, where

$$\Sigma = [\dot{P}'M\dot{P}]^{-1}\dot{P}'MVM\dot{P}[\dot{P}'M\dot{P}]^{-1}. \tag{64}$$

Assuming that V is nonsingular, $M = V^{-1}$ is the optimal choice and the resulting asymptotic variance $\Sigma = [\dot{P}'V^{-1}\dot{P}]^{-1}$ is minimum.

To reduce the computational burden and add flexibility in the generation of estimates that are BAN, Ferguson (1958) suggested getting the estimates as roots of linear forms in certain variables. Assuming that M does not depend on θ, the first order conditions of the minimization problem (63)

$$-n\dot{P}'M[T_n - P(\theta)] = 0 \tag{65}$$

are linear combinations of the statistic $[T_n - P(\theta)]$ with weights $\dot{P}'M$. Based on this observation, Ferguson (1958) considered equations using weights of general form. He went on to study the weights that produce BAN estimates. This is reminiscent of the earlier EF approach but also of the GMM approach that follows.

Example: (Optimality of the Pearson χ^2 statistic in estimation) Consider the multinomial probabilistic experiment with k classes and with probability of j-th class being $q_j(\theta)$, $j =$

$1, \ldots, k$; the dependence of each class probability on the unknown p-vector parameter θ is made explicit. In this case the matrix V is given by

$$V = D_0 - QQ', \tag{66}$$

where $D_0 = diag\{q_1(\theta), q_2(\theta), \ldots, q_k(\theta)\}$ and $Q' = (q_1(\theta), q_2(\theta), \ldots, q_k(\theta))$. Ferguson (1958) shows that if there is a nonsingular matrix V_0 such that

$$VV_0\dot{P} = \dot{P}, \tag{67}$$

then the asymptotic variance of $\hat{\theta}$ takes its minimum when the weights are given by the matrix $\dot{P}'V_0$ and the minimum value is $\Sigma = [\dot{P}'V_0\dot{P}]^{-1}$.

In this case, V in (66) is singular, but $D_0^{-1} = diag\{1/q_1(\theta), 1/q_2(\theta), \ldots, 1/q_k(\theta)\}$ satisfies condition (67). Indeed, setting $V_0 = D_0^{-1}$ we get

$$VD_0^{-1}\dot{Q} = \dot{Q}, \tag{68}$$

where $\dot{Q}$ is a $k \times p$ matrix of derivatives of Q with respect to θ. To see why the last equality holds true, note that the $1 \times p$ vector $Q'D_0^{-1}\dot{Q}$ has typical element $\sum_{j=1}^{k} \partial q_j(\theta)/\partial\theta_i$, which should be zero by the fact that $\sum_j q_j(\theta) = 1$. Pearson χ^2 statistic uses precisely $V_0 = D_0^{-1}$ and therefore the resulting estimators are efficient [see equation (11)]. Pearson χ^2 statistic is an early example of the optimal GMM approach to estimation.

Ferguson (1958) reformulated the minimum χ^2 estimation problem in terms of estimating equations. By interpreting the first order conditions of the problem as a linear combination of elementary estimating functions, he realized that the scope of consistent estimation can be greatly enhanced. In econometrics, Goldberger (1968, p.4) suggested the use of "analogy principle of estimation" as a way of forming the normal equations in the regression problem. However, the use of this principle (or MM) instead of the least squares was not greatly appreciated at that time.[21] It was only with GMM that econometricians gave new force to the message delivered by Ferguson (1958) and realized that IV estimation can be considered as a special case of a more general approach. Hansen (1982) looked on the problem of consistent estimation from the

[21] In reviewing Goldberger (1968), Katti (1970) commented on the analogy principle as: "This is cute and will help one remember the normal equations a little better. There is no hint in the chapter–and I believe no hints exist–which would help locate such criteria in other situations, for example, in the problem of estimating the parameters in the gamma distribution. Thus, this general looking method has not been shown to be capable of doing anything that is not already known. Ad hoc as this method is, it is of course difficult to prove that the criteria so obtained are unique or optimal." Manski (1988, p.ix) acknowledged that he became aware of Goldberger's analogy principle only in 1984 when he started writing his book.

MM point of view.[22] For instance, assuming that the covariance of each instrumental variable with the residual is zero provided that the parameter is at its true value, the sample analog of the covariance will be a consistent estimator of zero. Therefore, the root of the equation that equates sample covariance to zero should be close to the true parameter. On the contrary, in the case of regressors subjected to measurement errors, the normal equations $X'u = 0$ will fail to produce consistent estimates because the corresponding population moment condition is not zero.

Let $h(y; \theta)$ be a general $(q \times 1)$ function of data y and p-vector parameter θ, with the property that for $\theta = \theta_0$

$$E[h(y; \theta_0)] = 0. \tag{69}$$

We say that the function $h(y; \theta)$ satisfies a *moment* or an *orthogonality* condition that holds true only when θ is at its true value θ_0. Alternatively, there are q moment conditions that can be used to determine an estimate of the p-vector θ_0. If $q = p$ then

$$g(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} h(y_i; \theta) = 0$$

will yield roots $\hat{\theta}$ that are unique parameter estimates.

In the general case, in which $q \geq p$, one can utilize Sargan's idea and construct p "composite" moment conditions by creating p linear combinations of those initial ones. More specifically, Hansen proposed as GMM class of estimators the minimizers of the quadratic form

$$g(\theta)' A g(\theta), \tag{70}$$

where A is a $(q \times q)$ positive definite weighting matrix. The minimizer of (70) satisfies the p equations of the first order conditions:

$$\left[\frac{\partial g(\theta)}{\partial \theta'}\right]' A g(\theta) = 0. \tag{71}$$

That is, by applying GMM the estimation is based on p linear combinations of the moment conditions with weights given by the $p \times q$ matrix $[\partial g(\theta)/\partial \theta']' A$. The matrix A can be chosen from an efficiency standpoint of view. Here, Hansen (1982) followed Sargan's way of choosing

[22] From our conversation with Lars Hansen and Christopher Sims, we learned that a rudimentary form of GMM estimator existed in Sims' lecture notes. Hansen attended Sims' lectures as a graduate student at the University of Minnesota. Hansen formally wrote up the theories of GMM estimator and found its close connection to Sargan's (1958) IV estimator. To differentiate his paper, Hansen used a very general framework and established the properties of the estimator under mild assumptions. In the econometrics literature, Hansen (1982) has been the most influential paper to popularize the moment type and minimum χ^2 estimation techniques.

linear combinations that minimize the covariance matrix of asymptotic distribution of the GMM estimators. The result for the optimal A is known by now from Ferguson (1958) and the EF approach discussed earlier: within the GMM class of estimators, we obtain the efficient one for $A = S^{-1}$, the inverse of the covariance matrix of the moment conditions. Likewise, the efficient GMM estimator has covariance matrix equal to $[g_0' S^{-1} g_0]^{-1}$, where $g_0 = E[\partial h(y; \theta_0)/\partial\theta]$. In practice, S is unknown and the GMM estimator needs a first step for the calculation of a consistent estimate of S. Of course, the finite sample properties of the GMM estimator depend on the estimate of S.

Hansen's GMM has particular appeal to the economists who deal with a variety of moment or *orthogonality* conditions derived from the theoretical properties of their postulated economic models. Highly complex economic models make it difficult to write down a tractable likelihood function. Then, the GMM approach provides a way to estimate model parameters consistently. The methodology is very similar to the EF approach developed in the statistical literature. However, the initial development of EF approach concerned with the optimal estimation. In general, the use of optimal weighting matrices is impossible due to the presence of unknown parameters; Newey (1993) has discussion on the feasible approaches to the efficient estimation. Hansen's GMM allows any weighting matrix $[\partial g(\theta)/\partial\theta']' A$, thereby offering more flexibility in practice. Moreover, the use of the objective function (70) offers a selection criterion among multiple roots, a practical problem that is associated with the EF approach.

To see the connection of Hansen's (1982) GMM estimation and Sargan's (1958) IV estimation [see also Hall (1993)], in the context of the classical linear regression model, with notation similar to that utilized in the description of IV estimation, consider the q *orthogonality (moment) conditions*

$$E(z_k u) = 0 \qquad k = 1, 2, \ldots, q.$$

Using the GMM methodology, minimization of

$$\left(\frac{Z'u}{n}\right)' A \left(\frac{Z'u}{n}\right)$$

with respect to β, in view of $u = y - X\beta$, yields first order conditions

$$-2X'ZAZ'(y - X\beta) = 0. \tag{72}$$

If $q = p$, then (72) reduces to (58) which yields the simple IV estimator. Clearly, in case the orthogonality conditions are equal to the number of unknown parameters, the β is just identified and the choice of the weighting matrix A is inconsequential.

If $q > p$, the $p \times q$ matrix $X'ZA$ in (72) transforms the q (sample) moments $Z'(y - X\hat{\beta})$ into p linear combinations that can be used to produce a unique solution for $\hat{\beta}$. Now, different choices of A will yield consistent estimates with different asymptotic variances. The discussion earlier suggests that, in order to attain efficiency with the given set of q orthogonality conditions, beyond a constant of proportionality, we should pick $A = \Sigma_{ZZ}^{-1}$; in practice the consistent estimate $(n^{-1}Z'Z)^{-1}$ will be utilized. With that choice for A, the GMM estimator is identical to Sargan's generalized IV estimator (62). Now we turn our discussion to empirical likelihood (EL)-based estimation that has recently become popular in econometrics.

In practice, inference based on the GMM as suggested by Hansen (1982) and its asymptotically equivalent forms suffers from poor finite sample properties; see for instance, Hansen, Heaton and Yaron (1996) and, Pagan and Robertson (1997). The EL method proposed by Owen (1988) has been fruitfully adapted to offer an alternative approach to inference in econometric models: the empirical likelihood is maximized subject to the moment restrictions that data satisfy in the population. Effectively, these methods impose the moment restrictions by appropriately reweighting the data. The resulting estimator of the unknown parameter is first-order asymptotically equivalent to the efficient GMM estimator. Moreover, the sampling properties of the estimators are similar to those supported by bootstrap, thus offering an increased chance of finite-sample properties that are in accordance with asymptotic theory.

Let θ be the unknown parameter of primary interest. The moment condition $E[h(y;\theta)] = 0$ holds at θ_0, the true value of the parameter. Qin and Lawless (1994) proposed estimating the parameter θ by solving the problem:

$$\max_{\pi_i,\theta} \sum_{i=1}^{n} ln\pi_i \qquad \text{s.t.} \qquad \sum_{i=1}^{n} \pi_i = 1, \qquad \sum_{i=1}^{n} \pi_i h(y_i;\theta) = 0, \tag{73}$$

where each π_i is restricted to be non-negative. This is equivalent to minimizing the entropy measure (18) satisfying given moment conditions. The restriction $\sum_i \pi_i h(y_i;\theta) = 0$ can be interpreted as the empirical analog of the moment condition with π_i being the probability of the random variable y to take the value y_i. This empirical analog differs from the one that is used in the classical MM (or Hansen's GMM) in that in the latter we assume $\pi_i = n^{-1}$ for every observation y_i. In the absence of the moment restriction $\sum_i \pi_i h(y_i;\theta) = 0$ the above optimization program is the non-parametric maximum log-likelihood that has the solution $\hat{\pi}_i = n^{-1}$; each sample observation is equally weighted. In the presence of the constraint the differentiation of the lagrangian function

$$\mathcal{L} = \sum_{i=1}^{n} ln\pi_i + \mu \left(1 - \sum_{i=1}^{n} \pi_i\right) - n\tau' \sum_{i=1}^{n} \pi_i h(y_i;\theta) \tag{74}$$

yields estimated probabilities

$$\hat{\pi}_{el,i} = \frac{1}{n[1 + \hat{\tau}_{el}' h(y_i;\hat{\theta}_{el})]}. \tag{75}$$

The estimated probabilities (75) differ from the uniform distribution, which assigns to each observation equal weight n^{-1}, to the degree that the Lagrange multiplier vector of the moment restriction $\hat{\tau}_{el}$, or tilting parameter, is away from zero or to the degree that the moment restriction hold true. The EL approach for parameter estimation works by finding the multinomial distribution (or, alternatively, the empirical probability distribution) $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ that accommodates the moment conditions. Now, the estimated probabilities are functions of θ.

Manipulation of the first-order conditions from the lagrangian (74) gives an alternative characterization of the solution $(\hat{\theta}_{el}, \hat{\tau}_{el})$ that connects with the estimating function approach. The estimators satisfy the equations $\sum_{i=1}^{n} \rho_{el}(y_i; \hat{\theta}_{el}, \hat{\tau}_{el}) = 0$, where

$$\rho_{el}(y; \theta, \tau) = \begin{pmatrix} \tau' \dfrac{\partial h(y;\theta)}{\partial \theta'} [1 + \tau' h(y;\theta)]^{-1} \\ h(y;\theta)[1 + \tau' h(y;\theta)]^{-1} \end{pmatrix}. \tag{76}$$

The dimension of the multiplier τ is equal to the number of moment conditions given by the dimension of the function $h(y;\theta)$. It can be shown that, under regularity conditions and for the given set of moment conditions, $\hat{\theta}_{el}$ is asymptotically equivalent to the efficient estimator within the class of the GMM estimators. The efficient GMM estimator usually necessitates a two-step procedure for the estimation of the best weighting matrix. In contrast, the empirical likelihood approach requires only one-step, and this is expected to result in improved finite-sample properties of estimators [see for example, Mittelhammer and Judge (2001)]. In other words, as it is also emphasized in Mittelhammer et al. (2000, ch.12), the empirical likelihood approach offers an operational way to optimally combine estimating equations.

Golan and Judge (1996), Imbens (1997), Imbens et al. (1998), and, Qin and Lawless (1994) proposed a different objective function motivated by entropy or the *Kullback-Leibler information criterion* (KLIC). This suggestion leads to the so-called exponential tilting (ET) estimator θ_{et} of θ, which is based on the optimization problem [see also equation (20) and the discussion that follows]

$$\min_{\pi_i, \theta} \sum_{i=1}^{n} \pi_i ln \pi_i \qquad \text{s.t.} \qquad \sum_{i=1}^{n} \pi_i = 1, \qquad \sum_{i=1}^{n} \pi_i h(y_i;\theta) = 0. \tag{77}$$

We denote the estimated Lagrange multiplier for the moment restrictions by $\hat{\tau}_{et}$. Again, the solution yields estimated probabilities that are different from n^{-1}:

$$\hat{\pi}_{et,i} = \frac{\exp\{\hat{\tau}_{et}' h(y_i; \hat{\theta}_{et})\}}{\sum_{i=1}^{n} \exp\{\hat{\tau}_{et}' h(y_i; \hat{\theta}_{et})\}}. \tag{78}$$

The estimators in this problem satisfy the estimating equations $\sum_{i=1}^{n} \rho_{et}(y_i; \hat{\theta}_{et}, \hat{\tau}_{et}) = 0$, where

$$\rho_{et}(y; \theta, \tau) = \begin{pmatrix} \tau' \dfrac{\partial h(y;\theta)}{\partial \theta'} \exp\{\tau' h(y_i;\theta)\} \\ h(y;\theta) \exp\{\tau' h(y_i;\theta)\} \end{pmatrix}. \tag{79}$$

The ET estimator was also suggested by Kitamura and Stutzer (1997) who handle the important case of weakly dependent data. Their proposal utilizes the framework of linear inverse problems and that allows them to propose convenient computational procedures. They show that the ET estimators for (θ, τ) can be characterized by the optimization program

$$(\hat{\theta}_{et}, \hat{\tau}_{et}) = \arg\max_{\theta} \min_{\tau} \sum_{i=1}^{n} \exp\{\tau' h(y_i; \theta)\}. \tag{80}$$

From (80), it is easy to obtain the estimating functions given by (79). Similar characterization can be made for the EL problem thereby yielding the set of just identified equations (76).

In the case of serially correlated observations Kitamura and Stutzer (1997) propose to smooth the observations before the optimization and, in particular, to replace $h(y_i; \theta)$ in (80) by

$$\bar{h}_i(\theta) = \sum_{k=-K}^{K} \frac{1}{2K+1} h(y_{i-k}, \theta),$$

where $K^2/n \to 0$, and $K \to \infty$ as $n \to \infty$.

The limiting distributions of the EL estimator $\hat{\theta}_{el}$ and the ET estimator $\hat{\theta}_{et}$ are identical; see Imbens et al. (1998). They are both $\sqrt{n}$–consistent and asymptotically normal $N(0, \Sigma)$, where

$$\Sigma = \left\{ E\left(\frac{\partial h(y;\theta_0)}{\partial \theta}\right) \left[E(h(y;\theta_0)h(y;\theta_0)')\right]^{-1} E\left(\frac{\partial h(y;\theta_0)}{\partial \theta}\right) \right\}^{-1}, \tag{81}$$

where the expectations are evaluated at the true value of parameter θ_0. A consistent estimate of Σ can be constructed by using the empirical analogs of each expectation involved. Efficiency dictates that each observation y_i is weighted by the estimated probability $\hat{\pi}_i$. For instance, when working with the exponential tilting estimator $E(h(y;\theta_0)h(y;\theta_0)')$ will be estimated by $\sum_i \hat{\pi}_{et,i} h(y_i; \hat{\theta}_{et}) h(y_i; \hat{\theta}_{et})'$. However, consistency is not lost by using the uniform weights n^{-1}.

Imbens et al. (1998) also show that the two estimators can be nested in a class of estimators that arise from the minimization of the Cressie-Read power-divergence criterion (14) [Cressie and Read (1984), Read and Cressie (1988)]. Consider two discrete distributions with common support $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ and rewrite the Cressie-Read power-divergence criterion as

$$I_\lambda(p, q) = \frac{1}{\lambda(1+\lambda)} \sum_{i=1}^{n} p_i \left[\left(\frac{p_i}{q_i}\right)^{\lambda} - 1 \right], \tag{82}$$

where λ is a given constant, and therefore, we can talk about a class of divergence criteria. For given λ, the estimators for θ are defined so that the distance of the estimated probabilities $\hat{\pi}_i$ from the empirical distribution n^{-1} is minimized subject to the moment restrictions. More specifically, let $p_i = n^{-1}$ and $q_i = \pi_i$. For fixed constant λ, we consider the minimization program

$$\min_{\pi_i, \theta} I_\lambda(\mathbf{i}/n, \pi) \qquad \text{s.t.} \qquad \sum_{i=1}^{n} \pi_i = 1, \qquad \sum_{i=1}^{n} \pi_i h(y_i; \theta) = 0. \tag{83}$$

For $\lambda \to 0$, we obtain the empirical likelihood based estimators $(\hat{\pi}_{el}, \hat{\theta}_{el})$. When $\lambda \to -1$, the exponential tilting estimators $(\hat{\pi}_{et}, \hat{\theta}_{et})$ arise. It would be an interesting problem to combine Choi et al. (2000) empirical tilting model method and the above empirical likelihood method. A possible optimization problem would be to maximize $\sum_{i=1}^{n} \pi_i \ln f(y_i; \theta)$ subject to $\sum_{i=1}^{n} \pi_i = 1$, $\sum_{i=1}^{n} \pi_i h(y_i; \theta) = 0$ and $I_\lambda(\mathbf{i}/n, \pi) = \delta$ for given small value of δ.

An interesting special case of the minimization program (83) is the log Euclidean likelihood estimator $\hat{\theta}_{lel}$, obtained when $\lambda = -2$; see Owen (1991), Qin and Lawless (1994). Imbens et al. (1998) show that $\hat{\theta}_{lel}$, for given estimate of the multiplier τ, is characterized by the estimating equation

$$\left[\sum_{i=1}^{n} \frac{\partial h(y_i; \hat{\theta}_{lel})}{\partial \theta'} \left(1 + \tau' h(y_i; \hat{\theta}_{lel})\right)\right]' \left[\sum_{i=1}^{n} h(y; \hat{\theta}_{lel}) h(y_i; \hat{\theta}_{lel})'\right]^{-1} \sum_{i=1}^{n} h(y_i; \hat{\theta}_{lel}) = 0. \qquad (84)$$

When $\partial h(y; \theta)/\partial \theta$ does not depend on data y, the term $(1 + \tau' h(y_i; \theta))$ factors out and the equation (84) is identical to the one that characterizes the iterated GMM estimator [see Hansen et al. (1996)].

The characterization of EL and ET by systems of just identified estimating equations, as well as the fact that GMM can be regarded as a special case of Cressie-Read criterion, suggest the prevalence of the MM approach to estimation. Manski (1988) included many estimation approaches under the umbrella of analog estimation that estimates a parameter by the sample analog. In the majority of applications, "sample analog" was obtained by equal weighting of each observation. Recent methods have expanded the class of analog estimation by stressing efficient weighting schemes. In our view it will also be interesting to see how the moment restrictions can be accommodated in a likelihood framework that is less prone to misspecification of the density function.

6 Epilogue

We have reviewed important phases in the development of parameter estimation, both in the statistical and econometric literature. We tried to stress the historical continuity of the estimation methodology. In particular, our goal has been towards a synthesis of seemingly unrelated lines of thought. In the light of advancements in the course of the last century, estimation methods in current use in econometrics do not seem that distant from earlier developments in the statistical literature.

We started with Pearson's MM at 1894 and continued with Fisher's parametric likelihood at 1912 and 1922. Based on Pearson's χ^2 statistic a third competing method of estimation was developed under the name "minimum chi-squared." Subsequently, this statistic provided the basis for the development of a MM approach by Ferguson (1958), that was able to handle cases where the number of available moments exceeds the number of unknown parameters. Parallel lines of research were followed by Godambe (1960) and Durbin (1960) with the estimating function approach. More recently, in econometrics Hansen (1982) advocated the GMM procedure that bears close resemblance to the last two approaches by Ferguson and Godambe-Durbin.

Nowadays, econometricians are moving away from the maximum likelihood estimation. The nature of the conclusions derived from economic theorizing and the desire to avoid parametric assumptions push the applied researcher to variants of the method of moments. At the onset of the new century, after a long devotion to Fisher's likelihood, the ideas of Pearson look more useful than ever. It is our view that this interplay between Fisher's concept of likelihood and Pearson's method of moments has resulted in an amazing synthesis that provides efficient estimation with minimal model assumptions and very promising properties in finite samples.

These ideas have been blended fruitfully by the nonparametric method of empirical likelihood. By using the Kullback-Leibler information criterion as a measure of the distance between two distributions, we can also tie the method of moments to the concept of entropy. The current state of affairs calls for more work on these estimation strategies. Maybe this task also calls for a fresh reading of the old proposals; certainly, the close relationship between modern econometrics with the ideas of Pearson and Fisher points to this direction.

In this review of the development of estimation, we have occasionally separated ourselves from the riveting narrative and digressed to some side stories in statistics. In our view these stories are important. Sometimes seemingly unambiguous description of different estimation methodologies in econometrics textbooks brings temporary clarity to our confusions and helps veil some of the parallel, and often-times much earlier, developments in the statistics literature. It is worthwhile to record the initial motivation and historical progress in one place since the original lines of development and philosophies become clouded by time. Although our overall aim was too ambitious, this paper is essentially our first modest attempt, and some more research along these lines would be profitable for students of econometrics.

Acknowledgements

We are grateful to Amos Golan and two anonymous referees for many helpful suggestions that have considerably improved the paper. An earlier version of this paper under the title "A Brief History of Estimation" was presented at a conference in honor of George Judge held at the University of Illinois, May 1-2, 1999. We wish to acknowledge helpful comments from the participants, in particular, from George Judge and Arnold Zellner. Kostas Fokianos, Alastair Hall and Yuichi Kitamura kindly provided some extra references and/or detailed comments. We would also like to thank Tatyana Dubovyk for providing very competent research assistance and Naoko Miki for her help in preparing the manuscript. Special thanks are due to Amos Golan without whose encouragement and kind prodding the paper would not have been completed. However, we retain the responsibility for any remaining errors. Financial support from the Office of Research, College of Commerce and Business Administration, University of Illinois of Urbana-Champaign (for Anil Bera) and from University of Cyprus (for Yannis Bilias) is gratefully acknowledged. This paper is dedicated to George Judge for his lifelong and continuing contribution to econometrics.

References

Aldrich, J., 1997. R.A. Fisher and the making of maximum likelihood 1912-1922, Statistical Science 12, 162-176.

Back, K., Brown, D. P., 1993. Implied probabilities in GMM estimators, Econometrica 61, 971-975.

Bennett, T. L., 1907-1908. Errors of observation, Technical Lecture, Number 4, Survey Department, Ministry of Finance, National Printing Department, Cairo, Egypt.

Bera, A. K., 2000. Hypothesis testing in the 20th Century with a special reference to testing with misspecified models, in: Rao, C. R., Szekely, G. J. eds., Statistics for the 21st Century (Marcel Dekker, New York) 33-92.

Bera, A. K., Bilias, Y., 2000. The MM, ME, ML, EL, EF and GMM approaches to estimation: A synthesis, Working Paper, Department of Economics, University of Illinois.

Bera, A. K., Bilias, Y., 2001. Rao's score, Neyman's C(α) and Silvey's LM tests: An essay on historical developments and some new results, Journal of Statistical Planning and Inference 97, forthcoming.

Bera, A. K., Ghosh, A., 2001. Neyman's smooth test and its applications in econometrics, in: Ullah, A., Wan, A., Chaturvedi, A. T. eds., Handbook of Applied Econometrics and Statistical Inference (Marcel Dekker, New York), forthcoming.

Berger, J. O., 1985. Statistical Decision Theory and Bayesian Analysis (Springer-Verlag, New York).

Berkson, J., 1980. Minimum chi-square, not maximum likelihood!, Annals of Statistics 8, 457-487.

Box, J. F., 1978. R. A. Fisher: The Life of a Scientist (John Wiley & Sons, New York).

Choi, E., Hall, P., Presnell, B., 2000. Rendering parametric procedures more robust by empirically tilting the model, Biometrika 87, 453-465.

Cressie, N., Read, T. R. C., 1984. Multinomial goodness-of-fit tests, Journal of the Royal Statistical Society 46, Series B, 440-464.

Cowell, F. A., 1980. On the structure of additive inequality measures, Review of Economic Studies 47, 521-531.

Crowder, M., 1986. On consistency and inconsistency of estimating equations, Econometric Theory 2, 305-330.

David, H. A., 1995. First (?) occurrence of common terms in mathematical statistics, American Statistician 49, 121-133.

Durbin, J., 1954. Errors in variables, Review of Institute of International Statistics 22, 23-54.

Durbin, J., 1960. Estimation of parameters in time-series regression models, Journal of the Royal Statistical Society 22, Series B, 139-153.

Edgeworth, F. Y., 1908. On the probable errors of frequency-constants, Journal of the Royal Statistical Society 71, 381-397, 499-512, 651-678.

Edgeworth, F. Y., 1909. Addendum on "Probable errors of frequency-constants", Journal of the Royal Statistical Society 72, 81-90.

Edwards, A. W. F., 1997a. Three early papers on efficient parametric estimation, Statistical Science 12, 35-47.

Edwards, A. W. F., 1997b. What did Fisher mean by 'inverse probability' in 1912-1922?, Statistical Science 12, 177-184.

Engledow, F. L., Yule, G. U., 1914. The determination of the best value of the coupling-ratio from a given set of data, Proceedings of the Cambridge Philosophical Society 17, 436-440.

Ferguson, T. S., 1958. A method of generating best asymptotically normal estimates with application to estimation of bacterial densities. Annals of Mathematical Statistics 29, 1046-1062.

Ferguson, T. S., 1996. A Course in Large Sample Theory (Chapman & Hall, London).

Fisher, R. A., 1912. On an absolute criterion for fitting frequency curves, Messenger of Mathematics 41, 155-160.

Fisher, R. A., 1921. On the "probable error" of a coefficient of correlation deduced from a small sample, Metron 1, 3-32.

Fisher, R. A., 1922. On the mathematical foundations of theoretical statistics, Philosophical Transactions of the Royal Society of London 222, Series A, 309-368.

Fisher, R. A., 1924a. The conditions under which χ^2 measures the discrepancy between observation and hypothesis, Journal of the Royal Statistical Society 87, 442-450.

Fisher, R. A., 1924b. On a distribution yielding the error functions of several well known statistics, Proceedings of the International Congress of Mathematics (Toronto) 2, 805-813.

Fisher, R. A., 1928. Statistical Methods for Research Workers (Oliver and Boyd, Edinburgh).

Fisher, R. A., 1937. Professor Karl Pearson and the method of moments, Annals of Eugenics 7, 303-318.

Freeman, M. F., Tukey, J. W., 1950. Transformations related to the angular and the square root, Annals of Mathematical Statistics 21, 607-611.

Galton, F., 1885. Regression towards mediocrity in hereditary stature, Journal of the Anthropological Institute 15, 246-263.

Geary, R. C., 1948. Studies in the relations between economic time series, Journal of the Royal Statistical Society 10, Series B, 140-158.

Geary, R. C., 1949. Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown, Econometrica 17, 30-58.

Geisser, S., 1980. Basic theory of the 1922 mathematical statistics paper, in: Fienberg, S. E., Hinkley, D. V. eds., R. A. Fisher: An Appreciation (Springer-Verlag, New York) 59-66.

Godambe, V. P., 1960. An optimum property of regular maximum likelihood estimation, Annals of Mathematical Statistics 31, 1208-1212.

Godambe, V. P., 1985. The foundations of finite sample estimation in stochastic processes, Biometrika 72, 419-428.

Godambe, V. P., ed., 1991. Estimating Functions (Oxford Science Publications, Oxford).

Godambe, V. P., Heyde, C. C., 1987. Quasi-likelihood and optimal estimation, International Statistical Review 55, 231-244.

Gokhale, D. V., 1975. Maximum entropy characterizations of some distributions, in: Patil, G. P., Kotz, S., Ord, J. K. eds., Statistical Distributions in Scientific Work, Volume 3 – Characterizations and Applications (D. Reidel Publishing Company, Dordrecht) 299-304.

Golan, A., Judge, G., 1996. A maximum entropy approach to empirical likelihood estimation and inference, University of California ARE Working Paper.

Golan, A., Judge, G., Miller, D., 1996. Maximum Entropy Econometrics: Robust Estimation with Limited Data (John Wiley & Sons, New York).

Golan, A., Judge, G., Miller, D., 1997. The maximum entropy approach to estimation and inference: An overview, in: Fomby, T.B., Hill, R. C. eds., Advances in Econometrics: Applying Maximum Entropy to Econometric Problems (JAI Press, Greenwich) 3-24.

Golan, A., Judge, G., Miller, D., 1998. Information recovery in simultaneous-equations statistical models, in: Ullah A., Giles, D. E. A. eds., Handbook of Applied Economic Statistics (Marcel Dekker, New York) 365-381.

Goldberger, A. S., 1968. Topics in Regression Analysis (The Macmillan Company, New York).

Goldberger, A. S., 1972. Structural equation methods in the social sciences, Econometrica 40, 979-1001.

Grandy, W. T., Schick, L. H. eds., 1991. Maximum Entropy and Bayesian Methods (Kluwer Academic Publishers, Dordrecht).

Haberman, S. J., 1984. Adjustment by minimum discriminant information, Annals of Statistics 12, 971-988.

Hacking, I., 1984. Trial by number, Science 84, 67-70.

Hald, A., 1998. A History of Mathematical Statistics form 1750 to 1930 (John Wiley & Sons, New York).

Hald, A., 1999. On the history of maximum likelihood in relation to inverse probability and least squares, Statistical Science 14, 214-222.

Haldane, J. B. S., 1919a. The probable errors of calculated linkage values, and the most accurate methods of determining gametic from certain zygotic series, Journal of Genetics 8, 291-297.

Haldane, J. B. S., 1919b. The combination of linkage values, and the calculation of distances between the loci of linked factors, Journal of Genetics 8, 299-309.

Hall, A., 1993. Some aspects of generalized method of moments estimation, in: Maddala, G. S., Rao, C. R., Vinod, H. D. eds., Handbook of Statistics, Volume 11 (North Holland, Amsterdam) 393-417.

Hall, A., 2001. Generalized method of moments, in: Baltagi, B., ed., A Companion to Theoretical Econometrics, (Blackwell Publishers, Oxford) 230-250.

Hansen, L., 1982. Large sample properties of generalized method of moments estimators, Econometrica 50, 1029-1054.

Hansen, L., Heaton, J., Yaron, A., 1996. Finite sample properties of some alternative GMM estimators, Journal of Business and Economic Statistics 14, 262-280.

Harris, J. A., 1912. A simple test of the goodness of fit of Mandelian ratios, American Naturalist 46, 741-745.

Heyde, C. C., 1997. Quasi-likelihood And Its Applications: A General Approach to Optimal Parameter Estimation (Springer, New York).

Huber, P. J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions, in: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, Berkeley), Volume 1, 221-233.

Imbens, G. W., 1997. One-step estimators for over-identified generalized method of moments models, Review of Economic Studies 64, 359-383.

Imbens, G. W., Spady, R. H., Johnson, P., 1998. Information theoretic approaches to inference in moment condition models, Econometrica 66, 333-357.

Jaynes, E. T., 1957. Information theory and statistical mechanics, Physical Review 106, 620-630.

Kagan, A. M., Linnik, Y. V., Rao, C. R., 1973. Characterization Problems in: Mathematical Statistics (John Wiley & Sons, New York).

Katti, S. K., 1970. Review of "Topics in Regression Analysis", Econometrica 38, 945-946.

Kent, J. T., 1982. Robust properties of likelihood ratio tests, Biometrika 69, 19-27.

Kitamura, Y., Stutzer, M., 1997. An information theoretic alternate to generalized method of moments estimation, Econometrica 65, 861-874.

Koenker, R., 2000. Galton, Edgeworth, Frisch, and prospect for quantile regression in econometrics, Journal of Econometrics 95, 347-374.

Le Cam, L., 1990. Maximum likelihood: An introduction, International Statistical Review 58, 153-171.

Lehmann, E. L., 1999. "Student" and small-sample theory, Statistical Science 14, 418-426.

Li, D. X., Turtle, H. J., 2000. Semiparametric ARCH models: An estimating function approach. Journal of Business and Economic Statistics , 175-186.

Lindsay, B. G., 1994. Efficiency versus robustness: The case for minimum Hellinger distance and related methods, Annals of Statistics 22, 1081-1114.

Maasoumi, E., 1993. A compendium to information theory in economics and econometrics, Econometric Reviews 12, 137-181.

Manski, C., 1988. Analog Estimation Methods in Econometrics (Chapman and Hall, London).

Mardia, K. V., 1975. Characterization of directional distributions, in: Patil, G. P., Kotz, S., Ord, J. K. eds., Statistical Distributions in Scientific Work, Volume 3 – Characterizations and Applications (D. Reidel Publishing Company, Dordrecht) 365-385.

McLeish, D. L., 1984. Estimation for aggregate models: The aggregate Markov chain, Canadian Journal of Statistics 12, 265-282.

Mensch, R., 1980. Fisher and the method of moments, in: Fienberg, S.E., Hinkley, D. V. eds., R. A. Fisher: An Appreciation (Springer-Verlag, New York) 67-74.

Mittelhammer, R. C., Judge, G. G., 2001. Finite sample performance of empirical likelihood under endogeneity, Discussion Paper, University of California, Berkeley.

Mittelhammer, R. C., Judge, G. G., Miller, D. J., 2000. Econometric Foundations (Cambridge University Press, Cambridge).

Newey, W. K., 1993. Efficient estimation of models with conditional moment restrictions, in: Maddala, G. S., Rao, C. R., Vinod, H. D., eds., Handbook of Statistics, Volume 11 (North Holland, Amsterdam) 419-454.

Neyman, J., 1937. "Smooth test" for goodness of fit, Skand. Akturarietidskr 20, 150-199.

Neyman, J., 1949. Contribution to the theory of the χ^2 test, in: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, Berkeley) 329-273.

Neyman, J., 1967. R.A. Fisher (1890-1962): An appreciation, Science 156, 1456-1460.

Neyman, J., Pearson, E. S., 1928. On the use and interpretation of certain test criteria for purpose of statistical inference, Part II, Biometrika 20, 263-294.

Owen, A., 1988. Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75, 237-249.

Owen, A., 1990. Empirical likelihood confidence regions, Annals of Statistics 18, 90-120.

Owen, A., 1991. Empirical likelihood for linear models, Annals of Statistics 19, 1725-1747.

Pagan, A. R., Robertson, J., 1997. GMM and its problems, manuscript, Australian National University.

Pearson, E. S., 1936. Karl Pearson: An appreciation of some aspects of his life and work, Biometrika 28, 193-257.

Pearson, K., 1894. Contribution to the mathematical theory of evolution, Philosophical Transactions of the Royal Society of London 185, Series A, 71-110.

Pearson, K., 1895. Skew variation in homogeneous material, Philosophical Transactions of the Royal Society of London 186, Series A, 343-414.

Pearson, K., 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philosophical Magazine Series 50, 157-175.

Pearson, K., 1902. On the systematic fitting of curves to observations and measurements, Parts I and II, Biometrika 1, 265-303; 2, 1-23.

Pearson, K., 1936. Method of moments and method of maximum likelihood, Biometrika 28, 34-59.

Phillips, P. C. B., 1985. ET interviews: Professor J. D. Sargan, Econometric Theory 1, 119-139.

Pompe, B., 1994. On some entropy methods in data analysis, Chaos, Solitons and Fractals 4, 83-96.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations, Annals of Statistics 22, 300-325.

Read, T. R. C., Cressie, N., 1988. Goodness-of-Fit Statistics for Discrete Multivariate Data (Springer-Verlag, New York).

Rényi, A., 1961. On measures of entropy and information, Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics, and Probability (University of California Press, Berkeley) Volume 1, 547-561.

Reiersøl, O., 1941. Confluence analysis by means of lag moments and other methods of confluence analysis, Econometrica 9, 1-23.

Reiersøl, O., 1945. Confluence Analysis by Means of Sets of Instrumental Variables (Almquist & Wiksell, Uppsala).

Sargan, J. D., 1958. The estimation of economic relationships using instrumental variables, Econometrica 26, 393-415.

Sargan, J. D., 1959. The estimation of relationships with auto-correlated residuals by the use of instrumental variables, Journal of the Royal Statistical Society 21, Series B, 91-105.

Savage, L. J., 1976. On rereading R. A. Fisher, Annals of Statistics 4, 441-483.

Schützenberger, M. P., (1954). Contribution aux applications statistiques de la théorie de l' information, Publication of the Institute of Statistics, University of Paris.

Sen, A. K., 1975. On Economic Inequality (Claredon Press, Oxford).

Shannon, C. E., Weaver, W., 1949. The Mathematical Theory of Communication (University of Illinois Press, Urbana).

Shenton, L. R., 1950. Maximum likelihood and the efficiency of the method of moments, Biometrika 37, 111-116.

Shenton, L. R., 1958. Moment estimators and maximum likelihood, Biometrika 45, 411-420.

Shenton, L. R., 1959. The distribution of moment estimators, Biometrika 46, 296-305.

Shorrocks, A. F., 1980. The class of additively decomposable inequality measures, Econometrica 48, 613-625.

Smith, K., 1916. On the 'best' values of the constants in the frequency distributions, Biometrika 11, 262-276.

Steven, S. S., (1975). Psychophysics (John Wiley, New York).

Stigler, S. M., 1976. Comment on "On rereading R. A. Fisher", Annals of Statistics 4, 498-500.

Student, 1908. The probable error of a mean, Biometrika 6, 1-25.

Ullah, A., 1996. Entropy, divergence and distance measures with econometric applications, Journal of Statistical Planning and Inference 49, 137-162.

Urzúa, C. M., 1988. A class of maximum-entropy multivariate distributions, Communications in Statistics: Theory and Method 17, 4039-4057.

Urzúa, C. M., 1997. Omnibus tests for multivariate normality based on a class of maximum entropy distributions, in: Fomby, T. B., Hill, R. C. eds., Advances in Econometrics: Applying Maximum Entropy to Econometric Problems (JAI Press, Greenwich) 341-358.

Vinod, H. D., 1998. Foundations of statistical inference based on numerical roots of robust pivot functions, Journal of Econometrics 86, 387-396.

Wedderburn, R. W. M., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, Biometrika 61, 439-447.

White, H., 1982. Maximum likelihood estimation of misspecified models, Econometrica 50, 1-26.

Wilks, S. S., 1943. Mathematical Statistics (Princeton University Press, Princeton)

Wright, S., 1920. The relative importance of heredity and environment in determining the birth weight of guinea pigs, Proceedings of the National Academy of Sciences 6, 320-332.

Wright, S., 1921. Correlation and causation, Journal of Agricultural Research 20, 557-585.

Wright, S., 1925. Corn and hog correlations, U.S. Department of Agriculture Bulletin 1300, Washington.

Wright, S., 1928. Appendix to Wright, P. The Traffic on Animal and Vegetable Oils (Macmillan, New York).

Zellner, A., 1991. Bayesian methods and entropy in economics and econometrics, in: Grandy, W. T., Schick, L. H. eds., Maximum Entropy and Bayesian Methods (Kluwer Academic Publishers, Dordrecht) 17-31.

Zellner, A., 1997. Bayesian method of moments (BMOM): Theory and applications, in: Fomby, T. B., Hill, R. C. eds., Advances in Econometrics: Applying Maximum Entropy to Econometric Problems (JAI Press, Greenwich) 85-105.

Zellner, A., Highfield, R., 1988. Calculation of maximum entropy distributions and approximation of marginal posterior distributions, Journal of Econometrics 37, 195-209.

SELECTED RECENT PUBLICATIONS

Bera A. K. and Yannis Bilias, Rao´s Score, Neyman´s C (α) and Silvey´s LM Tests: An Essay on Historical Developments and Some New Results, *Journal of Statistical Planning and Inference,* forthcoming.

Bertaut C. and M. Haliassos, Precautionary Portfolio Behavior from a Life - Cycle Perspective, *Journal of Economic Dynamics and Control,* 21, 1511-1542, 1997.

Bilias Y., Minggao Gu and Zhiliang Ying, Towards a General Asymptotic Theory for the Cox model with Staggered Entry, *The Annals of Statistics*, 25, 662-682, 1997.

Blundell R., P. Pashardes and G. Weber, What Do We Learn About Consumer Demand Patterns From Micro-Data?, *American Economic Revie*w, 83, 570-597, 1993.

Bougheas S., P. Demetriades and T. P. Mamouneas, Infrastructure, Specialization and Economic Growth, *Canadian Journal of Economics,* forthcoming.

Caporale W., C. Hassapis and N. Pittis, Unit Roots and Long Run Causality: Investigating the Relationship between Output, Money and Interest Rates, *Economic Modeling*, 15(1), 91-112, January 1998.

Caporale G. and N. Pittis, Efficient estimation of cointegrated vectors and testing for causality in vector autoregressions: A survey of the theoretical literature, *Journal of Economic Surveys*, forthcoming.

Caporale G. and N. Pittis, Unit root testing using covariates: Some theory and evidence, *Oxford Bulletin of Economics and Statistics*, forthcoming.

Caporale G. and N. Pittis, Causality and Forecasting in Incomplete Systems, *Journal of Forecasting,* 16, 6, 425-437, 1997.

Clerides K. S., Lach S. and J.R. Tybout, Is Learning-by-Exporting Important? Micro-Dynamic Evidence from Colombia, Morocco, and Mexico, *Quarterly Journal of Economics* 113(3), 903- 947, August 1998.

Cukierman A., P. Kalaitzidakis, L. Summers and S. Webb, Central Bank Independence, Growth, Investment, and Real Rates", Reprinted in Sylvester Eijffinger (ed), Independent Central Banks and Economic Performance, Edward Elgar, 416-461, 1997.

Dickens R., V. Fry and P. Pashardes, Non-Linearities and Equivalence Scales, *The Economic Journa*l, 103, 359-368, 1993.

Demetriades P. and T. P. Mamuneas, Intertemporal Output and Employment Effects of Public Infrastructure Capital: Evidence from 12 OECD Economies, *Economic Journal,* July 2000.

Eicher Th. and P. Kalaitzidakis, The Human Capital Dimension to Foreign Direct Investment: Training, Adverse Selection and Firm Location". In Bjarne Jensen and Kar-yiu

Wong (eds), Dynamics,Economic Growth, and International Trade, The University of Michigan Press, 337-364, 1997.

Fry V. and P. Pashardes, Abstention and Aggregation in Consumer Demand, *Oxford Economic Paper*s, 46, 502-518, 1994.

Gatsios K., P. Hatzipanayotou and M. S. Michael, International Migration, the Provision of Public Good and Welfare, *Journal of Development Economics*, 60/2, 561-577, 1999.

Haliassos M. and C. Hassapis, Non-expected Utility, saving, and Portfolios, *The Economic Journal*, 110, 1-35, January 2001.

Haliassos M. and J. Tobin, The Macroeconomics of Government Finance, reprinted in J.Tobin, Essays in Economics, vol. 4, Cambridge: MIT Press, 1996.

Haliassos M. and C. Bertaut, Why Do So Few Hold Stocks?, *The Economic Journal*, 105, 1110- 1129, 1995.

Haliassos M., On Perfect Foresight Models of a Stochastic World, *Economic Journal*, 104, 477-491, 1994.

Hassapis C., N. Pittis and K. Prodromidis, Unit Roots and Granger Causality in the EMS Interest Rates: The German Dominance Hypothesis Revisited, *Journal of International Money and Finance*, 18(1), 47-73, 1999.

Hassapis C., S. Kalyvitis and N. Pittis, Cointegration and Joint Efficiency of International Commodity Markets”, *The Quarterly Review of Economics and Finance,* 39, 213-231, 1999.

Hassapis C., N. Pittis and K. Prodromides, EMS Interest Rates: The German Dominance Hypothesis or Else?” in European Union at the Crossroads: A Critical Analysis of Monetary Union and Enlargement, Aldershot, UK., Chapter 3, 32-54, 1998. Edward Elgar Publishing Limited.

Hatzipanayotou P., and M. S. Michael, General Equilibrium Effects of Import Constraints Under Variable Labor Supply, Public Goods and Income Taxes, *Economica*, 66, 389-401, 1999.

Hatzipanayotou, P. and M.S. Michael, Public Good Production, Nontraded Goods and Trade Restriction, *Southern Economic Journal,* 63, 4, 1100-1107, 1997.

Hatzipanayotou, P. and M. S. Michael, Real Exchange Rate Effects of Fiscal Expansion Under Trade Restrictions, *Canadian Journal of Economics,* 30-1, 42-56, 1997.

Kalaitzidakis P., T. P. Mamuneas and Th. Stengos, A Nonlinear Sensitivity Analysis of Cross-Country Growth Regressions, *Canadian Journal of Economics,* forthcoming.

Kalaitzidakis P., T. P. Mamuneas and Th. Stengos, European Economics: An Analysis Based on Publications in Core Journals, *European Economic Revie*w, 43, 1150-1168, 1999.

Kalaitzidakis P., On-the-job Training Under Firm-Specific Innovations and Worker Heterogeneity, *Industrial Relations,* 36, 371-390, July 1997.

Ludvigson S. and A. Michaelides, Does Buffer Stock Saving Explain the Smoothness and Excess Sensitivity of Consumption?, *American Economic Review*, forthcoming

Lyssiotou Panayiota, Dynamic Analysis of British Demand for Tourism Abroad, *Empirical Economics,* forthcoming, 2000.

Lyssiotou P., P. Pashardes and Th. Stengos, Testing the Rank of Engel Curves with Endogenous Expenditure, *Economics Letters,* 64, 61-65, 1999.

Lyssiotou P., P. Pashardes and Th. Stengos, Preference Heterogeneity and the Rank of Demand Systems, *Journal of Business and Economic Statistics*, 17 (2), 248-252, April 1999.

Lyssiotou Panayiota, Comparison of Alternative Tax and Transfer Treatment of Children using Adult Equivalence Scales, *Review of Income and Wealth*, 43 (1), 105-117, March 1997.

Mamuneas, Theofanis P., Spillovers from Publicly – Financed R&D Capital in High-Tech Industries, *International Journal of Industrial Organization,* 17(2), 215-239, 1999.

Mamuneas, T. P. and Nadiri M. I., R&D Tax Incentives and Manufacturing-Sector R&D Expenditures, in *Borderline Case: International Tax Policy, Corporate Research and Development, and Investmen*t, James Poterba (ed.), National Academy Press, Washington D.C., 1997. Reprinted in *Chemtec*h, 28(9), 1998.

Mamuneas, T. P. and Nadiri M. I., Public R&D Policies and Cost Behavior of the US Manufacturing Industries, *Journal of Public Economics*, 63, 57-81, 1996.

Michaelides A. and Ng, S., Estimating the Rational Expectations Model of Speculative Storage: A Monte Carlo Comparison of three Simulation Estimators, *Journal of Econometrics*, 96(2), 231-266, 2000.

Pashardes Panos, Equivalence Scales in a Rank-3 Demand System, *Journal of Public Economic*s, 58, 143-158, 1995.

Pashardes Panos, Bias in Estimating Equivalence Scales from Grouped Data, *Journal of Income Distributio*n, Special Issue: Symposium on Equivalence Scales, 4, 253-264,1995.

Pashardes Panos., Bias in Estimation of the Almost Ideal Demand System with the Stone Index Approximation, *Economic Journa*l, 103, 908-916, 1993.

Spanos Aris, Revisiting Date Mining: ´Hunting´ With or Without a License, *Journal of Methodology,* July 2000.

Spanos Aris, On Normality and the Linear Regression Model, *Econometric Reviews,* 14,195-203, 1995.

Spanos Aris, On Theory Testing in Econometrics: Modeling with nonexperimental Data, *Journal of Econometrics,* 67, 189-226, 1995.

Spanos Aris, On Modeling Heteroscedasticity: The Student's *t* and Elliptical Linear Regression Models, *Econometric Theor*y, 10, 286-315, 1994.