Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

ESTIMATING FUNCTIONS: A SYNTHESIS OF LEAST SQUARES AND MAXIMUM LIKELIHOOD METHODS

V.P. Godambe
University of Waterloo

ABSTRACT

The development of the modern theory of estimating functions is traced from its inception. It is shown that this development has brought about a synthesis of the two historically important methodologies of estimation namely, the 'least squares' and the 'maximum likelihood'.

Key Words: Estimating functions; likelihood; score function.

1 Introduction

In common with most of the historical investigations, it is difficult to trace the origin of the subject of this conference: 'Estimating Functions'. However, in the last two centuries clearly there are three important precursors of the modern theory of estimating functions (EF): In the year 1805, Legendre introduced the least squares (LS) method. At the turn of the last century Pearson proposed the method of moments and in 1925 Fisher put forward the maximum likelihood (ML) equations. Of these three, the method of moments faded out in time because of its lack of any sound theoretical justification. However the other two methods namely the LS and the ML even at present play an important role in the statistical methodology. These two methods would also concern us in the following. The LS method was justified by what today is called the Gauss-Markoff (GM) theorem: The estimates obtained from LS equations are 'optimal' in the sense that they have minimum variance in the class of linear unbiased estimates. This was a *finite sample* justification. At about the same time Laplace provided a different 'asymptotic justification' for the method. Fisher justified the ML estimation, for it produced estimates which are *asymptotically* unbiased with smallest variance. This left open the question, is there a finite sample justification for the ML estimation corresponding to the GM theorem justification for the LS estimation?

The modern EF theory provided such a justification. According to the 'optimality criterion' of the EF theory, the score function (SF) is 'optimal'.

2 SF Optimality

To state the just mentioned result formally we introduce briefly some notation. Let $\mathcal{X} = \{x\}$ be the sample (observations) space and a class of possible distributions (densities) on $\mathcal{X}$ be given by $\{f(\cdot|\theta), \theta \in \Omega\}$, Ω being the parameter space, which we assume here to be the real line. If the function f is completely specified up to the (unknown) parameter θ, $f(\cdot|\theta)$ is called a *parametric model.* For this model the (score function) $SF = \partial \log f(\cdot|\theta)/\partial\theta$. Any real function of x and θ say $g(x,\theta)$ is called an *estimating function*, (EF). It is said to be unbiased if its mean value for $\theta \in \Omega$, is zero; $\mathcal{E}g = 0$. Further, for reasons which would be clear later, corresponding to every EF g we define a *standardized version* $g/\{\mathcal{E}(\frac{\partial g}{\partial\theta})\}$. Now in a class $\mathcal{G} = \{g\}$ of unbiased estimating functions, g^* is said to be 'optimal' if the variance of the standardized EF g, is minimized for $g = g^*$:

$$\mathcal{E}(g^*)^2/\{\mathcal{E}(\frac{\partial g^*}{\partial\theta})\}^2 \leq \mathcal{E}(g)^2/\{\mathcal{E}(\frac{\partial g}{\partial\theta})\}^2, \quad \theta \in \Omega,\ g \in \mathcal{G}. \tag{2.1}$$

SF Theorem (Godambe, 1960). For a parametric model $f(\cdot|\theta)$, granting some regularity conditions, in the class of all unbiased EFs, the optimal estimating function is given by the SF i.e.

$$g^* = \partial \log f(\cdot|\theta)/\partial\theta.$$

The optimality of the SF given by the above Theorem should be distinguished from the optimality of the LS estimates based on the GM theorem. The SF optimality (though with some additional assumptions implies asymptotic optimality of the ML estimate) is essentially *optimality* of the '*estimating function*' while the LS optimality is *optimality of the 'estimate'.* The concept underlying optimality criterion of the EF theory became more vivid and compelling in relation to the problem of *nuisance parameters.*

3 Conditional SF Optimality

Now let the parameter θ consist of two components θ_1 and θ_2, $\theta = (\theta_1, \theta_2)$ and the parametric model be $f(\cdot|\theta_1, \theta_2)$ where θ_1 is real and θ_2 is a vector; $\theta \in \Omega$, $\theta_1 \in \Omega_1$, $\theta_2 \in \Omega_2$ and $\Omega = \Omega_1 \times \Omega_2$. Further suppose we want to estimate only θ_1 (the interesting parameter) *ignoring* θ_2 (the nuisance parameter). How to proceed? To this question the ML estimation provides no satisfactory answer. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are jointly ML estimates for θ_1 and θ_2, as is well known, the estimate $\hat{\theta}_1$ can be inconsistent (unacceptable) in case the dimensionality of the parameter θ_2 goes on increasing with the number of observations (cf. Neyman-Scott, 1948). The EF theory, for the present situation implies, restricting to that part of likelihood function which is governed by the interesting parameter θ_1 only. Formally for the parametric model $f(\cdot|\theta_1,\theta_2)$, let $\mathcal{G}_1$ be the class of all unbiased EFs $g(x,\theta_1)$, that is functions of x and θ_1 only:

$$\mathcal{G}_1 = \{g : \; g = g(x,\theta_1), \; \mathcal{E}(g) = 0, \; \theta \in \Omega\}.$$

Further let t be a complete sufficient statistic for the parameter θ_2, for every fixed θ_1. Assuming the statistic t is *independent* of the parameter θ_1 we have **Conditional SF Theorem** (Godambe, 1976). Granting some regularity conditions, in the class of EFs $\mathcal{G}_1$, the 'optimal' EF g^* is given by the conditional SF i.e. $g^* = \partial \log f(\cdot|t;\theta_1)/\partial\theta_1$.

Note in the above theorem the definition of optimality is obtained from (2.1) just by replacing in it $\mathcal{G}$ by $\mathcal{G}_1$ and consequently $\mathcal{E}(\partial g/\partial\theta)$ by $\mathcal{E}(\partial g/\partial\theta_1)$. That is the criterion of optimality is *unconditional.* In the case of the Neyman-Scott example, unlike the ML estimate $\hat{\theta}$, the equation 'conditional SF = 0' provides a consistent estimate of θ_1. Further the EF optimality criterion suggests a definition of 'conditional SF' in case the statistic t depends on the parameter θ_1. If $t(\theta_{10})$ is the value of t at $\theta_1 = \theta_{10}$ then we *define* the conditional SF by g^* where

$$g^* = \{\partial \log f(\cdot|t(\theta_{10}),\theta_1,\theta_2)/\partial\theta_1\}_{\theta_{10}=\theta_1}. \tag{3.1}$$

This definition is motivated as follows. The EF g^* in (3.1) $\in \mathcal{G}_1$ though it depends on θ_2. It further is 'optimal' in $\mathcal{G}_1$ though only *locally* at θ_2 (Lindsay, 1982). Unlike the previous situation, when the sufficient statistic t, was independent of θ_1, now no universally optimal g^* (i.e. for all $\theta_2 \in \Omega_2$) exists in $\mathcal{G}_1$. Further though the EF g^* in (3.1) depends on θ_2, it is *orthogonal* to the marginal SF of the sufficient statistic t, hence the substitution of an estimate $\hat{\theta}_2$ derived from the latter, in the former would still leave the former nearly optimal for *large samples*, (Lindsay, 1982; Godambe, 1991; Small and McLeish, 1994; Liang and Zeger, 1995). The equation

$$\{g^*\}_{\hat{\theta}_2} = 0,$$

would provide a (nearly optimal) consistent estimate of θ_1.

Note in the forgoing discussion, *conditioning* is used just as a 'technique' to obtain (unconditionally) 'optimum' EFs; *but it is not used as a principle of inference.* In fact, without invoking any conditioning at all, Godambe and Thompson (1974) established, in case of the normal distribution $N(\theta_1,\theta_2)$, the optimality of the EF $(s^2 - \theta_2)$, for the interest parameter θ_2, ignoring the nuisance parameter θ_1. How this (unconditional) optimality leads to a very 'flexible conditioning' will be discussed later.

For a general perspective on the topic of conditioning and optimality we refer to Small and McLeish (1988), Lindsay and Waterman (1991) and Lindsay and Li (1995).

Lloyd (1987) and Bhapkar (1991) have given results concerning optimality of 'marginal SF' under 'conditional completeness'.

From the above discussion it is clear that the EF theory has corrected a major deficiency of the ML estimation in case of the nuisance parameters.

Some earlier references in respect of the nuisance parameters are Bartlett (1936), Cox (1958), Barnard (1963), Kalbfleisch and Sprott (1970), Barndorff-Nielsen (1973) and others. Some of these authors tried to obtain conditions under which the marginal distribution of t does not contain any information about θ_1 the parameter of interest. As we have seen the optimality criterion of the EF theory yields such a condition in terms of 'completeness of the statistic t'. Though not universally applicable (as none can be, I suppose) it by now has been commonly used for its mathematical manageability. It also carries with it greater conviction for it is derived from an optimality criterion which has proved to be fruitful very generally.

In the following we would show that the EF theory, just as it corrected ML estimation, also corrects some major inadequacies of the LS estimation and the GM theorem.

4 Quasi-Score Function

We now replace the abstract (observation) sample in the discussion by n real variates $x_i : i = 1, ..., n$ which are assumed to be independently distributed with means $\mu_i(\theta)$ and variances $v_i(\theta)$ (μ_i and v_i being some specified functions of θ) $i = 1, ..., n$. For simplicity let θ be a scalar parameter. Initially we consider the special case where μ_i are linear functions of θ and v_i are independent of θ. Here the LS equation is given by $\sum_1^n (x_i - \mu_i)(\frac{\partial \mu_i}{\partial \theta})/v_i = 0$. The solution of the equation, as said before, according to GM theorem, has smallest variance in the class of all linear unbiased estimates of θ; hence is 'optimal'. The estimating function $\sum (x_i - \mu_i)(\frac{\partial \mu_i}{\partial \theta})/v_i$ is also 'optimal' according to criterion (2.1), in the class of all EFs of the form

$$g = \sum_1^n (x_i - \mu_i) a_i \tag{4.1}$$

where a_i can be any arbitrary functions of θ. (Actually here we minimize $\mathcal{E}(g^2)$ subject to holding $\mathcal{E}(\partial g/\partial \theta) =$ const.. This will explain the standardization of EF mentioned earlier.) Note this EF optimality implies more than the GM optimality, for the solutions corresponding to all the equations $g = 0$ include not only all linear unbiased estimates of θ but many more.

Now let the means μ_i and variances v_i be arbitrarily specified functions of θ. Here the LS equation is given by $\overline{g} + B = 0$ where

$$\overline{g} = \sum_1^n (x_i - \mu_i) \frac{(\partial \mu_i / \partial \theta)}{v_i}$$

and (4.2)

$$B = \sum_1^n (x_i - \mu_i)^2 \frac{(\partial v_i / \partial \theta)}{v_i^2}.$$

Clearly in (4.2), $\mathcal{E}(\overline{g}) = 0$ and $\mathcal{E}(B) = \sum_1^n \partial \log v_i / \partial \theta$. Note for large n, $(\overline{g}/n) \simeq 0$ while (B/n) could still be very large. Hence because of the bias term (B) the LS equation $\overline{g} + B = 0$ would generally lead to an *inconsistent* estimate. On the other hand according to the optimality criterion (2.1) of the EF theory, in the class of EFs given by (4.1) for different functions $a_i(\theta)$, $i = 1, ..., n$, $\overline{g}$ given by (4.2) is 'optimal'. Generally the equation $\overline{g} = 0$ would lead to a consistent solution. (Here GM theorem cannot be of any avail for the solution of $\overline{g} = 0$ would generally lead to a biased estimate.) For reasons to be explained soon, we would call the estimating function $\overline{g}$ a *quasi-score function, (Quasi-SF).*

Interestingly the EF optimality of the quasi-SF $\overline{g}$ was first established in a wider setting of discrete stochastic processes with a martingale structure. **Quasi-SF Theorem** (Godambe 1985). If μ_i and v_i denote the means and variances of x_i conditional on the past observations $x_{i-1}, ..., x_0$ i.e. $\mu_i = \mu_i(\theta, x_0, ..., x_{i-1})$ and $v_i = v_i(\theta, x_0, ..., x_{i-1})$ for $i = 1, ..., n$ then

$$\overline{g} = \sum_{i=1}^n (x_i - \mu_i) \frac{\partial \mu_i / \partial \theta}{v_i}. \tag{4.3}$$

Here $\overline{g}$ is the optimal EF in the class of the EFs given by (4.1) where a_i are functions of $x_{i-1}, ..., x_0$ in addition to θ.

Among precursors to the EF $\overline{g}$ in (4.3) above are included the following: Durbin (1960) gave a GM theorem analogue for linear time series model. Klimko and Nelson (1978) obtained conditional LS equations. Kalbfleisch and Lawless (1983) suggested a special case of $\overline{g}$ for Markov models.

For further generalizations of the EF optimality results, relating to $\overline{g}$ in (4.3) we refer to Godambe (1985), Godambe and Heyde (1987), Godambe and Thompson (1989).

Returning back for simplicity to the case where the variates x_i, $i = 1, .., n$ are independently distributed, we summarize important properties of the EF $\overline{g}$ given by (4.2). The 'optimality' of $\overline{g}$ is for the semi-parametric model defined by means $\mu_i(\theta)$ and variances $v_i(\theta)$. As a special case when μ_i is linear in θ and v_i is independent of θ the EF optimality of $\overline{g}$ implies the GM optimality of the LS estimates. In this special case, if the underlying distribution is normal, the LS estimates coincide with the ML estimates. Generally, for the exponential family distributions the SF coincides with the optimal EF $\overline{g}$ given by (4.2). Even outside the exponential family of distributions, $\overline{g}$ satisfies a very general property of the SF; $\mathcal{E}(SF)^2 = -\mathcal{E}(\partial SF/\partial\theta)^2$ similarly we have $\mathcal{E}(\overline{g})^2 = -\mathcal{E}(\partial\overline{g}/\partial\theta)^2$. Further if SF denotes a generic 'score function' for the class of distributions consistent with the semi-parametric model mentioned above then $\mathcal{E}(\overline{g} - SF)^2 \leq \mathcal{E}(g - SF)^2$, (the expectation being taken w.r.t. the distribution that corresponds to the SF), for all the EFs g given by (4.1). These are the properties which justify the previously introduced term 'quasi-SF' for $\overline{g}$. (Even before the EF optimality of $\overline{g}$ was discovered the term *quasi-likelihood* was commonly used in the literature on *generalized linear models*, McCullagh and Nelder 1983, 1989).

As we have seen previously the EF theory corrected a major deficiency in ML estimation relating to nuisance parameters. The above discussion points to yet another accomplishment of the EF theory. It brought about, via quasi-score function $\overline{g}$, a kind of *synthesis* of two historically distinct methods of estimation: LS for semi-parametric models and ML for parametric models. The same criterion of the EF optimality namely (2.1) is satisfied in case of the latter by the SF and in case of the former by the quasi-SF $\overline{g}$ in (4.2). Only the classes of competing EFs are different. They are taken appropriate to the model (see Godambe and Thompson, 1989 Appendix). Of course, the forgoing discussion also shows that the quasi-SF $\overline{g}$ does provide, not only a unification (of the two methods LS and ML) but much more. It provides a *generalization* to deal with problems outside the scope of both LS and ML methods.

As a further contribution of the EF theory to statistics, below we briefly outline a very 'flexible conditioning' that the theory permits and the consequent incorporation of the Bayesian factor within its methodology.

5 A Generalization

It was mentioned earlier that within the framework of martingales and corresponding filtering, the EF theory suggested use of weighted conditional least square estimation, on grounds of its optimality property. But to deal with general *spatial processes* one needs more flexible conditioning than used before; this was provided by Godambe and Thompson (1989): Let as before $\mathcal{X} = \{x\}$ be an abstract sample space and $\mathcal{F} = \{F\}$ be a class of distributions on $\mathcal{X}$. Further let θ be a real parameter, defined on $\mathcal{F}$; $\{\theta(F), F \in \mathcal{F}\} = \Omega$. Now suppose h_j is a real function on $\mathcal{X} \times \Omega$ and $\mathcal{X}_j$ a specified partition (or a σ-field generated by a partition) of $\mathcal{X}$ such that

$$\mathcal{E}(h_j|\mathcal{X}_j) = 0, \qquad j = 1, ..., k. \tag{5.1}$$

The functions h_j, $j = 1, ..., k$ are called the elementary EFs; they are not exhaustive. Their choice is determined by the problem at hand. Now suppose the elementary EFs, $h_1, ..., h_k$ are mutually orthogonal (Def. Godambe and Thompson 1989) and the class of underlying distributions $\mathcal{F}$ satisfy certain conditions. Then in the class of all EFs g of the form

$$g = \sum_{j=1}^{k} h_j q_j \tag{5.2}$$

where q_j are some real functions on $\mathcal{X} \times \Omega$ which are measurable on $\mathcal{X}_j$, $j = 1, ..., k$ the 'optimal' one is given by

$$g^* = \sum_{j=1}^{k} h_j q_j^* \tag{5.3}$$

where $q_j^* = \{\mathcal{E}(\partial h_j / \partial\theta | \mathcal{X}_j)\} / \{\mathcal{E}(h_j^2 | \mathcal{X}_j)\}$. Here the *criterion* of optimality as always is *unconditional*, given by (2.1) for a real parameter θ (or its appropriate version if θ is a vector); the expectation is taken with respect to $F \in \mathcal{F}$.

Up to the above results the EF theory was 'restricted' to the classical setup where distributions on the sample space $\mathcal{X}$ for some *fixed* values of the parameters are considered. But the formalism of the EF optimality criterion is flexible enough and the just mentioned 'restriction' can be set aside if we know something about the prior distribution of θ; for instance its mean (θ_0) and variance (v_0). Under such Bayesian setup the only changes that are required to be done are as follows: (i) In (5.1) now $\mathcal{X}_j$ is not necessarily a partition of just the sample space $\mathcal{X}$, but it can be a partition of $\mathcal{X} \times \Omega$, Ω as before being the parameter space. (ii) Some elementary EFs h_j, $j = 1, ..., k$ can now be functions exclusively of the parameter θ. (iii) All expectations in the optimality criterion (2.1) are now with respect to the *joint* distributions of (x, θ) (and not as before with respect to distributions of x *given* θ). Following is an illustration.

Let the partitions of the sample space $\mathcal{X}_j$ and the elementary estimating functions hj, $j = 1, ..., k$ be the same as in (5.1). Further, as suggested before, let the mean value (θ_0) and the variance (v_0) of the prior distribution of θ be known. Now, to the set of elementary estimating functions $h_1, ..., h_k$ we add one more, namely $h_{k+1} = \theta - \theta_0$. In this case the optimal EF is given by $g^* + (\theta - \theta_0)/v_0$, where g^* is the same as in (5.3). Similarly now the quasi-SF $\overline{g}$ in (4.2), which was obtained under the assumption 'θ is fixed' will now have to be replaced by $\overline{g} - (\theta - \theta_0)/v_0$ (Godambe, 1994).

The 'optimality' of the EF given by the derivative of the logarithm of the posterior density was established in a 'parametric setup' by Ferreira (1982) and Ghosh (1993). Naik-Nimbalkar and Rajarshi (1995) have established some optimality results in 'semi-parametric Bayesian setup'.

6 Other Topics

Now, following are a few remarks (possibly only tangential) about the likelihoods: empirical, partial, profile, quasi and the like. Basically when the likelihood function is precisely known, with no nuisance parameters, likelihood ratio test is 'optimal' in the conventional sense of the term. Also the SF satisfies the EF criterion of 'optimality'. Now the various likelihoods, empirical partial, quasi just mentioned, try to 'approximate' the underlying (true, precise) likelihood in situations of nuisance parameters and/or of semiparametric models. In similar situations EF theory tries to 'approximate' the (true) underlying SF. However, unlike the former, the latter 'approximation' can be assessed with a plausible *finite sample* criterion. Suppose $g(x, \theta)$ is a real function of the sample x and the parameter of interest θ such that the expectation $\mathcal{E}(g) = 0$ for all possible underlying distributions F i.e. for $F \in \mathcal{F}$. Let further SF be a score function corresponding to F in $\mathcal{F}$. Then the finite sample criterion of assessing the approximation g for SF is given by $\mathcal{E}(g - SF)^2$, for all $F \in \mathcal{F}$. This criterion as said before leads to the 'optimality' criterion (2.1) of the EF theory. As I have previously shown optimal or approximately optimum EFs are found in many practical problems and in fact by now they are in common use. Now while optimum EFs and approximations thereof can provide a handy instrument for constructing confidence intervals and related tests cf. Rao's test (Rao 1947, Basawa 1991), for some other problems some kind of 'approximate likelihood' would be more handy. I think, to be safer, construction of such approximate likelihoods should be tied to the optimum EFs, whenever possible. It is good to note already a strong trend in that direction (Qin and Lawless, 1994).

An often asked question (cf. Liang and Zeger 1995) is how does the EF optimality relate to the properties of the corresponding estimate? *How good is the estimate*? Usually the answer is given in terms of the '*error*' of the estimate. Now this 'error' is somewhat of an involved concept. Certainly, error is not just a square root of an arbitrary (unbiased or nearly so) estimate of variance. However for a parametric model the concept is clear. *The error is derived from the conditional (or the natural estimate of) variance of the SF.* Thus error is the inverse of the square root of observed Fisher information (Efron and Hinkley, 1978). This methodology is formalized and extended by the EF theory. Consider the confidence intervals, $\hat{\theta} \pm \text{const(error)}$, where the estimate $\hat{\theta}$ is obtained from the unbiased estimating equation $g(\hat{\theta}) = 0$. Here a more direct way of obtaining the confidence intervals is by inverting the distribution of the standardized version (cf. Godambe 1991, eq. 40) of the EF g around $\hat{\theta}$. These intervals, compared to former ones, are easier to compute. Also *if g is the optimal EF, the corresponding intervals are shortest* compared to that of any other unbiased EF, (Godambe and Heyde 1987).

The standardizing factor of the EF g directly leads to the computation of the 'error' for the estimate $\hat{\theta}$ (Godambe, 1995).

For important previous review articles on the subject we refer to Heyde (1989) and Godambe and Kale (1991). The present review highlights some more recent developments and presents older results with different emphasis and interpretations. A further reference along this line is Desmond (1997).

References

Barnard, G.A. (1963). Some logical aspects of the fiducial argument. *J.R. Statist. Soc. B*, 25, 111-114.

Barndorff-Nielsen, O.E. (1973). On M-ancillarity. *Biometrika* 60, 447-455.

Bartlett, M.S. (1936). The information available in small samples. *Proc. Camb. Phil. Soc.*, 34, 33-40.

Basawa, I.V. (1991). Generalized score tests for composite hypotheses. *Estimating functions.* (ed. V.P. Godambe), Oxford Univ. Press, Oxford. 121-131.

Bhapkar, V.P. (1991). Sufficiency, ancillarity and information in estimating functions. *Estimating Functions.* (Ed. V.P. Godambe), Oxford Univ. Press, Oxford. 240-254.

Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* 29, 357-372.

Desmond, A.F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling (with discussion). *J. Stat. Plan. Inf.* 60, 77-121.

Durbin, J. (1960). Estimation of parameters in time series regression models. *J. Roy. Statist. Soc. B*, 22, 139-153.

Ferreira, P.E. (1982). Multiparametric estimating equations. *Ann. Stat. Math.* 34, 423-431.

Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Cambridge Phil. Soc.* 22, 700-706.

Ghosh, M. (1990). On a Bayesian analog of the theory of estimating functions. *C.G. Khatri Memorial Volume of Gujard Statistical Review*, 17A, 47-52.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* 31, 1208-1212.

Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63, 277-284.

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

Godambe, V.P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika* 78, 143-151.

Godambe, V.P. (1994). Linear Bayes and optimal estimation. Tech. Report STAT-94-11, University of Waterloo.

Godambe, V.P. (1995). Discussion of the paper, 'Inference Based on estimating functions in the presence of nuisance parameters' by Liang, K.Y. and Zeger, S.L. *Statistical Science* 10, 173-174.

Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *Int. Stat. Rev.* 55, 231-244.

Godambe, V.P. and Kale, B.K. (1991). Estimating functions: an overview. *Estimating Functions.* (Ed. V.P. Godambe), Oxford University Press, Oxford. 1-20.

Godambe, V.P. and Thompson, M.E. (1974). Estimating equations in presence of nuisance parameters. *Ann. Stat.* 2, 568-571.

Godambe, V.P. and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (With Discussion). *J. Stat. Plan. Inf.* 22, 137-172.

Heyde, C.C. (1989). Quasi-likelihood and optimality of estimating functions: some current unifying themes. *Bull. Int. Stat. Inst.* Book 1, 19-29.

Kalbfleisch, J.D. and Sprott, D.A. (1970). Applications of likelihood methods to models involving large number of parameters. (With Discussion). *J.R. Statist. B*, 32, 175-208.

Kalbfleisch, J.D., Lawless, J.F. and Vollmer, W.M. (1983). Estimation in Markov models from aggregate data. *Biometrics* 39, 907-919.

Klimko, L.A. and Nelson, P.I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* 6, 629-642.

Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes.* Paris: Courcier.

Liang, K.Y. and Zeger, S.L. (1995). Inference based on estimating functions in the presence of nuisance parameters. (With Discussion). *Statistical Science* 10, 158-195.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika* 69, 503-512.

Lindsay, B. and Waterman, R.P. (1991). Extending Godambe's method in nuisance parameter problems. *Proceedings of a Symposium in honour of Prof. V.P. Godambe.* University of Waterloo, 1-43.

Lindsay, B.G. and Li, B. (1995). Discussion of the paper, 'Inference based on estimating functions in the presence of nuisance parameters' by Liang, K.Y. and Zeger, S.L. *Statistical Science* 10, 175-177.

Lloyd, C.J. (1987). Optimality of marginal likelihood estimating equations. *Comm. Stat. Theory and Meth.* 16, 1733-1741.

McCullagh, P. and Nelder, J.A. (1983, 1989). *Generalized linear models* (1st and 2nd editions). Chapman and Hall, London.

Naik-Nimbalkar, U.V. and Rajarshi, M.B. (1995). Filtering and smoothing via estimating functions. *J. Amer. Statist. Asso.* 90, 301-306.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, 16, 1-32.

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

Rao, C.R. (1947). Large sample tests for statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Camb. Phil. Soc.* 44, 50-57.

Small, C. and McLeish, D.L. (1988). *The theory and applications of statistical inference functions.* Lecture Notes in statistics No. 44, Springer Verlag. Heidelberg, New York, London.

Small, C. and McLeish, D.L. (1994). *Hilbert space methods in probability and statistical inference.* John Wiley and Sons, Inc. New York.