Ann. Inst. Statist. Math.
Vol. 46, No. 1, 29–41 (1994)

THE KULLBACK-LEIBLER RISK OF THE STEIN ESTIMATOR AND THE CONDITIONAL MLE

TAKEMI YANAGIMOTO

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan

(Received October 29, 1992; revised June 22, 1993)

Abstract. The decomposition of the Kullback-Leibler risk of the maximum likelihood estimator (MLE) is discussed in relation to the Stein estimator and the conditional MLE. A notable correspondence between the decomposition in terms of the Stein estimator and that in terms of the conditional MLE is observed. This decomposition reflects that of the expected log-likelihood ratio. Accordingly, it is concluded that these modified estimators reduce the risk by reducing the expected log-likelihood ratio. The empirical Bayes method is discussed from this point of view.

Key words and phrases: Conditional inference, empirical Bayes method, expected log-likelihood ratio, exponential dispersion model, maximum likelihood estimator, Stein estimator.

1. Introduction

As Neyman and Scott (1948) noted first, the maximum likelihood estimator (MLE) does not perform well, when a number of parameters to be estimated is large. As we will discuss later in familiar examples, such a number could be very small, say 2 or 3. Since a model containing many parameters becomes popular, this fact is more important in practice than it was believed. To specify our situation we concentrate on the simultaneous estimation of the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ and a single dispersion parameter θ. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a sample vector of size n having the density function $\prod p(x_i; \mu_i, \theta) = p(\boldsymbol{x}; \boldsymbol{\mu}, \theta)$.

There are various approaches to improving the MLE. The conditional MLE is probably the most classical one, which was begun by Fisher (1935) and advocated later by Andersen (1970). The practical application of the conditional MLE is often seen in the estimation of the dispersion parameter (Yanagimoto and Anraku (1989)). When the conditional distribution given an estimator $\hat{\boldsymbol{\mu}}$ is free from $\boldsymbol{\mu}$, the conditional MLE $\hat{\theta}_c$ is defined by maximizing the conditional likelihood. Suitable conditions for recommending the conditional MLE are discussed by many authors including Barndorff-Nielsen (1978), Lindsay (1982), Godambe (1984), Cox and Reid (1987) and Yanagimoto (1987).

A striking estimator of μ under the normality assumption was proposed by James and Stein (1961), which dominates the MLE $\hat{\mu}_u = \boldsymbol{x}$ with respect to a loss. This estimator was first regarded as a rather pathological one, but its relation with the empirical Bayes method as noted in Efron and Morris (1973) attracts our attention more to this estimator. Because of its wide applicability, the empirical Bayes method is becoming a familiar technique in various fields, such as in nonparametric regression estimation, see Morris (1983), Cassella (1985) and Yanagimoto and Yanagimoto (1987), for example.

The aim of the present paper is to obtain a common feature of these two estimators through the decomposition of the risk induced from the Kullback-Leibler separator (Kullback and Leibler (1951)). The Kullback-Leibler loss of an estimator $(\hat{\boldsymbol{\mu}}(\boldsymbol{x}), \hat{\theta}(\boldsymbol{x}))$, which will be called the KL loss and be written as $(\hat{\mu}, \hat{\theta})$ for simplicity, is given by

$$KL(\hat{\boldsymbol{\mu}}, \hat{\theta}; \boldsymbol{\mu}, \theta) = \int \log \left\{ \prod p(z_j; \hat{\mu}_j, \hat{\theta}) \Big/ \prod p(z_j; \mu_j, \theta) \right\} \prod p(z_j; \hat{\mu}_j, \hat{\theta}) \prod dz_j. \tag{1.1}$$

The KL risk, $RKL(\hat{\boldsymbol{\mu}}, \hat{\theta}; \boldsymbol{\mu}, \theta)$, is given by the expectation of $KL(\hat{\boldsymbol{\mu}}, \hat{\theta}; \boldsymbol{\mu}, \theta)$ with respect to $p(\boldsymbol{x}; \boldsymbol{\mu}, \theta)$. The formal extension of the KL loss to that for a pair of estimators $(\hat{\boldsymbol{\mu}}_1, \hat{\theta}_1)$ and $(\hat{\boldsymbol{\mu}}_2, \hat{\theta}_2)$ is possible by $KL(\hat{\boldsymbol{\mu}}_1, \hat{\theta}_1; \hat{\boldsymbol{\mu}}_2, \hat{\theta}_2)$ in (1.1) and therefore $RKL(\hat{\boldsymbol{\mu}}_1, \hat{\theta}_1; \hat{\boldsymbol{\mu}}_2, \hat{\theta}_2)$ also can be defined. In the theory of differential geometry the KL loss is -1-divergence (Amari (1985), Section 3.5). Our terminology is convenient for distinguishing the loss and the risk clearly. The KL risk will be of our primary concern.

When the density function is a member of the exponential family, it is known (Kullback (1959)) that

$$KL(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_u; \boldsymbol{\mu}, \theta) = \log\{p(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_u, \hat{\theta}_u)/p(\boldsymbol{x}; \boldsymbol{\mu}, \theta)\} \tag{1.2}$$

for the MLE, $(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_u)$. We will write the right-hand side of (1.2) as $LR(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_u; \boldsymbol{\mu}, \theta)$, and write the expectation of it as $ELR(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_u; \boldsymbol{\mu}, \theta)$.

The most familiar decomposition formula in the statistical theory would be

$$\sum (x_i - \mu)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \tag{1.3}$$

Suppose $p(\boldsymbol{x}; \mu, \theta)$ is the density function of $N_n(\mu \mathbf{1}, \theta I)$ where $\mathbf{1} = (1, \ldots, 1)'$ and I stands for the $n \times n$ identity matrix. Then dividing (1.3) by 2θ we have $KL(\boldsymbol{x}, \theta; \mu \mathbf{1}, \theta) = KL(\boldsymbol{x}, \theta; \hat{\mu} \mathbf{1}, \theta) + KL(\hat{\mu} \mathbf{1}, \theta; \mu \mathbf{1}, \theta)$ with $\hat{\mu} = \bar{x}$. Thus this derives a decomposition of the KL loss. Geometrically, this is a Pythagorean relation in the space of the probability distributions. The Pythagorean relation of the KL loss has been extensively investigated; see Simon (1973), Amari (1985) and Saville and Wood (1991).

The main result of this paper is to show that a well designed estimator $(\hat{\mu}, \hat{\theta})$ can satisfy a decomposition

$$RKL(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_u; \boldsymbol{\mu}, \theta) = RKL(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_u; \hat{\boldsymbol{\mu}}, \hat{\theta}) + RKL(\hat{\boldsymbol{\mu}}, \hat{\theta}; \boldsymbol{\mu}, \theta).$$

For simplicity of the statement we will suppress the words, "for every μ and θ". Since the KL risk is positive in our examples, this equality implies the inequality $RKL(\hat{\mu}_u, \hat{\theta}_u; \mu, \theta) > RKL(\hat{\mu}, \hat{\theta}; \mu, \theta)$, yielding inadmissibility of the MLE. An advantage of this approach is to obtain an inequality showing inadmissibility of the MLE through an equality regarding the decomposition of the KL risk. In addition, the equality $RKL(\hat{\mu}, \hat{\theta}; \mu, \theta) = ELR(\hat{\mu}, \hat{\theta}; \mu, \theta)$ holds in our examples. Since this equality holds also for the MLE, it follows that $RKL(\hat{\mu}_u, \hat{\theta}_u; \mu, \theta) - RKL(\hat{\mu}, \hat{\theta}; \mu, \theta) = ELR(\hat{\mu}_u, \hat{\theta}_u; \mu, \theta) - ELR(\hat{\mu}, \hat{\theta}; \mu, \theta)$ $(= ELR(\hat{\mu}_u, \hat{\theta}_u; \hat{\mu}, \hat{\theta}))$. Consequently, we can conclude that such an estimator reduces the KL risk by reducing the expected log-likelihood ratio.

The examples of the estimator satisfying the decomposition include the conditional MLE of the variance of the normal distributions and the Stein estimator of the means of the independent normal distributions. Our results present a common reason of favorable performance of these estimators. In the light of their derivations from completely different ideas, the observed notable correspondence looks surprising.

In Section 2, simple forms of the KL loss under the exponential dispersion model are derived, and then fundamental properties are obtained in Section 3. The main results on the Stein estimator and on the conditional MLE are given in Sections 4 and 5. Section 6 treats the empirical Bayes method in relation to the Stein estimator.

2. The exponential dispersion model

To obtain a common feature of the decomposition, our attention focuses on a restricted family of the exponential dispersion model (Jorgensen (1987)). The exponential dispersion model was introduced to describe a wide class of error distributions in the generalized linear model (Nelder and Wedderburn (1972), McCullagh and Nelder (1989)). Properties of this model and other related models were extensively studied by many authors including Barndorff-Nielsen (1978), Morris (1982), Blaesild and Jensen (1985), and Jorgensen (1987). Since our interest is in the comparison of estimators, we will not pursue much the mathematical refinement of the regularity assumptions and an extension of the model.

We introduce here a slightly different notation for the restricted model to emphasize the role of the mean parameter. The density function having mean μ and dispersion parameter θ is expressed as

$$p(x; \mu, \theta) = \exp\left\{\frac{c(\mu)(x-\mu) + C(\mu) - C(x)}{\theta} + b(\theta) + a(x)\right\} \tag{2.1}$$

where the function $C(\mu)$ is a primitive of $c(\mu)$. The function $c(\mu)$ is the canonical link function in the generalized linear model. This model is a member of the exponential family, having the natural parameters $c(\mu)/\theta$ and $1/\theta$. In a usual expression of the exponential dispersion model the parameters $\theta^* = c(\mu)$ and $\lambda = 1/\theta$ were employed. The parameters μ and θ are then orthogonal, that is, the Fisher information matrix is diagonal. The variance is $\theta/c'(\mu)$ and therefore, we call θ the dispersion parameter. Three familiar distributions, the normal, the

gamma and the inverse Gaussian, belong to this family. The explicit forms of the functions appearing in (2.1) are given in Table 1.

Let $\boldsymbol{x}$ be a sample vector from a population having the density function (2.1). Then the MLE of $\boldsymbol{\mu}$, $\hat{\boldsymbol{\mu}}_u$, is $\boldsymbol{x}$. Next suppose that all the means are common, that is, $\boldsymbol{\mu} = \mu\mathbf{1}$. Then the MLE of μ is $\bar{x}$ and that of θ the solution of the equation $C(\bar{x}) - \sum C(x_i)/n + h(\theta) = 0$ with $h(\theta) = -b'(\theta)\theta^2$. The family of distributions is reproductive; more explicitly the sample mean $\bar{x}$ has the density function $p(\bar{x}; \mu, \theta/n)$. This yields the following factorization of the density function.

$$
\begin{aligned}
(2.2) \qquad p(\boldsymbol{x}; \boldsymbol{\mu}, \theta) = \exp&\left\{\frac{c(\mu)(\bar{x}-\mu) + C(\mu) - C(\bar{x})}{\theta/n} + b(\theta/n) + a(\bar{x})\right\} \\
&\cdot \exp\left\{\frac{nC(\bar{x}) - \sum C(x_i)}{\theta}\right. \\
&\qquad \left. + nb(\theta) - b(\theta/n) + \sum a(x_i) - a(\bar{x})\right\} \\
&(= L(\bar{x}; \mu, \theta/n) LC(\boldsymbol{x}; \theta \mid \bar{x})).
\end{aligned}
$$

Note that the second factor in (2.2) is free from μ. The conditional MLE of θ, $\hat{\theta}_c$, is obtained by maximizing $LC(\boldsymbol{x}; \theta \mid \bar{x})$ by discarding the former factor $L(\bar{x}; \mu, \theta/n)$.

Now we give explicit forms of the log-likelihood ratio and the KL loss in this family. First we present the former one, $LR(\hat{\boldsymbol{\mu}}, \hat{\theta}; \boldsymbol{\mu}, \theta) = \log\{p(\boldsymbol{x}; \hat{\boldsymbol{\mu}}, \hat{\theta})/p(\boldsymbol{x}; \boldsymbol{\mu}, \theta)\}$, which is expressed as

$$
\begin{aligned}
(2.3) \quad \sum &\left\{\frac{c(\hat{\mu}_i)(x_i - \hat{\mu}_i) + C(\hat{\mu}_i) - C(x_i)}{\hat{\theta}} - \frac{c(\mu_i)(x_i - \mu_i) + C(\mu_i) - C(x_i)}{\theta}\right\} \\
&+ n(b(\hat{\theta}) - b(\theta)).
\end{aligned}
$$

The expectation of this statistic with respect to the population distribution will be written as $ELR(\hat{\boldsymbol{\mu}}, \hat{\theta}; \boldsymbol{\mu}, \theta)$.

Next we give an explicit form of the KL loss. Using the equality $E(\partial \log p(\boldsymbol{x}; \boldsymbol{\mu}, \theta)/\partial\theta) = 0$, we obtain that $E(C(x)) = C(\mu) - b'(\theta)\theta^2 = C(\mu) + h(\theta)$. Thus the KL loss in a general form is then expressed as

$$
\begin{aligned}
(2.4) \qquad KL(\hat{\boldsymbol{\mu}}, \hat{\theta}; \boldsymbol{\mu}, \theta) = &\sum\left\{-\frac{c(\mu_i)(\hat{\mu}_i - \mu_i) + C(\mu_i) - C(\hat{\mu}_i)}{\theta}\right\} \\
&+ n\left\{b(\hat{\theta}) - \frac{h(\hat{\theta})}{\hat{\theta}} - b(\theta) + \frac{h(\hat{\theta})}{\theta}\right\}.
\end{aligned}
$$

This loss is simplified when only a part of parameters are of interest. In fact we get

$$
\begin{aligned}
KL(\hat{\boldsymbol{\mu}}, \theta; \boldsymbol{\mu}, \theta) &= \sum\left\{-\frac{c(\mu_i)(\hat{\mu}_i - \mu_i) + C(\mu_i) - C(\hat{\mu}_i)}{\theta}\right\}, \\
KL(\hat{\boldsymbol{\mu}}, \hat{\theta}; \hat{\boldsymbol{\mu}}, \theta) &= KL(\boldsymbol{\mu}, \hat{\theta}; \boldsymbol{\mu}, \theta) = n\left\{b(\hat{\theta}) - \frac{h(\hat{\theta})}{\hat{\theta}} - b(\theta) + \frac{h(\hat{\theta})}{\theta}\right\}.
\end{aligned}
$$

Table 1. The explicit forms of the functions in the exponential distribution model with the expression (2.1) for the three distributions.

Distribution	Density	$c(\mu)$	$C(\mu)$	$b(\theta)$	$a(x)$
Normal	$\frac{1}{\sqrt{2\pi\theta}}e^{-(x-\mu)^2/2\theta}$	μ	$\frac{1}{2}\mu^2$	$-\frac{1}{2}\log\theta$	$-\frac{1}{2}\log 2\pi$
Inv. Gauss.	$\sqrt{\frac{1}{2\pi\theta x^3}}e^{-(x-\mu)^2/2\theta\mu^2 x}$	$-\frac{1}{2\mu^2}$	$\frac{1}{2\mu}$	$-\frac{1}{2}\log\theta$	$-\frac{1}{2}\log 2\pi x^3$
Gamma	$\frac{x^{1/\theta-1}}{\Gamma(1/\theta)(\theta\mu)^{1/\theta}}e^{-x/\theta\mu}$	$-\frac{1}{\mu}$	$-\log\mu$	$-\log\Gamma(1/\theta)-\frac{1}{\theta}(1+\log\theta)$	$-\log x$

Note that the KL loss (2.4) is decomposed into

$$KL(\hat{\mu}, \hat{\theta}; \mu, \theta) = KL(\hat{\mu}, \theta; \mu, \theta) + KL(\mu, \hat{\theta}; \mu, \theta).$$

This decomposition permits us to compare the KL risks of the estimators of μ and θ separately. Therefore, we will compare estimators of μ and θ separately.

3. Some properties

In this section we give elementary, important properties of the KL risk and the expected log-likelihood ratio. As noted above, we will compare those of estimators of μ and θ separately. We begin with the comparison of estimators of μ.

Let $\hat{\mu}$, $\hat{\mu}_1$ and $\hat{\mu}_2$ be estimators of μ, and $\hat{\theta}$ an estimator of θ. The first proposition presents a condition for the decomposition of the risk.

PROPOSITION 3.1. *The following three statements are equivalent*:
i) $RKL(\hat{\mu}_1, \theta, \mu, \theta) = RKL(\hat{\mu}_1, \theta; \hat{\mu}_2, \theta) + RKL(\hat{\mu}_2, \theta; \mu, \theta)$,
ii) $RKL(\hat{\mu}_1, \hat{\theta}; \mu, \theta) = RKL(\hat{\mu}_1, \theta; \hat{\mu}_2, \theta) + RKL(\hat{\mu}_2, \hat{\theta}; \mu, \theta)$,
iii) $E\{\sum(c(\hat{\mu}_{2i}) - c(\mu_i))(\hat{\mu}_{1i} - \hat{\mu}_{2i})\} = 0$.
When either i) *or* ii) *holds*, $\hat{\mu}_2$ *is superior to* $\hat{\mu}_1$.

The following proposition is concerned with a condition for the equivalence of the KL risk and the expected log-likelihood ratio. Here we assume that $\hat{\mu}$ and $\hat{\theta}$ are independent.

PROPOSITION 3.2. *The following three statements are equivalent*:
i) $RKL(\hat{\mu}, \theta; \mu, \theta) = ELR(\hat{\mu}, \theta; \mu, \theta)$,
ii) $RKL(\hat{\mu}, \hat{\theta}; \mu, \hat{\theta}) = ELR(\hat{\mu}, \hat{\theta}; \mu, \hat{\theta})$,
iii) $E\{\sum(c(\hat{\mu}_i) - c(\mu_i))(x_i - \hat{\mu}_i)\} = 0$.

The conditions iii) in both the propositions look close. In fact suppose that $\hat{\mu}_1 = \boldsymbol{x}$ $(= \hat{\mu}_u)$ and $\hat{\mu}_2 = \hat{\mu}$ satisfy that of Proposition 3.1 and that $\hat{\mu}$ satisfies that of Proposition 3.2. These conditions are then equivalent. It holds also that $ELR(\hat{\mu}, \theta; \mu, \theta) = RKL(\boldsymbol{x}, \theta; \mu, \theta) - RKL(\boldsymbol{x}, \theta; \hat{\mu}, \theta)$.

Next we discuss the comparison of estimators of θ. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be estimators of θ, and $\hat{\mu}$ be an estimator of μ independent of the estimators of θ. The following two propositions correspond with Propositions 3.1 and 3.2.

PROPOSITION 3.3. *The following four statements are equivalent*:
i) $RKL(\mu, \hat{\theta}_1; \mu, \theta) = RKL(\mu, \hat{\theta}_1; \mu, \hat{\theta}_2) + RKL(\mu, \hat{\theta}_2; \mu, \theta)$,
ii) $RKL(\hat{\mu}, \hat{\theta}_1; \hat{\mu}, \theta) = RKL(\hat{\mu}, \hat{\theta}_1; \hat{\mu}, \hat{\theta}_2) + RKL(\hat{\mu}, \hat{\theta}_2; \hat{\mu}, \theta)$,
iii) $RKL(\hat{\mu}, \hat{\theta}_1; \mu, \theta) = RKL(\hat{\mu}, \hat{\theta}_1; \hat{\mu}, \hat{\theta}_2) + RKL(\hat{\mu}, \hat{\theta}_2; \mu, \theta)$,
iv) $E\{(h(\hat{\theta}_1) - h(\hat{\theta}_2))(1/\hat{\theta}_2 - 1/\theta)\} = 0$.

PROPOSITION 3.4. *The following two statements are equivalent*:
i) $RKL(\hat{\mu}, \hat{\theta}; \hat{\mu}, \theta) = ELR(\hat{\mu}, \hat{\theta}; \hat{\mu}, \theta)$,

ii) $E\{\sum(c(\hat{\mu}_i)(x_i - \hat{\mu}_i) + C(\hat{\mu}_i) - C(x_i) + h(\hat{\theta}))(1/\hat{\theta} - 1/\theta)\} = 0.$

This equivalence still holds when $\hat{\boldsymbol{\mu}}$ is replaced by $\boldsymbol{\mu}$ in i) and ii). Conditions iv) in Proposition 3.3 and ii) in Proposition 3.4 become equivalent when $\hat{\theta}_1$ is the MLE given $\hat{\boldsymbol{\mu}}$ and $\hat{\theta} = \hat{\theta}_2$. This is shown by evaluating $\partial \log p(\boldsymbol{x}; \hat{\boldsymbol{\mu}}, \theta)/\partial\theta$ at $\theta = \hat{\theta}_1$.

It should be noted here that the properties of the KL loss and the log-likelihood corresponding to the above four propositions also hold without taking the expectation. In the following two sections, we will show that Propositions 3.1 and 3.2 present favorable properties of the Stein estimator and that Propositions 3.3 and 3.4 those of the conditional MLE of θ.

4. Stein estimator

James and Stein (1961) showed that a shrinkage estimator $\hat{\boldsymbol{\mu}}_s = (1 - (n - 2)\theta/\|\boldsymbol{x}\|^2)\boldsymbol{x}$ dominates $\hat{\boldsymbol{\mu}}_u = \boldsymbol{x}$ when the loss is the sum of squared differences, that is, the KL loss. This result stimulated various fields of the statistical method such as the ridge estimator and the simultaneous estimation of many parameters. Stein (1956), Brown (1979), Berger (1980) and Dey *et al.* (1987) developed a series of techniques in improving naive simultaneous estimators. Recent successful development of smoothing techniques is also heavily influenced by this estimator, as noted in the Introduction.

Propositions 3.1 and 3.2 are applicable in this situation. Set $\hat{\boldsymbol{\mu}}_k = (1 - k\theta/\|\boldsymbol{x}\|^2)\boldsymbol{x}$. Then $\hat{\boldsymbol{\mu}}_{n-2} = \hat{\boldsymbol{\mu}}_s$ and $\hat{\boldsymbol{\mu}}_0 = \hat{\boldsymbol{\mu}}_u$. The following proposition extends the results due to James and Stein (1961).

PROPOSITION 4.1. *Set* $E_\mu = \{\hat{\boldsymbol{\mu}}_k; k \neq 0\}$. *Then the following five statements for* $\hat{\boldsymbol{\mu}}_k \in E_\mu$ *are equivalent*:

i) $\hat{\boldsymbol{\mu}}_k = \hat{\boldsymbol{\mu}}_s$,
ii) $RKL(\hat{\boldsymbol{\mu}}_u, \theta; \boldsymbol{\mu}, \theta) = RKL(\hat{\boldsymbol{\mu}}_u, \theta; \hat{\boldsymbol{\mu}}_k, \theta) + RKL(\hat{\boldsymbol{\mu}}_k, \theta; \boldsymbol{\mu}, \theta)$,
iii) $RKL(\hat{\boldsymbol{\mu}}_k, \theta; \boldsymbol{\mu}, \theta) = ELR(\hat{\boldsymbol{\mu}}_k, \theta; \boldsymbol{\mu}, \theta)$,
iv) $E\{\sum(\hat{\mu}_{ki} - \mu_k)(x_i - \hat{\mu}_{ki})\} = 0$,
v) $\hat{\boldsymbol{\mu}}_k$ *minimizes* $RKL(\hat{\boldsymbol{\mu}}_k, \theta; \boldsymbol{\mu}, \theta)$.

PROOF. The equivalence of i) and v) is a result by Stein (1956), and that of ii) and v) was stated essentially in Brandwein and Strawderman (1990). The equivalence of iii) and iv) is seen in Proposition 3.2. Consequently, it suffices to show that of iv) and v). The statement iv) is written as

$$E\left[k\theta\left\{\sum \frac{1}{\|\boldsymbol{x}\|^2}\left(\frac{\|\boldsymbol{x}\|^2 - k\theta}{\|\boldsymbol{x}\|^2}x_i^2 - x_i\mu_i\right)\right\}\right]$$
$$= E\left[k\theta\left\{\frac{\|\boldsymbol{x}\|^2 - k\theta}{\|\boldsymbol{x}\|^2} - \frac{\sum x_i\mu_i}{\|\boldsymbol{x}\|^2}\right\}\right] = 0.$$

This statement is that for v) as in James and Stein (1961). □

The statements ii) and iii) provide us with a deeper understanding of the Stein estimator. Statement ii) can be regarded as a type of Pythagorean relation.

If statement ii) holds for the loss KL instead of the risk RKL, then the equality means the Pythagorean relation among the distributions $p(\boldsymbol{x};\mu,\theta)$, $p(\boldsymbol{x};\hat{\mu}_u,\theta)$ and $p(\boldsymbol{x};\hat{\mu}_s,\theta)$. The statement ii) is weaker; it means that the Pythagorean relation holds with taking expectation. Thus the risk $RKL(\hat{\mu}_u,\theta;\hat{\mu}_s,\theta)$ is unexpected and should be avoided. Now the relation (1.2) and these two statements iii) yield

$$RKL(\hat{\mu}_u,\theta;\hat{\mu}_s,\theta) = ELR(\hat{\mu}_u,\theta;\hat{\mu}_s,\theta). \tag{4.1}$$

This equality shows that the difference between the KL risk of the Stein estimator and that of the MLE is equal to that of the expected log-likelihood ratio. Thus if this condition is satisfied, a lower average of the log-likelihood ratio results in a lower KL risk. This fact looks striking, since it superficially contradicts the familiar criterion of the maximum likelihood method. In this regard, it should be noted that the MLE minimizes $KL(\boldsymbol{x},\theta;\hat{\mu},\theta)$ $(= \log p(\boldsymbol{x};\hat{\mu},\theta) + nb(\theta) + \sum a(x_i))$ in this setup, while our aim is to make $RKL(\hat{\mu},\theta;\mu,\theta)$ small.

Various other extensions are possible, some of which will be discussed in the later sections. In here we note that the assumption of the common variance is not essential. Suppose that $x_i \sim N(\mu_i,\theta_i)$, $i = 1,\ldots,n$ and θ_i are known. Then it is possible to apply the Stein estimator to $x_i/\sqrt{\theta_i}$. Next suppose that there exists an unbiased estimator of θ independent of $\boldsymbol{x}$ and that $d\hat{\theta}/\theta$ has the chi-square distribution with d degrees of freedom. As in Stein (1962), the extended Stein estimator becomes

$$\hat{\mu}_s = \left\{1 - \frac{(n-2)d\hat{\theta}}{(d+2)\|\boldsymbol{x}\|^2}\right\}\boldsymbol{x}.$$

Using i) or ii) of Proposition 3.1 and i) of Proposition 3.2, we can extend Proposition 4.1.

5. The conditional MLE

When a factorization property (2.2) holds, we can derive the conditional MLE of θ by maximizing the conditional likelihood $LC(\boldsymbol{x};\theta \mid \bar{x})$. The restriction of our attention to the density function (2.1) yields a simple explicit form of the estimating equation for the conditional MLE, $C(\bar{x}) - \sum C(x_i)/n + h(\theta) - nh(\theta/n) = 0$. An explicit form of $\hat{\theta}_c$ is possible in the cases of the normal and the inverse Gaussian distributions, where the function $h(\theta)$ is $\theta/2$. The conditional MLE is $\sum(x_i - \bar{x})^2/(n-1)$ for the normal distribution, and $\sum(1/x_i - 1/\bar{x})/(n-1)$ for the inverse Gaussian. Thus it holds $\hat{\theta}_c = n\hat{\theta}_u/(n-1)$ in both cases. The factor $n/(n-1)$ is greater than 1, and therefore is regarded as an expansion factor in contrast to a shrinkage one in the Stein estimator.

We discuss first the above two cases. Let $\hat{\theta}_k = k\hat{\theta}_u$ and $E_\theta = \{\hat{\theta}_k \mid k > 0 \text{ but } k \neq 1\}$. The following proposition presents a good correspondence between the conditional MLE and the Stein estimator.

PROPOSITION 5.1. *Suppose that the underlying distribution is normal or inverse Gaussian, and that $\hat{\theta}_k \in E_\theta$. Then the following five statements are equivalent:*

i) $\hat{\theta}_k = \hat{\theta}_c$,
ii) $RKL(\bar{x}\mathbf{1}, \hat{\theta}_u; \mu\mathbf{1}, \theta) = RKL(\bar{x}\mathbf{1}, \hat{\theta}_u; \bar{x}\mathbf{1}, \hat{\theta}_k) + RKL(\bar{x}\mathbf{1}, \hat{\theta}_k; \mu\mathbf{1}, \theta)$,
iii) $RKL(\bar{x}\mathbf{1}, \hat{\theta}_k; \mu\mathbf{1}, \theta) = ELR(\bar{x}\mathbf{1}, \hat{\theta}_k; \mu\mathbf{1}, \theta)$,
iv) $E(\hat{\theta}_u - \hat{\theta}_k)(1/\theta - 1/\hat{\theta}_k) = 0$,
v) $\hat{\theta}_k$ *minimizes* $RKL(\bar{x}\mathbf{1}, \hat{\theta}_k; \mu\mathbf{1}, \theta)$.

PROOF. The equivalence of the statements i), iii) and v) is due to Yanagimoto (1991). In addition, the proof follows from Propositions 3.3 and 3.4.

We observe the good correspondence between Propositions 4.1 and 5.1. Again the statement ii) can be regarded as a type of Pythagorean relation. In addition the equality (4.1) holds also by replacing $(\hat{\mu}_u, \theta; \hat{\mu}_s; \theta)$ with $(\bar{x}\mathbf{1}, \hat{\theta}_u; \bar{x}\mathbf{1}, \hat{\theta}_c)$. Thus we can conclude that the reduction of the KL risk of the conditional MLE is realized by reducing the expected log-likelihood ratio.

The relaxation of the assumption on the distribution is desirable for applications. Consider the case of the gamma distribution. Unfortunately, $\hat{\theta}_u/\hat{\theta}_c$ is not constant in this case. The simulation study by Yanagimoto and Anraku (1989) shows the condition iii) holds approximately, and therefore we can guess the inequality $RKL(\bar{x}\mathbf{1}, \hat{\theta}_u; \mu\mathbf{1}, \theta) > RKL(\bar{x}\mathbf{1}, \hat{\theta}_c; \mu\mathbf{1}, \theta)$ holds. This inequality is important for applications. Next, consider the one way design of the normal population. Let x_{ij} be a sample of size n_i (> 0) from $N(\mu_i; \theta)$, $i = 1, \ldots, n$. Set $N = \sum n_i$. The MLE of μ_i is $\bar{x}_i$ and that of θ is $\sum(x_{ij} - \bar{x}_i)^2/N$. The estimators, $\hat{\mu}_{iu}$, $i = 1, \ldots, n$ and $\hat{\theta}_u$ are mutually independent. On the other hand the conditional MLE of θ given $\bar{x}_i$ is written as $\hat{\theta}_c = N\hat{\theta}_u/(N-n)$ when $N - n > 0$. Recall that $(N-n)\hat{\theta}_c/\theta$ follows the chi-square distribution with $(N-n)$ degrees of freedom. Thus we can apply an extended form of the Stein estimator discussed below Proposition 4.1, and consequently the Stein estimator of μ_i becomes

$$\hat{\mu}_{is} = \left\{ 1 - \frac{(n-2)(N-n)\hat{\theta}_c}{(N-n+2)\sum n_i \bar{x}_i^2} \right\} \bar{x}_i.$$

To apply Propositions 4.1 and 5.1 we consider the N-dimensional parameter vector $\boldsymbol{\mu}$ which takes the value μ_i from the $(N_{i-1}+1)$-th component to the N_i-th with N_i being the sum of n_i up to i. Using the estimators of μ_i, we define $\hat{\boldsymbol{\mu}}_u$ and $\hat{\boldsymbol{\mu}}_s$ in a similar way. Then Proposition 5.1 is applicable to compare $(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_u)$ and $(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_c)$, and an extended version of Proposition 4.1 is applicable to compare $(\hat{\boldsymbol{\mu}}_u, \hat{\theta}_c)$ and $(\hat{\boldsymbol{\mu}}_s, \hat{\theta}_c)$. Consequently, a combination of the Stein estimator and the conditional MLE derives the reasonable estimator $(\hat{\boldsymbol{\mu}}_s, \hat{\theta}_c)$, which dominates the MLE. Again the reduction of the KL risk of this estimator to that of the MLE reflects the reduction of the expected log-likelihood ratio.

The simultaneous estimation of many parameters is an attractive, useful problem. A practical way to obtain an estimator is to apply the empirical Bayes method. The decomposition of the KL risk provides us with new insight into the empirical Bayes method, which will be discussed in the following section.

6. Empirical Bayes method

In spite of its fine performance, the empirical Bayes method does not look very familiar. For simplicity, our interest focuses on the estimation of the mean of $p(\boldsymbol{x};\boldsymbol{\mu},\theta)$ with a known dispersion parameter θ in the case of the normal or the gamma distribution. In addition, we assume the conjugate prior; μ_i and $\tau_i = 1/\mu_i$ are samples of size n from a hyperpopulation having the normal and the gamma distributions with mean λ and the dispersion parameter δ, respectively. Then the posterior mean, or equivalently the posterior mode, becomes $\hat{\boldsymbol{\mu}} = (\delta\boldsymbol{x}+\theta\lambda\mathbf{1})/(\delta+\theta)$ in the normal distribution, and $\hat{\boldsymbol{\mu}} = (\lambda\delta\boldsymbol{x}+\theta\mathbf{1})/\lambda(\delta+\theta)$ in the gamma distribution. The derivation of the latter will be given in the proof of the proposition below. A decomposition formula is given under the Bayesian framework.

PROPOSITION 6.1. *Under the situation stated above we get*

$$E\{RKL(\boldsymbol{x},\theta;\boldsymbol{\mu},\theta)\} = E\{RKL(\boldsymbol{x},\theta;\hat{\boldsymbol{\mu}},\theta) + RKL(\hat{\boldsymbol{\mu}},\theta;\boldsymbol{\mu},\theta)\} \tag{6.1}$$

where the expectation is taken with respect to the prior distribution.

PROOF. The case of the normal distribution is obvious, and only the case of the gamma is shown here. The marginal likelihood of λ and δ to $\boldsymbol{x}$ is

$$L(\boldsymbol{x};\lambda,\delta) = \prod \frac{\Gamma(1/\theta+1/\delta)}{\Gamma(1/\theta)\Gamma(1/\delta)} \frac{\theta^{1/\delta}(\delta\lambda)^{1/\theta}x_i^{1/\theta-1}}{(\theta+\lambda\delta x_i)^{1/\theta+1/\delta}}. \tag{6.2}$$

Thus the posterior density of $\boldsymbol{\tau} = (\ldots, 1/\mu_i, \ldots)$ is

$$pp(\boldsymbol{x};\boldsymbol{\tau}) = \prod \frac{1}{\Gamma\left(\frac{1}{\theta}+\frac{1}{\delta}\right)} \left\{\frac{\theta+\lambda\delta x_i}{\lambda\delta\theta}\right\}^{(1/\theta+1/\delta)} \cdot \tau_i^{(1/\theta+1/\delta-1)} \exp\left\{-\tau_i\left(\frac{x_i}{\theta}+\frac{1}{\lambda\delta}\right)\right\}$$

which yields the posterior mean, that is, $\hat{\tau}_i = \lambda(\theta+\delta)/(\theta+\lambda\delta x_i)$.

Applying the equivalence of i) and iii) of Proposition 3.1, we may show

$$E\left\{\sum \frac{\lambda x_i - 1}{\lambda\delta x_i + \theta} - \sum \frac{\tau_i(\lambda x_i - 1)}{\lambda(\theta+\delta)}\right\} = 0$$

where the expectation is taken with respect to the distribution of x_i and a prior distribution of τ_i. By evaluating the mean of $\partial \log L(\boldsymbol{x};\lambda,\delta)/\partial\lambda$ we can show that the mean of the former term vanishes. It is easy to show the case of the latter term. This completes the proof. □

The result under the Bayesian framework may be discouraging for a frequentist, since the hyperparameters λ and δ are assumed known. A simple way to avoid this assumption is to estimate the hyperparameters by using the likelihood

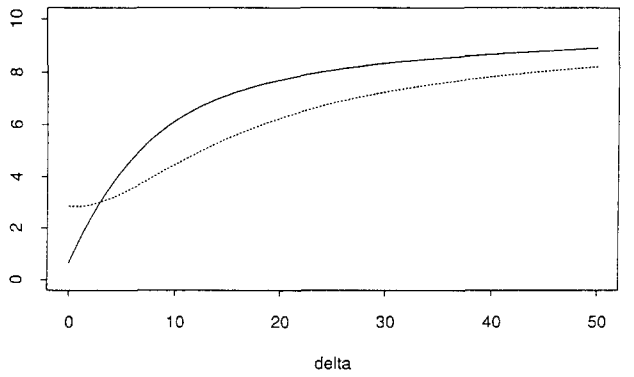

Fig. 1. The graphs of $2RKL(\hat{\mu}_e, \theta; \mu, \theta)$ in the solid line and $2ELR(\hat{\mu}_e, \theta; \mu, \theta)$ in the dotted line with $\delta = \|\mu\|^2$: Case $n = 10$.

of λ and δ as in (6.2). This treatment is a type of empirical Bayes method. Then we obtain an estimator $\hat{\mu}(\hat{\lambda}, \hat{\delta})$ free from the hyperparameters. When the sample size is large, the estimates $\hat{\lambda}$ and $\hat{\delta}$ are expected to be close to their true values. Thus we conjecture that the decomposition (6.1) holds approximately when the sample size is large.

To discuss this conjecture and to pursue the relation with the Stein estimator, we consider the case of the normal distribution with λ being assigned as 0. Since $\boldsymbol{x} \sim N(0, (\delta + \theta)I)$, the maximum likelihood estimator of δ is $[\|\boldsymbol{x}\|^2 - n\theta]^+/n$, where the symbol $[y]^+$ denotes $\max(0, y)$. Then the resulting estimator of μ is written as

$$\hat{\mu}_e = \frac{[\|x\|^2 - n\theta]^+}{\|\boldsymbol{x}\|^2}\boldsymbol{x}.$$

This is close to the positive part Stein estimator, $[\hat{\mu}_s]^+$, which is known to dominate the original Stein estimator. The difference of $\hat{\mu}_e$ and $[\hat{\mu}_s]^+$ is in the coefficients n and $(n-2)$. Recall that the coefficient $(n-2)$ is optimal only in the original Stein estimator. Accordingly, this estimator is appealing, since it is routinely derived by applying the empirical Bayes method. Note that the distribution of $\hat{\mu}_e$ depends on μ only through its norm. As a result, the performance of the estimator does not depend heavily on the assumed prior distribution.

Consider an asymptotic situation where $\|\mu\|^2/n$ converges to a positive constant. It is shown that the quantity $\{RKL(\boldsymbol{x}, \theta; \mu, \theta) - RKL(\boldsymbol{x}, \theta; \hat{\mu}_e, \theta) - RKL(\hat{\mu}_e, \theta; \mu, \theta)\}/n$ tends to zero. Recall that this quantity is written also as $\{ELR(\hat{\mu}_e, \theta; \mu, \theta) - RKL(\hat{\mu}_e, \theta; \mu, \theta)\}/n$. The evaluation of this quantity is useful in practice, since $(n/2)\log(2\pi\theta) + n/2 - \log p(\boldsymbol{x}; \theta; \hat{\mu}_e, \theta)$ is an unbiased estimator of the expected log-likelihood ratio. The close relation of $\hat{\mu}_e$ with the Stein estimator suggests that this quantity is not large for a moderate sample size. To evaluate the difference, we conduct a numerical study. Figures 1 and 2 present $2RKL(\hat{\mu}_e, \theta; \mu, \theta)$ and $2ELR(\hat{\mu}_e, \theta; \mu, \theta)$ for $\theta = 1$ and $\delta = \|\mu\|^2$ in the cases of $n = 10$ and 40. We observe that the KL risk is less than the expected log-likelihood ratio for a small δ, but is greater for a moderate or a large one. The difference between the former quantity and the latter ranges from -5.50 to 2.36 when $n = 40$. The ratio of the

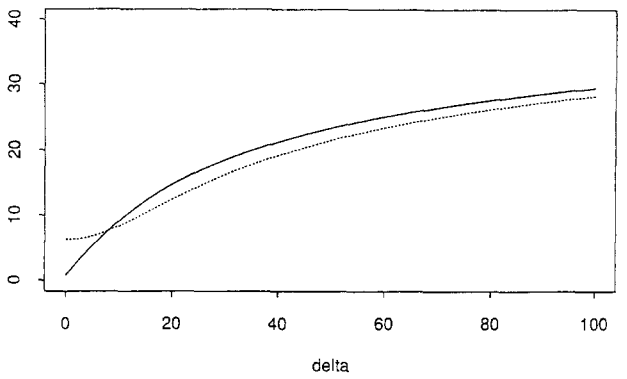

Fig. 2. Case $n = 40$.

difference to n is small, though the absolute difference is not small. It appears that the fine performance of the empirical Bayes method in this example comes from the decomposition property of the Bayes estimator.

Acknowledgements

The author is grateful to the referees for their constructive comments, which lead to substantial improvement in presentation.

References

Amari, S-I. (1985). *Differential-Geometrical Methods in Statistics*, Springer, Berlin.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators, *J. Roy. Statist. Soc. Ser. B*, **32**, 283–301.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*, Wiley, New York.

Berger, J. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters, *Ann. Statist.*, **8**, 545–571.

Blaesild, P. and Jensen, J. L. (1985). Saddlepoint formulas for reproductive exponential models, *Scand. J. Statist.*, **12**, 193–202.

Brandwein, A. C. and Strawderman, W. (1990). Stein estimation: The spherically symmetric case, *Statist. Sci.*, **5**, 356–369.

Brown, L. D. (1979). A heuristic method for determining admissibility of estimators—with applications, *Ann. Statist.*, **7**, 960–994.

Cassella, G. (1985). An introduction to empirical Bayes data analysis, *Amer. Statist.*, **39**, 83–87.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion), *J. Roy. Statist. Soc. Ser. B*, **49**, 1–39.

Dey, D. K., Ghosh, M. and Srinivasan, M. (1987). Simultaneous estimation of parameters under entropy loss, *J. Statist. Plann. Inference*, **15**, 347–363.

Efron, B. and Morris, C. (1973). Stein's estimation rules and its competitors—An empirical Bayes approach, *J. Amer. Statist. Assoc.*, **68**, 117–130.

Fisher, R. A. (1935). The logic of inductive inference, *J. Roy. Statist. Soc.*, **98**, 39–54.

Godambe, V. P. (1984). On ancillarity and Fisher information in the presence of a nuisance parameter, *Biometrika*, **71**, 626–629.

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proc. Fourth Berkley Symp. on Math. Statist. Prob.*, Vol. 1, 361–380, Univ. of California Press, Berkeley.

Jorgensen, B. (1987). Exponential dispersion model (with discussion), *J. Roy. Statist. Soc. Ser. B*, **49**, 127–162.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79–86.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions, *Ann. Statist.*, **10**, 65–80.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion), *J. Amer. Statist. Assoc.*, **78**, 47–65.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear model (with discussion), *J. Roy. Statist. Soc. Ser. A*, **34**, 370–384.

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

Saville, D. J. and Wood, G. R. (1991). *Statistical Methods*: *The Geometric Approach*, Springer, New York.

Simon, G. (1973). Additivity of information in exponential family probability laws, *J. Amer. Statist. Assoc.*, **68**, 478–482.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proc. Third Berkeley Symp. on Math. Statist. Prob.*, Vol. 1, 197–206, Univ. of California Press, Berkeley.

Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution (with discussion), *J. Roy. Statist. Soc. Ser. B*, **24**, 265–296.

Yanagimoto, T. (1987). A notion of an obstructive residual likelihood, *Ann. Inst. Statist. Math.*, **39**, 247–261.

Yanagimoto, T. (1991). Estimating a model through the conditional MLE, *Ann. Inst. Statist. Math.*, **43**, 735–746.

Yanagimoto, T. and Anraku, K. (1989). Possible superiority of the conditional MLE over the unconditional MLE, *Ann. Inst. Statist. Math.*, **41**, 269–278.

Yanagimoto, T. and Yanagimoto, M. (1987). The use of the marginal likelihood for a diagnostic test for the goodness of fit of the simple regression model, *Technometrics*, **29**, 95–101.