Ann Inst Stat Math (2014) 66:789–809
DOI 10.1007/s10463-013-0421-1

Permissible boundary prior function as a virtually proper prior density

Takemi Yanagimoto · Toshio Ohnishi

Received: 4 September 2012 / Revised: 12 March 2013 / Published online: 18 August 2013

Abstract Regularity conditions for an improper prior function to be regarded as a virtually proper prior density are proposed, and their implications are discussed. The two regularity conditions require that a prior function is defined as a limit of a sequence of proper prior densities and also that the induced posterior density is derived as a smooth limit of the sequence of corresponding posterior densities. This approach is compared with the assumption of a degenerated prior density at an unknown point, which is familiar in the empirical Bayes method. The comparison study extends also to the assumption of an improper prior function discussed separately from any proper prior density. Properties and examples are presented to claim potential usefulness of the proposed notion.

Keywords Degenerated prior · Logarithmic divergence · Marginal density · Non-informative prior · Weakly informative prior

1 Introduction

Bayesian theory is in principle based on the assumption of a proper prior, and it is always desired to assume a proper prior, if it is possible. On the other hand, recent widespread applications of Bayesian methods to various fields require various extensions of an assumed prior. Such extensions make a Bayesian method to meet practical

T. Yanagimoto (✉)
Department of Industrial and Systems Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
e-mail: yanagmt@indsys.chuo-u.ac.jp

T. Ohnishi
Faculty of Economics, Kyushu University, 6-19-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

conditions in actual applications. It is our understanding that an improper prior function is expected to be related directly to a family of proper prior densities in Bayesian theory. In this view, an improper prior function is an alternative to a weakly informative proper prior density and is hoped to be regarded as a virtually proper prior density. Our aim here is to enhance the wide use of an improper function satisfying regularity conditions. Recently, McCullagh and Han (2011) discussed this subject from a different point of view and gave references.

An improper prior function appears in Bayesian analysis, when a suitable informative prior density is unavailable and a non-informative prior function becomes improper. A non-informative prior function is widely employed in practice, since it is not rare that our prior knowledge about a parameter is not enough to choose an informative prior density. Familiar non-informative prior functions include the reference prior Berger et al. (2009) and the Jeffreys prior Jeffreys (1961), which are elicited only from a family of sampling densities. When a non-informative prior density is employed, it is often emphasized to avoid an unnecessary subjective prior density. Our aim is completely different from this line. We pursue the relation with a family of proper prior densities, since an improper prior function is hoped to be closely related with a weakly informative prior density. We will find that the posterior density induced from an improper prior function and that induced form a suitably chosen proper prior density are close to each other. In such a case an improper prior function can be treated as an alternative to a proper prior density. It may be possible to assume an improper prior function within the framework of the traditional Bayesian method.

A competitor of an improper prior function in practice is to assume a degenerated prior density at an unknown point, which is widely employed in the conventional empirical Bayes method. A degenerated prior density is a probability density, but it contains a hyperparameter to be estimated. Examples of the use of such a prior density are seen in the smoothing method, see for example Wahba (1985) and Yanagimoto and Yanagimoto (1987) and lasso Tibshirani (1996). The unknown point represents a parameter in the sampling density and is treated as a hyperparameter in the empirical Bayes model. The parameter is estimated by maximizing the marginal likelihood. The use of the marginal likelihood was employed in the Lindley paradox (Lindley 1957), and it was formulated as familiar Bayesian factor for comparing multiple candidate Bayesian models, see Kass and Raftery (1995) for example. The marginal density makes sense under a degenerated prior density at an unknown point. This fact is probably a key reason why such a degenerated density is employed in the empirical Bayes method.

Our primary aim is to claim the preference of a combination of an improper prior function and the posterior density to that of a degenerated prior density at an unknown point and the marginal density. In this concern, we will raise critiques of the use of the marginal density to claim the important role of an improper prior function. Recall that a Bayesian method is based primarily on the posterior density induced from a proper prior density rather than the marginal density (see Aitkin 2009, Chapter 2) for this discussion. It seems to us that further developments of Bayesian methods will not be based on the marginal density but on the posterior density. Existing methods based on the posterior density include DIC in Spiegelhalter et al. (2002) and its modifications discussed in Plummer (2008) and Yanagimoto and Ohnishi (2009a).

To impose regularity conditions on an improper prior function, we consider first a family of proper prior densities. Then the improper prior function is defined as the limit of the sequence of proper prior densities multiplied by suitable constants. A suitable multiplier is necessary, since the formal limit of the original sequence does not converge weakly to an improper prior function. Next, we require a smoothness property of the family of posterior densities induced from the proper prior densities. A notion of permissibility was introduced in terms of the logarithmic divergence in Berger et al. (2009) to define a regularity condition on the reference prior function. We modify the notion and employ a stronger condition by taking into account of the dual structure of the logarithmic divergence.

To explain our idea specifically, let $\{p(x|\theta)|\,\theta \in \Theta\}$ be a family of sampling densities, and let $\mathcal{P} = \{\pi(\theta;\ g, c)|\, g, c \in (0, \infty)\}$ be a family of proper prior densities for θ. Denote by $\pi_N(\theta)$ the prior function taking the value 0 for every $\theta \in \Theta$, and suppose that $\pi(\theta;\ g, c)$ converges weakly to $\pi_N(\theta)$, that is, the sequence $\{\pi(\theta;\ g_n, c)\}$ converges weakly to $\pi_N(\theta)$ as g_n tends to 0. On the other hand, we write a degenerated density at a point c in terms of the Dirac function as $\delta_D(\theta - c)$, and assume that $\pi(\theta;\ g, c)$ converges in probability to $\delta_D(\theta - c)$ as g tends to ∞. We discuss regularity conditions on an improper prior function $b(\theta)$ so that it is interpreted as a boundary of the family at $g = 0$, and can be used as a substitute for $\pi(\theta;\ g, c)$ for a very small value of g. In other words, $b(\theta)$ can be regarded as a virtually proper prior density. In contrast, the competitor $\delta_D(\theta - \theta_h)$ causes various inconveniences, though it is superficially another boundary prior density. A reason is that θ_h is an unknown hyperparameter to be estimated, and the other is that the posterior density is the same as the prior density. Consequently, no strong relation is observed between $\delta_D(\theta - \theta_h)$ and $\mathcal{P}$. Our problem is to compare $b(\theta)$ with $\delta_D(\theta - \theta_h)$ in the context of Bayesian inference.

To reconfirm the important role of an improper function, we present the following example of a familiar improper function for parameters in the normal sampling density. It is usually regarded as a non-informative prior and yields a reasonable estimate. In contrast, we find that a degenerated prior density at an unknown point is less appealing.

Example 1 (Normal sampling density) Consider the normal sampling density $p(x;\ \mu, 1/\tau)$ with mean μ and reciprocal variance τ, which consist of θ. Let a sample of size n be $\boldsymbol{x} = (x_1, \cdots, x_n)$, and set $s^2 = \sum(x_i - \bar{x})^2/(n-1)$. The following prior function and density are of our primary concern here: one is

$$b(\mu, \tau) \propto \frac{1}{\tau}. \tag{1}$$

and the other is $\delta_D(\theta - \theta_h) = \delta_D(\mu - \mu_h)\delta_D(\tau - \tau_h)$. The former function is known as the reference prior. Both are used, when a suitable proper prior is unavailable because of the lack of our prior knowledge about θ.

The posterior density is formally induced from $b(\mu, \tau)$ in (1). The posterior means of μ and τ are written as $\hat{\mu} = \bar{x}$ and $\hat{\tau} = 1/s^2$, respectively. The estimate $\hat{\tau}$ is equivalent to the conditional maximum likelihood estimate of τ given $\bar{x}$ in the frequentist context, which is regarded as a reasonable estimate. The posterior density induced

from $\delta_D(\theta - \theta_h)$ is the same as the prior density. The marginal density is written in the same form as the sampling density with mean μ_h and reciprocal variance τ_h. The maximum marginal likelihood estimate is the same as the crude maximum likelihood estimate in the frequentist context. It is known for $n \geq 4$ that the conditional maximum likelihood estimator $(\hat{\mu}, \hat{\tau})$ is superior to the maximum likelihood estimator $(\hat{\mu}_M, \hat{\tau}_M) = (\hat{\mu}, n\hat{\tau}/(n-1))$ under the loss based on the logarithmic divergence in the frequentist context Yanagimoto (1991). This means that $b(\mu, \tau)$ in (1) yields a more favorable estimate than $\delta_D(\mu - \mu_h)\delta_D(\tau - \tau_h)$ in the normal model. This example will be discussed further in Sect. 6.

The present paper is constructed as follows: a permissible boundary prior function is defined, and its preliminary properties and examples are presented in Sect. 2, which is followed by basic properties and two examples of general families in Sect. 3. Sections 4 and 5 are devoted to showing restrictive role of the marginal density to indicate the important role of procedures derived from the posterior density. Further examples of various families of prior densities are presented in Sect. 6. In the final section, supplemental explanations of examples are given and a recommendation is discussed.

2 Definition and preliminaries

Consider a family of sampling densities on $\mathcal{X}$, $\mathcal{M} = \{p(x|\theta)|\,\theta \in \Theta\}$. Let $\mathcal{G}(\subset R^+)$ be an open connected space and assume that its closure contains the origin 0. Consider also a family of proper prior densities indexed by $\mathcal{G}$, $\mathcal{P} = \{\pi(\theta;\, g)|\, g \in \mathcal{G}\}$. To focus our attentions on the main problems, we employ the following strong regularity conditions on the sampling and the prior densities: throughout this paper, we will assume that all the sampling densities in $\mathcal{M}$ have the common support $\mathcal{X}$ and also that they satisfy usual smoothness conditions with respect to x and θ. Further, $p(x|\theta)$ is assumed to be bounded as a function of θ for every x. Thus when a prior density is proper, the posterior density exists. Further, we will assume that a proper prior density in $\mathcal{P}$ takes a positive value for the interior of Θ, and the function $\pi(\theta;\, g)$ varies smoothly with θ and g. For notational simplicity, we will treat scalar cases of an observation and a parameter, unless stated otherwise. Straightforward extensions are possible to vector cases. The sequence of functions $\{\pi_n(\theta)\}$ is called to converge weakly to a function $\pi(\theta)$ if for an arbitrary bounded continuous function $f(\theta)$ the integral $\int f(\theta)\pi_n(\theta)\mathrm{d}\theta$ converges to $\int f(\theta)\pi(\theta)\mathrm{d}\theta$.

Let $b(\theta)$ be a σ-additive positive function on Θ. Though we allow that it is improper, that is, the integral of $b(\theta)$ over Θ is infinity, we assume that the integral of $p(x|\theta)b(\theta)$ over Θ, $m(x) = \int p(x|\theta)b(\theta)\mathrm{d}\theta$, exists for every x. This regularity condition yields a formal definition of a probability density

$$\pi_b(\theta|x) = \frac{p(x|\theta)b(\theta)}{m(x)}. \tag{2}$$

which can be treated as a posterior density induced from a prior function $b(\theta)$. Our aim is to discuss conditions that $b(\theta)$ is defined in a direct relation to a family of proper prior densities $\mathcal{P}$ and also that it is treated as $\pi(\theta; 0)$. We begin with giving a weaker condition, which is a formal statement of a regularity condition on an improper prior

function widely accepted in the literature, see Jaynes (2003) and McCullagh and Han (2011) for example.

Definition 1 A σ-additive function $b(\theta)$ is called the boundary prior function to $\mathcal{P}$, when for any sequence $\{g_n \in \mathcal{G};\ n = 1, 2\cdots\}$ tending to 0 there exists a sequence of positive numbers $\{a_n;\ n = 1, 2\cdots\}$ such that $\{a_n\pi(\theta;\ g_n)\}$ converges weakly to $b(\theta)$.

When $b(\theta)$ is a probability density, we may choose $a_n = 1$. Otherwise, we need to rescale the sequence $\{\pi(\theta;\ g_n)\}$ so that it converges weakly to $b(\theta)$. When $b(\theta)$ is improper, the sequence $\{a_n\}$ diverges to ∞ as n tends to ∞.

Set $\bar{\mathcal{G}} = \mathcal{G}\cup\{0\}$ and $\tilde{\mathcal{P}} = \mathcal{P}\cup\{b(\theta)\}$. For a fixed x denote by $\mathcal{D} = \{\pi(\theta|x;\, g)|\, g \in \mathcal{G}\}$ and $\bar{\mathcal{D}} = \{\pi(\theta|x;\, g)|\, g \in \bar{\mathcal{G}}\}$ the families of posterior densities induced from the family of the sampling densities $\mathcal{M} = \{p(x|\,\theta)|\,\theta \in \Theta\}$ and the families of the prior densities $\mathcal{P}$ and $\tilde{\mathcal{P}}$, respectively. Under suitable conditions it is expected that the family $\bar{\mathcal{D}}$ is a unified combination of the two subfamilies, $\mathcal{D}$ and $\{\pi(\theta|\,x;\, 0)\}$, but the family $\tilde{\mathcal{P}}$ is simply a union of the two separated subfamilies, $\mathcal{P}$ and $\{b(\theta)\}$. The point to be discussed here is to specify such regularity conditions. The unified nature of $\bar{\mathcal{D}}$ allows us to treat a boundary prior function $b(\theta)$ as a virtually proper prior density, so far as our inferential procedure is based on the posterior density. This approach will be compared with the use of a degenerated prior density at an unknown point.

To present a definition of such a suitable condition, we apply the logarithmic divergence between two probability densities, $D\big(\pi_1(\theta),\ \pi_2(\theta)\big)$, is given by $E\{\log\big(\pi_1(\theta)/\pi_2(\theta)\big);\ \pi_1(\theta)\}$, where $E\{f(\theta);\ \pi(\theta)\}$ denotes the expectation of $f(\theta)$ under a probability density $\pi(\theta)$. It is referred to also as the Kullback–Leibler separator and the relative entropy. A different notation $\kappa(\pi_1(\theta)|\,\pi_2(\theta))$ was employed in Berger et al. (2009), which is equivalent to $D(\pi_2(\theta),\ \pi_1(\theta))$. Inspired by Definition 9 in Berger et al. (2009), we explicitly define a smoothness property of the logarithmic divergence between posterior densities in $\bar{\mathcal{D}}$ with respect to g. The regularity condition we intend to impose on the family $\bar{\mathcal{D}}$ is the continuity of $D(\pi_1(\theta|\,x;\, g),\ \pi_2(\theta|\,x;\, g_0))$ and $D(\pi_1(\theta|\,x;\, g_0),\ \pi_2(\theta|\,x;\, g))$ for a fixed g_0.

Definition 2 Suppose that a function $b(\theta)$ is a boundary prior function to $\mathcal{P}$. It is called a permissible boundary prior function to $\mathcal{P}$ and for $\mathcal{M}$, if the following two conditions hold: (1) $\pi(\theta|\,x;\, 0)$ is the same as $\pi_b(\theta|x)$, and (2). For an arbitrarily fixed $g_0 \in \bar{\mathcal{G}}$ both the logarithmic divergences, $D(\pi(\theta|\,x;\, g),\ \pi(\theta|\,x;\, g_0))$ and $D(\pi(\theta|\,x;\, g_0),\ \pi(\theta|\,x;\, g))$, are continuous in $g \in \bar{\mathcal{G}}$.

The above definition requires that the posterior density $\pi(\theta|\,x;\, g)$ varies smoothly with g in $\bar{\mathcal{G}}$. Special attentions are paid to the case of $g_0 = 0$, when we check the regularity condition.

There are two differences between our definition of permissibility and that in Berger et al. (2009). One is that the divergence $D(\pi(\theta|\,x;\, g),\ \pi(\theta|\,x;\, g_0))$ and its dual divergence are both taken into account in Definition 2, and the other is that we do not take the expectation of the logarithmic divergence under the assumed model $p(x|\theta)\pi(\theta;\, g)$. Their definition pertains only to the reference prior $\pi_R(\theta)$, and they considered a family of strictly increasing compact subsets Θ_n converging to Θ, and

defined $\pi_n(\theta) = \pi_R(\theta) \cdot I(\theta|\,\Theta_n)$, where $I(\theta|\,\Theta_n)$ denotes the indicator function of Θ_n. Thus their interest focused only on $D(\pi_n(\theta),\ \pi_R(\theta))$. Another technical reason of their definition concerns the fact that the dual divergence $D(\pi_R(\theta), \pi_n(\theta))$ does not exist for every n. Next, we discuss reasons why the expectation is not taken. Apart from analytical simplicity, a primary reason is because a Bayesian procedure is in principle constructed for a given observation x, without taking the expectation over x. In this concern, they commented in their paper that "it might seem odd (to take the expectation)". Another reason pertains to the difference of the aims to introduce the notion of permissibility. Our aim is to pursue a relation of an improper prior function to a family of proper prior densities, and their aim is to justify the use of the reference prior when a family of sampling densities $\mathcal{M}$ is assumed.

To aid our understanding of the definition, we discuss a permissible boundary prior function to a general family of proper prior densities.

Example 2 (A general family) Consider a family of proper prior densities

$$\mathcal{P} = \{\pi(\theta; g, c) = \exp\{-gd(\theta, c)\}b(\theta) \cdot K(g, c)|\, g \in R^+\} \tag{3}$$

where $d(\theta, c)$ and $b(\theta)$ are a suitable distance (or divergence) between θ and c in Θ and an improper prior function, respectively. When g takes a moderate value, the primary term of (3) is in the exponent. Since the limit of $K(g, c)$ at $g = 0$ for a fixed c vanishes, $\pi(\theta; g, c)$ and $\pi(\theta; g, c)/K(g, c)$ converge weakly to $\pi_N(\theta)$ and $b(\theta)$, respectively. Thus $b(\theta)$ becomes a primary term of $\pi(\theta; g, c)$, when g is very small. Our definition of permissibility is designed for a condition that the function $b(\theta)$ can be treated in Bayesian inference as a substitute for a proper prior density $\pi(\theta; g_s)$ with g_s being a very small value in $\mathcal{G}$.

This example indicates that our definition is advantageous because of the direct relation of $b(\theta)$ to a family of proper prior densities $\mathcal{P}$. Recall that an improper prior function has been discussed only in relation to $\mathcal{M}$. It is often interpreted as a supporting function or as a Jacobian. The present approach allows us to choose $b(\theta)$ in the reverse way; we first explore a family of proper prior densities $\pi(\theta; g, c)$, and we assume an improper prior function as a secondary option when our prior knowledge about θ is not enough to specify g and c. This view suggests the possibility that we can choose a suitable choice of an improper prior function $b(\theta)$ by examining families $\mathcal{P}$ in (3) for various $b(\theta)$'s. A suggestion is that we choose $b(\theta)$ so that $K(g, c)$ is independent of c. This requirement yields that the parameters c and g in $\mathcal{P}$ are orthogonal. Note that c and g denote the center of our prior information on θ and the strength of our plausibility of the center, respectively. Thus the orthogonality condition on the components c and g is expected to be helpful for choosing a suitable proper prior density. To discuss this point in an explicit way, we return back to Example 1.

Example 1 (Continued) The family of normal-gamma prior densities for the normal sampling density is easy to be understood, and is conveniently tractable. Writing a normal-gamma prior density as

$$\pi(\mu, \tau;\ g_1, g_2, c_1, c_2) = \frac{\sqrt{g_1}}{\sqrt{2\pi}} \exp\left\{-\frac{g_1}{2}(\mu - c_1)^2\right\} \cdot \frac{1}{\Gamma(g_2)} c_2^{g_2} \tau^{g_2 - 1} \times \exp(-g_2 c_2 \tau) \tag{4}$$

we set $\mathcal{P} = \{\pi(\mu, \tau;\ g_1,\ g_2,\ c_1,\ c_2) | \, c_1 \in R,\ c_2,\ g_1 \text{ and } g_2 \in R^+\}$. Then the prior function $b(\mu, \tau)$ proportional to $1/\tau$ in (1) is a permissible boundary prior function to $\mathcal{P}$.

Our interest is placed also on a degenerated prior density at an unknown point θ_h, $\delta_D(\theta - \theta_h)$, which is treated as a reference. Similarly to an improper prior function, this prior density is employed when our prior information about θ is not enough to assume a proper prior density. Though the roles of these two priors are close to each other in practical applications, their theoretical properties are largely different. First, we note that the assumption of a degenerated prior density is theoretically equivalent to that of a family of degenerated prior densities

$$\mathcal{P}_D = \{\pi(\theta;\ \theta_h) = \delta_D(\theta - \theta_h) | \, \theta_h \in \Theta\}. \tag{5}$$

To pursue similarities and dissimilarities between $b(\theta)$ and $\delta_D(\theta - \theta_h)$, we introduce an enlarged family of $\mathcal{P}$

$$\mathcal{P}^E = \{\pi(\theta; g, \theta_h) = \exp\{-gd(\theta, \theta_h)\} b(\theta) \cdot K(g, \theta_h) | \, g \in R^+,\ \theta_h \in \Theta\}. \tag{6}$$

Then we can assert that $b(\theta)$ is a boundary prior function to $\mathcal{P}^E$ in the sense that for a fixed θ_h $K(g, \theta_h)\pi(\theta; g, \theta_h)$ converges weakly to $b(\theta)$, which is independent of θ_h. On the other hand, a degenerated density $\delta_D(\theta - \theta_h)$ can be regarded as another boundary prior density to $\mathcal{P}^E$ at the reverse side. In fact, $\pi(\theta;\ g, \theta_h)$ converges in probability to $\delta_D(\theta - \theta_h)$ as g tends to infinity. However, the limit contains an unknown hyperparameter θ_h. As a result, we cannot specify the sequence of proper prior densities $\{\pi(\theta;\ g_n, \theta_n)\}$ in the family $\mathcal{P}^E$ in (6) such that the sequence converges to $\delta_D(\theta - \theta_h)$, unless θ_h is known. This indicates that the assumption of a degenerated prior density lacks a close relation to that of a proper prior density.

Another serious defect of the assumption of a degenerated prior density is in the discontinuity in the posterior densities. The logarithmic divergence between a proper prior density and either $\pi_N(\theta)$ or $\delta_D(\theta - \theta_h)$ does not take a finite value. Recall that the logarithmic divergence between posterior densities in $\bar{\mathcal{D}}$ takes a finite value, when $b(\theta)$ is a permissible boundary prior function. On the other hand, the posterior density induced from a degenerated prior density is the same as the prior density, which yields that the logarithmic divergence between a posterior density in $\mathcal{D}$ and $\delta_D(\theta - \theta_h)$ does not take a finite value. Thus the permissibility condition is not satisfied. This fact indicates that the assumption of a degenerated prior density is inconvenient for constructing procedures based on the posterior density.

The following example shows that a familiar prior function induces a degenerated posterior density:

Example 3 (Poisson sampling density) Suppose that x is an observation from a Poisson sampling distribution with mean λ. A familiar prior function is of the form $b(\lambda;\ a) =$

λ^{a-1} for $a \geq 0$. The prior function is improper for every a. Two widely employed values of a in the literature (Bolstad 2007 for example) are 0.5 and 1, which are referred to as the Jeffreys or the reference prior and the uniform prior, respectively. The other familiar choice of a is 0. It yields the posterior mean equivalent to the maximum likelihood estimate and is interpreted as the uniform prior density for the canonical parameter $\log \lambda$.

The posterior density $\pi(\lambda|\, x;\ a)$ can be defined for $a > 0$ and $x \geq 0$. Write the gamma density on λ of the form $\pi(\lambda;\ g, a) = (ag)^a \lambda^{a-1} \exp(-ag\lambda)/\Gamma(a)$ as $\mathrm{Ga}(1/g,\ a)$ with $g > 0$. For a positive a we assume a gamma prior density with $\mathrm{Ga}(1/g,\ a)$ for λ. This conjugate prior density derives the posterior density with $\mathrm{Ga}\big((x+a)/(1+ag),\ x+a\big)$. It is easily shown that $b(\lambda; a)$ is the boundary prior function to $\mathcal{P} = \{\pi(\lambda;\ g, a)|\, g > 0\}$ for every a and also that $\pi(\lambda|\, x;\ g, a)$ converges weakly to $\pi(\lambda|\, x;\ a)$ as g tends to 0.

The situation is largely different in the case of $a = 0$. The posterior density $\pi(\lambda|\, x; 0)$ for $x \neq 0$ can be formally defined as the limit of $\pi(\lambda|\, x; g)$ as g tends to 0. However, careful treatments are required in the case of $x = 0$. The posterior density $\pi(\lambda|\, 0; 0)$ can be regarded as the limiting density of $\pi(\lambda|\, 0;\ a)$ at $a = 0$, which becomes $\delta_D(\lambda - 0)$. The logarithmic divergence $D\big(\pi(\lambda|\, 0; 0), \pi(\lambda|\, 0; a)\big)$, however, does not exist for every $a > 0$. Let the usual predictor $p(y|\, 0; a)$ be the posterior mean of the density $p(y|\, \lambda)$, where y denotes a future (or unobserved) variable. Then $p(y|\, 0; a)$ follows the negative binomial density for every $a > 0$, and the density $p(y|\, 0; 0)$ is degenerated at 0. These facts yield that the logarithmic divergence between $p(y|\, 0; 0)$ and $p(y|\, 0; a)$ does not exist for every $a > 0$.

Another point to be discussed pertains to the possibility that a formal posterior density is improper. An improper posterior function is more discouraging than an improper prior function. Speckman and Sun (2003) pointed out that improper posterior functions appear rather often in the empirical Bayes model. We give an example where the assumption of a degenerated prior density is associated with an improper posterior function. An additional example pertaining to the smoothing method will be given also in Example 11.

Example 4 (Improper posterior function) Let $\boldsymbol{y} \in R^p$ be an observation from the normal model $\mathrm{N}(\boldsymbol{\eta},\ (1/\tau)\mathrm{I})$. Assume that a prior density for $\boldsymbol{\eta}$, $\pi_3(\boldsymbol{\eta};\ \boldsymbol{\eta}_h)$, is given as $\delta_D(\boldsymbol{\eta} - \boldsymbol{\eta}_h)$ where $\boldsymbol{\eta}_h$ is a hyperparameter. Let $z \in R^1$ be another observation from the normal model $\mathrm{N}(\mu,\ 1/\tau)$. Assume that a prior density for μ, $\pi_1(\mu)$, follows $\mathrm{N}(m,\ 1/d)$, and also that a prior density for τ, $\pi_2(\tau; k)$, is expressed as

$$\pi_2(\tau; k) = \frac{k}{(1+\tau)^{k+1}} \tag{7}$$

with $p/2 > k > 0$. This distribution is referred to as the Pareto distribution of type II or the Lomax distribution. Write the sampling density of $x = (z, \boldsymbol{y}')$ as $p(z, \boldsymbol{y}|\mu, \tau, \boldsymbol{\eta})$, and the prior density for $(\mu, \boldsymbol{\eta}, \tau)$ as $\pi(\mu, \boldsymbol{\eta}, \tau; d, k, \boldsymbol{\eta}_h) = \pi_1(\mu; d)\pi_2(\tau; k)\pi_3(\boldsymbol{\eta}; \boldsymbol{\eta}_h)$. Then the posterior density of μ given $(x, \tau, \boldsymbol{\eta})$ follows $\mathrm{N}(\hat{\mu}, 1/(\tau+d))$ with $\hat{\mu} = (\tau z + dm)/(\tau + d)$. The posterior density of τ given $(x, \boldsymbol{\eta})$ is proportional to

$$\tau^{\frac{p+2}{2}} \frac{1}{\sqrt{\tau+d}} \exp\left\{-\frac{\tau}{2}\|\boldsymbol{y}-\boldsymbol{\eta}\|^2 - \frac{\tau d}{2(\tau+d)}(z-m)^2\right\} \cdot \frac{k}{(1+\tau)^{k+1}} \cdot \delta_D(\boldsymbol{\eta}-\boldsymbol{\eta}_h). \tag{8}$$

It is shown that the integral of (8) with respect to τ given $\boldsymbol{y}$ and $\boldsymbol{\eta}$ does not exist when $\boldsymbol{\eta} = \boldsymbol{y}$. Note that this exceptional case is important in the empirical Bayes model, since $\boldsymbol{\eta}_h$ is estimated by $\boldsymbol{y}$ in the conventional empirical Bayes method. The marginal likelihood $m(x; d, k, \boldsymbol{\eta}_h)$ exists in this case, and can be used for obtaining an estimate of $(d, k, \boldsymbol{\eta}_h)$ by maximizing it.

Recall that a hyperparameter in the empirical Bayes model is often estimated in terms of the marginal density instead of the posterior density. This is probably the key reason why a degenerated prior density is widely assumed rather than an improper prior function. The comparison between the posterior density and the marginal density in the empirical Bayes model is a difficult problem to be solved. However, we will attempt brief comparison studies to assert advantages of a permissible boundary prior function in Sections 4 and 5.

3 Basic properties

We first present three basic properties. Suppose that $b(\theta)$ is a permissible boundary prior function to $\mathcal{P} = \{\pi(\theta; g) | \, g \in \mathcal{G}\}$ and for $\mathcal{M} = \{p(x|\theta) | \, \theta \in \Theta\}$.

Proposition 1 (Invariance property) *Suppose that θ is written as $f(\eta)$ for a strictly monotone function with a first derivative. Then $b(f(\eta))|f'(\eta)|$ is a permissible boundary prior function to $\mathcal{P}_f = \{\pi_f(\eta; g) = p(f(\eta); g)|f'(\eta)| \, \big| \, g \in \mathcal{G}\}$.*

Proposition 2 *1) Let $h(\theta)$ be a function satisfying the condition that there exists a positive constant M such that $1/M < h(\theta) < M$ for every θ. Consider a family of prior densities, $\mathcal{P}_h = \{\pi_h(\theta; g) = \pi(\theta; g)h(\theta)K(g) \big| \, g \in \mathcal{G}\}$ with the normalizing constant $K(g)$. Then $b(\theta)h(\theta)$ is a permissible boundary prior function to the family of prior densities $\mathcal{P}_h$.*

2) Suppose that a positive function $f(x)$ derives another family of sampling densities $\mathcal{M}_f = \{p(x|\theta)f(x)K(\theta)\}$ with the normalizing constant $K(\theta)$. We assume that there exists a positive constant M such that $1/M < K(\theta) < M$ for every θ. Then $b(\theta)$ is a permissible boundary prior function to $\mathcal{P}$ and for $\mathcal{M}_f$.

When we attempt to assume an improper prior function $b(\theta)$ for a parameter in a family of sampling densities $\mathcal{M}$ in practical applications, it is necessary to construct explicitly a family of proper prior densities $\mathcal{P}$ to which the prior function $b(\theta)$ is permissible boundary. By applying the power family in Ibrahim and Chen (2000), we discuss a general method under weak regularity conditions. Recall that $p(x|\theta)$ is assumed to be bounded as a function of θ for every x in Sect. 2.

Proposition 3 *Suppose that there exist a sample value $x_0 \in \mathcal{X}$ and a positive number $c(x_0)$ such that the integral $\int p^g(x|\theta)b(\theta)\mathrm{d}\theta$ exists for $0 < g < c(x_0)$. Set*

$$\mathcal{P} = \{\pi(\theta; g) = p^g(x_0|\theta)b(\theta)K(g, x_0) | \, 0 < g < c(x_0)\} \tag{9}$$

with $K(g, x_0)$ being the normalizing constant. Then $b(\theta)$ is a permissible boundary prior function to $\mathcal{P}$ and for $\mathcal{M}$.

Proof Writing the upper bound of $p(x|\theta)$ as $B(x)$, we can show that $\int p^h(x_0|\theta)b(\theta)\mathrm{d}\theta$ is less than $B^{h-g}(x_0)/K(g, x_0)$. Thus we may choose ∞ as $c(x_0)$. Further it follows that the expectation $E\{p(x|\theta);\ \pi(\theta; g)\}$ exists for every x, since $p(x|\theta)$ is less or equal to $B(x)$. This expectation yields the posterior density $p(\theta|\, x)$ under a prior density $\pi(\theta; g)$. It is expressed as $p(x|\theta)p^g(x_0|\theta)b(\theta)K(g, x_0, 1, x)$, where $K(g, x_0, h, x)$ is the normalizing constant.

Setting $a_n = 1/K(1/n, x_0)$, we obtain that $a_n\pi(\theta; 1/n)$ converges weakly to $b(\theta)$. In order to evaluate the logarithmic divergence between two posterior densities, we give an explicit form of the divergence for non-negative values g and h that

$$D(\pi(\theta|x; g), \pi(\theta|x; h)) = (g-h)E\{\log p(x|\theta);\ p(\theta|\, x)\} + \log\{K(g, x_0, 1, x)/K(h, x_0, 1, x)\}. \tag{10}$$

The posterior mean of $\exp t\{\log p(x|\theta)\}$ is written as $K(g, x_0, 1, x)/K(g, x_0, 1+t, x)$, which is a moment generating function. Thus the posterior mean in the right-hand side in (10) is obtained by the partial derivative of the posterior mean with respect to t at $t = 0$. This takes a finite value, since $K(g, x_0, h, x)$ exists for every $h > 0$. Thus the former term vanishes, as $g - h$ tends to 0. The second term also vanishes, since $K(g, x_0, 1, x)$ is continuous in g. □

The assumption of a conjugate prior density yields a closed form of the posterior density. As usual, it provides us with useful and tractable examples:

Example 5 (Conjugate prior) Suppose that the sampling density is in the regular exponential family $\mathcal{M} = \{p(\boldsymbol{x}|\,\theta) = \exp[n(\bar{x}\theta - M(\theta))]a(\boldsymbol{x})|\,\theta \in \Theta\}$. Consider an exponential dispersion family of prior densities of the form

$$\mathcal{P} = \{\pi(\theta; g, m) = \exp\big[g\{m\theta - M(\theta) - N(m)\}\big]b(\theta)\cdot K(g, m)|\, g > 0, m \in C\} \tag{11}$$

where C is the image of $M'(\theta)$. This family is widely known as the conjugate prior density and includes familiar prior densities. The conjugacy property yields that the posterior density is written as

$$\pi(\theta|x; g, m) = \pi(\theta;\ n+g, \tilde{\mu}) \tag{12}$$

where $\tilde{\mu} = (n\bar{x} + gm)/(n+g)$ with $\mu = M'(\theta)$. Setting $a_g = 1/K(g, m)$ for a fixed m, we can show that $b(\theta)$ is a boundary prior function to $\mathcal{P}$. Note that the expectations of θ and $M(\theta)$ under $\pi(\theta; g, m)$ exist for every g and m when the sampling density is in the regular exponential family, and also that the logarithmic divergence $D(\pi(\theta; g, m),\ \pi(\theta; g', m'))$ is continuous in (g, m) and (g', m'). Thus $b(\theta)$ is a permissible boundary prior function to $\mathcal{P}$ and for $\mathcal{M}$. The form (12) indicates that the value g represents amounts of information on our prior knowledge contained in $\pi(\theta; g)$,

which corresponds to the sample size in the sampling density. Consider two very small values of g, g_s and g_s'. Assumptions of these two prior densities $\pi(\theta; g_s, m)$ and $\pi(\theta; g_s', m)$ reflect very small amounts of information on our prior knowledge about θ. Thus these two prior densities play almost equivalent roles in practical applications, and the derived posterior densities are almost the same.

A non-informative prior for the location family of the sampling densities has been studied extensively in the objective Bayesian approach. Special attentions are paid to the uniform prior function, but we discuss the prior function in relation to the location-scale family of the prior densities.

Example 6 (Location-scale family) Another general class of families of prior densities is given for the sampling density in the location family $\mathcal{M} = \{p(\boldsymbol{x}|\theta) = \prod p(x_i - \theta)\}$. Let $b(\theta)$ be a positive function on $\Theta = R^1$, and consider a family of proper prior densities of the form

$$\mathcal{P} = \{p(g(m-\theta))b(\theta) \cdot K(g,m) | \, g > 0\}. \tag{13}$$

A naive and appealing prior function $b(\theta)$ is a uniform prior function on Θ. Berger et al. (2009) gave a regularity condition for the reference prior function $b(\theta)$ to be permissible in their definition.

Since our interest is in the requirement that the constant $K(g,m)$ is independent of m, we examine again whether the family $\mathcal{P}$ for a fixed g is complete or not. A proposition will be given at the end of this section.

A more general family of sampling densities is given as $\mathcal{M} = \{p(\boldsymbol{x}|\theta,\tau) = \prod \tau p\big(\tau(x_i - \theta)\big)\}$ by introducing a scale parameter τ. It is reasonable to assume a prior density given a fixed τ in the family $\{p(g_1(m-\theta))K(g_1) | \, g_1 > 0\}$. There is no widely accepted prior density for τ. A naive choice of the density may be the gamma one with mean $1/t$ and variance $1/(g_2 t^2)$. Another choice of a prior density for (θ, τ) is the normal-gamma density. In each case, $b(\theta,\tau) = 1/\tau$ is a permissible boundary prior function.

In Examples 5 and 6, we claimed that orthogonality condition in a family $\{\pi(\theta; g, m)\}$ with respect to g and m may be useful for specifying $b(\theta)$. We state formally the uniqueness of the specification through the orthogonality condition.

Proposition 4 *Suppose that for an arbitrarily fixed g the family $\mathcal{P}$ is complete and also that $K(g,m)$ is independent of m. Then the permissible boundary prior function $b(\theta)$ is unique up to a constant multiplier.*

A regular exponential family satisfies the completeness property, and the reference prior function (1) in Example 1 is the unique permissible boundary prior function up to a constant multiplier. When $b(\theta)$ is a uniform prior function, $K(g,m)$ is independent of m.

Proposition 5 *Consider the location family of the sampling densities of the form $\mathcal{P}$ in (13). A sufficient condition for $K(g,m)$ to be independent of m is that $b(\theta)$ is a uniform prior function. Then $K(g,m)$ is expressed as ag for a positive constant a.*

Combining the above two propositions, we find that a uniform function is the unique permissible boundary prior function to the families in Example 6 up to a constant multiplier.

4 Marginal density

The Bayesian prediction theory provides us with a flexible and perspective view of Bayesian inferential method. The standard Bayesian predictive density due to Aitchison (1975) is $p_m(y|x) = E\{p(y|\theta);\ \pi(\theta|x)\}$ with $y \in \mathcal{X}$. Corcuera and Giummole (1999) extends the predictive density in terms α-mixture for $-1 \le \alpha \le 1$, where the original definition corresponds to $\alpha = -1$. In the differential geometric context this mixture is called also m-mixture. The dual version of the predictor $p_e(y\,|\,x)$ will be used in the next section.

The posterior Bayes factor $p_m(x|x)$ was proposed in Aitkin (1991), but was criticized because of the double use of an observation x in $p(x|\theta)$ and also in $\pi(\theta|\,x)$. An easy way to dissolve this possible over-fitness is to apply the cross-validation treatment (Stone 1977). To express the cross-validation method, we assume that the sampling density is in the $i.i.d.$ case. Write an observation in R^n as $\boldsymbol{x}$ and the sub-vector of $\boldsymbol{x}$ being dropped off the i-th component as $\boldsymbol{x}_{-i}$. Then it holds that

$$p(x_i\,|\,\boldsymbol{x}_{-i})\left(=\ E\{p(x_i\,|\,\theta)\,\theta);\ \pi(\theta|\boldsymbol{x}_{-i})\}\right)\ =\ \frac{m(\boldsymbol{x})}{m(\boldsymbol{x}_{-i})}.$$

Note that the left-hand side dissolves the problem of the double use of an observation. Note also that it is written as the posterior mean of $p(x_i|\theta)$. This equality yields an expression of a cross-validated version of the posterior Bayes factor.

Proposition 6 *Set* $\mathrm{CVC} = -2\sum_{i=1}^{n}\log p_m(x_i\,|\,\boldsymbol{x}_{-i})$. *Then it holds that*

$$\mathrm{CVC}\ =\ -2\sum\log\frac{m(\boldsymbol{x})}{m(\boldsymbol{x}_{-i})}. \tag{14}$$

We discuss the equation (14) in the cases of $\pi(\theta;\,g)$ for $g \in \mathcal{G}$, where $\mathrm{CVC}(g)$ depends on g. Suppose that $b(\theta)$ is a permissible boundary prior function to $\mathcal{P} = \{\pi(\theta;\ g)|\,g \in \mathcal{G}\}$. Since each term of $\mathrm{CVC}(g)$ is the logarithmic transformation of a posterior mean, we can define $\mathrm{CVC}(0)$ as the limit of $\mathrm{CVC}(g)$ at $g = 0$. When $b(\theta)$ is a permissible boundary prior function to $\mathcal{P}$, $\mathrm{CVC}(0)$ is derived also from the posterior density induced from $b(\theta)$ and a family of the sampling densities $\mathcal{M}$.

We will call $-2\log m(x)$ the marginal likelihood criterion MLC, since it is used as a model selection criterion. Kass and Raftery (1995) gave the following decomposition of MLC in a rather narrative way. Write the sample vector consisting of the first i components as $\boldsymbol{x}_i = (x_1, \ldots, x_i)'$. Their decomposition in the section 3.2 is formally expressed in terms of our notation as

$$\mathrm{MLC}(g)\ =\ -2\sum_{i=2}^{n}\log p_m(x_i\,|\,\boldsymbol{x}_{i-1})\ -\ 2\log m(x_1) \tag{15}$$

which will be written as $\mathrm{MLC}_1(g) + \mathrm{MLC}_2(g)$. We find that all the terms in $\mathrm{MLC}_1(g)$ are written as the posterior means but that the marginal density appears in $\mathrm{MLC}_2(g)$. The former term $\mathrm{MLC}_1(g)$ consists of $(n-1)$ terms but the latter term $\mathrm{MLC}_2(g)$ has only one term. When $b(\theta)$ is a permissible boundary prior function to $\mathcal{P}$, the main term $\mathrm{MLC}_1(g)$ in (15) can be defined for $g = 0$. However, the latter term $\mathrm{MLC}_2(g)$ tends to ∞ as g tends to 0.

Another view of the marginal density is given through the following known expression:

$$\frac{1}{m(x)} = E\left\{\frac{1}{p(x|\theta)};\ \pi(\theta|x)\right\}.$$

We observe that $p(x|\theta)$ and $\pi(\theta|x)$ in the right-hand side appear in the numerator and the denominator, respectively. This form looks confusing, since we favor larger values of these terms in Bayesian modeling. In fact, the predictor due to Aitchison (1975) is expressed as $p_m(y|x) = E\{p(y|\theta);\ \pi(\theta|x)\}$, where neither $p(x|\theta)$ nor $\pi(\theta|x)$ appears in the denominator.

5 Marginal density and DIC

We discuss here the empirical Bayes model, where a prior density $\pi(\theta;\delta)$ contains a hyperparameter δ to be estimated. Spiegelhalter et al. (2002) introduced a criterion based on the posterior density. We compare it with $\mathrm{MLC}(\delta)$ to examine the role of a permissible boundary prior function.

The e-mixture ($(\alpha =)1$-mixture) predictive density is given as

$$p_e(y|x;\delta) = \frac{1}{c(x;\delta)} \exp\big[E\{\log p(y|\theta);\ \pi(\theta\,|\,x;\delta)\}\big]. \tag{16}$$

When the sampling density is in the exponential family and θ is the canonical parameter, $p_e(y|x;\delta)$ is expressed as $p(y|\hat{\theta};\delta)$. The two predictors, $p_e(y|x;\delta)$ and $p_m(y|x;\delta)$, are optimum under the logarithmic divergence loss functions, $D(p(y|x;\delta), p(y|\theta))$ and $D(p(y|\theta), p(y|x;\delta))$, respectively.

When θ is the canonical parameter in the exponential family, DIC is expressed as

$$\mathrm{DIC}(\delta) = -2\log p_e(x\,|\,x;\delta) + p_D + p_D \tag{17}$$

where $p_D = 2E\left\{\mathrm{D}\big(p_e(y|x;\delta),\ p(y;\theta)\big)\,|\,\pi(\theta\,|\,x;\delta)\right\}$. The second and the third terms are written separately for the easier comparison with other criteria. This criterion is modified so as to satisfy an unbiasedness condition

$$E\{\mathrm{uDIC} + 2\log[p_e(y|x;\delta)];\quad p(x|\theta)p(y|\theta)\pi(\theta;\delta)\} = 0$$

for every δ. This modification yields

$$\mathrm{uDIC}(\delta) = -2\log p_e(x\,|\,x;\delta) + p_D + q_D \tag{18}$$

where $q_D = 2E\left\{\mathrm{D}\big(p(y|\theta),\ p_e(y|x;\delta)\big);\ \pi(\theta\,|\,x;\delta)\right\}$ Yanagimoto and Ohnishi (2009a). This unbiasedness condition is satisfied even when that $\mathcal{M}$ is not in the exponential family. The following proposition presents an expression of MLC corresponding to an expression of DIC in (17) and that of uDIC in (18).

Proposition 7 *It holds that*

$$\mathrm{MLC}(\delta) = -2\log p_e(x\,|\,x;\delta) + p_D + 2D(\pi(\theta|\,x;\delta),\ \pi(\theta;\ \delta)). \tag{19}$$

Proof The divergence between the posterior and a prior densities is written as

$$D(\pi(\theta|\,x;\delta),\ \pi(\theta;\ \delta)) = E\left\{\log\frac{p(x|\,\theta)}{m(x;\ \delta)};\ \pi(\theta|\,x;\delta)\right\}.$$

As discussed in the equality (5.1) in Yanagimoto and Ohnishi (2009a), the following identity holds:

$$E\left\{\log\frac{p_e(x|\,x;\delta)}{p(x|\,\theta)};\ \pi(\theta|\,x;\delta)\right\} = E\left\{\mathrm{D}\big(p_e(y|\,x;\delta),\ p(y|\,\theta)\big);\ \pi(\theta\,|\,x;\delta)\right\}.$$

Noting that the right-hand side in the above equality is $p_D/2$, we obtain the necessary equality (19). □

It is widely believed that performance of MLC(δ) is largely different from that of DIC(δ). However, we find surprisingly close relationships among their analytical expressions (17), (18) and (19). The first two terms of the three criteria are common, and the third terms differ from each other. Among the three the third term of MLC(δ) is largely different from the other two. The third term of MLC(δ) appears in Berger et al. (2009), which is interpreted as amount of information added by the observation x. When a prior density $\pi(\theta;\ \delta)$ is weakly informative, this quantity is likely to take a large value. Thus this criterion is highly sensitive with a weakly informative prior density.

The third terms, p_D of DIC(δ) and q_D of uDIC(δ), are expressed as the posterior means of the logarithmic divergence between the optimum predictor and the sampling density. Spiegelhalter et al. (2002) interpreted the term p_D as the complexity of the model. They are not sensitive with a small value of δ.

As a corollary of Proposition 7, we obtain an expression of the difference between two MLC(δ)'s for different values of δ. Consider two prior densities $\pi(\theta;\ \delta)$ and $\pi(\theta;\ \delta_0)$. The latter prior function will be treated as a reference one.

Proposition 8 *It follows from the definitions of the marginal densities that*

$$\mathrm{MLC}(\delta) - \mathrm{MLC}(\delta_0) = 2\log\left[E\left\{\frac{\pi(\theta;\ \delta_0)}{\pi(\theta;\ \delta)};\ \pi(\theta|\,x;\delta)\right\}\right]. \tag{20}$$

The expression of the difference in (20) indicates that it is sensitive with the ratio of $\pi(\theta;\ \delta_0)/\pi(\theta;\ \delta)$. Note that the ratio takes a large value, when $\pi(\theta;\ \delta)$ takes a small

value. This expression indicates that the criterion is not sensitive with a large value of $\pi(\theta;\ \delta)$ but is sensitive with a small value of $\pi(\theta;\ \delta)$. The latter fact is simply due to the fact that $\pi(\theta;\delta)$ converges weakly to $\pi_N(\theta)$. Again, we observe that MLC(δ) is very sensitive with a weakly informative prior density.

6 Further examples

Further examples are presented here to supplement the important role of an improper function relating a family of proper prior densities. The first example follows up the case of a normal-gamma prior density in Example 1 to the case of Stein type estimator of a mean vector. The subsequent two examples discuss general families in addition to Examples 5 and 6. We present also three additional examples concerning the linear model, the model used in the Lindley paradox, and the empirical Bayes method for the smoothing model.

Example 1 (Continued) Consider a following family of prior densities of the form:

$$\pi(\mu,\tau;\ \delta,\lambda,m,t)\ =\ \frac{\sqrt{\delta}}{\sqrt{2\pi}}\exp\left\{-\frac{\delta}{2}(\mu-m)^2\right\}\cdot\frac{\lambda^{\lambda}\tau^{\lambda-1}}{\Gamma(\lambda)t^{\lambda}}\exp\left\{-\frac{\lambda\tau}{t}\right\}. \quad (21)$$

Here (δ,λ) corresponds to (g_1,g_2) in (4). Setting $k(\delta_n,\lambda_n,t)\ =\ \sqrt{2\pi}\Gamma(\lambda_n)t^{\lambda_n}/\{\sqrt{\delta_n}\lambda_n{}^{\lambda_n}\}$, we can show that the sequence of prior densities $\{k(\delta_n,\lambda_n,t)\pi(\mu,\tau;\ \delta_n,\lambda_n,m,t)\}$ converges weakly to $b(\mu,\tau)$ as both δ_n and λ_n tend to 0. In addition, it follows that the induced posterior density $\pi(\mu,\tau|\boldsymbol{x};\ \delta,\lambda)$ varies smoothly with $\delta(\geq 0)$ and $\lambda(\geq 0)$.

On the other hand, the sequence $\{\pi(\mu,\tau;\ \delta_n,\lambda_n,\mu_h,\tau_h)\}$ converges weakly to $\delta_D(\theta-\theta_h)$, as both δ_n and λ_n tend to ∞. However, the prior density contains the unknown hyperparameters μ_h and τ_h. In addition, the logarithmic divergence between an induced posterior density $\pi(\mu,\tau|\boldsymbol{x};\ \delta_n,\lambda_n,\mu_h,\tau_h)$ and the corresponding posterior density $\delta_D(\theta-\theta_h)$ does not exist.

The posterior density $\pi(\mu,\tau\,|\,\boldsymbol{x};\delta,\lambda)$ is proportional to

$$\frac{\sqrt{\delta+n\tau}}{\sqrt{2\pi}}\exp-\frac{\delta+n\tau}{2}(\mu-\hat{\mu})^2\cdot\frac{\tau^{\lambda+\frac{n}{2}-1}}{\sqrt{\delta+n\tau}}\exp-\tau\left\{\frac{\lambda}{t}+\frac{n-1}{2}s^2+\frac{n\delta}{2(\delta+n\tau)}(\bar{x}-m)^2\right\} \quad (22)$$

with $\hat{\mu}\ =\ (\delta m+n\tau\bar{x})/(\delta+n\tau)$. The former factor of (22) denotes the normal posterior density given $\boldsymbol{x}$ and τ $\pi(\mu|\,\boldsymbol{x},\tau)$. The integral of the second factor of the variable τ exists, and we write it as $K(\delta,\lambda)$. We write the marginal posterior density of τ as $\pi(\tau\,|\,\boldsymbol{x})$. After some calculations, we obtain that the logarithmic divergence between $\pi(\mu,\tau\,|\,\boldsymbol{x};\delta,\lambda)$ and $\pi(\mu,\tau\,|\,\boldsymbol{x};\delta',\lambda')$ is expressed as the expectation of

$$\frac{\delta'-\delta}{2(\delta+n\tau)}\left\{1+\frac{n^2\tau^2}{\delta'+n\tau}\right\}\ +\ (\lambda'-\lambda)\left\{\frac{\tau}{t}-\log\tau\right\}\ +\ \log\frac{K(\delta',\lambda')}{K(\delta,\lambda)} \quad (23)$$

under $\pi(\tau|\boldsymbol{x})$. To obtain that this divergence vanishes as (δ', λ') approaches to (δ, λ), it is enough to show the continuity of the normalizing constant $K(\delta, \lambda)$ and the existence of expectations of the functions in (23). These follow from the fact that the posterior density $\pi(\tau|\boldsymbol{x})$ decays sharply as τ tends to either 0 or ∞.

Another potential prior density is the Jeffreys prior function, which is proportional to $1/\sqrt{\tau}$ instead of $1/\tau$ in (1). Modifying the previous family in (21), we introduce the family of prior densities of the form

$$\pi(\mu, \tau;\ \delta, \lambda, m, t) = \frac{\sqrt{\delta\tau}}{\sqrt{2\pi}} \exp\left\{-\frac{\delta\tau}{2}(\mu - m)^2\right\} \cdot \frac{\lambda^\lambda \tau^{\lambda-1}}{\Gamma(\lambda) t^\lambda} \exp\left\{-\frac{\lambda\tau}{t}\right\}. \quad (24)$$

Write the right-hand side (24) as $\pi(\mu|\tau;\ \delta, m) \cdot \pi(\tau;\ \lambda, t)$. This family is the conjugate family to the normal sampling density. In fact, the posterior density $\pi(\mu, \tau|\boldsymbol{x};\ \delta, \lambda, m, t)$ is expressed as

$$\pi(\mu|\tau;\ n\tau + \delta, \hat{\mu}) \cdot \pi\left(\tau;\ \frac{n}{2} + \lambda, \frac{\frac{n}{2} + \lambda}{\lambda/t + s^2/2 + n\delta(\bar{x} - m)^2/(n + \delta)}\right) \quad (25)$$

with $\hat{\mu} = (n\bar{x} + \delta m)/(n + \delta)$. Since an explicit expression of the posterior densities is available, it is easy to show that the Jeffreys prior density is a permissible boundary prior function to $\mathcal{P}$ and for $\mathcal{M}$.

Example 7 (Location-dispersion family) Consider a location family of the sampling densities on R^1, $\mathcal{M} = \{p(\boldsymbol{x}|\theta) = \prod \exp -d(x_i - \theta)\}$ and assume a location-dispersion family of prior densities

$$\mathcal{P} = \{\pi(\theta;\ g, m) = \exp\big[-gd(m - \theta)\big]K(g)\big|\ g > 0\}. \quad (26)$$

Note that this family is close to but is different from the location-scale family in Example 6. An extension of the family of prior densities in (26) is possible by replacing $K(g)$ by $b(\theta)K(g, m)$, but we will focus our attention here on the above restricted case. Superficially, it looks that a prior density in the location-scale family is familiar and simple, but the location-dispersion family provides us with a wide variety of distributions, including the logarithmic gamma, Gumbel and Fréchet distributions. The dispersion parameter is usually easier to be handled than the scale parameter. Further, the exponential dispersion family is often in the exponential family or in the curved exponential family. The basic properties of this family can be obtained in Jorgensen (1997).

Suppose that the marginal density $m(\boldsymbol{x};\ g, m)$ exists for $g \geq 0$, which is satisfied if the function $d(\theta)$ decays in a moderate rate as θ tends to ∞. Differentiating an prior density $\pi(\theta;\ g, m)$ with respect to g, we obtain that the expectation of $d(m - \theta)$ under $\pi(\theta;\ g, m)$ exists and is equal to $-\partial \log K(g)/\partial g$. The posterior density is written as

$$\pi(\theta|\boldsymbol{x};\ g, m) = \frac{\exp\{\sum -d(x_i - \theta) - gd(m - \theta)\}}{m(\boldsymbol{x};\ g, m)} K(g).$$

This implies the logarithmic divergence $D(\pi(\theta|\boldsymbol{x}; g, m),\ \pi(\theta|\boldsymbol{x}; g', m))$ is given by the expectation of

$$-(g-g')d(\theta-m)\ +\ \log\left\{\frac{K(g)\ m(\boldsymbol{x}; g', m)}{K(g')\ m(\boldsymbol{x}; g, m)}\right\}$$

under $\pi(\theta|\boldsymbol{x}; g, m)$. Thus the logarithmic divergence exists for positive g and g'. It follows from the definition of the marginal density that $m(\boldsymbol{x}; g, m)/K(g)$ is independent of g. Further, the logarithmic divergence is continuous in g and g'. Consequently, we obtain that a uniform prior function is permissible boundary to $\mathcal{P}$ in (26) and for $\mathcal{M}$.

Proposition 3 gave a general method for constructing a family of proper prior density yielding a permissible boundary prior function. Fortunately, the method is applicable to the power family.

Example 8 (Power family) A familiar class of families of prior densities derived from the sampling density is given by

$$\mathcal{P}\ =\ \left\{\pi(\theta; g)=\frac{p^g(x_0|\theta)}{p^g\left(x_0|\hat{\theta}_M(x_0)\right)}b(\theta)K(g, \hat{\theta}_M(x_0))\ \middle|\ g\in\mathcal{G}\right\} \tag{27}$$

where $\hat{\theta}_M(x_0)$ is the maximum likelihood estimate. This family was discussed below Proposition 3, where x_0 is treated as a known value in the notation. Note that x_0 can be arbitrarily chosen in most practical examples. This family is covered by the general family in (3) in Example 2 by setting $c=\hat{\theta}_M(x_0)$. We discuss here reasons why this family was employed in Proposition 3 when a prior function $b(\theta)$ and a family of sampling densities were given. A reason is its simple multiplicative forms of the member of $\mathcal{P}$, and another is its simple form of the logarithmic divergence in the proof. When the sampling density is in the exponential family, the normalizing constant in (27) is written in terms of the logarithmic divergence as $\exp\{-gD\left(p(y|\hat{\theta}_M(x_0)),\ p(y|\theta)\right)\}$. By setting $m=M'\left(\hat{\theta}_M(x_0)\right)$ in (11), we obtain this family.

Ibrahim and Chen (2000) extensively explored this family. They suggested the use of a proper prior density $\pi(\theta; g)$ in (27). We agree with their suggestion, when such a proper prior is available. Otherwise, we suggest the use of $\pi(\theta; g)$ in (27) for a suitably chosen small value of g. A permissible boundary prior function $b(\theta)$ may be chosen when g is considered as a very small value, but we find practical difficulties in specifying it.

Example 9 (Linear model) Consider a linear regression model, where an observation $\boldsymbol{x}$ follows the normal distribution $\mathrm{N}(Z\theta, (1/\tau)\mathrm{I})$ with Z being the $n\times p$ $(1\le p<n)$ design matrix of the rank p. Then a family of the normal-gamma prior densities discussed in Example 1 provides us with another view of a permissible boundary prior function. This choice of a permissible boundary prior function performs better than that of a degenerated prior density, as was observed in Example 1.

Our special interest here is in the case where p is not small, say 5 or 6, and $n - p$ is moderately large. In light of the widely known Stein effect, an empirical Bayes approach is promising in this case. We assume a naive normal-gamma prior density is

$$\pi(\theta, \tau;\ \boldsymbol{g}, \boldsymbol{c}, t) = \frac{\prod \sqrt{g_i}^p}{\sqrt{2\pi}^p} \exp\left\{-\frac{1}{2}\sum g_i(\theta_i - c_i)^2\right\} \cdot \frac{g_0^{g_0}\tau^{g_0-1}}{\Gamma(g_0)t^{g_0}} \exp\left\{-\frac{g_0\tau}{t}\right\} \cdot \frac{1}{\tau} \tag{28}$$

where $\boldsymbol{g} = (g_0, g_1, \ldots, g_p)$ and $\boldsymbol{c} = (c_1, \ldots, c_p)$. The prior function $1/\tau$ is a permissible boundary prior function to this family.

Suppose that a statistician wonders as to which of an improper prior function and a weakly informative prior density is suitable for the analysis. Specifically, the problem lies in choosing an improper prior function proportional to $1/\tau$ or the prior function of the form (28) for small values of g_i's. Since our definition of permissibility was designed for the prior function $1/\tau$ to be treated as a substitute for a very weakly informative prior density, $\pi(\theta, \tau;\ \boldsymbol{g}_s, \boldsymbol{c}, t)$, posterior densities induced from these prior function and density are close to each other. Thus the problem is not serious so far as the analysis is based the posterior density.

In contrast, the problem becomes severely tough, when one wonders as to which of a degenerated prior density at an unknown point and a weakly informative prior density is suitable for the analysis. It is not easy to find a density among prior densities in (28) close to a degenerated prior density $\delta_D(\boldsymbol{\eta} - \boldsymbol{\eta}_h) \cdot \delta_D(\tau - \tau_h)$, as was noted previously. If we choose a large value of g_i's, a difficult problem to specify suitable values of $\boldsymbol{c}$ and t arises. Next, consider the case where we choose a small value of g_i's. In this case the problem of specifying $\boldsymbol{c}$ and t becomes less important. On the other hand, the choice of a small value of g_i's contradicts Proposition 3.2, since $\pi(\theta, \tau;\ \boldsymbol{g}, \boldsymbol{c}, t)$ converges weakly to $\delta_D(\boldsymbol{\eta} - \boldsymbol{c}) \cdot \delta_D(\tau - t)$ as all the components of $\boldsymbol{g}$ tend to ∞.

The above discussions indicate that an inferential procedure based on a permissible boundary prior function to $\mathcal{P}$ is associated with that based on a very weakly informative prior density in $\mathcal{P}$. On the contrary, an inferential procedure based on a degenerated prior density at an unknown point is separated from that based on a proper prior density.

Example 10 (Model used in the Lindley paradox) Suppose x is sample of size 1 from the normal population $\mathrm{N}(\mu, 1/\tau_0)$, and let g_s be a very small positive value. We compare a very weakly informative prior density $\pi_1(\mu;\ g_s) \sim \mathrm{N}(\mu_0, 1/(g_s\tau_0))$ and a degenerated prior density at an unknown point $\pi_2(\mu) = \delta_D(\mu - \mu_1)$. Set $S = \{x | \sqrt{\tau_0}|x - \mu_0| < 10\}$ and $\mathcal{P} = \{\pi_1(\mu;\ g_s) | g_s > 0\}$. Then it follows that $\Pr(S;\ \mathrm{N}(\mu_0, 1/\tau_0))$ is very close to 1. When $\mu_1 = \mu_0 + (a + 10)/\sqrt{\tau_0}$, it holds for $x \in S$ that

$$m_1(x) = \frac{\sqrt{g_s\tau_0}}{\sqrt{2\pi(1+g_s)}} \exp -\frac{g_s\tau_0}{2(1+g_s)}(x - \mu_0)^2 < \frac{\sqrt{g_s\tau_0}}{\sqrt{2\pi}}$$

and also that

$$m_2(x) = \frac{\sqrt{\tau_0}}{\sqrt{2\pi}} \exp -\frac{\tau_0}{2}(x - \mu_0 - \mu_1)^2 > \frac{\sqrt{\tau_0}}{\sqrt{2\pi}} \exp -\frac{1}{2}a^2.$$

Thus $m_1(x) < m_2(x)$ holds, when $\exp(-a^2) \geq g_s$. When we set $g_s = 10^{-20}$ as a very small value, the value $a_0 = \sqrt{20 \log 10}$ satisfies this inequality. The marginal likelihood criterion MLC indicates that the prior density $\delta_D(\mu - \mu_1)$ with $\mu_1 = \mu_0 + (10 + a_0)/\sqrt{\tau_0}$ is superior to the prior density $\mu \sim \mathrm{N}\big(\mu_0, 10^{20}/\tau_0\big)$, when x is in S.

The posterior density $\pi_1(\mu|x; g_s)$ follows $\mathrm{N}\big(\hat{\mu}, 1/\tau_0(1 + g_s)\big)$ with $\hat{\mu} = (x + g_s\mu_0)/(1 + g_s)$, which is well approximated by $\mathrm{N}(x, 1/\tau_0)$. The approximated posterior distribution is induced from a permissible boundary prior function to $\mathcal{P}$, which is a uniform prior function in this model. The probability of the interval $(0, \mu_1)$ under the posterior density is extremely close to 1. This result indicates that the criterion MLC is sensitive with an assumption of a weakly informative prior density, as was discussed below Proposition 7. The above result is discouraging for the present authors, since the assumption of two weakly informative prior densities is believed to affect little to the induced posterior densities. It is our understanding that Lindley's paradox is associated with this property of MLC. Critiques of MLC are seen in Chacon et al. (2007) and Yanagimoto and Ohnishi (2009b).

Example 11 (Smoothing model) Let $\boldsymbol{x} = (x(1), \cdots, x(n))'$ be a sample vector from the n-dimensional normal population $\mathrm{N}(\boldsymbol{\mu}, (1/\tau)\mathrm{I})$ with $\boldsymbol{\mu} = (\mu(1), \cdots, \mu(n))'$. Let D_k be the difference matrix of order k. Denote by M_k the closure of $\{\boldsymbol{\mu} | \, D_k\boldsymbol{\mu} \neq \mathbf{0}_{n-k}\}$ where $\mathbf{0}_{n-k}$ is the $(n-k)$ dimensional 0 vector, and write its orthogonal complement as $M_k^{\perp}$. Then a parameter $\boldsymbol{\mu}$ in M is decomposed into the direct sum $\boldsymbol{\mu}_1 \oplus \boldsymbol{\mu}_2$ with $\boldsymbol{\mu}_1 \in M_k$ and $\boldsymbol{\mu}_2 \in M_k^{\perp}$. Write also the Moore–Penrose g-inverse of D_k as D_k^-. Then a prior distribution for $\boldsymbol{\mu}_1$ in M_k is written as

$$\pi_a(\boldsymbol{\mu}_1; \gamma) \sim \mathrm{N}(\mathbf{0}_n, (1/\tau\gamma)D_k^-). \tag{29}$$

This prior density is to be compared with

$$\pi_b(\boldsymbol{\mu}_1; \delta) \sim \mathrm{N}(\mathbf{0}_n, (1/\delta)D_k^-). \tag{30}$$

Assume that a prior density for $\boldsymbol{\mu}_2$ and τ, $\pi(\boldsymbol{\mu}_2, \tau)$ is proportional to $1/\tau$. Then it is shown that $\pi_a(\boldsymbol{\mu}_1; \gamma)\pi(\boldsymbol{\mu}_2, \tau)$ yields a posterior density for a fixed γ, but that $\pi_b(\boldsymbol{\mu}_1; \delta)\pi(\boldsymbol{\mu}_2, \tau)$ does not yield a proper posterior density for a fixed δ. This means that we should be careful about the behavior of the posterior density, as was emphasized in Speckman and Sun (2003). On the other hand, the marginal density does not distinguish the two models (29) and (30). The parameters (τ, δ) and (τ, γ) are treated simply as two equivalent ones in the conventional empirical Bayes method.

7 Conclusion

We attempted to defend the assumption of an improper prior function and raised strong reservations about the assumption of a degenerated prior density at an unknown point. Though the assumption of a degenerated prior density is associated with the use of the marginal likelihood, Bayesian inference is in principle based on the posterior density

induced from an assumed proper prior density. This indicates that an improper prior function is hoped to be assumed in relation to a proper prior density and also to the posterior density.

A serious defect of the marginal likelihood is that it takes a very small value when a prior density is close to an improper prior function. Consequently, the marginal likelihood curiously favors a degenerated prior density to a weakly informative prior density. Analytical aspects of this fact are given in Propositions 7 and 8. This view elucidates the erroneous assertion in Lindley's paradox, as in Example 10.

It is not rare that a posterior density does not exist, as Speckman and Sun (2003) remarked. Example 4 shows that the marginal likelihood can exist even in such a case. Example 11 gave an educational example in a smoothing model, which is familiar in the empirical Bayes method. The example presents the two superficially similar models containing different improper prior functions; one of the two induces a proper posterior density but the other does not. These prior densities are not clearly distinguished in practical applications, so far as our own experiences in Yanagimoto and Yanagimoto (1987) and Yanagimoto and Kashiwagi (1990) concern.

Other examples show that improper prior functions in the existing literature mostly satisfy regularity conditions for a permissible boundary prior function to a family of proper prior densities.

A researcher, who plans to assume an improper prior function, is advised to consider a family of proper prior densities $\mathcal{P}$ to which it is a permissible boundary prior function. We expect that such an effort enhances to assume a proper prior density instead of an improper prior function. When the researcher considers the assumption of a proper prior density, a suitable one can be found in the family $\mathcal{P}$.

Acknowledgments The authors are grateful to the two reviewers for their keen comments on the original version, which elucidated points to be clarified.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, *62*, 547–554.
Aitkin, M. (1991). Posterior Bayes factors (with discussions). *Journal of the Royal Statistical Society, Series B*, *53*, 111–142.
Aitkin, M. (2009). *Statisitical inference: an intergrated Bayesian/likelihood approach*. Boca Raton: CRC Press.
Berger, J.O., Bernardo, J.M., Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, *37*, 905–938.
Bolstad V.W. (2007). *Introduction to Bayesian statistics*. Wiley, New York.
Chacon, J.E., Montanero, J., Nogales, A.G., Perez, P. (2007). On the use of Bayes factor in the frequentist testing of a precise hypothesis. *Communications in Statististics -Theory and Methods*, *36*, 2251–2261.
Corcuera, J.M., Giummole, F. (1999). A generalized Bayes rule for prediction. *Scandinavian Journal of Statistics*, *26*, 265–279.
Ibrahim, J.G., Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, *15*, 46–60.
Jaynes, E. T. (2003). *Probability theory: the logic of science*. New York: Cambridge University Press.
Jeffreys, H. (1961). *Theory of probability (third edition)*. Oxford: Oxford University Press.
Jorgensen, B. (1997). *The theory of dispersion models*. London: Chapman & Hall.
Kass, R.E., Raftery, A.E. (1995). Bayes factors. *The Journal of American Statistical Association*, *90*, 773–795.
Lindley, D.V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

McCullagh, P., Han, H. (2011). On Bayes's theorem for improper mixtures. *The Annals of Statistics*, *39*, 2007–2020.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatisitcs*, *9*, 523–539.

Speckman, B.P., Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, *90*, 289–302.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussions). *Journal of the Royal Statistical Society, Series B*, *64*, 583–639.

Stone, M.(1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B*, *39*, 44–47.

Tibshirani, T. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, *13*, 1378–1402.

Yanagimoto, T. (1991). Estimating a model through the conditional MLE. *Annals of the Institute of Statistical Mathematics*, *43*, 735–746.

Yanagimoto, T., Kashiwagi, N. (1990). Empirical Bayes methods for smoothing data and simultaneous estimation of many parameters. *Environmental Health Perspectives*, *87*, 109–114.

Yanagimoto, T., Ohnishi, T. (2005). Standardized posterior mode for the flexible use of a conjugate prior. *Journal of Statistical Planning and Inference*, *131*, 253–269.

Yanagimoto, T., Ohnishi, T. (2009a). Bayesian prediction of a density function in terms of *e*-mixture. *Journal of Statistical Planning and Inference*, *139*, 3064–3075.

Yanagimoto, T., Ohnishi, T. (2009b). Predictive credible region for Bayesian diagnosis of a hypothesis with applications. *Journal of the Japan Statistical Society*, *39*, 111–131.

Yanagimoto, T., Ohnishi, T. (2011). Saddlepoint condition on a predictor to reconfirm the need for the assumption of a prior distribution. *Journal of Statistical Planning and Inference*, *41*, 1990–2000.

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