Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–25

Matching Prior Pairs Connecting Maximum A Posteriori Estimation and Posterior Expectation

Michiko Okudo* and Keisuke Yano†

Abstract. Bayesian statistics has two common measures of central tendency of a posterior distribution: posterior means and Maximum A Posteriori (MAP) estimates. In this paper, we discuss a connection between MAP estimates and posterior means. We derive an asymptotic condition for a pair of prior densities under which the posterior mean based on one prior coincides with the MAP estimate based on the other prior. A sufficient condition for the existence of this prior pair relates to α-flatness of the statistical model in information geometry. We also construct a matching prior pair using α-parallel priors. Our result elucidates an interesting connection between regularization in generalized linear regression models and posterior expectation.

Keywords: Bayesian inference, generalized linear regression, information geometry, prior selection.

1 Introduction

In Bayesian statistics, two common measures of central tendency of a posterior distribution are posterior mean and Maximum A Posteriori (MAP) estimate. Posterior mean is the Bayes estimate, an estimate minimizing the expected loss for a squared-error loss function. This is usually computed by the expectation using Markov chain Monte Carlo (MCMC). MAP estimate lies in the literature of penalized likelihood estimate or regularized maximum likelihood estimate. This is obtained by the optimization. Although the computational schemes of two estimates are different, the celebrated Bernstein–von-Mises theorem tells that in the first-order asymptotic regime of the sample size, the posterior shape becomes Gaussian with the center equal to the MAP estimate. Thus the posterior mean and the MAP estimate become the same in the asymptotic regime. Yet, practical behaviours of these estimates are quite different (e.g., Pananos and Lizotte, 2020). Recent studies (Gribonval, 2011; Gribonval and Machart, 2013; Louchet and Moisan, 2013; Burger and Lucka, 2014) highlight differences and connections between these estimates in several statistical models. In particular, Gribonval and Machart (2013) reveals that in Gaussian linear inverse problems, although the posterior mean and the MAP estimate for the same prior may be different, every posterior mean based on a prior is also the MAP estimate based on a different prior.

*Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN, okudo@mist.i.u-tokyo.ac.jp

†The Institute of Statistical Mathematics, 10-3 Midori cho, Tachikawa City, Tokyo, 190-8562, Japan, yano@ism.ac.jp

 https://doi.org/10.1214/24-BA1500

To elucidate a further connection between MAP estimates and posterior means in general statistical models, this paper derives the asymptotic condition for a pair of priors $(\pi, \widetilde{\pi})$ under which the posterior mean derived from one prior π coincides with the MAP estimate based on the other prior $\widetilde{\pi}$. We call this pair of priors matching prior pair. From our discovery of matching prior pairs, we see that in a generalized linear regression model, although the posterior mean based on a Gaussian prior may be different from the ridge regression (the MAP estimate based on a Gaussian prior), the matching prior pair of the Gaussian prior can deliver the MAP estimate closer to the posterior mean based on the Gaussian prior in asymptotic regimes. Although our main discovery is theoretical, it can also have practical implications, in particular, in the computation of each estimate. When there exists a difficulty in optimizing the log posterior density to obtain a MAP estimate for a prior, we can utilize the posterior mean based on another prior that forms a matching pair with the given prior. In contrast, when it is hard to build an MCMC for computing a posterior mean, we can instead evaluate a MAP estimate that matches to the posterior mean. This point will be further clarified by numerical experiments in Section 3.

The existence of a matching prior pair has an information-geometrical flavor (Amari, 1985). The information geometry presents a class of α-connections ($\alpha \in \mathbb{R}$) concerning the manifold of probability distributions. We show that a matching prior pair exists for an α-affine parameterization, that is, the parameterization with α-connection equal to 0. Further, we also provide an explicit construction of the matching prior pair using α-parallel priors (Takeuchi and Amari, 2005). This information-geometrical notation appears because the posterior expectation elicits the information about the flatness of the statistical model with respect to the (-1)-connection as observed in Komaki (1996) and in Okudo and Komaki (2021).

There is a literature on bridging the gap between the MAP estimation and the posterior expectation. In the objective Bayesian literature, a prior yielding the posterior mean asymptotically equal to the maximum likelihood estimate (MLE) is called the moment matching prior. Ghosh and Liu (2011) derives a formula for constructing a moment matching prior. Hashimoto (2019) extends the construction to non-regular statistical models. Yanagimoto and Miyata (2023) extends the moment matching prior to the conditional inference. Our matching prior pair includes the moment matching prior and naturally extends its idea to the MAP estimate based on a non-uniform distribution. Gribonval and Machart (2013) reveals an elegant construction of an exact matching prior pair for linear inverse problems. Polson and Scott (2016) proposes an exact prior pair that matches a density with a MAP estimate plugged-in and a marginal density. These results are exact in the sense that it holds even in the finite regime of sample size but are limited to several models. Although our construction relies on the asymptotics with respect to the sample size, it elucidates the connection in general statistical models using information geometry.

The rest of this paper is structured as follows. Section 2 delivers the main result, an information-geometrical construction of a matching prior pair. Section 2.3 displays analytical examples that examine the main result. Section 3 presents numerical examples using synthetic and real data. All technical proofs are presented in Section 4.

2 Matching prior pairs

We first prepare several notations for the theory and then present the construction of the matching prior pair.

2.1 Preparation

Let $\mathcal{Y}$ be a sample space and let dy be a base measure. Assume that we have observations $y^n = \{y(1), y(2), \dots, y(n)\}$ $(y(1), \dots, y(n) \in \mathcal{Y})$ independently distributed according to a probability distribution with a density function $p(y\,;\,\theta)$ that belongs to a statistical model parameterized by θ:

$$\mathcal{P} = \{p(y\,;\,\theta) \mid \theta \in \Theta\} \text{ with } \Theta \subset \mathbb{R}^d.$$

We denote by E_θ the expectation with respect to the density with θ.

For the theory, we first introduce several information-geometric notations; for details, see Amari (1985). Components of the Fisher information matrix $g = (g_{ab})_{a,b=1,\dots,d}$ are defined as

$$g_{ab}(\theta) := \int p(y\,;\,\theta)\{\partial_a \log p(y\,;\,\theta)\}\{\partial_b \log p(y\,;\,\theta)\}dy = \int \frac{\partial_a p(y\,;\,\theta)\partial_b p(y\,;\,\theta)}{p(y\,;\,\theta)}dy,$$

where $\partial_a = \partial/\partial\theta^a$. For $a, b = 1, \dots, d$, let g^{ab} be a component of the inverse matrix of the Fisher information matrix g. For $a, b, c = 1, \dots, d$, the m-connection coefficient (the (-1)-connection coefficient) and e-connection coefficient (the 1-connection coefficient) are defined as

$$\begin{aligned} \overset{\mathrm{m}}{\Gamma}_{abc} &:= \int \frac{\partial_a\partial_b p(y\,;\,\theta)\partial_c p(y\,;\,\theta)}{p(y\,;\,\theta)}dy \quad \text{and} \\ \overset{\mathrm{e}}{\Gamma}_{abc} &:= \int p(y;\theta)\{\partial_a\partial_b \log p(y\,;\,\theta)\}\{\partial_c \log p(y\,;\,\theta)\}dy, \end{aligned} \tag{1}$$

respectively. For $a, b, c = 1, \dots, d$, let

$$T_{abc} := \overset{\mathrm{m}}{\Gamma}_{abc} - \overset{\mathrm{e}}{\Gamma}_{abc} = \int p(y;\theta)\{\partial_a \log p(y\,;\,\theta)\}\{\partial_b \log p(y\,;\,\theta)\}\{\partial_c \log p(y\,;\,\theta)\}dy.$$

Further, the α-connection coefficient for $\alpha \in \mathbb{R}$ is defined as

$$\overset{\alpha}{\Gamma}_{abc} := \overset{\mathrm{m}}{\Gamma}_{abc} - \frac{1+\alpha}{2}T_{abc} = \overset{\mathrm{e}}{\Gamma}_{abc} + \frac{1-\alpha}{2}T_{abc} \quad (a, b, c = 1, \dots, d). \tag{2}$$

These connections form dual connections, that is,

$$\partial_a g_{bc}(\theta) = \overset{\alpha}{\Gamma}_{abc}(\theta) + \overset{-\alpha}{\Gamma}_{acb}(\theta) \quad \text{for any } \alpha \in \mathbb{R}. \tag{3}$$

To ease the notation, we use the Einstein summation convention: if an index occurs twice in any one term, once as an upper and once as a lower index, summation over that index is implied. Let

$$T_a := T_{abc}g^{bc} \quad \text{and} \quad \overset{\alpha}{\Gamma}{}_{ab}^{\;c} := \overset{\alpha}{\Gamma}_{abe}g^{ce} \quad \text{for } a, b, c = 1, \dots, d.$$

We then prepare notions of flatness in the information geometry. For given $\alpha \in \mathbb{R}$, the statistical model $\mathcal{P}$ is called α-flat if and only if there exists a parameterization with $\overset{\alpha}{\Gamma}_{abc} = 0$ for all $a, b, c = 1, \ldots, d$; e.g., p. 47 of Amari (1985). The parameterization with $\overset{\alpha}{\Gamma}_{abc} = 0$ for $a, b, c = 1, \ldots, d$ is called an α-affine coordinate. Further, the statistical model is said to be statistically equi-affine when $\partial_a T_b = \partial_b T_a$ for $\theta \in \Theta$; e.g., Definition 2 of Takeuchi and Amari (2005). The concept of statistical equi-affinity is important to the existence of the subsequent α-parallel priors and we have a handy sufficient condition for the statistically equi-affinity as described below.

Lemma 2.1 (Lauritzen (1987); Propositions 3 and 4 of Takeuchi and Amari (2005))**.** *If the model is α-flat for a certain $\alpha \neq 0$, it is statistically equi-affine.*

We thirdly introduce α-parallel priors. For $\alpha \in \mathbb{R}$, an α-parallel prior π_α proposed by Takeuchi and Amari (2005) is defined by

$$\partial_a \log \pi_\alpha(\theta) = \overset{\alpha}{\Gamma}{}_{ab}^{\ \ b}(\theta) \ \ (a = 1, \ldots, d) \tag{4}$$

if it exists. This class contains several non-informative priors proposed in the objective Bayesian literature (e.g., Tanaka, 2023). First, it includes the well-known Jeffreys prior $\pi_{\mathrm{J}}(\theta) := |g(\theta)|^{1/2}$ with the determinant $|\cdot|$ as the 0-parallel prior:

$$\partial_a \log \pi_{\mathrm{J}}(\theta) = \frac{1}{2}\partial_a \log |g|(\theta) = \frac{1}{2} g^{bc}(\theta)\partial_a g_{bc}(\theta) = \overset{0}{\Gamma}{}_{ab}^{\ \ b}(\theta) \ \ (a = 1, \ldots, d).$$

Second, it has the χ^2-prior $\pi_{\chi^2}(\theta)$ proposed by Liu et al. (2014) as 1/2-parallel prior:

$$\partial_a \log \pi_{\chi^2}(\theta) = \overset{1/2}{\Gamma}{}_{ab}^{\ \ b}(\theta) \ \ (a = 1, \ldots, d),$$

which is recently pointed out by Tanaka (2023). For $\alpha = 1$, we call this e-parallel prior π_{e}, and for $\alpha = -1$, we call this m-parallel prior π_{m}. Takeuchi and Amari (2005) find the following lemmas for the existence of α-parallel priors.

Lemma 2.2 (Proposition 2 of Takeuchi and Amari (2005))**.** *A statistically equi-affine statistical model has α-parallel priors for any $\alpha \in \mathbb{R}$. Otherwise, the model has only the* 0*-parallel prior.*

Lemma 2.3 (Proposition 5 of Takeuchi and Amari (2005))**.** *If a statistical model is α_0-flat for a certain $\alpha_0 \neq 0$, then the model has α-parallel priors for arbitrary $\alpha \in \mathbb{R}$.*

Example 2.1 (Exponential family)**.** *Consider an exponential family with a sufficient statistic $T(y) \in \mathbb{R}^d$:*

$$\{\exp\{\theta^\top T(y) - \psi(\theta)\}m(y) \mid \theta \in \Theta\},$$

where $\psi(\theta)$ is the potential function and $m(y)$ is a given function. The exponential family is e- *&* m-*flat (±1-flat), that is, it has parameterizations θ and $\eta = \mathrm{E}_\theta[T(Y)]$ with $\overset{\mathrm{e}}{\Gamma}{}_{ab}^{\ \ c}(\theta) = 0$ and $\overset{\mathrm{m}}{\Gamma}{}_{ab}^{\ \ c}(\eta) = 0$, respectively. Thus, Lemma 2.3 implies that exponential families have α-parallel priors for arbitrary α. In particular, the* e*-parallel prior is a uniform prior with respect to θ, and the* m*-parallel prior is a uniform prior with respect to η.*

We conclude this subsection by introducing a condition on the existence of the solution of a certain partial differential equation. For a differentiable function $h(\theta) = (h_1(\theta), \dots, h_d(\theta))^\top$, consider the partial differential equation of $u(\theta)$ given by

$$\partial_a u(\theta) = h_a(\theta) \quad (a = 1, \dots, d).$$

Lemma 2.4 (p. 348 of Zwillinger (1997); Section 1-3 of Matsuda (1976))**.** *If $\partial_b h_a(\theta) = \partial_a h_b(\theta)$ for $1 \le a < b \le d$, then $u(\theta)$ exists locally.*

2.2 Main results

We present information-geometrical condition and construction of matching prior pairs. We begin with an asymptotic condition of matching prior pairs for general models, and then we derive a simple form for an α-affine coordinate in a statistically equi-affine model. Lastly, we derive matching prior pairs for other statistics including variance and higher-order moments. In the rest of the paper, we assume regularity conditions for asymptotic expansions of posterior mean estimators and MAP estimators, and we consider only the class of priors that satisfy the regularity conditions. For the details of the regularity conditions, see Okudo and Yano (2024).

The following theorem delivers an asymptotic condition to match a posterior mean based on a prior π_{PM} and a MAP estimate based on a prior π_{MAP} except for terms of $\mathrm{o}_p(n^{-1})$. The proof is given in Section 4.

Theorem 2.1. *The posterior mean $\hat{\theta}_{\mathrm{PM}}$ based on a prior π_{PM} and the MAP estimate $\hat{\theta}_{\mathrm{MAP}}$ based on a prior π_{MAP} coincide except for terms of $o_p(n^{-1})$ when the prior pair $(\pi_{\mathrm{PM}}, \pi_{\mathrm{MAP}})$ satisfies, for $a = 1, \dots, d$,*

$$\partial_a \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{MAP}}}(\hat{\theta}_{\mathrm{MLE}}) - \left(\partial_a \log \pi_{\mathrm{J}}(\hat{\theta}_{\mathrm{MLE}}) + \frac{1}{2} g^{cd}(\hat{\theta}_{\mathrm{MLE}}) \overset{\mathrm{e}}{\Gamma}_{cda}(\hat{\theta}_{\mathrm{MLE}}) \right) = \mathrm{o}_p(n^{-1}), \qquad (5)$$

where $\hat{\theta}_{\mathrm{MLE}}$ is the MLE.

Remark 2.1. *Theorem 2.1 includes the construction of the moment matching prior proposed by Ghosh and Liu (2011). The moment matching prior is the prior that yields the posterior mean asymptotically equal to the MLE. Tanaka (2023) rewrites the partial differential equation for the moment matching prior π_{MM} in an information-geometrical way:*

$$\partial_a \log \pi_{\mathrm{MM}}(\theta) - \left(\partial_a \log \pi_{\mathrm{J}}(\theta) + \frac{1}{2} g^{cd}(\theta) \overset{\mathrm{e}}{\Gamma}_{cda}(\theta) \right) = 0 \;\; (a = 1, \dots, d).$$

As the prior that yields MLE as the MAP estimate is a uniform prior, the moment matching prior π_{MM} satisfies (5) as the matching prior pair of a uniform prior.

Remark 2.2. *From the objective Bayesian perspective, the usage of the MAP estimate is somewhat controversial as the MAP estimate is not invariant with respect to the parameterization; see Druilhet and Marin (2007) for the literature. To resolve this issue, Druilhet and Marin (2007) proposes JMAP estimate, that is, the MAP estimate obtained from the original prior π divided by the Jeffreys prior π_{J}. Our result also tells the* *connection between the posterior mean based on a prior* π_{PM} *and the JMAP estimate based on a prior* π_{JMAP}. *Putting* $\pi_{\mathrm{JMAP}}/\pi_{\mathrm{J}}$ *to* π_{MAP}, *equation (5) becomes*

$$\partial_a \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{JMAP}}}(\hat{\theta}_{\mathrm{MLE}}) - \frac{1}{2} g^{cd}(\hat{\theta}_{\mathrm{MLE}}) \overset{\mathrm{e}}{\Gamma}_{cda}(\hat{\theta}_{\mathrm{MLE}}) = o_p(n^{-1}),$$

which implies seeking a prior pair between the posterior mean and the JMAP estimate is more directly related to the behavior of e*-connection coefficients.*

Remark 2.3. *In connection with the invariance in the previous remark, we should mention the lack of invariance in a matching prior pair. In fact, equation (5) is not invariant with respect to parameterization, which implies that an explicit form of a matching prior pair depends on parameterization. Consider changing the parameterization* θ *to* ξ. *By using change of variables, equation (5) for the new parameterization* ξ *becomes*

$$\frac{\partial \theta^a}{\partial \xi^{a'}}(\hat{\theta}_{\mathrm{MLE}}) \left\{ \frac{\partial}{\partial \theta^a} \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{MAP}}}(\hat{\theta}_{\mathrm{MLE}}) - \left(\frac{\partial}{\partial \theta^a} \log \pi_{\mathrm{J}}(\hat{\theta}_{\mathrm{MLE}}) + \frac{1}{2} g^{cd}(\hat{\theta}_{\mathrm{MLE}}) \overset{\mathrm{e}}{\Gamma}_{cda}(\hat{\theta}_{\mathrm{MLE}}) \right) \right.$$
$$\left. - \frac{1}{2} g^{cd}(\hat{\theta}_{\mathrm{MLE}}) g_{ab}(\hat{\theta}_{\mathrm{MLE}}) \frac{\partial \xi^{c'}}{\partial \theta^c}(\hat{\theta}_{\mathrm{MLE}}) \frac{\partial \xi^{d'}}{\partial \theta^d}(\hat{\theta}_{\mathrm{MLE}}) \frac{\partial^2 \theta^b}{\partial \xi^{c'} \partial \xi^{d'}}(\hat{\theta}_{\mathrm{MLE}}) \right\} = o_p(n^{-1}).$$

So, even the existence of a matching prior pair depends on parameterization. This is mainly because the connection coefficient is not a tensor, and is reasonable because the form of the posterior mean itself changes according to the parameterization. Given this fact, we shall discuss the existence and the construction of a matching prior pair below.

In a one-dimensional statistical model, an explicit construction of a matching prior pair is easy. Consider the following ordinary differential equation:

$$\frac{d}{d\theta} \log \frac{\pi_{\mathrm{PM}}(\theta)}{\pi_{\mathrm{MAP}}(\theta)\pi_{\mathrm{J}}(\theta)} = \frac{1}{2} g^{11}(\theta) \overset{\mathrm{e}}{\Gamma}_{111}(\theta).$$

The integration with respect to θ yields, for arbitrary $\theta_0 \in \Theta$,

$$\frac{\pi_{\mathrm{PM}}(\theta)}{\pi_{\mathrm{MAP}}(\theta)} \propto \pi_{\mathrm{J}}(\theta) \exp \left\{ \int_{\theta_0}^{\theta} \left(\frac{1}{2} g^{11}(\theta') \overset{\mathrm{e}}{\Gamma}_{111}(\theta') d\theta' \right) \right\}.$$

Yet, in a multi-dimensional statistical model, even the existence of a matching prior pair is non-trivial. We then seek a sufficient condition for the existence of a matching prior pair and an explicit construction of the pair. The following corollary provides a sufficient condition of the existence and a explicit construction using information geometry.

Corollary 2.1. *Assume that the model is statistically equi-affine and* θ *is* α*-affine for a certain* $\alpha \in \mathbb{R}$. *Then, a matching prior pair exists. Further, the prior pair* $(\pi_{\mathrm{PM}}, \pi_{\mathrm{MAP}})$ *satisfying*

$$\frac{\pi_{\mathrm{PM}}(\theta)}{\pi_{\mathrm{MAP}}(\theta)} \propto \pi_{\mathrm{J}}(\theta) \left(\frac{\pi_{\mathrm{e}}(\theta)}{\pi_{\mathrm{m}}(\theta)} \right)^{(1-\alpha)/4} \tag{6}$$

is a matching prior pair; that is, the posterior mean based on π_{PM} *and the MAP estimate based on* π_{MAP} *coincide except for terms of* $o_p(n^{-1})$.

Proof. Observe that we have $\overset{\mathrm{e}}{\Gamma}_{abc}(\theta) = -\{(1-\alpha)/2\}T_{abc}(\theta)$ $(a, b, c = 1, \ldots, d)$ for an α-affine coordinate θ. Together with (2), this implies that for an α-affine coordinate θ, the condition (5) becomes

$$\partial_a \left(\log \frac{\pi_{\mathrm{PM}}(\theta)}{\pi_{\mathrm{MAP}}(\theta)\pi_{\mathrm{J}}(\theta)} \right) = -\frac{1-\alpha}{4} T_a(\theta) + \mathrm{o}_p(n^{-1}) \quad \text{for } \theta = \hat{\theta}_{\mathrm{MLE}}.$$

Then, consider the following partial differential equation:

$$\partial_a \left(\log \frac{\pi_{\mathrm{PM}}(\theta)}{\pi_{\mathrm{MAP}}(\theta)\pi_{\mathrm{J}}(\theta)} \right) = -\frac{1-\alpha}{4} T_a(\theta). \tag{7}$$

Lemma 2.4 tells that this equation has a solution if

$$\partial_b T_a(\theta) = \partial_a T_b(\theta) \quad (1 \leq a < b \leq d),$$

which is equal to the statistical equi-affinity of the model and thus a matching prior pair exists for an α-affine coordinate θ in a statistically equi-affine model.

Further, by the definition of the e-&m-parallel priors (4), we have

$$\partial_a \log \frac{\pi_{\mathrm{m}}(\theta)}{\pi_{\mathrm{e}}(\theta)} = T_a(\theta) \ \ (a = 1, \ldots, d),$$

and obtain

$$\partial_a \left(\log \frac{\pi_{\mathrm{PM}}(\theta)}{\pi_{\mathrm{MAP}}(\theta)\pi_{\mathrm{J}}(\theta)} \right) = \partial_a \log \left(\frac{\pi_{\mathrm{e}}(\theta)}{\pi_{\mathrm{m}}(\theta)} \right)^{(1-\alpha)/4} \qquad (a = 1, \ldots, d).$$

So, the prior pair (6) is a matching prior pair, which completes the proof. □

From Lemma 2.1, α-affine coordinates satisfy the assumption above. In the following subsections, we shall give several such examples in submodels of exponential families. However, readers may consider practical applications beyond submodels of exponential families. Although finding the matching prior pair may be difficult in general statistical models, in another direction, one-step calibration between the MAP estimate and the posterior expectation based on a prior is possible using the following corollary.

Corollary 2.2. *For a prior π, the posterior expectation $\hat{\theta}_{\mathrm{PM}}$ is calculated by*

$$\hat{\theta}^a_{\mathrm{PM}} = \hat{\theta}^a_{\mathrm{MAP}} + \frac{1}{2n} g^{ab}(\hat{\theta}_{\mathrm{MAP}}) g^{cd}(\hat{\theta}_{\mathrm{MAP}}) \left\{ \frac{1}{n} \sum_{t=1}^{n} \partial_{bcd} \log p(y(t)\,;\, \hat{\theta}_{\mathrm{MAP}}) \right\} + o_p(n^{-1}) \tag{8}$$

for $a = 1, \ldots, d$.

This calibration formula is derived from several ingredients of the proof of the main theorem. It can also be obtained through the asymptotic expansion in Miyata (2004) (see also Yanagimoto and Miyata, 2023); however, it has not been used for our specific purpose, that is, calibrating the posterior expectation and the MAP. In practice, the higher-order derivatives in the formula can be efficiently computed using automatic differentiation (cf. Iri, 1984; Baydin et al., 2018). The supplementary material Okudo and Yano (2024) checks the validity of this calibration.

We conclude this section with the following extension of matching prior pairs, that is, matching prior pairs for other statistics including higher-order moments.

Proposition 2.1. *Let $f(\theta) = (f_1(\theta), \dots, f_d(\theta))$ be a third-times differentiable function. If two priors π_{PM} and π_{MAP} satisfy*

$$\partial_a f_i(\theta) \left\{ \partial_b \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{MAP}}}(\theta) - \partial_b \log \pi_{\mathrm{J}}(\theta) - \frac{1}{2} g^{cd}(\theta) \overset{\mathrm{e}}{\Gamma}_{cdb}(\theta) \right\} - \frac{1}{2} \partial_a \partial_b f_i(\theta) = \mathrm{o}_p(n^{-1})$$

at $\theta = \hat{\theta}_{\mathrm{MLE}}$ for $i = 1, \dots, d$ and $a = 1, \dots, d$, the posterior expectation $f_{\pi_{\mathrm{PM}}}$ of f based on π_{PM} and the MAP-plugged-in estimate $f(\hat{\theta}_{\mathrm{MAP}})$ based on π_{MAP} coincide except for $\mathrm{o}_p(n^{-1})$-terms.

2.3 Examples

In this section, we present matching prior pairs (6) in a submodel of an exponential family:

$$\{\exp\{\theta^\top(\xi) T(y) - \psi(\theta(\xi))\} m(y) \mid \xi \in \Xi\},$$

where θ is the p-dimensional canonical/natural parameter of the exponential family, and ξ is a d-dimensional model parameter. This includes the exponential family itself and the generalized linear regression model with a canonical link function. As the exponential family is e-&m-flat and has α-parallel priors, their e- or m-flat submodels also have α-parallel priors and so we confine ourselves to e- or m-flat submodels of an exponential family.

Generalized linear models with canonical links and regression coefficients

We first consider a generalized linear regression model (GLM) with a canonical link function:

$$\{\exp(\theta^\top(\beta) T(y) - \psi(\theta(\beta))) m(y) \mid \beta \in \Xi\} \quad \text{with } \theta = X\beta,$$

where β is an unknown regression coefficient of d dimension ($p \geq d$), and X is a given full rank matrix $X \in \mathbb{R}^{p \times d}$. In this case, since θ is the e-affine coordinate and $\partial\theta/\partial\beta = X^\top$, the e-connection coefficient with respect to β also vanishes $\overset{\mathrm{e}}{\Gamma}_{abc}(\beta) = 0$, it suffices to seek the prior pair satisfying

$$\partial_a \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{MAP}}}(\beta) - (\partial_a \log \pi_{\mathrm{J}}(\beta)) = 0, \tag{9}$$

which is equal to

$$\frac{\pi_{\mathrm{PM}}(\beta)}{\pi_{\mathrm{MAP}}(\beta)} \propto \pi_{\mathrm{J}}(\beta). \tag{10}$$

As a simple example, consider a Gaussian model with mean zero and unknown variance: the data $y^n = \{y(1), \dots, y(n)\}$ independently come from a Gaussian distribution $N(0, \sigma^2)$. Here we employ an inverse-gamma prior

$$\sigma^2 \sim \mathrm{InvGamma}(a, b) \quad \text{with } a, b > 0.$$

Consider the canonical parameter $\theta = \sigma^{-2}$ and the posterior mean of θ. In this parameterization, the prior becomes

$$\theta \sim \mathrm{Gamma}(a, b).$$

Let π_{PM} denote its density. The posterior distribution of θ is $\mathrm{Gamma}(a + n/2, b + \sum_{i=1}^n y(i)^2/2)$, and then the posterior mean based on π_{PM} is

$$\hat{\theta}_{\mathrm{PM}} = \frac{a + n/2}{b + \sum_{i=1}^n y(i)^2/2}.$$

By using (10), we set $\pi_{\mathrm{MAP}}(\theta) \propto \pi_{\mathrm{PM}}(\theta)/\pi_{\mathrm{J}}(\theta)$, where $\pi_{\mathrm{J}}(\theta) \propto \theta^{-1}$. Since the log posterior density based on π_{MAP} is

$$\log p(y^n\,;\,\theta) + \log \pi_{\mathrm{MAP}}(\theta) = (a + n/2 - 1)\log\theta - (b + \sum\nolimits_{i=1}^n y(i)^2/2)\theta + \log\theta + C$$

with the constant C independent from θ, the MAP estimate based on π_{MAP} is

$$\hat{\theta}_{\mathrm{MAP}} = \frac{a + n/2}{b + \sum_{i=1}^n y(i)^2/2},$$

which implies the exact matching $\hat{\theta}_{\mathrm{PM}} = \hat{\theta}_{\mathrm{MAP}}$.

m-flat submodels and m-affine parameters

We proceed to a linear submodel with respect to the expectation parameter $\eta = \mathrm{E}_\theta[T(Y)]$ of the exponential family:

$$\{\exp(\theta^\top(\xi)T(y) - \psi(\theta(\xi)))m(y) \mid \xi \in \Xi\} \;\; \text{with } \eta = \eta(\theta) = X\xi,$$

where $\eta \in \mathbb{R}^p$, ξ is a model parameter of d dimension ($p \geq d$), and X is a given full rank matrix $X \in \mathbb{R}^{p\times d}$. In this case, the m-connection coefficients $\overset{\mathrm{m}}{\Gamma}_{abc}(\xi)$ $(a, b, c = 1, \dots, d)$ vanish and

$$\partial_a \log \pi_{\mathrm{J}}(\xi) = \overset{0}{\Gamma}{}^b_{ab}(\xi) = \left(\overset{\mathrm{m}}{\Gamma}{}^b_{ab}(\xi) - \frac{1}{2}T_a(\xi)\right) = -\frac{1}{2}T_a(\xi) = -\frac{1}{2}\partial_a \log\frac{\pi_{\mathrm{m}}(\xi)}{\pi_{\mathrm{e}}(\xi)},$$

which implies

$$\frac{\pi_{\mathrm{e}}(\xi)}{\pi_{\mathrm{m}}(\xi)} \propto \{\pi_{\mathrm{J}}(\xi)\}^2.$$

Then, the condition (6) becomes

$$\frac{\pi_{\mathrm{PM}}(\xi)}{\pi_{\mathrm{MAP}}(\xi)} \propto \pi_{\mathrm{J}}(\xi)\left\{(\pi_{\mathrm{J}}(\xi))^2\right\}^{1/2} = \{\pi_{\mathrm{J}}(\xi)\}^2. \tag{11}$$

As a simple example, consider a Poisson model: the data $y^n = \{y(1), \ldots, y(n)\}$ independently come from a Poisson distribution Poisson(λ). Let us consider a Gamma prior $\lambda \sim \mathrm{Gamma}(a, b)$ with $a, b > 0$ for λ and denote the density by π_{PM}. Then the posterior distribution is the Gamma distribution $\mathrm{Gamma}(a + \sum_{i=1}^n y(n), b + n)$, and the posterior mean based on π_{PM} is

$$\hat{\lambda}_{\mathrm{PM}} = \frac{a + \sum_{i=1}^n y(n)}{b + n}.$$

Since λ is the expectation parameter and $\pi_{\mathrm{J}}(\lambda) \propto \lambda^{-1/2}$, the matching prior pair satisfies

$$\frac{\pi_{\mathrm{PM}}(\lambda)}{\pi_{\mathrm{MAP}}(\lambda)} \propto \lambda^{-1}.$$

We set $\pi_{\mathrm{MAP}}(\lambda) = \lambda \pi_{\mathrm{PM}}(\lambda)$. Since the log posterior density based on π_{MAP} is

$$\log p(y^n\,;\,\lambda) + \log \pi_{\mathrm{MAP}}(\lambda) = \left(a + \sum_{i=1}^n y(i) - 1 \right) \log \lambda - (b + n)\lambda + \log \lambda + C$$

with the constant C independent from λ, the MAP estimate based on π_{MAP} is

$$\hat{\lambda}_{\mathrm{MAP}} = \frac{a + \sum_{i=1}^n y(n)}{b + n},$$

which implies the exact matching $\hat{\lambda}_{\mathrm{PM}} = \hat{\lambda}_{\mathrm{MAP}}$.

3 Numerical experiments

In this section, we examine the theory using the Bayesian logistic regression model and the Poisson shrinkage model.

3.1 The Bayesian logistic regression

Bayesian logistic regression model is a Bayesian version of the popular logistic regression model. By putting a Gaussian prior on regression coefficients, the working model becomes

$$\begin{aligned} Y(i) \mid X(i)\,,\, \beta &\sim \mathrm{Bernoulli}(\sigma(X(i)\beta)) \ (i = 1, \ldots, n), \\ \beta &\sim \mathrm{Normal}(0, I), \end{aligned}$$

where $\sigma(x) = 1/\{1 + \exp(-x)\}$. This Bayesian model has been sometimes related to the logistic ridge regression:

$$\arg\max_{\beta} \left\{ \sum_{i=1}^n l(y(i), x(i)\,;\,\beta) - \frac{\|\beta\|^2}{2} \right\}$$

with $l(y(i), x(i)\,;\,\beta) := y(i) \log \sigma(x(i)\beta) + (1 - y(i)) \log\{1 - \sigma(x(i)\beta)\}$. Our theory tells a gap between the posterior mean based on the Gaussian prior $\pi_{\mathrm{PM}}(\beta)$ and the logistic ridge regression (the MAP estimate based on π_{PM}). Also, the matching prior pair

(5) gives another prior π_{MAP} yielding the MAP estimate asymptotically equal to the posterior mean of π_{PM}

$$\pi_{\mathrm{MAP}}(\beta) \propto \left|\sum_{i=1}^{n}(X(i))^{\top}X(i)\left[\sigma(X(i)\beta)\{1-\sigma(X(i)\beta)\}\right]\right|^{-1/2}\pi_{\mathrm{PM}}(\beta),$$

inducing the following optimization:

$$\arg\max_{\beta}\sum_{i=1}^{n} l(y(i),x(i);\beta)-\frac{\|\beta\|^2}{2}-\frac{1}{2}\log\left|\sum_{i=1}^{n}(X(i))^{\top}X(i)\left[\sigma(X(i)\beta)\{1-\sigma(X(i)\beta)\}\right]\right|.$$

First, we check the behaviour of the matching prior pair by using the following two synthetic data:

$$X^{(1)}(i)=i/n,$$
$$Y^{(1)}(i)\mid X^{(1)}(i)\sim \mathrm{Bernoulli}(\sigma(X^{(1)}(i)+0.0)),$$

and

$$(Y^{(2)}(i),X^{(2)}(i))=\begin{cases}(1,i/n) & \text{if } i>n/2,\\ (0,i/n) & \text{if otherwise.}\end{cases}$$

The logistic regression model with 2-dimensional parameters (slope and intercept) applied to the former data $\{(Y^{(1)}(i),X^{(1)}(i)) : i=1,\dots,n\}$ is correctly-specified, while the model applied to the latter data $\{(Y^{(2)}(i),X^{(2)}(i)) : i=1,\dots,n\}$ is misspecified. The former data is random and so we take the mean of the performance using 50 repetitions. For the calculation of the posterior mean, we use the 10000 Markov chain Monte Carlo samples after the 10000 burnin samples by conducting the Pólya-Gamma augmentation (Polson and Scott, 2016). We vary the sample size in $\{2^t : t=4,5,6,7,8,9\}$ for the former case and in $\{2^t : t=4,5,6,7,8,9,10,11\}$ for the latter case, respectively.

Figures 1 and 2 display the results. From Figure 1, we see that under the correctly-specified model, the gap between the logistic ridge regression and the posterior mean based on the Gaussian prior is larger than the gap between the MAP estimate based on the matching prior pair and the posterior mean based on the Gaussian prior. Both gaps become smaller as the sample size gets larger. Figure 2 showcases the performance under the misspecified model. The performance with small sample sizes seems random but with moderate or large sample sizes, the MAP estimate based on the matching prior pair gets closer to the posterior mean based on the Gaussian prior. In both cases, there exists a gap between the logistic ridge regression and the posterior mean based on the Gaussian prior, and the matching prior pair reduces this gap.

We further examine the computational time for obtaining the posterior mean and the MAP estimates based on the matching prior pair in the synthetic dataset as in Figure 1. We calculate the mean and standard deviation using 10 repetitions. Table 1 indicates that the optimization is relatively fast compared to the MCMC algorithm,

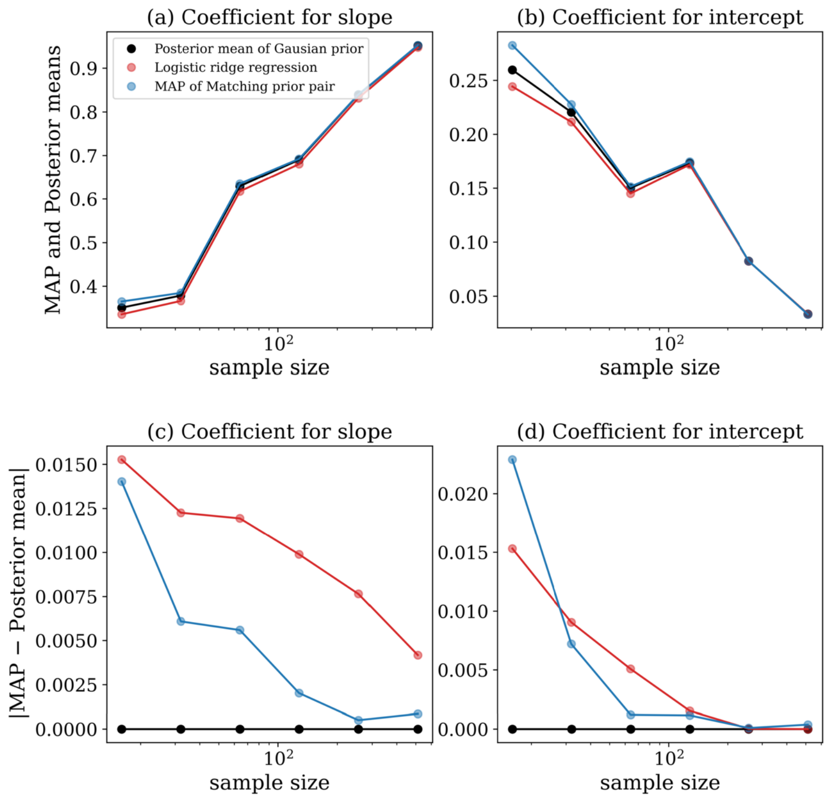

Figure 1: The posterior mean based on the Gaussian prior (colored in black), the logistic ridge regression (colored in red), and the MAP estimate based on the matching prior pair (colored in blue) under the correctly specified logistic regression model. (a)–(b) estimates themselves; (c)–(d) the differences with respect to the posterior mean based on the Gaussian prior. The horizontal axis adopts a logarithmic scale.

Sample size n	Posterior mean	MAP of Matching prior pair
2^4	4.72 s ($\pm$0.07 s)	0.01 s ($\pm$0.00 s)
2^5	5.13 s ($\pm$0.05 s)	0.01 s ($\pm$0.00 s)
2^6	6.12 s ($\pm$0.07 s)	0.02 s ($\pm$0.00 s)
2^7	8.15 s ($\pm$0.08 s)	0.05 s ($\pm$0.00 s)
2^8	12.7 s ($\pm$0.07 s)	0.08 s ($\pm$0.01 s)
2^9	45.8 s ($\pm$2.72 s)	0.17 s ($\pm$0.00 s)

Table 1: Computational time for obtaining the posterior mean and the MAP estimates based on the matching prior pair. The set-up is the same as in Figure 1.

particularly in regimes with large sample sizes. Therefore, the MAP estimate based on the matching prior pair can serve as a useful approximation of the posterior mean while providing fast computational times.

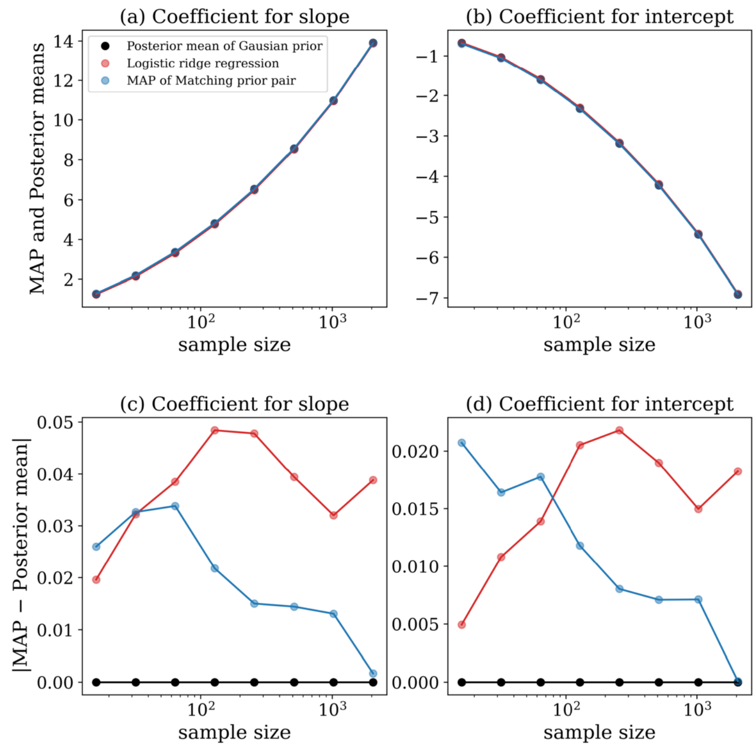

Figure 2: The posterior mean based on the Gaussian prior (colored in black), the logistic ridge regression (colored in red), and the MAP estimate based on the matching prior pair (colored in blue) under the misspecified logistic regression model. (a)–(b) estimates themselves; (c)–(d) the differences with respect to the posterior mean based on the Gaussian prior. The horizontal axis adopts a logarithmic scale.

Next, we check the performance of the matching prior pair by the banknote authentication data from UCI Machine Learning Repository (Dua and Graff, 2017). The banknote authentication data set classifies genuine and forged banknote-like specimens based on four image features (Variance, Skewness, Curtosis, and Entropy). The number of unknown parameters in this case is 4. We check the performance for sample sizes of $\{2^t : t = 4, 5, 6, 7, 8, 9, 10\}$. For each sample size, we take indices randomly taken 50 times and then take the average of the performance.

Figure 3 displays the result. For all four variables, the matching prior pair reduces the gap with respect to the posterior mean based on the Gaussian prior except for small sample sizes. Although there seem some biases for the coefficients of Skewness and Curtosis, we note that there exist deviations from the theoretical values of the posterior means due to the randomness in MCMC. Overall, the calibration based on the matching prior pair works well for the logistic regression model.

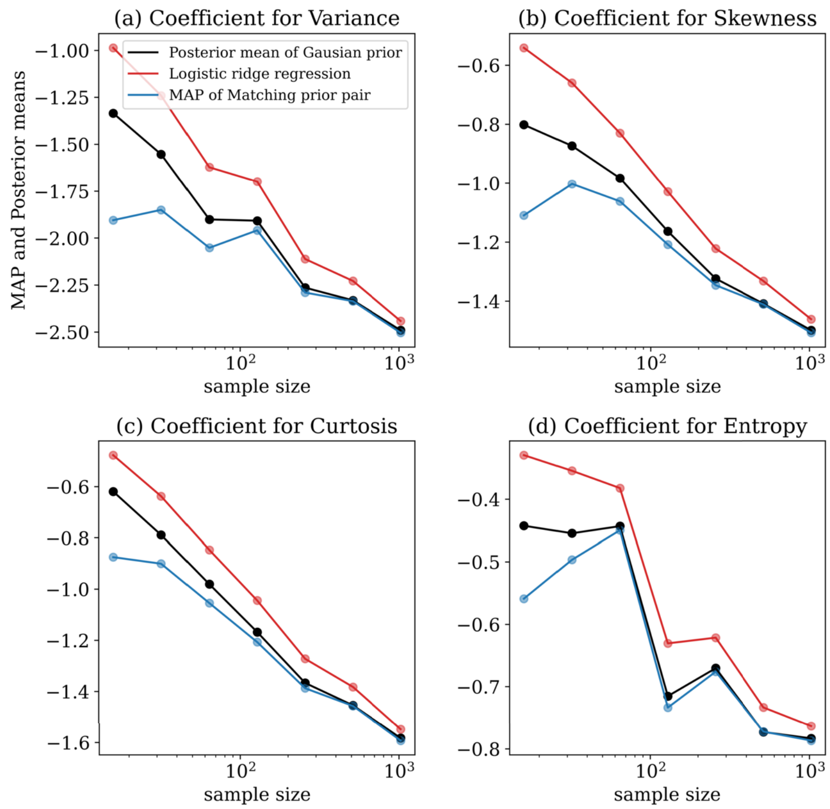

Figure 3: The Posterior mean based on the Gaussian prior (colored in black), the logistic ridge regression (colored in red), and the MAP estimate based on the matching prior pair (colored in blue) for the banknote authentification data.

3.2 The Poisson shrinkage model

Poisson sequence model is a canonical model for count-data analysis. Recently, incorporating the high-dimensional structure with Poisson sequence model has been well investigated (Komaki, 2004; Datta and Dunson, 2016; Yano et al., 2021; Hamura et al., 2022).

The working model here is

$$Y(t) = \begin{pmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_d(t) \end{pmatrix} \quad \mid \quad \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_d \end{pmatrix} \quad \sim \quad \otimes_{i=1}^{d} \mathrm{Poisson}(\lambda_i) \quad (t = 1, \ldots, n),$$

where t is an index for the observation, i is an index for the coordinate, and $\otimes$ denotes the product of measures. In this model, each observation $Y(t)$ given λ follows from the

d-dimensional independent Poisson distribution. In the application to spatio-temporal count-data analysis, t may be the index for the year and i may be the index for the observation site such as a district; see Datta and Dunson (2016); Yano et al. (2021); Hamura et al. (2022) for the details.

We investigate the calibration based on the matching prior pair in high-dimension and under an improper prior. We work with the improper shrinkage prior proposed by Komaki (2006):

$$\pi(\lambda) = \frac{\lambda_1^{\beta_1 - 1} \cdots \lambda_d^{\beta_d - 1}}{(\lambda_1 + \cdots + \lambda_d)^{\alpha}},$$

where $\alpha > 0$ and $\beta = (\beta_1, \ldots, \beta_d)$. The reason of the prior choice is as follows. The optimization in finding the MAP estimate based on this prior is a bit tricky due to the singularity around $\lambda = 0$, while we can easily access the posterior expectation as the efficient Gibbs sampling algorithm is available. So, the matching prior pair can offer useful surrogates of the MAP estimate. The number of dimension d is 100 for synthetic data analysis and is 99 for real data analysis, respectively. We set $\beta = (3, \ldots, 3)$ and $\alpha = \sum_{j=1}^{d} \beta_j - 1$. In order to avoid the singularity issue in the optimization finding the MAP estimate, we restrict the parameter space to $[10^{-3}, \infty)^d$ on the basis of the try-and-error.

We begin with displaying the numerical experiment using the following synthetic data:

$$Y(t) = \begin{pmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_{100}(t) \end{pmatrix} \quad \Big| \quad \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{100} \end{pmatrix} \quad \sim \quad \otimes_{i=1}^{100} \mathrm{Poisson}(\lambda_i) \quad (t = 1, \ldots, n),$$

$$\lambda_j = \begin{cases} 0.001 & \text{if } j \text{ is odd,} \\ 2 & \text{if otherwise} \end{cases} \quad (j = 1, \ldots, 100).$$

In this experiment, we display the result for one realization because the result is not so much dependent on realization. For the calculation of the posterior mean, we use 10000 MCMC samples.

Figure 4 showcases the MAP estimate based on the shrinkage prior (colored in black), the posterior mean based on the shrinkage prior (colored in red), and the posterior mean based on the matching prior pair (colored in blue). From (c)–(d) of Figure 4, we see that the posterior mean based on the matching prior pair of the shrinkage prior can get closer to the MAP estimate based on the shrinkage prior than that based on the shrinkage prior. Surprisingly, even for high dimensional cases such as $(n, d) = (1, 100)$ and $(n, d) = (10, 100)$, the matching prior pair works well. One of potential reasons for this success in high dimension is that a Laplace approximation of the posterior distribution might still work in certain high-dimensional set-ups (e.g., Panov and Spokoiny, 2015; Yano and Kato, 2020; Kasprzaki et al., 2023). Further, we measure the computational time (CPU times) for the MAP estimate and the posterior expectation based on the matching

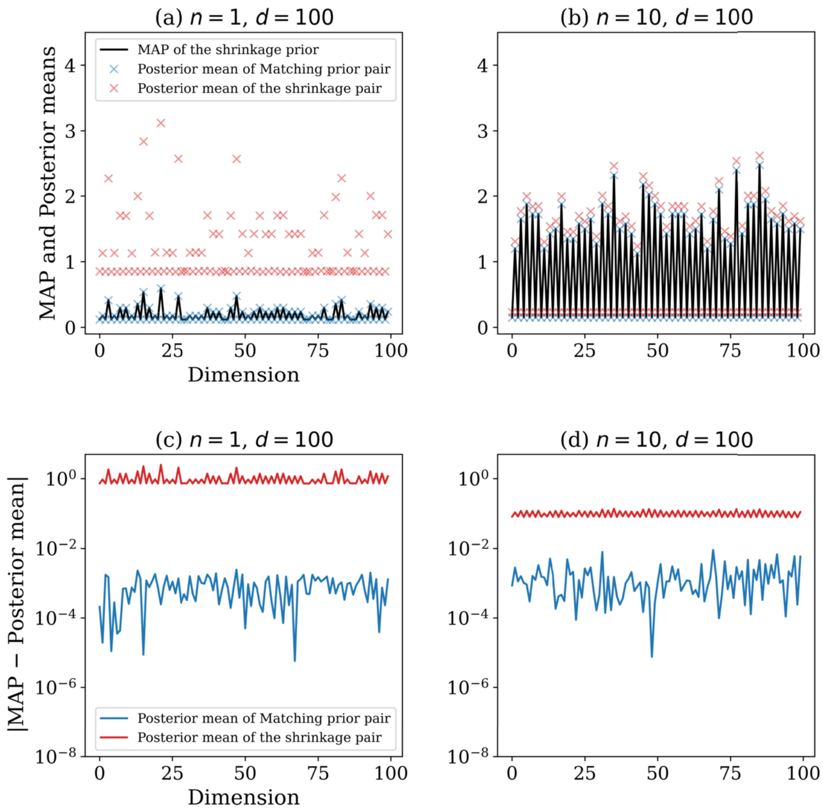

Figure 4: The MAP estimate based on the shrinkage prior (colored in black), the posterior mean based on the shrinkage prior (colored in red), and the posterior mean based on the matching prior pair (colored in blue) under the synthetic Poisson sequence model.

Sample size n	MAP	Posterior means based on Matching prior pair
1	6.38 s	0.07 s
10	3.53 s	0.06 s
100	5.12 s	0.06 s
1000	12.5 s	0.06 s

Table 2: Computational time for obtaining the MAP estimates and the posterior expectation based on the matching prior pair. The set-up is the same as in Figure 4.

prior pair. In this example, due to the singularity issue and the high-dimensionality, the optimization for the MAP estimate is relatively slow compared to the MCMC algorithm as in Table 2. Thus, this implies that the Bayesian computation using the matching prior pair can offer a good surrogate for the MAP estimate if we have an efficient MCMC algorithm and the optimization is slow or difficult.

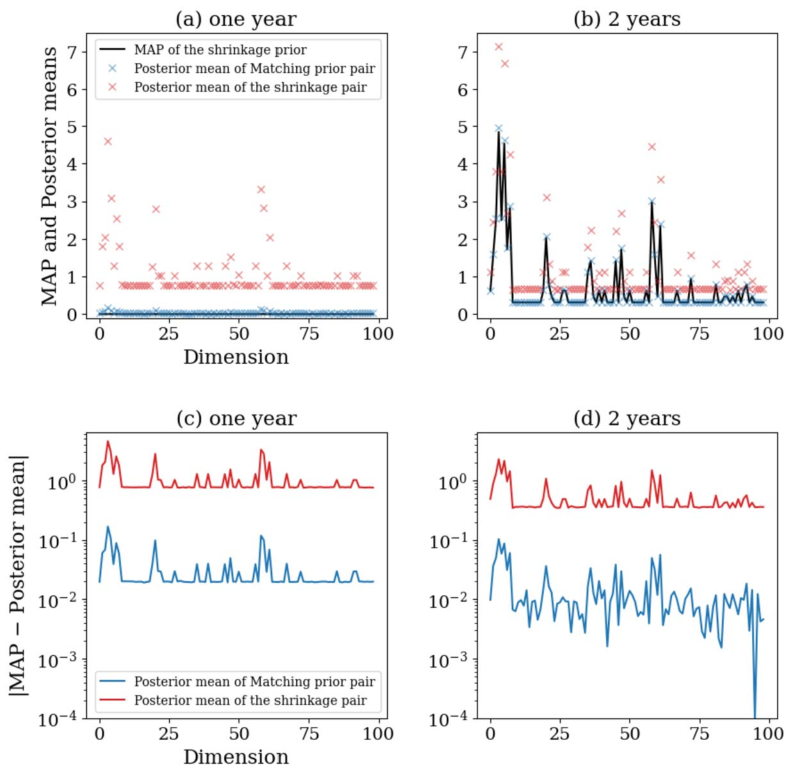

Figure 5: The MAP estimate based on the shrinkage prior (colored in black), the posterior mean based on the shrinkage prior (colored in red), and the posterior mean based on the matching prior pair (colored in blue) for pickpockets in Chuo ward, Tokyo, Japan. (a) estimates using only data for the entire year of 2012; (b) estimates using data from the two-year period of 2012–2013; (c)–(d) the differences with respect to the MAP estimate based on the shrinkage prior.

We proceed to an application to Japanese pickpocket data from Tokyo Metropolitan Police Department (2023). This data reports the total numbers of pickpockets in each year in Tokyo Prefecture, and are classified by town and also by the type of crimes. We use pickpocket data from 2012 to 2013 at 99 towns in Chuo ward. We work with the Poisson sequence model ($d = 99$, $n \leq 2$; n is the number of years we use in the analysis) and report how the matching prior pair calibrates the shrinkage prior so as to get the posterior mean closer to the MAP estimate based on the shrinkage prior. For the calculation of the posterior mean, we use 10000 MCMC samples.

Figure 5 showcases the MAP estimate based on the shrinkage prior π (colored in black), the posterior mean based on the shrinkage prior π (colored in red), and the posterior mean based on the matching prior pair of π (colored in blue) for pickpockets in Chuo ward, Tokyo, Japan. Figure 5 shows that for pickpocket data, the MAP estimate and the posterior mean based on the same shrinkage prior are different although the difference gets smaller as the sample size becomes larger, and the matching prior pair successfully yields the posterior mean closer to the MAP estimate based on the improper shrinkage prior.

4 Proofs

This section provides the proof of the main results.

Proof of Theorem 2.1. The proof employs the following asymptotic expansions for a posterior mean and a MAP estimate.

Lemma 4.1. *The posterior mean of θ based on a prior $\pi_{\mathrm{PM}}(\theta)$ is expanded as*

$$\hat{\theta}^a_{\mathrm{PM}} = \hat{\theta}^a_{\mathrm{MLE}} + \frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\left(\partial_b \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{J}}}(\hat{\theta}_{\mathrm{MLE}}) + \frac{T_b(\hat{\theta}_{\mathrm{MLE}})}{2}\right) + \frac{g^{bc}(\hat{\theta}_{\mathrm{MLE}})}{2n}\left(-\overset{\mathrm{m}}{\Gamma}{}_{bc}^{\ a}(\hat{\theta}_{\mathrm{MLE}})\right) + \mathrm{o}_p(n^{-1}) \quad (a = 1, \dots, d). \tag{12}$$

Lemma 4.2. *The MAP estimate of θ based on a prior $\pi_{\mathrm{MAP}}(\theta)$ is expanded as*

$$\hat{\theta}^a_{\mathrm{MAP}} = \hat{\theta}^a_{\mathrm{MLE}} + \frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\partial_b \log \pi_{\mathrm{MAP}}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}) \quad (a = 1, \dots, d). \tag{13}$$

The proofs of these lemmas are given right after the main proof.

These lemmas give the following condition under which the posterior mean $\hat{\theta}_{\mathrm{PM}}$ (12) and the MAP estimate $\hat{\theta}_{\mathrm{MAP}}$ (13) coincide except for $o_p(n^{-1})$ terms: for $a = 1, \dots, d$,

$$\frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\left(\partial_b \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{J}}}(\hat{\theta}_{\mathrm{MLE}}) + \frac{T_b(\hat{\theta}_{\mathrm{MLE}})}{2}\right) + \frac{g^{bc}(\hat{\theta}_{\mathrm{MLE}})}{2n}\left(-\overset{\mathrm{m}}{\Gamma}{}_{bc}^{\ a}(\hat{\theta}_{\mathrm{MLE}})\right)$$
$$= \frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\partial_b \log \pi_{\mathrm{MAP}}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}).$$

This is rewritten as

$$\begin{aligned}
& g^{ab}(\hat{\theta}_{\mathrm{MLE}})\partial_b \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{MAP}}}(\hat{\theta}_{\mathrm{MLE}}) \\
&= g^{ab}(\hat{\theta}_{\mathrm{MLE}})\left(\partial_b \log \pi_{\mathrm{J}}(\hat{\theta}_{\mathrm{MLE}}) - \frac{T_b(\hat{\theta}_{\mathrm{MLE}})}{2}\right) + \frac{g^{bc}(\hat{\theta}_{\mathrm{MLE}})}{2}\overset{\mathrm{m}}{\Gamma}{}_{bc}^{\ a}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}) \\
&= g^{ab}(\hat{\theta}_{\mathrm{MLE}})\left[\partial_b \log \pi_{\mathrm{J}}(\hat{\theta}_{\mathrm{MLE}}) - \frac{1}{2}\left\{\overset{\mathrm{m}}{\Gamma}_{cdb}(\hat{\theta}_{\mathrm{MLE}}) - \overset{\mathrm{e}}{\Gamma}_{cdb}(\hat{\theta}_{\mathrm{MLE}})\right\} g^{cd}(\hat{\theta}_{\mathrm{MLE}})\right] \\
&\quad + \frac{g^{bc}(\hat{\theta}_{\mathrm{MLE}})}{2}\overset{\mathrm{m}}{\Gamma}{}_{bc}^{\ a}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1})
\end{aligned}$$

$$= g^{ab}(\hat{\theta}_{\mathrm{MLE}})\partial_b \log \pi_{\mathrm{J}}(\hat{\theta}_{\mathrm{MLE}}) + \frac{g^{bc}(\hat{\theta}_{\mathrm{MLE}})}{2} \overset{\mathrm{e}}{\Gamma}{}_{bc}^{\ \ a}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}), \tag{14}$$

where the second identity follows since $T_{abc} = \overset{\mathrm{m}}{\Gamma}_{abc} - \overset{\mathrm{e}}{\Gamma}_{abc}$. This completes the proof. □

Proof of Lemma 4.1. In the proof, we consider an approximation of the posterior expectation of arbitrary third-times differentiable function $f : \Theta \to \mathbb{R}$. Setting $f(\theta) = \theta^a$ $(a = 1, \dots, d)$ gives the approximation of the posterior mean of θ. The following proof of Lemma 4.1 proceeds closely following the proof of Theorem III.1 in Okudo and Komaki (2021). The first step is to employ the Laplace approximation of integrals to get an approximation of the posterior expectation. The second step is to arrange terms in information-geometrical notations.

Step 1: Laplace approximation. Observe that the posterior expectation of a third-times differentiable function $f(\theta)$ based on a prior $\pi(\theta)$ is written as

$$f_\pi(y^n) = \frac{\int f(\theta)p(y^n\,;\,\theta)\pi(\theta)d\theta}{\int p(y^n\,;\,\theta)\pi(\theta)d\theta} = \frac{\int f(\theta)\exp(n\bar{L}(\theta))\pi(\theta)d\theta}{\int \exp(n\bar{L}(\theta))\pi(\theta)d\theta},$$

where $\bar{L}(\theta) = (1/n)\sum_{t=1}^n \log p(y(t)\,;\,\theta)$. We approximate this using the Laplace method (e.g., Theorem 4.6.1 of Kass and Vos (1997) and Tierney and Kadane (1986)). Consider an expansion of $\pi(\theta)\exp(n\bar{L}(\theta))$ around $\theta = \hat{\theta}_{\mathrm{MLE}}$. In the following, for any function $g(\theta)$, we abbreviate the value $g(\hat{\theta}_{\mathrm{MLE}})$ to $\hat{g}$; e.g., $\hat{\pi} := \pi(\hat{\theta}_{\mathrm{MLE}})$. By rescaling θ as $\theta = \hat{\theta}_{\mathrm{MLE}} + \phi/\sqrt{n}$, we get

$$\begin{aligned}
&\pi(\theta)\exp(n\bar{L}(\theta)) \\
&= \left(\hat{\pi} + \frac{(\partial_a\hat{\pi})\phi^a}{\sqrt{n}} + \frac{(\partial_{ab}\hat{\pi})\phi^a\phi^b}{2n} + \frac{(\partial_{abc}\hat{\pi})\phi^a\phi^b\phi^c}{6n\sqrt{n}} + \mathrm{O}_p(n^{-2})\right) \\
&\quad\times \exp\left(n\hat{L} + \frac{(\partial_{ab}\hat{L})\phi^a\phi^b}{2} + \frac{(\partial_{abc}\hat{L})\phi^a\phi^b\phi^c}{6\sqrt{n}} + \frac{(\partial_{abcd}\hat{L})\phi^a\phi^b\phi^c\phi^d}{24n}\right. \\
&\qquad\qquad \left. + \frac{(\partial_{abcde}\hat{L})\phi^a\phi^b\phi^c\phi^d\phi^e}{120n\sqrt{n}} + \mathrm{O}_p(n^{-2})\right) \\
&= \hat{\pi}e^{n\hat{L}}e^{(\partial_{ab}\hat{L})\phi^a\phi^b/2}\left(1 + \frac{(\partial_a\hat{\pi})\phi^a}{\hat{\pi}\sqrt{n}} + \frac{(\partial_{ab}\hat{\pi})\phi^a\phi^b}{2\hat{\pi}n} + \frac{(\partial_{abc}\hat{\pi})\phi^a\phi^b\phi^c}{6\hat{\pi}n\sqrt{n}} + \mathrm{O}_p(n^{-2})\right) \\
&\quad\times\left(1 + \frac{(\partial_{abc}\hat{L})\phi^a\phi^b\phi^c}{6\sqrt{n}} + \frac{(\partial_{abcd}\hat{L})\phi^a\phi^b\phi^c\phi^d}{24n}\right. \\
&\qquad\qquad \left. + \frac{(\partial_{abc}\hat{L})(\partial_{a'b'c'}\hat{L})\phi^a\phi^b\phi^c\phi^{a'}\phi^{b'}\phi^{c'}}{72n} + \mathrm{O}_p(n^{-3/2})\right) \\
&= \hat{\pi}e^{n\hat{L}}e^{-\hat{J}_{ab}\phi^a\phi^b/2}\left(1 + \frac{(\partial_a\hat{\pi})\phi^a}{\hat{\pi}\sqrt{n}} + \frac{(\partial_{abc}\hat{L})\phi^a\phi^b\phi^c}{6\sqrt{n}} + \frac{(\partial_{ab}\hat{\pi})\phi^a\phi^b}{2\hat{\pi}n}\right. \\
&\quad + \frac{(\partial_a\hat{\pi})(\partial_{bcd}\hat{L})\phi^a\phi^b\phi^c\phi^d}{6\hat{\pi}n}
\end{aligned}$$

$$+\frac{(\partial_{abc}\hat{\pi})\phi^a\phi^b\phi^c}{6\hat{\pi}n\sqrt{n}}+\frac{(\partial_{ab}\hat{\pi})(\partial_{cde}\hat{L})\phi^a\phi^b\phi^c\phi^d\phi^e}{12\hat{\pi}n\sqrt{n}}+\frac{(\partial_{a}\hat{\pi})(\partial_{bcde}\hat{L})\phi^a\phi^b\phi^c\phi^d\phi^e}{24\hat{\pi}n\sqrt{n}}$$
$$+\frac{(\partial_a\hat{\pi})(\partial_{bcd}\hat{L})(\partial_{efg}\hat{L})\phi^a\phi^b\phi^c\phi^d\phi^e\phi^f\phi^g}{72\hat{\pi}n\sqrt{n}}+\frac{C_1}{n}+\mathrm{O}_p(n^{-2})\Bigg), \tag{15}$$

where the second equation follows from $\exp(x)=1+x+O(x^2)$, and in the third equation, we denote $-\partial_{ab}\hat{L}$ by $\hat{J}_{ab}$ and denote terms not depending on π and n by C_1, respectively.

Next, we integrate both sides of (15) with respect to θ. Let $(\hat{J}^{ab})$ be the inverse matrix of $(\hat{J}_{ab})$. By changing the variables from θ to ϕ, and by using the formula of moments of multivariate Gaussian distributions, we obtain

$$\int \pi(\theta)\exp(n\bar{L}(\theta))d\theta$$
$$= C_2\hat{\pi}\left(1+\frac{(\partial_{ab}\hat{\pi})}{2\hat{\pi}n}\int\phi^a\phi^b e^{-\hat{J}_{cd}\phi^c\phi^d/2}d\phi+\frac{(\partial_a\hat{\pi})(\partial_{bcd}\hat{L})}{6\hat{\pi}n}\int\phi^a\phi^b\phi^c\phi^d e^{-\hat{J}_{ef}\phi^e\phi^f/2}d\phi\right.$$
$$\left.+\frac{C_1}{n}+\mathrm{O}_p(n^{-2})\right)$$
$$= C_2\hat{\pi}\left(1+\frac{(\partial_{ab}\hat{\pi})\hat{J}^{ab}}{2\hat{\pi}n}+\frac{(\partial_a\hat{\pi})(\partial_{bcd}\hat{L})(\hat{J}^{ab}\hat{J}^{cd}+\hat{J}^{ac}\hat{J}^{bd}+\hat{J}^{ad}\hat{J}^{bc})}{6\hat{\pi}n}+\frac{C_1}{n}+\mathrm{O}_p(n^{-2})\right)$$
$$= C_2\hat{\pi}\left(1+\frac{(\partial_{ab}\hat{\pi})\hat{J}^{ab}}{2\hat{\pi}n}+\frac{(\partial_a\hat{\pi})(\partial_{bcd}\hat{L})\hat{J}^{ab}\hat{J}^{cd}}{2\hat{\pi}n}+\frac{C_1}{n}+\mathrm{O}_p(n^{-2})\right),$$

where C_2 is a constant not depending on π and n. Replacing $\pi(\theta)$ by $f(\theta)\pi(\theta)$ for an arbitrary third-times differentiable function $f:\mathbb{R}^d\to\mathbb{R}$, we have

$$\int f(\theta)\pi(\theta)\exp(n\bar{L}(\theta))d\theta$$
$$= C_2\hat{f}\hat{\pi}\left(1+\frac{\{\partial_{ab}(\hat{f}\hat{\pi})\}\hat{J}^{ab}}{2\hat{f}\hat{\pi}n}+\frac{\{\partial_a(\hat{f}\hat{\pi})\}(\partial_{bcd}\hat{L})\hat{J}^{ab}\hat{J}^{cd}}{2\hat{f}\hat{\pi}n}+\frac{C_1}{n}+\mathrm{O}_p(n^{-2})\right).$$

Therefore, the posterior expectation of $f(\theta)$ is expanded as

$$f_\pi = \frac{\int f(\theta)\exp(n\bar{L}(\theta))\pi(\theta)d\theta}{\int\exp(nL(\theta))\pi(\theta)d\theta}$$
$$= C_2\hat{f}\hat{\pi}\left(1+\frac{\partial_{ab}(\hat{f}\hat{\pi})\hat{J}^{ab}}{2\hat{f}\hat{\pi}n}+\frac{\{\partial_a(\hat{f}\hat{\pi})\}(\partial_{bcd}\hat{L})\hat{J}^{ab}\hat{J}^{cd}}{2\hat{f}\hat{\pi}n}+\frac{C_1}{n}+\mathrm{O}_p(n^{-2})\right)$$
$$\Bigg/ C_2\hat{\pi}\left(1+\frac{(\partial_{ab}\hat{\pi})\hat{J}^{ab}}{2\hat{\pi}n}+\frac{(\partial_a\hat{\pi})(\partial_{bcd}\hat{L})\hat{J}^{ab}\hat{J}^{cd}}{2\hat{\pi}n}+\frac{C_1}{n}+\mathrm{O}_p(n^{-2})\right)$$
$$= \hat{f}\left(1+\frac{\hat{J}^{ab}}{2n}\left(\frac{\partial_{ab}(\hat{f}\hat{\pi})}{\hat{\theta}_i\hat{\pi}}-\frac{\partial_{ab}\hat{\pi}}{\hat{\pi}}\right)+\frac{\hat{J}^{ab}\hat{J}^{cd}\partial_{bcd}\hat{L}}{2n}\left(\frac{\partial_a(\hat{f}\hat{\pi})}{\hat{\theta}\hat{\pi}}-\frac{\partial_a\hat{\pi}}{\hat{\pi}}\right)+\mathrm{O}_p(n^{-2})\right)$$

$$= \hat{f} + \frac{\hat{J}^{ab}}{2n}\left(\partial_{ab}\hat{f} + \frac{2(\partial_a \hat{f})(\partial_b \hat{\pi})}{\hat{\pi}}\right) + \frac{\hat{J}^{ab}\hat{J}^{cd}\partial_{bcd}\hat{L}}{2n}\partial_a \hat{f} + \mathrm{O}_p(n^{-2}). \tag{16}$$

This completes Step 1.

Step 2: Rearrangement using the information-geometric notations. The law of large numbers yields $\hat{J}_{ab} = \hat{g}_{ab} + o_p(1)$ and $\partial_{bcd}\hat{L} = \mathrm{E}_\theta[\partial_{bcd}\hat{L}] + o_p(1)$. The Bartlett identity gives

$$\mathrm{E}_\theta[\partial_{bcd}\bar{L}] = -\partial_b g_{cd}(\theta) - \overset{\mathrm{e}}{\Gamma}_{cdb}(\theta) = -\partial_b g_{cd}(\theta) - \overset{\mathrm{m}}{\Gamma}_{cdb}(\theta) + T_{bcd}(\theta). \tag{17}$$

Together with the definition of 0-parallel prior $g^{cd}\partial_b g_{cd} = \partial_b \log(|g|) = 2\partial_b \log \pi_{\mathrm{J}}$, these give the following representation of the approximated posterior expectation of f:

$$\begin{aligned} f_\pi &= \hat{f} + \frac{\hat{g}^{ab}}{2n}\left(\partial_{ab}\hat{f} + 2\partial_a \hat{f}\partial_b \log \hat{\pi}\right) \\ &\quad + \frac{\hat{g}^{ab}}{2n}\left(-2\partial_b \log \hat{\pi}_J - \hat{g}^{cd}\overset{\mathrm{m}}{\Gamma}_{cdb}(\hat{\theta}_{\mathrm{MLE}}) + \hat{T}_b\right)\partial_a \hat{f} + \mathrm{o}_p(n^{-1}) \\ &= \hat{f} + \frac{\hat{g}^{ab}}{2n}\left(\partial_{ab}\hat{f} - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c}(\hat{\theta}_{\mathrm{MLE}})\partial_c \hat{f}\right) + \frac{\hat{g}^{ab}}{n}\left(\partial_b \log \frac{\hat{\pi}}{\hat{\pi}_J} + \frac{\hat{T}_b}{2}\right)\partial_a \hat{f} + \mathrm{o}_p(n^{-1}). \end{aligned}$$

Thus, replacing f by θ^a, we have

$$\hat{\theta}^a_{\mathrm{PM}} = \hat{\theta}^a + \frac{\hat{g}^{bc}}{2n}\left(-\overset{\mathrm{m}}{\Gamma}{}_{bc}^{a}(\hat{\theta}_{\mathrm{MLE}})\right) + \frac{\hat{g}^{ab}}{n}\left(\partial_b \log \frac{\hat{\pi}}{\hat{\pi}_J} + \frac{T_b(\hat{\theta}_{\mathrm{MLE}})}{2}\right) + \mathrm{o}_p(n^{-1}). \qquad \square$$

Proof of Lemma 4.2. Observe the definition of the MAP estimate $\hat{\theta}_{\mathrm{MAP}}$:

$$n\partial_a \bar{L}(\hat{\theta}_{\mathrm{MAP}}) + \partial_a \log \pi(\hat{\theta}_{\mathrm{MAP}}) = 0 \ (a = 1, \ldots, d),$$

where $\bar{L}(\theta) = (1/n)\sum_{t=1}^n \log p(y(t)\,;\,\theta)$. Letting $\delta = \hat{\theta}_{\mathrm{MAP}} - \hat{\theta}_{\mathrm{MLE}}$, the Taylor expansion around $\hat{\theta}_{\mathrm{MLE}}$ yields, for $a = 1, \ldots, d$,

$$\begin{aligned} 0 &= \partial_a \bar{L}(\hat{\theta}_{\mathrm{MAP}}) + \frac{1}{n}\partial_a \log \pi(\hat{\theta}_{\mathrm{MAP}}) \\ &= \partial_a \bar{L}(\hat{\theta}_{\mathrm{MLE}}) + \delta^b \partial_{ab}\bar{L}(\hat{\theta}_{\mathrm{MLE}}) + \frac{1}{n}\partial_a \log \pi(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{O}_p(\|\delta\|^2), \\ &= \delta^b \partial_{ab}\bar{L}(\hat{\theta}_{\mathrm{MLE}}) + \frac{1}{n}\partial_a \log \pi(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{O}_p(\|\delta\|^2), \end{aligned}$$

where the last equation follows since $\partial_a \bar{L}(\hat{\theta}_{\mathrm{MLE}}) = 0$ for $a = 1, \ldots, d$. Because the law of large numbers and the central limit theorem give

$$\partial_{ab}\bar{L}(\hat{\theta}_{\mathrm{MLE}}) = -g_{ab}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(1) \quad \text{and} \quad \partial_{ab}\bar{L}(\hat{\theta}_{\mathrm{MLE}}) + g_{ab}(\hat{\theta}_{\mathrm{MLE}}) = \mathrm{O}_p(1/\sqrt{n}),$$

we get

$$\delta^b g_{ab}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{O}_p(\|\delta\|/\sqrt{n}) = \frac{1}{n}\partial_a \log \pi(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{O}_p(\|\delta\|^2)$$

and completes the proof. $\square$

Proof of Corollary 2.2. Using (17) and the definition of 0-parallel prior,

$$
\begin{aligned}
g^{ab}g^{cd}\mathrm{E}_\theta[\partial_{bcd}\bar{L}] &= -g^{ab}g^{cd}\partial_b g_{cd}(\theta) - g^{ab}g^{cd}\overset{\mathrm{m}}{\Gamma}_{cdb}(\theta) + g^{ab}g^{cd}T_{bcd}(\theta) \\
&= -2g^{ab}\partial_b \log \pi_{\mathrm{J}}(\theta) - g^{cd}\overset{\mathrm{m}}{\Gamma}{}^a_{cd}(\theta) + g^{ab}T_b(\theta).
\end{aligned}
$$

From (12) and (13), we have for $a = 1, \dots, d$,

$$
\begin{aligned}
&\hat{\theta}^a_{\mathrm{PM}} - \hat{\theta}^a_{\mathrm{MAP}} \\
&= \frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\left(\partial_b \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{J}}}(\hat{\theta}_{\mathrm{MLE}}) + \frac{T_b(\hat{\theta}_{\mathrm{MLE}})}{2}\right) + \frac{g^{bc}(\hat{\theta}_{\mathrm{MLE}})}{2n}\left(-\overset{\mathrm{m}}{\Gamma}{}_{bc}^{\ \ a}(\hat{\theta}_{\mathrm{MLE}})\right) \\
&\quad - \frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\partial_b \log \pi_{\mathrm{MAP}}(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}) \\
&= \frac{g^{ab}(\hat{\theta}_{\mathrm{MAP}})}{n}\partial_b \log \frac{\pi_{\mathrm{PM}}}{\pi_{\mathrm{MAP}}}(\hat{\theta}_{\mathrm{MAP}}) + \frac{1}{2n}g^{ab}(\hat{\theta}_{\mathrm{MAP}})g^{cd}(\hat{\theta}_{\mathrm{MAP}})\partial_{bcd}\bar{L}(\hat{\theta}_{\mathrm{MAP}}) + \mathrm{o}_p(n^{-1}).
\end{aligned}
$$

In the last identity, we used $\partial_{bcd}\bar{L}(\hat{\theta}_{\mathrm{MLE}}) = \mathrm{E}_\theta[\partial_{bcd}\bar{L}(\hat{\theta}_{\mathrm{MLE}})] + \mathrm{o}_p(1)$ and $\hat{\theta}_{\mathrm{MLE}} = \hat{\theta}_{\mathrm{MAP}} + \mathrm{o}_p(1)$. When $\pi_{\mathrm{PM}} = \pi_{\mathrm{MAP}}$, we have

$$
\hat{\theta}^a_{\mathrm{PM}} - \hat{\theta}^a_{\mathrm{MAP}} = \frac{1}{2n^2}g^{ab}(\hat{\theta}_{\mathrm{MAP}})g^{cd}(\hat{\theta}_{\mathrm{MAP}})\sum_{t=1}^{n}\partial_{bcd}\log p(y(t)\,;\,\hat{\theta}_{\mathrm{MAP}}) + \mathrm{o}_p(n^{-1}). \qquad \square
$$

Proof of Proposition 2.1. This proposition is proved simply by changing Lemmas 4.1–4.2 to the following lemmas. Their proofs are straightforward and omitted.

Lemma 4.3. *For $i = 1, \dots, d$, the posterior mean of $f_i(\theta)$ based on a prior $\pi(\theta)$ is expanded as*

$$
\begin{aligned}
(f_\pi)_i =& f_i(\hat{\theta}_{\mathrm{MLE}}) + \frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\left(\partial_b \log \frac{\pi}{\pi_J}(\hat{\theta}_{\mathrm{MLE}}) + \frac{T_b(\hat{\theta}_{\mathrm{MLE}})}{2}\right)\partial_a f_i(\hat{\theta}_{\mathrm{MLE}}) \\
&+ \frac{g^{bc}(\hat{\theta}_{\mathrm{MLE}})}{2n}\left(\partial_b\partial_c f_i(\hat{\theta}_{\mathrm{MLE}}) - \overset{\mathrm{m}}{\Gamma}{}_{bc}^{\ \ a}(\hat{\theta}_{\mathrm{MLE}})\partial_a f_i(\hat{\theta}_{\mathrm{MLE}})\right) + \mathrm{o}_p(n^{-1}).
\end{aligned}
$$

Lemma 4.4. *For $i = 1, \dots, d$, a plugin of the MAP estimate $\hat{\theta}_{\mathrm{MAP}}$ of θ based on a prior $\pi(\theta)$ into a statistic $f_i(\theta)$ is expanded as*

$$
f_i(\hat{\theta}_{\mathrm{MAP}}) = f_i(\hat{\theta}_{\mathrm{MLE}}) + \frac{g^{ab}(\hat{\theta}_{\mathrm{MLE}})}{n}\partial_b \log \pi(\hat{\theta}_{\mathrm{MLE}})\partial_a f_i(\hat{\theta}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}). \qquad \square
$$

Acknowledgments

The authors thank the Editor, the handling editor, and two referees for their constructive comments that have improved the quality of this paper. The authors thank Kaoru Irie for helpful comments to the early version of this work. The authors thank Ryoya Kaneko for sharing his python codes of data preprocessing. The authors thank Takemi Yanagimoto for helpful discussions.

Funding

This work is supported by JSPS KAKENHI (JP19K20222, JP20K23316, JP21H05205, JP21K12067, JP22H00510, JP23K11024), MEXT (JPJ010217), and "Strategic Research Projects" grant (2022-SRP-13) from ROIS (Research Organization of Information and Systems).

Supplementary Material

Supplement to "Matching prior pairs connecting Maximum A Posteriori estimation and posterior expectation" (DOI: 10.1214/24-BA1500SUPP; .pdf). Python codes used in this paper (DOI: 10.5281/zenodo.13854194).

References

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer-Verlag, New York. MR0788689. doi: https://doi.org/10.1007/978-1-4612-5056-2. 2, 3, 4

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43. MR3800512. 8

Burger, M. and Lucka, F. (2014). Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators. *Inverse Problems*, 30. MR3274588. doi: https://doi.org/10.1088/0266-5611/30/11/114004. 1

Datta, J. and Dunson, D. (2016). Bayesian inference on quasi-sparse count data. *Biometrika*, 103:971–983. MR3620451. doi: https://doi.org/10.1093/biomet/asw053. 14, 15

Druilhet, P. and Marin, J. (2007). Invariant HPD credible sets and MAP estimators. *Bayesian Analysis*, 2:681–691. MR2361970. doi: https://doi.org/10.1214/07-BA227. 5

Dua, D. and Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml. 13

Ghosh, M. and Liu, R. (2011). Moment matching priors. *Sankhyā: The Indian Journal of Statistics*, 73:185–201. MR2887093. doi: https://doi.org/10.1007/s13171-011-0012-2. 2, 5

Gribonval, R. (2011). Should penalized least squares regression be interpreted as Maximum A Posteriori estimation? *IEEE Transaction on Signal Processing*, 59:2405–2410. MR2816500. doi: https://doi.org/10.1109/TSP.2011.2107908. 1

Gribonval, R. and Machart, P. (2013). Reconciling "priors" & "priors" without prejudice? *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013).* 1, 2

Hamura, H., Irie, K., and Sugasawa, S. (2022). On global-local shrinkage priors for count data. *Bayesian Analysis*, 17:545–564. MR4483230. doi: https://doi.org/10.1214/21-ba1263. 14, 15

Hashimoto, S. (2019). Moment matching priors for non-regular models. *Journal of Statistical Planning and Inference*, 203:169–177. MR3950602. doi: https://doi.org/10.1016/j.jspi.2019.03.009. 2

Iri, M. (1984). Simultaneous computation of functions, partial derivatives and estimates of rounding errors – complexity and practicality. *Japan Joumal of Applied Mathematics*, 1:223–252. MR0840797. doi: https://doi.org/10.1007/BF03167059. 8

Kasprzaki, M., Giordano, R., and Broderick, T. (2023). How good is your Laplace approximation of the Bayesian posterior? Finite-sample computable error bounds for a variety of useful divergences. arXiv:2209.14992. 15

Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley, New York. MR1461540. doi: https://doi.org/10.1002/9781118165980. 19

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, 83:299–313. MR1439785. doi: https://doi.org/10.1093/biomet/83.2.299. 2

Komaki, F. (2004). Simultaneous prediction of independent Poisson observables. *Annals of Statistics*, 32:1744–1769. MR2089141. doi: https://doi.org/10.1214/009053604000000445. 14

Komaki, F. (2006). Shrinkage priors for Bayesian prediction. *Annals of Statistics*, 34:808–819. MR2283393. doi: https://doi.org/10.1214/009053606000000010. 15

Lauritzen, S. (1987). *Statistical manifolds*, chapter 4. Institute of Mathematical Statistic. 4

Liu, R., Chakrabarti, A., Samanta, T., Ghosh, J., and Ghosh, M. (2014). On divergence measures leading to Jeffreys and other reference priors. *Bayesian Analysis*, 9:331–370. MR3216999. doi: https://doi.org/10.1214/14-BA862. 4

Louchet, C. and Moisan, L. (2013). Posterior expectation of the total variation model: Properties and experiments. *SIAM JOurnal of Imaging Science*, 6:2640–2684. MR3143828. doi: https://doi.org/10.1137/120902276. 1

Matsuda, M. (1976). *Theory of exterior differential forms*. Iwanami shoten. In Japanese. 5

Miyata, Y. (2004). Fully exponential Laplace approximations using asymptotic modes. *Journal of the American Statistical Association*, 99:1037–1049. MR2109493. doi: https://doi.org/10.1198/016214504000001673. 7

Okudo, M. and Komaki, F. (2021). Bayes extended estimators for curved exponential families. *IEEE Transactions on Information Theory*, 67:1088–1098. MR4232003. doi: https://doi.org/10.1109/TIT.2020.3035027. 2, 19

Okudo, M. and Yano, K. (2024). Supplement to "Matching prior pairs connecting Maximum A Posterior estimation and posterior expectation." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1500SUPP. 5, 8

Pananos, A. and Lizotte, D. (2020). Comparisons between Hamiltonian Monte Carlo and Maximum A Posteriori for a Bayesian model for Apixaban induction dose & dose personalization. *Proceedings of Machine Learning Research*, 126:1–20. 1

Panov, M. and Spokoiny, V. (2015). Finite sample Bernstein–von Mises theorem for semiparametric problems. *Bayesian Analysis*, 10:665–710. MR3420819. doi: https://doi.org/10.1214/14-BA926. 15

Polson, N. and Scott, J. (2016). Mixtures, envelopes and hierarchical duality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):701–727. MR3534347. doi: https://doi.org/10.1111/rssb.12130. 2, 11

Takeuchi, J. and Amari, S. (2005). α-parallel prior and its properties. *IEEE transactions on information theory*, 51(3):1011–1023. MR2237966. doi: https://doi.org/10.1109/TIT.2004.842703. 2, 4

Tanaka, F. (2023). Geometric properties of noninformative priors based on the chi-square divergence. *Frontiers in Applied Mathematics and Statistics*, 9. 4, 5

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86. MR0830567. doi: https://doi.org/10.1080/01621459.1986.10478240. 19

Tokyo Metropolitan Police Department. (2023). The number of crimes in tokyo prefecture by town and type. https://www.keishicho.metro.tokyo.lg.jp/about_mpd/jokyo_tokei/jokyo/. 17

Yanagimoto, T. and Miyata, Y. (2023). A pair of novel priors for improving and extending the conditional MLE. *Journal of Statistical Planning and Inference*, in press:106–117. MR4669626. doi: https://doi.org/10.1016/j.jspi.2023.106117. 2, 7

Yano, K. and Kato, K. (2020). On frequentist coverage errors of Bayesian credible sets in moderately high dimensions. *Bernoulli*, 26:616–641. MR4036046. doi: https://doi.org/10.3150/19-BEJ1142. 15

Yano, K., Kaneko, R., and Komaki, F. (2021). Minimax predictive density for sparse count data. *Bernoulli*, 27:1212–1238. MR4255232. doi: https://doi.org/10.3150/20-bej1271. 14, 15

Zwillinger, D. (1997). *Handbook of Differential Equations*. Academic Press, third edition. MR1149384. 5