Sankhyā : The Indian Journal of Statistics
2011, Volume 73-A, Part 2, pp. 185-201

Moment Matching Priors

Malay Ghosh
University of Florida, Gainesville, USA
Ruitao Liu
University of Iowa, Iowa City, USA

Abstract

There are various proposals for the selection of the so-called "objective" or "default" priors in Bayesian analysis. The paper introduces a new criterion, the moment matching criterion, which requires the matching of the posterior mean with the maximum likelihood estimator up to a high order of approximation.

A complete characterization of such priors in the one or multi-parameter case is provided. In the process, many new priors are derived. One interesting finding is that even in the absence of nuisance parameters, it is possible to find priors different from Jeffreys' prior for a real valued parameter based on our criterion.

AMS (2000) *subject classification.* Primary 62F15.
Keywords and phrases. Asymptotic expansion, exponential family, location-scale family, maximum likelihood, posterior mean, proper dispersion models.

1 Introduction

Bayesian methods are increasingly used in these days in the theory and practice of statistics. One key component in any Bayesian inference is the selection of priors. With enough historical data, it is often possible to elicit a suitable prior. But even in its absence, Bayesian methods can often be very effective by the use of so-called 'default' or 'objective' priors.

It is needless to say that over the years, the catalog of 'default' or 'objective' priors has become quite large, and one is often confronted with the selection of an appropriate one in a given context. Thus, it becomes almost imperative to invoke a meaningful criterion of 'objectivity' for prior selection.

In our assessment, so far, there are three major objectivity criteria. The first is to use a suitable divergence measure between the prior and the posterior, and select a prior which maximizes this divergence. This idea was first put forward by Lindley (1956) and was subsequently highly popularized in several articles of Bernardo (1979) and Berger and Bernardo (1989, 1992a, 1992b). These priors, now well-known as 'reference priors', are based on maximizing the 'Kullback-Leibler' divergence between the prior and the posterior. A rigorous derivation of these priors is given in the interesting articles of Clarke and Barron (1990, 1994).

The second, the so-called 'left' and 'right' invariant priors have also a rich history. The notion of priors invariant under one-to-one reparameterization was introduced by Jeffreys (1961) partly to alleviate the criticism raised earlier against the uniform priors of Bayes (1763) and Laplace (1776) which do not remain invariant under one-to-one reparameterization. For many problems of interest, the left-invariant priors are the so-called Jeffreys' general rule prior which are positive square root of the determinant of the Fisher information matrix (Datta and Ghosh, 1995a). However, in most instances, more appropriate are the right-invariant priors which provide many frequentist optimality properties. For example, for the location-scale family of distributions with location parameter μ and scale parameter σ, the left-invariant prior for (μ, σ) is σ^{-2}, while the right-invariant prior is σ^{-1}, the latter being usually referred to as Jeffreys' independent prior. Some recent articles on left and right invariant priors are due to Datta and Ghosh (1995a), Severini, Mukerjee and Ghosh (2002) and Eaton and Sudderth (2004).

The third is the well-known 'probability matching' criterion originally introduced by Welch and Peers (1963) and Peers (1965), and more recently popularized in several articles by Stein (1985), Tibshirani (1989), Ghosh and Mukerjee (1992), Mukerjee and Dey (1993), Datta and Ghosh (1995a), Datta and Ghosh (1995c, 1996), Berger and Sun (2007), Sweeting (2008) among others. The basic idea is to find a one-sided Bayesian credible interval the coverage probability of which matches asymptotically the frequentist coverage probability of the credible interval. There is an overlap between such priors and the right-invariant priors, and indeed many right-invariant priors achieve exact matching of Bayesian and frequentist coverage probability (Severini, Mukerjee and Ghosh, 2002, Eaton and Sudderth, 2004 and Berger and Sun, 2008).

The probability matching priors have been derived mostly for regular family of distributions, namely those whose support does not depend on parameters, plus other conditions such as differentiability and change in the order of differentiation and integration (or summation) etc. hold. A notable exception is the paper of Ghosal (1999) who found probability matching priors for the uniform$(0, \theta)$ and other related non-regular distributions.

The objective of this article is to introduce a new matching criterion which we will refer to as the 'moment matching criterion'. For a regular family of distributions, the classic article of Bernstein and von-Mises (see e.g. Ferguson, 1967, p. 141, or Ghosh, Delampady and Samanta, 2006, p. 104) proved the asymptotic normality of the posterior of a parameter vector centered around the maximum likelihood estimator or the posterior mode and variance equal to the inverse of the observed Fisher information matrix evaluated at the maximum likelihood estimator or the posterior mode. More general results involving the asymptotic expansion of the posterior density are due to Johnson (1970) and Ghosh, Sinha and Joshi (1982). A convenient source for this is the recent book by Ghosh, Delampady and Samanta (2006).

In this article we use a result of Ghosh, Sinha and Joshi (1982) (see also Ghosh, 1994 and Ghosh, Delampady and Samanta, 2006) to find priors which can provide high order matching of the moments of the posterior distribution and those of the maximum likelihood estimator. For simplicity of exposition, we shall primarily confine ourselves to priors which achieve the matching of the first moments, although, it is easy to see how higher order moment matching is equally possible.

The motivation for moment matching priors stems from several considerations. First, these priors lead to posterior means which share the asymptotic optimality of the MLE's up to a high order. In particular, if one is interested in asymptotic bias or MSE reduction of the MLE's through some adjustment, the same adjustment applies directly to the posterior means. In this way, it is possible to achieve Bayes-frequentist synthesis of point estimates. The second important aspect of these priors is that they provide new viable alternatives to Jeffreys' prior even for real-valued parameters in the absence of nuisance parameters motivated from the proposed criterion. A third motivation, which will be made clear in Section 4, is that with moment matching priors, it is possible to construct credible regions for parameters of interest based only on the posterior mean and the posterior variance, which match the maximum likelihood based confidence intervals to a high order of approximation.

The proposed approach bears resemblance to, but is different from the one taken in Hartigan (1998). Like ours, Hartigan wanted to find a prior based on a high order asymptotic agreement between a Bayes and a maximum likelihood estimator. However, his Bayes estimator was based on a truncated Kullback-Leibler loss, and was different from the posterior mean. The motivation for such a loss was apparently to justify the "asymptotically locally invariant prior" obtained earlier in Hartigan (1964). We will see in later sections through examples that our approach can yield priors different from the ones of Hartigan (1998). In particular, for the one-parameter exponential family, while Hartigan's approach yields a constant prior for the canonical parameter, our approach will yield Jeffreys' prior. Another point to note is that while Hartigan's approach involves fairly complicated Taylor series expansion due to the complexity of the loss, ours is a more transparent and direct approach based only on a well-known asymptotic expansion of the posterior, and yet yielding the same order of approximation as in Hartigan (1998).

The organization of the remaining sections is as follows. In Section 2 of this paper, for the regular one-parameter family of distributions including but not limited to the regular one-parameter exponential family, we characterize the class of priors which achieves first order moment matching. Examples are given to illustrate the main theorem. Surprisingly, even in this case, the derived prior may be different from Jeffreys' prior. Extensions of these results to the multi-parameter case are given in Section 3, and some partial differential equations are set up which provide simultaneously necessary and sufficient conditions for the existence as well as construction of such priors. One example is provided to show non-existence of such priors for some regular multi-parameter exponential family of distributions. Section 4 deals with construction of credible intervals based on the posterior mean and the posterior variance which match asymptotically the coverage probabilities of the corresponding confidence intervals based on the MLE to a high order of approximation. A few final remarks are made in Section 5.

A different kind of moment matching in the small area estimation context was proposed by Datta, Rao and Smith (2005) and Ganesh and Lahiri (2008). The objective there was to have asymptotic matching of the posterior variance of a hierarchical Bayes procedure in an area level model to a corresponding empirical Bayes procedure.

2 Priors for The Regular One-Parameter Family of Distributions

Let $X_1, X_2, \cdots, X_n|\theta$ be independent and identically distributed with common density function $f(x|\theta)$, where $\theta \in \Theta$, some interval in the real line. Consider a general class of priors $\pi(\theta)$, $\theta \in \Theta$ for θ. Throughout, it is assumed that both f and π satisfy all the needed regularity conditions as required in Theorem 4.3 of Ghosh, Delampady and Samanta (2006, p. 106).

Let $\hat{\theta}_n$ denote the maximum likelihood estimator of θ. The Bernstein-von Mises Theorem asserts that under some regularity conditions, the posterior distribution of $T_n = \sqrt{n}(\theta - \hat{\theta}_n)$ is asymptotically normal with mean 0 and variance $\hat{I}_n^{-1}$, where $\hat{I}_n = -n^{-1} \sum_{i=1}^n \frac{\partial^2 \log f(X_i|\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}_n}$. Under the prior π, we denote the posterior mean of θ by $\hat{\theta}^B_{n,\pi}$. Let $S_n = \sqrt{n}(\theta - \hat{\theta}_n)\hat{I}_n^{1/2}$. Then the posterior of S_n is asymptotically $N(0,1)$. The following theorem (Johnson, 1970, Ghosh, Sinha and Joshi, 1982) gives an asymptotic expansion of the posterior up to a certain order.

Theorem 2.1. *Assume all the regularity conditions needed in Theorem 4.3 of Ghosh, Delampady and Samanta (2006). In addition, it is assumed that the logarithm of the pdf is five times differentiable in θ and that a prior π is thrice differentiable and is continuous in the interior of the parameter space. Then an asymptotic expansion for the posterior π_n of S_n is given by*

$$\pi_n(s|X_1, \cdots, X_n) = \phi(s)\Bigg[1 + n^{-\frac{1}{2}}\left\{\frac{a_3 s^3}{6\hat{I}_n^{3/2}} + \frac{s}{\hat{I}_n^{1/2}}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\right\}$$
$$+ n^{-1}\left\{\frac{a_4 s^4}{24\hat{I}_n^2} + \frac{a_3^2 s^6}{72\hat{I}_n^3} + \frac{s^2}{2\hat{I}_n^2}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{a_3 s^4}{6\hat{I}_n^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_4}{8\hat{I}_n^2} - \frac{5a_3^2}{24\hat{I}_n^3}\right.$$
$$\left. - \frac{1}{2\hat{I}_n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_3}{2\hat{I}_n^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\right\} + o_p(n^{-1})\Bigg], \tag{2.1}$$

where

$$a_3 = n^{-1}\sum_{i=1}^n \frac{\partial^3 \log f(X_i|\theta)}{\partial \theta^3}\Big|_{\theta=\hat{\theta}_n}, \quad a_4 = n^{-1}\sum_{i=1}^n \frac{\partial^4 \log f(X_i|\theta)}{\partial \theta^4}\Big|_{\theta=\hat{\theta}_n}$$

and $\phi(s)$ is the $N(0,1)$ density function evaluated at s.

REMARK 2.1. The above result is given in Datta and Mukerjee (2004, p. 13) and Ghosh, Delampady and Samanta (2006, p. 108) without stating explicitly all the needed regularity conditions. One of the beauties of the above expansion is that although formally resembling an Edgeworth expansion, this expansion is quite different from a regular Edgeworth expansion. In particular, while an Edgeworth expansion will involve cumulants of the X_i, the one given in Theorem 2.1 involves only derivatives of log-likelihood, and typically is much simpler to evaluate. This is especially so in the multiparameter case where an Edgeworth expansion involves several mixed cumulants.

Our goal in this section is to find a prior such that the difference between the resulting posterior mean and the MLE converges to zero at the $O(n^{-3/2})$ rate as n, the sample size, tends to infinity. To this end we use the following simple theorem which follows from Theorem 2.1. The formal proof makes use of the moments of the $N(0,1)$ distribution. A rigorous proof is given in Ghosh, Sinha and Joshi (1982, p. 425). The result is also stated in Ghosh (1994, p. 47) and Ghosh, Delampady and Samanta (2006, p. 109). We have corrected minor typos in all these sources.

THEOREM 2.2. *Assume the same regularity conditions as in Theorem 2.1. Then*

$$\hat{\theta}^B_{n,\pi} - \hat{\theta}_n = n^{-1}\left(\frac{a_3}{2\hat{I}_n^2} + \frac{1}{\hat{I}_n}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\right) + O_p(n^{-3/2}).$$

Next observe that by the laws of large numbers and consistency of the maximum likelihood estimator, conditional on θ,

$$n(\hat{\theta}^B_{n,\pi} - \hat{\theta}_n) \xrightarrow{P} \left(\frac{g_3(\theta)}{2I^2(\theta)} + \frac{1}{I(\theta)}\frac{\pi'(\theta)}{\pi(\theta)}\right),$$

where

$$g_3(\theta) = E\left[\frac{\partial^3 \log f(X_1|\theta)}{\partial\theta^3}|\theta\right] \text{ and } I(\theta) \text{ is Fisher information number.}$$

With the choice $\frac{\pi'(\theta)}{\pi(\theta)} = -\frac{g_3(\theta)}{2I(\theta)}$, i.e.

$$\pi(\theta) = \exp\left[-\frac{1}{2}\int^{\theta}\frac{g_3(t)}{I(t)}dt\right], \tag{2.2}$$

one gets $(\hat{\theta}^B_{n,\pi} - \hat{\theta}_n) = O_p(n^{-3/2})$. We shall denote this prior π by $\pi_M(\theta)$.

We now provide a result which provides a general expression for the matching prior $\pi_M(\phi)$, where ϕ is a one-to-one function of θ.

THEOREM 2.3. *Suppose $\pi_M(\theta)$ is a moment matching prior for θ, and ϕ is a one-to-one function of θ. Then the moment matching prior $\pi^*_M(\phi) = \pi_M(\theta)\left(\frac{d\theta}{d\phi}\right)^{\frac{3}{2}}$.*

Proof. Define $g_3(\phi) = E\left[\frac{\partial^3 \log f(X_1|\phi)}{\partial\phi^3}|\phi\right]$ and $I(\phi)$ as the Fisher information number for ϕ. By (2.2), one gets

$$\pi_M(\phi) = \exp\left[-\frac{1}{2}\int \frac{g_3(\phi)}{I(\phi)}d\phi\right].$$

First rewrite

$$I(\phi) = I(\theta)\left(\frac{d\theta}{d\phi}\right)^2. \tag{2.3}$$

Next use the Bartlett identity (Bartlett, 1953)

$$0 = g_3(\phi) + 3E\left[\left(\frac{d^2 \log f}{d\phi^2}\right)\left(\frac{d \log f}{d\phi}\right)\right] + E\left(\frac{d\log f}{d\phi}\right)^3. \tag{2.4}$$

and

$$I^{'}(\phi) = -g_3(\phi) - E\left[\left(\frac{d^2 \log f}{d\phi^2}\right)\left(\frac{d\log f}{d\phi}\right)\right]. \tag{2.5}$$

From (2.3) – (2.5),

$$\begin{aligned} g_3(\phi) &= -\frac{3}{2}I^{'}(\phi) + \frac{1}{2}E\left[\frac{d\log f}{d\phi}\right]^3 \\ &= -\frac{3}{2}\left[I^{'}(\theta)\left(\frac{d\theta}{d\phi}\right)^3 + 2I(\theta)\frac{d\theta}{d\phi}\frac{d^2\theta}{d\phi^2}\right] + \frac{1}{2}E\left[\frac{d\log f}{d\theta}\right]^3\left(\frac{d\theta}{d\phi}\right)^3. \end{aligned} \tag{2.6}$$

Since

$$g_3(\theta) = -\frac{3}{2}I^{'}(\theta) + \frac{1}{2}E\left[\frac{d\log f}{d\theta}\right]^3,$$

(2.6) leads to

$$g_3(\phi) = g_3(\theta)\left(\frac{d\theta}{d\phi}\right)^3 - 3I(\theta)\frac{d\theta}{d\phi}\frac{d^2\theta}{d\phi^2}. \tag{2.7}$$

Hence, from (2.7),

$$\begin{aligned} \pi_M^*(\phi) &= \exp\left[-\frac{1}{2}\int\frac{g_3(\phi)}{I(\phi)}d\phi\right] \\ &= \exp\left[-\frac{1}{2}\int\frac{g_3(\theta)}{I(\theta)}\frac{d\theta}{d\phi}d\phi\right] \times \exp\left[\frac{3}{2}\int\frac{I(\theta)\frac{d\theta}{d\phi}\frac{d^2\theta}{d\phi^2}}{I(\theta)\left(\frac{d\theta}{d\phi}\right)^2}d\phi\right] \\ &= \pi_M(\theta)\exp\left[\frac{3}{2}\int\frac{d^2\theta/d\phi^2}{d\theta/d\phi}d\phi\right] \\ &= \pi_M(\theta)\exp\left[\frac{3}{2}\log\left(\frac{d\theta}{d\phi}\right)\right] = \pi_M(\theta)\left(\frac{d\theta}{d\phi}\right)^{\frac{3}{2}} \end{aligned}$$

□

REMARK 2.2. It is now clear that a fundamental difference between priors obtained by matching probabilities and those obtained by matching moments is the lack of invariance of the latter under one-to-one reparameterization. It may be interesting to find conditions under which the moment matching prior agree with Jeffreys' prior $I^{1/2}(\theta)$ or the uniform constant prior. The former holds if and only if $g_3(\theta) = -I'(\theta)$, while the latter holds if and only if $g_3(\theta) = 0$.

The if part of the above results are immediate from (4). To prove the only if parts, note that if $\pi(\theta) = I^{1/2}(\theta)$ in (4), first taking logarithms, and then differentiating with respect to θ, one gets $\frac{I'(\theta)}{2I(\theta)} = -\frac{g_3(\theta)}{2I(\theta)}$ so that $g_3(\theta) = -I'(\theta)$. On the other hand, if $\pi(\theta) = c$, then taking logarithms followed by differentiation with respect to θ, one gets $g_3(\theta) = 0$.

We now consider several examples to illustrate these results.

EXAMPLE 2.1. Consider the regular one-parameter exponential family of densities given by

$$f(x|\theta) = \exp[\theta x - \psi(\theta) + h(x)].$$

For the canonical parameter θ, noting that $g_3(\theta) = -\psi'''(\theta) = -I'(\theta)$, one gets $\pi(\theta) = I^{1/2}(\theta)$ which is Jeffreys' prior, and is different from Hartigan's prior $I(\theta)$. On the other hand, for the population mean $\phi = \psi^{'}(\theta)$ which is a strictly increasing function of θ (since $\psi^{''}(\theta) = V(X|\theta) > 0$), by Theorem 2.3,

$$\pi^*_M(\phi) = \pi_M(\theta)\left(\frac{d\theta}{d\phi}\right)^{\frac{3}{2}} = I^{1/2}(\theta)\left\{\frac{1}{\psi^{''}(\theta)}\right\}^{3/2} = \left(\psi^{''}(\theta)\right)^{-1} = I(\phi),$$

since $I(\phi) = I(\theta)\left(\frac{d\theta}{d\phi}\right)^2 = \psi^{''}(\theta)\left(\psi^{''}(\theta)\right)^{-2} = \left(\psi^{''}(\theta)\right)^{-1}$. Here, instead of Jeffreys' prior, namely the square root of the Fisher information, one gets the Fisher information itself. In particular, for the binomial proportion p, this leads to the Haldane prior $\pi_H(p) \propto p^{-1}(1-p)^{-1}$ which is the same as Hartigan's maximum likelihood prior.

EXAMPLE 2.2. Consider next the general one-parameter location family of distribution with $f(x|\theta) = p(x-\theta)$, where p is a probability density function. Then $I(\theta) = \int (p^{'}(x)/p(x))^2 p(x)dx = -\int \frac{d}{dx}(p^{'}(x)/p(x))dx = c_2$ (say) and $g_3(\theta) = -\int \frac{d^2}{dx^2}(p^{'}(x)/p(x))p(x)dx = c_3$ (say) so that $\pi_M(\theta) = \exp(-c_3\theta/2c_2)$. In the special case of symmetric location family, $c_3 = 0$ so that $\pi_M(\theta) = 1$. Consider now a nonsymmetric extreme-value distribution with $f(x|\theta) = \exp[x - \theta - \exp(x-\theta)]$. Now $p(x) = \exp[x - \exp(x)]$ so that $p^{'}(x)/p(x) = 1 - \exp(x)$. Then, $I(\theta) = \int_{-\infty}^{+\infty} \exp[2x - \exp(x)]dx = 1$ and $g_3(\theta) = -\int \exp[2x - \exp(x)]dx = -1$. Hence, $\pi_M(\theta) = \exp(\frac{1}{2}\theta)$. Jeffreys' prior $\pi_J(\theta) = 1$ in this case.

EXAMPLE 2.3. Next, we consider the general scale-family of distributions with $f(x|\theta) = \theta^{-1}p(x/\theta)$, where p is a probability density function and $\theta > 0$. Then, $\log f(x|\theta) = -\log\theta + h(x/\theta)$, where $h(x) = \log p(x)$. Hence,

$$\frac{d\log f(x|\theta)}{d\theta} = -\theta^{-1} - x\theta^{-2}h^{'}(x/\theta);$$

$$\frac{d^2 \log f(x|\theta)}{d\theta^2} = \theta^{-2} + 2x\theta^{-3}h^{'}(x/\theta) + x^2\theta^{-4}h^{''}(x/\theta);$$

$$\frac{d^3 \log f(x|\theta)}{d\theta^3} = -2\theta^{-3} - 6x\theta^{-4}h^{'}(x/\theta) - 6x^2\theta^{-5}h^{''}(x/\theta) - x^3\theta^{-6}h^{'''}(x/\theta).$$

Hence, $I(\theta) = c_2/\theta^2$ and $g_3(\theta) = c_3/\theta^3$, where

$$c_2 = -\left[1 + 2\int xh^{'}(x)p(x)dx + \int x^2h^{''}(x)p(x)dx\right]$$

and

$$c_3 = -\left[2 + 6\int xh^{'}(x)p(x)dx + 6\int x^2h^{''}(x)p(x)dx + \int x^3h^{'''}(x)p(x)dx\right]$$

Hence, $\pi_M(\theta) = \theta^{-\frac{c_3}{2c_2}}$.

As special cases, for the $N(0,\theta^2)$ family of distributions, since $h^{'}(x) = -x$, $h^{''}(x) = -1$ and $h^{'''}(x) = 0$, $c_2 = 2$ and $c_3 = 10$. Hence, $\pi_M(\theta) = \theta^{-\frac{5}{2}}$. For the exponential distribution with scale parameter θ, $h^{'}(x) = -1$ and $h^{''}(x) = h^{'''}(x) = 0$ so that $c_2 = 1$ $c_3 = 4$. Hence, $\pi_M(\theta) = \theta^{-2}$. For the double exponential distribution with median 0 and scale parameter θ, $h^{'}(x) = -\text{sgn}(x)$ and $h^{''}(x) = h^{'''}(x) = 0$ for almost all x. Hence, $c_2 = 1$ $c_3 = 4$ and once again $\pi_M(\theta) = \theta^{-2}$. For the Cauchy distribution with location parameter 0 and scale parameter θ, $c_2 = 1/2$ and $c_3 = 3/2$. This leads to $\pi_M(\theta) = \theta^{-\frac{3}{2}}$.

Next we consider multi-parameter extension of the results of Section 2.

3 Multiparameter Extension

Suppose now $X_1, \cdots, X_n|\boldsymbol{\theta}$ be independent and identically distributed with common probability density function $f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$ is a vector-valued parameter. Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ denote the maximum likelihood estimator of $\boldsymbol{\theta}$ and $\boldsymbol{T_n} = \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}})$. Then an asymptotic expansion of the posterior $\pi_n(\boldsymbol{t}|X_1, \cdots, X_n)$ of $\boldsymbol{T_n}$ (cf. Datta and Mukerjee, 2004, p. 15) under the prior π is

$$\pi_n^*(\boldsymbol{t}|X_1, \cdots, X_n) = (2\pi)^{-\frac{p}{2}} \exp\left[-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{I_n}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}})\boldsymbol{t}\right] \times \left[1 + n^{-\frac{1}{2}}\left\{R_1(\boldsymbol{t}) + \frac{1}{6}R_3(\boldsymbol{t})\right\} + O_p(n^{-1})\right],$$

where

$$\boldsymbol{I_n}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) = -n^{-1}\sum_{i=1}^{n} \frac{d^2 \log f(X_i|\boldsymbol{\theta})}{d\boldsymbol{\theta} d\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}}$$

and

$$R_1(\boldsymbol{t}) = \sum_{j=1}^{p} t_j \left(\frac{\partial \log \pi}{\partial \theta_j}\right)\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}},$$

$$R_3(\boldsymbol{t}) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{r=1}^{p} \sum_{s=1}^{p} t_j t_r t_s \left. \frac{\partial^3 \log f(X_i|\boldsymbol{\theta})}{\partial\theta_j \partial\theta_r \partial\theta_s} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}}.$$

This leads to

$$\begin{aligned} E(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}|X_1, \cdots, X_n) &= n^{-\frac{1}{2}} E(\boldsymbol{T}_{\boldsymbol{n}}|X_1, \cdots, X_n) \\ &= n^{-1} \left(\boldsymbol{U}_{\boldsymbol{n}} + \frac{1}{2} \boldsymbol{V}_{\boldsymbol{n}} \right) + o_p(n^{-1}), \end{aligned}$$

where

$$\boldsymbol{U}_{\boldsymbol{n}} = (U_{1n}, \cdots, U_{pn})^T,$$

$$\begin{aligned} U_{jn} &= \int t_j \sum_{k=1}^{p} t_k \left. \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial\theta_k} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}} \exp\left(-\frac{1}{2} \boldsymbol{t}^T \boldsymbol{I}_{\boldsymbol{n}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) \boldsymbol{t} \right) d\boldsymbol{t} \\ &= \sum_{k=1}^{p} \left. \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial\theta_k} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}} I^{jk}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}), \end{aligned}$$

$I^{jk}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}})$ being the (j,k)th element of $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}})$; $\boldsymbol{V}_{\boldsymbol{n}} = (V_{1n}, \cdots, V_{pn})^T$,

$$V_{jn} = \sum_{k=1}^{p} \sum_{r=1}^{p} \sum_{s=1}^{p} I^{jk}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) I^{rs}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) \left. \frac{\partial^3 \log f(x|\boldsymbol{\theta})}{\partial\theta_k \partial\theta_r \partial\theta_s} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}}.$$

Accordingly,

$$nE(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}|X_1, \cdots, X_n) \xrightarrow{P} \boldsymbol{U} + \frac{1}{2}\boldsymbol{V}, \text{ where } \boldsymbol{U} = (U_1, \cdots, U_p)^T,$$

where

$$U_j = \sum_{k=1}^{p} \frac{\partial \log \pi}{\partial\theta_k} I^{jk} \text{ and } \boldsymbol{V} = (V_1, \cdots, V_p)^T,$$

$$V_j = \sum_{k=1}^{p} \sum_{r=1}^{p} \sum_{s=1}^{p} I^{jk}(\boldsymbol{\theta}) I^{rs}(\boldsymbol{\theta}) E\left[\partial^3 \log f(x|\boldsymbol{\theta}) / \partial\theta_k \partial\theta_r \partial\theta_s \right].$$

We choose a prior π which satisfies $\boldsymbol{U} + \frac{1}{2}\boldsymbol{V} = \boldsymbol{0}$, thereby leading to

$$\boldsymbol{I}^{-1}(\boldsymbol{\theta}) \nabla \log \pi = -\frac{1}{2} \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \boldsymbol{b}, \text{ where } \quad \boldsymbol{b} = (b_1, \cdots, b_p)^T,$$

$$b_k = \sum_{r=1}^{p} \sum_{s=1}^{p} E\left[\frac{\partial^3 \log f(x|\boldsymbol{\theta})}{\partial\theta_k \partial\theta_r \partial\theta_s} \right] I^{rs}(\boldsymbol{\theta}), \tag{3.1}$$

and ∇ is the gradient vector. This leads to the solution $\nabla \log \pi = -\frac{1}{2}\boldsymbol{b}$ when it exists.

We first illustrate the above general result with a few examples.

Example 3.1. Consider the multiparameter exponential family of distributions with pdf $f(\boldsymbol{X_n}|\boldsymbol{\theta}) = \exp\left[\boldsymbol{\theta}^T\boldsymbol{X_n} - \psi(\boldsymbol{\theta}) + h(\boldsymbol{X_n})\right]$, where $\boldsymbol{X_n} = (X_1, \cdots, X_n)$. Then

$$-\frac{\partial^2 \log f(\boldsymbol{X_n}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \frac{\partial^2\psi}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \boldsymbol{I}(\boldsymbol{\theta})$$

so that $E(\partial^3 \log f(X|\theta)/\partial\theta_j\partial\theta_r\partial\theta_s) = -\frac{\partial I_{rs}(\boldsymbol{\theta})}{\partial\theta_j}$. Hence,

$$\frac{\partial \log \pi}{\partial\theta_j} = \frac{1}{2}\sum_{r=1}^{p}\sum_{s=1}^{p}\frac{\partial I_{rs}(\boldsymbol{\theta})}{\partial\theta_j}\frac{B_{rs}(\boldsymbol{\theta})}{|\boldsymbol{I}(\boldsymbol{\theta})|} \tag{3.2}$$

where $B_{rs}(\boldsymbol{\theta})$ is the cofactor of $I_{rs}(\boldsymbol{\theta})$. By Lemma A.4.5 of Anderson (1986, p. 598), the right hand side of (3.2) equals to $\frac{1}{2}\frac{\partial \log|\boldsymbol{I}(\boldsymbol{\theta})|}{\partial\theta_j}$ which leads to $\pi(\boldsymbol{\theta}) = |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2}$, Jeffreys' general rule prior.

Next observe that $E(\boldsymbol{X_n}|\boldsymbol{\theta}) = \frac{\partial\psi}{\partial\boldsymbol{\theta}} = \boldsymbol{\phi}$(say), and $\boldsymbol{\phi}$ is one-to-one function of $\boldsymbol{\theta}$ since $\frac{d\boldsymbol{\phi}}{d\boldsymbol{\theta}} = \frac{\partial^2\psi}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \boldsymbol{I}(\boldsymbol{\theta})$ which is positive definite. Also, $\boldsymbol{I}(\boldsymbol{\phi}) = \boldsymbol{I}^{-1}(\boldsymbol{\theta})$. Also,

$$E\left[\frac{\partial^3 \log f(x|\boldsymbol{\phi})}{\partial\phi_k\partial\phi_r\partial\phi_s}\right] = \left[-\frac{\partial I_{rs}(\boldsymbol{\phi})}{\partial\phi_k} - \frac{\partial I_{rk}(\boldsymbol{\phi})}{\partial\phi_s}\right],$$

where $I_{rs}(\boldsymbol{\phi})$ is the rsth element of $\boldsymbol{I}(\boldsymbol{\phi})$. For multiparameter exponential family of distributions, it can be checked that $\frac{\partial I_{rs}(\boldsymbol{\phi})}{\partial\phi_k} = \frac{\partial I_{rk}(\boldsymbol{\phi})}{\partial\phi_s}$. So, equation (3.1) becomes to

$$b_k = -2\sum_{r=1}^{p}\sum_{s=1}^{p}\frac{\partial I_{rs}(\boldsymbol{\phi})}{\partial\phi_k}I^{rs}(\boldsymbol{\phi}) = -2\frac{\partial \log|\boldsymbol{I}(\boldsymbol{\phi})|}{\partial\phi_k},$$

where $I^{rs}(\boldsymbol{\phi})$ is the rs_{th} element of $\boldsymbol{I}^{-1}(\boldsymbol{\phi})$.

Then, by solving the differential equation

$$\frac{\partial \log \pi}{\partial\phi_k} = \frac{\partial \log|\boldsymbol{I}(\boldsymbol{\phi})|}{\partial\phi_k},$$

one gets $\pi(\boldsymbol{\phi}) = |\boldsymbol{I}(\boldsymbol{\phi})|$. This is a generalization of the one-parameter result proved in the previous section.

Example 3.2. Consider next the proper dispersion model introduced by Jorgensen (1997). This general two-parameter class of probability density functions is given by

$$f(x|\mu,\lambda) = a(\lambda)c(x)\exp[\lambda t(x,\mu)].$$

Now observe that

$$\frac{\partial \log f}{\partial\mu} = \lambda\frac{\partial t}{\partial\mu}, \quad \frac{\partial \log f}{\partial\lambda} = u(\lambda) + t(x,\mu),$$

where $u(\lambda) = a^{'}(\lambda)/a(\lambda)$. Accordingly,

$$\frac{\partial^2 \log f}{\partial \mu^2} = \lambda \frac{\partial^2 t}{\partial \mu^2}, \quad \frac{\partial^2 \log f}{\partial \mu \partial \lambda} = \frac{\partial t}{\partial \mu}, \quad \frac{\partial^2 \log f}{\partial \lambda^2} = u^{'}(\lambda).$$

Since $E\left(\frac{\partial t}{\partial \mu}|\mu, \lambda\right) = 0$, the Fisher information matrix is

$$I(\mu, \lambda) = \text{ Diag } [I_{\mu\mu}(\mu, \lambda), -u^{'}(\lambda)],$$

where $I_{\mu\mu}(\mu, \lambda) = -\lambda E\left[\frac{\partial^2 t}{\partial \mu^2}|\mu, \lambda\right]$. Thus μ and λ are orthogonal in the sense of Cox and Reid (1987). Further, $\frac{\partial^3 \log f}{\partial \mu^3} = \lambda \frac{\partial^3 t}{\partial \mu^3}$, $\frac{\partial^3 \log f}{\partial \mu^2 \partial \lambda} = \frac{\partial^2 t}{\partial \mu^2}$, $\frac{\partial^3 \log f}{\partial \mu \partial \lambda^2} = 0$, $\frac{\partial^3 \log f}{\partial \lambda^3} = u^{''}(\lambda)$. Now, $E\left[\frac{\partial^3 \log f}{\partial \mu^2 \partial \lambda}|\mu, \lambda\right] = E\left[\frac{\partial^2 t}{\partial \mu^2}|\mu, \lambda\right] = -\lambda^{-1} I_{\mu\mu}(\mu, \lambda)$. Thus the moment matching prior π (when it exists) must satisfy the differential equations

$$\text{(i)} \quad \frac{\partial \log \pi}{\partial \mu} = \frac{1}{2} \frac{E\left[\frac{\partial^3 t}{\partial \mu^3}|\mu, \lambda\right]}{E\left[\frac{\partial^2 t}{\partial \mu^2}|\mu, \lambda\right]}$$

and

$$\text{(ii)} \quad \frac{\partial \log \pi}{\partial \lambda} = \frac{1}{2}\left[\frac{u^{''}(\lambda)}{u^{'}(\lambda)} + \frac{1}{\lambda}\right].$$

Solving (ii), one gets $\pi = [-\lambda u^{'}(\lambda)]^{1/2} g(\mu)$, where g is an arbitrary function of μ. Now, by (i), in order that a matching prior exists,

$$\exp\left[\frac{1}{2}\int\left\{\frac{E\left(\frac{\partial^3 t}{\partial \mu^3}|\mu, \lambda\right)}{E\left(\frac{\partial^2 t}{\partial \mu^2}|\mu, \lambda\right)}\right\}\right] \tag{3.3}$$

factor into $g(\mu)h(\lambda)$ for some g and an arbitrary h. If this happens, π is of the form

$$\pi(\mu, \lambda) = g(\mu)\left[-\lambda u^{'}(\lambda)\right]^{\frac{1}{2}},$$

where g is arbitrary function of μ and $h(\lambda) = \left[-\lambda u^{'}(\lambda)\right]^{\frac{1}{2}}$.

We consider now several special cases of the above general result.

EXAMPLE 3.3. Consider the two-parameter gamma probability density function

$$f(x|\mu, \lambda) = \exp\left(-\frac{\lambda}{\mu}x\right)\frac{\lambda^\lambda x^{\lambda-1}}{\mu^\lambda}\frac{1}{\Gamma(\lambda)}.$$

With this particular parameterization, $E(x|\mu, \lambda) = \mu$. Here $a(\lambda) = \lambda^\lambda/\Gamma(\lambda)$ so that $u(\lambda) = 1 + \log \lambda - \psi(\lambda)$, where $\psi(\lambda) = \frac{d}{d\lambda}\log \Gamma(\lambda)$ is the digamma function. Thus

$u^{'}(\lambda) = \lambda^{-1} - \psi^{-1}(\lambda)$. Moreover, $t(x,\mu) = -\frac{x}{\mu} + \log\left(\frac{x}{\mu}\right)$ so that $\frac{\partial t}{\partial \mu} = \frac{x}{\mu^2} - \frac{1}{\mu}$, $\frac{\partial^2 t}{\partial \mu^2} = -\frac{2x}{\mu^3} + \frac{1}{\mu^2}$, $\frac{\partial^3 t}{\partial \mu^3} = \frac{6x}{\mu^4} - \frac{2}{\mu^3}$. Hence,

$$E\left(\frac{\partial^2 t}{\partial \mu^2}|\mu,\lambda\right) = \frac{1}{\mu^2}$$

and

$$E\left(\frac{\partial^3 t}{\partial \mu^3}|\mu,\lambda\right) = \frac{4}{\mu^3}.$$

Hence (3.3) simplifies to $\exp(-2/\mu)$. Then the moment matching prior $\pi(\mu,\lambda)$ is

$$\pi(\mu,\lambda) = \mu^{-2}[\lambda\psi^{'}(\lambda) - 1]^{\frac{1}{2}}$$

Example 3.4. (Inverse Gaussian) The probability density function

$$f(x|\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right].$$

Here $a(\lambda) = \lambda^{1/2}$ and $t(x,\mu) = -\frac{(x-\mu)^2}{2\mu^2 x} = -\frac{1}{2}\left(\frac{x}{\mu^2} - \frac{2}{\mu} + \frac{1}{x}\right)$, $u(\lambda) = \frac{1}{2\lambda}$, $u^{'}(\lambda) = -\frac{1}{2\lambda^2}$ and $u^{''}(\lambda) = \frac{1}{\lambda^3}$ so that $\frac{\partial t}{\partial \mu} = \frac{x}{\mu^3} - \frac{1}{\mu^2}$, $\frac{\partial^2 t}{\partial \mu^2} = -\frac{3x}{\mu^4} + \frac{2}{\mu^3}$ and $\frac{\partial^3 t}{\partial \mu^3} = \frac{12x}{\mu^5} - \frac{6}{\mu^6}$. Then $E\left(\frac{\partial^2 t}{\partial \mu^2}|\mu,\lambda\right) = -\frac{1}{\mu^3}$, $E\left(\frac{\partial^3 t}{\partial \mu^3}|\mu,\lambda\right) = \frac{6}{\mu^4}$. So (3.3) simplifies to $\exp(-3\log\mu) = \mu^{-3}$. Therefore, the moment matching prior $\pi(\mu,\lambda) = \mu^{-3}\lambda^{-1/2}$.

Example 3.5. (Fisher von-Mises) The probability density function

$$f(x|\mu,\lambda) = \frac{\exp[\lambda\cos(x-\mu)]}{2\pi I_0(\lambda)}, \quad \text{where } I_0(\lambda) = \frac{1}{2\pi}\int \exp(\lambda\cos x)dx.$$

Then $t(x,\mu) = \cos(x-\mu)$, $a(\lambda) = I_0^{-1}(\lambda)$. Hence, $\frac{\partial t}{\partial \mu} = \sin(x-\mu)$, $\frac{\partial^2 t}{\partial \mu^2} = -\cos(x-\mu)$, $\frac{\partial^3 t}{\partial \mu^3} = \sin(x-\mu)$, so that $E\left(\frac{\partial^3 t}{\partial \mu^3}|\mu,\lambda\right) = E\left(\frac{\partial t}{\partial \mu}|\mu,\lambda\right) = 0$. Further $u(\lambda) = -\frac{I_0^{'}(\lambda)}{I_0(\lambda)}$ and $u^{'}(\lambda) = -\frac{d}{d\lambda}\left[\frac{I_0^{'}(\lambda)}{I_0(\lambda)}\right]$. Hence,

$$\pi(\mu,\lambda) = \lambda\left\{\frac{d}{d\lambda}\left[\frac{I_0^{'}(\lambda)}{I_0(\lambda)}\right]\right\}^{\frac{1}{2}} = \lambda\left\{\frac{d^2}{d\lambda^2}\log I_0(\lambda)\right\}^{\frac{1}{2}}.$$

Example 3.6. Consider general symmetric location-scale family of distributions with probability density function $f(x|\mu,\sigma) = \frac{1}{\sigma}p(\frac{x-\mu}{\sigma})$ where $p(x) = p(-x)$. Writing $h(x) = \log p(x)$ and noting that $h^{'}(x) = -h^{'}(-x)$, $h^{''}(x) = h^{''}(-x)$ and $h^{'''}(x) = -h^{'''}(-x)$, one gets

$$E\left[\frac{\partial^2 \log f}{\partial \mu^2}\bigg|\mu,\sigma\right] = -\sigma^{-2}\int h^{''}(x)p(x)dx, \quad E\left[\frac{\partial^2 \log f}{\partial \mu \partial \sigma}\bigg|\mu,\sigma\right] = 0,$$

$$E\left[\frac{\partial^2 \log f}{\partial\sigma^2}\bigg|\mu,\sigma\right] = -\sigma^{-2}\left[1 + 2\int xh^{'}(x)p(x)dx + \int x^2h^{''}(x)p(x)dx\right],$$

$$E\left[\frac{\partial^3 \log f}{\partial\mu^3}\bigg|\mu,\sigma\right] = E\left[\frac{\partial^3 \log f}{\partial\mu\partial\sigma^2}\bigg|\mu,\sigma\right] = 0,$$

$$E\left[\frac{\partial^3 \log f}{\partial\mu^2\partial\sigma}\bigg|\mu,\sigma\right] = -\sigma^{-3}\left[2\int h^{''}(x)p(x)dx + \int xh^{'''}(x)p(x)dx\right]$$

and

$$E\left[\frac{\partial^3 \log f}{\partial\sigma^3}\bigg|\mu,\sigma\right] =$$
$$-\sigma^{-3}\left[2 + 6\int xh^{'}(x)p(x)dx + 6\int x^2h^{''}(x)p(x)dx + \int x^3h^{'''}(x)p(x)dx\right]$$

Hence, the prior π is found by solving the equations $\partial \log \pi/\partial\mu = 0$ and $\partial \log \pi/\partial\sigma = -\frac{c}{2\sigma}$, where

$$\begin{aligned} c &= \frac{2\int h^{''}(x)p(x)dx + \int xh^{'''}(x)p(x)dx}{\int h^{''}(x)p(x)dx} \\ &+ \frac{2 + 6\int xh^{'}(x)p(x)dx + 6\int x^2h^{''}(x)p(x)dx + \int x^3h^{'''}(x)p(x)dx}{1 + 2\int xh'(x)p(x)dx + \int x^2h''(x)p(x)dx} \end{aligned}$$

Hence, $\pi(\mu,\sigma) \propto \sigma^{-\frac{1}{2}c}$.

As special cases, recall that for the $N(\mu,\sigma^2)$ distribution, $h^{'}(x) = -x$, $h^{''}(x) = -1$ and $h^{'''}(x) = 0$. Hence, $c = 2 + \frac{2-6-6}{1-2-1} = 7$ so that $\pi(\mu,\sigma) \propto \sigma^{-\frac{7}{2}}$. For the Cauchy distribution with location parameter μ and scale parameter σ, recall that

$$h^{'}(x) = -\frac{2x}{1+x^2}, \quad h^{''}(x) = -\frac{2}{1+x^2} + \frac{4x^2}{(1+x^2)^2}$$

and

$$h^{'''}(x) = \frac{12x}{(1+x^2)^2} - \frac{16x^3}{(1+x^2)^3}.$$

Hence, after some simplification, $c = 4$ so that $\pi(\mu,\sigma) \propto \sigma^{-2}$ which is Jeffreys's general rule prior.

As stated before, for the multi-parameter case, the moment matching priors should satisfy the equations

$$\nabla \log \pi = -\frac{1}{2}\boldsymbol{b}. \tag{3.4}$$

If there is no solution to these equations, then moment matching priors do not exist. Here is an example for non-existence of such priors.

EXAMPLE 3.7. Consider gamma distribution with traditional parametrization which has the following density:

$$f(x|\alpha,\beta) = \exp\left(-\frac{x}{\beta}\right)\frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}.$$

Then, one gets

$$E\left[\frac{\partial^2 \log f}{\partial\alpha^2}\bigg|\alpha,\beta\right] = -\frac{d^2\log\Gamma(\alpha)}{d\alpha^2}, \quad E\left[\frac{\partial^2 \log f}{\partial\alpha\partial\beta}\bigg|\alpha,\beta\right] = -\frac{1}{\beta},$$

$$E\left[\frac{\partial^2 \log f}{\partial\beta^2}\bigg|\alpha,\beta\right] = -\frac{\alpha}{\beta^2}, \quad E\left[\frac{\partial^3 \log f}{\partial\alpha^3}\bigg|\alpha,\beta\right] = -\frac{d^3\log\Gamma(\alpha)}{d\alpha^3},$$

$$E\left[\frac{\partial^3 \log f}{\partial\alpha^2\partial\beta}\bigg|\alpha,\beta\right] = 0, \quad E\left[\frac{\partial^3 \log f}{\partial\alpha\partial\beta^2}\bigg|\alpha,\beta\right] = \frac{1}{\beta^2}$$

and

$$E\left[\frac{\partial^3 \log f}{\partial\beta^3}\bigg|\alpha,\beta\right] = \frac{4\alpha}{\beta^3}.$$

So, from (3.4), the moment matching prior should satisfy the following equations:

$$\begin{aligned}\frac{\partial\log\pi}{\partial\alpha} &= T_1(\alpha),\\ \frac{\partial\log\pi}{\partial\beta} &= \frac{1}{\beta}T_2(\alpha),\end{aligned}$$

where

$$T_1(\alpha) = \frac{\frac{d^2\log\Gamma(\alpha)}{d\alpha^2} - \alpha\frac{d^3\log\Gamma(\alpha)}{d\alpha^3}}{\alpha\frac{d^2\log\Gamma(\alpha)}{d\alpha^2} - 1}, \quad T_2(\alpha) = \frac{4\alpha\frac{d^2\log\Gamma(\alpha)}{d\alpha^2} - 1}{\alpha\frac{d^2\log\Gamma(\alpha)}{d\alpha^2} - 1}.$$

Clearly, there is no solution to the above equations. Therefore, the moment matching priors do not exist for gamma(α,β) distribution.

4 Asymptotic Properties of Credible Intervals

We show here how the results of the previous sections can be used for the construction of credible intervals based on the posterior mean and the posterior variance which match asymptotically the coverage probabilities of the corresponding confidence intervals based on the MLE. To this end, first note that for the one-parameter family of distributions satisfying the regularity conditions of Theorem 2.1,

$$\begin{aligned}V_\pi(\theta|X_1,\cdots,X_n) &= V_\pi(\theta-\hat{\theta}_n|X_1,\cdots,X_n)\\ &= E_\pi[(\theta-\hat{\theta}_n)^2|X_1,\cdots,X_n] - (\hat{\theta}^B_{n,\pi}-\hat{\theta}_n)^2.\end{aligned}$$

From (2.1), after some simplifications, it follows that

$$V_\pi(\theta|X_1, \cdots, X_n) = \frac{1}{n\hat{I}_n} + \frac{1}{n^2}\left[\frac{a_4}{2\hat{I}_n^3} + \frac{19a_3^2}{36\hat{I}_n^4} + \frac{1}{\hat{I}_n^3}\frac{\pi^{''}(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{3a_3}{\hat{I}_n^3}\frac{\pi^{'}(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{1}{\hat{I}_n^2}\left\{\frac{\pi(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\right\}^2\right] + o_p(n^{-2}).$$

Hence, for every prior π satisfying the needed regularity conditions, the leading term of $V_\pi(\theta|X_1, \cdots, X_n)$ converges in probability to $I^{-1}(\theta)$. Also it is easy to check that

$$(\theta - E(\theta|\boldsymbol{X}_n))/V^{1/2}(\theta|\boldsymbol{X}_n) = n^{1/2}(\theta - \hat{\theta}_n)\hat{I}_n^{1/2} + O_p(n^{-1}).$$

Hence, from the asymptotic normality of the MLE under less regularity conditions, it is easy to see that a two sided credible interval for θ centered at the posterior mean and scaled by the posterior standard deviation will have the same asymptotic frequentist coverage probability as the one centered at the MLE and scaled by the square root of the reciprocal of Fisher information, and this agreement holds to a high order of approximation.

5 Summary and Conclusion

The paper derives some new priors which are designed to achieve asymptotic equivalence of the posterior mean and the maximum likelihood estimator to a high order of approximation. The class of priors is characterized when the parameter of interest is real-valued as well as when it is multidimensional. One surprising finding is that even in the one-parameter case without any nuisance parameters, the proposed approach can lead to priors other than Jeffreys' prior. An important extension of these results should be in the context of prediction problem.

As suggested by a Co-editor, it may be worthwhile to generalize the definition of matching by requiring the posterior mean to be as close as possible to the MLE. In that sense, we may demand the right hand side of (3) to be quite small, but not exactly equal to zero. It is not clear though whether this will lead to an interesting extended class of priors.

Acknowledgement. This research was partially supported by a National Security Agency Grant MSPF-07G-097. The paper has benefitted much from the comments of a Co-editor and a referee.

References

Anderson, T.W. (1986). *An Introduction to Multivariate Statistical Analysis.* 2nd Edition. John Wiley, New York.

BARTLETT, M.S. (1953). Approximate Confidence Intervals. *Biometrika*, **40**, 12–19.

BERGER, J.O. and BERNARDO, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.*, **84**, 200–207.

BERGER, J.O. and BERNARDO, J.M. (1992a). On the development of the reference priors. *Bayesian Statist.*, **4**, 35–60.

BERGER, J.O. and BERNARDO, J.M. (1992b). Ordered group reference priors with application to the multinomial problem. *Biometrika*, **79**, 25–37.

BERGER, J.O. and SUN, D. (2008). Objective priors for the bivariate normal model. *Ann. Statist.*, **36**, 963–982.

BERNARDO, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, **41**, 113–147.

BICKEL, P.J. and GHOSH, J.K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction - a Bayesian argument. *Ann. Statist.*, **18**, 1070–1090.

CLARKE, B. and BARRON, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, **36**, 453–471.

CLARKE, B. and BARRON, A. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inference*, **41**, 37–60.

COX, D.R. and REID, N. (1987). Parameter orthogonality and approximate conditional inferece (with discussion). *J. Roy. Statist. Soc. B*, **49**, 1–39.

DATTA, G.S. and MUKERJEE, R. (2004). *Probability Matching Priors: Higher Order Asymptotics.* Springer, New York.

DATTA, G.S. and GHOSH, J.K. (1995a). On priors providing frequentist validity in Bayesian inference. *Biometrika*, **82**, 37–45.

DATTA, G.S. and GHOSH, J.K. (1995b). Noninformative priors for maximal invariant parameter in group models. *Test*, **4**, 95–114.

DATTA, G.S. and GHOSH, M. (1995c). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.*, **90**, 1357–1363.

DATTA, G.S. and GHOSH, M. (1996). On the invariance of noninformative priors. *Ann. Statist.*, **24**, 141–159.

DATTA, G.S., RAO, J.N.K. and SMITH,D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**, 183–196.

EATON, M.L. and SUDDERTH, W. D. (2004). Properties of right Haar predictive inference. *Sankhya*, **66**, 487–512.

FERGUSON, T.S. (1967). *Mathmetical Statistics: A Decision-Theoretic Approach.* Academic Press, New York.

GANESH, N. and LAHIRI, P. (2008). A new class of average moment matching priors. *Biometrika*, **95**, 514–520.

GHOSH, J.K. (1994). *Higher Order Asymptotics.* NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 4. Institute of Mathematical Statistics, Hayward, California.

GHOSH, J.K. and MUKERJEE, R. (1992). Non-informative priors (with discussion). In *Bayesian Statistics 4*, (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.), 195–210. Oxford University Press, New York.

GHOSH, J.K. DELAMPADY, M. and SAMANTA, T. (2006). *An Introduction to Bayesian Analysis.* Springer, New York.

GHOSH, J.K. SINHA, B.K. and JOSHI, S.N. (1982). Expansion for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III,* **1**, (S.S. Gupta and J.O. Berger, eds.), 403-456. Academic Press, New York.

GHOSAL, S. (1999). Probability matching priors for non-regular cases. *Biometrika*, bf 86, 956–964.

HARTIGAN, J.A. (1964). Invariant prior densities. *Ann. Math. Statist.*, **35**, 836–845.

HARTIGAN, J.A. (1998). The maximum likelihood prior. *Ann. Statist.*, **26**, 2083–2103.

JEFFREYS, H. (1961). *Theory of Probability.* (3rd edition.) Clarendon Press, Oxford.

JOHSON, R.A. (1970). Asymptotic expansions associated with posterior distribution. *Ann. Math. Statist.*, **41**, 851–864.

JORGENSEN, B. (1997). *The Theory of Dispersion Models.* Chapman and Hall, New York.

LINDLEY, D.V. (1956). On the measure of the information provided by an experiment. *Ann. Math. Statist.*, **27**, 986–1005.

MUKERJEE, R. and DEY, D.K. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: Higher order asymptotics. *Biometrika*, **80**, 499–505.

PEERS, H.W. (1965). On confidence sets and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. B*, **27**, 9–16.

SEVERINI, T. A., MUKERJEE, R. and GHOSH, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika*, **89**, 952–957.

STEIN, C. (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics*, (R. Zielinski, ed.), 485–514. Banach Center Publications, **16**. Polish Scientific Publishers, Warsaw.

SWEETING, T.J. (2008). On predictive probability matching priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, (B. Clarke and S. Ghoshal, eds.), 46–59. Institute of Mathematical Statistics Collections, **3**. Institute of Mathematical Statistics, Beachwood.

TIBISHIRANI, R.J. (1989). Noninformative priors for one parameter of many. *Biometrika*, **76**, 604–608.

WELCH, B.L. and PEERS, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihood. *J. Roy. Statist. Soc. B*, **25**, 318–329.

MALAY GHOSH
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
102 GRIFFIN-FLOYD HALL
P.O. BOX 118545
GAINESVILLE, FL 32611, USA
E-mail: ghoshm@stat.ufl.edu

RUITAO LIU
DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF IOWA
241 SCHAEFFER HALL
IOWA CITY, IA 52242-1409, USA
E-mail: ruitao-liu@uiowa.edu

Paper received December 2008; revised August 2010.