On invariant moment matching priors for Bayesian point prediction

Shintaro Hashimoto
Department of Mathematics, Hiroshima University

May 28, 2026

Abstract

This paper derives objective priors that asymptotically match the mean of the Bayesian predictive distribution with that of the frequentist plug-in predictive distribution. This moment matching criterion was originally proposed by Ghosh and Liu (2011, Sankhya A) for estimation problems; the resulting priors are referred to as moment matching priors. In the predictive context, while the derived priors take a slightly different form from those for estimation, they exhibit a desirable invariance property under one-to-one parameter transformations. This is a feature not typically attained in estimation frameworks. Furthermore, this study characterizes asymptotically unbiased priors for point prediction.

Keywords: Higher-order asymptotics; Invariance; Matching priors, Objective priors, Point prediction, Unbiasedness

Mathematics Subject Classification: Primary 62F15; Secondary 62F10.

1 Introduction

In Bayesian statistics, the prior distribution serves not only a vehicle for incorporating prior knowledge into data analysis but also a profound mathematical object that elucidates the relationship between Bayesian and frequentist frameworks. For instance, the Jeffreys prior is renowned for its invariance under one-to-one transformations and its remarkable property of ensuring that Bayesian credible intervals achieve frequentist coverage probability up to a higher order under regularity conditions (Datta and Mukerjee, 2004). Furthermore, Ghosh and Liu (2011) investigated priors that match the maximum likelihood estimator (MLE) with the Bayesian estimator, specifically the posterior mean, under squared error loss up to a higher order; these are referred to as moment matching priors. It has been demonstrated that such priors generally differ from the Jeffreys prior and lack invariance under reparameterization.

This paper extends the concept of matching priors, typically formulated within estimation problems, to the domain of predictive inference. In this study, we focus specifically on point prediction and investigate priors that bridge the frequentist and Bayesian frameworks. While existing literature has discussed decision-theoretic properties such as second-order admissibility for Bayesian point predictions (e.g., Chang and Mukerjee, 2004), to the best of our knowledge, the development and analysis of moment matching priors in this context have not yet been explored. Specifically, we consider a predictive

version of the moment matching prior, a challenge proposed in the discussion of Ghosh and Liu (2011) that has remained unaddressed until now. We derive a prior distribution that matches the mean of the frequentist plug-in predictive distribution with the mean of the Bayesian predictive distribution up to a higher order, and we show its invariance properties alongside concrete examples. Furthermore, we consider priors for second-order unbiased Bayesian predictive means.

The motivation for matching plug-in predictive means with Bayesian predictive means is that these priors lead to Bayesian predictive means which share the asymptotic property of the plug-in predictive means based on the MLE up to a high order. If one is interested in the asymptotic bias reduction of the plug-in estimates through specific adjustments, the same adjustments apply directly to the Bayesian predictive means. In this way, it is possible to achieve a *Bayes-frequentist synthesis* of point predictions.

The remainder of this paper is organized as follows. In Section 2, after describing the basic setup, we introduce predictive moment matching priors and show the invariance property. In Section 3, building upon the ideas developed in Section 2, we provide the conditions under which the Bayesian predictive mean becomes asymptotically unbiased.

2 Predictive moment matching priors

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random vectors with a common density $f(x;\theta)$, where $\theta = (\theta_1, \ldots, \theta_p)^\top$. Assume that a future observation X_{n+1} has the same density $f(x;\theta)$ and we assume that X_{n+1} and $X = (X_1, \ldots, X_n)^\top$ are conditionally independent given θ. Consider the Bayesian prediction of a future observation X_{n+1} based on X. The Bayesian predictive density of X_{n+1} given X is defined by

$$\tilde{\pi}(x_{n+1}|X) = \int f(x_{n+1};\theta)\pi(\theta|X)d\theta,$$

where $\pi(\theta \mid X)$ is the posterior density of θ given X under the prior $\pi(\cdot)$. Let $\hat{\theta}$ be the maximum likelihood estimator of θ. We assume the same regularity conditions as Datta et al. (2000) to ensure the validity of the asymptotic expansion of the posterior distribution.

Let $\ell_n(\theta) = n^{-1}\sum_{i=1}^n \log f(X_i;\theta)$. With $\partial_j \equiv \partial/\partial\theta_j$, let $a_{jr} = \partial_j\partial_r\ell_n(\theta)|_{\theta=\hat{\theta}}$, $a_{jrs} = \partial_j\partial_r\partial_s\ell_n(\theta)|_{\theta=\hat{\theta}}$, $c_{jr} = -a_{jr}$, $\pi_j(\theta) = \partial_j\pi(\theta)$, $f_j(x;\theta) = \partial_j f(x;\theta)$, and $f_{jr}(x;\theta) = \partial_j\partial_r f(x;\theta)$. Let $C = (c_{jr})$ be the positive definite matrix, and $C^{-1} = (c^{jr})$ be the inverse matrix of C. From the result of Komaki (1996) and Datta et al. (2000), the asymptotic expansion of the Bayesian predictive density is given by

$$\tilde{\pi}(x_{n+1}|X) = f(x_{n+1};\hat{\theta}) + \frac{1}{2n}\left[c^{st}\left\{c^{jr}a_{jrs} + \frac{2\pi_s(\hat{\theta})}{\pi(\hat{\theta})}\right\}f_t(x_{n+1};\hat{\theta}) + c^{jr}f_{jr}(x_{n+1};\hat{\theta})\right] + o_p(n^{-1}) \tag{2.1}$$

as $n \to \infty$, where $f(x_{n+1};\hat{\theta})$ is the plug-in predictive density. Throughout this paper, unless otherwise specified, we adopt the summation convention over indices ranging from 1 to p. Based on this expansion, Datta et al. (2000) derived the probability matching prior for Bayesian prediction, which ensures that the posterior coverage probability of a Bayesian predictive region asymptotically matches the corresponding frequentist coverage

probability.

We consider the moment matching prior in Bayesian prediction problem. Such priors asymptotically match the mean of the Bayesian predictive density with the mean of the frequentist plug-in predictive density. Let $\mu(\theta) = \mathrm{E}_\theta[X] = \int x f(x;\theta)dx$. Let $\hat{\mu}_{\mathrm{Bayes}} := \hat{\mu}_{\mathrm{Bayes}}(X) = \int x_{n+1}\tilde{\pi}(x_{n+1} \mid X)dx_{n+1}$ and $\hat{\mu}_{\mathrm{plug}} := \mu(\hat{\theta}) = \int_{\mathbb{R}} x_{n+1} f(x_{n+1};\hat{\theta})dx_{n+1}$. Taking the expectation of (2.1), the difference between them is given by

$$n(\hat{\mu}_{\mathrm{Bayes}} - \hat{\mu}_{\mathrm{plug}}) = \frac{1}{2}\left[c^{st}\left\{c^{jr}a_{jrs} + \frac{2\pi_s(\hat{\theta})}{\pi(\hat{\theta})}\right\}\mu_t(\hat{\theta}) + c^{jr}\mu_{jr}(\hat{\theta})\right] + o_p(1),$$

where $\mu_t(\theta) := \int u f_t(u;\theta)du$, $\mu_{jr}(\theta) := \int u f_{jr}(u;\theta)du$. By the continuous mapping theorem and the consistency of the MLE, we have

$$n(\hat{\mu}_{\mathrm{Bayes}} - \hat{\mu}_{\mathrm{plug}}) \xrightarrow{p} \frac{1}{2}\left[I^{st}\left\{I^{jr}A_{jrs} + \frac{2\pi_s(\theta)}{\pi(\theta)}\right\}\mu_t(\theta) + I^{jr}\mu_{jr}(\theta)\right] \tag{2.2}$$

as $n \to \infty$, where I^{jr} $(j, r = 1, \ldots, p)$ is the (j, r)-th element of the matrix I^{-1} with $I = ((I_{jr}))$ and $I_{jr} = \mathrm{E}_\theta[(\partial_j \ell)(\partial_r \ell)]$, and $A_{jrs} = \mathrm{E}_\theta[\partial_j \partial_r \partial_s \ell]$. The right hand side of (2.2) equals 0 if and only if

$$I^{st}\left\{I^{jr}A_{jrs} + \frac{2\pi_s(\theta)}{\pi(\theta)}\right\}\mu_t(\theta) + I^{jr}\mu_{jr}(\theta) = 0. \tag{2.3}$$

A prior distribution $\pi(\theta)$ that satisfies the partial differential equation (2.3) is defined as the predictive moment matching (PMM) prior, denoted by π_{PMM}. This prior ensures that the mean of the Bayesian predictive distribution asymptotically matches the mean of the frequentist plug-in predictive distribution up to the order of $o(n^{-1})$. In the scalar case where $p = 1$, the prior satisfying (2.3) is then given by

$$\pi_{\mathrm{PMM}}(\theta) \propto \exp\left\{-\frac{1}{2}\int^{\theta}\left(\frac{a_3(t)}{I(t)} + \frac{\mu''(t)}{\mu'(t)}\right)dt\right\}, \tag{2.4}$$

where $I(\theta) = \mathrm{E}_\theta[(\ell')^2]$, $a_3(\theta) = \mathrm{E}_\theta[\ell'''']$ with $\ell = \ell(\theta) = \log f(x;\theta)$. Note that the prior (2.4) differs slightly from that of Ghosh and Liu (2011), which was developed in the estimation framework. Ghosh and Liu (2011)'s moment matching priors focus on parameter estimation. Its objective is to asymptotically match the posterior mean of a parameter θ with its maximum likelihood estimator (MLE). In a multi-parameter setting where $\theta = (\theta_1, \ldots, \theta_p)^\top$, the moment matching (MM) prior $\pi_{\mathrm{MM}}(\theta)$ is defined as a solution to the following system of p partial differential equations:

$$I^{jr}A_{jrs} + \frac{2\pi_s(\theta)}{\pi(\theta)} = 0 \quad \text{for } s = 1, \ldots, p. \tag{2.5}$$

There are two fundamental structural differences between the MM prior and the proposed PMM prior. First, the matching equation (2.5) for the MM prior constitutes a simultaneous *system* of p separate differential equations, whereas that for the PMM prior is a *single scalar* equation. Second, the PMM partial differential equation (2.3) uniquely features the

term $I^{jr}(\theta)\mu_{jr}(\theta)$. For a scalar θ, the MM prior is given by

$$\pi_{\mathrm{MM}}(\theta) \propto \exp\left\{-\frac{1}{2}\int^{\theta} \frac{a_3(t)}{I(t)}dt\right\}. \tag{2.6}$$

If $\mu'(\theta) \equiv$ constant, the prior given by (2.4) reduces to the MM prior (2.6) for estimation. The condition that $\mu'(\theta)$ is constant implies that $\mu(\theta)$ is a linear function of θ. In such a case, the identity $\mu(\mathrm{E}[\theta|X]) = \mathrm{E}[\mu(\theta)|X]$ holds by the linearity of the expectation. Consequently, the PMM prior formulated for the predictive problem becomes equivalent to the conventional MM prior established for the estimation problem.

A prior construction rule is said to be invariant under a one-to-one reparameterization $\phi = g(\theta)$ if applying the rule directly in the ϕ-space yields the same prior $\pi^*(\phi)$ as applying the change-of-variables formula to the prior $\pi(\theta)$ derived in the θ-space. An invariant prior is highly desirable because it ensures that posterior inferences and predictive results remain identical regardless of the chosen parameterization. A prominent example of such an invariant prior is the celebrated Jeffreys prior, defined as $\pi(\theta) \propto \sqrt{|I(\theta)|}$, where $I(\theta)$ denotes the Fisher information matrix. Importantly, being a proper density does not guarantee this invariance; for instance, a uniform prior on a bounded support is proper, yet its uniformity is lost under non-linear transformations. Furthermore, many objective priors, such as matching priors, are improper (i.e., $\int \pi(\theta)d\theta = \infty$), which underscores the critical importance of verifying whether the transformed prior remains well-defined and retains its objective properties. It is worth noting that even under an improper prior, formal Bayesian inference remains perfectly valid as long as the resulting posterior distribution is proper. The moment matching priors of Ghosh and Liu (2011) do not possess the invariance property under one-to-one transformations of the parameter θ. Indeed, as shown in Ghosh and Liu (2011), if $\phi = g(\theta)$ is a one-to-one transformation, the transformed prior becomes $\pi^*_{\mathrm{MM}}(\phi) = \pi_{\mathrm{MM}}(\theta)(d\theta/d\phi)^{3/2}$. For non-regular models, the moment matching prior provided by Hashimoto (2019) in the estimation framework also lacks this invariance. Surprisingly, however, it can be demonstrated that the moment matching prior does satisfy the invariance property when formulated within the framework of prediction.

Theorem 2.1. *Suppose that $\phi = g(\theta)$ is a one-to-one transformation of θ. Then the predictive moment matching prior $\pi_{\mathrm{PMM}}(\theta)$ is invariant under this transformation.*

The full details of the proof is given in the Supplementary Material; briefly, the invariance is established because the higher-order terms arising from the chain rule under the parameter transformation perfectly cancel each other out.

The term μ''/μ' can be interpreted as a geometric compensation that precisely cancels the distortion induced by reparameterization. Ghosh and Liu (2011)'s approach generally results in a prior that is coordinate-dependent. In contrast, by shifting the focus to predictive targets, we demonstrate that the inherent non-linearity of the predictive mean $\mathrm{E}_\theta[X]$ provides a natural mechanism for ensuring invariance under reparameterization.

Example 2.1 (One-parameter exponential family)**.** Consider the regular one-parameter exponential family with the density given by $f(x;\theta) = \exp[\theta x - \psi(\theta) + h(x)]$, where $\psi(\theta) = \log\left(\int \exp[\theta x + h(x)]dx\right)$. For the natural parameter θ, we note that $a_3(\theta) = -\psi'''(\theta) = -I'(\theta)$ and $\mathrm{E}_\theta[X] = \mu(\theta) = \psi'(\theta)$. Further, we have $\mu'(\theta) = \psi''(\theta) = I(\theta)$ and $\mu''(\theta) = \psi'''(\theta) = I'(\theta)$. Then the predictive moment matching prior for the natural

parameter θ is given by $\pi_{\mathrm{PMM}}(\theta) \propto 1$, which corresponds to the uniform prior. The result differs from $\pi_{\mathrm{MM}}(\theta) \propto I^{1/2}(\theta)$ derived by Ghosh and Liu (2011). On the other hand, for the expectation parameter $\phi = \psi'(\theta)$, it follows from Theorem 2.1 that $\pi^*_{\mathrm{PMM}}(\phi) = \pi_{\mathrm{PMM}}(\theta)|d\theta/d\phi| = (\psi''(\theta))^{-1} \propto I(\phi)$. Interestingly, this result coincides with the MM prior $\pi^*_{\mathrm{MM}}(\phi) \propto I(\phi)$ derived by Ghosh and Liu (2011).

We present examples where $\mu(\theta)$ is a non-linear function of θ, while ensuring that the posterior and predictive distributions remain analytically tractable.

Example 2.2 (Log-normal distribution)**.** Consider the log-normal distribution $\mathrm{LN}(\theta, \sigma^2)$, where $\theta \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is known. Without loss of generality, let $\sigma^2 = 1$. The density function is given by

$$f(x;\theta) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\log x - \theta)^2}{2}\right), \quad x > 0.$$

Since the log-likelihood function is $\ell(\theta) = \text{const.} + \theta \log x - (\theta^2/2)$, it follows that $\ell' = \log x - \theta$, $\ell'' = -1$, $I(\theta) = -\mathrm{E}_\theta[\ell''] = 1$, $\ell''' = 0$ and $a_3(\theta) = 0$. The mean of the distribution is $\mu(\theta) = \exp(\theta + 1/2)$, which yields $\mu'(\theta) = \mu''(\theta) = \exp(\theta + 1/2)$. Hence, $\pi_{\mathrm{PMM}}(\theta) \propto e^{-\theta/2}$, which is an improper prior. Consider an i.i.d. sample $\{X_1, \ldots, X_n\}$ from $\mathrm{LN}(\theta, 1)$, and let $Y_i = \log X_i$ and $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$. Note that Y_i is the sufficient statistic for θ. Since $Y_i \sim \mathcal{N}(\theta, 1)$, it follows that $\bar{Y} \sim \mathcal{N}(\theta, n^{-1})$. The posterior distribution of θ given $\bar{Y}$ is $\mathcal{N}(\bar{Y} - (1/2n), 1/n)$. Consequently, the posterior distribution is proper, and the posterior mean of θ is $\mathrm{E}[\theta|\bar{Y}] = \bar{Y} - (1/2n)$. Meanwhile the maximum likelihood estimator (MLE) of θ is $\hat{\theta}_{\mathrm{MLE}} = \bar{Y}$. Thus, we have $\mathrm{E}[\theta|\bar{y}] - \hat{\theta}_{\mathrm{MLE}} = -1/(2n)$. Furthermore, the mean of Bayesian predictive density is calculated as

$$\hat{\mu}_{\mathrm{Bayes}} = \int \mathrm{E}_\theta[X_{n+1}]\pi(\theta|\bar{Y})d\theta = \mathrm{E}\left[\exp(\theta + 1/2)|\bar{Y}\right] = \exp(\bar{Y} + 1/2).$$

The mean of the plug-in predictive distribution is $\hat{\mu}_{\mathrm{plug}} = \mu(\hat{\theta}_{\mathrm{MLE}}) = \exp(\bar{Y} + 1/2)$. In this example, exact matching is achieved such that $\hat{\mu}_{\mathrm{plug}} = \hat{\mu}_{\mathrm{Bayes}}$.

Example 2.3 (Pareto distribution)**.** Consider the Pareto distribution $\mathrm{Pa}(\theta, \gamma)$, where $\theta > 1$ is the unknown shape parameter and $\gamma > 0$ is a scale parameter. The density function is given by

$$f(x;\theta) = \theta\gamma^\theta x^{-(\theta+1)}, \quad x \geq \gamma.$$

In the prediction problem, the target is the mean $\mu(\theta) = \mathrm{E}_\theta[X] = \theta\gamma/(\theta - 1)$ for $\theta > 1$. Since the log-likelihood is $\ell(\theta) = \log\theta + \theta\log\gamma - (\alpha + 1)\log x$, we have $\ell'(\theta) = \theta^{-1} + \log\gamma - \log x$, $\ell''(\theta) = -\theta^{-2}$, and $\ell'''(\theta) = 2\theta^{-3}$. It follows that $I(\theta) = \theta^{-2}$ and $a_3(\theta) = \mathrm{E}_\theta[\ell''')] = 2\theta^{-3}$. Furthermore, we have $\mu''(\theta)/\mu'(\theta) = -2/(\theta - 1)$. Hence, the predictive moment matching prior is obtained by $\pi_{\mathrm{PMM}}(\theta) \propto (\theta - 1)/\theta$ for $\theta > 1$. Although the prior is improper, the corresponding posterior distribution is proper. For i.i.d. observations $X = (X_1, \ldots, X_n)$, the posterior distribution of θ given sufficient statistic T is expressed as $\pi(\theta|T) = C_\gamma(\theta - 1)\theta^{n-1}e^{-\theta T}$, where the normalizing constant is $C_\gamma = T^{n+1}/\{\Gamma(n + 1, T) - T \cdot \Gamma(n, T)\}$ with the upper incomplete gamma function $\Gamma(a, s) = \int_s^\infty u^a e^{-u} du$. The maximum likelihood estimator of θ is $\hat{\theta}_{\mathrm{MLE}} = n/T$, and the mean of plug-in predictive distribution is $\hat{\eta}_{\mathrm{plug}} = \mu(\hat{\theta}_{\mathrm{MLE}}) = \gamma(n/T)/\{(n/T) - 1\} = \gamma/\{1 - (T/n)\}$. The mean of

Bayesian predictive distribution is calculated as

$$\hat{\mu}_{\text{Bayes}} = \int_1^\infty \mathrm{E}_\theta[X_{n+1}]\pi(\theta|T)d\theta = C_\gamma \gamma \int_1^\infty \theta^n e^{-\theta T} d\theta = \frac{\gamma}{1 - T\frac{\Gamma(n,T)}{\Gamma(n+1,T)}}.$$

Since $\Gamma(n,T)/\Gamma(n+1,T) = n^{-1} + O(n^{-2})$ for large n, it follows that $\hat{\mu}_{\text{Bayes}}$ is asymptotically equivalent to $\hat{\mu}_{\text{plug}}$.

For vector θ, the moment matching prior is the solution of the first-order linear partial differential equation given in (2.3). Similar PDEs appear in the contexts of probability matching priors (see e.g., Datta et al., 2000).

Example 2.4 (Multi-parameter exponential family)**.** Consider the p-dimensional exponential family with the density function given by $f(x;\theta) = \exp\left\{\theta^\top x - \psi(\theta) + h(x)\right\}$. Extending the argument from Example 2.1 to the multivariate case, we find that the PMM prior is $\pi_{\text{PMM}}(\theta) \propto 1$. For the expectation parameter $\phi = \nabla\psi(\theta)$, the prior becomes $\pi^*_{\text{PMM}}(\phi) \propto \det(I(\phi))$, which is identical in form to the result obtained in the one-dimensional case.

Example 2.5 (Log-normal distribution with unknown ξ and σ^2)**.** Consider the log-normal distribution $\mathrm{LN}(\xi, \sigma^2)$. The log-likelihood function is given by $\ell(\xi, \sigma^2) = -\log(2\pi\sigma^2)/2 - (y-\xi)^2/(2\sigma^2)$, where$y = \log x$. The Fisher information matrix $I = (I_{jk})$ and its inverse $I^{-1} = (I^{jk})$ are $I = \mathrm{diag}(1/\sigma^2, 1/(2\sigma^4))$ and $I^{-1} = \mathrm{diag}(\sigma^2, 2\sigma^4)$, respectively. The target parameter is $\mu(\xi, \sigma^2) = \exp(\xi + (\sigma^2/2))$ and it is calculated by $\mu_1 = \mu$, $\mu_2 = \mu/2$, $\mu_{11} = \mu$, $\mu_{22} = \mu/4$, and $\mu_{12} = \mu/2$. Regarding the third order derivatives of the log-likelihood, we have $A_{111} = 0$, $A_{112} = \sigma^{-4}$, $A_{221} = 0$, and $A_{222} = 2\sigma^{-6}$. Using the equation (2.3), we have

$$\frac{\partial}{\partial\xi}\log\pi(\xi,\sigma^2) + \sigma^2\frac{\partial}{\partial\sigma^2}\log\pi(\xi,\sigma^2) = -3 - \frac{1}{4}\sigma^2. \tag{2.7}$$

Since (2.7) is a single first-order quasi-linear partial differential equation with smooth coefficients, a solution locally exists by the method of characteristics. Specifically, for any differentiable function $G(\cdot)$, a solution is given by $\pi_{\text{PMM}}(\xi,\sigma^2) \propto \exp(-3\xi - (\sigma^2/4))G(\sigma^2 e^{-\xi})$. Provided that G is appropriately selected such that the posterior distribution becomes proper, the mean of the predictive distribution can be computed via Markov chain Monte Carlo (MCMC). Further discussion on this point, however, is beyond the scope of this paper.

If the scale parameter γ in Example 2.3 is treated as unknown, the model becomes a two-parameter model. However, since an unknown γ results in a non-regular model to which the theory developed in this paper does not apply, we do not pursue this case here. The extension to non-regular models remains an intriguing subject for future research (see e.g., Hashimoto, 2019, 2021).

While the results in Section 2 are framed under the independent and identically distributed (iid) setting, the proposed PMM prior framework can be naturally extended to non-iid structures, and its invariance property remains fully valid. In particular, by adapting the regression-based asymptotic framework of Datta and Mukerjee (2003), we provide detailed formulations and invariance proofs for regression models in the Supplementary Material.

3 Asymptotically unbiased Bayesian predictive mean

This section investigates the conditions under which a prior leads to an asymptotically unbiased Bayesian predictive means. Asymptotically unbiased priors for estimation problems have been studied in the literature (e.g., Hartigan, 1965; Sakai et al., 2025). Let $C(\theta, \pi)$ denote the right-hand side of (2.2) in Section 2. While convergence in probability describes the behavior of the estimators' realized values, it does not automatically guarantee the convergence of their expectations. To ensure that $\mathrm{E}_\theta[n(\hat{\mu}_{\mathrm{Bayes}} - \hat{\mu}_{\mathrm{plug}})] \to C(\theta, \pi)$ as $n \to \infty$, we require the sequence of random variables $\{n(\hat{\mu}_{\mathrm{Bayes}} - \hat{\mu}_{\mathrm{plug}})\}$ to be uniformly integrable. In practical terms, this condition ensures that the tails of the distribution of $n(\hat{\mu}_{\mathrm{Bayes}} - \hat{\mu}_{\mathrm{plug}})$ do not carry excessive mass as $n \to \infty$, preventing outliers from dominating the expectation. Under the condition, we consider a prior that the Bayesian predictive mean $\hat{\mu}_{\mathrm{Bayes}}$ is asymptotically unbiased. To this end, we rewrite the relation $\mathrm{E}_\theta[n(\hat{\mu}_{\mathrm{Bayes}} - \hat{\mu}_{\mathrm{plug}})] \to C(\theta, \pi)$ as

$$n \cdot \mathrm{Bias}(\hat{\mu}_{\mathrm{Bayes}}) = C(\theta, \pi) + n \cdot \mathrm{Bias}(\hat{\mu}_{\mathrm{plug}}) + o(1), \tag{3.1}$$

where $\mathrm{Bias}(\delta) = \mathrm{E}_\theta[\delta] - \mu(\theta)$ denotes the bias of δ. Since $\hat{\mu}_{\mathrm{plug}} = \mu(\hat{\theta})$, its bias can be expanded as

$$\mathrm{E}_\theta[\mu(\hat{\theta})] - \mu(\theta) = \mu_j \mathrm{E}_\theta(\hat{\theta}^j - \theta) + \frac{1}{2}\mu_{jr}\mathrm{E}_\theta\left[(\hat{\theta}^j - \theta)(\hat{\theta}^r - \theta)\right] + o(n^{-1}).$$

From the asymptotic theory of the maximum likelihood estimator (see e.g., Cox and Snell, 1968) and Taylor's expansion, it holds that

$$\mathrm{E}_\theta\left[(\hat{\theta}^j - \theta)(\hat{\theta}^r - \theta)\right] = \frac{1}{n}I^{jr}(\theta) + o(n^{-1}), \quad \mathrm{E}_\theta(\hat{\theta}^j - \theta) = \frac{1}{n}B^j(\theta) + o(n^{-1}),$$

where $B^j(\theta) = I^{js}I^{rt}(A^{(1)}_{srt} + \frac{1}{2}A_{srt})$, $A_{srt} = \mathrm{E}_\theta[\partial_s\partial_r\partial_t\ell]$, and $A^{(1)}_{srt} = \mathrm{E}_\theta[(\partial_s\ell)(\partial_r\partial_t\ell)]$. From (3.1), when $C(\theta, \pi) + n \cdot \mathrm{Bias}(\hat{\mu}_{\mathrm{plug}}) = 0$, the Bayesian predictive mean $\hat{\mu}_{\mathrm{Bayes}}$ is asymptotically unbiased. The condition is equivalent to

$$I^{st}\mu_t\left\{\frac{\pi_s}{\pi} + I^{jr}(A_{jrs} + A^{(1)}_{jrs})\right\} + I^{jr}\mu_{jr} = 0.$$

Remark 3.1. The prior satisfying (3.1) can also be interpreted as a predictive moment matching prior such that the Bayesian predictive mean matches the bias-corrected plug-in predictive mean. The bias-corrected (BC) plug-in estimator is defined as $\hat{\mu}_{\text{BC-plug}} = \hat{\mu}_{\mathrm{plug}} - n^{-1}b(\hat{\theta})$, where $b(\theta) = I^{ts}I^{jr}(A^{(1)}_{jrs} + \frac{1}{2}A_{jrs})\mu_t + \frac{1}{2}I^{jr}\mu_{jr}$, and $b(\hat{\theta})$ is the $O(n^{-1})$ bias of the standard plug-in estimator evaluated at the MLE.

For a scalar θ, the asymptotically unbiased prior is simply expressed as

$$\pi_{\mathrm{AUP}}(\theta) \propto \exp\left\{-\int^\theta \left(\frac{a_3(t) + a_{1,2}(t)}{I(t)} + \frac{\mu''(t)}{\mu'(t)}\right) dt\right\}, \tag{3.2}$$

where $a_3(\theta) = \mathrm{E}_\theta[\ell'''']$ and $a_{1,2}(\theta) = \mathrm{E}_\theta[\ell'\ell'']$ with $\ell = \ell(\theta) = \log f(x;\theta)$.

Theorem 3.1. *Suppose that $\phi = g(\theta)$ is a one-to-one transformation of θ. Then the asymptotically unbiased prior for prediction $\pi_{\mathrm{AUP}}(\theta)$ given by* (3.2) *is invariant under this transformation.*

The proof of Theorem 3.1 is provided in the Supplementary Material.

Example 3.1 (One-parameter exponential family)**.** Since $a_{1,2}(\theta) = 0$ in the natural parameterization, we have $\pi_{\mathrm{AUP}}(\theta) \propto 1$ which coincides with $\pi_{\mathrm{PMM}}(\theta) \propto 1$ derived in Example 2.1. According to Theorem 3.1, the prior for the expectation parameter $\phi = \psi'(\theta)$ is also $\pi^*_{\mathrm{AUP}}(\phi) \propto I(\phi)$.

Example 3.2 (Log-normal distribution)**.** Consider the same model as in Example 2.2. Given that $a_{1,2}(\theta) = 0$, the prior in (3.2) is obtained as $\pi_{\mathrm{AUP}}(\theta) \propto e^{-\theta}$. Under this prior, the posterior distribution of θ given $\bar{Y}$ is $\mathcal{N}(\bar{Y} - (1/n), 1/n)$. Hence, we have

$$\hat{\mu}_{\mathrm{Bayes}} = \mathrm{E}\left[\exp(\theta + 1/2)|\bar{Y}\right] = \exp\left(\bar{Y}\right) \exp\left(\frac{1}{2} - \frac{1}{2n}\right).$$

Since $\bar{Y} \sim \mathcal{N}(\mu, 1/n)$, the expectation of $\hat{\mu}_{\mathrm{Bayes}}$ is calculated by $\mathrm{E}_\theta[\hat{\mu}_{\mathrm{Bayes}}] = \exp(\mu + 1/2)$. Hence, $\hat{\mu}_{\mathrm{Bayes}}$ is an exactly unbiased point predictor of X_{n+1}.

Since solving PDEs is generally difficult in multidimensional cases, an alternative approach is to solve the differential equations numerically within an MCMC algorithm (e.g., Sweeting, 2005). Such approaches are likely to be useful for complex statistical models where analytical solutions are often intractable.

Acknowledgements

The author would like to thank two anonymous reviewers for their helpful comments and suggestions to improve the quality of this article. This work was supported by Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science (JSPS), under Contract Number 25K07131.

References

Chang, I. H. and R. Mukerjee (2004). Asymptotic results on the frequentist mean squared error of generalized bayes point predictors. *Statistics & probability letters 67*(1), 65–71.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological) 30*(2), 248–265.

Datta, G. S., M. Ghosh, R. Mukerjee, and T. J. Sweeting (2000). Bayesian prediction with approximate frequentist validity. *The Annals of Statistics 28*(5), 1414–1426.

Datta, G. S. and R. Mukerjee (2003). Probability matching priors for predicting a dependent variable with application to regression models. *Annals of the Institute of Statistical Mathematics 55*(1), 1–6.

Datta, G. S. and R. Mukerjee (2004). *Probability Matching Priors: Higher Order Asymptotics: Higher Order Asymptotics*, Volume 178. Springer Science & Business Media.

Ghosh, M. and R. Liu (2011). Moment matching priors. *Sankhya A 73*(2), 185–201.

Hartigan, J. (1965). The asymptotically unbiased prior distribution. *The Annals of Mathematical Statistics*, 1137–1152.

Hashimoto, S. (2019). Moment matching priors for non-regular models. *Journal of Statistical Planning and Inference 203*, 169–177.

Hashimoto, S. (2021). Predictive probability matching priors for a certain non-regular model. *Statistics & Probability Letters 174*, 109096.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika 83*(2), 299–313.

Sakai, M., T. Matsuda, and T. Kubokawa (2025). Priors for second-order unbiased bayes estimators. *Biometrika 112*(4), asaf068.

Sweeting, T. J. (2005). On the implementation of local probability matching priors for interest parameters. *Biometrika 92*(1), 47–57.

Supplementary material for "On invariant moment matching priors for Bayesian point prediction"

Shintaro Hashimoto

Department of Mathematics, Hiroshima University, Japan

In this supplementary material, we provide proofs of Theorem 2.1 and 3.1 in the main text. We also discuss an extension of the framework presented in the main text to a non-i.i.d. setting, where the future covariate X_{n+1} shares common parameters with the observed data $X_1, \ldots, X_n$ but follows a different distribution. Since the asymptotic expansion of conditional predictive distributions has been investigated in the pioneering work of Datta and Mukerjee (2003), we adopt and operate within their established framework.

S1 Proof of Theorem 2.1 (One-Dimensional Case)

Proof. Let $J = d\theta/d\phi$ be the Jacobian of the one-to-one transformation $\phi = g(\theta)$, and denote its second derivative with respect to ϕ by $\theta_{\phi\phi} = d^2\theta/d\phi^2$.

First, using the standard transformation rules and the chain rule for the derivatives of the log-likelihood function and the predictive mean μ, the components in the ϕ-space are related to those in the θ-space as follows:

$$\begin{aligned} I(\phi) &= I(\theta)J^2, \\ a_3(\phi) &= a_3(\theta)J^3 - 3I(\theta)J\theta_{\phi\phi}, \\ \mu'(\phi) &= \mu'(\theta)J, \\ \mu''(\phi) &= \mu''(\theta)J^2 + \mu'(\theta)\theta_{\phi\phi}. \end{aligned} \tag{S1}$$

By substituting these transformation rules into the integrand of the ϕ-space, we obtain

$$\begin{aligned} \frac{a_3(\phi)}{I(\phi)} &= \frac{a_3(\theta)J^3 - 3I(\theta)J\theta_{\phi\phi}}{I(\theta)J^2} = \frac{a_3(\theta)}{I(\theta)}J - 3\frac{\theta_{\phi\phi}}{J}, \\ \frac{\mu''(\phi)}{\mu'(\phi)} &= \frac{\mu''(\theta)J^2 + \mu'(\theta)\theta_{\phi\phi}}{\mu'(\theta)J} = \frac{\mu''(\theta)}{\mu'(\theta)}J + \frac{\theta_{\phi\phi}}{J}. \end{aligned}$$

Summing these two terms yields the following simplified expression for the integrand in the ϕ-space:

$$\frac{a_3(\phi)}{I(\phi)} + \frac{\mu''(\phi)}{\mu'(\phi)} = \left(\frac{a_3(\theta)}{I(\theta)} + \frac{\mu''(\theta)}{\mu'(\theta)}\right)J - 2\frac{\theta_{\phi\phi}}{J}. \tag{S2}$$

Next, we evaluate the predictive moment matching prior directly constructed in the ϕ-space, denoted by $\pi^*_{\mathrm{PMM}}(\phi)$. By applying the change-of-variable technique with $d\phi = J^{-1}d\theta$ and substituting (S2), the integral can be rewritten as

$$\begin{aligned} \pi^*_{\mathrm{PMM}}(\phi) &\propto \exp\left\{-\frac{1}{2}\int\left(\frac{a_3(\phi)}{I(\phi)} + \frac{\mu''(\phi)}{\mu'(\phi)}\right)d\phi\right\} \\ &= \exp\left[-\frac{1}{2}\int\left\{\left(\frac{a_3(\theta)}{I(\theta)} + \frac{\mu''(\theta)}{\mu'(\theta)}\right)J - 2\frac{\theta_{\phi\phi}}{J}\right\}J^{-1}d\theta\right] \end{aligned}$$

$$= \exp\left[-\frac{1}{2}\int\left(\frac{a_3(\theta)}{I(\theta)} + \frac{\mu''(\theta)}{\mu'(\theta)}\right)d\theta + \int\frac{\theta_{\phi\phi}}{J^2}d\theta\right].$$

Using the linearity of the integral, we can separate the exponent into the product of two exponential functions:

$$\begin{aligned}\pi^*_{\mathrm{PMM}}(\phi) &\propto \exp\left\{-\frac{1}{2}\int\left(\frac{a_3(\theta)}{I(\theta)} + \frac{\mu''(\theta)}{\mu'(\theta)}\right)d\theta\right\}\exp\left(\int\frac{\theta_{\phi\phi}}{J^2}d\theta\right)\\ &= \pi_{\mathrm{PMM}}(\theta)\exp\left(\int\frac{\theta_{\phi\phi}}{J^2}d\theta\right).\end{aligned}$$

To evaluate the remaining integral in the second exponent, we apply the chain rule to differentiate the log-absolute Jacobian $\log|J|$ with respect to θ:

$$\frac{d}{d\theta}\log|J| = \frac{d}{d\phi}\log|J|\cdot\frac{d\phi}{d\theta} = \left(\frac{1}{J}\frac{dJ}{d\phi}\right)\cdot J^{-1} = \frac{\theta_{\phi\phi}}{J}\cdot J^{-1} = \frac{\theta_{\phi\phi}}{J^2}.$$

Therefore, by the fundamental theorem of calculus, the integral simplifies directly to $\log|J|$:

$$\int\frac{\theta_{\phi\phi}}{J^2}d\theta = \int\left(\frac{d}{d\theta}\log|J|\right)d\theta = \log|J|.$$

Consequently, we arrive at

$$\pi^*_{\mathrm{PMM}}(\phi) \propto \pi_{\mathrm{PMM}}(\theta)\exp(\log|J|) = \pi_{\mathrm{PMM}}(\theta)|J| = \pi_{\mathrm{PMM}}(\theta)\left|\frac{d\theta}{d\phi}\right|.$$

This continuous density transformation strictly satisfies the geometric invariance property of probability density functions under one-to-one transformations. This completes the proof. □

S2 Proof of Theorem 2.1 (General Proof)

We provide a general proof of Theorem 2.1 in multidimensional case.

S2.1 Preliminaries on Differential Geometry and Tensor Analysis

Throughout this section, unless stated otherwise, any index appearing twice in a single term—once as a superscript (contravariant component) and once as a subscript (covariant component)—implies summation over the parameter dimensions from 1 to p (Einstein summation convention). For example, if a common index a appears as both a superscript and a subscript, X^aY_a denotes $\sum_{a=1}^p X^aY_a$.

Let the original parameter space be denoted by $\theta = (\theta^1, \ldots, \theta^p)^\top$, which is mapped to a new parameter space $\phi = (\phi^1, \ldots, \phi^p)^\top$ via a smooth, one-to-one coordinate transformation $\phi = g(\theta)$. We define the components of the Jacobian matrix as $J^i_a = \frac{\partial\theta^i}{\partial\phi^a}$ and the components of its inverse as $(J^{-1})^a_i = \frac{\partial\phi^a}{\partial\theta^i}$. By definition, these components satisfy the following geometric relationship:

$$J^i_a(J^{-1})^a_k = \frac{\partial\theta^i}{\partial\phi^a}\frac{\partial\phi^a}{\partial\theta^k} = \frac{\partial\theta^i}{\partial\theta^k} = \delta^i_k, \tag{S1}$$

where δ_k^i is the Kronecker delta. Furthermore, we denote the second-order nonlinear components associated with the coordinate transformation by $\theta_{ab}^j = \frac{\partial^2 \theta^j}{\partial \phi^a \partial \phi^b}$. By applying the product rule for differentiation to the identity $\frac{\partial}{\partial \phi^b}[J_a^i (J^{-1})_i^c] = 0$, we obtain the following crucial tensor identity for the derivative of the inverse Jacobian matrix:

$$\frac{\partial (J^{-1})_i^c}{\partial \phi^b} = -(J^{-1})_j^c (J^{-1})_i^a \theta_{ab}^j. \tag{S2}$$

Here, we introduce Jacobi's formula regarding the derivative of a determinant. The logarithmic partial derivative of the absolute multivariate Jacobian $|J| = \left|\det(J_a^i)\right|$ with respect to ϕ^s always satisfies the following identity, which is derived from equation (S2) and the differential properties of determinants:

$$\frac{\partial \log |J|}{\partial \phi^s} = (J^{-1})_m^j \frac{\partial J_j^m}{\partial \phi^s} = (J^{-1})_m^j \theta_{js}^m. \tag{S3}$$

S2.2 General Proof

Proof. By applying the chain rule and the transformation laws of tensors, the geometric components defined in the θ-space are transformed into the ϕ-space as follows:

$$I^{ab}(\phi) = I^{kl}(\theta)(J^{-1})_k^a (J^{-1})_l^b, \tag{S4}$$

$$A_{abc}(\phi) = A_{jrs}(\theta) J_a^j J_b^r J_c^s - I_{jr}(\theta) \left(\theta_{ab}^j J_c^r + \theta_{bc}^j J_a^r + \theta_{ca}^j J_b^r \right), \tag{S5}$$

$$\mu_a(\phi) = \mu_t(\theta) J_a^t, \tag{S6}$$

$$\mu_{ab}(\phi) = \mu_{jr}(\theta) J_a^j J_b^r + \mu_j(\theta) \theta_{ab}^j. \tag{S7}$$

These identities are multidimensional version of those of (S1) in the proof in one-dimensional case.

Let $\mathcal{E}_\theta(\pi)$ and $\mathcal{E}_\phi(\pi^*)$ denote the differential operators representing the left-hand side of the PMM equations evaluated in the θ-space and ϕ-space, respectively. Specifically, in the ϕ-space, the operator is evaluated for an arbitrary transformed prior $\pi^*(\phi)$ as follows:

$$\mathcal{E}_\phi(\pi^*) = I^{st}(\phi) \left\{ I^{jr}(\phi) A_{jrs}(\phi) + 2 \frac{\partial \log \pi^*(\phi)}{\partial \phi^s} \right\} \mu_t(\phi) + I^{jr}(\phi) \mu_{jr}(\phi). \tag{S8}$$

By definition, the PMM prior $\pi_{\mathrm{PMM}}(\theta)$ in the original space uniquely satisfies $\mathcal{E}_\theta(\pi_{\mathrm{PMM}}) = 0$.

To establish the structural invariance of this framework without any a priori assumptions, we treat $\pi^*(\phi)$ as an initially unconstrained, independent prior density natively defined in the ϕ-space. Under a smooth one-to-one transformation $\phi = g(\theta)$, we expand the geometric components of (S8) via their tensor and non-tensor transformation rules to determine the exact relationship required between $\pi_{\mathrm{PMM}}(\theta)$ and $\pi^*(\phi)$ to maintain the invariance of the PMM system.

Using the inverse metric transformation (S4) and the corrected cumulant transformation (S5), we evaluate the contraction of the third-order cumulant:

$$\begin{aligned} I^{jr}(\phi) A_{jrs}(\phi) &= \left[I^{kl}(\theta)(J^{-1})_k^j (J^{-1})_l^r \right] \left[A_{mnq}(\theta) J_j^m J_r^n J_s^q - I_{mn}(\theta) \left(\theta_{jr}^m J_s^n + \theta_{rs}^m J_j^n + \theta_{sj}^m J_r^n \right) \right] \\ &= I^{kl}(\theta) A_{klq}(\theta) J_s^q - I^{kl}(\theta) I_{mn}(\theta) (J^{-1})_k^j (J^{-1})_l^r \left(\theta_{jr}^m J_s^n + \theta_{rs}^m J_j^n + \theta_{sj}^m J_r^n \right). \end{aligned}$$

Applying the orthogonality relation of the Fisher information matrix $I^{kl}(\theta)I_{mn}(\theta) = \delta^k_m \delta^l_n$ to eliminate the metric components, the indices contract cleanly. This yields the exact expression for the contracted cumulant:

$$I^{jr}(\phi)A_{jrs}(\phi) = I^{kl}(\theta)A_{klm}(\theta)J^m_s - (J^{-1})^j_k (J^{-1})^r_l \theta^k_{jr} J^l_s - 2(J^{-1})^j_m \theta^m_{js}. \tag{S9}$$

Next, we multiply equation (S9) by $I^{st}(\phi)\mu_t(\phi) = I^{uw}(\theta)\mu_w(\theta)(J^{-1})^s_u$. Exploiting the orthogonality of the Jacobian $J^m_s (J^{-1})^s_u = \delta^m_u$, the first term of the PMM equation (S8) simplifies to:

$$\begin{aligned} I^{st}(\phi)I^{jr}(\phi)A_{jrs}(\phi)\mu_t(\phi) = I^{uw}(\theta)I^{kl}(\theta)A_{klu}(\theta)\mu_w(\theta) - I^{uw}(\theta)\mu_w(\theta)(J^{-1})^j_k (J^{-1})^r_u \theta^k_{jr} \\ - 2I^{uw}(\theta)\mu_w(\theta)(J^{-1})^s_u (J^{-1})^j_m \theta^m_{js}. \end{aligned} \tag{S10}$$

For the second term in (S8) containing the partial derivative of the unknown density $\pi^*(\phi)$, the chain rule yields:

$$2I^{st}(\phi)\frac{\partial \log \pi^*(\phi)}{\partial \phi^s}\mu_t(\phi) = 2I^{uw}(\theta)(J^{-1})^s_u \frac{\partial \log \pi^*(\phi)}{\partial \phi^s}\mu_w(\theta). \tag{S11}$$

For the third term in (S8) containing the second derivative of the predictive mean, we correctly distribute the inverse metric $I^{jr}(\phi) = I^{uw}(\theta)(J^{-1})^j_u (J^{-1})^r_w$ across both the tensor and non-tensor parts of (S7), which yields:

$$I^{jr}(\phi)\mu_{jr}(\phi) = I^{kl}(\theta)\mu_{kl}(\theta) + I^{uw}(\theta)\mu_k(\theta)(J^{-1})^j_u (J^{-1})^r_w \theta^k_{jr}. \tag{S12}$$

Combining the expanded terms (S10), (S11), and (S12), we reconstruct the full expression for $\mathcal{E}_\phi(\pi^*)$. By swapping the dummy indices of the first non-tensor residual term ($w \to k$ and $k \to w$) and utilizing the symmetry of the metric $I^{uw}(\theta) = I^{wu}(\theta)$, the geometric components involving θ^k_{jr} in the first and third terms cancel each other out directly:

$$-I^{uw}(\theta)\mu_w(\theta)(J^{-1})^j_k (J^{-1})^r_u \theta^k_{jr} + I^{uw}(\theta)\mu_k(\theta)(J^{-1})^j_u (J^{-1})^r_w \theta^k_{jr} = 0.$$

Consequently, the full differential operator in the ϕ-space can be systematically related to the original operator by adding and subtracting the original gradient term $2I^{uw}(\theta)\frac{\partial \log \pi_{\mathrm{PMM}}(\theta)}{\partial \theta^u}\mu_w(\theta)$:

$$\begin{aligned} \mathcal{E}_\phi(\pi^*) = \mathcal{E}_\theta(\pi_{\mathrm{PMM}}) - 2I^{uw}(\theta)\frac{\partial \log \pi_{\mathrm{PMM}}(\theta)}{\partial \theta^u}\mu_w(\theta) \\ - 2I^{uw}(\theta)\mu_w(\theta)(J^{-1})^s_u (J^{-1})^j_m \theta^m_{js} + 2I^{uw}(\theta)(J^{-1})^s_u \frac{\partial \log \pi^*(\phi)}{\partial \phi^s}\mu_w(\theta). \end{aligned} \tag{S13}$$

Substituting Jacobi's formula $(J^{-1})^j_m \theta^m_{js} = \frac{\partial \log |J|}{\partial \phi^s}$ into the second line of (S13) and factoring out the common metric and predictive mean terms yields the following exact structural identity:

$$\mathcal{E}_\phi(\pi^*) = \mathcal{E}_\theta(\pi_{\mathrm{PMM}}) + 2I^{uw}(\theta)\mu_w(\theta)(J^{-1})^s_u \left\{ \frac{\partial \log \pi^*(\phi)}{\partial \phi^s} - \frac{\partial \log \pi_{\mathrm{PMM}}(\theta)}{\partial \theta^u} J^u_s - \frac{\partial \log |J|}{\partial \phi^s} \right\}, \tag{S14}$$

where we have utilized the identity $\frac{\partial \log \pi_{\mathrm{PMM}}(\theta)}{\partial \theta^u} = \frac{\partial \log \pi_{\mathrm{PMM}}(\theta)}{\partial \theta^v}\delta^v_u = \frac{\partial \log \pi_{\mathrm{PMM}}(\theta)}{\partial \theta^v} J^v_s (J^{-1})^s_u$.

Since $\pi_{\mathrm{PMM}}(\theta)$ is the natively derived prior in the θ-space, it satisfies $\mathcal{E}_\theta(\pi_{\mathrm{PMM}}) = 0$. For the predictive moment matching framework to be geometrically invariant under coordinate transformations, the corresponding prior $\pi^*(\phi)$ must simultaneously satisfy $\mathcal{E}_\phi(\pi^*) = 0$ for any arbitrary statistical model. From (S14), this invariance holds if and only if the non-tensor residual term enclosed in the braces vanishes identically:

$$\frac{\partial \log \pi^*(\phi)}{\partial \phi^s} = \frac{\partial \log \pi_{\mathrm{PMM}}(\theta)}{\partial \theta^u} \frac{\partial \theta^u}{\partial \phi^s} + \frac{\partial \log |J|}{\partial \phi^s}.$$

Integrating this differential relation with respect to ϕ^s directly yields the algebraic condition:

$$\log \pi^*(\phi) = \log \pi_{\mathrm{PMM}}(\theta(\phi)) + \log |J| + C,$$

where C is the integration constant. Hence, we obtain

$$\pi^*(\phi) \propto \pi_{\mathrm{PMM}}(\theta(\phi))|J|.$$

This proves that the operational rule required to maintain the system's validity across coordinate changes is precisely the standard probability density transformation law. Thus, the PMM prior is rigorously invariant under smooth one-to-one coordinate transformations. □

S3 Proof of Theorem 3.1

We provide the proof of Theorem 3.1 in the main text.

Proof. Let $J = d\theta/d\phi$ be the Jacobian of the transformation and denote $\theta_{\phi\phi} = d^2\theta/d\phi^2$.

Using the transformation rules for the derivatives of the log-likelihood function, we have the following relation:

$$a_{1,2}(\phi) = a_{1,2}(\theta)J^3 + I(\theta)J\theta_{\phi\phi}.$$

By combining this with the transformation rules for $a_3(\phi)$, $I(\phi)$, and the derivatives of the predictive mean $\mu(\phi)$, and by tracking the cancellation of the non-linear padding terms involving $\theta_{\phi\phi}$, it follows that

$$\frac{a_3(\phi) + a_{1,2}(\phi)}{I(\phi)} + \frac{\mu''(\phi)}{\mu'(\phi)} = \left(\frac{a_3(\theta) + a_{1,2}(\theta)}{I(\theta)} + \frac{\mu''(\theta)}{\mu'(\theta)} \right) J - \frac{\theta_{\phi\phi}}{J}.$$

Now, we evaluate the prior in the ϕ-space. By performing the change-of-variable technique with $d\phi = J^{-1}d\theta$, the integral in the exponent can be rewritten as

$$\begin{aligned}
\pi^*_{\mathrm{AUP}}(\phi) &\propto \exp\left[-\int \left(\frac{a_3(\phi) + a_{1,2}(\phi)}{I(\phi)} + \frac{\mu''(\phi)}{\mu'(\phi)} \right) d\phi \right] \\
&= \exp\left[-\int \left\{ \left(\frac{a_3(\theta) + a_{1,2}(\theta)}{I(\theta)} + \frac{\mu''(\theta)}{\mu'(\theta)} \right) J - \frac{\theta_{\phi\phi}}{J} \right\} J^{-1} d\theta \right] \\
&= \exp\left[-\int \left(\frac{a_3(\theta) + a_{1,2}(\theta)}{I(\theta)} + \frac{\mu''(\theta)}{\mu'(\theta)} \right) d\theta + \int \frac{\theta_{\phi\phi}}{J^2} d\theta \right].
\end{aligned}$$

By the linearity of the integral, this factors into the product of two exponential terms:

$$\begin{aligned}\pi^*_{\mathrm{AUP}}(\phi) &\propto \exp\left[-\int\left(\frac{a_3(\theta)+a_{1,2}(\theta)}{I(\theta)}+\frac{\mu''(\theta)}{\mu'(\theta)}\right)d\theta\right]\exp\left(\int\frac{\theta_{\phi\phi}}{J^2}d\theta\right)\\ &= \pi_{\mathrm{AUP}}(\theta)\exp\left(\int\frac{\theta_{\phi\phi}}{J^2}d\theta\right).\end{aligned}$$

Note that the integral in the second exponent evaluates to $\log|J|$ because, by the chain rule, the derivative of $\log|J|$ with respect to θ is given by

$$\frac{d}{d\theta}\log|J| = \frac{d}{d\phi}\log|J|\cdot\frac{d\phi}{d\theta} = \frac{\theta_{\phi\phi}}{J}\cdot J^{-1} = \frac{\theta_{\phi\phi}}{J^2}.$$

Consequently, by the fundamental theorem of calculus, we obtain

$$\pi^*_{\mathrm{AUP}}(\phi) \propto \pi_{\mathrm{AUP}}(\theta)\exp(\log|J|) = \pi_{\mathrm{AUP}}(\theta)|J| = \pi_{\mathrm{AUP}}(\theta)\left|\frac{d\theta}{d\phi}\right|.$$

This completes the proof. □

We provided the proof for the one-dimensional case; the multi-dimensional case follows analogously by applying the arguments from the general proof of Theorem 2.1.

S4 Predictive Moment Matching Priors in the Regression Framework

S4.1 Derivation of the PMM Prior

Following the asymptotic expansion framework established by Komaki (1996) and Datta et al. (2000), we extend the predictive problem to a regression setting. Let X be an independent variable (covariate) and Y be a dependent variable, having a joint density $f(x,y;\theta) = g(x;\theta)h(y|x;\theta)$, where $\theta = (\theta_1,\ldots,\theta_p)^\top$. Here, $g(x;\theta)$ represents the marginal density of the covariate X, and $h(y \mid x;\theta)$ denotes the conditional density of Y given X.

Following Datta and Mukerjee (2003), based on the conditional predictive density with the higher-order asymptotic expansion, the posterior predictive density of y_{n+1} given the observed data $d = \{(x_i,y_i)\}_{i=1}^n$ and the future covariate x_{n+1} under the prior $\pi(\theta)$ is expressed as

$$\begin{aligned}&\pi^*(y_{n+1} \mid d, x_{n+1})\\ &= \frac{\tilde{\pi}(x_{n+1},y_{n+1} \mid d)}{\int_{-\infty}^{\infty}\tilde{\pi}(x_{n+1},y_{n+1} \mid d)dy_{n+1}}\\ &= h(y_{n+1} \mid x_{n+1};\hat{\theta})\\ &\quad+\frac{1}{2n}\left[c^{st}\left\{c^{jr}a_{jrs}+\frac{2\pi_s(\hat{\theta})}{\pi(\hat{\theta})}\right\}h_t(y_{n+1} \mid x_{n+1};\hat{\theta})+c^{jr}b_{jr}(y_{n+1} \mid x_{n+1};\hat{\theta})\right]+o_p(n^{-1}),\end{aligned} \tag{S1}$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE), c^{jr} denotes the (j,r)-element of the inverse matrix C^{-1}, and a_{jrs} represents the likelihood-based high-order terms evaluated at the MLE. Throughout this framework, we adopt the summation convention over indices

$t, j, r, s \in \{1, \ldots, p\}$. The partial derivatives of the conditional density are denoted by $h_t = \partial_t h$, and the regression-specific adjustment term b_{jr} is defined as

$$b_{jr}(y_{n+1} \mid x_{n+1}; \theta) = \frac{f_{jr}(x_{n+1}, y_{n+1}; \theta) - g_{jr}(x_{n+1}; \theta) h(y_{n+1} \mid x_{n+1}; \theta)}{g(x_{n+1}; \theta)}$$

(see also Datta and Mukerjee, 2003).

To derive the predictive moment matching (PMM) prior, we calculate the conditional Bayesian predictive mean by multiplying (S1) by y_{n+1} and integrating over $(-\infty, \infty)$ with respect to y_{n+1}:

$$\begin{aligned}
&\mathrm{E}_\pi[y_{n+1} \mid d, x_{n+1}] \\
&= \int_{-\infty}^{\infty} y_{n+1} \pi^*(y_{n+1} \mid d, x_{n+1}) dy_{n+1} \\
&= \mu(x_{n+1}; \hat{\theta}) + \frac{1}{2n} \left[c^{st} \left\{ c^{jr} a_{jrs} + \frac{2\pi_s(\hat{\theta})}{\pi(\hat{\theta})} \right\} \mu_t(x_{n+1}; \hat{\theta}) + c^{jr} B_{jr}(x_{n+1}; \hat{\theta}) \right] + o_p(n^{-1}),
\end{aligned}$$

where we define the conditional expectation of y_{n+1}, its derivative, and the integrated adjustment term as follows:

$$\begin{aligned}
\mu(x_{n+1}; \theta) &= \int_{-\infty}^{\infty} y_{n+1} h(y_{n+1} \mid x_{n+1}; \theta) dy_{n+1}, \\
\mu_t(x_{n+1}; \theta) &= \int_{-\infty}^{\infty} y_{n+1} h_t(y_{n+1} \mid x_{n+1}; \theta) dy_{n+1} = \partial_t \mu(x_{n+1}; \theta), \\
B_{jr}(x_{n+1}; \theta) &= \int_{-\infty}^{\infty} y_{n+1} b_{jr}(y_{n+1} \mid x_{n+1}; \theta) dy_{n+1}.
\end{aligned}$$

In the regression context, the bias of the predictive estimator varies depending on the specific realization of the future covariate x_{n+1}. Therefore, the PMM prior is designed to cancel out the $O(n^{-1})$ bias terms on average across the entire covariate space. By taking the expectation with respect to the marginal distribution of X, denoted by $g(x; \theta)$, and applying the continuous mapping theorem alongside the consistency of the MLE, we require the expected asymptotic discrepancy between the Bayesian predictive mean and the plug-in predictive mean to vanish identically. Following the same derivation as in Section 2 of the main text, we have the fundamental partial differential equation (PDE) for the PMM prior $\pi_{\mathrm{PMM}}(\theta)$, which directly mirrors the equation (3) of the main text:

$$I^{st} \left\{ I^{jr} A_{jrs} + \frac{2\pi_s(\theta)}{\pi(\theta)} \right\} \mathrm{E}_X[\mu_t(X; \theta)] + I^{jr} \mathrm{E}_X[B_{jr}(X; \theta)] = 0, \tag{S2}$$

where $\mathrm{E}_X[\cdot]$ is the expectation with respect to the marginal distribution $g(x; \theta)$.

When the parameter is a scalar ($p = 1$), all indices reduce to a single dimension, and the matrices simplify to scalar functions. Following the notation of the main text for the scalar case, the higher-order likelihood term A_{111} is denoted by $a_3(\theta) = \mathrm{E}_\theta[\ell'''']$, the integrated regression adjustment term B_{11} is denoted by $b_2(x; \theta)$, and the inverse Fisher information element I^{11} becomes $I(\theta)^{-1}$. By substituting $p = 1$ into the general matching

PDE (S2), we obtain the following ordinary differential equation (ODE):

$$I(\theta)^{-1}\left\{I(\theta)^{-1}a_3(\theta)+\frac{2\pi'(\theta)}{\pi(\theta)}\right\}\mathrm{E}_X[\mu'(X;\theta)]+I(\theta)^{-1}\mathrm{E}_X[b_2(X;\theta)]=0.$$

Then we have

$$\pi_{\mathrm{PMM}}(\theta)\propto\exp\left\{-\frac{1}{2}\int\left(\frac{a_3(\theta)}{I(\theta)}+\frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]}\right)d\theta\right\}. \tag{S3}$$

S4.2 Invariance

We now formally establish that the predictive moment matching (PMM) prior derived in the regression framework satisfies the invariance property under arbitrary one-to-one transformations.

Theorem S4.1 (Invariance of the Regression-Based PMM Prior)**.** *Let $\phi = g(\theta)$ be a smooth, one-to-one reparameterization, and let $J = d\theta/d\phi$ denote the Jacobian of the transformation. Then, the predictive moment matching prior $\pi_{\mathrm{PMM}}(\theta)$ defined in* (S3) *is invariant under this transformation, meaning that the prior constructed directly in the ϕ-space, denoted by $\pi^*_{\mathrm{PMM}}(\phi)$, satisfies $\pi^*_{\mathrm{PMM}}(\phi) = \pi_{\mathrm{PMM}}(\theta)|d\theta/d\phi|$.*

Proof. By definition, the PMM prior constructed directly under the reparameterized parameter ϕ is given by

$$\pi_{\mathrm{PMM}}(\phi)\propto\exp\left\{-\frac{1}{2}\int\left(\frac{a_3(\phi)}{I(\phi)}+\frac{\mathrm{E}_X[b_2(X;\phi)]}{\mathrm{E}_X[\partial_\phi\mu(X;\phi)]}\right)d\phi\right\}, \tag{S4}$$

where each component is evaluated explicitly within the ϕ-space.

To rigorously evaluate the transformation rule for $b_2(X;\phi)$, we first simplify its definition in the θ-space. By substituting the joint density expansion

$$\partial_\theta^2 f(x,y;\theta)=\{\partial_\theta^2 g(x;\theta)\}h(y\mid x;\theta)+2\{\partial_\theta g(x;\theta)\}\{\partial_\theta h(y\mid x;\theta)\}+g(x;\theta)\{\partial_\theta^2 h(y\mid x;\theta)\}$$

into the definition of $b_{11}(y\mid x;\theta)$, the second-order derivative of the marginal covariate density cancels out inside the numerator:

$$b_{11}(y\mid x;\theta)=2\frac{\partial_\theta g(x;\theta)}{g(x;\theta)}\partial_\theta h(y\mid x;\theta)+\partial_\theta^2 h(y\mid x;\theta). \tag{S5}$$

Multiplying by y and integrating with respect to y, and noting that $\int y\partial_\theta h\,dy=\mu'(x;\theta)$ and $\int y\partial_\theta^2 h\,dy=\mu''(x;\theta)$, we obtain the explicit identity for the integrated adjustment term:

$$b_2(x;\theta)=2\frac{\partial_\theta g(x;\theta)}{g(x;\theta)}\mu'(x;\theta)+\mu''(x;\theta). \tag{S6}$$

We now consider the corresponding term $b_2(x;\phi)$ defined in the ϕ-space. Applying the first- and second-order chain rules of differentiation ($\partial_\phi = J\partial_\theta$ and $\partial_\phi^2 = J^2\partial_\theta^2+\theta_{\phi\phi}\partial_\theta$) to the identity (S6) yields

$$b_2(x;\phi)=2\left(\frac{J\partial_\theta g(x;\theta)}{g(x;\theta)}\right)\big(J\mu'(x;\theta)\big)+\big(J^2\mu''(x;\theta)+\theta_{\phi\phi}\mu'(x;\theta)\big)$$

$$= J^2 \left(2\frac{\partial_\theta g(x;\theta)}{g(x;\theta)}\mu'(x;\theta) + \mu''(x;\theta) \right) + \mu'(x;\theta)\theta_{\phi\phi}$$
$$= b_2(x;\theta)J^2 + \mu'(x;\theta)\theta_{\phi\phi}.$$

Taking the expectation with respect to the marginal covariate distribution $g(x;\theta)$, and noting that the expectation operator $\mathrm{E}_X[\cdot]$ is linear and operates strictly on the covariate X, we immediately establish the required numerator transformation:

$$\mathrm{E}_X[b_2(X;\phi)] = \mathrm{E}_X[b_2(X;\theta)]J^2 + \mathrm{E}_X[\mu'(X;\theta)]\theta_{\phi\phi}. \tag{S7}$$

For the denominator, a direct application of the first-order chain rule leads to

$$\mathrm{E}_X[\partial_\phi \mu(X;\phi)] = \mathrm{E}_X[\mu'(X;\theta)]\frac{d\theta}{d\phi} = \mathrm{E}_X[\mu'(X;\theta)]J. \tag{S8}$$

By substituting the transformation relations (S7) and (S8) into the second component of the integrand, we can simplify the ratio as follows:

$$\begin{aligned}\frac{\mathrm{E}_X[b_2(X;\phi)]}{\mathrm{E}_X[\partial_\phi \mu(X;\phi)]} &= \frac{\mathrm{E}_X[b_2(X;\theta)]J^2 + \mathrm{E}_X[\mu'(X;\theta)]\theta_{\phi\phi}}{\mathrm{E}_X[\mu'(X;\theta)]J} \\ &= \frac{\mathrm{E}_X[b_2(X;\theta)]J^2}{\mathrm{E}_X[\mu'(X;\theta)]J} + \frac{\mathrm{E}_X[\mu'(X;\theta)]\theta_{\phi\phi}}{\mathrm{E}_X[\mu'(X;\theta)]J} \\ &= \frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]}J + \frac{\theta_{\phi\phi}}{J}.\end{aligned}$$

Next, we combine this with the standard transformation rule for the high-order likelihood adjustment term, which is given by

$$\frac{a_3(\phi)}{I(\phi)} = \frac{a_3(\theta)}{I(\theta)}J - 3\frac{\theta_{\phi\phi}}{J}.$$

Summing the two components inside the full integrand of (S4) leads to

$$\begin{aligned}\frac{a_3(\phi)}{I(\phi)} + \frac{\mathrm{E}_X[b_2(X;\phi)]}{\mathrm{E}_X[\partial_\phi \mu(X;\phi)]} &= \left(\frac{a_3(\theta)}{I(\theta)}J - 3\frac{\theta_{\phi\phi}}{J} \right) + \left(\frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]}J + \frac{\theta_{\phi\phi}}{J} \right) \\ &= \left(\frac{a_3(\theta)}{I(\theta)} + \frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]} \right) J - 2\frac{\theta_{\phi\phi}}{J}.\end{aligned}$$

We now perform the integration with respect to ϕ by applying the integration-by-substitution rule, where $d\phi = J^{-1}d\theta$:

$$\begin{aligned}&\int \left(\frac{a_3(\phi)}{I(\phi)} + \frac{\mathrm{E}_X[b_2(X;\phi)]}{\mathrm{E}_X[\partial_\phi \mu(X;\phi)]} \right) d\phi \\ &= \int \left\{ \left(\frac{a_3(\theta)}{I(\theta)} + \frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]} \right) J - 2\frac{\theta_{\phi\phi}}{J} \right\} J^{-1} d\theta \\ &= \int \left(\frac{a_3(\theta)}{I(\theta)} + \frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]} \right) d\theta - 2\int \frac{\theta_{\phi\phi}}{J^2} d\theta.\end{aligned}$$

Since the derivative of $\log|J| = \log|d\theta/d\phi|$ with respect to θ is exactly given by $\partial_\theta \log|J| = \theta_{\phi\phi}/J^2$, the second integral evaluates directly to $2\log|J|$. Substituting this back into the

exponent of (S4) yields

$$\begin{aligned}\pi_{\mathrm{PMM}}(\phi) &\propto \exp\left\{-\frac{1}{2}\left[\int\left(\frac{a_3(\theta)}{I(\theta)}+\frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]}\right)d\theta - 2\log|J|\right]\right\}\\ &= \exp\left\{-\frac{1}{2}\int\left(\frac{a_3(\theta)}{I(\theta)}+\frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]}\right)d\theta\right\}\cdot\exp\left(\log|J|\right)\\ &= \pi_{\mathrm{PMM}}(\theta)|J|\\ &= \pi_{\mathrm{PMM}}(\theta)|d\theta/d\phi|.\end{aligned}$$

This completes the proof. □

S4.3 Examples

Example S4.1 (Simple Linear Regression with Nonlinear Mean Structure)**.** To demonstrate the practical applicability and explicit calculation of the components in (S3), we consider a simple linear regression model where the predictor has a non-linear target structure. Let X be a scalar covariate distributed according to a known marginal density $g(x;\theta) = g(x)$ independent of θ. Given $X = x$, the dependent variable Y follows a normal distribution:

$$Y \mid X = x \sim \mathcal{N}(\mu(x;\theta), \sigma^2),$$

where σ^2 is known, and the conditional expectation $\mu(x;\theta)$ has a non-linear impact with respect to the scalar parameter θ, expressed as $\mu(x;\theta) = x\psi(\theta)$ for some thrice continuously differentiable function $\psi(\theta)$ (e.g., $\psi(\theta) = e^\theta$).

Under this Gaussian specification, the conditional density $h(y \mid x;\theta)$ is given by

$$h(y \mid x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y - x\psi(\theta))^2}{2\sigma^2}\right\}.$$

We now explicitly evaluate the integrated terms required for the PMM prior. First, the first-order derivative of the conditional mean function with respect to θ is

$$\mu'(x;\theta) = x\psi'(\theta) \quad\Longrightarrow\quad \mathrm{E}_X[\mu'(X;\theta)] = \psi'(\theta)\mathrm{E}_X[X].$$

Second, we calculate $b_2(x;\theta)$. By definition, $b_2(x;\theta) = \int_{-\infty}^{\infty} y\, b_{11}(y \mid x;\theta)dy$. Because the covariate distribution $g(x)$ does not depend on θ, the term $b_{11}(y \mid x;\theta)$ simplifies strictly to the normalized second derivative of the conditional density, i.e., $b_{11}(y \mid x;\theta) = \partial_\theta^2 h(y \mid x;\theta)$. Using the identity that the integration of the second derivative of a density weighted by y reduces to the second derivative of its conditional expectation, we have:

$$b_2(x;\theta) = \int_{-\infty}^{\infty} y\,\partial_\theta^2 h(y \mid x;\theta)dy = \partial_\theta^2\left(\int_{-\infty}^{\infty} y\, h(y \mid x;\theta)dy\right) = \mu''(x;\theta).$$

Differentiating $\mu(x;\theta) = x\psi(\theta)$ twice with respect to θ immediately yields:

$$b_2(x;\theta) = x\psi''(\theta).$$

Taking the expectation of $b_2(X;\theta)$ with respect to the marginal distribution of the covari-

ate X, we obtain:

$$\mathrm{E}_X[b_2(X;\theta)] = \psi''(\theta)\mathrm{E}_X[X].$$

Consequently, the ratio of the two expectations simplifies elegantly to a form independent of the covariate's higher-order moments:

$$\frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]} = \frac{\psi''(\theta)\mathrm{E}_X[X]}{\psi'(\theta)\mathrm{E}_X[X]} = \frac{\psi''(\theta)}{\psi'(\theta)}.$$

Finally, we combine the likelihood and predictive ratios to construct the PMM prior defined in (S3). The integrand inside the exponent is given by

$$\frac{a_3(\theta)}{I(\theta)} + \frac{\mathrm{E}_X[b_2(X;\theta)]}{\mathrm{E}_X[\mu'(X;\theta)]} = -3\frac{\psi''(\theta)}{\psi'(\theta)} + \frac{\psi''(\theta)}{\psi'(\theta)} = -2\frac{\psi''(\theta)}{\psi'(\theta)}.$$

Substituting this into the definition yields the explicit algebraic form of the PMM prior:

$$\begin{aligned}
\pi_{\mathrm{PMM}}(\theta) &\propto \exp\left\{-\frac{1}{2}\int\left(-2\frac{\psi''(\theta)}{\psi'(\theta)}\right)d\theta\right\} \\
&= \exp\left\{\int\frac{\psi''(\theta)}{\psi'(\theta)}d\theta\right\} \\
&= \exp\left(\log|\psi'(\theta)|\right) = |\psi'(\theta)|.
\end{aligned}$$

This highlights that under typical regression frameworks, the regression-based PMM prior can be reduced to a surprisingly clean, intrinsic form that is entirely free from both the experimental noise σ^2 and the higher-order moments of the covariate distribution.

Derivation of $I(\theta)$ and $a_3(\theta)$ in the Nonlinear Mean Structure Example

The conditional log-likelihood function for a single observation (x, y) given $X = x$ under the normal regression model $Y \mid X = x \sim \mathcal{N}(x\psi(\theta), \sigma^2)$ is given by

$$\ell(\theta) = -\frac{1}{2\sigma^2}(y - x\psi(\theta))^2.$$

By applying the chain rule, the first three derivatives of $\ell(\theta)$ with respect to θ are sequentially calculated as follows:

$$\begin{aligned}
\ell' &= \frac{\partial\ell(\theta)}{\partial\theta} = \frac{1}{\sigma^2}(y - x\psi(\theta))x\psi'(\theta), \\
\ell'' &= \frac{\partial^2\ell(\theta)}{\partial\theta^2} = \frac{1}{\sigma^2}\left[-x^2(\psi'(\theta))^2 + (y - x\psi(\theta))x\psi''(\theta)\right], \\
\ell''' &= \frac{\partial^3\ell(\theta)}{\partial\theta^3} = \frac{1}{\sigma^2}\left[-2x^2\psi'(\theta)\psi''(\theta) - x^2\psi'(\theta)\psi''(\theta) + (y - x\psi(\theta))x\psi'''(\theta)\right] \\
&= \frac{1}{\sigma^2}\left[-3x^2\psi'(\theta)\psi''(\theta) + (y - x\psi(\theta))x\psi'''(\theta)\right].
\end{aligned}$$

By definition, the Fisher information is the expectation of the squared first derivative of the log-likelihood function with respect to the joint distribution of X and Y. Then we

have

$$I(\theta) = \mathrm{E}_{X,Y}\left[(\ell')^2\right] = \mathrm{E}_X\left[\mathrm{E}_{Y|X}\left[\left(\frac{1}{\sigma^2}(Y - X\psi(\theta))X\psi'(\theta)\right)^2 \middle| X\right]\right]$$
$$= \mathrm{E}_X\left[\frac{1}{\sigma^4}X^2(\psi'(\theta))^2\mathrm{E}_{Y|X}\left[(Y - X\psi(\theta))^2 \middle| X\right]\right].$$

Since the conditional distribution of Y given X is Gaussian with mean $\mu(X;\theta) = X\psi(\theta)$ and variance σ^2, the conditional expectation of the squared residual is exactly the conditional variance:

$$\mathrm{E}_{Y|X}\left[(Y - X\psi(\theta))^2 \middle| X\right] = \mathrm{Var}(Y \mid X) = \sigma^2.$$

Substituting this relation into the expectation yields:

$$I(\theta) = \mathrm{E}_X\left[\frac{1}{\sigma^4}X^2(\psi'(\theta))^2 \cdot \sigma^2\right] = \frac{(\psi'(\theta))^2}{\sigma^2}\mathrm{E}_X[X^2].$$

Next, we evaluate the expectation of the third-order derivative ℓ'''. The joint expectation can be decomposed by condition on X:

$$a_3(\theta) = \mathrm{E}_{X,Y}\left[\ell'''\right] = \mathrm{E}_X\left[\mathrm{E}_{Y|X}\left[\frac{1}{\sigma^2}\left(-3X^2\psi'(\theta)\psi''(\theta) + (Y - X\psi(\theta))X\psi'''(\theta)\right) \middle| X\right]\right]$$
$$= \mathrm{E}_X\left[-\frac{3X^2\psi'(\theta)\psi''(\theta)}{\sigma^2} + \frac{X\psi'''(\theta)}{\sigma^2}\mathrm{E}_{Y|X}\left[Y - X\psi(\theta) \middle| X\right]\right].$$

Under the true model specification, the conditional expectation of the error term (the first-order residual) strictly vanishes:

$$\mathrm{E}_{Y|X}\left[Y - X\psi(\theta) \middle| X\right] = \mathrm{E}_{Y|X}[Y \mid X] - X\psi(\theta) = 0.$$

Consequently, the second term inside the expectation becomes zero, and the expression simplifies exclusively to the expectation over the covariate X:

$$a_3(\theta) = \mathrm{E}_X\left[-\frac{3X^2\psi'(\theta)\psi''(\theta)}{\sigma^2} + 0\right] = -\frac{3\psi'(\theta)\psi''(\theta)}{\sigma^2}\mathrm{E}_X[X^2].$$

This completes the explicit derivation of both baseline likelihood components.

Example S4.2 (Logistic Regression Model)**.** To further illustrate the framework and demonstrate its immediate applicability to generalized linear models (GLMs), we consider a simple logistic regression setup. Let $X \in \mathbb{R}$ be a scalar covariate distributed according to a known marginal density $g(x;\theta) = g(x)$ independent of θ. Given $X = x$, the binary response variable $Y \in \{0, 1\}$ follows a Bernoulli distribution:

$$Y \mid X = x \sim \mathrm{Bernoulli}(p(x;\theta)),$$

where the success probability $p(x;\theta)$ is modeled via the standard logistic link function with a scalar parameter θ:

$$p(x;\theta) = \mu(x;\theta) = \mathrm{E}[Y \mid X = x] = \frac{1}{1 + e^{-\theta x}}.$$

Under this specification, the conditional PMF $h(y \mid x; \theta)$ for $y \in \{0, 1\}$ is written as

$$h(y \mid x; \theta) = \{p(x; \theta)\}^y \{1 - p(x; \theta)\}^{1-y}.$$

The first- and second-order derivatives of the conditional expectation $\mu(x; \theta) = p(x; \theta)$ with respect to θ are readily obtained by the properties of the logistic function:

$$\begin{aligned}
\mu'(x; \theta) &= xp(x; \theta)\{1 - p(x; \theta)\}, \\
\mu''(x; \theta) &= x^2 p(x; \theta)\{1 - p(x; \theta)\}\{2p(x; \theta) - 1\}.
\end{aligned}$$

Taking the expectation over the covariate space with respect to $g(x)$, the denominator component for the PMM prior is given by

$$\mathrm{E}_X[\mu'(X; \theta)] = \int_{-\infty}^{\infty} xp(x; \theta)\{1 - p(x; \theta)\}g(x)dx.$$

Next, we evaluate the term $b_2(x; \theta)$. Because the covariate distribution $g(x)$ does not depend on θ, the general reduction identity holds universally regardless of whether the response is continuous or discrete, yielding $b_2(x; \theta) = \mu''(x; \theta)$. For clarity in the discrete Bernoulli case, this can be cross-verified via direct summation over the support $y \in \{0, 1\}$:

$$b_2(x; \theta) = \sum_{y=0}^{1} y\, h_{11}(y \mid x; \theta) = 1 \cdot h_{11}(1 \mid x; \theta) + 0 \cdot h_{11}(0 \mid x; \theta) = \partial_\theta^2 p(x; \theta) = \mu''(x; \theta).$$

Substituting the explicit second derivative of the logistic curve, we obtain the exact relation for $b_2(x; \theta)$:

$$b_2(x; \theta) = x^2 p(x; \theta)\{1 - p(x; \theta)\}\{2p(x; \theta) - 1\}.$$

We now take the expectation of $b_2(X; \theta)$ with respect to the marginal distribution of the covariate X:

$$\mathrm{E}_X[b_2(X; \theta)] = \int_{-\infty}^{\infty} x^2 p(x; \theta)\{1 - p(x; \theta)\}\{2p(x; \theta) - 1\}g(x)dx.$$

Consequently, the core predictive component ratio inside the PMM prior integral is explicitly expressed as the ratio of two heavily weighted one-dimensional integrals over the covariate space:

$$\frac{\mathrm{E}_X[b_2(X; \theta)]}{\mathrm{E}_X[\mu'(X; \theta)]} = \frac{\int_{-\infty}^{\infty} x^2 p(x; \theta)\{1 - p(x; \theta)\}\{2p(x; \theta) - 1\}g(x)dx}{\int_{-\infty}^{\infty} xp(x; \theta)\{1 - p(x; \theta)\}g(x)dx}.$$

We now introduce the standard likelihood-based components to complete the specification of the PMM prior. Let $\ell(\theta) = \log h(y \mid x; \theta)$ be the conditional log-likelihood for a single observation. Using the fundamental derivative property of the logistic curve, the first three successive derivatives with respect to θ are verified as:

$$\begin{aligned}
\ell' &= x(y - p(x; \theta)), \\
\ell'' &= -x^2 p(x; \theta)(1 - p(x; \theta)), \\
\ell''' &= x^3 p(x; \theta)(1 - p(x; \theta))(2p(x; \theta) - 1).
\end{aligned}$$

Under the true model specification, the conditional variance of the Bernoulli response is $\mathrm{E}_{Y|X}[(Y - p(X;\theta))^2] = p(X;\theta)(1 - p(X;\theta))$. By the law of total expectation, the Fisher information $I(\theta) = \mathrm{E}_{X,Y}[(\ell')^2]$ is explicitly evaluated as the following weighted integral over the covariate space:

$$I(\theta) = \int_{-\infty}^{\infty} x^2 p(x;\theta)\{1 - p(x;\theta)\}g(x)dx.$$

Furthermore, because the third-order derivative ℓ''' does not depend on the response variable Y, its joint expectation $a_3(\theta) = \mathrm{E}_{X,Y}[\ell'''] $ simplifies directly to the marginal expectation over X:

$$a_3(\theta) = \int_{-\infty}^{\infty} x^3 p(x;\theta)\{1 - p(x;\theta)\}(2p(x;\theta) - 1)g(x)dx.$$

Taking the ratio of these two likelihood-based components yields the first half of the PMM integrand:

$$\frac{a_3(\theta)}{I(\theta)} = \frac{\int_{-\infty}^{\infty} x^3 p(x;\theta)\{1 - p(x;\theta)\}(2p(x;\theta) - 1)g(x)dx}{\int_{-\infty}^{\infty} x^2 p(x;\theta)\{1 - p(x;\theta)\}g(x)dx}.$$

Remarkably, both the likelihood ratio $a_3(\theta)/I(\theta)$ and the predictive ratio $\mathrm{E}_X[b_2]/\mathrm{E}_X[\mu']$ share an identical kernel structure based on the logistic weights $x^k p(x;\theta)\{1 - p(x;\theta)\}$, differing only by a single order of degree shift in the polynomial terms of x. The PMM prior for the logistic regression model is completely and explicitly determined as:

$$\pi_{\mathrm{PMM}}(\theta) \propto \exp\left\{ -\frac{1}{2} \int \left(\frac{\int_{-\infty}^{\infty} x^3 p(x;\theta)\{1 - p(x;\theta)\}(2p(x;\theta) - 1)g(x)dx}{\int_{-\infty}^{\infty} x^2 p(x;\theta)\{1 - p(x;\theta)\}g(x)dx} + \frac{\int_{-\infty}^{\infty} x^2 p(x;\theta)\{1 - p(x;\theta)\}\{2p(x;\theta) - 1\}g(x)dx}{\int_{-\infty}^{\infty} x p(x;\theta)\{1 - p(x;\theta)\}g(x)dx} \right) d\theta \right\}.$$

This formulation highlights a key practical advantage of the regression-based PMM prior regarding its implementation in Markov chain Monte Carlo (MCMC) algorithms. Although the outer integral over θ cannot be evaluated in closed form, the prior density (or its logarithm) at any specific parameter value can be computed efficiently during simulation. Specifically, to evaluate the prior within an MCMC sampler, one only needs to compute simple, one-dimensional numerical integrals over the covariate space at each iteration. This task can be performed instantaneously and with high precision using standard numerical quadrature (e.g., the `integrate` function in R or custom quadrature nodes in Stan) for any given marginal covariate distribution $g(x)$.