Ann. Inst. Statist. Math.
36 (1984), Part A, 199-206

A CHARACTERIZATION OF SECOND ORDER EFFICIENCY IN A CURVED EXPONENTIAL FAMILY

SHINTO EGUCHI

(Received Apr. 20, 1983; revised Nov. 22, 1983)

Summary

The optimality of estimation method is investigated in a curved exponential family. A risk function, which is an extension of a residual sum of squares in regression analysis, is introduced. It is shown that second order efficiency of an estimation method is equivalent to attain the minimum among limiting risks of all estimation methods.

1. The main results

Let $\mathcal{F}$ be an n-dimensional exponential family of densities on the sample space R^n with respect to a carrier measure ω. The family $\mathcal{F}$ is expressed as

$$f(x|\theta)\equiv\exp(\langle x, \theta\rangle-\psi(\theta))$$

by the natural parameter $\theta\equiv(\theta^1,\cdots,\theta^n)$ with the usual inner product $\langle\ ,\ \rangle$ of R^n. The expectation parameter $\eta\equiv(\eta_1,\cdots,\eta_n)$ of $\mathcal{F}$ is defined by the transformation of θ into η:

$$\eta[\theta]\equiv\mathrm{E}\,X\,.$$

Then the MLE of η or θ based on a sample $(x_1,\cdots,x_n)$ is given by

$$\bar{x}\equiv\frac{1}{N}(x_1+\cdots+x_N)$$

or $\bar{\theta}\equiv\eta^{-1}[\bar{x}]$, respectively. An m-dimensional curved exponential family is denoted by $\tilde{\mathcal{F}}$. That is,

$$\tilde{\mathcal{F}}\equiv\{f(x|\theta(u)):u\in U\}\ ,$$

where U is an open set in R^m and the image of the mapping $\theta(\cdot)$ of u

Key words: Exponential family, Kullback-Leibler divergence, minimum contrast estimator, maximum likelihood method, second order efficiency.

into θ with the Jacobian matrix of rank m is non-flat. Notice that the m-component parameter u is used only to name $\tilde{\mathcal{F}}$. The family $\tilde{\mathcal{F}}$ is invariant under transformations of parameter spaces.

Assume that $(x_1, \cdots, x_N)$ is a sample from a density of $\tilde{\mathcal{F}}$. We concern with a class of all estimation methods which correctly map under parameter transformations of $\tilde{\mathcal{F}}$. That is, for any one-to-one parameter transformation g of u into τ, the estimator $\hat{\tau}$ of τ by the method of estimation satisfies

$$\hat{\tau}=g(\hat{u}) \tag{1.1}$$

with the estimator $\hat{u}$ of u by the same method. The maximum likelihood method enjoys (1.1) (see Efron [2]). The property is equivalent that the summary:

$$\hat{f}\equiv f(\cdot\,|\,\theta(\hat{u}))$$

based on the method is invariant under transformations of parameters. All minimum contrast methods have the property (1.1) (cf. Drossos and Philippou [3]). Notice that many minimum contrast methods are also invariant under one-to-one transformations of the sample space. We compare these methods of estimation, identifying estimators which map correctly each other as an estimation method. There is a confusing aspect. The superiority relations between two estimators does not generally conserve by comparing their mean square errors for various parameter spaces with one-to-one correspondance. Should we consider which method is superior to others? This phenomenon is caused owing that the risk function, i.e. mean square error in this case, is not invariant under parameter transformations.

In a linear regression analysis, the sum of squares plays an important role for inference and enables us to associate a simple geometric interpretation. So we extend a residual sum of squares to the curved model $\tilde{\mathcal{F}}$. That is,

$$S_N(\hat{u}, u)=\frac{1}{2}(\bar{x}-\eta(\hat{u}))'G(\eta(u))(\bar{x}-\eta(\hat{u}))$$

for an estimator $\hat{u}$ of u, where $\eta\equiv\eta[\theta(u)]$ and $G(\eta)$ is the Fisher information matrix of η. This quantity $S_N(\hat{u}, u)$ gives an account of the difference between the canonical sufficient statistic $\bar{x}$ and the fitted value $\eta(\hat{u})$. So we adopt

$$R_N(\hat{u}, u)\equiv \mathrm{E}\, S_N(\hat{u}, u)$$

as a risk function. The risk function R_N is invariant under transformations of parameter spaces.

We investigate the relations between the risk R_N and asymptotic efficiencies. The information loss in reducing from the sample to a statistic T is defined by the difference between Fisher information matrices of the sample and T, say $\Delta_N(T, u)$. A Fisher-consistent estimator $\hat{u}$ of u is said to be first order efficient if

$$\lim_{N\to\infty} \frac{1}{N} \Delta_N(\hat{u}, u) = O \,, \tag{1.2}$$

where O is a zero matrix. Note that the condition (1.2) is equivalent to the definition of BAN for $\hat{u}$. Furthermore the estimator $\hat{u}$ is said to be second order efficient if the limiting information loss due to $\hat{u}$ is a minimum among those of all other first order efficient estimators in the sense of nonnegative definitness of matrices (cf. Ghosh and Subramanyam [6]).

Henceforth we identify an estimator with a method of estimation in the above text. The following theorems will be proved in the next section.

THEOREM 1. *First order efficiency of a Fisher-consistent estimator $\hat{u}$ is equivalent to each of the following conditions* (i), (ii) *and* (iii):

(i) $\lim_{N\to\infty} N[R_N(\tilde{u}, u) - R_N(\hat{u}, u)] \geqq 0$

for any Fisher-consistent estimator $\tilde{u}$.

(ii) $\lim_{N\to\infty} (N/2)\, \mathrm{E}\{(\eta(\hat{u}) - \eta(u))' G(\theta(u))(\eta(\hat{u}) - \eta(u))\} = m/2.$

(iii) $\lim_{N\to\infty} N[R_N(\hat{u}, u)] = (n-m)/2.$

Furthermore

THEOREM 2. *A first order efficient estimator $\hat{u}$ is second order efficient if and only if*

$$\lim_{N\to\infty} N^2[R_N(\tilde{u}, u) - R_N(\hat{u}, u)] \geqq 0$$

for all first order efficient estimators $\tilde{u}$.

Theorems 1 and 2 tell us that the risk R_N exactly discriminates the first and second order efficiencies of estimators, respectively. Theorem 1 is only a geometrical version of famous χ^2-decomposition theorems (cf. Figure 1).

We choose the expected Kullback-Leibler divergence

$$R_N^*(\hat{u}, u) \equiv \mathrm{E}\, \rho_{\mathrm{KL}}(f_{\bar{\theta}}, f_{\theta(\hat{u})}) \,,$$

as a risk function, where

$$\rho_{\mathrm{KL}}(f_1, f_2) \equiv \int f_1(x)[\log f_1(x) - \log f_2(x)]\, \omega\{dx\} \,.$$

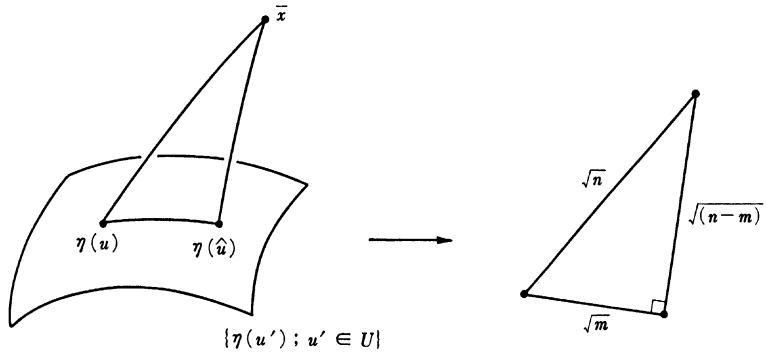

Fig. 1. The right triangle.

Let $\hat{u}$ be first order efficient estimator of u. The triangle with sides of $\sqrt{N}|\bar{x}-\eta(\hat{u})|$, $\sqrt{N}|\eta(\hat{u})-\eta(u)|$ and $\sqrt{N}|\eta(u)-\bar{x}|$ converges to the right triangle with $\sqrt{(n-m)}$, $\sqrt{m}$ and $\sqrt{N}$, where $|\eta|^2$ denotes $\eta' G(\theta(u))\eta$.

The risk R_N^* is invariant under transformations of both parameter spaces and sample spaces. However one may not regard the risk R_N^* as a reasonable measure of optimality in the class of estimators since the maximum likelihood estimator is nothing but the minimizer of the Kullback-Leibler divergence ρ_{KL}.

This question is answered by the assertion:

Theorems 1 and 2 are valid for the risk R_N^* in place of R_N.

The function

$$R_N^*(\tilde{u}, u)-R_N^*(\hat{u}, u)$$

is closely related to the discrimination rate of ρ_{KL}, introduced by Kuboki [7], in the case including a sufficient statistic.

The author introduced a class of second order efficient estimators which are defined by minimizers of the corresponding contrast functions ρ's (cf. Eguchi [4]). In practice the above assertion holds for all these contrast functions ρ in addition to ρ_{KL}. Let $\mathcal{L}(u)$ be the likelihood function of the parameter u. It follows that

$$\log \mathcal{L}(u_1)-\log \mathcal{L}(u_2)=N\{\rho_{\mathrm{KL}}(\bar{\theta}, u_2)-\rho_{\mathrm{KL}}(\bar{\theta}, u_1)\} \tag{1.3}$$

for all u_1 and u_2 in U, where

$$\rho_{\mathrm{KL}}(\theta, u)\equiv\rho_{\mathrm{KL}}(f_\theta, f_{\theta(u)}) .$$

From (1.3) it follows

COROLLARY 1. *A Fisher-consistent estimator $\hat{u}$ is first order efficient if and only if*

$$\lim_{N\to\infty} \mathrm{E}\left[\frac{\mathcal{L}(\hat{u})}{\mathcal{L}(\tilde{u})}\right] \leqq 1$$

for all Fisher-consistent estimators $\tilde{u}$. *Furthermore the estimator* $\hat{u}$ *is second order efficient if and only if*

$$\lim_{N\to\infty} \mathrm{E}\left[\frac{\mathcal{L}(\hat{u})}{\mathcal{L}(\tilde{u})}\right]^N \leqq 1$$

for all first order efficient estimators $\tilde{u}$.

Corollary 1 may be extended to a smooth parametric family with some regularity conditions.

2. Proof of Theorem 2

Theorem 2 will be shown in terms of differential geometric approach, which is originated by S. Amari. Refer Amari [1] with respect to the information metric g, the exponential connection $\overset{e}{\Gamma}$ and the mixture $\overset{m}{\Gamma}$. We state only an outline of the proof owing to avoiding tedious calculuses (cf. Eguchi [5] for the detail proof).

Take an $n\times(n-m)$ matrix $B^{\perp}(u)$ of full rank to satisfy

$$B(u)'\bar{G}(\theta(u))B^{\perp}(u)=O\ ,$$

where $B(u)$ is the Jacobian matrix $\theta(u)$ and $\bar{G}(\theta)$ is the Fisher information matrix of θ. The tangent space T_f of $\mathcal{F}$ at f in $\tilde{\mathcal{F}}$ is decomposed into the tangent and the normal spaces, say $\tilde{T}_f$ and $T_f^{\perp}$, of which bases are expressed as

$$\{e_i(u)\equiv \bar{x}_i-\eta_i(u)\}_{i=1,\cdots,n}\ ,$$

$$\{e_a(u)\equiv B_a^i(u)e_i(u)\}_{a=1,\cdots,m}$$

and

$$\{e_\lambda(u)\equiv B_\lambda^i(u)e_i(u)\}_{\lambda=m+1,\cdots,n}\ ,$$

respectively, where $B_a^i(u)$ and $B_\lambda^i(u)$ are elements of matrices $B(u)$ and $B^{\perp}(u)$, respectively. The induced components of g to $\tilde{T}_f$ and $T_f^{\perp}$ are expressed as

$$\tilde{g}_{ab}(u)=\mathrm{E}\{e_a(u)e_b(u)\} \qquad \text{and} \qquad \tilde{g}_{\lambda\mu}(u)=\mathrm{E}\{e_\lambda(u)e_\mu(u)\}\ ,$$

respectively.

The second fundamental tensor of $\tilde{\mathcal{F}}$ with respect to $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$ are denoted $\overset{m}{H}$ and $\overset{e}{H}$, respectively. The components of $\overset{m}{H}$ and $\overset{e}{H}$ are ex-

pressed as

$$\overset{m}{H}_{ab\lambda}(u)=B^i_\lambda(u)\partial_a[B^j_b(u)g_{ij}(\theta(u))]$$

and

$$\overset{e}{H}_{ab\lambda}(u)=B^i_\lambda(u)g_{ij}(\theta(u))\partial_a B^j_b(u)$$

with respect to the parameter u with $\partial_a\equiv\partial/\partial u^a$. Henceforth we omit the arguments of the above geometric quantities at the true value and freely raise and lower their indices, e.g.

$$e^\lambda\equiv\tilde{g}^{\lambda\mu}(u)e_\mu(u)\qquad \overset{e}{H}{}^a_{b\lambda}\equiv\overset{e}{H}_{cb\lambda}(u)\tilde{g}^{ac}(u)\ ,$$

where $\tilde{g}^{ab}(u)$ and $\tilde{g}^{\lambda\mu}(u)$ are the inverses of $\tilde{g}_{ba}(u)$ and $\tilde{g}_{\mu\lambda}(u)$, respectively. For a first order efficient estimator $\hat{u}$, the set

$$\{f(\cdot\,|\,\theta);\ \hat{u}(\theta)=u\}$$

is called the ancillary subspace by Amari [1], of which the second fundamental tensor at $f=f(\cdot\,|\,\theta(u))$ with respect to $\overset{m}{\Gamma}$ is denoted by $\hat{H}$. Then we can rewrite Theorem 7 in Amari [1] in the convenient form:

THEOREM A. *Let $\hat{u}$ be a first order efficient estimator with the second fundamental tensor $\hat{H}$ of the ancillary subspace. Then*

$$\hat{u}^a-u^a=e^a-\frac{1}{2}\overset{m}{\Gamma}{}^a_{bc}e^be^c+\overset{e}{H}{}^a_{b\lambda}e^be^\lambda-\frac{1}{2}\hat{H}^a_{\lambda\mu}e^\lambda e^\mu+O(\|e\|^3)\ . \tag{2.1}$$

Furthermore the estimator $\hat{u}$ is second order efficient if and only if the tensor $\hat{H}$ vanishes over $\tilde{\mathcal{F}}$.

Theorem A is the same as Theorem 1 (ii) in Ghosh and Subramanyam [6] in the case of one parameter.

To prove Theorem 2, we prepare

LEMMA. *Let $\hat{u}$ be a first order efficient estimator with the tensor $\hat{H}$. It holds that*

$$\begin{aligned}\lim_{N\to\infty}N\Big[NR_N(\hat{u},u)-\frac{n-m}{2}\Big]&\\ =-\frac{1}{8}\|\overset{m}{\Gamma}\|^2+\frac{1}{2}\|\overset{e}{H}\|^2+\frac{1}{8}\|\hat{H}\|^2-(\overset{e}{H},\overset{m}{H})-\frac{1}{2}(\overset{m}{H},T)\ ,&\end{aligned} \tag{2.2}$$

where

$$\|\overset{m}{\Gamma}\|^2\equiv\overset{m}{\Gamma}{}^a_{bc}\overset{m}{\Gamma}{}^d_{ef}\tilde{g}_{ad}\tilde{g}^{be}\tilde{g}^{cf}\ ,$$

$$\|\overset{e}{H}\|^2\equiv\overset{e}{H}{}^a_{b\lambda}\overset{e}{H}{}^c_{d\mu}\tilde{g}_{ac}\tilde{g}^{bd}\tilde{g}^{\lambda\mu}\ ,$$

$$\|\hat{H}\|^2 \equiv \hat{H}^a_{\lambda\mu}\hat{H}^b_{\nu\xi}\tilde{g}_{ab}\tilde{g}^{\lambda\xi}\tilde{g}^{\mu\nu} ,$$

$$(\overset{e}{H}, \overset{m}{H}) \equiv \overset{e}{H}{}^a_{b\lambda}\overset{m}{H}{}^c_{d\mu}\tilde{g}_{ac}\tilde{g}^{bd}\tilde{g}_{\lambda\mu} ,$$

$$(\overset{m}{H}, T) \equiv \overset{m}{H}{}^a_{b\lambda}T^c_{d\mu}\tilde{g}_{ac}\tilde{g}^{bd}\tilde{g}^{\lambda\mu} ,$$

with the tensor $T \equiv \overset{m}{\Gamma} - \overset{e}{\Gamma}$.

PROOF. The statistic $e_i(u)$ is expanded as

$$e_i(\hat{u}) = B_{\lambda i}e^\lambda + B_{ai}\Delta^a - \frac{1}{2}\partial_b B_{ai}e^a e^b - \partial_b B_{ai}e^a \Delta^b - \frac{1}{6}\partial_c\partial_b B_{ai}e^a e^b e^c + O(\|e\|^4) \tag{2.3}$$

by Taylor's theorem, where $\Delta^a \equiv e^a - (\hat{u}^a - u^a)$. It follows from Lemma that the substitution of (2.1) into (2.3) leads (2.2) by taking the expectation since the identity

$$g_{ij} = B_{ai}\tilde{g}^{ab}B_{bj} + B_{\lambda i}\tilde{g}^{\lambda\mu}B_{\mu j}$$

holds. This completes the proof.

The proof of Theorem 2 is easily seen from Lemma. Let $\hat{u}$ be second order efficient estimator. It holds for any first order efficient estimator $\hat{u}$ with the second fundamental tensor $\hat{H}$ that

$$\lim_{N\to\infty} N^2[R_N(\hat{u}, u) - R_N(\tilde{u}, u)] = \frac{1}{8}\|\tilde{H}\|^2 \geqq 0$$

since the term

$$M(u) \equiv -\frac{1}{8}\|\overset{m}{\Gamma}\|^2 + \frac{1}{2}\|\overset{e}{H}\|^2 - (\overset{e}{H}, \overset{m}{H}) - \frac{1}{2}(\overset{m}{H}, T)$$

is common among all first order efficient estimators. The inverse assertion is clear since $\|\hat{H}\| = 0$ implies $\hat{H} = 0$.

Remark. Consider a submodel of a multinormal family with known covariance matrix Σ. Since the connection $\overset{m}{\Gamma}$ coinsides with $\overset{e}{\Gamma}$,

$$\mathrm{E}\, e(\hat{u})'\Sigma^{-1}e(\hat{u}) = \frac{n-m}{N} - \frac{1}{N^2}\left(\frac{\|\Gamma\|^2}{4} + \|H\|^2\right) + O(N^{-3}) ,$$

with $\Gamma \equiv \overset{m}{\Gamma} = \overset{e}{\Gamma}$ and $H \equiv \overset{m}{H} = \overset{e}{H}$. In this model the expectation of sum of squares is always smaller than $(n-m)/N$ at any true value for sufficiently large N.

Acknowledgements

I wish to thank the referee for his valuable comments. I am also grateful to Professor M. Okamoto and Dr. Y. Toyooka of Osaka University for their encouragements.

OSAKA UNIVERSITY*

REFERENCES

[1] Amari, S. (1982). Differential geometry of curved exponential families-curvature and information loss, *Ann. Statist.*, **10**, 357-385.

[2] Drossos, C. A. and Phillipou, A. N. (1980). A note on minimum distance estimates, *Ann. Inst. Statist. Math.*, **32**, A, 121-123.

[3] Efron, B. (1982). Maximum likelihood and decision theory, *Ann. Statist.*, **10**, 340-356.

[4] Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family, *Ann. Statist.*, **11**, 793-803.

[5] Eguchi, S. (1984). A property of second order efficiency, *Technical Report*, No. 132, Statistical Research Group, Hiroshima University.

[6] Ghosh, J. K. and Subramanyam, K. (1974). Second order efficiency of maximum likelihood estimators, *Sankhyā*, A, **36**, 325-358.

[7] Kuboki, H. (1982). Unbiased estimations in the sense of Lehmann and their discrimination rates, *Ann. Inst. Statist. Math.*, A, **34**, 19-37.

* Now at Hiroshima University.