Chin. Ann. of Math.
16B: 4(1995),487-498.

GEOMETRIC METHOD OF SEQUENTIAL ESTIMATION RELATED TO MULTINOMIAL DISTRIBUTION MODELS***

Wei Bocheng* Li Shouye**

Abstract

In 1980's, differential geometric methods are successfully used to study curved exponential families and normal nonlinear regression models. This paper presents a new geometric structure to study multinomial distribution models which contain a set of nonlinear parameters. Based on this geometric structure, the authors study several asymptotic properties for sequential estimation. The bias, the variance and the information loss of the sequential estimates are given from geometric viewpoint, and a limit theorem connected with the observed and expected Fisher information is obtained in terms of curvature measures. The results show that the sequential estimation procedure has some better properties which are generally impossible for nonsequential estimation procedures.

Keywords Multinomial distribution model, Statistical curvature, Sequential estimation, Stopping rule, Fisher information, Information loss.
1991 MR Subject Classification 62F.

§1. Introduction

Multinomial distributions are widely used in statistical inference. In this paper, we present a differential geometric method to study a kind of multinomial distribution models in which a set of nonlinear parameters are of interest. The differential geometric method in statistical inference had been well studied in 1980's and a nice review was given by Kass[1]. There are two approaches and two kinds of models which are widely accepted and proved to be successful in statistical analysis. One was given by Efron[2] and Amari[3] for curved exponential families by introducing a Riemannian geometric framework; the other was given by Bates and Watts[4] for nonlinear regression models by proposing two kinds of curvature measures in Euclidean space. The geometry and models discussed in this paper are different from those of Efron and Amari (EA) and Bates and Watts (BW). The models we study are close to those of EA, but here the multinomial distribution model can not be regarded as a full, regular and minimally represented exponential family and may cause some problems from EA viewpoint[5]. The geometry we present is close to that of BW, but we introduce a Fisher information inner product as a metric to Euclidean space. Besides, the models, and the problems of sequential maximum likelihood estimates we shall study in this paper, are substantially different from nonlinear regression models studied by BW. We try to combine

Manuscript received May 13, 1993.
*Department of Mathematics, Southeast University, Nanjing 210018, China.
**Jiangsu Business Management College, Nanjing 210000, China.
***Project supported by the National Natural Science Foundation of China.

the advantages of both EA and BW approaches, and apply our geometric method to study asymptotics of sequential estimation, which have not been seen in the literature.

In Section 2, we first review some basic results for multinomial distributions, and then propose a differential geometric framework in Euclidean inner product space for these models. In Section 3, the stochastic expansions for the sequential estimates are obtained from geometric viewpoint. Section 4 studies the information loss, the bias and the asymptotic variance of the sequential estimate. A limit theorem connected with the observed and expected Fisher information is also given in terms of curvature measures.

§2. Geometry of Multinomial Distribution Models

Let $\{\mathbf{X}^k\}$ be a set of n independent $m \times 1$ vectors where each $\mathbf{X}^k = ({x_1}^k, \cdots, {x_m}^k)^T$ satisfies

$$P(x_i^k = 1) = \pi_i, \quad P(x_i^k = 0) = 1 - \pi_i, \quad 0 < \pi_i < 1,$$

$$\sum_i \pi_i = 1, \quad \sum_i x_i^k = 1 \ (k = 1, \cdots, n; \ i = 1, \cdots, m).$$

Then the joint density function of $\mathbf{X} = (\mathbf{X}^1, \cdots, \mathbf{X}^n)^T$ has the form

$$P(x, \pi) = \prod_{i=1}^m \pi_i^{y_i} \quad , y_i = \sum_{i=1}^m x_i^k, \tag{2.1}$$

where $\mathbf{Y} = (y_1, \cdots, y_m)^T = \sum_k \mathbf{X}^k$ is a sufficient statistic. It is easily seen that

$$E(\mathbf{X}^k) = \pi, \quad \mathrm{Var}(\mathbf{X}^k) = \Phi, \quad \Phi = g - \pi\pi^T, \tag{2.2}$$

where $\pi = (\pi_1, \cdots, \pi_m)^T$, $g = \mathrm{diag}(\pi_1, \cdots, \pi_m)$. Let the log likelihood of $\mathbf{X}$ be $l(\pi) = \log[P(x, \pi)]$. Then the score function $\dot{l}$ and the observed information $-\ddot{l}$ of $\mathbf{X}$ for model (2.1) satisfy

$$\dot{l}(\pi) = g^{-1}\mathbf{Y}, \quad -\ddot{l}(\pi) = g^{-2}\mathrm{diag}(y_1, \cdots, y_m),$$

where $\mathbf{Y} = (y_1, \cdots, y_m)^T$. In what follows, dots over the functions will denote the derivatives. Note that since there is constraint on π: $\sum_i \pi_i = 1$, equations $E[\dot{l}(\pi)] = 0$ and $E[-\ddot{l}(\pi)] = \mathrm{Var}[\dot{l}(\pi)]$ do not hold. This is not the regular case. Now we assume that for model (2.1), π is a function of a vector parameter $\theta = (\theta_1, \cdots, \theta_p)^T$ of interest ($p \leq m - 1$). This is a commonly encountered situations (see, for example, [6] and [1]). In this case, model (2.1) can be denoted by

$$P(x, \pi(\theta)) = \prod_{i=1}^m \{\pi_i(\theta)^{y_i}\}, \tag{2.3}$$

$$\pi = \pi(\theta), \quad \sum_{i=1}^m \pi_i(\theta) = 1. \tag{2.4}$$

Our discussions will be based on this model. We assume that $\pi(\theta)$ is thrice continuously differentiable with respect to θ in some neighborhood $\Theta_\circ$ of parameter space Θ. The first two derivatives of $\pi(\theta)$ are denoted by

$$D(\theta) = \frac{\partial \pi}{\partial \theta^T}, \quad W = \frac{\partial^2 \pi}{\partial \theta \partial \theta^T},$$

where $D(\theta)$ is an $m \times p$ full rank matrix and $W(\theta)$ is an $m \times p \times p$ array with elements $D_{ia} = \frac{\partial \pi_i}{\partial \theta_a}$ and $W_{iab} = \frac{\partial^2 \pi_i}{\partial \theta_a \partial \theta_b}$ respectively, where $i = 1, \cdots, m$ and $a, b = 1, \cdots, p$. By equation (2.4), $D(\theta)$ and $W(\theta)$ have the following specific properties:

$$\pi^T g^{-1} D = \mathbf{1}^T D = 0, \qquad [\pi^T g^{-1}][W] = [\mathbf{1}^T][W] = 0, \tag{2.5}$$

where $\mathbf{1} = (1, \cdots, 1)^T$ is an n-vector, $[\cdot][\cdot]$ indicates array multiplication as defined in [4] (1980). For the sake of simplicity, we denote $l(\pi(\theta))$ and $g(\pi(\theta))$ by $l(\theta)$ and $g(\theta)$ respectively. Similar notation are used for some other quantities in the rest of this paper. It follows from (2.3)–(2.5) that the score function $\dot{l}(\theta)$ and the observed information -$\ddot{l}(\theta)$ of $\mathbf{X}$ for model (2.1) satisfy

$$\dot{l} = D^T g^{-1}(\theta) r(\theta), \quad E[\dot{l}(\theta)] = 0, \tag{2.6}$$

$$r(\theta) = \mathbf{Y} - n\pi(\theta), \quad (\overline{\mathbf{Y}} - \pi(\theta)) \to 0 \quad \text{(a.s.)}, \tag{2.7}$$

$$n^{-\frac{1}{2}} r(\theta) \xrightarrow{L} N(0, \Phi), \quad \Phi = g - \pi\pi^T \quad \text{as} \quad n \to \infty,$$

where $\overline{\mathbf{Y}} = n^{-1}\mathbf{Y}$, "a.s." denotes the convergence almost surely, "L" denotes the convergence by law. Thus from the above equations we have

$$-\ddot{l}(\theta) = nD^T g^{-1} D - [r^T g^{-1}][W - \Gamma], \tag{2.8}$$

$$\Gamma = D^T g^{-1} G g^{-1} D, \tag{2.9}$$

where G is an $m \times m \times m$ array with $G_{iii} = \pi_i$ and zeros elsewhere.

Now let $\hat{\theta}$ be the maximum likelihood estimate of θ based on $\mathbf{X}$. We assume that -$\ddot{l}(\theta)$ is positive definite in a neighborhood $\Theta_\circ$ of $\hat{\theta}$ in Θ. It follows from equations (2.6) that $\hat{\theta}$ satisfies

$$D^T(\hat{\theta}) g^{-1}(\hat{\theta}) r(\hat{\theta}) = 0. \tag{2.10}$$

The geometric interpretation of (2.10) can be described as follows: In Euclidean space R^m, the "residual vector" $\hat{r} = r(\hat{\theta})$ is orthogonal to all column vectors of $D(\hat{\theta})$ with respect to the matrix $g^{-1}(\hat{\theta})$ inner product.

Now we can introduce a geometric framework for model (2.1) based on this interpretation.

Take $\pi = (\pi_1, \cdots, \pi_m)^T$ as a coordinate in Euclidean space R^m. Then $z_n = n\pi(\theta)$ denotes a p-dimensional surface in R^m, which is denoted by M_n and may be called the solution locus (see [4]). It is easily seen that the tangent space T_θ of M_n at θ is spanned by the columns of $D(\theta)$. For any two vectors a and b in R^m, we define an inner product as $\langle a, b \rangle = a^T g^{-1}(\theta) b$. Under this inner product, the corresponding normal space is denoted by T'_θ. Now we can define curvatures for the solution locus $z = n\pi(\theta)$. To this aim, we choose orthonormal basis for spaces T_θ and T'_θ. Suppose that the QR decomposition of $D(\theta)$ is given by

$$D(\theta) = (Q, N) \begin{pmatrix} R \\ 0 \end{pmatrix} = QR,$$

where R and $L = R^{-1}$ are $p \times p$ nonsingular upper triangular matrices and the columns of Q and N are orthonormal basis for the tangent space and normal space of solution locus $z_n = n\pi(\theta)$ at θ in R^m. The matrices Q and N satisfy $Q^T g^{-1} Q = I_p$, $Q^T g^{-1} N = 0$, $N^T g^{-1} N = I_{m-p}$, where I_m and I_{m-p} are identity matrices of order p and $m-p$ respectively. For the solution locus $z_n = n\pi(\theta)$, we have $D_n = \frac{\partial (n\pi(\theta))}{\partial \theta^T} = nD = QR_n$, $R_n = nR$;

$W_n = \frac{\partial^2(n\pi(\theta))}{\partial\theta\partial\theta^T} = nW$. Denoting $U_n = (L_n)^T W_n L_n$, where $L_n = (R_n)^{-1}$, we define the intrinsic curvature array A_n^I and the parameter-effects curvature array A_n^P as follows

$$A_n^I = [N^T g^{-1}][U_n], \quad A_n^P = [Q^T g^{-1}][U_n].$$

It follows that

$$A_n^I = n^{-1}A^I, \quad A_n^P = n^{-1}A^P, \tag{2.11}$$

$$A^I = [N^T g^{-1}][U], \quad A^P = [Q^T g^{-1}][U], \tag{2.12}$$

where $U = L^T W L$. A^I and A^P are curvature arrays for the surface $z = \pi(\theta)$. If the sample size n is fixed, the solution locus $z_n = n\pi(\theta)$ can be regarded as a similar transformation of the surface $z = \pi(\theta)$. It is easy to see from (2.5) that the unit vector $\pi(\theta)$ is always in normal space T'_θ. Then we can set $N = (\pi(\theta), N_1)$, $\pi^T g^{-1} N_1 = 0$.

By (2.10), the residual vector $\hat{r}$ is in normal space of solution locus at $\hat{\theta}$ so that $\hat{r}$ can be represented as

$$\hat{r} = \mathbf{Y} - n\pi(\hat{\theta}) = nN(\hat{\theta})\hat{\lambda} \ , \ \ \overline{\mathbf{Y}} - \pi(\hat{\theta}) = N(\hat{\theta})\hat{\lambda}, \tag{2.13}$$

where $n\hat{\lambda}$ is the coordinate of $\hat{r}$ in normal space in which the columns of $N(\hat{\theta})$ are an orthonormal basis. So we may extend (2.13) in following way (see [15]).

Let u be an arbitrary point in R^m by the solution locus M_n. Then under some conditions there exists a point $n\pi(\theta)$ on M_n such that $u - n\pi(\theta) = nN(\theta)\lambda$. Denoting $\omega = (\theta, \lambda)$, $\hat{\omega} = (\hat{\theta}, \hat{\lambda})$, $\omega_\circ = (\theta, 0)$, we have

$$\begin{aligned} &u = nh(\omega) = n\pi(\theta) + nN(\theta)\lambda, \\ &\overline{\mathbf{Y}} = h(\hat{\omega}) = h(\hat{\theta}, \hat{\lambda}), \quad \pi(\theta) = h(\omega_0) = h(\theta, 0). \end{aligned} \tag{2.14}$$

In this paper, we suppose that the following conditions hold:

(A) λ is defined on an open set Λ which contains $\lambda = 0$. Denote the closure of Λ by $\overline{\Lambda}$.

(B) $\pi(\theta)$ and $N(\theta)$ are continuous on Θ and thrice differentiable with respect to θ in Θ and $\partial h(\omega)/\partial\omega^T$ is a nonsingular matrix.

Under the above assumptions, ω can be uniquely represented as $\omega = g(u)$, where $g(\cdot)$ is the inverse function of $h(\cdot)$. In particular, we have

$$\hat{\omega} = g(\overline{\mathbf{Y}}), \quad \omega_0 = g(\pi(\theta)). \tag{2.15}$$

It follows from (2.7) that

$$(\hat{\theta}, \hat{\lambda}) \to (\theta, 0), \quad \text{(a.s.)} \quad \text{as } n \to \infty.$$

Lemma 2.1. *Let $F = (F_{via}) = \frac{\partial N_{iv}}{\partial\theta_a}$ and $N = (N_{iv})$ $(i = 1, \cdots, m; \ \ a = 1, \cdots, p; \ v = 1, \cdots, m - p)$. Then under the conditions stated above we have*

$$D^T g^{-1} F = -R^T \widetilde{A}^I R, \tag{2.16}$$

$$\widetilde{A}^I = A^I - \Gamma^I, \quad \Gamma^I = [N^T g^{-1}][L^T \Gamma L]. \tag{2.17}$$

Proof. Denote $D = (D_{ia})$. Then $D^T g^{-1} N = 0$, which implies

$$\sum_{i=1}^{m} D_{ia} N_{iv} \pi_i^{-1} = 0$$

for any a and v. Differentiating these equations with respect to θ and denoting $W = (W_{iab})$ give

$$\sum_{i=1}^{m}(W_{iab}N_{iv}\pi_i^{-1} + D_{ia}F_{vib}\pi^{-1} - \pi_i^2 D_{ia}D_{ib}N_{iv}) = 0,$$

this equation implies

$$[N^T g^{-1}][W] + D^T g^{-1}F - [N^T g^{-1}][D^T g^{-1}Gg^{-1}D] = 0.$$

Thus we obtain

$$D^T g^{-1}F = -R(A^I - \Gamma^I)R.$$

Note that $\widetilde{A} = A^I - \Gamma^I$ is also an intrinsic curvature. In fact, let $\gamma_i(\theta) = \log\pi_i(\theta)$, $(i = 1, \cdots, m)$, $\gamma(\theta) = (\gamma_1(\theta), \cdots, \gamma_m(\theta))^T$, then $\gamma(\theta)$ can be regarded as the dual parameter of $\pi(\theta)$ (see [5]). Now take γ as a coordinate in R^m and define an inner product for any two vectors a and b as $\langle a, b\rangle = a^T gb$. The solution locus M_γ is defined as $\gamma = \gamma(\pi(\theta))$ in this space. Then the tangent space of M_γ at θ is spanned by columns of $D_\gamma = \partial\gamma/\partial\theta^T$. We can define curvature arrays as follows. Suppose that the QR decomposition of D_γ with respect to above inner product is $D_\gamma = (Q_\gamma, N_\gamma)(R_\gamma^T, 0) = Q_\gamma R_\gamma$ which satisfies $Q_\gamma^T gQ_\gamma = I_p$, $Q_\gamma^T gN_\gamma = 0$, $N_\gamma^T gN_\gamma = I_{n-p}$ respectively. Then A_γ^I and A_γ^P are respectively defined as

$$A_\gamma^I = [N_\gamma^T g][U_\gamma], \quad A_\gamma^P = [Q_\gamma^T g][U_\gamma], \quad U_\gamma = L_\gamma^T W_\gamma L_\gamma,$$

where $L_\gamma = R_\gamma^{-1}$ and $W_\gamma = \frac{\partial^2\gamma}{\partial\theta\partial\theta^T}$. By direct calculations we have

$$D_\gamma = g^{-1}D, \quad W_\gamma = [g^{-1}][W - \Gamma], \quad Q_\gamma = g^{-1}Q,$$

$$R_\gamma = R, \quad N_\gamma = g^{-1}N, \quad U_\gamma = [g^{-1}][U - L^T\Gamma L].$$

Therefore the curvature arrays A_γ^I and A_γ^P can be expressed as

$$\begin{aligned} A_\gamma^I &= \tilde{A}^I = A^I - \Gamma^I, \quad \Gamma^I = [N^T g^{-1}][L^T\Gamma L], \\ A_\gamma^P &= A^P - \Gamma^P, \quad \Gamma^P = [Q^T g^{-1}][L^T\Gamma L]. \end{aligned} \tag{2.18}$$

§3. Sequential Estimation and Stochastic Expansions

Sequential estimation procedures can be defined in two stages (a) definition of a stopping rule, and (b) definition of the estimation procedure once the stopping rule is determined. Following [8] and [9], we adopt a stopping rule by which the random sample size n satisfies

$$E_\theta(n) = K\nu_1(\pi(\theta)) + O(1),$$

where K is a large number playing the role of the average number of observations and $\nu_1(\cdot)$ is a smooth positive scalar function (see [7],[8] and [9]). Similar to [9], we assume that the number n of observations is determined by our stopping rule such that

$$n(K) = K\nu_1(\overline{\mathbf{Y}}) + c_1(\overline{\mathbf{Y}}) + \varepsilon$$

holds, where $c_1(\overline{\mathbf{Y}})$ is a function of order 1 and ε is a small order term asymptotically independent of $\overline{\mathbf{Y}}$ satisfying $E(\varepsilon) = o(1)$.

By equation (2.14) and assumptions stated above, $n(K)$ can be represented as

$$n(K) = K\nu(\hat{\theta}, \hat{\lambda}) + c(\hat{\theta}, \hat{\lambda}) + \varepsilon, \tag{3.1}$$

where $\nu(\hat{\theta},\hat{\lambda}) = \nu_1(h(\hat{\theta},\hat{\lambda}))$ and $c(\hat{\theta},\hat{\lambda}) = c_1(h(\hat{\theta},\hat{\lambda}))$ and $\hat{\theta}$ is the maximum likelihood estimator of θ.

From geometric point of view, when sample size is fixed, the surface $z_n = n\pi(\theta)$ is a similar transformation of $z = \pi(\theta)$. Thus by (2.11) the curvatures of the surface $z_n = n\pi(\theta)$ are $1/n$ of that of $z = \pi(\theta)$. In sequential case, however, the sample size n is a random variable so that the surface $z_n = n\pi(\theta)$ is the conformal transformation of $z = \pi(\theta)$ and it is possible from geometrical viewpoint to determine n such that the sequential estimation procedure has some better properties which are generally impossible for nonsequential estimation procedure.

By (2.1) and (2.7), $r(\theta)$ can be represented as

$$r(\theta) = \sum_{i=1}^{n} T_i = \sum_{i=1}^{n} (\mathbf{X}^i - \pi(\theta)). \tag{3.2}$$

It is easy to see from Wald identity[10] that

$$E(r) = 0, \quad \mathrm{Var}(r) = K\nu[\pi + O(K^{-1})], \tag{3.3}$$

$$J(\mathbf{X}) = E[-\ddot{l}(\theta)] = K\nu[D^T g^{-1} D + O(K^{-1})]. \tag{3.4}$$

Lemma 3.1. *Under the assumptions stated above we have*

$$K^{-1}n(K) \xrightarrow{p} \nu(\theta, 0) \quad \textit{and} \quad e = [K\nu(\theta,0)]^{-\frac{1}{2}} r \xrightarrow{L} N(0, \Phi) \tag{3.5}$$

as $K \to \infty$, where "p" denotes the convergence in probability.

Proof. Let $\Omega = \{\mathbf{X} = (\mathbf{X}^1, \cdots, \mathbf{X}^n, \cdots,): \inf_{\mathbf{X}} [\nu(\hat{\theta},\hat{\lambda})] \geq 0\}$. It follows from (2.7) and (2.14) that

$$\overline{\mathbf{Y}} \to \pi(\theta) \ \ \text{(a.s.)}, \qquad \nu(\hat{\theta},\hat{\lambda}) \to \nu(\theta, 0) > 0 \ \ \text{(a.s.)} \quad \text{as} \ \ n \to \infty,$$

which implies that $P_\theta(\Omega) = 1$. Hence it is easy to see from (3.1) that $n(K) \to \infty$ and $K^{-1}n(K) \xrightarrow{p} \nu(\theta, 0)$ as $K \to \infty$. Then we obtain (3.5) by using Lemma 4.3.6 in [11].

It is easily seen from (3.5) that

$$e = O_p(1), \quad (K\nu)^{\frac{1}{2}}[\overline{\mathbf{Y}} - \pi(\theta)] = O_p(1),$$

and it follows from (2.15) that

$$(\hat{\theta},\hat{\lambda}) - (\theta, 0) = g(\overline{\mathbf{Y}}) - g(\pi(\theta)) = O_p(K^{-\frac{1}{2}}),$$

$$\overline{\Delta\theta} = (K\nu)^{\frac{1}{2}}(\theta - \hat{\theta}) = O_p(1), \quad \overline{\lambda} = (K\nu)^{\frac{1}{2}}\hat{\lambda} = O_p(1).$$

Now we derive the stochastic expansions for the sequential ML estimator $\hat{\theta}$ from geometrical viewpoint.

Theorem 3.1. *Under the assumptions stated above, the second order expansion of $\hat{\theta}$ may be given by*

$$\Delta\theta = (K\lambda)^{-\frac{1}{2}} L\tau + (K\lambda)^{-1} L\{[\eta^T][\tilde{A}^I]\tau - \frac{1}{2}\tau^T A^P \tau - (\tau^T \bar{s_1})\tau - (\eta^T s_2)\tau\} + O_p(K^{-\frac{3}{2}}), \tag{3.6}$$

where $\tau = Q^T g^{-1} e$ and $\eta = N^T g^{-1} e$ are uncorrelated and τ has asymptotically normal distribution $N(0, I_p)$; $\eta^T = (\eta_1, \eta_2^T)$, $\eta_1 = 0$ (a.s.) and η_2 has asymptotically normal distribution $N(0, I_{m-p-1})$; $\bar{s}_1 = L^T s_1$, $s_1 = \frac{\partial(\log\nu)}{\partial\theta}$; $s_2 = \frac{\partial(\log\nu)}{\partial\lambda}$; all the quantities are evaluated at $(\theta, 0)$.

Proof. It follows from (3.5) and (2.13) that

$$\begin{aligned} e &= n(K\nu)^{-\frac{1}{2}}\{\overline{Y} - \pi(\hat{\theta}) + \pi(\hat{\theta}) - \pi(\theta)\} \\ &= n(K\nu)^{-1}\{N(\hat{\theta})\overline{\lambda} + D(\theta)\overline{\Delta\theta} + \frac{1}{2}(K\nu)^{-\frac{1}{2}}(\overline{\Delta\theta})^T W(\overline{\Delta\theta}) + O_p(K^{-1})\}. \end{aligned}$$

Expanding $N(\hat{\theta})$ and (3.1) we obtain

$$\begin{aligned} N(\hat{\theta}) &= N(\theta) + (K\nu)^{-\frac{1}{2}}F(\theta)\overline{\Delta\theta} + O_p(K^{-1}), \\ n(K\nu)^{-1} &= 1 + (K\nu)^{-\frac{1}{2}}(\overline{\Delta\theta}^T s_1 + \overline{\lambda}^T s_2) + O_p(K^{-1}). \end{aligned} \tag{3.7}$$

Substituting these two equations into the expression of e gives

$$\begin{aligned} e = N\overline{\lambda} + D\overline{\Delta\theta} + (K\nu)^{-\frac{1}{2}}[\overline{\lambda}^T][F]\overline{\Delta\theta} + \frac{1}{2}(K\nu)^{-\frac{1}{2}}(\overline{\Delta\theta})^T W(\overline{\Delta\theta}) \\ + (K\nu)^{-\frac{1}{2}}(\overline{\Delta\theta}^T \bar{s_1} + \overline{\lambda}^T s_2)(N\overline{\lambda} + D\overline{\Delta\theta}) + O_p(K^{-1}). \end{aligned} \tag{3.8}$$

Multiplying this equation by $Q^T g^{-1}$ and $N^T g^{-1}$ respectively gives

$$\overline{\Delta\theta} = L\tau + O_p(K^{-\frac{1}{2}}), \quad \overline{\lambda}^T = \eta + O_p(K^{-\frac{1}{2}}). \tag{3.9}$$

Substituting these equations into quadratic terms of (3.8) gives

$$\begin{aligned} e = N\overline{\lambda} + D\overline{\Delta\theta} + (K\nu)^{-\frac{1}{2}}[\eta^T][F]L\tau + \frac{1}{2}(K\nu)^{-\frac{1}{2}}\tau^T U\tau \\ + (K\nu)^{-\frac{1}{2}}(\tau^T L^T s_1 + \eta^T s_2)(N\eta + Q\tau) + O_p(K^{-1}). \end{aligned}$$

Multiplying this equation by $Q^T g^{-1}$ gives

$$\begin{aligned} \tau = R\overline{\Delta\theta} + (K\nu)^{-\frac{1}{2}}[\eta^T][Q^T g^{-1}F]L\tau + \frac{1}{2}(K\nu)^{-\frac{1}{2}}(Q^T g^{-1})(\tau^T U\tau) \\ + (K\nu)^{-\frac{1}{2}}(\tau^T L^T s_1 + \eta^T s_2)\tau + O_p(K^{-1}). \end{aligned}$$

Using (2.12) and (2.16), we can obtain (3.6) from the above equations. The asymptotic normality of τ and η can be obtained from (3.5). The asymptotic variances of τ and η are given by

$$\begin{aligned} \mathrm{Var}(\tau) &= Q^T g^{-1}(g - \pi\pi^T)g^{-1}Q = I_p, \\ \mathrm{Var}(\eta) &= N^T g^{-1}(g - \pi\pi^T)g^{-1}N = I_{m-p} - N^T g^{-1}\pi\pi^T g^{-1}N \\ &= I_{m-p} - \mathrm{diag}(1, 0, \cdots, 0). \end{aligned}$$

Hence we have $\mathrm{Var}(\eta_1) = 0$, $\eta_1 = 0$ (a.s.). Similarly, we have $\mathrm{Cov}(\tau, \eta) = 0$ so that τ and η are asymptotically independent. Then the theorem is proved.

§4. Some Characteristics of Sequential Estimates Related to Statistical Curvatures

By Theorem 3.1 we have

(1) $(K\nu)^{\frac{1}{2}}(\hat{\theta} - \theta) = \overline{\Delta\theta} = (D^T g^{-1}D)^{-1}D^T g^{-1}e + O_p(K^{-\frac{1}{2}})$ has asymptotically normal distribution $N(0, (D^T g^{-1}D)^{-1})$.

(2) The asymptotic distribution of $\hat{\lambda}$ does not depend on θ so that $\hat{\lambda}$ is an asymptotic ancillary statistic.

(3) The bias of MLE $\hat{\theta}$ can be given by

$$\mathrm{bias}(\hat{\theta}) = E(\hat{\theta} - \theta) = -(2K\nu)^{-1}L\{tr[A^P] + 2L^T s_1\} + O_p(K^{-\frac{3}{2}}),$$

where $\mathrm{tr}[A^P] = (\mathrm{tr}(A_1^P), \cdots, \mathrm{tr}(A_p^P))^T$, and A_i is the i-th face of A^P $(i = 1, \cdots, p)$.

Now we study some asymptotic properties of sequential MLE $\hat{\theta}$ related to the statistical curvatures.

For our models, the information loss of $\hat{\theta}$ is defined as

$$\Delta J(\hat{\theta}) = J(\mathbf{X}) - J(\hat{\theta}),$$

where $J(\mathbf{X})$ and $J(\hat{\theta})$ are Fisher information matrices of the observed vector $\mathbf{X}$ and the MLE $\hat{\theta}$ respectively. It can be shown that $\Delta J(\hat{\theta})$ can be expressed as ([2], [3])

$$\Delta J(\hat{\theta}) = E_\theta\Big\{\mathrm{Var}_\theta\Big(\frac{\partial l}{\partial \theta}\Big|\hat{\theta}\Big)\Big\}.$$

Theorem 4.1. *Under the assumptions stated above, the information loss of $\hat{\theta}$ may be given by*

$$\Delta J(\hat{\theta}) \simeq R^T A_{IS} R, \tag{4.1}$$

$$A_{IS} = \sum_{i=2}^{m-p} \{(\widetilde{A}_i^I)^2 + s_{21}^2 I_p - 2 s_{2i} \widetilde{A}_i^I\}, \tag{4.2}$$

where $\widetilde{A}_i^I$ is the i-th face of $\widetilde{A}^I$, and ${}_s2i$ is the i-th component of s_2.

Proof. Now we first derive the stochastic expansion for $\frac{\partial l}{\partial \theta}$. It follows from (2.6) and (3.5) that

$$\frac{\partial l}{\partial \theta} = D^T g^{-1} r = (K\nu)^{\frac{1}{2}} D^T g^{-1} e.$$

The substitution of (3.8) into this equation gives

$$\begin{aligned}
\frac{\partial l}{\partial \theta} &= (K\nu)^{\frac{1}{2}} D^T g^{-1} D\overline{\Delta\theta} + [\overline{\lambda}^T][D^T g^{-1} F]\overline{\Delta\theta} + \frac{1}{2} D^T g^{-1}\{(\overline{\Delta\theta})^T W(\overline{\Delta\theta})\} \\
&\quad + (\overline{\Delta\theta}^T s_1 + \overline{\lambda}^T s_2) D^T g^{-1} D\overline{\Delta\theta} + O_p(K^{-\frac{1}{2}}) \\
&= [\eta^T][-R^T \tilde{A}^I R]\overline{\Delta\theta} + s_2 R^T R\overline{\Delta\theta} + (K\nu)^{\frac{1}{2}} R^T R\overline{\Delta\theta} \\
&\quad + (\overline{\Delta\theta})^T s_1 R^T R\overline{\Delta\theta} - \frac{1}{2} D^T g^{-1}\{(\overline{\Delta\theta})^T W(\overline{\Delta\theta})\} + O_p(K^{-\frac{1}{2}}).
\end{aligned}$$

Substituting this equation into (4.11) gives

$$\begin{aligned}
\mathrm{Var}\Big(\frac{\partial l}{\partial \theta}\Big|\hat{\theta}\Big) &\simeq \mathrm{Var}\{[\eta^T][-R^T \widetilde{A}^I R]\overline{\Delta\theta} + \eta^T s_2 R^T R\overline{\Delta\theta})|\hat{\theta}\} \\
&\simeq \mathrm{Var}\Big\{\Big(\sum_{i=2}^{m-p} R^T \widetilde{A}_i^I R\overline{\Delta\theta}\eta_i\Big)\Big|\hat{\theta}\Big\} + \mathrm{Var}\Big\{\Big(\sum_{i=2}^{m-p} R^T R\overline{\Delta\theta} s_{2i}\eta_i\Big)\Big|\hat{\theta}\Big\} \\
&\quad - \mathrm{Cov}\Big\{\Big(\sum_{i=2}^{m-p} R^T \widetilde{A}_i^I R\overline{\Delta\theta}\eta_i, \sum_{i=2}^{m-p} R^T R\overline{\Delta\theta} s_{2i}\eta_i\Big)\Big|\hat{\theta}\Big\} \\
&\quad - \mathrm{Cov}\Big\{\Big(\sum_{i=2}^{m-p} R^T R\overline{\Delta\theta} s_{2i}\eta_i, \sum_{i=2}^{m-p} R^T \widetilde{A}_i^I R\overline{\Delta\theta}\eta_i\Big)\Big|\hat{\theta}\Big\} \\
&\simeq \sum_{i=2}^{m-p} \{R^T \widetilde{A}_i^I R\overline{\Delta\theta}(\overline{\Delta\theta})^T R^T \widetilde{A}_i^I R + R^T R\overline{\Delta\theta}(\overline{\Delta\theta})^T R^T R s_{2i}^2 \\
&\quad - R^T \widetilde{A}_i^I R\overline{\Delta\theta}(\overline{\Delta\theta})^T s_{2i}^2 - R^T R\overline{\Delta\theta}(\overline{\Delta\theta})^T R^T \widetilde{A}_i^I R s_{2i}\}.
\end{aligned} \tag{4.3}$$

Substituting (3.9) into this equation we obtain (4.2) and (4.3). So the theorem is proved.

Theorem 4.1 shows that in the sequential case, the information loss not only depends on the intrinsic curvature $\tilde{A}^I$ but also depends on s_2 which is determined by the stopping rule so that it is possible to minimize the loss of information in some sense by adapting a suitable stopping rule.

Theorem 4.2. *The trace of relative loss of information is minimized by the stopping rule satisfying* $s_{2i} = \tilde{a}_i^I = p^{-1}\mathrm{tr}(\widetilde{A}_i^I), \quad (i = 2, \cdots, m-p)$.

Proof. From (3.4) and (4.3) we have

$$\begin{aligned} K\nu\mathrm{tr}\{\Delta J(\hat{\theta})J^{-1}(\mathbf{X})\} &\simeq \mathrm{tr}\{R^T A_{IS} R(R^T R)^{-1}\} \simeq \mathrm{tr}(A_{IS}) \\ &= \sum_{i=2}^{m-p}\{\mathrm{tr}(\widetilde{A}_i^I)^2 + ps_{2i}^2 - 2ps_{2i}\tilde{a}_i^I\} \\ &= \sum_{i=2}^{m-p}\{\mathrm{tr}(\widetilde{A}_i^I)^2 + p(s_{2i} - \tilde{a}_i^I)^2 - p(\tilde{a}_i^I)^2\}. \end{aligned}$$

Obviously, the trace attains its minimum when $s_{2i} = \tilde{a}_i^I$. This completes the proof.

It can be shown that the loss of information of nonsequential MLE is given by $\sum\limits_{i=2}^{m-p}\mathrm{tr}(\widetilde{A}_i^I)^2$. Hence $p(\tilde{a}_i^I)^2$ corresponds to the information recovered by the sequential estimation. This shows that the sequential procedure is better than nonsequential procedure for our model from the viewpoint of reducing loss of information.

Corollary 4.1. *If* θ *is a one-dimensional parameter, then when* $s_{2i} = \mathrm{tr}(\tilde{A}_i^I)$, $(i = 2, \cdots, m-1)$, *the information loss vanishes approximately.*

The relationship between the observed information and the expected information is a commonly concerned problem in statistical inference, which has been studied by Efron[12], Efron and Hinkley[13], Amari[14], Wei[15] and so on. For our models, let

$$\Omega = (K\nu)^{1/2}\{-\ddot{l}(\hat{\theta})J^{-1}(\mathbf{X}) - I_p\}_{\theta=\hat{\theta}} \tag{4.4}$$

which can be rewritten as $(K\nu)^{1/2}\{-\ddot{l}(\hat{\theta}) - J(\mathbf{X})\}J^{-1}(\mathbf{X})$, and represents the relative difference between the observed information and the expected information of θ contained in $\mathbf{X}$. For the sake of convenience, let $\mathrm{vec}[A]$ be a $p^2 \times k$ matrix whose i-th column is $\mathrm{vec}(A_i)$ for a $k \times p \times p$ array A, where $\mathrm{vec}(A_i)$ is the vectorization of A_i, $i = 1, \cdots, k$. Then we have

Theorem 4.3. *Under the notation and assumptions stated above,* $\mathrm{vec}(\Omega)$ *and* $\mathrm{tr}(\Omega)$ *are asymptotically normal and satisfy*

$$\mathrm{vec}(\Omega) \xrightarrow{L} N(0, \Sigma), \tag{4.5}$$

$$\mathrm{tr}(\Omega) \xrightarrow{L} N(0, \sigma^2), \tag{4.6}$$

where $\Sigma = V\overline{I}_{m-p-1}V^T$, $\sigma^2 = v^T\overline{I}_{m-p-1}v$, $V = \{(L \otimes R^T(\mathrm{vec}[\widetilde{A^I}]) - \mathrm{vec}(I_p) \otimes s_2^T\}$, $v = \mathrm{tr}[\widetilde{A^I}] - ps_2$, *and* $\overline{I}_{m-p-1} = \mathrm{diag}(0, 1, \cdots, 1)$ *is an* $(m-p) \times (m-p)$ *matrix.*

Proof. Expanding (3.1) gives

$$n = K\nu + (K\nu)^{\frac{1}{2}}s_2^T\overline{\lambda} + O_p(1),$$

where ν, s_2 are evaluated at $(\hat{\theta}, 0)$. Hence we have

$$(K\nu)^{\frac{1}{2}}\{(\frac{n}{K\nu})I_p - I_p\} = s_2^T(\hat{\theta}, 0)\overline{\lambda}I_p + O_p(K^{\frac{1}{2}}). \tag{4.7}$$

Substituting (2.8) and (3.4) into (4.4) gives

$$\Omega = (K\nu)^{\frac{1}{2}}\{(\frac{n}{K\nu})I_p - (K\nu)^{-1}[rg^{-1}][W-\Gamma](D^Tg^{-1}D)^{-1} - I_p\}_{\theta=\hat{\theta}} + O_p(K^{-\frac{1}{2}}).$$

From $\hat{\theta}-\theta = O_p(K^{-\frac{1}{2}})$, we have $D(\hat{\theta}) = D(\theta) + O_p(K^{-\frac{1}{2}})$. Similar relations hold for W, g and Γ. Substituting these and (4.7) into Ω given above we have

$$\Omega = s_2^T\overline{\lambda}I_p - [\hat{e}^Tg^{-1}][W-\Gamma]LL^T + O_p(K^{-\frac{1}{2}}),$$

where $\hat{e} = (K\nu)^{-\frac{1}{2}}\hat{r}$ and it follows from (3.8) that $\hat{e} = N\overline{\lambda} + O_p(K^{-\frac{1}{2}})$. Hence it follows from (2.12), (2.17) and (3.9) that

$$\Omega = s_2^T\eta I_p - R^T[\eta^T][\widetilde{A}^I]L^T + O_p(K^{-\frac{1}{2}}), \tag{4.8}$$

$$\begin{aligned}\mathrm{Vec}(\Omega) &= s_2^T\eta\mathrm{Vec}(I_p) - (L\otimes R^T\mathrm{Vec}([\eta^T][\widetilde{A}^I]) + O_p(K^{-\frac{1}{2}})\\ &= \{\mathrm{Vec}(I_p)\otimes s_2^T - (L\otimes R^T)(\mathrm{Vec}[\widetilde{A}^I])\}\eta + O_p(K^{-\frac{1}{2}}).\end{aligned}$$

Thus we obtain (4.5) from the asymptotic normality of η.

From (4.8) we have

$$\begin{aligned}\mathrm{tr}(\Omega) &= ps_2^T\eta - \mathrm{tr}([\eta^T][\widetilde{A}^I]) + O_p(K^{-\frac{1}{2}})\\ &= (ps_2 - \mathrm{tr}[\widetilde{A}^I])^T\eta + O_p(K^{-\frac{1}{2}}),\end{aligned}$$

which results in (4.6).

Corollary 4.2. *If $s_{2i} = p^{-1}\mathrm{tr}(\widetilde{A}_i^I)$ $(i = 2,\cdots,m-p)$, then $\mathrm{tr}(\Omega)\simeq 0$(a.s.). In particular, for a one-dimensional parameter θ, if $s_{2i} = \widetilde{A}_i^I (i = 2,\cdots,m-p)$, then $\Omega\simeq 0$ (a.s.).*

This corollary shows that in the sequential case the observed information is approximately equal to the expected information in some sense while it is generally impossible for the fixed sample case.

Now we calculate the asymptotic variance of $\hat{\theta}$.

Lemma 4.1. *Under the assumptions stated above, we have*

$$E(\tau_a\tau_b) = \delta_{ab} + O_p(K^{-\frac{1}{2}}),$$

$$E(\tau_a\tau_b\tau_c) = (K\nu)^{-\frac{1}{2}}(\Gamma^P_{abc} + \Delta^s_{abc}) + O_p(K^{-\frac{1}{2}}), \tag{4.9}$$

$$E(\tau_a\tau_b\tau_c\tau_d) = (\delta_{ab}\delta_{cd} + \delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}) + O_p(K^{-\frac{1}{2}}) \tag{4.10}$$

for $a,b,c,d = 1,\cdots,p$, where $\delta_{ab} = 1(a=b), \delta_{ab} = 0(a\neq b)$, $\Delta^s_{abc} = s_{1a}\delta_bc + s_{1b}\delta_{ac} + s_{1c}\delta_{ab}$, $\Delta^s = (\Delta^s_{abc})$ is a $p\times p\times p$ array, and Γ^P is given in (2.18).

Proof. It is easily seen that for our model the fundamental lemma of [10] holds. It is similar to Lemma 2.1 of [8] that we have

$$\tau = \sum_{i=1}^n z_i, \qquad z_i = Q^Tg^{-1}\frac{(X^i - \pi(\theta))}{(K\nu)^{\frac{1}{2}}}, \tag{4.11}$$

$$\begin{aligned}E(\tau_a\tau_b\tau_c) &= E(n)E(z_{1a}z_{1b}z_{1c}) + E(nz_{1a})E(z_{1b}z_{1c})\\ &\quad + E(nz_{1b})(z_{1a}z_{1c}) + E(nz_{1c})E(z_{1a}z_{1b}),\end{aligned} \tag{4.12}$$

where z_{1a}, z_{1b}, and z_{1c} are the components of z_1. Denote $H = Q^Tg^{-1}$. Then the moment-generating function of z_1 is

$$M(t) = \sum_{i=1}^m \pi_i e^{A_i}, \quad A_i = \sum_{a=1}^p (K\nu)^{-\frac{1}{2}}\sum_{a=1}^p H_{ai}t_a,$$

and hence the first four derivatives of $M(t)$ are given by

$$\begin{aligned}
E(z_{1a}) &= M'(0) = \sum_{i=1}^{m} \pi_i \frac{H_{ai}}{K\nu^{\frac{1}{2}}} = 0,\\
E(z_{1a}z_{1b}) &= M''_{ab}(0) = (K\nu)^{-1} \sum_{i=1}^{m} \pi_i H_{ai}H_{bi} = (K\nu)^{-1}\delta_{ab},\\
E(z_{1a}z_{1b}z_{1c}) &= M^{(3)}_{abc}(0) = (K\nu)^{-\frac{3}{2}} \sum_{i=1}^{m} \pi_i H_{ai}H_{bi}H_{ci}\\
&= (K\nu)^{-\frac{3}{2}}\Gamma^p_{abc},\\
E(z_{1a}z_{1b}z_{1c}z_{1d}) &= M^{(4)}_{abcd}(0) = O(K^{-2}).
\end{aligned}$$

It follows from $E(n) = K\nu + O_p(1)$ that

$$\begin{aligned}
n(K\nu)^{-1} &= 1 + (\overline{s}_1^T\tau + s_2^T\eta)/(K\nu)^{1/2} + O_p(K^{-1}),\\
E(nz_{1a})E(z_{1b}z_{1c}) &= \delta_{bc}E\{z_{1a} + (\overline{s}_1^T\tau z_{1a} + s_2^T\eta z_{1a})/(K\nu)^{1/2}\} + O_p(K^{-1})\\
&= \delta_{bc}E\Big\{\sum_{t=1}^{p}\sum_{i=1}^{n} \overline{s}_{1t}z_{1t}z_{1a}/(K\nu)^{1/2}\Big\} = O(K^{-1})\\
&= \delta_{bc}\sum_{t=1}^{p} \overline{s}_{1t}^T E(n)E(z_{1t}z_{1a})/(K\nu)^{1/2} + O(K^{-1})\\
&= \delta_{bc}\sum_{t=1}^{p} \overline{s}_{1t}\delta_{at}/(K\nu)^{1/2} + O(K^{-1})\\
&= \overline{s}_{1t}\delta_{bc}/(K\nu)^{1/2} + O(K^{-1}).
\end{aligned}$$

Substituting the above equations into (4.12) gives (4.9) and (4.10) can be obtained by the derivation similar to the Lemma 2.1 of [8].

Theorem 4.4. *Under the assumptions stated above, the asymptotic variance of $\hat{\theta}$ can be approximately represented as*

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) \simeq\ & J^{-1}(\mathbf{X}) + J^{-1}(\mathbf{X})\Delta J(\hat{\theta})J^{-1}(\mathbf{X})\\
&+ (K\nu)^{-2}L(V_p + V_s + V_{ps} + \overline{V}_{ps})L^T,
\end{aligned} \tag{4.13}$$

where

$$\begin{aligned}
(V_p)_{ij} &= \frac{1}{2}\mathrm{tr}(A_i^pA_j^p - A_i^p\Gamma_j^p - A_j^p\Gamma_i^p),\\
V_s &= (s_1^TLL^Ts_1)I_p + L^Ts_1s_1^TL - 2[s_1^TL][\Gamma^p + \Delta^s],\\
V_{ps} &= A^pL^Ts_1 + (A^pL^Ts_1)^T,\\
(\overline{V}_{ps})_{ij} &= -\frac{1}{2}\mathrm{tr}(A_i^p\Delta_j^s + A_j^p\Delta_i^s),
\end{aligned}$$

$(V_p)_{ij}$ and $(\overline{V}_{ps})_{ij}$ denote the elements of V_p and $\overline{V}_{ps}$ at (i,j) position; A_i^P, Γ_i^P and Δ_i^s are the $i-$th face of arrays A^P, Γ^P, and Δ^s respectively.

Proof. To calculate $\mathrm{Var}(\hat{\theta})$, we use the following well-known formula

$$\mathrm{Var}(\hat{\theta}) = E[\mathrm{Var}(\hat{\theta}|\tau)] + \mathrm{Var}[E(\hat{\theta}|\tau)].$$

By Theorem 3.1, η and τ are asymptotically normal and asymptotically independent. Thus

from (3.6) we have

$$\mathrm{Var}(\hat{\theta}|\tau) \simeq (K\nu)^{-2} L\mathrm{Var}\{([\eta^T][\widetilde{A}^I]\tau - \eta^T s_2\tau)|\tau\}L^T,$$
$$E\{\mathrm{Var}(\hat{\theta}|\tau)\} \simeq (K\nu)^{-2} LA_{Is}L^T \simeq J^{-1}(\mathbf{X})\Delta J(\hat{\theta})J^{-1}(\mathbf{X}).$$

It follows from (3.6) that

$$\mathrm{Var}\{E(\hat{\theta}|\tau)\} = L\mathrm{Var}\{(K\nu)^{-\frac{1}{2}}\tau - \frac{1}{2}(K\nu)^{-1}\tau^T A^p\tau - (K\nu)^{-1}\tau^T \bar{s_1}\tau\}L^T. \tag{4.14}$$

The element of $\mathrm{Cov}\{\tau(K\nu)^{-\frac{1}{2}},\ (-2K\nu)^{-1}\tau^T A^P\tau\}$ at (i,j) is

$$\begin{aligned}
&-\frac{1}{2}(K\nu)^{-\frac{3}{2}} E\Big\{\sum_{k=1}^{p}\sum_{l=1}^{p} A^p_{jkl}\tau_i\tau_k\tau_l\Big\} \\
&= -\frac{1}{2}(K\nu)^{-2}\sum_{k=1}^{p}\sum_{l=1}^{p} A^p_{jkl}(\Gamma^p_{ikl} + \Delta^s_{ikl}) + O(K^{-\frac{5}{2}}) \\
&= -\frac{1}{2}(K\nu)^{-2}\mathrm{tr}(A^p_j\Gamma^p_i + A^p_j\Delta^s_i) + O(K^{-\frac{5}{2}}).
\end{aligned}$$

By similar derivations we can get other terms in (4.14) and the details are omitted here. Combining all above results, we can get (4.13) to complete the proof of the theorem.

Theorem 4.4 shows that the first term of $\mathrm{Var}(\hat{\theta})$ is C-R lower bound and the second term indicates the relationship between the variance of $\hat{\theta}$ and the information loss of $\hat{\theta}$. This term depends only on the intrinsic curvature. The other terms of $\mathrm{Var}(\hat{\theta})$ depend on the parameter-effects curvature and the stopping rule.

References

[1] Kass, R. E., The geometry of asymptotic inference, *Statist. Sci.,* **4** (1989), 188-219.
[2] Efron, B., Defining the curvature of a statistical problem (with application to second order efficiency), *Ann. Statist.,* **3** (1975), 1189-1142.
[3] Amari, S., Differential geometry of curved exponential family-curvatures and information loss, *Ann. Statist.,* **10** (1982), 375-385.
[4] Bates, D. M. & Watts, D. G., Relative curvature measures of nonlinearity, *J. R. Statist. Soc. Ser. B*, **42** (1980), 1-25.
[5] Amari, S., Differential geometrical method in statistics, Lecture notes in Statistics **28**, 1985, Springer, Berlin.
[6] Rao, C. R., Linear statistical inference and its applications, John Wiley, New York, 1973..
[7] Takeuchi, K. & Akahira, M., Second order asymptotic efficiency in terms of asymptotic variance of the sequential maximum likelihood estimation procedures, Statistical Theory and Data Analysis **1**: Proceeding of Second Pacific Area Statistical Conference, North-Holland, 1988, 191-196.
[8] Akahira, M. & Takeuchi, K., Third order asymptotic efficiency of the sequential maximum likelihood estimation procedure, *Sequential Anal.,* **8** (1989), 333-359.
[9] Okamoto, I., Amari, S. & Takeuchi, K., Asymptotic theory of sequential estimation: Differential geometric approach, *Ann. Statist.,* **19** (1991), 961-981.
[10] Wald, A., Sequential analysis, Dover Publications, Inc., New York, 1973.
[11] Cheng, X. R., Introduction to mathematical statistics, Sci. Press of China, Beijing, 1981.
[12] Efron, B., The geometry of exponential families, *Ann. Statist.,* **6** (1978), 362-376.
[13] Hinkley, D. V., Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, *Biometrika*, **65** (1978), 457-487.
[14] Amari, S., Geometrical theory of asymptotic ancillarity and conditional inference, *Biometrika,* **69** (1982), 1-17.
[15] Wei, B. C., Some second order asymptotics in nonlinear regression, *Aust. J. Statist.,* **33** (1991), 75-84.