数理解析研究所講究録 1439 巻 2005 年 202-220

Dual structure in the conjugate analysis of curved exponential families

統計数理研究所　大西 俊郎 (Toshio Ohnishi)　柳本 武美 (Takemi Yanagimoto)
The Institute of Statistical Mathematics

Abstract

Curved exponential families admitting conjugate prior densities are introduced and explored. Introducing extended versions of the mean and the canonical parameters, we expand the conjugate analysis to these curved exponential families. Emphasis is put on dual structures. In fact, we derive the dual Pythagorean relationships with respect to posterior risks, each of which makes it clear how the Bayes estimator dominates other estimators. We also show that the conjugate prior density is the least informative.

Key Words: closure under sampling, conjugacy, duality, least information, Legendre transformation, linearity, proper dispersion model, Pythagorean relationship, standardized posterior mode

1. Introduction

The conjugate analysis is one of the most important fields in Bayesian inference. It has attracted interests of many researchers including Consonni and Veronese (1992, 2001), Gutiérrez-Peña (1992, 1997) and Gutiérrez-Peña and Smith (1997). Simplicity in calculating the posterior mean, or the Bayes estimator, is characteristic of the conjugate analysis. A minimax property of the conjugate prior density was shown by Morris (1983) and Consonni and Veronese (1992). Recently, extensions of the conjugate prior density have been studied by such authors as Ibrahim and Chen (1998, 2000) and Yanagimoto and Ohnishi (2005a). The dual structure is elegantly observed in the exponential families and the curved exponential families (Barndorff-Nielsen 1978a, Amari and Nagaoka 2000). In fact, the importance of the curved exponential families owes largely to the dual structure. That in the conjugate analysis was pursued in naive ways by Yanagimoto and Ohnishi (2005ab).

The original definition of conjugacy is closure under sampling, i.e., that the prior and the posterior densities belong to the same family of distributions, which was defined by Raiffa and Schlaifer (1961, pp.43–57). In this paper we mean closure under sampling by conjugacy according to their definition. It is known that this definition produces ambiguity. Take a sampling density in a natural exponential family

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

for instance. The prior density $\pi(\eta;\, m,\, \delta) \propto \exp\left[\delta\,\{m\eta - \psi(\eta)\}\right] b(\eta)$ is conjugate, that is, closed under sampling, and we cannot specify the type of the supporting measure $b(\eta)$ by conjugacy alone. Diaconis and Ylvisaker (1979) characterized the choice $b(\eta) = 1$ by linearity of the posterior mean of the mean parameter and defined conjugacy by this linearity. The

reason why the present authors adopt such an ambiguous definition is a conjecture that closure under sampling in itself implies a certain optimum property. This will be shown affirmatively in Section 3.

The conjugate analysis is not restricted to the natural exponential family case. Mardia and El-Atoum (1976) showed that the von Mises distribution, which is in the curved exponential families, has a conjugate prior density. For the sampling density

$$p_{\text{vM}}(x;\, \mu,\, \tau) = \frac{1}{2\pi I_0(\tau)} \exp\{\tau \cos(x-\mu)\}, \tag{1.2}$$

where $I_0(\tau)$ is the modified Bessel function of the first kind, the von Mises prior density $p_{\text{vM}}(\mu;\, m,\, \delta)$ is conjugate. This prior density was employed by Guttorp and Lockhart (1988) and Rodrigues *et al.* (2000). Here the linearity of the posterior mean of μ does not hold in the sense of Diaconis and Ylvisaker (1979), although Rodrigues *et al.* (2000) pointed out that a type of linearity holds.

This paper has the following two aims. One is to reveal an essential aspect of the conjugate analysis. We consider the following sampling density

$$p(\boldsymbol{x};\, \boldsymbol{\mu}) = \exp\{-d(\boldsymbol{x},\, \boldsymbol{\mu})\}\, a(\boldsymbol{x}), \tag{1.3}$$

where $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are p-dimensional, and $d(\boldsymbol{a},\, \boldsymbol{t})$ is expressed through the $(2p+2)$ functions, $f_k(\boldsymbol{t})$'s and $h_k(\boldsymbol{t})$'s, as

$$d(\boldsymbol{a},\, \boldsymbol{t}) = \sum_{j=1}^{p+1} h_j(\boldsymbol{a})\{f_j(\boldsymbol{t}) - f_j(\boldsymbol{a})\}.$$

In general, the density (1.3) belongs to the curved exponential families. As will be seen in the subsequent sections, the sampling density (1.3) with $p=1$ covers the natural exponential family (1.1) and the von Mises distribution (1.2). Thus, a unified discussion is possible. We will show that the prior density of the form $\pi(\boldsymbol{\mu};\, \boldsymbol{m},\, \delta) \propto \exp\{-\delta d(\boldsymbol{m},\, \boldsymbol{\mu})\}\, c(\boldsymbol{\mu})$ is conjugate for the sampling density (1.3). We will also prove that the conjugate prior has the minimum information among a certain set of prior densities. This property implies a type of superiority of the conjugate analysis over non-conjugate ones. It seems to be closely related to the minimax property of the conjugate prior density shown by Morris (1983) and Consonni and Veronese (1992).

The other, but main aim is to show dual structure of the conjugate analysis. We will assume two types of prior densities which have dual properties, and discuss conjugate analyses separately. The loss functions we adopt are also dual to each other. We derive the dual Pythagorean relationships with respect to posterior risks. These relationships make it clear how the Bayes estimator dominates other ones. The dual structure we will show is similar to the one with respect to the mean and the canonical parameters in the exponential families, which Barndorff-Nielsen (1978a) and Amari and Nagaoka (2000) pointed out. It is a substantial extension of previous results by the authors to the curved exponential family (1.3).

The organization of this paper is as follows. Section 2 introduces certain curved exponential families admitting the conjugate analysis. Extended versions of the mean and the canonical parameters are defined under some regularity conditions. Section 3 shows conjugacy of the assumed prior density. An optimum property of the conjugate prior density is also proved. Sections 4 and 5 reveal dual structure of the conjugate analysis. We derive the dual Pythagorean relationships with respect to posterior risks. Extended versions of the dual

Pythagorean relationships are also obtained. Section 6 discusses the conjugate analysis under weaker regularity conditions, which covers the von Mises case.

2. Extended mean and canonical parameters

In this section we introduce certain curved exponential families for which we can discuss the conjugate analysis. Counterparts of the mean and the canonical parameters in the exponential families are defined. We will learn that these parameters are useful in understanding the dual structure of the conjugate analysis. The two propositions and the two lemmas are obtained, the proofs of which are given in Appendix.

We investigate the conjugate analysis of the curved exponential family

$$\mathcal{F} = \{\, p(\boldsymbol{x};\, \boldsymbol{\mu}) \mid p(\boldsymbol{x};\, \boldsymbol{\mu}) = \exp\{-d(\boldsymbol{x},\, \boldsymbol{\mu})\}\, a(\boldsymbol{x}) \,\}, \tag{2.1}$$

where $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are p-dimensional, $a(\boldsymbol{x})$ is the supporting measure and

$$d(\boldsymbol{a},\, \boldsymbol{t}) = \sum_{j=1}^{p+1} h_j(\boldsymbol{a})\{f_j(\boldsymbol{t}) - f_j(\boldsymbol{a})\}. \tag{2.2}$$

In the above we assume the following three regularity conditions:

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

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

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

The function $d(\boldsymbol{a},\, \boldsymbol{t})$ is the deviance function introduced in Jørgensen (1997, p.4). The regularity condition (C.3) implies that

$$\frac{\partial}{\partial \boldsymbol{t}} d(\boldsymbol{a},\, \boldsymbol{t})\bigg|_{\boldsymbol{t}=\boldsymbol{a}} = \boldsymbol{0} \quad \text{for any } \boldsymbol{a}. \tag{2.3}$$

The family $\mathcal{F}$ covers the exponential family case. In fact, set $h_{p+1}(\boldsymbol{x}) = 1$ in the sampling density in (2.1). Then the density is written as $p(\boldsymbol{x};\, \boldsymbol{\mu}) = \exp\{-\sum_{j=1}^{p} h_j(\boldsymbol{x}) f_j(\boldsymbol{\mu}) - f_{p+1}(\boldsymbol{\mu})\}\, \tilde{a}(\boldsymbol{x})$, where $\tilde{a}(\boldsymbol{x}) = \exp\{\sum_{j=1}^{p} h_j(\boldsymbol{x}) f_j(\boldsymbol{x}) + f_{p+1}(\boldsymbol{x})\}\, a(\boldsymbol{x})$. This is a density in an exponential family.

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 $F_{p,p}(\boldsymbol{t})$ denote the $p \times p$ matrix whose (i,j)th component is $\partial f_j(\boldsymbol{t})/\partial t_i$ $(1 \leq i,\, j \leq p)$. In addition to (C.1)–(C.3) we assume the following regularity condition:

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

The case where this *non-singularity condition* is not satisfied will be discussed in the final section. Here we show that $h_{p+1}(\boldsymbol{a}) \neq 0$ for any $\boldsymbol{a}$. Suppose that $h_{p+1}(\boldsymbol{a}_0) = 0$ for some $\boldsymbol{a}_0$. The equality (2.3) can be rewritten as

$$F_{p,p}(\boldsymbol{a})\, \boldsymbol{h}(\boldsymbol{a}) = -h_{p+1}(\boldsymbol{a})\, \frac{\partial}{\partial \boldsymbol{a}} f_{p+1}(\boldsymbol{a}),$$

where $\boldsymbol{h}(\boldsymbol{a}) = \big(h_1(\boldsymbol{a}), \ldots, h_p(\boldsymbol{a})\big)^T$. This set of linear equations, together with (C.4), gives that $\boldsymbol{h}(\boldsymbol{a}_0) = \boldsymbol{0}$ and therefore that $d(\boldsymbol{a}_0,\, \boldsymbol{t}) = 0$ for any $\boldsymbol{t}$, which contradicts (C.3). Thus, we assume without loss of generality that

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

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

$$\eta_j = -f_j(\boldsymbol{\mu}) \tag{2.4}$$

for $j = 1, \ldots, p$. 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*. The parameter vector $\boldsymbol{\eta}$ is the very canonical one in the exponential family case.

We regard $f_{p+1}(\boldsymbol{\mu})$ as a function of $\boldsymbol{\eta}$ and set

$$\psi(\boldsymbol{\eta}) = f_{p+1}(\boldsymbol{\mu}). \tag{2.5}$$

This function becomes the cumulant function in the exponential family case. Although the cumulant function is convex, the convexity is not obvious in the curved exponential family $\mathcal{F}$. We show in the following lemma that convexity also holds true for $\mathcal{F}$.

Lemma 2.1.
The function $\psi(\boldsymbol{\eta})$ defined by (2.5) *is convex.*

Using the Legendre transformation, we define another parameter $\boldsymbol{\theta}$ and another convex function $\phi(\boldsymbol{\theta})$ conjugate to $\boldsymbol{\eta}$ and $\psi(\boldsymbol{\eta})$, respectively. We set $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ as $\theta_j = (\partial/\partial\eta_j)\psi(\boldsymbol{\eta})$ for $j = 1, \ldots, p$. As is given by (A.4) in Appendix, we have

$$\theta_j = \frac{h_j(\boldsymbol{\mu})}{h_{p+1}(\boldsymbol{\mu})}. \tag{2.6}$$

The following lemma clarifies the meaning of $\boldsymbol{\theta}$. We may call $\boldsymbol{\theta}$ the *extended mean parameter*.

Lemma 2.2.
It holds for $j = 1, \ldots, p$ that

$$\mathrm{E}\big[\, h_j(\boldsymbol{x}) - \theta_j h_{p+1}(\boldsymbol{x}) \;\big|\; p(\boldsymbol{x};\, \boldsymbol{\mu}) \,\big] = 0.$$

The convex function conjugate to $\psi(\boldsymbol{\eta})$ is expressed as $\phi(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\boldsymbol{\eta} - \psi(\boldsymbol{\eta})$ 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 function $\phi(\boldsymbol{\theta})$ has the following representation as a function of $\boldsymbol{\mu}$:

$$\phi(\boldsymbol{\theta}) = -\sum_{j=1}^{p} \frac{h_j(\boldsymbol{\mu})}{h_{p+1}(\boldsymbol{\mu})} f_j(\boldsymbol{\mu}) - f_{p+1}(\boldsymbol{\mu}). \tag{2.7}$$

The definition of $\phi(\boldsymbol{\theta})$ yields that

$$L(\boldsymbol{\mu}_1,\, \boldsymbol{\mu}_2) = \phi(\boldsymbol{\theta}_1) + \psi(\boldsymbol{\eta}_2) - \boldsymbol{\theta}_1^T\boldsymbol{\eta}_2 \tag{2.8}$$

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

$$L(\boldsymbol{\mu}_1,\, \boldsymbol{\mu}_3) - L(\boldsymbol{\mu}_1,\, \boldsymbol{\mu}_2) - L(\boldsymbol{\mu}_2,\, \boldsymbol{\mu}_3) = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3) \tag{2.9}$$

holds, which will play a key role in subsequent discussions.

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. Note that the function $d(\boldsymbol{x}, \boldsymbol{\mu})$ of $\boldsymbol{\mu}$ given data $\boldsymbol{x}$ becomes the normed log-likelihood function, i.e., $d(\boldsymbol{x}, \boldsymbol{\mu}) = \max_{\boldsymbol{\mu}}\{\log p(\boldsymbol{x}; \boldsymbol{\mu})\} - \log p(\boldsymbol{x}; \boldsymbol{\mu})$. A calculation using the formulas (2.4) through (2.7) gives

$$d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = h_{p+1}(\boldsymbol{\mu}_1)\, L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2). \tag{2.10}$$

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

$$\mathrm{KL}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \mathrm{E}\big[\, h_{p+1}(\boldsymbol{x}) \;\big|\; p(\boldsymbol{x}; \boldsymbol{\mu}_1) \,\big]\, L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2). \tag{2.11}$$

These two expressions (2.10) and (2.11) reveal the relation among $d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$, $L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and $\mathrm{KL}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. Modification of the loss functions $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ and $L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ will be dealt with in Sections 4 and 5.

The following two examples give calculations of the extended mean and the extended canonical parameters. We deal with the natural exponential family and the hyperbola distribution.

Example 2.1. Let us consider the case of the natural exponential family (1.1). Let μ be the mean parameter and $\phi(\mu)$ the convex function conjugate 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 (1.1) as

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

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 the extended canonical parameters is obtained by setting $f_1(\mu) = \psi(\phi'(\mu))$, $f_2(\mu) = -\phi'(\mu)$, $h_1(x) = 1$ and $h_2(x) = x$. If we adopt this parameterization in the gamma distribution, the derived dual convex functions are the same as those in the Poisson distribution under the ordinary parameterization. This is directly related to the fact that the gamma prior density is conjugate for both the sampling distributions.

Example 2.2. We discuss the hyperbola distribution having the density

$$p_{\mathrm{H}}(x; \mu, \tau) = \frac{1}{2K_0(\tau)} \exp\{-\tau \cosh(x - \mu)\}, \tag{2.12}$$

where $K_0(\tau)$ 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 the extended canonical parameters are given by $\theta = \tanh\mu$ and $\eta = \sinh\mu$, respectively. This sampling density allows us the conjugate analysis 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).

3. Conjugacy with the least information property

Consider the prior density

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

on the extended canonical parameter $\boldsymbol{\eta}$ where $b(\boldsymbol{\eta})$ is a non-negative function and $\exp\{K(\boldsymbol{m},\, \delta)\}$ is the normalizing constant. We prove that this prior density is conjugate for the sampling density in (2.1). Comparing with non-conjugate prior densities, we also show the least information property of the conjugate prior density.

First, we give a proof of the conjugacy in terms of the duality of the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}(\boldsymbol{\mu})$ denote the p-dimensional vector with the jth component $\theta_j = \theta_j(\boldsymbol{\mu})$ in (2.6). In this paper we employ the *standardized posterior mode* $\hat{\boldsymbol{\mu}}_{smap}$, which is a modified posterior mode of $\boldsymbol{\mu}$ derived by discarding the Jacobian factor $b(\boldsymbol{\eta})$ in Yanagimoto and Ohnishi (2005b). In our case it is given by

$$\hat{\boldsymbol{\mu}}_{smap} = \arg\min_{\boldsymbol{\mu}}\{d(\boldsymbol{x},\, \boldsymbol{\mu}) + \delta d(\boldsymbol{m},\, \boldsymbol{\mu})\}. \tag{3.2}$$

It should be noted that the estimation procedure is invariant with respect to a parameter transformation.

The regularity conditions (C.4) and (C.5) yield that the standardized posterior mode is uniquely determined for any $\boldsymbol{x}$, $\boldsymbol{m}$ and δ. Actually, a calculation using (2.8) and (2.10) gives the expression of the standardized posterior mode $\hat{\boldsymbol{\theta}}_{smap}$ as

$$\hat{\boldsymbol{\theta}}_{smap} = \frac{h_{p+1}(\boldsymbol{x})\boldsymbol{\theta}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})\boldsymbol{\theta}(\boldsymbol{m})}{h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})}.$$

Noting that $\hat{\boldsymbol{\theta}}_{smap} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}}_{smap})$ and recalling the equality (2.6), we obtain the componentwise expression

$$\frac{h_j(\hat{\boldsymbol{\mu}}_{smap})}{h_{p+1}(\hat{\boldsymbol{\mu}}_{smap})} = \frac{h_j(\boldsymbol{x}) + \delta h_j(\boldsymbol{m})}{h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})} \quad (1 \le j \le p). \tag{3.3}$$

We can see a type of linearity of the standardized posterior mode in $\boldsymbol{\theta}$. It is interesting to compare this linearity holding for any $b(\boldsymbol{\eta})$ with the posterior linearity by which Diaconis and Ylvisaker (1979) characterized the constant supporting measure on the canonical parameter.

Theorem 3.1.

The prior density (3.1) *is conjugate. The posterior density is expressed as* $\pi(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$ *where* $\hat{\boldsymbol{\mu}}_{smap}$ *is the standardized posterior mode* (3.2) *and*

$$\delta^* = \frac{h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})}{h_{p+1}(\hat{\boldsymbol{\mu}}_{smap})}. \tag{3.4}$$

Proof. The posterior density is proportional to $\exp\{-d(\boldsymbol{x}, \boldsymbol{\mu}) - \delta d(\boldsymbol{m}, \boldsymbol{\mu})\} b(\boldsymbol{\eta})$. The expression (2.2) of $d(\boldsymbol{a}, \boldsymbol{t})$ gives

$$\begin{aligned} & d(\boldsymbol{x}, \boldsymbol{\mu}) + \delta d(\boldsymbol{m}, \boldsymbol{\mu}) - d(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_{smap}) - \delta d(\boldsymbol{m}, \hat{\boldsymbol{\mu}}_{smap}) \\ & = \sum_{j=1}^{p+1} \{h_j(\boldsymbol{x}) + \delta h_j(\boldsymbol{m})\}\{f_j(\boldsymbol{\mu}) - f_j(\hat{\boldsymbol{\mu}}_{smap})\}. \end{aligned} \tag{3.5}$$

It follows from (3.3) and (3.4) that

$$h_j(\boldsymbol{x}) + \delta h_j(\boldsymbol{m}) = \delta^* h_j(\hat{\boldsymbol{\mu}}_{smap})$$

for $j = 1, \ldots, p$. Thus, using (2.2) again, we see that the left-hand side of (3.5) reduces to $\delta^* d(\hat{\boldsymbol{\mu}}_{smap}, \boldsymbol{\mu})$, which completes the proof. □

Next, we show that the conjugate prior density has the least information property. For this purpose we make comparison with a non-conjugate prior. Let $\pi(\boldsymbol{\eta})$ denote an arbitrary prior density, and write the corresponding posterior density as $\pi(\boldsymbol{\eta}|\boldsymbol{x})$ for a given $\boldsymbol{x}$. Then we consider the family $\mathcal{P}(\boldsymbol{x}, \boldsymbol{m}, \delta)$ of prior densities satisfying

$$\mathrm{E}\big[\, (\boldsymbol{\eta}^T, \psi(\boldsymbol{\eta})) \;\big|\; \pi(\boldsymbol{\eta}|\boldsymbol{x}) \,\big] = \mathrm{E}\big[\, (\boldsymbol{\eta}^T, \psi(\boldsymbol{\eta})) \;\big|\; \pi(\boldsymbol{\eta}, \hat{\boldsymbol{\mu}}_{smap}, \delta^*) \,\big]. \tag{3.6}$$

Since $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \phi(\hat{\boldsymbol{\theta}}) + \psi(\boldsymbol{\eta}) - \hat{\boldsymbol{\theta}}^T \boldsymbol{\eta}$, this condition is equivalent to the condition that the equality

$$\mathrm{E}\big[\, L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \;\big|\; \pi(\boldsymbol{\eta}|\boldsymbol{x}) \,\big] = \mathrm{E}\big[\, L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \;\big|\; \pi(\boldsymbol{\eta}, \hat{\boldsymbol{\mu}}_{smap}, \delta^*) \,\big]$$

holds for any estimate $\hat{\boldsymbol{\mu}}$. To be specific, any prior density in $\mathcal{P}(\boldsymbol{x}, \boldsymbol{m}, \delta)$ has the identical Bayes estimate and the identical posterior risk of the Bayes estimate. Thus, it is reasonable to compare the amount of information contained among the prior densities in $\mathcal{P}(\boldsymbol{x}, \boldsymbol{m}, \delta)$.

The following theorem gives a Pythagorean relationship holding for the conjugate prior density. See Figure 1. The least information property is obtained as a corollary.

Theorem 3.2.
Let $\pi(\boldsymbol{\eta})$ be any prior density in $\mathcal{P}(\boldsymbol{x}, \boldsymbol{m}, \delta)$ defined by the condition (3.6)*, and write the corresponding posterior density as $\pi(\boldsymbol{\eta}|\boldsymbol{x})$. Then, the following Pythagorean relationship*

$$\begin{aligned} \mathrm{KL}\big(\pi(\boldsymbol{\eta}|\boldsymbol{x}), \pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)\big) &= \mathrm{KL}\big(\pi(\boldsymbol{\eta}|\boldsymbol{x}), \pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{smap}, \delta^*)\big) \\ &\quad + \mathrm{KL}\big(\pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{smap}, \delta^*), \pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)\big) \end{aligned} \tag{3.7}$$

holds for any hyperparameters $\boldsymbol{m}_1$ and δ_1.

Proof. Note that

$$\begin{aligned} & \mathrm{KL}\big(\pi(\boldsymbol{\eta}|\boldsymbol{x}), \pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)\big) - \mathrm{KL}\big(\pi(\boldsymbol{\eta}|\boldsymbol{x}), \pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{smap}, \delta^*)\big) \\ & = \mathrm{E}\left[\log \frac{\pi(\boldsymbol{\eta}; \hat{\boldsymbol{\mu}}_{smap}, \delta^*)}{\pi(\boldsymbol{\eta}; \boldsymbol{m}_1, \delta_1)} \,\middle|\, \pi(\boldsymbol{\eta}|\boldsymbol{x}) \right]. \end{aligned}$$

If we replace $\pi(\boldsymbol{\eta}|\boldsymbol{x})$ with $\pi(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$ in the right-hand side, the expected value becomes the Kullback-Leibler separator from $\pi(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$ to $\pi(\boldsymbol{\eta};\, \boldsymbol{m}_1,\, \delta_1)$. Thus, it is sufficient to show that this replacement does not change the above expected value. It follows that

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

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

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

Since the posterior density $\pi(\boldsymbol{\eta}|\boldsymbol{x})$ satisfies (3.6) by definition, the required result is obtained. □

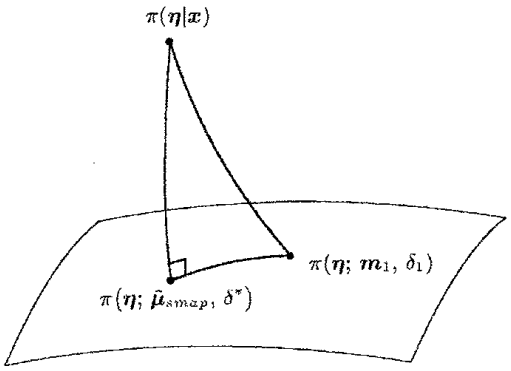

Figure 1: The Pythagorean relationship holding for the conjugate prior.

Now, we solve the minimization problem of the following functional

$$G[\pi(\boldsymbol{\eta})] = \mathrm{KL}\big(\, \pi(\boldsymbol{\eta}|\boldsymbol{x}),\, \pi(\boldsymbol{\eta};\, \boldsymbol{x},\, 1)\, \big).$$

Recall that the factor $b(\boldsymbol{\eta})$ in the prior density (3.1) is discarded when deriving the standardized posterior mode (3.2). Since we may look upon the sampling density $p(\boldsymbol{x};\, \boldsymbol{\mu}) = \exp\{-d(\boldsymbol{x},\, \boldsymbol{\mu})\}\, a(\boldsymbol{x})$ as the prior density $\pi(\boldsymbol{\eta};\, \boldsymbol{x},\, 1)$, the functional $G[\pi(\boldsymbol{\eta})]$ can be regarded as the information contained in the prior density $\pi(\boldsymbol{\eta})$. The following corollary gives the minimizer of $G[\pi(\boldsymbol{\eta})]$.

Corollary 3.3.
The conjugate prior density (3.1) *minimizes the functional* $G[\pi(\boldsymbol{\eta})] = KL\big(\, \pi(\boldsymbol{\eta}|\boldsymbol{x}),\, \pi(\boldsymbol{\eta};\, \boldsymbol{x},\, 1)\, \big)$ *among the family* $\mathcal{P}(\boldsymbol{x},\, \boldsymbol{m},\, \delta)$ *of prior densities defined by the condition* (3.6).

Proof. Set $\boldsymbol{m}_1 = \boldsymbol{x}$ and $\delta_1 = 1$ in Theorem 3.2, and we have

$$G[\pi(\boldsymbol{\eta})] = G[\pi(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta)] + \mathrm{KL}\big(\,\pi(\boldsymbol{\eta}|\boldsymbol{x}),\, \pi(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)\,\big).$$

This equality completes the proof. □

Note that this corollary is closely related to discussions on the minimax property of the conjugate prior density employed by Morris (1983) and Consonni and Veronese (1992).

We close this section by emphasizing to a potential relation between the conjugate analysis and the generalized linear model (GLM). Conjugate priors for the GLM were studied by Chen and Ibrahim (2003). The GLM is based on the sampling density $p(x;\, \mu)$ with mean μ in the one-parameter exponential family. It is known to hold that $\log\{p(x;\, \hat{\mu}_{\mathrm{ML}})/p(x;\, \mu)\} = \mathrm{KL}\big(p(y;\, \hat{\mu}_{\mathrm{ML}}),\, p(y;\, \mu)\big)$ where $\hat{\mu}_{\mathrm{ML}} = x$ is the maximum likelihood estimator. This is formally rewritten as

$$\mathrm{KL}\big(\delta(y - \hat{\mu}_{\mathrm{ML}}),\, p(y;\, \mu)\big) = \mathrm{KL}\big(\delta(y - \hat{\mu}_{\mathrm{ML}}),\, p(y;\, \hat{\mu}_{\mathrm{ML}})\big) + \mathrm{KL}\big(p(y;\, \hat{\mu}_{\mathrm{ML}}),\, p(y;\, \mu)\big),$$

where $\delta(y - x)$ is the Dirac's delta function. A similar Pythagorean relationship holds approximately in the GLM. Comparing with the Pythagorean relationship (3.7) in Theorem 3.2, we learn that a type of similarity lies between the conjugate analysis and the GLM.

4. A Pythagorean relationship

In this and the following sections the dual Pythagorean relationships are derived, each of which manifests how the standardized posterior mode dominates other estimators. The loss functions we adopt in the two cases are dual to each other. Assuming the two conjugate prior densities, or the two types of $b(\boldsymbol{\eta})$, we discuss the conjugate analysis separately.

First, we pursue an optimality of the estimator 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 $b_c(\boldsymbol{\eta})$ such that

$$\frac{\partial}{\partial \boldsymbol{m}} \int \exp\{-\delta_1 L(\boldsymbol{m},\, \boldsymbol{\mu})\}\, b_c(\boldsymbol{\eta})\, d\boldsymbol{\eta} = \mathbf{0}. \tag{4.1}$$

We set the integral in (4.1) as $\exp\{-K(\delta_1)\}$. The density $\exp\{-\delta_1 L(\boldsymbol{m},\, \boldsymbol{\mu}) + K(\delta_1)\}\, b_c(\boldsymbol{\eta})$ belongs to the proper dispersion model introduced in Jørgensen (1997, p.5). Setting $\delta_1 = \delta h_{p+1}(\boldsymbol{m})$, we assume the prior density

$$\pi_c(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta) = \exp\big\{-\delta d(\boldsymbol{m},\, \boldsymbol{\mu}) + K(\delta h_{p+1}(\boldsymbol{m}))\big\}\, b_c(\boldsymbol{\eta}). \tag{4.2}$$

It should be noted that the normalizing constant depends on $\boldsymbol{m}$ and δ only through the product $\delta h_{p+1}(\boldsymbol{m})$.

The conjugate prior density (4.2) has the following property with respect to the expectation of the extended canonical parameter.

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

$$\mathrm{E}\big[\boldsymbol{\eta} - \boldsymbol{\eta}(\boldsymbol{m}) \;\big|\; \pi_c(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta)\,\big] = \mathbf{0},$$

where $\boldsymbol{\eta}(\boldsymbol{m}) = -\big(f_1(\boldsymbol{m}), \ldots, f_p(\boldsymbol{m})\big)^T$. *Further, the posterior density corresponding to* $\pi_c(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta)$ *satisfies*

$$\mathrm{E}\big[\,\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_{smap} \;\big|\; \pi_c(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)\,\big] = \mathbf{0}.$$

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

$$\int \{\boldsymbol{\eta}(\boldsymbol{m}) - \boldsymbol{\eta}\} \exp\{-\delta_1 L(\boldsymbol{m},\, \boldsymbol{\mu})\}\, b_c(\boldsymbol{\eta})\, d\boldsymbol{\eta} = 0 \tag{4.3}$$

for any $\boldsymbol{m}$ and $\delta_1 > 0$. Setting $\delta_1 = \delta h_{p+1}(\boldsymbol{m})$, we obtain the former part.

Theorem 3.1 yields that the corresponding posterior density is expressed as $\pi_c(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$. Noting that $\hat{\boldsymbol{\eta}}_{smap} = \boldsymbol{\eta}(\hat{\boldsymbol{\mu}}_{smap})$, we see that the latter part follows from the former part. □

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. This extension is realized by introducing $\boldsymbol{\eta}$ suitably.

We clarify implications of Proposition 4.1 through the following example where the sampling density is in the natural exponential family (1.1).

Example 4.1. Set $f_j(\mu)$ and $h_j(x)$ $(i = 1, 2)$ in the natural exponential family (1.1) as in the former part of Example 2.1. Suppose that the assumption (4.1) is satisfied, that is, the normalizing constant in (4.2) depends only on δ. Then, the posterior mean of $\eta = \phi'(\mu)$ is $\phi'(\hat{\mu}_{smap})$ with $\hat{\mu}_{smap} = (x + \delta m)/(1 + \delta)$.

Next, we deal with the case where the sampling density is defined on $\mathbb{R}^+$ and set $f_j(\mu)$ and $h_j(x)$ $(i = 1, 2)$ as in the latter part of Example 2.1. The assumption (4.1) is equivalent to the one that the normalizing constant in (4.2) is a function of δm. Under this assumption the posterior mean of $\psi(\eta) = \psi(\phi'(\mu))$ is $\psi(\phi(\hat{\mu}_{smap}))$.

Now, let us derive a Pythagorean relationship with respect to posterior risks.

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

$$\mathrm{E}\big[\,L(\hat{\boldsymbol{\mu}},\, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}}_{smap},\, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}},\, \hat{\boldsymbol{\mu}}_{smap}) \;\big|\; \pi_c(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)\,\big] = 0 \tag{4.4}$$

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

Proof. It follows from the identity (2.9) that

$$L(\hat{\boldsymbol{\mu}},\, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}}_{smap},\, \boldsymbol{\mu}) - L(\hat{\boldsymbol{\mu}},\, \hat{\boldsymbol{\mu}}_{smap}) = \{\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\theta}}_{smap}\}^T (\hat{\boldsymbol{\eta}}_{smap} - \boldsymbol{\eta}). \tag{4.5}$$

Note that $\boldsymbol{\theta}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\theta}}_{smap}$ is constant in $\boldsymbol{\eta}$. Thus, the latter part of Proposition 4.1 yields the Pythagorean relationship (4.4). The optimum property of $\hat{\boldsymbol{\mu}}_{smap}$ follows from this Pythagorean relationship. □

We derive an extended version of the Pythagorean relationship in Proposition 4.2. This is done by modifying the loss function $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ for an appropriate choice of $b(\boldsymbol{\eta})$ in the prior density (3.1). Suppose that there exist a positive function $I(\boldsymbol{m})$ and a non-negative function $\tilde{b}_c(\boldsymbol{\eta})$ such that

$$\frac{\partial}{\partial \boldsymbol{m}} \int \exp\{-\delta_1 I(\boldsymbol{m}) L(\boldsymbol{m}, \boldsymbol{\mu})\}\, \tilde{b}_c(\boldsymbol{\eta})\, d\boldsymbol{\eta} = \mathbf{0}, \tag{4.6}$$

and we write the integral in (4.6) as $\exp\{-\tilde{K}(\delta_1)\}$. The assumption (4.6) is weaker than (4.1), since the former allows $I(\boldsymbol{m})$.

The prior density we assume on $\boldsymbol{\eta}$ is of the form

$$\tilde{\pi}_c(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta) = \exp\left\{ -\delta d(\boldsymbol{m}, \boldsymbol{\mu}) + \tilde{K}\left(\frac{\delta h_{p+1}(\boldsymbol{m})}{I(\boldsymbol{m})} \right) \right\} \tilde{b}_c(\boldsymbol{\eta}).$$

Theorem 3.1 means that the corresponding posterior density is expressed as $\tilde{\pi}_c(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$. A modified Pythagorean relationship is derived under the loss $I(\hat{\boldsymbol{\mu}})L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$. It should be noted that the posterior risk difference is expressed through the Kullback-Leibler separator between the two (prior) densities.

Proposition 4.3.
Under the assumption (4.6) *set* $\pi_I(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta_1) = \exp\{-\delta_1 I(\boldsymbol{m}) L(\boldsymbol{m}, \boldsymbol{\mu}) + \tilde{K}(\delta_1)\}\, \tilde{b}_c(\boldsymbol{\eta})$. *The following modified Pythagorean relationship*

$$\begin{aligned} &\mathrm{E}\big[\, I(\hat{\boldsymbol{\mu}}) L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) - I(\hat{\boldsymbol{\mu}}_{smap}) L(\hat{\boldsymbol{\mu}}_{smap}, \boldsymbol{\mu}) \;\big|\; \tilde{\pi}_c(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*) \,\big] \\ &= \frac{1}{\delta_1^*} \mathrm{KL}\big(\pi_I(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta_1^*),\, \pi_I(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}},\, \delta_1^*) \big) \end{aligned} \tag{4.7}$$

holds for any estimator $\hat{\boldsymbol{\mu}}$ *where* $\delta_1^* = \{h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})\}/I(\hat{\boldsymbol{\mu}}_{smap})$. *Consequently, the standardized posterior mode* $\hat{\boldsymbol{\mu}}_{smap}$ *is optimum under the loss* $I(\hat{\boldsymbol{\mu}})L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$.

Proof. A calculation of the right-hand side of (4.7) gives

$$\begin{aligned} &\frac{1}{\delta_1^*} \mathrm{KL}\big(\pi_I(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta_1^*),\, \pi_I(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}},\, \delta_1^*) \big) \\ &= \mathrm{E}\big[\, I(\hat{\boldsymbol{\mu}}) L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) - I(\hat{\boldsymbol{\mu}}_{smap}) L(\hat{\boldsymbol{\mu}}_{smap}, \boldsymbol{\mu}) \;\big|\; \pi_I(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta_1^*) \,\big] \end{aligned}$$

The equality (2.10) and the expression (3.4) of δ^*, together with the expression of δ_1^* in Proposition 5.1, give

$$\begin{aligned} \delta_1^* I(\hat{\boldsymbol{\mu}}_{smap}) L(\hat{\boldsymbol{\mu}}_{smap}, \boldsymbol{\mu}) &= \delta^*\, d(\hat{\boldsymbol{\mu}}_{smap}, \boldsymbol{\mu}), \\ \tilde{K}(\delta_1^*) &= \tilde{K}\left(\frac{\delta^* h_{p+1}(\hat{\boldsymbol{\mu}}_{smap})}{I(\hat{\boldsymbol{\mu}}_{smap})} \right). \end{aligned}$$

Thus, we see that the posterior density $\tilde{\pi}_c(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$ is equal to $\pi_I(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta_1^*)$, which completes the proof. □

Another expression of the term $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{smap})$ in Proposition 4.2 is obtained as

$$L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{smap}) = \frac{1}{\delta_1^*} \mathrm{KL}\big(\pi_1(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta_1^*),\, \pi_1(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}},\, \delta_1^*) \big)$$

where $\pi_1(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta_1) = \exp\{-\delta_1 L(\boldsymbol{m},\, \boldsymbol{\mu}) + K(\delta_1)\}\, b_e(\boldsymbol{\eta})$ and $\delta_1^* = h_{p+1}(\boldsymbol{x}) + \delta h_{p+1}(\boldsymbol{m})$.

The hyperbola density (2.12) provides us with an illustrative example of Proposition 4.3, where a modified loss function $I(\hat{\mu})L(\hat{\mu},\, \mu)$ is more familiar than the original one $L(\hat{\mu},\, \mu)$.

Example 4.2. The dual convex functions are $\psi(\eta) = \cosh(\sinh^{-1}\eta)$ and $\phi(\theta) = \theta \sinh(\tanh^{-1}\theta) - \cosh(\tanh^{-1}\theta)$ in the hyperbola density $p_{\mathrm{H}}(x;\, \mu,\, \tau)$ in (2.12). Thus, the loss function $L(\hat{\mu},\, \mu)$ is of the form $L(\hat{\mu},\, \mu) = \{\cosh(\hat{\mu} - \mu) - 1\}/\cosh\hat{\mu}$. A familiar loss function in the literature is $I(\hat{\mu})L(\hat{\mu},\, \mu) = \cosh(\hat{\mu} - \mu) - 1$, which is obtained by setting $I(\mu) = \cosh\mu$. If we choose $b(\eta)$ as $\tilde{b}_e(\eta) = d\mu/d\eta = 1/\cosh(\sinh^{-1}\eta)$, then the integral

$$\int_{-\infty}^{\infty} \exp\{-\delta_1 I(m) L(m,\, \mu)\}\tilde{b}_e(\eta)\, d\eta = \int_{-\infty}^{\infty} \exp\{-\delta_1 \cosh(m - \mu)\}\, d\mu$$

is independent of m. Note that the Kullback-Leibler separator from $p_{\mathrm{H}}(\mu;\, m_1,\, \delta)$ to $p_{\mathrm{H}}(\mu;\, m_2,\, \delta)$ is calculated as

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

For an arbitrary estimator $\hat{\mu}$ Proposition 4.3 gives the following modified Pythagorean relationship

$$\mathrm{E}\big[\, \cosh(\hat{\mu} - \mu) - \cosh(\hat{\mu}_{smap} - \mu) \;\big|\; p_{\mathrm{H}}(\mu;\, \hat{\mu}_{smap},\, \delta_1^*)\, \big] = \frac{1}{\delta_1^*}\mathrm{KL}\big((\hat{\mu}_{smap},\, \delta_1^*),\, (\hat{\mu},\, \delta_1^*)\big),$$

where $\tanh\hat{\mu}_{smap} = \{\tau \sinh x + \delta \sinh m\}/\{\tau \cosh x + \delta \cosh m\}$ and $\delta_1^* = \{\tau^2 + \delta^2 + 2\tau\delta \cosh(x - m)\}^{1/2}$.

5. A dual version of the Pythagorean relationship

We move to the case of an alternative loss function $L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}})$, dual to $L(\hat{\boldsymbol{\mu}},\, \boldsymbol{\mu})$. Another conjugate prior density which is in a sense dual to $\pi_e(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta)$ in (4.2) is dealt with. Setting $b(\boldsymbol{\eta}) = 1$, we assume the prior density

$$\pi_m(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta) \propto \exp\{-\delta d(\boldsymbol{m},\, \boldsymbol{\mu})\} \tag{5.1}$$

with respect to the Lebesgue measure on $\boldsymbol{\eta}$. When the sampling density is in the regular natural exponential family, this prior density reduces to what is called the DY prior density.

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

$$\lim_{\eta_j \to \overline{\eta}_j} d(\boldsymbol{m},\, \boldsymbol{\mu}) = \infty \quad \text{and} \quad \lim_{\eta_j \to \underline{\eta}_j} d(\boldsymbol{m},\, \boldsymbol{\mu}) = \infty \quad \text{for } j = 1, \ldots, p. \tag{5.2}$$

In the above $\overline{\eta}_j = \overline{\eta}_j(\boldsymbol{\eta}_{(j)})$ and $\underline{\eta}_j = \underline{\eta}_j(\boldsymbol{\eta}_{(j)})$ are respectively the upper and the lower boundary point when $\boldsymbol{\eta}_{(j)} = (\eta_1, \ldots, \eta_{j-1}, \eta_{j+1}, \ldots, \eta_p)^T$ are fixed. Roughly speaking, this assumption implies that the density vanishes at the boundary. The following proposition claims that the prior density (5.1) has a property dual to the one in Proposition 4.1.

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

$$\mathrm{E}\big[\, \boldsymbol{\theta} - \boldsymbol{\theta}(\boldsymbol{m}) \;\big|\; \pi_m(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta)\, \big] = \mathbf{0}.$$

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

$$\mathrm{E}\big[\, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{smap} \;\big|\; \pi_m(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*) \,\big] = \mathbf{0}.$$

Proof. It follows from (5.2) that

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

for $j = 1, \ldots, p$. We have from (2.6) and (A.4)

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

Combining these, we obtain the former part.

The proof of the latter part is parallel to that of the latter part of Proposition 4.1. □

Now, let us derive a Pythagorean relationship with respect to the loss function $L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}}) = d(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}})/h_{p+1}(\boldsymbol{\mu})$. Note that the loss function and the property of the prior density are dual to those in the previous Pythagorean relationship (4.4).

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

$$\mathrm{E}\big[\, L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}}) - L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}}_{smap}) - L(\hat{\boldsymbol{\mu}}_{smap},\, \hat{\boldsymbol{\mu}}) \;\big|\; \pi_m(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*) \,\big] = 0 \tag{5.3}$$

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

Proof. The proof is parallel to that of Proposition 4.2. Instead of the identity (4.5), we use

$$L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}}) - L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}}_{smap}) - L(\hat{\boldsymbol{\mu}}_{smap},\, \hat{\boldsymbol{\mu}}) = \big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{smap}\big)^T \big(\hat{\boldsymbol{\eta}}_{smap} - \hat{\boldsymbol{\eta}}\big),$$

where $\hat{\boldsymbol{\eta}}$ is the estimator equivalent to $\hat{\boldsymbol{\mu}}$. □

Next, a modification of the Pythagorean relationship (5.3) is dealt with. 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

$$\tilde{\pi}_m(\boldsymbol{\eta};\, \boldsymbol{m},\, \delta) \propto \exp\{-\delta d(\boldsymbol{m},\, \boldsymbol{\mu})\}/J(\boldsymbol{\eta}). \tag{5.4}$$

It follows from Theorem 3.1 that the above prior density is also conjugate, and also that the posterior density is given as $\tilde{\pi}_m(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$. Here again we assume the regularity condition (5.2). We learn that a modified Pythagorean relationship holds under the loss $J(\boldsymbol{\eta})L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}})$. Note that the third term in the posterior expectation in the following proposition is not $J(\hat{\boldsymbol{\eta}}_{smap})L(\hat{\boldsymbol{\mu}}_{smap},\, \hat{\boldsymbol{\mu}})$, but $J(\boldsymbol{\eta})L(\hat{\boldsymbol{\mu}}_{smap},\, \hat{\boldsymbol{\mu}})$.

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

$$\mathrm{E}\big[\, J(\boldsymbol{\eta})L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}}) - J(\boldsymbol{\eta})L(\boldsymbol{\mu},\, \hat{\boldsymbol{\mu}}_{smap}) - J(\boldsymbol{\eta})L(\hat{\boldsymbol{\mu}}_{smap},\, \hat{\boldsymbol{\mu}}) \;\big|\; \tilde{\pi}_m(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)\,\big] = 0 \qquad (5.5)$$

holds for any estimator $\hat{\boldsymbol{\mu}}$. *Thus, the standardized posterior mode* $\hat{\boldsymbol{\mu}}_{smap}$ *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 $J(\boldsymbol{\eta})\,\tilde{\pi}_m(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*) \propto \pi_m(\boldsymbol{\eta};\, \hat{\boldsymbol{\mu}}_{smap},\, \delta^*)$ as functions of $\boldsymbol{\eta}$. The modified Pythagorean relationship (5.5) is a rewritten version of the original one (5.3). □

Interestingly, the standardized posterior mode is optimum for all the loss functions in Propositions 4.2, 4.3, 5.2 and 5.3.

Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^T$ be a new parameter vector 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 $\exp\{-\delta d(\boldsymbol{m},\, \boldsymbol{\mu})\}$ with respect to the Lebesgue measure on $\boldsymbol{\xi}$. Especially when the sampling density is in the exponential family, this prior density is called standard conjugate by Consonni and Veronese (1992). The prior density is equivalent to (5.4) with $1/J(\boldsymbol{\eta}) = \left|\, \partial\boldsymbol{\xi}/\partial\boldsymbol{\eta} \,\right|$.

The following example gives implications of Propositions 5.2 and 5.3 to the natural exponential family (1.1).

Example 5.1. Let us assume that the natural exponential family (1.1) is regular, i.e., that its canonical space is assumed to be open. This assumption implies that

$$\lim_{\eta \to \overline{\eta}} \mathrm{KL}(m,\, \mu) = \infty \quad \text{and} \quad \lim_{\eta \to \underline{\eta}} \mathrm{KL}(m,\, \mu) = \infty,$$

where $\mathrm{KL}(\mu_1,\, \mu_2)$ is the Kullback-Leibler separator from $p(x;\, \mu_1)$ to $p(x;\, \mu_2)$. Thus, the assumption (5.2) is satisfied. It is known that the DY prior density exists for a regular natural exponential family. It is of the form

$$\pi_m(\eta;\, m,\, \delta) = \pi_{\mathrm{DY}}(\eta;\, m,\, \delta) \propto \exp\{-\delta \mathrm{KL}(m,\, \mu)\}$$

with respect to the Lebesgue measure on η. Then, the standardized posterior mode $\hat{\mu}_{smap} = (x + \delta m)/(1 + \delta)$ is optimum with respect to the loss $\mathrm{KL}(\mu,\, \hat{\mu})$.

Next, we introduce a new parameter $\xi = \xi(\eta) = \xi(\phi'(\mu))$, and consider the prior density

$$\tilde{\pi}_m(\eta;\, m,\, \delta) \propto \exp\{-\delta \mathrm{KL}(m,\, \mu)\} \left|\, \frac{d\xi}{d\eta} \,\right|.$$

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 the function $v(\mu)$ denotes the variance function.

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	$\mathrm{KL}(\mu, \hat{\mu})$	
μ	$\dfrac{\mathrm{KL}(\mu, \hat{\mu})}{v(\mu)}$	
$\log \mu$	$\dfrac{\mu \mathrm{KL}(\mu, \hat{\mu})}{v(\mu)}$	$\mu > 0$
$\psi(\eta)$	$\dfrac{\mathrm{KL}(\mu, \hat{\mu})}{\mu}$	$\mu > 0$
$\phi(\mu)$	$\dfrac{\mathrm{KL}(\mu, \hat{\mu})}{\eta v(\mu)}$	$\eta > 0$
$\log \eta$	$\eta \mathrm{KL}(\mu, \hat{\mu})$	$\eta > 0$

6. Examination of the non-singularity condition

The aim of this section is to make regularity conditions weaker. Our discussions in Sections 2 through 5 were based on the non-singularity condition (C.4). However, the conjugate analysis is possible without this regularity condition to some extent. An example is the von Mises distribution, the conjugate analysis of which was studied by Mardia and El-Atoum (1976).

Let $F_{p,p+1}(\boldsymbol{t})$ denote the $p \times (p+1)$ matrix whose (i,j)th component is $\partial f_j(\boldsymbol{t})/\partial t_i$ $(1 \leq i \leq p,\ 1 \leq j \leq p+1)$. In place of (C.4) requiring the non-singularity of $F_{p,p}(\boldsymbol{t})$ and (C.5), we here assume the following regularity condition

(C.4′) $\quad \mathrm{rank}\, F_{p,p+1}(\boldsymbol{t}) = p$ for any $\boldsymbol{t}$.

In order to make the difference between (C.4) and (C.4′) clear, we consider the von Mises case. Whether we set $f_1(t) = -\cos t$ or $f_1(t) = -\sin t$, the condition (C.4) is not satisfied. However, the rank of the 1×2 matrix $(\sin t, -\cos t)$ is equal to one for any t, that is, (C.4′) is satisfied.

Since it seems difficult to define the extended canonical parameter, we assume prior densities on the parameter $\boldsymbol{\mu}$. The assumed prior density has the form

$$\pi(\boldsymbol{\mu};\, \boldsymbol{m},\, \delta) \propto \exp\{-\delta d(\boldsymbol{m},\, \boldsymbol{\mu})\}\, c(\boldsymbol{\mu}), \tag{6.1}$$

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* (6.1) *is conjugate.*

Proof. The proof is similar to that of Theorem 3.1. We prove that the right-hand side of (3.5) is proportional to $d(\hat{\boldsymbol{\mu}}_{smap},\, \boldsymbol{\mu})$. It suffices to show that the two vectors $\tilde{\boldsymbol{h}}(\boldsymbol{x}) + \delta\tilde{\boldsymbol{h}}(\boldsymbol{m})$ and $\tilde{\boldsymbol{h}}(\hat{\boldsymbol{\mu}}_{smap})$ are proportional where $\tilde{\boldsymbol{h}}(\boldsymbol{t})$ denote the $(p+1)$-dimensional vector $\big(h_1(\boldsymbol{t}), \ldots, h_{p+1}(\boldsymbol{t})\big)^T$. By definition, the standardized posterior mode $\hat{\boldsymbol{\mu}}_{smap}$ satisfies

$(\partial/\partial\boldsymbol{\mu})\{d(\boldsymbol{x}, \boldsymbol{\mu}) + \delta d(\boldsymbol{m}, \boldsymbol{\mu})\} \big|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}_{smap}} = \mathbf{0}$. This is expressed in a matrix representation as $F_{p,p+1}(\hat{\boldsymbol{\mu}}_{smap}) \{\tilde{\boldsymbol{h}}(\boldsymbol{x}) + \delta\tilde{\boldsymbol{h}}(\boldsymbol{m})\} = \mathbf{0}$. The equality (2.3) with $\boldsymbol{a} = \hat{\boldsymbol{\mu}}_{smap}$ is rewritten as $F_{p,p+1}(\hat{\boldsymbol{\mu}}_{smap})\, \tilde{\boldsymbol{h}}(\hat{\boldsymbol{\mu}}_{smap}) = \mathbf{0}$. Note that the matrix $F_{p,p+1}(\hat{\boldsymbol{\mu}}_{smap})$ is of full rank. It follows from the theory of linear algebra that there exists δ^* such that

$$\tilde{\boldsymbol{h}}(\boldsymbol{x}) + \delta\tilde{\boldsymbol{h}}(\boldsymbol{m}) = \delta^*\, \tilde{\boldsymbol{h}}(\hat{\boldsymbol{\mu}}_{smap}). \tag{6.2}$$

Thus, the desired proportionality

$$d(\boldsymbol{x}, \boldsymbol{\mu}) + \delta d(\boldsymbol{m}, \boldsymbol{\mu}) - d(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_{smap}) - \delta d(\boldsymbol{m}, \hat{\boldsymbol{\mu}}_{smap}) = \delta^* d(\hat{\boldsymbol{\mu}}_{smap}, \boldsymbol{\mu})$$

is obtained. The existence assumption of $\hat{\boldsymbol{\mu}}_{smap}$ guarantees that $\delta^* > 0$. Thus, we see that the posterior density is expressed as $\pi(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_{smap}, \delta^*)$. □

Discussions similar to those in Propositions 4.2 and 4.3 hold true under the weaker regularity condition (C.4′) in place of (C.4) and (C.5). We assume the following prior density

$$\pi_0(\boldsymbol{\mu}; \boldsymbol{m}, \delta) \propto \exp\{-\delta d(\boldsymbol{m}, \boldsymbol{\mu})\}\, c_0(\boldsymbol{\mu})$$

under the assumption that there exist a positive function $\tilde{I}(\boldsymbol{m})$ and a non-negative function $c_0(\boldsymbol{\mu})$ such that

$$\frac{\partial}{\partial \boldsymbol{m}} \int \exp\{-\delta_2 \tilde{I}(\boldsymbol{m}) d(\boldsymbol{m}, \boldsymbol{\mu})\}\, c_0(\boldsymbol{\mu})\, d\boldsymbol{\mu} = \mathbf{0}. \tag{6.3}$$

Proposition 6.2.
Under the assumption (6.3) *set* $\tilde{\pi}_0(\boldsymbol{\mu}; \boldsymbol{m}, \delta_2) \propto \exp\{-\delta_2 \tilde{I}(\boldsymbol{m}) d(\boldsymbol{m}, \boldsymbol{\mu})\}\, c_0(\boldsymbol{\mu})$. *The following modified Pythagorean relationship*

$$\begin{aligned} &\mathrm{E}\big[\, \tilde{I}(\hat{\boldsymbol{\mu}}) d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) - \tilde{I}(\hat{\boldsymbol{\mu}}_{smap}) d(\hat{\boldsymbol{\mu}}_{smap}, \boldsymbol{\mu}) \;\big|\; \pi_0(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_{smap}, \delta^*) \,\big] \\ &= \frac{1}{\delta_2^*} \mathrm{KL}\big(\tilde{\pi}_0(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_{smap}, \delta_2^*), \tilde{\pi}_0(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}, \delta_2^*) \big) \end{aligned}$$

holds for any estimator $\hat{\boldsymbol{\mu}}$ *where* δ^* *is the constant given in* (6.2) *and* $\delta_2^* = \delta^* / \tilde{I}(\hat{\boldsymbol{\mu}}_{smap})$. *Consequently, the standardized posterior mode* $\hat{\boldsymbol{\mu}}_{smap}$ *is optimum under the loss* $\tilde{I}(\hat{\boldsymbol{\mu}}) d(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$.

Proof. The proof is parallel to that of Proposition 4.3. The key is the equality $\pi_0(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_{smap}, \delta^*) = \tilde{\pi}_0(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}_{smap}, \delta_2^*)$. □

Here we investigate the von Mises case in order to explain the above proposition.

Example 6.1. Consider the von Mises density $p_{\mathrm{vM}}(x; \mu, \tau)$ in (1.2). If we set $\tilde{I}(m) = 1$ and $c_0(\mu) = 1$, the integral

$$\int_0^{2\pi} \exp\{-\delta_2 \tilde{I}(m) d(m, \mu)\} c_0(\mu)\, d\mu = \int_0^{2\pi} \exp\big[-\delta_2\{1 - \cos(m - \mu)\}\big]\, d\mu$$

is independent of m. Since the condition (6.3) is satisfied, we can apply Proposition 6.2. We obtain the following modified Pythagorean relationship

$$\mathrm{E}\big[\cos(\hat{\mu}_{smap}-\mu)-\cos(\hat{\mu}-\mu)\;\big|\;p_{\mathrm{vM}}(\mu;\,\hat{\mu}_{smap},\,\delta_2^*)\big]=\frac{1}{\delta_2^*}\frac{I_1(\delta_2^*)}{I_0(\delta_2^*)}\{1-\cos(\hat{\mu}-\hat{\mu}_{smap})\},$$

where $\hat{\mu}_{smap}=\arg\max_\mu\{\tau\cos(x-\mu)+\delta\cos(m-\mu)\}$ and $\delta_2^*=\{\tau^2+\delta^2+2\tau\delta\cos(x-m)\}^{1/2}$. This result is to be compared with Example 4.2.

Although we succeed in extending Propositions 4.2 and 4.3, it seems difficult to develop the arguments parallel to those in Propositions 5.2 and 5.3. This is due to severity in defining the extended canonical parameter without the regularity condition (C.4).

References

Amari, S.-I. & Nagaoka, H. (2000). *Methods of information geometry.* American Mathematical Society.

Bagchi, P. (1994). Empirical Bayes estimation in directional data. *J. Appl. Stat.* **21**, 317–326.

Barndorff-Nielsen, O. E. (1978a). *Information and exponential families in statistical theory.* J. Wiley & Sons, New York.

Barndorff-Nielsen, O. (1978b). Hyperbolic distribution and distribution on hyperbolae. *Scand. J. Statist.* **5**, 151–157.

Chen, M.-H. & Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statist. Sinica*, **13**, 461–476.

Consonni, G. & Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions. *J. Amer. Statist. Assoc.* **87**, 1123–1127.

Consonni, G. & Veronese, P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors. *Scand. J. Statist.* **28**, 377–406.

Diaconis, P. & Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–281.

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. In *Bayesian statistics* **4** (eds. J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith), 669–674, Oxford University Press, Oxford.

Gutiérrez-Peña, E. (1997). Moments for the canonical parameter of an exponential family under a conjugate distribution. *Biometrika* **84**, 727–732.

Gutiérrez-Peña, E. & Smith, A. F. M. (1997). Exponential and Bayesian conjugate families: Review and extensions (with discussion). *Test* **6**, 1–90.

Guttorp, P. & Lockhart, R. A. (1988). Finding the location of a signal: A Bayesian analysis. *J. Amer. Statist. Assoc.* **83**, 322–330.

Ibrahim, J. G. & Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazard model. *Sankhyā Ser. B* **60**, 48–64.

Ibrahim, J. G. & Chen, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15**, 46–60.

Jensen, J. L. (1981). On the hyperboloid distribution. *Scand. J. Statist.* **8**, 193–206.

Jørgensen, B. (1997). *The theory of dispersion models.* Chapman and Hall, London.

Mardia, K. V. & 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. *Ann. Statist.* **11**, 515–529.

Raiffa, H. & Schlaifer, R. (1961). *Applied statistical decision theory.* Graduate School of Business Administration, Harvard Univ., Boston.

Rodrigues, J., Leite, J. G. & Milan, L. A. (2000). An empirical Bayes inference for the von Mises distribution. *Aust. N. Z. J. Stat.* **42**, 433–440.

Yanagimoto, T. & Ohnishi, T. (2005a). Extensions of a conjugate prior through the Kullback-Leibler separators, *J. Multivariate Anal.* **92**, 116–133.

Yanagimoto, T. & Ohnishi, T. (2005b). Standardized posterior mode for the flexible use of a conjugate prior, *J. Statist. Plann. Inference* **131**, 253–269.

Appendix

Proofs of Lemmas 2.1 and 2.2.
The chain rule for partial differentiation gives

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

and

$$\delta_{jl} = -\frac{\partial}{\partial \eta_j} f_l(\boldsymbol{\mu}) = -\sum_{k=1}^{p} \frac{\partial}{\partial \mu_k} f_l(\boldsymbol{\mu}) \frac{\partial \mu_k}{\partial \eta_j}, \tag{A.2}$$

where δ_{jk} is Kronecker's delta. It follows from the kth component of the equality (2.3) that

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

Combining (A.1), (A.2) and (A.3), we have

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

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

$$\mathrm{E}\left[h_j(\boldsymbol{x}) - \frac{h_j(\boldsymbol{\mu})}{h_{p+1}(\boldsymbol{\mu})} h_{p+1}(\boldsymbol{x}) \,\middle|\, p(\boldsymbol{x}; \boldsymbol{\mu}) \right] = 0, \tag{A.5}$$

which is the required result of Lemma 2.2.

Again, differentiating both sides of (A.5) with respect to η_k, we see that

$$\begin{aligned}
&\mathrm{E}\left[\, h_{p+1}(\boldsymbol{x}) \mid p(\boldsymbol{x};\, \boldsymbol{\mu}) \,\right] \frac{\partial^2}{\partial \eta_k \partial \eta_j} \psi(\boldsymbol{\eta}) \\
&= \mathrm{E}\left[\left\{ h_j(\boldsymbol{x}) - \frac{h_j(\boldsymbol{\mu})}{h_{p+1}(\boldsymbol{\mu})} h_{p+1}(\boldsymbol{x}) \right\} \left\{ h_k(\boldsymbol{x}) - \frac{h_k(\boldsymbol{\mu})}{h_{p+1}(\boldsymbol{\mu})} h_{p+1}(\boldsymbol{x}) \right\} \middle| \, p(\boldsymbol{x};\, \boldsymbol{\mu}) \right].
\end{aligned}$$

This implies the convexity of $\psi(\boldsymbol{\eta})$, which completes the proof of Lemma 2.1. □