DUALITY INDUCED FROM CONJUGACY IN THE CURVED EXPONENTIAL FAMILY

Toshio Ohnishi* and Takemi Yanagimoto**

A class of curved exponential families whose likelihood function admits the conjugate analysis is derived, and its duality is explored. We show that conjugacy yields the existence of sufficient statistics as well as duality. Extended versions of the mean and the canonical parameters can be defined, which shed a new light on duality and the conjugate analysis in the exponential family. As a result, an essential reason is revealed as to why a common prior density can be conjugate for different sampling densities, as in the case of a gamma prior density which is conjugate for the Poisson and the gamma sampling densities. The least information property of the conjugate analysis is explained, which is compatible with the minimax property of the generalized linear model. We also derive dual Pythagorean relationships with respect to posterior risks to show the optimality of the Bayes estimator.

Key words and phrases: Closure under sampling, duality, least information, Legendre transformation, linearity, proper dispersion model, Pythagorean relationship, standardized posterior mode, sufficient statistic.

1. Introduction

Likelihood functions with excellent properties are of great importance in statistical inference. A typical example is the generalized linear model (GLM), which has been used in a variety of fields. Its inference is based on an exponential family distribution. Duality is a desirable property of the exponential family, which was described elegantly in terms of the canonical and the mean parameters by Barndorff-Nielsen (1978a) and Amari and Nagaoka (2000).

The subject matter of this paper is the curved exponential family. It possesses various analytical conveniences so that it is interesting from viewpoints of both theory and application. We attempt here to derive a class of curved exponential families whose likelihood functions have excellent properties for statistical inference. A key notion is a condition that the sampling density admits a conjugate prior density. This approach was employed by Ohnishi and Yanagimoto (2007), who characterized such location families admitting the conjugate analysis. The derived families belong to the curved exponential families.

The conjugate analysis provides us with a basic scheme of Bayesian inference even in the recent rapid development in computational skills. It has attracted the interest of many researchers including Morris (1983), Consonni and Veronese (1992, 2001), Gutiérrez-Peña (1997), Gutiérrez-Peña and Smith (1997), Ibrahim

Received July 24, 2009. Revised January 6, 2010. Accepted April 21, 2010.

*Faculty of Economics, Kyushu University, 6-19-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.
Email: ohnishi@en.kyushu-u.ac.jp

**Department of Industrial and Systems Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. Email: yanagmt@indsys.chuo-u.ac.jp

and Chen (2000), Yanagimoto and Ohnishi (2005a, b) and Ohnishi and Yanagimoto (2007). An advantage of the conjugate analysis is that its simple inferential scheme visualizes an essence of Bayesian inference. Another advantage is a minimax property of the conjugate prior density, which was shown by Morris (1983) and Consonni and Veronese (1992).

This paper has the following two aims. One is to reveal duality which is induced from the conjugacy condition in that class of curved exponential families. This duality, an extension of the one observed in the exponential family, casts a new light on the exponential family and its conjugate analysis. Actually, extended versions of the mean and the canonical parameters are defined. For instance, take a sampling density in an exponential family

$$(1.1) \quad p(x; \eta) = \exp\{\eta x - \psi(\eta)\}a(x).$$

In addition to the ordinary set (μ, η) of the mean and the canonical parameters, we will show that $(1/\mu, -\psi(\eta))$ can be regarded as the extended versions of the mean and the canonical parameters. This extension makes it clear why a gamma prior is conjugate both for the gamma and the Poisson sampling densities.

We also learn that the existence of sufficient statistics and conjugacy are two sides of the same coin. This is one of the most interesting results obtained by employing a discussion of the curved exponential families. In the case of the exponential families the following two facts are so obvious that it seems tricky to see the relationship between them: (1) The conjugate analysis is possible; (2) there exists a sufficient statistic.

The other aim of this paper is to show that duality also holds in the conjugate analysis of the class of curved exponential families. We derive the dual Pythagorean relationships with respect to posterior risks, each of which makes it clear how the Bayes estimator dominates others. A Pythagorean relationship is very useful in showing dominance or superiority, and it has been used by such authors as Stein (1981), Ohnishi and Yanagimoto (2003, 2007) and Yanagimoto and Ohnishi (2005a, b, 2009).

An implication of this paper is a potential relationship between GLM and the conjugate analysis. They have the least information property in common. As Tsubaki (1988) pointed out, the minimax property of the inference in the GLM is a direct consequence of its least information property. We will show the conjugate prior density is the least informative, which seems to be closely related to the minimax property of the conjugate prior density stated above. The least information property implies a type of superiority of the conjugate analysis over non-conjugate ones.

2. Extended mean and canonical parameters

In order to develop a general and unified discussion, we investigate the following class of curved exponential families

$$(2.1) \quad p(\mathbf{x}; \boldsymbol{\mu}, \tau_0) = \exp\{-\tau_0 d(\mathbf{x}, \boldsymbol{\mu})\} a(\mathbf{x}; \tau_0),$$

$$d(\mathbf{s}, \mathbf{t}) = \sum_{j=1}^{p+1} h_j(\mathbf{s}) \{f_j(\mathbf{t}) - f_j(\mathbf{s})\},$$

where $\mathbf{x}$ and $\boldsymbol{\mu}$ are p -dimensional, $\tau_0 > 0$ is a dispersion parameter, $f_k(t)$'s and $h_k(t)$'s are appropriate functions, and $a(\mathbf{x}; \tau_0)$ is a supporting measure. Since the two parameters $\boldsymbol{\mu}$ and τ_0 are orthogonal with each other, we will focus on the Bayesian estimation of $\boldsymbol{\mu}$ and assume that τ_0 is known.

The following three regularity conditions on the density (2.1) will be useful for our later discussions.

(C.1) $h_1(\mathbf{t}), \dots, h_{p+1}(\mathbf{t})$ are linearly independent.

(C.2) $1, f_1(\mathbf{t}), \dots, f_{p+1}(\mathbf{t})$ are linearly independent.

(C.3) $d(\mathbf{s}, \mathbf{t}) \geq 0$ and $d(\mathbf{s}, \mathbf{t}) = 0$ if and only if $\mathbf{s} = \mathbf{t}$.

Here the functions $h_1(\mathbf{t}), \dots, h_{p+1}(\mathbf{t})$ are said to be linearly independent if the following holds: $\sum_{j=1}^{p+1} a_j h_j(\mathbf{t}) = 0$ for any $\mathbf{t}$ in an open set if and only if $a_1 = \dots = a_{p+1} = 0$. A difference between (C.1) and (C.2) is that one of the functions $h_j(\mathbf{t})$'s can be identically constant, but that is not true for $f_j(\mathbf{t})$'s. We also assume an appropriate smoothness condition.

The class (2.1) covers the exponential family case. In fact, if we set $h_{p+1}(\mathbf{x}) = 1$ and $\tau_0 = 1$, then the density is written as $p(\mathbf{x}; \boldsymbol{\mu}, 1) = \exp\{-\sum_{j=1}^p h_j(\mathbf{x}) f_j(\boldsymbol{\mu}) - f_{p+1}(\boldsymbol{\mu})\} \tilde{a}(\mathbf{x})$ with $\tilde{a}(\mathbf{x}) = \exp\{\sum_{j=1}^p h_j(\mathbf{x}) f_j(\mathbf{x}) + f_{p+1}(\mathbf{x})\} a(\mathbf{x}; 1)$, which is in an exponential family. Also, the class (2.1) is derived from a $(p+1)$ -parameter exponential-dispersion family by imposing a condition. Consider the following density

$$\exp \left[\tau_0 \left\{ \sum_{j=1}^{p+1} \eta_j h_j(\mathbf{x}) - \kappa(\eta_1, \dots, \eta_{p+1}) \right\} \right] c(\mathbf{x}; \tau_0)$$

where τ_0 is a known dispersion parameter, $\eta_1, \dots, \eta_{p+1}$ are the canonical parameters, $\kappa(\eta_1, \dots, \eta_{p+1})$ is the cumulant function, and $c(\mathbf{x}; \tau_0)$ is the supporting measure. If we happen to find functions $f_1(\boldsymbol{\mu}), \dots, f_{p+1}(\boldsymbol{\mu})$ such that $\kappa(-f_1(\boldsymbol{\mu}), \dots, -f_{p+1}(\boldsymbol{\mu}))$ is constant, then we obtain (2.1) by replacing η_j with $-f_j(\boldsymbol{\mu})$ for $j = 1, \dots, p+1$.

We explain the reason why we choose the density (2.1). Let us consider the following function

$$d_K(\mathbf{s}, \mathbf{t}) = \sum_{j=1}^K h_j(\mathbf{s}) \{f_j(\mathbf{t}) - f_j(\mathbf{s})\}$$

in place of $d(\mathbf{s}, \mathbf{t})$ in (2.1). Here $h_j(\mathbf{t})$'s and $f_j(\mathbf{t})$'s are assumed to satisfy the conditions (C.1) and (C.2) with $p+1$ being replaced with K .

Combining the following two propositions, we have $K = p + 1$. Their proofs are given in the Appendix.

PROPOSITION 2.1. *If $d_K(\mathbf{s}, \mathbf{t})$ satisfies the condition (C.3), then $K \geq p + 1$.*

PROPOSITION 2.2. *Suppose that there exists a function $\rho(\mathbf{s}_1, \mathbf{s}_2, \delta)$ such that the identity with respect to $\mathbf{t}$*

$$d_K(\mathbf{s}_1, \mathbf{t}) + \delta d_K(\mathbf{s}_2, \mathbf{t}) = \rho(\mathbf{s}_1, \mathbf{s}_2, \delta) d_K(\hat{\mathbf{t}}_M, \mathbf{t}) + d_K(\mathbf{s}_1, \hat{\mathbf{t}}_M) + \delta d_K(\mathbf{s}_2, \hat{\mathbf{t}}_M)$$

holds for any $\mathbf{s}_1, \mathbf{s}_2$ and δ , where $\hat{\mathbf{t}}_M$ is the solution to the equation $(\partial/\partial \mathbf{t})\{d_K(\mathbf{s}_1, \mathbf{t}) + \delta d_K(\mathbf{s}_2, \mathbf{t})\} = \mathbf{0}$. Then, $K \leq p + 1$.

The condition in Proposition 2.2 implies the existence of sufficient statistics. Suppose that $d(\mathbf{s}, \mathbf{t})$ satisfies the condition. Then, a mathematical induction gives that there exists $\rho_M = \rho_M(\mathbf{x}_1, \dots, \mathbf{x}_n)$ such that

$$\sum_{i=1}^n d(\mathbf{x}_i, \boldsymbol{\mu}) = \rho_M d(\hat{\boldsymbol{\mu}}_M, \boldsymbol{\mu}) + \sum_{i=1}^n d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_M),$$

where $\hat{\boldsymbol{\mu}}_M$ is the maximum likelihood estimator (MLE), which is a solution to the equation $(\partial/\partial \boldsymbol{\mu}) \sum d(\mathbf{x}_i, \boldsymbol{\mu}) = \mathbf{0}$. This equality implies that $\hat{\boldsymbol{\mu}}_M$ and ρ_M are sufficient statistics. In the case of the exponential family, ρ_M becomes the data size n . Note that Bernardo and Smith (2000, pp. 265–266) gave a definition of the conjugate prior on the basis of the sufficient statistic.

Now, we define extended versions of the mean and the canonical parameters in order to develop discussions similar to those in the exponential family case. Let $F_{p,p}(\mathbf{t})$ denote the $p \times p$ matrix whose (i, j) th component is $\partial f_j(\mathbf{t})/\partial t_i$ ($1 \leq i, j \leq p$). In addition to (C.1)–(C.3) we assume the following non-singularity condition:

$$(C.4) \quad \det F_{p,p}(\mathbf{t}) \neq 0 \text{ for any } \mathbf{t}.$$

An examination of this condition will be discussed in the final section.

We show that $h_{p+1}(\mathbf{s}) \neq 0$ for any $\mathbf{s}$. Suppose that $h_{p+1}(\mathbf{s}_0) = 0$ for some $\mathbf{s}_0$ and write $\mathbf{h}(\mathbf{s}) = (h_1(\mathbf{s}), \dots, h_p(\mathbf{s}))^T$. The condition (C.3) implies that

$$(2.2) \quad \left. \frac{\partial}{\partial \mathbf{t}} d(\mathbf{s}, \mathbf{t}) \right|_{\mathbf{t}=\mathbf{s}} = \mathbf{0} \quad \text{for any } \mathbf{s},$$

which can be rewritten as

$$F_{p,p}(\mathbf{s}) \mathbf{h}(\mathbf{s}) = -h_{p+1}(\mathbf{s}) \frac{\partial f_{p+1}(\mathbf{s})}{\partial \mathbf{s}}.$$

This set of linear equations, together with (C.4), yields that $\mathbf{h}(\mathbf{s}_0) = \mathbf{0}$ and therefore that $d(\mathbf{s}_0, \mathbf{t}) = 0$ for any $\mathbf{t}$, which contradicts (C.3). Thus, we assume without loss of generality that

$$(C.5) \quad h_{p+1}(\mathbf{t}) > 0 \text{ for any } \mathbf{t}.$$

We introduce a new parameter vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_p)^T$ as

$$(2.3) \quad \eta_j = -f_j(\boldsymbol{\mu}) \quad (j = 1, \dots, p).$$

It follows from the inverse function theorem that (C.4) guarantees the one-to-one correspondence between $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$. We may call $\boldsymbol{\eta}$ the *extended canonical parameter*. In the exponential family case where $h_{p+1}(\mathbf{t}) = 1$ and $\tau_0 = 1$, the parameter vector $\boldsymbol{\eta}$ is the canonical one.

LEMMA 2.1. *Regard $f_{p+1}(\boldsymbol{\mu})$ as a function of $\boldsymbol{\eta}$ and set*

$$(2.4) \quad \psi(\boldsymbol{\eta}) = f_{p+1}(\boldsymbol{\mu}).$$

The function $\psi(\boldsymbol{\eta})$ is convex.

PROOF. See the Appendix. $\square$

The Legendre transformation yields the parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ and the convex function $\phi(\boldsymbol{\theta})$ which are dual to $\boldsymbol{\eta}$ and $\psi(\boldsymbol{\eta})$, respectively. As is given by (A.4) in the Appendix, we have

$$(2.5) \quad \theta_j = \frac{h_j(\boldsymbol{\mu})}{h_{p+1}(\boldsymbol{\mu})}.$$

The convex function $\phi(\boldsymbol{\theta})$ is expressed as $\boldsymbol{\theta}^T \boldsymbol{\eta} - \psi(\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ is the parameter value corresponding to $\boldsymbol{\theta}$. Note that the convexity of $\psi(\boldsymbol{\eta})$ guarantees the one-to-one correspondence between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$.

The following lemma clarifies the meaning of $\boldsymbol{\theta}$. We may call $\boldsymbol{\theta}$ the *extended mean parameter*. In this paper we will mean by $E[f(x) \mid p(x)]$, the expectation of a function $f(x)$ with respect to a density $p(x)$.

LEMMA 2.2. *It holds for $j = 1, \dots, p$ that*

$$(2.6) \quad E[h_j(\mathbf{x}) - \theta_j h_{p+1}(\mathbf{x}) \mid p(\mathbf{x}; \boldsymbol{\mu}, \tau_0)] = 0.$$

PROOF. See the Appendix. $\square$

We see that the quantity $L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ defined by

$$(2.7) \quad L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \phi(\boldsymbol{\theta}_1) + \psi(\boldsymbol{\eta}_2) - \boldsymbol{\theta}_1^T \boldsymbol{\eta}_2$$

is positive where $\boldsymbol{\mu}_i$, $\boldsymbol{\eta}_i$ and $\boldsymbol{\theta}_i$ are equivalent to one another ($i = 1, 2$). It seems to be natural to adopt $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ or $L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ as a loss function. It should be noted that the following identity

$$(2.8) \quad L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_3) - L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) - L(\boldsymbol{\mu}_2, \boldsymbol{\mu}_3) = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3)$$

holds, which will play a key role in Sections 4 and 5.

An interesting result is found in the relation among $d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$, $L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and the Kullback-Leibler separator. A calculation using the formulas (2.3)–(2.7) gives

$$(2.9) \quad d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = h_{p+1}(\boldsymbol{\mu}_1) L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2).$$

Also, the Kullback-Leibler separator from $p(\mathbf{x}; \boldsymbol{\mu}_1, \tau_0)$ to $p(\mathbf{x}; \boldsymbol{\mu}_2, \tau_0)$ is calculated as

$$(2.10) \quad \text{KL}_{\tau_0}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \tau_0 \mathbb{E}[h_{p+1}(\mathbf{x}) \mid p(\mathbf{x}; \boldsymbol{\mu}_1, \tau_0)] L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2).$$

The following examples give calculations of the extended mean and the extended canonical parameters.

Example 2.1. In the exponential family (1.1) let μ be the mean parameter and $\phi(\mu)$ the convex function dual to the cumulant function $\psi(\eta)$. Noting that $\eta = \phi'(\mu)$ and $\phi(x) = x\phi'(x) - \psi(\phi'(x))$, we obtain another expression of the density as

$$p(x; \mu) = \exp[-x\{-\phi'(\mu) + \phi'(x)\} - \{\psi(\phi'(\mu)) - \psi(\phi'(x))\}] e^{\phi(x)} a(x).$$

If we set $f_1(\mu) = -\phi'(\mu)$, $f_2(\mu) = \psi(\phi'(\mu))$, $h_1(x) = x$ and $h_2(x) = 1$, then we obtain the mean and the canonical parameters in the ordinary sense.

When the sampling density is defined on $\mathbb{R}^+$, another choice is possible. The pair $(1/\mu, -\psi(\eta))$ of the extended mean and canonical parameters is obtained by setting $f_1(\mu) = \psi(\phi'(\mu))$, $f_2(\mu) = -\phi'(\mu)$, $h_1(x) = 1$ and $h_2(x) = x$. Let us adopt this parameterization in the following gamma density

$$p(x; \mu, \tau_0) = \frac{\tau_0^{\tau_0}}{\Gamma(\tau_0)} \frac{x^{\tau_0-1}}{\mu^{\tau_0}} \exp\left(-\tau_0 \frac{x}{\mu}\right),$$

where the dispersion parameter τ_0 is known. The extended mean and canonical parameters are given as $\theta = 1/\mu$ and $\eta = -\log \mu$, respectively. Also, the derived dual convex functions are

$$\tilde{\psi}(\eta) = e^\eta \quad \text{and} \quad \tilde{\phi}(\theta) = \theta \log \theta - \theta,$$

which are the same as those in the Poisson distribution under the ordinary parameterization. We see from (2.7) that these two sampling distributions have the common function $L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. This is directly related to the fact that a gamma prior density is conjugate for both of the sampling distributions. As will be seen in Section 3, $L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ essentially determines the conjugate prior density.

Example 2.2. We discuss the hyperbola distribution having the density

$$(2.11) \quad p(x - \mu; \tau_0) = \frac{1}{2K_0(\tau_0)} \exp\{-\tau_0 \cosh(x - \mu)\},$$

where $K_0(\tau_0)$ is the modified Bessel function of the third kind. The addition formula for the hyperbolic cosine function gives

$$\cosh(x - \mu) - 1 = \sinh x(-\sinh \mu + \sinh x) + \cosh x(\cosh \mu - \cosh x).$$

The regularity conditions (C.4) and (C.5) are satisfied if we set $f_1(\mu) = -\sinh \mu$, $f_2(\mu) = \cosh \mu$, $h_1(x) = \sinh x$ and $h_2(x) = \cosh x$. The extended mean and canonical parameters are given by $\theta = \tanh \mu$ and $\eta = \sinh \mu$, respectively. This sampling density allows us to conduct the conjugate analysis in the same manner as the von Mises density does. A close relationship between this density and the von Mises one was pointed out by Barndorff-Nielsen (1978b) and Jensen (1981).

The hyperbola distribution (2.11) is a special case of the density function

$$(2.12) \quad \begin{aligned} p(x - \mu; \tau_0) &\propto \exp\{-\tau_0 d(x - \mu)\}, \\ d(t) &= \frac{1}{\kappa + \gamma} \left(\frac{e^{\kappa t} - 1}{\kappa} + \frac{e^{-\gamma t} - 1}{\gamma} \right), \end{aligned}$$

where $\kappa > 0$ and $\gamma > 0$ are known constants. This density was characterized as a class of location families admitting the conjugate analysis by Ohnishi and Yanagimoto (2007). They showed that

$$d(x - \mu) = \tilde{d}(x)\{d(-\mu) - d(-x)\} + d'(x)\{d'(-\mu) - d'(-x)\},$$

where $\tilde{d}(t) = (\gamma e^{\kappa t} + \kappa e^{-\gamma t})/(\kappa + \gamma)$. Set $f_1(\mu) = d'(-\mu)$, $f_2(\mu) = d(-\mu)$, $h_1(x) = d'(x)$ and $h_2(x) = \tilde{d}(x)$. Then (C.4) and (C.5) are satisfied, and we have $\theta = (e^{\kappa\mu} - e^{-\gamma\mu})/(\gamma e^{\kappa\mu} + \kappa e^{-\gamma\mu})$ and $\eta = (e^{\gamma\mu} - e^{-\kappa\mu})/(\kappa + \gamma)$.

The density (2.12) provides us with another example in which a common prior density is conjugate for different sampling densities. Ohnishi and Dunn (2009) showed that a prior density which has the same form as the likelihood in (2.12) is conjugate for both the sampling density (2.12) and a GLM based on the Tweedie distribution. For the Tweedie distribution, see Jørgensen (1997, Chapter 4).

3. Conjugacy with the least information property

In this paper we adopt closure under sampling as the definition of conjugacy, following the original definition by Raiffa and Schlaifer (1961, pp. 43–57). It is known that this definition produces ambiguity. In the exponential family (1.1) the prior density $\pi(\eta; \mathbf{m}, \delta) \propto \exp[\delta\{m\eta - \psi(\eta)\}]b(\eta)$ is closed under sampling, and we cannot specify the type of the supporting measure $b(\eta)$ by closure under sampling. A reason why such a rather ambiguous definition is adopted is that closure under sampling in itself implies a certain optimum property. Another reason is that invariant estimation regardless of the choice of $b(\eta)$ is possible.

We assume on the extended canonical parameter $\boldsymbol{\eta}$ a prior density

$$(3.1) \quad \pi(\boldsymbol{\eta}; \mathbf{m}, \delta) = \exp\{-\delta d(\mathbf{m}, \boldsymbol{\mu}) + K(\mathbf{m}, \delta)\}b(\boldsymbol{\eta}),$$

where $b(\boldsymbol{\eta})$ is a non-negative function and $\exp\{K(\boldsymbol{m}, \delta)\}$ is the normalizing constant.

In this paper we will employ the following estimator:

$$(3.2) \quad \hat{\boldsymbol{\mu}}_B = \arg \min_{\boldsymbol{\mu}} \{\tau_0 d(\boldsymbol{x}, \boldsymbol{\mu}) + \delta d(\boldsymbol{m}, \boldsymbol{\mu})\}.$$

This estimator is a modified version of the posterior mode, which was proposed as the *standardized posterior mode* by Yanagimoto and Ohnishi (2005b). Since we discard the Jacobian factor $b(\boldsymbol{\eta})$ in maximizing the posterior density, the estimation procedure is invariant with respect to a parameter transformation.

Setting $\boldsymbol{\theta}(\boldsymbol{\mu})$ as the extended mean parameter corresponding to $\boldsymbol{\mu}$, we show that an expression of $\hat{\boldsymbol{\mu}}_B$ is given by

$$(3.3) \quad \boldsymbol{\theta}(\hat{\boldsymbol{\mu}}_B) = \frac{\tau_0 h_{p+1}(\boldsymbol{x}) \boldsymbol{\theta}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m}) \boldsymbol{\theta}(\boldsymbol{m})}{\tau_0 h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})}.$$

It follows from (2.7) and (2.9) that

$$(3.4) \quad \begin{aligned} & \tau_0 d(\boldsymbol{x}, \boldsymbol{\mu}) + \delta d(\boldsymbol{m}, \boldsymbol{\mu}) \\ &= \{\tau_0 h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})\} \psi(\boldsymbol{\eta}) \\ & \quad - \{\tau_0 h_{p+1}(\boldsymbol{x}) \boldsymbol{\theta}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m}) \boldsymbol{\theta}(\boldsymbol{m})\}^T \boldsymbol{\eta} \\ & \quad + \tau_0 h_{p+1}(\boldsymbol{x}) \phi(\boldsymbol{\theta}(\boldsymbol{x})) + \delta h_{p+1}(\boldsymbol{m}) \phi(\boldsymbol{\theta}(\boldsymbol{m})). \end{aligned}$$

Computing the derivative of the right-hand side of (3.4) with respect to $\boldsymbol{\eta}$ and using the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ through the Legendre transform, we see that (3.4) is minimized at $\boldsymbol{\theta} = \{\tau_0 h_{p+1}(\boldsymbol{x}) \boldsymbol{\theta}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m}) \boldsymbol{\theta}(\boldsymbol{m})\} / \{\tau_0 h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})\}$. It should be noted that the assumption (C.5) guarantees $\tau_0 h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m}) > 0$.

Recalling (2.5), we obtain the component-wise expression

$$(3.5) \quad \frac{h_j(\hat{\boldsymbol{\mu}}_B)}{h_{p+1}(\hat{\boldsymbol{\mu}}_B)} = \frac{\tau_0 h_j(\boldsymbol{x}) + \delta h_j(\boldsymbol{m})}{\tau_0 h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})} \quad (1 \leq j \leq p).$$

We can see a type of linearity of the standardized posterior mode. In the exponential family (1.1) Diaconis and Ylvisaker (1979) characterized the choice $b(\boldsymbol{\eta}) = 1$ by linearity of the posterior mean of the mean parameter and defined conjugacy by this linearity. It is interesting to compare this linearity holding for any $b(\boldsymbol{\eta})$, with the posterior linearity in Diaconis and Ylvisaker (1979).

PROPOSITION 3.1. *The posterior density is expressed as $\pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)$ where $\hat{\boldsymbol{\mu}}_B$ is the standardized posterior mode (3.2) and*

$$(3.6) \quad \rho_B = \rho_B(\boldsymbol{x}, \boldsymbol{m}, \tau_0, \delta) = \frac{\tau_0 h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})}{h_{p+1}(\hat{\boldsymbol{\mu}}_B)}.$$

Therefore the prior density (3.1) is conjugate.

PROOF. The posterior density is proportional to $\exp\{-\tau_0 d(\mathbf{x}, \boldsymbol{\mu}) - \delta d(\mathbf{m}, \boldsymbol{\mu})\} b(\boldsymbol{\eta})$. The expression of $d(\mathbf{s}, \mathbf{t})$ in (2.1) gives

$$(3.7) \quad \begin{aligned} & \tau_0 d(\mathbf{x}, \boldsymbol{\mu}) + \delta d(\mathbf{m}, \boldsymbol{\mu}) - \tau_0 d(\mathbf{x}, \hat{\boldsymbol{\mu}}_B) - \delta d(\mathbf{m}, \hat{\boldsymbol{\mu}}_B) \\ &= \sum_{j=1}^{p+1} \{\tau_0 h_j(\mathbf{x}) + \delta h_j(\mathbf{m})\} \{f_j(\boldsymbol{\mu}) - f_j(\hat{\boldsymbol{\mu}}_B)\}. \end{aligned}$$

It follows from (3.5) and (3.6) that $\tau_0 h_j(\mathbf{x}) + \delta h_j(\mathbf{m}) = \rho_B h_j(\hat{\boldsymbol{\mu}}_B)$ for $j = 1, \dots, p$. Thus, we see that the left-hand side of (3.7) reduces to $\rho_B d(\hat{\boldsymbol{\mu}}_B, \boldsymbol{\mu})$, which completes the proof. $\square$

Next, we discuss the least information property of the conjugate prior density in the following sense. Let $\pi(\boldsymbol{\eta})$ denote an arbitrary prior density, and write the corresponding posterior density as $\pi(\boldsymbol{\eta} \mid \mathbf{x})$ for a given $\mathbf{x}$. We regard the functional

$$G[\pi(\boldsymbol{\eta})] = \text{KL}(\pi(\boldsymbol{\eta} \mid \mathbf{x}), \pi(\boldsymbol{\eta}; \mathbf{x}, \tau_0))$$

as an amount of information contained in $\pi(\boldsymbol{\eta})$. This may be considered to be a difference between the posterior density $\pi(\boldsymbol{\eta} \mid \mathbf{x})$ and the sampling likelihood $p(\mathbf{x}; \boldsymbol{\mu}, \tau_0)$, or an amount of change in the likelihood caused by $\pi(\boldsymbol{\eta})$. Here we may look upon $p(\mathbf{x}; \boldsymbol{\mu}, \tau_0)$ as the density $\pi(\boldsymbol{\eta}; \mathbf{x}, \tau_0)$. Recall that $b(\boldsymbol{\eta})$ in (3.1) is discarded when deriving the standardized posterior mode (3.2).

Here we add two comments on the above interpretation. First, the derivation of the posterior density in the conjugate analysis of interest is expressed symbolically as

$$(\mathbf{x}, \tau_0) \quad \text{and} \quad (\mathbf{m}, \delta) \Rightarrow (\hat{\boldsymbol{\mu}}_B, \rho_B).$$

The formulas (3.3) and (3.6) show that both $(\mathbf{x}, \tau_0)$ and $(\mathbf{m}, \delta)$ are treated equally. This is a reason why we look upon $p(\mathbf{x}; \boldsymbol{\mu}, \tau_0)$ as the density $\pi(\boldsymbol{\eta}; \mathbf{x}, \tau_0)$.

Secondly, our idea of minimizing the Kullback-Leibler separator from the posterior density to the sampling likelihood seems to be similar to the idea found in the definition of the reference prior proposed by Bernardo (1979). The expected value of the Kullback-Leibler separator from the posterior density to the prior density with respect to the marginal density is called the missing information, and Bernardo and Smith (2000, Section 5.4.2) state that the reference prior is defined to be that prior which maximizes the missing information.

We consider the family $\mathcal{P}(\mathbf{x}, \mathbf{m}, \tau_0, \delta)$ of prior densities satisfying

$$(3.8) \quad \mathbb{E}[(\boldsymbol{\eta}^T, \psi(\boldsymbol{\eta})) \mid \pi(\boldsymbol{\eta} \mid \mathbf{x})] = \mathbb{E}[(\boldsymbol{\eta}^T, \psi(\boldsymbol{\eta})) \mid \pi(\boldsymbol{\eta}, \hat{\boldsymbol{\mu}}_B, \rho_B)].$$

Since $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \phi(\hat{\boldsymbol{\theta}}) + \psi(\boldsymbol{\eta}) - \hat{\boldsymbol{\theta}}^T \boldsymbol{\eta}$, it holds for any estimate $\hat{\boldsymbol{\mu}}$ that

$$\mathbb{E}[L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \mid \pi(\boldsymbol{\eta} \mid \mathbf{x})] = \mathbb{E}[L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \mid \pi(\boldsymbol{\eta}, \hat{\boldsymbol{\mu}}_B, \rho_B)].$$

Figure 1. The Pythagorean relationship holding for the conjugate prior.

Thus, it is reasonable to compare the amount of information contained among the prior densities in $\mathcal{P}(\boldsymbol{x}, \boldsymbol{m}, \tau_0, \delta)$.

The following Proposition gives a Pythagorean relationship holding for the conjugate prior density. See Figure 1.

PROPOSITION 3.2. *Let $\pi(\boldsymbol{\eta})$ be any prior density in $\mathcal{P}(\boldsymbol{x}, \boldsymbol{m}, \tau_0, \delta)$ defined by (3.8), and write the corresponding posterior density as $\pi(\boldsymbol{\eta} | \boldsymbol{x})$. Then, the following Pythagorean relationship*

$$(3.9) \quad \text{KL}(\pi(\boldsymbol{\eta} | \boldsymbol{x}), \pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)) = \text{KL}(\pi(\boldsymbol{\eta} | \boldsymbol{x}), \pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)) \\ + \text{KL}(\pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B), \pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1))$$

holds for any $\boldsymbol{m}_1$ and $\delta_1 > 0$.

PROOF. Note that

$$(3.10) \quad \text{KL}(\pi(\boldsymbol{\eta} | \boldsymbol{x}), \pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)) - \text{KL}(\pi(\boldsymbol{\eta} | \boldsymbol{x}), \pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)) \\ = \text{E} \left[\log \frac{\pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)}{\pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)} \middle| \pi(\boldsymbol{\eta} | \boldsymbol{x}) \right].$$

A calculation gives

$$\log \frac{\pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)}{\pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)} = \boldsymbol{a}_1^T \boldsymbol{\eta} + a_2 \psi(\boldsymbol{\eta}) + a_3,$$

where $\boldsymbol{a}_1$, a_2 and a_3 are independent of $\boldsymbol{\eta}$. They are explicitly represented as

$$\begin{aligned} \boldsymbol{a}_1 &= \rho_B h_{p+1}(\hat{\boldsymbol{\mu}}_B) \boldsymbol{\theta}(\hat{\boldsymbol{\mu}}_B) - \delta_1 h_{p+1}(\boldsymbol{m}_1) \boldsymbol{\theta}(\boldsymbol{m}_1), \\ a_2 &= \delta_1 h_{p+1}(\boldsymbol{m}_1) - \rho_B h_{p+1}(\hat{\boldsymbol{\mu}}_B), \\ a_3 &= \delta_1 h_{p+1}(\boldsymbol{m}_1) \phi(\boldsymbol{\theta}(\boldsymbol{m}_1)) - \rho_B h_{p+1}(\hat{\boldsymbol{\mu}}_B) \phi(\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}_B)) - K(\boldsymbol{m}_1, \delta_1) + K(\hat{\boldsymbol{\mu}}_B, \rho_B). \end{aligned}$$

Since the posterior density $\pi(\boldsymbol{\eta} | \boldsymbol{x})$ satisfies (3.8), the right-hand side of (3.10) is equal to $\text{KL}(\pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B), \pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1))$. $\square$

The least information property of the conjugate prior density is given in the following corollary.

COROLLARY 3.1. *The conjugate prior density (3.1) minimizes the functional*

$$G[\pi(\boldsymbol{\eta})] = \text{KL}(\pi(\boldsymbol{\eta} \mid \mathbf{x}), \pi(\boldsymbol{\eta}; \mathbf{x}, \tau_0))$$

among the family $\mathcal{P}(\mathbf{x}, \mathbf{m}, \tau_0, \delta)$ of prior densities defined by (3.8).

PROOF. Set $\mathbf{m}_1 = \mathbf{x}$ and $\delta_1 = \tau_0$ in Proposition 3.2, and we have

$$G[\pi(\boldsymbol{\eta})] = G[\pi(\boldsymbol{\eta}; \mathbf{m}, \delta)] + \text{KL}(\pi(\boldsymbol{\eta} \mid \mathbf{x}), \pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)),$$

which completes the proof. $\square$

We close this section by emphasizing a potential relationship between the conjugate analysis and the GLM. The GLM is based on the exponential family (1.1). Write the density as $p(x; \mu)$ in terms of the mean parameter μ . It is known to hold that $\log\{p(x; \hat{\mu}_M)/p(x; \mu)\} = \text{KL}(p(y; \hat{\mu}_M), p(y; \mu))$ where $\hat{\mu}_M = x$ is the MLE. This is formally rewritten as

$$\text{KL}(\delta(y - \hat{\mu}_M), p(y; \mu)) = \text{KL}(\delta(y - \hat{\mu}_M), p(y; \hat{\mu}_M)) + \text{KL}(p(y; \hat{\mu}_M), p(y; \mu)),$$

where $\delta(y - x)$ is the Dirac's delta function. Comparing with the Pythagorean relationship (3.9), we learn that a type of similarity lies between the conjugate analysis and the GLM. Conjugate priors for the GLM were studied by Chen and Ibrahim (2003) and Ohnishi and Dunn (2009).

4. A Pythagorean relationship

In this and the following sections we derive dual Pythagorean relationships. First, we pursue an optimality under the loss function $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})/h_{p+1}(\hat{\boldsymbol{\mu}})$, when there exists a non-negative function $b_0(\boldsymbol{\eta})$ such that

$$(4.1) \quad \frac{\partial}{\partial \mathbf{m}} \int \exp\{-\tilde{\delta}L(\mathbf{m}, \boldsymbol{\mu})\} b_0(\boldsymbol{\eta}) d\boldsymbol{\eta} = \mathbf{0} \quad \text{for any } \tilde{\delta}.$$

By setting the above integral as $\exp\{-K(\tilde{\delta})\}$, the prior density we assume is expressed as

$$(4.2) \quad \pi_0(\boldsymbol{\eta}; \mathbf{m}, \delta) = \exp\{-\delta d(\mathbf{m}, \boldsymbol{\mu}) + K(\delta h_{p+1}(\mathbf{m}))\} b_0(\boldsymbol{\eta}).$$

It should be noted that the normalizing constant depends on $\mathbf{m}$ and δ only through the product $\delta h_{p+1}(\mathbf{m})$. The density proportional to $\exp\{-\tilde{\delta}L(\mathbf{m}, \boldsymbol{\mu})\} b_0(\boldsymbol{\eta})$ belongs to the proper dispersion model introduced in Jørgensen (1997, p. 5).

PROPOSITION 4.1. *Under the assumption (4.1) it holds for any $\mathbf{m}$ and $\delta > 0$ that*

$$\mathbb{E}[\boldsymbol{\eta} - \boldsymbol{\eta}(\mathbf{m}) \mid \pi_0(\boldsymbol{\eta}; \mathbf{m}, \delta)] = \mathbf{0},$$

where $\boldsymbol{\eta}(\mathbf{m}) = -(f_1(\mathbf{m}), \dots, f_p(\mathbf{m}))^T$. Further, the posterior density corresponding to $\pi_0(\boldsymbol{\eta}; \mathbf{m}, \delta)$ satisfies

$$\mathbb{E}[\boldsymbol{\eta} - \boldsymbol{\eta}(\hat{\boldsymbol{\mu}}_B) \mid \pi_0(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)] = \mathbf{0}.$$

PROOF. Recall (2.7). Differentiating the integral in (4.1) with respect to $\boldsymbol{\theta}(\mathbf{m})$, we have

$$\int \{\boldsymbol{\eta}(\mathbf{m}) - \boldsymbol{\eta}\} \exp\{-\tilde{\delta}L(\mathbf{m}, \boldsymbol{\mu})\} b_0(\boldsymbol{\eta}) d\boldsymbol{\eta} = \mathbf{0}$$

for any $\mathbf{m}$ and $\tilde{\delta} > 0$. Setting $\tilde{\delta} = \delta h_{p+1}(\mathbf{m})$, we obtain the former part.

The latter part follows from the former part since Proposition 3.1 gives the corresponding posterior density as $\pi_0(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)$. $\square$

This proposition is an extension of Proposition 4.5 (ii) in Yanagimoto and Ohnishi (2005a), where the sampling density is restricted to be in the natural exponential family. They called it the ϵ -preservation property. This extension is realized by suitably introducing $\boldsymbol{\eta}$.

Example 4.1. We revisit the exponential family case in Example 2.1 to clarify implications of Proposition 4.1. In the former part the extended canonical parameter coincides with the ordinary canonical parameter η , and $h_2(m) = 1$. If the assumption (4.1) is satisfied, that is, the normalizing constant of the prior density depends only on δ , then Proposition 4.1 yields that the posterior mean of $\eta = \phi'(\mu)$ is $\phi'(\hat{\mu}_B)$ with $\hat{\mu}_B = (x + \delta m)/(1 + \delta)$. In the latter part the extended canonical parameter is $-\psi(\eta) = -\psi(\phi'(\mu))$, and $h_2(m) = m$. If the normalizing constant of the prior density is a function of δm , then the posterior mean of $\psi(\eta)$ is obtained as $\psi(\phi(\hat{\mu}_B))$. These results were given as Theorem 4.2 in Yanagimoto and Ohnishi (2005b).

Now, let us derive a Pythagorean relationship with respect to posterior risks, which is an extension of the result in Section 5 in Yanagimoto and Ohnishi (2005a).

PROPOSITION 4.2. *Under the assumption (4.1) the Pythagorean relationship*

$$(4.3) \quad \mathbb{E}[L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}}_B, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_B) \mid \pi_0(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)] = 0$$

holds for any estimator $\hat{\boldsymbol{\mu}}$. Thus, the standardized posterior mode $\hat{\boldsymbol{\mu}}_B$ is optimum under the loss $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$.

PROOF. It follows from the identity (2.8) that

$$(4.4) \quad L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}}_{\text{B}}, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{\text{B}}) = \{\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}) - \boldsymbol{\theta}(\hat{\boldsymbol{\mu}}_{\text{B}})\}^T \{\boldsymbol{\eta}(\hat{\boldsymbol{\mu}}_{\text{B}}) - \boldsymbol{\eta}\}.$$

Note that $\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}) - \boldsymbol{\theta}(\hat{\boldsymbol{\mu}}_{\text{B}})$ is constant in $\boldsymbol{\eta}$. Thus, the latter part of Proposition 4.1 yields the Pythagorean relationship (4.3). The optimum property of $\hat{\boldsymbol{\mu}}_{\text{B}}$ follows from this Pythagorean relationship. $\square$

An extended version of (4.3) is derived by modifying the loss function for an appropriate choice of $b(\boldsymbol{\eta})$. Suppose that there exist a positive function $I(\mathbf{m})$ and a non-negative function $b_1(\boldsymbol{\eta})$ such that

$$(4.5) \quad \frac{\partial}{\partial \mathbf{m}} \int \exp\{-\tilde{\delta}I(\mathbf{m})d(\mathbf{m}, \boldsymbol{\mu})\}b_1(\boldsymbol{\eta})d\boldsymbol{\eta} = \mathbf{0}.$$

Assumption (4.1) is obtained as a special case by setting $I(\mathbf{m}) = 1/h_{p+1}(\mathbf{m})$ and $b_1(\boldsymbol{\eta}) = b_0(\boldsymbol{\eta})$.

Set $\delta = \tilde{\delta}I(\mathbf{m})$ and assume the prior density $\pi_1(\boldsymbol{\eta}; \mathbf{m}, \delta) \propto \exp\{-\delta d(\mathbf{m}, \boldsymbol{\mu})\}b_1(\boldsymbol{\eta})$. Proposition 3.1 means that the corresponding posterior density is expressed as $\pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{\text{B}}, \rho_{\text{B}})$.

PROPOSITION 4.3. *Under the assumption (4.5) the following modified Pythagorean relationship*

$$(4.6) \quad \begin{aligned} & \mathbb{E}[I(\hat{\boldsymbol{\mu}})d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) - I(\hat{\boldsymbol{\mu}}_{\text{B}})d(\hat{\boldsymbol{\mu}}_{\text{B}}, \boldsymbol{\mu}) \mid \pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{\text{B}}, \rho_{\text{B}})] \\ &= \frac{I(\hat{\boldsymbol{\mu}}_{\text{B}})}{\rho_{\text{B}}} \text{KL}(\pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{\text{B}}, \rho_{\text{B}}), \pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}, \tilde{\rho}_{\text{B}})) \end{aligned}$$

holds for any estimator $\hat{\boldsymbol{\mu}}$, where $\tilde{\rho}_{\text{B}} = \rho_{\text{B}}I(\hat{\boldsymbol{\mu}})/I(\hat{\boldsymbol{\mu}}_{\text{B}})$. Consequently, the standardized posterior mode $\hat{\boldsymbol{\mu}}_{\text{B}}$ is optimum under the loss $I(\hat{\boldsymbol{\mu}})d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$.

PROOF. Combining (4.5) with the expression of $\tilde{\rho}_{\text{B}}$, we see that the two densities $\pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{\text{B}}, \rho_{\text{B}})$ and $\pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}, \tilde{\rho}_{\text{B}})$ have the common normalizing constant. A calculation of the Kullback-Leibler separator from $\pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{\text{B}}, \rho_{\text{B}})$ to $\pi_1(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}, \tilde{\rho}_{\text{B}})$ gives (4.6). $\square$

It should be noted that the posterior risk difference is expressed through the Kullback-Leibler separator between the two prior densities. We see that $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{\text{B}})$ in (4.3) is a special case.

Example 4.2. An illustrative example of this proposition is Example 4.3 in Ohnishi and Yanagimoto (2007), where a modified loss function is more familiar than $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$. Consider the hyperbola density $p(x - \mu; \tau_0)$ in (2.11). It follows from Example 2.2 that the dual convex functions are $\psi(\eta) = \cosh(\sinh^{-1} \eta)$ and $\phi(\theta) = \theta \sinh(\tanh^{-1} \theta) - \cosh(\tanh^{-1} \theta)$. Thus, the loss function $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ is given as $\{\cosh(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - 1\} / \cosh \hat{\boldsymbol{\mu}}$. A familiar loss function in the literature is $d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \cosh(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - 1$.

We assume a hyperbola prior density $p(m - \mu; \delta)$ on μ by setting $b_1(\eta) = d\mu/d\eta$. The condition (4.5) is satisfied when we set $I(m) = 1$ since

$$\int_{-\infty}^{\infty} \exp\{-\delta I(m)d(m, \mu)\} b_1(\eta) d\eta = \int_{-\infty}^{\infty} \exp[-\delta\{\cosh(m - \mu) - 1\}] d\mu$$

is independent of m . Proposition 3.1 gives the posterior density as $p(\hat{\mu}_B - \mu; \rho_B)$, where $\tanh \hat{\mu}_B = \{\tau_0 \sinh x + \delta \sinh m\} / \{\tau_0 \cosh x + \delta \cosh m\}$ and $\rho_B = \{\tau_0 \cosh x + \delta \cosh m\} / \cosh \hat{\mu}_B$.

Note that the Kullback-Leibler separator from $p(m_1 - \mu; \delta)$ to $p(m_2 - \mu; \delta)$ is calculated as

$$\text{KL}_\delta(m_1, m_2) = \delta \frac{K_1(\delta)}{K_0(\delta)} \{\cosh(m_1 - m_2) - 1\}.$$

Proposition 4.3 gives the following modified Pythagorean relationship

$$\mathbb{E}[\cosh(\hat{\mu} - \mu) - \cosh(\hat{\mu}_B - \mu) \mid p(\hat{\mu}_B - \mu; \rho_B)] = \frac{1}{\rho_B} \frac{K_1(\rho_B)}{K_0(\rho_B)} \{\cosh(\hat{\mu} - \hat{\mu}_B) - 1\}.$$

5. A dual version of the Pythagorean relationship

We move to the case of an alternative loss function $L(\mu, \hat{\mu})$. Setting $b(\eta) = 1$, we assume the following prior density with respect to the Lebesgue measure on η :

$$(5.1) \quad \pi_2(\eta; \mathbf{m}, \delta) \propto \exp\{-\delta d(\mathbf{m}, \mu)\}.$$

We attempt here to extend Theorem 2 in Diaconis and Ylvisaker (1979) in various ways. For this purpose we assume that

$$(5.2) \quad \lim_{\eta_j \rightarrow \bar{\eta}_j} d(\mathbf{m}, \mu) = \infty \quad \text{and} \quad \lim_{\eta_j \rightarrow \underline{\eta}_j} d(\mathbf{m}, \mu) = \infty \quad \text{for } j = 1, \dots, p.$$

Here $\bar{\eta}_j = \bar{\eta}_j(\boldsymbol{\eta}_{(j)})$ and $\underline{\eta}_j = \underline{\eta}_j(\boldsymbol{\eta}_{(j)})$ are respectively the upper and the lower boundary points when $\boldsymbol{\eta}_{(j)} = (\eta_1, \dots, \eta_{j-1}, \eta_{j+1}, \dots, \eta_p)^T$ is fixed. Roughly speaking, this assumption implies that the density vanishes at the boundary.

PROPOSITION 5.1. *Under the assumption (5.2) it holds for any $\mathbf{m}$ and $\delta > 0$ that*

$$\mathbb{E}[\boldsymbol{\theta} - \boldsymbol{\theta}(\mathbf{m}) \mid \pi_2(\eta; \mathbf{m}, \delta)] = \mathbf{0}.$$

In addition, the posterior density corresponding to $\pi_2(\eta; \mathbf{m}, \delta)$ satisfies

$$\mathbb{E}[\boldsymbol{\theta} - \boldsymbol{\theta}(\hat{\mu}_B) \mid \pi_2(\eta; \hat{\mu}_B, \rho_B)] = \mathbf{0}.$$

PROOF. It follows from (5.2) that

$$\int_{\underline{\eta}_j}^{\bar{\eta}_j} \frac{\partial}{\partial \eta_j} \exp\{-\delta d(\mathbf{m}, \boldsymbol{\mu})\} d\eta_j = 0$$

for $j = 1, \dots, p$. We have from (2.5)

$$\frac{\partial}{\partial \eta_j} d(\mathbf{m}, \boldsymbol{\mu}) = h_{p+1}(\mathbf{m}) \{\theta_j - \theta_j(\mathbf{m})\}.$$

Combining these, we obtain the former part. The latter part follows from the former. $\square$

A Pythagorean relationship which is dual to (4.3) is derived.

PROPOSITION 5.2. *Under the assumption (5.2) the Pythagorean relationship*

$$(5.3) \quad \mathbb{E}[L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) - L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_B) - L(\hat{\boldsymbol{\mu}}_B, \hat{\boldsymbol{\mu}}) \mid \pi_2(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)] = 0$$

holds for any estimator $\hat{\boldsymbol{\mu}}$. Therefore, the standardized posterior mode $\hat{\boldsymbol{\mu}}_B$ is optimum under the loss $L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$.

PROOF. The proof is parallel to that of Proposition 4.2. We use the identity $L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) - L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_B) - L(\hat{\boldsymbol{\mu}}_B, \hat{\boldsymbol{\mu}}) = \{\boldsymbol{\theta} - \boldsymbol{\theta}(\hat{\boldsymbol{\mu}}_B)\}^T \{\boldsymbol{\eta}(\hat{\boldsymbol{\mu}}_B) - \boldsymbol{\eta}(\hat{\boldsymbol{\mu}})\}$ instead of (4.4). $\square$

In order to derive a modification of (5.3), we adopt a loss function $J(\boldsymbol{\eta})L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ with $J(\boldsymbol{\eta})$ being a positive function. The prior density we assume is of the form

$$(5.4) \quad \pi_3(\boldsymbol{\eta}; \mathbf{m}, \delta) \propto \exp\{-\delta d(\mathbf{m}, \boldsymbol{\mu})\} / J(\boldsymbol{\eta}).$$

It follows from Proposition 3.1 that the above prior density is also conjugate, and also that the posterior density is given as $\pi_3(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)$. Here again we assume the regularity condition (5.2).

PROPOSITION 5.3. *Under the assumption (5.2) the modified Pythagorean relationship*

$$(5.5) \quad \mathbb{E}[J(\boldsymbol{\eta})L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) - J(\boldsymbol{\eta})L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_B) - J(\boldsymbol{\eta})L(\hat{\boldsymbol{\mu}}_B, \hat{\boldsymbol{\mu}}) \mid \pi_3(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)] = 0$$

holds for any estimator $\hat{\boldsymbol{\mu}}$. Thus, the standardized posterior mode $\hat{\boldsymbol{\mu}}_B$ is optimum under the loss $J(\boldsymbol{\eta})L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$.

PROOF. Comparing the two prior densities (5.1) and (5.4), we see that $J(\boldsymbol{\eta})\pi_3(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B) \propto \pi_2(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_B, \rho_B)$ as functions of $\boldsymbol{\eta}$. The modified Pythagorean relationship (5.5) is a rewritten version of (5.3). $\square$

Table 1. Examples of the parameter ξ and the loss function $J(\eta)L(\mu, \hat{\mu})$ in the natural exponential family.

ξ	Loss function	Necessary assumption
η	$\text{KL}(\mu, \hat{\mu})$	
μ	$\frac{\text{KL}(\mu, \hat{\mu})}{v(\mu)}$	
$\log \mu$	$\frac{\mu \text{KL}(\mu, \hat{\mu})}{v(\mu)}$	$\mu > 0$
$\psi(\eta)$	$\frac{\text{KL}(\mu, \hat{\mu})}{\mu}$	$\mu > 0$
$\phi(\mu)$	$\frac{\text{KL}(\mu, \hat{\mu})}{\eta v(\mu)}$	$\eta > 0$
$\log \eta$	$\eta \text{KL}(\mu, \hat{\mu})$	$\eta > 0$

Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^T$ be a new parameter which has a one-to-one correspondence with $\boldsymbol{\eta}$. We write the Jacobian of the parameter transformation as $\partial \boldsymbol{\xi} / \partial \boldsymbol{\eta}$. Consider the prior density proportional to $\exp\{-\delta d(\boldsymbol{m}, \boldsymbol{\mu})\}$ with respect to the Lebesgue measure on $\boldsymbol{\xi}$. This is called the standard conjugate prior density by Consonni and Veronese (1992) when the sampling density is in the exponential family. The prior density is equivalent to (5.4) with $1/J(\boldsymbol{\eta}) = |\partial \boldsymbol{\xi} / \partial \boldsymbol{\eta}|$.

Example 5.1. Implications of Propositions 5.2 and 5.3 are clarified by considering the natural exponential family (1.1). We assume that it is regular, i.e., that its canonical space is open. This assumption implies that

$$\lim_{\eta \rightarrow \bar{\eta}} \text{KL}(m, \mu) = \infty \quad \text{and} \quad \lim_{\eta \rightarrow \underline{\eta}} \text{KL}(m, \mu) = \infty,$$

where $\text{KL}(\mu_1, \mu_2) = \phi(\mu_1) + \psi(\eta_2) - \mu_1 \eta_2$. Thus, (5.2) is satisfied. It is known that what is called the Diaconis-Ylvisaker prior density exists for a regular natural exponential family. The prior density is of the form $\pi_2(\eta; m, \delta) \propto \exp\{-\delta \text{KL}(m, \mu)\}$ with respect to the Lebesgue measure on η . Then, the standardized posterior mode $\hat{\mu}_B = (x + \delta m)/(1 + \delta)$ is optimum with respect to the loss $\text{KL}(\mu, \hat{\mu})$.

Defining a new parameter $\xi = \xi(\eta) = \xi(\phi'(\mu))$, we assume the prior density

$$\pi_3(\eta; m, \delta) \propto \exp\{-\delta \text{KL}(m, \mu)\} \left| \frac{d\xi}{d\eta} \right|.$$

The function $\xi(\eta)$ is assumed to be strictly increasing. Several cases of $\xi(\eta)$ and the corresponding loss function $J(\eta)L(\mu, \hat{\mu})$ are given in Table 1, where $v(\mu)$ denotes the variance function.

6. Examination of the non-singularity condition

We prove the conjugate analysis is possible under a condition weaker than (C.4), on which our discussions in Sections 2 through 5 were based.

The condition (C.4) is not satisfied in the von Mises density, which is given by

$$(6.1) \quad p(x - \mu; \tau_0) = \frac{1}{2\pi I_0(\tau_0)} \exp\{\tau_0 \cos(x - \mu)\},$$

where $I_0(\tau_0)$ is the modified Bessel function of the first kind. This is in the curved exponential families when the concentration parameter τ_0 is known. The von Mises prior density $p(m - \mu; \delta)$ is conjugate and was employed by Mardia and El-Atoum (1976), Guttorp and Lockhart (1988) and Rodrigues *et al.* (2000). Note that $1 - \cos(x - \mu) = -\cos x(\cos \mu - \cos x) - \sin x(\sin \mu - \sin x)$. Whether we set $f_1(t) = -\cos t$ or $f_1(t) = -\sin t$, there exists some t_0 such that $f'_1(t_0) = 0$.

Thus we assume the following regularity condition

$$(C.4') \quad \text{rank} F_{p,p+1}(\mathbf{t}) = p \text{ for any } \mathbf{t},$$

where $F_{p,p+1}(\mathbf{t})$ denotes the $p \times (p+1)$ matrix whose (i, j) th component is $\partial f_j(\mathbf{t})/\partial t_i$ ($1 \leq i \leq p, 1 \leq j \leq p+1$). The von Mises case satisfies this condition since the rank of the 1×2 matrix $(\sin t, -\cos t)$ is equal to one for any t .

It seems difficult to define the extended canonical parameter, and therefore we assume the following prior densities on the parameter $\boldsymbol{\mu}$ of the form

$$(6.2) \quad \pi(\boldsymbol{\mu}; \mathbf{m}, \delta) \propto \exp\{-\delta d(\mathbf{m}, \boldsymbol{\mu})\} c(\boldsymbol{\mu}),$$

where $c(\boldsymbol{\mu})$ is an appropriate non-negative function.

PROPOSITION 6.1. *Suppose that the standardized posterior mode (3.2) is uniquely determined. Then, the prior density (6.2) is conjugate.*

PROOF. The proof is similar to that of Proposition 3.1. We prove that the right-hand side of (3.7) is proportional to $d(\hat{\boldsymbol{\mu}}_B, \boldsymbol{\mu})$. It suffices to show that the two vectors $\tau_0 \tilde{\mathbf{h}}(\mathbf{x}) + \delta \tilde{\mathbf{h}}(\mathbf{m})$ and $\tilde{\mathbf{h}}(\hat{\boldsymbol{\mu}}_B)$ are proportional where $\tilde{\mathbf{h}}(\mathbf{t})$ denotes the $(p+1)$ -dimensional vector $(h_1(\mathbf{t}), \dots, h_{p+1}(\mathbf{t}))^T$. By definition, the standardized posterior mode $\hat{\boldsymbol{\mu}}_B$ satisfies

$$\left. \frac{\partial}{\partial \boldsymbol{\mu}} \{ \tau_0 d(\mathbf{x}, \boldsymbol{\mu}) + \delta d(\mathbf{m}, \boldsymbol{\mu}) \} \right|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}_B} = \mathbf{0}.$$

This is expressed in a matrix representation as $F_{p,p+1}(\hat{\boldsymbol{\mu}}_B) \{ \tilde{\mathbf{h}}(\mathbf{x}) + \delta \tilde{\mathbf{h}}(\mathbf{m}) \} = \mathbf{0}$. The equality (2.2) with $\mathbf{s} = \hat{\boldsymbol{\mu}}_B$ is rewritten as $F_{p,p+1}(\hat{\boldsymbol{\mu}}_B) \tilde{\mathbf{h}}(\hat{\boldsymbol{\mu}}_B) = \mathbf{0}$. Note that the matrix $F_{p,p+1}(\hat{\boldsymbol{\mu}}_B)$ is of full rank. It follows from the theory of linear algebra that there exists $\rho_B = \rho_B(\mathbf{x}, \mathbf{m}, \tau_0, \delta)$ such that

$$(6.3) \quad \tau_0 \tilde{\mathbf{h}}(\mathbf{x}) + \delta \tilde{\mathbf{h}}(\mathbf{m}) = \rho_B \tilde{\mathbf{h}}(\hat{\boldsymbol{\mu}}_B).$$

Thus, the desired proportionality

$$\tau_0 d(\mathbf{x}, \boldsymbol{\mu}) + \delta d(\mathbf{m}, \boldsymbol{\mu}) - \tau_0 d(\mathbf{x}, \hat{\boldsymbol{\mu}}_B) - \delta d(\mathbf{m}, \hat{\boldsymbol{\mu}}_B) = \rho_B d(\hat{\boldsymbol{\mu}}_B, \boldsymbol{\mu})$$

is obtained. The existence assumption of $\hat{\mu}_B$ guarantees that $\rho_B > 0$. Thus, we see that the posterior density is expressed as $\pi(\boldsymbol{\mu}; \hat{\mu}_B, \rho_B)$. $\square$

It should be noted that (6.3) is an extension of a type of linearity which was pointed out by Rodrigues *et al.* (2000) in the von Mises case.

Discussions similar to those in Propositions 4.2 and 4.3 hold true under (C.4'). We assume that there exist a positive function $I(\mathbf{m})$ and a non-negative function $c_1(\boldsymbol{\mu})$ such that

$$(6.4) \quad \frac{\partial}{\partial \mathbf{m}} \int \exp\{-\tilde{\delta} I(\mathbf{m}) d(\mathbf{m}, \boldsymbol{\mu})\} c_1(\boldsymbol{\mu}) d\boldsymbol{\mu} = \mathbf{0}.$$

PROPOSITION 6.2. *Assume the prior density $\pi_1(\boldsymbol{\mu}; \mathbf{m}, \delta) \propto \exp\{-\delta d(\mathbf{m}, \boldsymbol{\mu})\} c_1(\boldsymbol{\mu})$ under (6.4). The following modified Pythagorean relationship*

$$\begin{aligned} & \mathbb{E}[I(\hat{\boldsymbol{\mu}})d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) - I(\hat{\boldsymbol{\mu}}_B)d(\hat{\boldsymbol{\mu}}_B, \boldsymbol{\mu}) \mid \pi_1(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_B, \rho_B)] \\ &= \frac{I(\hat{\boldsymbol{\mu}}_B)}{\rho_B} \text{KL}(\pi_1(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_B, \rho_B), \pi_1(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}, \tilde{\rho}_B)) \end{aligned}$$

holds for any estimator $\hat{\boldsymbol{\mu}}$, where ρ_B is given in (6.3) and $\tilde{\rho}_B = \rho_B I(\hat{\boldsymbol{\mu}})/I(\hat{\boldsymbol{\mu}}_B)$. Consequently, the standardized posterior mode $\hat{\boldsymbol{\mu}}_B$ is optimum under the loss $I(\hat{\boldsymbol{\mu}})d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$.

PROOF. The proof is parallel to that of Proposition 4.3. The posterior density is $\pi_1(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_B, \rho_B)$ as shown in the proof of Proposition 6.1. The key is that the two densities $\pi_1(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_B, \rho_B)$ and $\pi_1(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}, \tilde{\rho}_B)$ have the common normalizing constant. $\square$

Example 6.1. Consider the von Mises density $p(x - \mu; \tau_0)$ in (6.1). A parallel discussion to that in Example 4.2 is applicable. We see that a von Mises prior $p(m - \mu; \delta)$, which is given by setting $I(m) = 1$ and $c_1(\mu) = 1$, satisfies (6.4) since

$$\int_0^{2\pi} \exp\{-\delta I(m) d(m, \mu)\} c_1(\mu) d\mu = \int_0^{2\pi} \exp[-\delta \{1 - \cos(m - \mu)\}] d\mu$$

is independent of m . Therefore we apply Proposition 6.2 to obtain the following modified Pythagorean relationship

$$\mathbb{E}[\cos(\hat{\mu}_B - \mu) - \cos(\hat{\mu} - \mu) \mid p(\hat{\mu}_B - \mu; \rho_B)] = \frac{I_1(\rho_B)}{I_0(\rho_B)} \{1 - \cos(\hat{\mu} - \hat{\mu}_B)\},$$

where $\hat{\mu}_B = \arg \max_{\mu} \{\tau_0 \cos(x - \mu) + \delta \cos(m - \mu)\}$ and $\rho_B = \{\tau_0^2 + \delta^2 + 2\tau_0\delta \cos(x - m)\}^{1/2}$.

Appendix A

PROOF OF PROPOSITION 2.1. For simplicity we deal with the case of $p = 1$. Suppose that $K < 2$, i.e., $K = 1$. It follows from (2.2) that $h_1(s)f'_1(s) = 0$ for any s . Thus we have $h_1(s) = 0$ or $f'_1(s) = 0$ for any s , which contradicts (C.1) or (C.2), respectively. $\square$

PROOF OF PROPOSITION 2.2. Again, we consider the one-dimensional case ($p = 1$) for notational simplicity. Suppose that $K > 2$, i.e., $K \geq 3$. Since $h_1(t)$ and $h_2(t)$ are linearly independent, there exists a pair (s_1, s_2) such that the two vectors $(h_1(s_1), h_1(s_2))$ and $(h_2(s_1), h_2(s_2))$ are linearly independent. Thus, there exist coefficients $a_j(s_1, s_2)$ and $b_j(s_1, s_2)$ such that

$$(A.1) \quad \begin{pmatrix} h_j(s_1) \\ h_j(s_2) \end{pmatrix} = \begin{pmatrix} h_1(s_1) & h_2(s_1) \\ h_1(s_2) & h_2(s_2) \end{pmatrix} \begin{pmatrix} a_j(s_1, s_2) \\ b_j(s_1, s_2) \end{pmatrix} \quad (j = 3, \dots, K).$$

It follows from the condition in this proposition that

$$\sum_{j=1}^K \{h_j(s_1) + \delta h_j(s_2)\} \{f_j(t) - f_j(\hat{t}_M)\} = \rho(s_1, s_2, \delta) \sum_{j=1}^K h_j(\hat{t}_M) \{f_j(t) - f_j(\hat{t}_M)\}.$$

From (C.2) we see that $h_j(s_1) + \delta h_j(s_2) = \rho(s_1, s_2, \delta) h_j(\hat{t}_M)$ for $j = 1, \dots, K$. This, together with (A.1), implies the following equality for any $\delta > 0$:

$$\begin{aligned} h_j(\hat{t}_M) &= \frac{1}{\rho(s_1, s_2, \delta)} [\{a_j(s_1, s_2)h_1(s_1) + b_j(s_1, s_2)h_2(s_1)\} \\ &\quad + \delta \{a_j(s_1, s_2)h_1(s_2) + b_j(s_1, s_2)h_2(s_2)\}] \\ &= a_j(s_1, s_2)h_1(\hat{t}_M) + b_j(s_1, s_2)h_2(\hat{t}_M) \quad (j = 3, \dots, K). \end{aligned}$$

Recall that $\hat{t}_M$ is dependent on δ while a_j and b_j are not. Therefore, we see that the equality contradicts (C.1), which completes the proof. $\square$

PROOFS OF LEMMAS 2.1 and 2.2. The chain rule for partial differentiation gives

$$(A.2) \quad \frac{\partial}{\partial \eta_j} f_{p+1}(\boldsymbol{\mu}) = \sum_{k=1}^p \frac{\partial f_{p+1}(\boldsymbol{\mu})}{\partial \mu_k} \frac{\partial \mu_k}{\partial \eta_j}.$$

It follows from (2.3) and the k th component of (2.2) that

$$(A.3) \quad \frac{\partial f_{p+1}(\boldsymbol{\mu})}{\partial \mu_k} = \frac{1}{h_{p+1}(\boldsymbol{\mu})} \sum_{l=1}^p h_l(\boldsymbol{\mu}) \frac{\partial \eta_l}{\partial \mu_k}.$$

Combining (A.2) and (A.3), we have

$$(A.4) \quad \frac{\partial \psi(\boldsymbol{\eta})}{\partial \eta_j} = \frac{h_j(\boldsymbol{\mu})}{h_{p+1}(\boldsymbol{\mu})}.$$

Note that $d(\mathbf{x}, \boldsymbol{\mu}) = -\sum_{j=1}^p \eta_j h_j(\mathbf{x}) + \psi(\boldsymbol{\eta}) h_{p+1}(\mathbf{x}) - \sum_{j=1}^{p+1} h_j(\mathbf{x}) f_j(\mathbf{x})$. Differentiating both sides of the equality $\int \exp\{-\tau_0 d(\mathbf{x}, \boldsymbol{\mu})\} a(\mathbf{x}; \tau_0) d\mathbf{x} = 1$ with respect to η_j , we have the required result (2.6) in Lemma 2.2.

Again, differentiating both sides of (2.6) with respect to η_k , we see that

$$\begin{aligned} & \mathbb{E}[h_{p+1}(\mathbf{x}) \mid p(\mathbf{x}; \boldsymbol{\mu}, \tau_0)] \frac{\partial^2 \psi(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_j} \\ &= \tau_0 \mathbb{E}[\{h_j(\mathbf{x}) - \theta_j h_{p+1}(\mathbf{x})\} \{h_k(\mathbf{x}) - \theta_k h_{p+1}(\mathbf{x})\} \mid p(\mathbf{x}; \boldsymbol{\mu}, \tau_0)], \end{aligned}$$

which implies the convexity of $\psi(\boldsymbol{\eta})$. $\square$

Acknowledgements

This work was supported by KAKENHIs (Grants-in-Aid for Scientific Research) 17700287 and 20700259. The authors are grateful to anonymous referees for their helpful comments, which improved the earlier version of this paper.

REFERENCES

- Amari, S.-I., Nagaoka, H. (2000). *Methods of Information Geometry*, American Mathematical Society.
- Barndorff-Nielsen, O. (1978a). *Information and Exponential Families in Statistical Theory*, New York, J. Wiley & Sons.
- Barndorff-Nielsen, O. (1978b). Hyperbolic distribution and distribution on hyperbolae, *Scand. J. Stat.*, **5**, 151–157.
- Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion), *J. Roy. Stat. Soc. B*, **41**, 113–147.
- Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*, Chichester, Wiley.
- Chen, M.-H. and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models, *Statistica Sinica*, **13**, 461–476.
- Consonni, G. and Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions, *J. Am. Stat. Assoc.*, **87**, 1123–1127.
- Consonni, G. and Veronese, P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors, *Scand. J. Stat.*, **28**, 377–406.
- Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families, *The Annals of Statistics*, **7**, 269–281. *Bayesian Statistics 4* (pp. 669–674). Oxford, Oxford University Press.
- Gutiérrez-Peña, E. (1997). Moments for the canonical parameter of an exponential family under a conjugate distribution, *Biometrika*, **84**, 727–732.
- Gutiérrez-Peña, E. and Smith, A. F. M. (1997). Exponential and Bayesian conjugate families: Review and extensions (with discussion), *Test*, **6**, 1–90.
- Guttorp, P. and Lockhart, R. A. (1988). Finding the location of a signal: A Bayesian analysis, *J. Am. Stat. Assoc.*, **83**, 322–330.
- Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models, *Statistical Science*, **15**, 46–60.
- Jensen, J. L. (1981). On the hyperboloid distribution, *Scand. J. Stat.*, **8**, 193–206.
- Jørgensen, B. (1997). *The Theory of Dispersion Models*, London, Chapman and Hall.
- Mardia, K. V. and El-Atoum, S. A. M. (1976). Bayesian inference for the von Mises-Fisher distribution, *Biometrika*, **63**, 203–206.
- Morris, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory, *The Annals of Statistics*, **11**, 515–529.

- Ohnishi, T. and Dunn, P. K. (2009). A Bayesian analysis of Tweedie generalized linear model using conjugate prior, *Research Memorandum*, The Institute of Statistical Mathematics, **1100**.
- Ohnishi, T. and Yanagimoto, T. (2003). Electrostatic views of Stein-type estimation of location vectors, *J. Jpn. Stat. Soc.*, **33**, 39–64.
- Ohnishi, T. and Yanagimoto, T. (2007). Conjugate location-dispersion families, *J. Jpn. Stat. Soc.*, **37**, 307–325.
- Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*, Graduate School of Business Administration, Harvard University, Boston.
- Rodrigues, J., Leite, J. G. and Milan, L. A. (2000). An empirical Bayes inference for the von Mises distribution, *Australian & New Zealand Journal of Statistics*, **42**, 433–440.
- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution, *Annal. Stat.*, **9**, 1135–1151.
- Tsubaki, H. (1988). An extension of the quasi-likelihood and its application to the linear statistical inference, Doctoral Dissertation, University of Tokyo. (in Japanese)
- Yanagimoto, T. and Ohnishi, T. (2005a). Extensions of a conjugate prior through the Kullback-Leibler separators, *J. Multivar. Anal.*, **92**, 116–133.
- Yanagimoto, T. and Ohnishi, T. (2005b). Standardized posterior mode for the flexible use of a conjugate prior, *J. Stat. Plann. Inference*, **131**, 253–269.
- Yanagimoto, T. and Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e -mixture, *J. Stat. Plann. Inference*, **139**, 3064–3075.