Asymptotical improvement of maximum likelihood estimators on Kullback-Leibler loss

Shinto Eguchi*, Takemi Yanagimoto
Institute of Statistical Mathematics and Graduate University for Advanced Studies,
Minami-Azabu, Minato-ku, Tokyo 106-8569 Japan

running title: Improvement of maximum likelihood estimator

Abstract

We discuss the general form of a first-order correction to the maximum likelihood estimator which is expressed in terms of the gradient of a function, which could for example be the logarithm of a prior density function. In terms of Kullback-Leibler divergence, the correction gives an asymptotic improvement over maximum likelihood under rather general conditions. The theory is illustrated for Bayes estimators with conjugate priors. The optimal choice of hyper-parameter to improve the maximum likelihood estimator is discussed. The results based on Kullback-Leibler risk are extended to a wide class of risk functions.

AMS Classification: primary 62F10; secondary 62F12

Key words : Bayes estimator; Conjugate prior; Differential geometric approach; Fisher consistency; Harmonic function; James-Stein estimator; Laplace-Beltrami operator; Maximum likelihood estimator; Kullback-Leibler risk; Second-order efficiency.

---

*Corresponding author. Fax: +81-3-54218728.

E-mail address: eguchi@ism.ac.jp (S. Eguchi).

1. Introduction

Let $\hat{\theta}$ be the maximum likelihood estimator for a vector parameter θ in a statistical model with probability density function $f(x;\theta)$. We consider alternative estimators of form

$$\hat{\theta}^* = \hat{\theta} + \frac{1}{n}h(\hat{\theta}) + o_p(n^{-1}), \tag{1.1}$$

where n is the sample size and h is an arbitrary smooth function; cf. Ghosh and Sinha (1981). A special choice of h is defined as the gradient of a function of θ. Such estimators arise, for example, in Bayesian estimation where h is the gradient of the log prior density. The Bayesian predictive distribution is discussed in a differential geometric framework similar to the present framework, cf. Komaki (1996). The differential geometric method for statistical inference is covered in the information geometry, see Amari and Nagaoka (2000) for the general discussion. Another approach for improving the maximum likelihood can be mentioned the local likelihood method, cf. Eguchi and Copas (1998) for the comparison of the maximum likelihood estimative distribution.

We note that any estimator of the form (1.1) has the same information as the maximum likelihood estimator in the limit because the two estimators $\hat{\theta}$ and $\hat{\theta}^*$ are in a one-to-one correspondence for a sufficiently large n. Hence the criterion of information loss makes no difference in this class of estimators. We take another criterion for our comparison among estimators of the form (1.1). Let $\tilde{\theta}$ be an estimator of θ. As a loss function for $\tilde{\theta}$, we take the Kullback-Leibler divergence

$$D(\tilde{\theta},\, \theta) = \int f(z;\tilde{\theta})\{\log f(z;\tilde{\theta}) - \log f(z;\theta)\}\mu(dz),$$

and define our risk function as $R(\tilde{\theta},\, \theta) = E_\theta\{D(\tilde{\theta},\, \theta)\}$, which we call the Kullback-Leibler risk function, see Clark and Barron (1990) for related discussion. By definition, $R(\tilde{\theta},\, \theta)$ is parametrization-invariant. In general, mean square error depends on the chosen parametrization. Thus the evaluation of an estimator based on the mean square error is not invariant to parametrization. In order to make our discussion invariant to parametrization, we now adopt Kullback-Leibler risk.

We will evaluate explicitly, in terms of a differential geometry viewpoint, a function $\delta_h(\theta)$ such that

$$\delta_h(\theta) = n^2[R(\hat{\theta},\, \theta) - R(\hat{\theta}^*,\, \theta) - E_\theta\{D(\hat{\theta},\, \hat{\theta}^*)\}] + o(1) \tag{1.2}$$

This leads immediately to a condition on h such that $\delta_h(\theta) \equiv 0$, and thus an estimator with smaller risk under the K-L loss function than the MLE. The James-Stein estimator falls within this theory. We can think of this in terms of a right-angled triangle formed by connecting the three points $\hat{\theta}$, $\hat{\theta}^*$ and θ when $\delta_h(\theta) = 0$, and so we call (1.2) the mean Pythagorean identity. See Yanagimoto (1994) for detailed discussion, also Hartigan (1998), Yanagimoto and Ohnishi (2005).

The paper is organized as follows. Section 2 gives the general asymptotic formula for the Kullback-Leibler risk for the class of estimators of the form (1.1). In § 3 we obtain the superiority of the Bayes estimator with conjugate prior in terms of a special choice for hyper-parameter, which includes a result on the multinomial parameter with conjugate Dirichlet priors. In § 4 the Kullback-Leibler risk is generalized to a class of risk functions with parametrization-invariance. A similar structure to that of Section 2 is shown by generalizing the definition of the relevant divergence operator.

2. Second-order structure of Kullback-Leibler risk

Let $f(x;\theta)$ be a probability density function with an unknown d-dimensional parameter θ in a parameter space Θ, and write the family $\mathcal{P} = \{f(x;\theta) : \theta \in \Theta\}$. Our approach to comparing performance of estimation is based not on a specific parameter but on invariant discussion with parametrization. For this we take a differential-geometric method; see Amari (1985) and Critchley, Marriot and Salmon (1994). We note that the first-order structure of the Kullback-Leibler risk in the class of estimators of the form (1.1) is simple: for any estimator $\hat{\theta}^*$ in this class

$$R(\hat{\theta}^*, \theta) = \frac{d}{2n} + O(\frac{1}{n^2}). \tag{2.1}$$

The common first-order term in $R(\hat{\theta}^*, \theta)$ is the lower bound in the class of consistent estimators. This result motivates the investigation of the second-order structure of $R(\hat{\theta}^*, \theta)$.

Let ϕ be a one-to-one transformation of θ into another parameter τ. We assumed for the estimator $\hat{\theta}^*$ such that the estimated density function $f(x; \hat{\theta}^*)$ is invariant under parameter transformations. Thus the estimator $\hat{\theta}^*$ is transformed into $\hat{\tau}^* = \phi(\hat{\theta}^*)$ as the corresponding estimator of τ, and the maximum likelihood estimator satisfies the same transformation rule, given as $\hat{\tau} = \phi(\hat{\theta})$.

We note that the estimator $\hat{\tau}^*$ has the following representation: $\hat{\tau}^* = \hat{\tau} + \frac{1}{n}\overline{h}(\hat{\tau}) +$

$o_p(n^{-1})$, where

$$\overline{h}(\hat{\tau}) = B(\hat{\theta})h(\hat{\theta}) + o_p(1)$$

with $B(\theta)$ being the Jacobi matrix of ϕ when evaluated at $\theta = \phi^{-1}(\tau)$. Thus the invariance argument naturally leads to the fact that for any reparametrization $\theta \to \tau$ there is a corresponding mapping of the function h into the $\overline{h}$ given, in other words, it works as a vector field on $\mathcal{P}$. We define the gradient of a function as a typical example of a vector field, which is a natural extension of this on a Euclidean space. The gradient of a function f on $\mathcal{P}$ is expressed as

$$(\text{grad } f(\theta))^i = \sum_{j=1}^{d} g^{ij}(\theta)\frac{\partial}{\partial\theta^j}f(\theta),$$

where $g^{ij}(\theta)$ is the inverse of the Fisher information $g_{ji}(\theta)$. Here we regard $\mathcal{P}$ as a Riemannian space with respect to the information metric g with components $g_{ij}(\theta)$; cf. Rao (1945). Similarly, the divergence of a vector field h is defined as

$$\text{div } (h) = \sum_i \frac{\partial h^i(\theta)}{\partial\theta_i} + \sum_{i,j,k} h^i(\theta)\, g^{jk}(\theta)\, \overline{\Gamma}_{ij,k}(\theta), \tag{2.2}$$

where $\{\overline{\Gamma}_{ij,k}(\theta)\}$ is the coefficients of the Levi-Civita connection with the metric g. It is noted that div (h) is a scalar, or parametrization-invariant function on $\mathcal{P}$ while h or grad f is subject to the covariant manner for parameter transformation. Thus, for example, the squared norm of h measured by the information metric g, i.e. $\| h \|^2 = \sum_{i,j} g_{ij}h^i h^j$, becomes invariant. The Laplace-Beltrami operator is then given by $\Delta = \text{div (grad)}$. A function f defined on $\mathcal{P}$ is said to be harmonic if $\Delta f = 0$ on $\mathcal{P}$ and superharmonic if $\Delta f \leq 0$.

We state the following main result on the second-order structure of the Kullback-Leibler risk:

THEOREM 1. *Let $\hat{\theta}$ be the maximum likelihood estimator of θ, and h a vector fields on $\mathcal{P}$. Define an estimator $\hat{\theta}$ by $\hat{\theta}^* = \hat{\theta} + \frac{1}{n}h(\hat{\theta})$. Then*

$$\delta_h(\theta) = -\{\text{div}(h) + \| h \|^2\},$$

where $\delta_h(\theta)$ is defined in (1.2).

The proof of Theorem 1 is given in Appendix 1. From Theorem 1 and $E_\theta D(\hat{\theta}, \hat{\theta}^*) = \frac{1}{2n^2} \parallel h \parallel^2 + o(n^{-2})$ it follows that the risk difference is expressed as

$$\gamma_h(\theta) = n^2\{R(\hat{\theta}, \theta) - R(\hat{\theta}^*, \theta)\} = -\{\mathrm{div}(h) + \frac{1}{2} \parallel h \parallel^2\} + o(1).$$

Thus we can say that $\hat{\theta}^*$ is better than $\hat{\theta}$ if and only if $\mathrm{div}(h) \leq -\frac{1}{2} \parallel h \parallel^2$. Further, if $\delta_h(\theta) = 0$, or $\gamma_h(\theta) = \frac{1}{2} \parallel h \parallel^2$, then $\gamma_h(\theta) - \gamma_{\alpha h}(\theta) = \frac{1}{2}(\alpha - 1)^2 \parallel h \parallel^2$. This concludes that the estimator with such an h has the greatest improvement of Kullback-Leibler risk in the class of estimators $\{\hat{\theta} + \frac{1}{n}\alpha h(\hat{\theta}) : \alpha \in R\}$.

COROLLARY. *Let u be a function defined on $\mathcal{P}$. Define an estimator $\hat{\theta}^*$ by*

$$\hat{\theta}^* = \hat{\theta} + \frac{1}{n}\mathrm{grad}\ u(\hat{\theta}). \tag{2.3}$$

Then the estimators $\hat{\theta}^*$ and $\hat{\theta}$ satisfy

$$\delta_{\mathrm{grad}\ u}(\theta) = -\exp(-u)\Delta\exp(u)$$

for any θ of Θ.

The proof follows from a direct application of Theorem 1 to the case of $h = \mathrm{grad}\ u$. Since $\Delta\exp(u) = \exp(u)\{\Delta u + \parallel \mathrm{grad}\ u \parallel^2\}$, we get $\delta_{\mathrm{grad}\ u} = -\exp(-u)\Delta\exp(u)$, which completes the proof of the Corollary.

From Corollary it follows that if $\exp(u)$ is superharmonic, then the estimator $\hat{\theta}^*$ is better than $\hat{\theta}$. Specifically, if $\exp(u)$ is harmonic, then the mean Pythagorean identity holds for the triangle connecting $\hat{\theta}$, $\hat{\theta}^*$ and θ since the right-angled triangle so formed $R(\hat{\theta}, \theta)$ is viewed as the squared length of the hypotenuse. Since a similar argument to the proof of Corollary yields

$$\gamma_h(\theta) = -\frac{2}{n^2}\exp(-\frac{1}{2}u)\Delta\exp(\frac{1}{2}u),$$

a necessary and sufficient condition for this superiority is that $\Delta\exp(\frac{1}{2}u)$ is superharmonic. Finally, we note that a vector field h is not always expressed as the gradient of a function; this needs to be checked by the condition: $\partial h_i/\partial\theta^j = \partial h_j/\partial\theta^i$ for $(\theta^i) = \theta$ and $h_i = \Sigma h^j g_{ij}$ in the Poicare lemma. See e.g. 2.3 in Abraham and Marsden (1978) for general discussion.

3. Bayes estimator for exponential models with conjugate priors

We now discuss Bayes estimators with a prior density function $p(\theta)$. The posterior density function is proportional to $L(\theta)p(\theta)$ with the likelihood function $L(\theta)$; cf. Bernardo and Smith (1994). By expanding the equation

$$\frac{\partial}{\partial\theta}\{\log L(\hat{\theta}^* + \log p(\hat{\theta}^*)\} = 0,$$

we obtain the result that the maximum posterior estimator of θ is of the form (1.1) with $h(\theta) = \text{grad } \log\ p(\theta)$. It follows from Corollary that the maximum posterior estimator with a harmonic or superharmonic prior density asymptotically improves the maximum likelihood estimator in terms of Kullback-Leibler risk.

We apply the general formula of the Kullback-Leibler risk given in Section 2 to an example of the Bayes estimator with conjugate prior. Consider the following exponential family model with the density, $f(x;\ \beta) = \exp\{x^T\beta - \psi(\beta)\}$ with the natural parameter β of dimension d. Then the prior density function conjugate to β is given by

$$\pi(\beta; m, \chi) = \exp[m\{\chi^T\beta - \psi(\beta)\}],$$

where m and χ denote hyper-parameters; see § 5.2 in Barnard and Smith (1994). For the expectation parameter η with the transformation $\eta = (\partial/\partial\beta)\psi(\beta)$ the maximum likelihood estimator $\hat{\eta}$ is just the canonical statistic $\overline{x} = (\Sigma x_i)/n$. The Bayes estimator $\hat{\eta}^*$ can be expressed as the convex combination of χ and $\hat{\eta}$ i.e.,

$$\hat{\eta}^* = \frac{m\chi + n\hat{\eta}}{n+m}, \tag{3.1}$$

which implies that $\hat{\eta}^*$ is of the form (1.1) with $h(\eta) = m(\chi - \eta)$. We note that h is expressed as the gradient of the logarithm of the prior density with respect to the coordinate system η. Thus it exactly holds that

$$\hat{\eta}^* = \hat{\eta} + \frac{1}{n+m}\,\text{grad } u(\hat{\eta}),$$

where $u(\eta) = \log\pi(\beta; m, \chi)$ through the inverse transformation of $\eta = (\partial/\partial\beta)\psi(\beta)$. A direct application of the theorem 1 in § 2 leads to

$$\delta_{\text{grad } u}(\eta) = -\frac{1}{n^2}\{-d + \langle m(\chi-\eta),\ m(\chi-\eta) - \frac{1}{2}t\rangle\} + o(n^{-2}). \tag{3.2}$$

where t is a vector having the i-th component,

$$t_i = \sum_{j,k=1}^{d} T_{ijk}\, g^{jk},$$

where g^{jk} is the inverse element of the Fisher information matrix $[(\partial^2/\partial\beta^j\partial\beta^k)\psi(\beta)]$ and T_{ijk} are components of the skewness tensor defined by

$$T_{ijk} = E_\beta\{(x_i-\eta_i)(x_j-\eta_j)(x_k-\eta_k)\} = \frac{\partial^3\psi(\beta)}{\partial\beta^i\partial\beta^j\partial\beta^k}.$$

We call t the skewness vector in this sense.

For example, let x be the number of successes in n independent Bernoulli trials with probability of success p. The Bayes estimator of p for a conjugate Beta prior with constants a and b is given by $\hat{p}^* = (x+a)/(n+a+b)$; cf. § 5.2 in Bernardo and Smith (1994). We get, from (3.1) and $t = 1-2p$,

$$\delta_h(p) = a+b+\frac{\{a-(a+b)p\}\{(1-a-b)p+a-\frac{1}{2}\}}{p(1-p)},$$

which becomes $1/n^2$ if $a = b = \frac{1}{2}$. Such a choice for a and b yields the uniform dominance of $\hat{p}^*$ over $\hat{p}$.

This result is extended to a multinomial distribution model with cell probabilities $(p_0, \ldots, p_d)$ in a d-dimensional simplex. When we take $p = (p_1, \ldots, p_d)^{\mathrm{T}}$ as the expectation parameter, the Bayes estimator of p for the conjugate Dirichlet prior is $\hat{p}^* = (n\hat{p}+m\chi)(n+m)$. We note the remarkable relation between the expectation parameter and the skewness vector as

$$t = \delta - (d+1)p$$

with $\delta = (1, \ldots, 1)^{\mathrm{T}}$. From this relation it follows that

$$\delta_h(p) = \frac{m}{n^2}\{d + \langle \chi - p,\, (\frac{d+1}{2} - m)p + m\chi - \frac{1}{2}\delta\rangle\},$$

which is reduced to $\frac{1}{2}d(d+1)$ by fixing $m = \frac{1}{2}(d+1)$ and $\chi = \{(d+1)\delta\}^{-1}$. In this way we can extend the above result for $d = 1$ to the multinomial distribution.

We next discuss two familiar classes of distributions. First, consider the family of exponential distributions with the density function $p(x,\beta) = \exp(-x\beta + \log\beta)$ on $\{x > 0\}$.

Then the expectation parameter η is $-1/\beta$, the information $g(\beta) = 1/\beta^2$ and the skewness $t = T(\beta)g(\beta)^{-1} = -2\beta^{-1}$. The Bayes estimator $\hat{\eta}^* = (m+n)^{-1}(n\hat{\eta} + m\chi)$ of η with a Gamma prior never dominates the maximum likelihood estimator $\hat{\eta}$ for any χ since we have

$$\delta_h(p) = m\{\frac{m(\chi-1)}{\eta}(1 - \frac{m(\chi-1)}{\eta}) + 1\},$$

which necessarily changes sign for m and χ as η varies on $(-\infty, 0)$. Secondly, consider a Poisson family with the probability $P\{X = x\} = (x!)^{-1}\eta^x \exp(-\eta)$ for an integer $x \geq 0$. Then it has the information $g(\beta) = e^\beta$ with $\beta = \log \eta$ and the skewness $t = T(\beta)g(\beta)^{-1} = 1$, and hence the Bayes estimator $\hat{\eta}^*$ for the Gamma prior satisfies

$$\delta_h(p) = \frac{m(\chi-\eta)}{\eta} - \frac{2m^2(\chi-\eta)^2}{\eta} - 2m.$$

The argument similar to the exponential distribution leads to no dominance of Bayes estimator over the maximum likelihood estimator. We heuristically consider an estimator having the form $\hat{\eta}^{**} = \hat{\eta} + n^{-1}a$, which can be viewed as a limit of a Bayes estimator in the following sense:

$$\hat{\eta}^{**} = \lim_{m \to 0} \frac{n\hat{\eta} + m\frac{a}{m}}{n+m}.$$

Then we get

$$R(\hat{\eta}, \eta) - R(\hat{\eta}^{**}, \eta) = \frac{1}{n^2\eta}\{-(a - \frac{1}{2})^2 + \frac{1}{4}\} + o(n^{-2}),$$

which attains a maximum at $a = \frac{1}{2}$ for any η. We note that if $a = \frac{1}{2}$, then the estimator satisfies the mean Pythagorean identity.

4. General risk structure

In the previous sections we discussed the asymptotical improvement of the maximum likelihood estimator with respect to the Kullback-Leibler risk. However, the improvement can also hold for other risk functions which are parametrization-invariant. A typical example is

$$D_\alpha(f_1, f_2) = \frac{4}{\alpha^2 - 1}\{1 - \int f_1(x)^{\frac{1-\alpha}{2}} f_2(x)^{\frac{1+\alpha}{2}} d\mu(x)\}$$

with $\mid \alpha \mid < 1$, which has been called the α-order divergence (Amari, 1982 a, b). It is noted that $\lim_{\alpha \downarrow -1} D_\alpha(f_1, f_2) = D(f_1, f_2)$ and $\lim_{\alpha \uparrow 1} D_\alpha(f_1, f_2) = D(f_2, f_1)$. Thus the family of α-order divergences includes the Kullback-Leibler divergence as a limit and also includes the squared Hellinger distance by setting $\alpha = 0$.

In this section we consider a class of risk functions of the form $R^{(\rho)}(\hat{\theta},\theta) = nE\rho(\hat{\theta},\theta)$, where the function ρ defined on $\mathcal{P}\times\mathcal{P}$ satisfies $\rho(\theta_1,\theta_2)\geq 0$ with equality if and only if $\theta_1=\theta_2$, and

$$\Big(\frac{\partial^2}{\partial\theta^i\partial\theta^j}\rho(\theta,\theta_0)\Big)_{\theta_0=\theta} = g_{ij}(\theta) \tag{4.1}$$

for any θ of Θ. We note that any α-order divergence satisfies (4.1). This requirement permits us to adopt $R^{(\rho)}$ as a measure of the first-order efficiency in the sense of (2.1). The function ρ generates a pair of affine connections $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$ with coefficients

$$\Gamma^{(\rho)}_{ij,k}(\theta) = \Big(-\frac{\partial^3}{\partial\theta_1^i\partial\theta_1^j\partial\theta_2^k}\rho(\theta_1,\theta_2)\Big)_{\substack{\theta_1=\theta\\ \theta_2=\theta}}$$

and

$${}^*\Gamma^{(\rho)}_{ij,k}(\theta) = \Big(-\frac{\partial^3}{\partial\theta_2^i\partial\theta_2^j\partial\theta_1^k}\rho(\theta_1,\theta_2)\Big)_{\substack{\theta_1=\theta\\ \theta_2=\theta}}$$

(Eguchi 1983, 1992). Note that, when the Kullback-Leibler divergence $D(\theta_1,\theta_2)$ is used as $\rho(\theta_1,\theta_2)$, $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$ are the mixture connection $\Gamma^{(m)}$ and the exponential connection $\Gamma^{(e)}$ respectively. We now define the following linear connection $\tilde{\Gamma}$ by

$$\tilde{\Gamma} = \Gamma^{(\rho)} + \frac{1}{2}{}^*\Gamma^{(\rho)} - \frac{1}{2}\Gamma^{(m)}$$

and the divergence with respect to $\tilde{\Gamma}$

$$(\widetilde{\mathrm{div}}\ h)(\theta) = \sum_i \frac{\partial h^i(\theta)}{\partial\theta^i} + \sum_{i,j,k}\tilde{\Gamma}_{ij,k}(\theta)h^i(\theta)g^{jk}(\theta), \tag{4.2}$$

and hence we write $\tilde{\Delta} = \widetilde{\mathrm{div}}\ \mathrm{grad}$.

Theorem 2. *Under the same assumption in Theorem 1,*

$$\delta_h^{(\rho)}(\theta) = -\frac{1}{n^2}\{\widetilde{\mathrm{div}}(h) + \parallel h \parallel^2\} + o(n^{-2}).$$

The proof will be given in Appendix 2.

By a similar argument to that of Theorem 2 we can extend Cororally of Theorem 1 to the following:

$$R^{(\rho)}(\hat{\theta},\theta) - R^{(\rho)}(\hat{\theta}^*,\theta) = E_\theta\{\rho(\hat{\theta},\,\hat{\theta}^*)\}$$

for any θ in Θ if and only if $\tilde{\Delta}\,\exp(u) = 0$ on Θ.

Theorem 2 provides us with a common structure for the risk functions satisfying (4.1). The corresponding connection in the case of the α-order divergence D_α is given by

$$\tilde{\Gamma} = \frac{1-\alpha}{4}\Gamma^{(m)} + \frac{3+\alpha}{4}\Gamma^{(e)},$$

or the $\frac{1}{2}(1+\alpha)$ connection, which implies

$$-\exp(-u)\tilde{\Delta}\exp(u) = -\exp(-u)\Delta\exp(u) + \frac{1+\alpha}{4}(\text{grad } u)^T t, \tag{4.3}$$

where t is the skewness vector field with the i-th component $t_i = \sum(\Gamma^{(m)}_{ij,k} - \Gamma^{(e)}_{ij,k})g^{jk}$ as the special form is seen in Section 3. In particular, if $\alpha = -1$, or equivalently $D_\alpha = D$, then $\tilde{\Gamma}$ becomes the Levi-Civita connection, so that Theorem 1 is regarded as a special case of Theorem 2. In view of (4.3) we get the condition, as in the following corollary, that the asymptotical improvement is conserved in the class of α-divergence risk functions.

COROLLARY. *Suppose that the estimator defined in (2.3) satisfies (2.4), and so improves maximum likelihood in terms of the Kullback-Leibler risk. Then the improvement is still valid for the α-order divergence risk for any $\alpha, |\alpha| \leq 1$, if $(\text{grad } u)^{\mathrm{T}} t > 0$ on Θ.*

We return to the example of the multinomial distribution model as in § 3. From the remarkable property that $t = \delta - (d+1)p$ it follows that

$$(\text{grad } u)^{\mathrm{T}} t = (d+1)^2 \parallel p - \frac{1}{(d+1)}\delta \parallel^2 .$$

Consequently, we conclude the dominance of the Bayes estimator for p over the maximum likelihood estimator for every α-order divergence risk.

5. Discussion

We consider a situation where maximum likelihood estimators can be improved by adding a vector field. By definition, maximum likelihood can be viewed as minimization of the Kullback-Leibler divergence, from a given sample distribution function to the statistical model. This implies not minimization of the risk suffered but another minimization based on the data. Hence the maximum likelihood method seems to overfit to the sample distribution or the data. In our main result the estimator $\hat{\theta} + n^{-1}h(\hat{\theta})$ asymptotically improves

the maximum likelihood estimator $\hat{\theta}$ by correction for overfiting made by a dynamical force associated with the vector field h. We have used the concepts of gradient, divergence and the Laplace-Beltrami operator as developed in electromagnetics or potential theory. In the Bayesian argument the prior density function leads to such a dynamical interpretation in terms of a potential function. For the multinomial distribution, for example, we obtain that the observed frequency vector is improved by shrinkage towards the average of cell probability vectors on the basis of the conjugate Dirichlet prior.

It is reasonable to expect that the mean Pythagorean identity holds in a finite sample setting in restricted cases. For example the James-Stein estimator satisfies (2.4) for $d \geq 3$ for any size of sample. Further research will pursue the question of to what extent, and under what conditions, the results developed here hold for small samples.

Acknowledgements

We are grateful to Hiroshi Nagaoka for his helpful suggestions in the proof of Theorem 1, and to John B. Copas and E. Gatierrez-Pena for valuable comments on the first draft of this paper, and to an associate editor and three referees for kind and instructive advice to revision.

Appendices

Appendix 1. *Proof of Theorem in § 2*

We use differential-geometric notation (Amari, 1985) for the probability density function $f(x;\theta)$, supposing the parametric family $\mathcal{P}$ to satisfy mild regularity conditions. The set of score functions

$$e_i = \frac{\partial}{\partial \theta^i} \log f(x;\theta) \qquad (i = 1, \ldots, d)$$

is regarded as a basis of the tangent space of $\mathcal{P}$ at $f(x;\theta)$. Thus the information metric g, the mixture connection $\Gamma^{(m)}$ and the exponential connection $\Gamma^{(e)}$ are defined componentwise by

$$g_{ij}(\theta) = E(e_i e_j),\ \Gamma^{(e)}_{ij,k}(\theta) = E(\frac{\partial e_i}{\partial \theta^j} e_k) \ \text{ and } \ \Gamma^{(m)}_{ij,k}(\theta) = \Gamma^{(e)}_{ij,k}(\theta) + T_{ijk}(\theta),$$

respectively, with respect to $\theta = (\theta^i)$, where $\{T_{ijk}(\theta) = E(e_i e_j e_k)\}$ are the components of the skewness tensor, see Dawid (1975) and Lauritzen (1987). The Levi-Civita connection

$\overline{\Gamma}$ with respect to g is expressed as the 0-connection

$$\overline{\Gamma} = \frac{1}{2}(\Gamma^{(m)} + \Gamma^{(e)}) \tag{A.1}$$

with the α-connection $\Gamma^{(\alpha)} = \frac{1}{2}(1-\alpha)\Gamma^{(m)} + \frac{1}{2}(1+\alpha)\Gamma^{(e)}$. Similarly, the tangent space on n-replicated form $p(y_n;\theta)$ of $f(x;\theta)$ is spanned by the basis

$$\overline{e}_i = \overline{e}_i(\theta) = \frac{1}{n}\frac{\partial}{\partial\theta^i}\log p(y_n;\theta) = \frac{1}{n}\sum_{a=1}^{n}\frac{\partial}{\partial\theta^i}\log f(x_a;\theta) \quad (i = 1, \ldots, d).$$

We next define

$$\overline{f}_{ij} = \overline{f}_{ij}(\theta) = \frac{\partial}{\partial\theta^i}\overline{e}_j - \Gamma_{ij}^{(e)k}\overline{e}_k + g_{ij},$$

where the summation convention is used for the index k and $\Gamma_{ij}^{(e)k} = \Gamma_{ij,l}^{(e)}g^{kl}$ with the elements g^{kl} of the inverse of the matrix (g_{lk}). Hereafter we use this convention of upper and lower indices as well as the summation convention. We note that $E(\overline{f}_{ij}) = 0$ and $\mathrm{Cov}(\overline{e}_i, \overline{f}_{jk}) = 0$. The following lemma for the asymptotic behavior of the maximum likelihood estimator will be necessary for the proof of Theorem.

LEMMA. *The maximum likelihood estimator $\hat{\theta}$ can be expressed as*

$$(\hat{\theta} - \theta)^i = \overline{e}^i + \overline{f}_j^i\overline{e}^j - \frac{1}{2}\Gamma_{jk}^{(m)i}\overline{e}^j\overline{e}^k + O_P(n^{-\frac{3}{2}}) \tag{A.2}.$$

Proof. The system of likelihood equations is $\overline{e}_i(\hat{\theta}) = 0$ $(i = 1, \ldots, d)$, which is expanded as

$$\overline{e}_i + \frac{\partial\overline{e}_i}{\partial\theta^j}\overline{\theta}^j + \frac{1}{2}\frac{\partial^2\overline{e}_i}{\partial\theta^j\partial\theta^k}\overline{\theta}^k\overline{\theta}^j = 0$$

to the second-order with $\overline{\theta} = \hat{\theta} - \theta$. This equation is rewritten as

$$\overline{e}_i - g_{ij}\overline{\theta}^j + (\frac{\partial\overline{e}_i}{\partial\theta^j} + g_{ij})\overline{\theta}^j + \frac{1}{2}E(\frac{\partial^2\overline{e}_i}{\partial\theta^j\partial\theta^k})\overline{\theta}^k\overline{\theta}^j = 0,$$

which is inverted to

$$\overline{\theta}^i = \overline{e}^i + g^{ij}(\frac{\partial\overline{e}_j}{\partial\theta^k} + g_{jk})\overline{e}^k + \frac{1}{2}g^{ij}E(\frac{\partial^2\overline{e}_i}{\partial\theta^k\partial\theta^l})\overline{e}^k\overline{e}^l. \tag{A.3}$$

Alternatively, we have

$$E(\frac{\partial^2\overline{e}_i}{\partial\theta^j\partial\theta^k}) = -(\Gamma_{ij,k}^{(e)} + \Gamma_{ki,j}^{(e)} + \Gamma_{jk,i}^{(m)}) \tag{A.4}$$

by differentiating both sides of $E(\overline{e}_i\overline{e}_j) = -E(\partial\overline{e}_i/\partial\theta^j)$ with respect to the k-th component θ^k. The substitution of (A.4) into (A.3) leads to (A.2), which completes the proof.

The Kullback-Leibler divergence satisfies a fundamental identity:

$$\begin{aligned} &D(\theta_1,\, \theta_3) - D(\theta_1\, \theta_2) - D(\theta_2,\, \theta_3) \\ &\qquad = \int \{f(x;\theta_1) - f(x;\theta_2)\}\{\log f(x;\theta_2) - \log f(x;\theta_3)\}\mu\{dx\} \end{aligned} \tag{A.5}$$

The proof of Theorem 1 uses Lemma 1 and Equation (A.5).

Proof of Theorem 1. From (A.5) it follows that

$$\begin{aligned} \Xi &= D(\hat{\theta},\, \theta) - D(\hat{\theta},\, \hat{\theta}^*) - D(\hat{\theta}^*,\, \theta) \\ &= \int \{f(x;\hat{\theta}) - f(x;\hat{\theta}^*)\}\{\log f(x;\hat{\theta}^*) - \log f(x;\theta)\}d\nu(x). \end{aligned}$$

We expand Ξ at $f(x;\theta)$ by neglecting $o_p(n^{-2})$:

$$\begin{aligned} \Xi = \int &(\hat{\theta} - \hat{\theta}^*)^i\{\frac{\partial}{\partial\theta^i}f(x;\theta) + \frac{\partial^2}{\partial\theta^i\partial\theta^j}f(x;\theta)(\hat{\theta}^* - \theta)^j\} \\ &\times \{(\hat{\theta}^* - \theta)^k\frac{\partial}{\partial\theta^k}\log\, f(x:\theta) + \frac{1}{2}(\hat{\theta}^* - \theta)^k(\hat{\theta}^* - \theta)^l\frac{\partial^2}{\partial\theta^k\partial\theta^l}\log\, f(x;\theta)\}\mu(dx) \\ &+ o_p(n^{-2}), \end{aligned}$$

which is, in differential-geometric notation, expressed as

$$\begin{aligned} \Xi = &(\hat{\theta} - \hat{\theta}^*)^i(\hat{\theta}^* - \theta)^j g_{ij} \\ &+ (\hat{\theta} - \hat{\theta}^*)^i(\hat{\theta}^* - \theta)^j(\hat{\theta}^* - \theta)^k(\frac{1}{2}\Gamma^{(e)}_{jk,i} + \Gamma^{(m)}_{ij,k}) + o_p(n^{-2}). \end{aligned}$$

From the lemma it follows that

$$\begin{aligned} \Xi = \frac{1}{n}[&\{-h^i - \frac{\partial}{\partial\theta^k}h^i\overline{e}^k\} \times \{\overline{e}^j + \overline{f}^j_l\overline{e}^l - \frac{1}{2}\Gamma^{(m)j}_{kl}\overline{e}^k\overline{e}^l + \frac{1}{n}h^j\}g_{ij} \\ &- h^i\overline{e}^j\overline{e}^k(\frac{1}{2}\Gamma^{(e)}_{jk,i} + \Gamma^{(m)}_{ij,k})] + o_p(n^{-2}), \end{aligned}$$

which is expressed as the final form

$$\begin{aligned} \Xi = -\frac{1}{n}\{&h^i(\overline{e}_i + \overline{f}_{ij}\overline{e}^j) - \frac{\partial}{\partial\theta^k}h^i\overline{e}^k\overline{e}_i \\ &- h^i\overline{e}^j\overline{e}^k(\frac{1}{2}\Gamma^{(e)}_{jk,i} + \frac{1}{2}\Gamma^{(m)}_{jk,i}) - \frac{1}{n}h^ih^jg_{ij}\} + o_p(n^{-2}). \end{aligned}$$

By taking the expectation under the supposed density $p(y_n;\theta)$ we obtain

$$E_\theta(\Xi) = -\frac{1}{n^2}\{\frac{\partial}{\partial\theta^i}h^i - h^i\overline{\Gamma}_{ij,k}g^{jk} - h^ih^jg_{ij}\} + o(n^{-2})$$

because of $(A.1)$. By the definition of the divergence operator at (2.2) in Section 2,

$$E_\theta(\Xi) = -\frac{1}{n^2}\{\text{div } h + \langle h,\, h\rangle\} + o(n^{-2}),$$

which completes the proof.

Appendix 2. *Proof of Theorem 2 in § 5*

In the notation of Section 5, a Taylor expansion leads to

$$\begin{aligned}\rho(\hat\theta^*,\theta) - \rho(\hat\theta,\theta) = {}& \frac{1}{2}\sum_{i,j} g_{ij}(\overline{\theta}_2^i\overline{\theta}_2^j - \overline{\theta}_1^i\overline{\theta}_1^j) + \frac{1}{6}\sum_{i,j,k}({}^*\Gamma^{(\rho)} + 2\Gamma^{(\rho)})_{ij,k}(\overline{\theta}_2^i\overline{\theta}_2^j\overline{\theta}_2^k - \overline{\theta}_1^i\overline{\theta}_1^j\overline{\theta}_1^k)\\ &+ \frac{1}{24}\sum_{i,j,k,l}\frac{\partial^4\rho(\theta,\theta_0)}{\partial\theta^i\partial\theta^j\partial\theta^k\partial\theta^l}_{|\theta_0=\theta}(\overline{\theta}_2^i\overline{\theta}_2^j\overline{\theta}_2^k\overline{\theta}_2^l - \overline{\theta}_1^i\overline{\theta}_1^j\overline{\theta}_1^k\overline{\theta}_1^l) \qquad (A.6)\end{aligned}$$

to the fourth order of $\overline{\theta}_i$ with $\overline{\theta}_1 = \hat\theta - \theta$ and $\overline{\theta}_2 = \hat\theta^* - \theta$. This expansion is derived from (3.1) and the equation

$$[\frac{\partial^3}{\partial\theta_1^i\partial\theta_1^j\partial\theta_1^k}\rho(\theta_1,\theta_2)]_{\substack{\theta_1=\theta\\ \theta_2=\theta}} = \Gamma^{(\rho)}_{ij,k}(\theta) + \Gamma^{(\rho)}_{ki,j}(\theta) +^* \Gamma^{(\rho)}_{jk,i}(\theta),$$

which is given by differentiating both the sides of

$$[\frac{\partial^2}{\partial\theta_1^j\partial\theta_1^k}\rho(\theta_1,\theta_2)]_{\substack{\theta_1=\theta\\ \theta_2=\theta}} = [-\frac{\partial^2}{\partial\theta_1^j\partial\theta_2^k}\rho(\theta_1,\theta_2)]_{\substack{\theta_1=\theta\\ \theta_2=\theta}}.$$

It follows from Lemma 1 that the first term in the right-side of (A.6) is expressed as

$$\begin{aligned}&\sum_{i,j}\frac{1}{2}g_{ij}(\overline{\theta}_2^i\overline{\theta}_2^j - \overline{\theta}_1^i\overline{\theta}_1^j)\\ &\quad= \sum_{i,j}\frac{1}{2}g_{ij}\{(\overline{e}^i + \overline{\delta}^i + \frac{h^i}{n} + \frac{1}{n}\sum_k\frac{\partial h^i}{\partial\theta^k}\overline{e}^k)(\overline{e}^j + \overline{\delta}^j + \frac{h^j}{n} + \frac{1}{n}\sum_k\frac{\partial h^j}{\partial\theta^k}\overline{e}^k)\}\\ &\qquad - \sum_{i,j}\frac{1}{2}g_{ij}\{(\overline{e}^i + \overline{\delta}^i + \frac{1}{n}\sum_k\frac{\partial h_1^i}{\partial\theta^k}\overline{e}^k)(\overline{e}^j + \overline{\delta}^j)\}\\ &\quad= \frac{1}{2n^2}\sum_{i,j}g_{ij}(h^ih^j) + \frac{1}{n}\sum_{i,j}\frac{\partial h^i}{\partial\theta^j}\overline{e}^j\overline{e}_i\\ &\qquad - \frac{1}{2n}\sum_{i,j,k}(\Gamma^{(m)} + 2\Gamma^{(e)})_{ij,k}\overline{e}^i\overline{e}^jh^k + o_p(n^{-2}).\end{aligned}$$

Thus the second term is

$$\sum_{i,j,k}({}^*\Gamma^{(\rho)}+2\Gamma^{(\rho)})_{ij,k}(\overline{\theta}_2^i\overline{\theta}_2^j\overline{\theta}_2^k-\overline{\theta}_1^i\overline{\theta}_1^j\overline{\theta}_1^k)=\sum_{i,j,k}3({}^*\Gamma^{(\rho)}+2\Gamma^{(\rho)})_{ij,k}\overline{e}^i\overline{e}^jh^k$$

and the last term vanishes to $o_p(n^{-2})$. Consequently, we get

$$\begin{aligned}\rho(\hat{\theta}^*,\theta)-\rho(\hat{\theta},\theta)=&\frac{1}{2n^2}\sum_{i,j}g_{ij}h^ih^j\\&+\frac{1}{n}\sum_{i,j}\frac{\partial h^i}{\partial\theta^j}\overline{e}^j\overline{e}_i+\frac{1}{n}\sum_{i,j}\frac{\partial\overline{e}_i}{\partial\theta^j}\overline{e}^jh^i\\&+\frac{1}{2n}\sum_{i,j,k}(-2\Gamma^{(e)}-\Gamma^{(m)}+2\Gamma^{(\rho)}+\Gamma^{(\rho)})_{ij,k}\overline{e}^i\overline{e}^jh^k+o_p(n^{-2}).\end{aligned}$$

By taking the expectation of both the sides of this equation we get

$$\begin{aligned}&R^{(\rho)}(\hat{\theta}^*,\theta)-R_n^{(\rho)}(\hat{\theta},\theta)\\&\quad=-\frac{1}{n^2}\{\sum_i\frac{\partial h^i}{\partial\theta^j}+\sum_{i,j,k}(-\frac{1}{2}\Gamma^{(m)}+{}^*\Gamma^{(\rho)}+\frac{1}{2}\Gamma^{(\rho)})h^ig^{jk}+\frac{1}{2}\parallel h\parallel^2\}+o(n^{-2})\\&\quad=-\frac{1}{n^2}\{\widetilde{\mathrm{div}}\ h+\frac{1}{2}\parallel h\parallel^2\}+o(n^{-2}).\end{aligned}$$

Thus we obtain

$$R^{(\rho)}(\hat{\theta},\theta)-E_\theta\rho(\hat{\theta},\hat{\theta}^*)-R^{(\rho)}(\hat{\theta}^*,\theta)=-\frac{1}{n^2}\{\widetilde{\mathrm{div}}(h+\langle h,\,h\rangle\}+o(n^{-2})$$

because

$$E_\theta\rho(\hat{\theta},\hat{\theta}^*)=\frac{1}{2n^2}\parallel h\parallel^2+o(n^{-2}).$$

The proof of Theorem 2 is now complete.

REFERENCES

Abraham, R. and Marsden, J. E. (1987). *Foundations of Mechanics.* Addison-Wesley, California.

Amari, S. (1982a). Differential geometry of curved exponential families - curvatures and information loss. *Ann. Statist.* **10**, 357-87.

Amari, S. (1982b). Geometric theory of asymptotic ancillarity and conditional inference. *Biometrika* **69**, 1-17.

Amari, S. (1985). *Differential-Geometrical Method in Statistics*. Lecture Note in Statistics **32**, Springer, New York.

Amari, S., and Nagaoka, H. (2000). *Methods of Information Geometry*. Translations of Mathematical Monographs **191**. Oxford University Press, Oxford.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, Wiley, New York.

Critchley, F., P.K. Marriott and M.H. Salmon (2000). An elementary account of Amari's expected geometry. P. K. Marriott and M. H. Salmon (Eds.), *Applications of Differential Geometry to Econometrics*, Cambridge University Press, 294 - 315.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Information theory* **36**, 453-71.

Dawid, A. P. (1975). Discussion to Efron's paper. *Ann. Statist.* **3**, 1231-4.

Diaconis, P and Ylvisker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269-81.

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

Eguchi, S. (1992). Geometry of minimum contrast. *Hiroshima Math. J.* **22**, 631-47.

Eguchi,S. and Copas,J. (1998). A class of local likelihood methods and near-parametric asymptotics. *J. Royal Statist. Soc.* B **60**, 709-724.

Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **122**, 700-25.

Ghosh, J.K. and Sinha, B.K. (1981). A necessary and sufficient condition for second order admissibility with applications to Berkson's bioassay problem. *Ann. Statist.* **9**, 1334-1338.

Hartigan, J.A. (1998) The Maximum Likelihood Prior. *Ann. Statist.* **26**, 2083-2103.

Komaki, F. (1996). On asymptotic properties of predictive distributions *Biometrika* **83**, 299-313.

James , W. and Stein, C. (1961). Estimation on quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1**, 361-80.

Lauritzen, S. (1987). Statistical manifold, *Differential geometry in statistical inference*, IMS Lecture Note Monograph Series, **10**, Hayward, California.

Nagaoka, H and Amari, S. (1982). Differential geometry of smooth families of probability distributions. METR, 82-7, University of Tokyo.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical Parameters *Bull. Calcutta Math. Soc.* **37**, 81-9.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution, ø*Ann. Statist.* **9**, 1135-51.

Yanagimoto, T. (1994). The Kullback-Leibler risk of the Stein estimator and the conditional MLE. *Ann. Inst. Statist. Math.* **46**, 29-41.

Yanagimoto, T. and Ohnishi, T. (2005). Standardized posterior mode for the flexible use of a conjugate prior. *J. Statist. Plann. Inf.* **131**, 253-269.