POSTERIOR COVARIANCE INFORMATION CRITERION FOR ARBITRARY LOSS FUNCTIONS

YUKITO IBA AND KEISUKE YANO

ABSTRACT. We propose a novel computationally low-cost method for estimating the predictive risks of Bayesian methods for arbitrary loss functions. The proposed method utilises posterior covariance and provides estimators of the Gibbs and the plugin generalization errors. We present theoretical guarantees of the proposed method, clarifying the connection between the widely applicable information criterion, the Bayesian sensitivity analysis, and the infinitesimal jackknife approximation of Bayesian leave-one-out cross validation. An application to differentially-private learning is also discussed.

Keywords: Bayesian statistics; predictive risk; Markov chain Monte Carlo; infinitesimal jackknife approximation; sensitivity; widely-applicable information criterion.

1. INTRODUCTION

Bayesian statistics has achieved great successes in many applied fields because it can represent complex shapes of distributions, naturally quantify uncertainties, and accommodate the prior information commonly accepted in applied fields. The development of the Markov chain Monte Carlo (MCMC) methods has improved the utilization of Bayesian statistics. Nowadays, advanced MCMC algorithms are available and utilised in applied fields. Several softwares implement MCMC and enhance the accessibility of Bayesian methods.

Once Bayesian models are fitted to the data, the goodness of fit is evaluated. Predictive performance is a method used to conduct this evaluation (e.g., Vehtari and Ojanen, 2012). It is also known as generalization ability, and it is an important concept in the machine learning literature. Many studies have estimated the predictive risks of Bayesian models using specific loss functions. To construct an asymptotically unbiased estimator of the posterior mean of an expected log-likelihood, Spiegelhalter et al. (2002) employed the difference between the posterior mean of the log-likelihood and the log-likelihood evaluated at the posterior mean, and proposed the deviance information criterion (DIC). Ando (2007) modified DIC to accommodate the model misspecification. As an alternative approach to constructing an asymptotically unbiased estimator, Watanabe (2010) employed the sample mean of the posterior variances of sample-wise log-likelihoods, and proposed the widely applicable information criterion (WAIC). These criteria successfully evaluate the posterior mean of the log-likelihood using samples from a single run of posterior simulation with primary data; additional simulations with leave-one-out data are unnecessary.

However, such methods that accommodate arbitrary loss functions are relatively scarce. Because the choice of a loss function is application-specific and intended to gain benefits from predicting the model, handling an arbitrary loss function is desirable. Cross-validation (Stone, 1974; Geisser, 1975) is a ubiquitous tool used for this purpose. Although the leave-one-out cross validation (LOOCV) is intuitive and accurate, the brute force implementation requires recomputing the posterior distribution repeatedly, and is almost prohibited. The importance sampling cross validation (IS-CV; Gelfand et al., 1992; Vehtari and Lampinen, 2003) estimates LOOCV without using leave-one-out posterior distributions. Vehtari (2002) proposed a generalization of DIC to an arbitrary loss function. Underhill and Smith (2016) proposed the Bayesian predictive score information criterion (BPSIC) using information matrices, and showed that BPSIC is asymptotically unbiased to generalization error.

In this study, we develop a novel generalization error estimate that accommodates arbitrary loss functions. The proposed method employs the posterior covariance to provide bias correction for empirical errors and presents an asymptotically unbiased estimator for generalization errors. It has several advantages over the previous methods. First, it avoids the importance sampling technique that is sensitive to
the presence of influential observations (Peruggia, 1997); therefore, it is numerically stable with respect to such observations. Second, it is theoretically supported by the asymptotic unbiasedness. Third, it avoids the computation of information matrices and their inverse matrices, which is often complicated and computationally unstable. These advantages are illustrated in applications of analytically tractable location-shift models and differential privacy-preserving learning (Dwork, 2006).

Why does the form of posterior covariance appear in the proposal? This question leads us to find an interesting connection between the Bayesian local sensitivity and the infinitesimal jackknife approximation of the LOOCV. The Bayesian sensitivity formula (e.g., Pérez et al., 2006; Millar and Stewart, 2007; Giordano et al., 2013) implies that the posterior covariance appears by the perturbation of the posterior distribution. By combining the Bayesian sensitivity formula and the infinitesimal jackknife approximation of LOOCV (e.g., Beirami et al., 2017; Giordano et al., 2019; Rad and Maleki, 2020), we find that the proposed method corresponds to an infinitesimal jackknife approximation of the Bayesian LOOCV, which is a reason for the appearance of the posterior covariance form. This aspect is reminiscent of the classical result on the asymptotic equivalence between LOOCV and Akaike information criterion (AIC; Akaike, 1973) discovered by Stone (1974). Also, it reduces to the asymptotic equivalence between WAIC and Bayesian LOOCV, given in Section 8 of Watanabe (2018), when the evaluation function is the log-likelihood.

2. PROPOSED METHOD

Suppose that we have independent and identically distributed (i.i.d.) current observations \( X^n = (X_1, \ldots, X_n) \) in a sample space \( \mathcal{X} \) in \( \mathbb{R}^d \) from an unknown sampling distribution \( Q \), and that our prediction target is \( X_{n+1} \) from the same sampling distribution \( Q \) and is independent of \( X^n \). Let \( \Theta \) be the parameter space in \( \mathbb{R}^p \).

2.1. Predictive evaluation. We work with the quasi-Bayesian approach using the quasi-posterior distribution:

\[
\pi(\theta; X^n) = \frac{\exp\{\sum_{i=1}^n s(X_i, \theta)\} \pi(\theta)}{\int \exp\{\sum_{i=1}^n s(X_i, \theta')\} \pi(\theta') d\theta'},
\]

where \( \pi(\theta) \) is a prior density on \( \Theta \) and \( s(x, \theta) \) is a score function that is the minus of a loss function. The score function may be different from the log-likelihood.

For the predictive evaluation of the quasi-Bayesian approach, consider an arbitrary loss function \( \nu(x, \theta) \) for a future observation \( x \in \mathcal{X} \) and a parameter vector \( \theta \in \Theta \). Examples include the mean squared error \( \nu(x, \theta) = \|x - \mathbb{E}[X | \theta]\|^2 \), the \( \ell_1 \) error \( \nu(x, \theta) = \|x - \mathbb{E}[X | \theta]\| \), the log likelihood \( \nu(x, \theta) = \log p(x | \theta) \), the likelihood \( \nu(x, \theta) = p(x | \theta) \), and the p-value \( \nu(x, \theta) = \int_x^\infty p(x' | \theta) dx' \), where \( \{p(x | \theta) : \theta \in \Theta\} \) is a parametric model and \( \mathbb{E}[\cdot | \theta] \) is the expectation with respect to \( p(x | \theta) \). On the basis of \( \nu(x, \theta) \), we consider two types of predictive measures: the Gibbs generalization error

\[
G_{G,n} = \mathbb{E}_{\text{pos}}[\nu(X_{n+1}, \theta)],
\]

and the plugin generalization error

\[
G_{P,n} = \nu(X_{n+1}, \mathbb{E}_{\text{pos}}[\theta]),
\]

where \( \mathbb{E}_{\text{pos}} \) is the quasi-posterior expectation.

To estimate the Gibbs and plugin generalization errors on the basis of current observations, we propose the Gibbs and the plugin posterior covariance information criteria PCIC\(_G\) and PCIC\(_P\):

\[
\text{PCIC}_G = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\text{pos}}[\nu(X_i, \theta)] - \frac{1}{n} \sum_{i=1}^n \text{Cov}_{\text{pos}}[\nu(X_i, \theta), s(X_i, \theta)] \quad \text{and} \quad (2)
\]

\[
\text{PCIC}_P = \frac{1}{n} \sum_{i=1}^n \nu(X_i, \mathbb{E}_{\text{pos}}[\theta]) - \frac{1}{n} \sum_{i=1}^n \text{Cov}_{\text{pos}}[\nu(X_i, \theta), s(X_i, \theta)], \quad (3)
\]
Table 1. Algorithm for the generalization error estimation.

| **Input:** Observations $X_1, \ldots, X_n$ and the number $M$ of posterior samples. |
| **Output:** An estimate of $\mathcal{G}_{G,n}$ and that of $\mathcal{G}_{P,n}$. |

**Step 1:**
Sample $\theta_1, \ldots, \theta_M$ from $\pi(\theta; X)$.  
**Step 2:**
Calculate $T_G = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{M} \sum_{k=1}^{M} \nu(X_i, \theta_k)$ and $T_P = \frac{1}{n} \sum_{i=1}^{n} \nu \left( X_i, \frac{1}{M} \sum_{k=1}^{M} \theta_k \right)$.  
Calculate also $V = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{M} \sum_{k=1}^{M} \{ \nu(X_i, \theta_k) s(X_i, \theta_k) \} - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{M} \sum_{k=1}^{M} \nu(X_i, \theta_k) \frac{1}{M} \sum_{k=1}^{M} s(X_i, \theta_k)$.  
**return** $(T_G - V)$ as an estimate of $\mathcal{G}_{G,n}$ and $(T_P - V)$ as an estimate of $\mathcal{G}_{P,n}$.  

where $\text{Cov}_{\text{pos}}$ is the quasi-posterior covariance. Table 2.1 summarises how we obtain estimates of the Gibbs and the plug-in generalization errors. The first terms correspond to the empirical errors

$$E_{G,n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\text{pos}}[\nu(X_i, \theta)] \quad \text{and} \quad E_{P,n} = \frac{1}{n} \sum_{i=1}^{n} \nu(X_i, \mathbb{E}_{\text{pos}}[\theta]);$$

Our proposed methods employ the posterior covariance as bias correction of empirical errors.

**Remark 2.1** (Computational complexity). An important point of the proposed criteria is that their computational complexity is controlled only by the sampling of posterior distributions. Typical generalization error estimates such as Takeuchi Information Criterion (TIC; Takeuchi, 1976), Regularization Information Criterion (RIC; Shibata, 1989), and Generalised Information Criterion (GIC; Konishi and Kitagawa, 1996) employ information matrices like $\hat{J}_s := (1/n) \sum_{i=1}^{n} \{ - \nabla_{\theta} \nabla_{\theta}^\top s(X_i, \theta_s) \}$ and their inverses, where $\nabla_{\theta}$ is the gradient with respect to $\theta$. However, the computation of $\hat{J}_s^{-1}$ is unstable and demanding; it requires $O(p^3)$ computation. The proposed method utilises posterior covariance to avoid such a demanding computation.

**Remark 2.2** (Connection to WAIC). When working with a parametric model $\{ p(x \mid \theta) : \theta \in \Theta \}$, we set the minus log-likelihood as the loss function, and consider the posterior distribution with learning rate $\beta > 0$. We then have $\nu(X, \theta) = -\log p(X \mid \theta)$ and $s(X, \theta) = \beta \log p(X \mid \theta)$. In this case, $\text{PCIC}_G$ is reduced to WAIC$_2$ given in Section 8.3 of Watanabe (2018):

$$\text{WAIC}_2 = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\text{pos}}[\log p(X_i \mid \theta)] + \frac{\beta}{n} \sum_{i=1}^{n} \mathbb{V}_{\text{pos}}[\log p(X_i \mid \theta)].$$

**Remark 2.3** (The presence of observation weights). Recent machine learning literature deals with predictive settings with observation weights; for example, transfer learning, including covariate shift adaptation (e.g. Shimodaira, 2000), and counterfactual prediction (e.g. Saito and Yasui, 2020). Iba and Yano (2020) proposed an extension of WAIC to accommodate such predictive settings. We can employ this extension for the Gibbs and plug-in generalization error estimation; for example, we obtain the following Gibbs generalization error estimate: Let $\{ w_i > 0 : i = 1, \ldots, n \}$ be the set of observation weights and let $\{ s_i(\cdot, \cdot) : i = 1, \ldots, n \}$ be the set of observation-wise score functions. Then, the weighted version of $\text{PCIC}_G$ is given by

$$\text{PCIC}_G = \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{E}_{\text{pos}}[\nu(X_i, \theta)] - \frac{1}{n} \sum_{i=1}^{n} w_i \text{Cov}_{\text{pos}}[\nu(X_i, \theta), s_i(X_i, \theta)].$$
2.2. Theoretical results. This subsection presents the theoretical support for the use of PCIC\(_G\). The same result holds for PCIC\(_P\), and it is omitted. To provide mathematically unblemished results, we set several conditions. In the conditions, we use the following additional notations: For \( w = (w_1, \ldots, w_n) \) with \( w_i \in [0, 1] \), we define the weighted quasi-posterior distribution

\[
\pi_w(\theta ; X^n) := \frac{\exp\{\sum_{i=1}^{n} w_i s(X_i, \theta)\} \pi(\theta)}{\int \exp\{\sum_{i=1}^{n} w_i s(X_i, \theta')\} \pi(\theta')} \tag{2.2}
\]

and denote by \( \mathbb{E}_{\text{pos}}^w \) the expectation with respect to \( \pi_w(\theta ; X^n) \). We denote using \( \mathbb{E}[\cdot] \) the expectation with respect to \( X^n \) from \( Q^{\circ n} \). Let \( W^{(-1)} := \{ w : w_1 \in [0, 1] \text{ and } w_j = 1, j \neq 1 \} \). The conditions in this paper are as follows:

(C1) The difference \( \mathbb{E}[G_{G,n} - G_{G,n}] \) is of the order \( \mathcal{O}(\mathbb{E}[G_{G,n}]) \);

(C2) The following relation holds:

\[
\mathbb{E} \left[ \sup_{w \in W^{(-1)}} \left| \mathbb{E}_{\text{pos}}^w \left[ \left\{ \nu(X_1, \theta) - \mathbb{E}_{\text{pos}}^w[\nu(X_1, \theta)] \right\} \{ s(X_1, \theta) - \mathbb{E}_{\text{pos}}^w[s(X_1, \theta)] \}^2 \right] \right| \right] = \mathcal{O}(\mathbb{E}[G_{G,n}]);
\]

(C3) There exists an integrable function \( M(\theta) \) such that for all \( w \in W^{(-1)} \),

\[
\left| \pi(\theta) \exp \left\{ w_1 s(X_1, \theta) + \sum_{k \neq 1} s(X_k, \theta) \right\} \nu^l(X_1, \theta) s^l(X_1, \theta) \right| \leq M(\theta), l = 0, 1, j = 0, 1.
\]

Remark 2.4 (Discussion on the conditions). First, consider Condition C1. Usually, the order of \( \mathbb{E}[G_{G,n}] \) is \( 1/n \), and this condition is satisfied. In regular statistical models and smooth loss functions, the result of Underhill and Smith (2016) implies that the order of \( \mathbb{E}[G_{G,n}] \) is \( 1/n \). For singular statistical models and log-likelihoods, the result of Watanabe (2010) implies that the order of \( \mathbb{E}[G_{G,n}] \) is \( 1/n \). Condition C2 is a condition for the residual. Usually, the order of the left hand side is \( n^{-3/2} \) and less than the order of \( \mathbb{E}[G_{G,n}] \). Therefore, this condition is satisfied. Finally, Condition C3 is a mild condition for assuming the existence of expectation.

The following is the main theorem stating the asymptotic unbiasedness of the proposed criteria.

**Theorem 2.5.** Under Conditions C1-C3, the criterion PCIC\(_G\) is asymptotically unbiased to the Gibbs generalization error:

\[
\mathbb{E}[\text{PCIC}_G] = \mathbb{E}[G_{G,n}] + \mathcal{O}(\mathbb{E}[G_{G,n}]).
\]

**Proof of Theorem 2.5** We start with the LOOCV

\[
\text{CV}_G := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\text{pos}}[\nu(X_i, \theta) | X^{-i}],
\]

where \( \mathbb{E}_{\text{pos}}[\cdot | X^{-i}] \) is the expectation with respect to the leave-one-out quasi-posterior distribution:

\[
\pi(\theta ; X^{-i}) = \frac{\exp\{\sum_{j \neq i} s(X_j, \theta)\} \pi(\theta)}{\int \exp\{\sum_{j \neq i} s(X_j, \theta')\} \pi(\theta') d\theta'}.
\]

We further define \( \mathcal{L}(X_i, w) := \mathbb{E}_{\text{pos}}^w[\nu(X_i, \theta)] \). Then we can rewrite \( \text{CV}_G \) as

\[
\text{CV}_G = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(X_i, \mathbb{1}_{-i}),
\]

where the \( j \)-th component of \( \mathbb{1}_{-i} \) is 0 if \( j = i \) and 1 otherwise.

Next, we approximate \( \mathcal{L}(X_i, \mathbb{1}_{-i}) \) by \( \mathcal{L}(X_i, \mathbb{1}) \) where \( \mathbb{1} \) is the all-one vector. We then have

\[
\mathcal{L}(X_i, \mathbb{1}_{-i}) = \mathcal{L}(X_i, \mathbb{1}) + \nabla_{w=1} \mathcal{L}(X_i, w) \delta_i + \frac{1}{2} \delta_i^\top \nabla_{w=1}^2 \mathcal{L}(X_i, w) \delta_i,
\]
where $\nabla_w$ is the gradient with respect to $w$, $\delta_i := 1_{-i} - 1$, and $w^*$ is a point between $1_{-i} \pm 1$. This is reduced to

$$\mathcal{L}(X_i, 1_{-i}) = \mathcal{L}(X_i, 1) + (-1) \frac{\partial}{\partial w_i} \Big|_{w_i=1} \mathcal{L}(X_i, w) + \frac{(-1)^2}{2} \frac{\partial^2}{\partial w_i^2} \Big|_{w_i=w_i^*} \mathcal{L}(X_i, w),$$

where $w_i^*$ is the $i$-th component of $w^*$.

Consider the quasi-posterior distribution

$$\pi(\theta; X^n) \propto \exp \left\{ -\beta \sum_{i=1}^n \| X_i - \theta \|^2 - \frac{\| \theta \|^2}{2\tau} \right\} \quad \text{with} \quad \beta > 0 \quad \text{and} \quad \tau > 0,$$

where $\| \cdot \|$ is the $\ell_2$ norm in $\mathbb{R}^d$; then, the quasi-posterior distribution is $\mathcal{N}(\hat{\theta}, S)$ with

$$\hat{\theta} = \frac{n\beta\tau}{n\beta\tau + 1} \bar{X} \quad \text{and} \quad S = \frac{1}{n\beta + 1/\tau} I_d$$

and the score function in this case is $s(x, \theta) = -(\beta/2)\| x - \theta \|^2$, where let $\bar{X} := \sum_{i=1}^n X_i/n$ and let $I_d$ be the $d \times d$ identity matrix.

Lemma 2.6. Under Condition C3, we have, for $k = 1, 2$,

$$\frac{\partial^k}{\partial w_i^k} \mathbb{E}_{\text{pos}}^{w}[\nu(X_i, \theta)] = \mathbb{E}_{\text{pos}}^{w}[\nu(X_i, \theta)] - \mathbb{E}_{\text{pos}}^{w}[\nu(X_i, \theta)] \left\{ \mathbb{E}_{\text{pos}}^{w}[s(X_i, \theta)] \right\}^k. $$

Using Lemma 2.6 we get

$$\mathcal{L}(X_i, 1_{-i}) = \mathcal{L}(X_i, 1) - \text{Cov}_{\text{pos}}[\nu(X_i, \theta), s(X_i, \theta)] + \frac{1}{2} \kappa^{w^*, i}_{3}[i], \quad \text{(4)}$$

where $w^*, i = (1, \ldots, 1, w_i^*, 1, \ldots, 1)$ and

$$\kappa^{w^*, i}_{3}[i] = \mathbb{E}_{\text{pos}}^{w^*}[\nu(X_i, \theta) - \mathbb{E}_{\text{pos}}^{w^*}[\nu(X_i, \theta)] s(X_i, \theta)] - \mathbb{E}_{\text{pos}}^{w^*, i}[s(X_i, \theta)]^2.$$

Summing up (4) with respect to $i$ yields

$$\text{CV}_G = \text{PCIC}_G + \frac{1}{2n} \sum_{i=1}^n \kappa^{w^*, i}_{3}[i], \quad \text{(5)}$$

and thus we get

$$\mathbb{E}[G_{G,n}] = \mathbb{E}[^{\text{PCIC}_G} + \{G_{G,n} - G_{G,n-1}\}_{=A} + \frac{1}{2n} \sum_{i=1}^n \mathbb{E}[^{\kappa^{w^*, i}_{3}[i]}]_{=B}].$$

Condition C1 makes $A = O(\mathbb{E}[G_{G,n}])$ and Condition C2 makes $B = o(\mathbb{E}[G_{G,n}])$, which completes the proof. □

3. Applications

This section presents applications of the proposed methods. Before studying the simulation, we analyse the behavior of the proposed method in a simple model.

3.1. Case study. Consider a simple location-shift model, where the observations $X^n = (X_1, \ldots, X_n)$ follow

$$X_i = \theta^* + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\theta^*$ is a vector in $\mathbb{R}^d$, and the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. from a possibly non-Gaussian distribution with mean zero and covariance matrix identity matrix. Consider the quasi-posterior distribution given by

$$\pi(\theta; X^n) \propto \exp \left\{ -\beta \sum_{i=1}^n \| X_i - \theta \|^2 - \frac{\| \theta \|^2}{2\tau} \right\} \quad \text{with} \quad \beta > 0 \quad \text{and} \quad \tau > 0,$$

where $\| \cdot \|$ is the $\ell_2$ norm in $\mathbb{R}^d$; then, the quasi-posterior distribution is $\mathcal{N}(\hat{\theta}, S)$ with

$$\hat{\theta} = \frac{n\beta\tau}{n\beta\tau + 1} \bar{X} \quad \text{and} \quad S = \frac{1}{n\beta + 1/\tau} I_d$$

and the score function in this case is $s(x, \theta) = -(\beta/2)\| x - \theta \|^2$, where let $\bar{X} := \sum_{i=1}^n X_i/n$ and let $I_d$ be the $d \times d$ identity matrix.
Consider the loss function \( \nu(x, \theta) = (x - \theta)^\top A (x - \theta) \) with a symmetric positive definite matrix \( A \in \mathbb{R}^{d \times d} \). Thus, the Gibbs generalization gap is given by

\[
\mathbb{E}[G_{G,n}] - \mathbb{E}[\mathcal{E}_{G,n}] = \frac{2}{n} \frac{n \beta \tau}{n \beta \tau + 1} \text{tr}(A),
\]

while the posterior covariance is given by

\[
\frac{1}{n} \sum_{i=1}^{n} \text{Cov}_\mathcal{G}[\nu(X_i, \theta), s(X_i, \theta)] = - \frac{2}{n} \frac{n \beta \tau}{n \beta \tau + 1} \sum_{i=1}^{n} \frac{\bar{X}_i^\top A \bar{X}_i}{n} - \frac{\beta}{(n \beta + 1/\tau)^2} \text{tr}(A),
\]

where \( \bar{X}_i := X_i - \bar{\theta} \) and the detailed calculi are given in Appendix A. Thus, we obtain

\[
\mathbb{E}[G_{G,n}] - \mathbb{E}[\mathcal{E}_{G,n}] = -\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \text{Cov}_\mathcal{G}[\nu(X_i, \theta), s(X_i, \theta)] \right] + \frac{2}{n} \frac{n \beta \tau}{n \beta \tau + 1} \frac{(\theta^*)^\top A \theta^*}{(n \beta \tau + 1)^2} + \text{rem},
\]

where we let

\[
\text{rem} = \frac{\beta \text{tr}(A)}{(n \beta + 1/\tau)^2} + \left\{ \left( \frac{n \beta \tau}{n \beta \tau + 1} \right)^3 - 2 \left( \frac{n \beta \tau}{n \beta \tau + 1} \right)^2 \right\} \frac{2 \text{tr}(A)}{n^2}.
\]

From (8), we conclude that in simple location-shift models, under the assumption

\[
(\theta^*)^\top A \theta^*/(n \beta \tau + 1)^2 = o(1),
\]

PCIC\(_C\) estimates well the Gibbs generalization error regardless of the dimension \( d \) and the distribution of the error term. For non-strong priors (\( \tau = O(1) \)), we can make the assumption in (9). For strong priors (\( 1/\tau \sim n \)), we cannot expect the validity of the assumption, and the bias term

\[
\frac{2}{n} \frac{n \beta \tau}{n \beta \tau + 1} \frac{(\theta^*)^\top A \theta^*}{(n \beta \tau + 1)^2}
\]

remains. Vehtari et al. (2017) pointed out such a bias for WAIC; see also Ninomiya (2021). Our quasi-Bayesian framework provides a simple modification to remove this bias. Add the log-prior density (up to constant) divided by \( n \) to \( s(x, \theta) \):

\[
s'(x, \theta) = s(x, \theta) - (1/n)||\theta||^2/(2\tau).
\]

Then, we have

\[
\mathbb{E}[G_{G,n}] - \mathbb{E}[\mathcal{E}_{G,n}] = -\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \text{Cov}_\mathcal{G}[\nu(X_i, \theta), s'(X_i, \theta)] \right] + \text{rem}_2
\]

with \( \text{rem}_2 \) an \( O(n^{-2}) \)-term that is independent of \( \theta^* \), which implies that the bias term can be removed.

3.2. Evaluation of differential private learning methods. Differential privacy is a notion of privacy that protects individual users’ private information and allows data analyses to be conducted with the desired statistical efficiency (Dwork, 2006). Dwork (2006) defines a notion of a differentially private learner, an \((\epsilon, \delta)\)-differentially private learner: For \( \epsilon, \delta \geq 0 \), an \((\epsilon, \delta)\)-differentially private learner is defined as a randomised estimator \( \hat{\theta} \) satisfying that, for any adjacent datasets \( D, D' \in X^n \), an inequality

\[
P \left( \hat{\theta}(D) \in A \right) \leq \exp(\epsilon) P \left( \hat{\theta}(D') \in A \right) + \delta
\]

holds for every measurable set \( A \). A common approach for attaining \((\epsilon, \delta)\)-differential privacy is a one-posterior-sample (OPS) estimator, that is, one sample drawn from a quasi-posterior distribution

\[
\pi_\beta(\theta; X^n) \propto \exp \left[ -\beta \sum_{i=1}^{n} L(X_i, \theta) \right],
\]

where \( L(\cdot, \cdot) \) is a user-specific training loss function and \( \beta \) is the hyperparameter controlled by the privacy level \((\epsilon, \delta)\), the loss function, and the sample size; see Wang et al. (2015); Minami et al. (2016).

Here, we demonstrate the application of PCIC\(_C\) to the predictive evaluation of OPS estimators. We use two sets of classification data from UCI Machine Learning Repository (Dua and Graff, 2017) namely, The Banknote authentication data set and The Adult data set. The Banknote authentication data set
classifies genuine and forged banknote-like specimens based on four image features. The Adult data set predicts whether income exceeds 50K/yr based on 14 features from census data. We work with the quasi-posterior distribution based on the logistic regression

$$
\pi_\beta(\theta; Y^n, X^n) \propto \exp \left[ \beta \sum_{i=1}^{n} \{ Y_i \log \sigma(X_i \theta) + (1 - Y_i) \log(1 - \sigma(X_i \theta)) \} \right] \exp\{-\theta^\top \theta/2\},
$$

where $\sigma(\cdot)$ is the sigmoid function: $\sigma(x) := 1/(1 + \exp(-x))$. We consider predictive evaluation of OPS estimators using three major classification losses, the Brier loss, the misclassification loss, and the spherical loss:

$$
\nu_{\text{Brier}}(x, p) = (x - p)^2,
$$

$$
\nu_{\text{misclass}}(x, p) = \begin{cases} 
    -1 & \text{if } x = 1 \text{ and } p > 1/2 \text{ or } x = 0 \text{ and } p < 1/2, \\
    0 & \text{if otherwise,}
\end{cases}
$$

$$
\nu_{\text{spherical}}(x, p) = \frac{xp + (1 - x)(1 - p)}{\sqrt{p^2 + (1 - p)^2}},
$$

where for the $i$ observation, $p$ is given by $\sigma(X_i \theta)$ using $\theta$. Note that, the misclassification loss is discontinuous for $p$. First, we randomly split the whole data into various training data sets, with a sample size 50. The test data set has a sample size of 10. We then calculate the empirical errors using the training data set, and the average of generalization errors using the test data set. We used 3980 MCMC samples after thinning out by 5 and a burnin period of length 100. We compare five generalization error
Figures 2 illustrate generalization error estimates in relation to the average generalization errors for the adult data set. The red line identifies the values of PCIC\textsubscript{G}. The blue line marks the values of IS-CV. The purple line corresponds to the DIC adapted for arbitrary loss functions (GDIC). The orange dot denotes the BPSIC value. The green line represents exact LOOCV. The black line indicates empirical errors.

(a) Generalization error estimates for the Brier loss. (b) Those for the misclassification loss. (c) Those for the spherical loss.

3.3. Application to Bayesian regression in the presence of influential observations. The performance of the proposed method and that of IS-CV do not differ from the previous example. However, the presence of influential observations impacts the variability of the case-deletion importance sampling weights, as discussed in Peruggia (1997), resulting in the instability of IS-CV.

Here, we focus on the performance comparison between our method and IS-CV in the presence of influential observations. We use the following simple Bayesian regression model: for a given $R$, $X_1, \ldots, X_n$ are fixed to

$$X_i = \begin{cases} 0.01i & \text{if } i < n, \\ R & \text{if } i = n, \end{cases}$$
FIGURE 3. Comparison of PCICG and IS-CV. We plot the differences relative to the average of the generalization error, which is why the values of the average of generalization errors are set to 0. The black dashed line denotes the values of the averages of generalization errors. The red dots denote the mean values of PCICG with the shaded region denoting the mean ± and the standard deviation. The blue dots denote the mean values of IS-CV, with the shaded region denoting the mean ± and the standard deviation. (a) PCICG and IS-CV for $\ell_2$. (b) Those for scaled $\ell_1$.

and then $Y_1, \ldots, Y_n$ are given as

$$Y_i = \beta_0 + X_i\beta_1 + \sigma\varepsilon_i,$$

where $\varepsilon_i$s are i.i.d. from the standard Gaussian distribution, and $\beta_0, \beta_1,$ and $\sigma$ are unknown parameters. We assign $(0,1,1)$ to the true values of $(\beta_0, \beta_1, \sigma)$. We set 50 to the sample size $n$. We work with the same prior setting as Peruggia (1997):

$$\beta \mid \Sigma \sim N(0, \Sigma),$$
$$\sigma^2 \sim \text{IG}(1, 1),$$
$$\Sigma \sim \text{IW}(4 I_{2 \times 2}, 4),$$

where IG denotes the inverse gamma distribution, IW denotes the inverse Wishart distribution, and $I_{2 \times 2}$ denotes the $2 \times 2$ identity matrix. For the loss functions, we use $\ell_2$ loss and scaled $\ell_1$ loss:

$$\nu_{\ell_2}(Y_i, X_i, \beta, \sigma) = |Y_i - \beta_0 - X_i\beta_1|^2$$
$$\nu_{\text{scaled }\ell_1}(Y_i, X_i, \beta, \sigma) = |Y_i - \beta_0 - X_i\beta_1|/\sigma.$$

For each $R$, we simulate the training data set 50 times and obtain 50 values of PCICG and IS-CV. We then calculate the average of the generalization errors using a test data set with a sample size of 10. We used 3980 MCMC samples after thinning out by 5 and a burnin period of length 100.

Figure 3 displays the comparison between PCICG and IS-CV. Consider $\nu_{\ell_2}$. For $R \leq 3$, the performance of the two methods does not differ. For $R \geq 4$, the bias of IS-CV relative to the average of the generalization errors, is larger than that of PCICG. Consider $\nu_{\text{scaled }\ell_1}$. In this case, for all $R$, the bias of IS-CV, relative to the average of the generalization errors, is larger than that of PCICG.
4. Discussions

4.1. Connection to local case-sensitivity. We begin by discussing the connection between the proposed method and local case-sensitivity. For Bayesian local-case sensitivity, Millar and Stewart (2007) discussed the use of curvature of the Kullback–Leibler divergence between the weighted posterior distribution \( \pi_w(\theta \mid X^n) \) and the posterior distribution \( \pi(\theta \mid X^n) \): For the \( i \)-th observation, the curvature is defined by

\[
I_i^{(2)} := \frac{\partial^2}{\partial w_i^2} \left[ \int \pi_w(\theta \mid X^n) \log \frac{\pi_w(\theta \mid X^n)}{\pi(\theta \mid X^n)} d\theta \right]_{w_j = 1, j = 1, \ldots, n}.
\]

Millar and Stewart (2007) also showed that, under certain conditions, this measure is calculated as

\[
I_i^{(2)} = \mathbb{V}_{\text{pos}}[\log p(X_i \mid \theta)],
\]

where \( \log p(x \mid \theta) \) is the log-likelihood.

![Figure 4](image.png)

**Figure 4.** The values of \( M_{\nu,i} \) divided by \( \sum_{i=1}^{n} M_{\nu,i} \) for the \( \ell_2 \) loss \( \nu_{\ell_2} \) and the scaled \( \ell_1 \) loss \( \nu_{\text{scaled} \ell_1} \).

From the perspective of estimating the generalization error, this measure expresses the contribution of each observation in filling the generalization gap when, using the log-likelihood and usual posterior distribution. Together with Theorem 2.5, this viewpoint suggests yet another measure of local influence of observation \( X_i \). It is given as follows:

\[
M_{\nu,i} := \text{Cov}_{\text{pos}}[\nu(X_i, \theta), s(X_i, \theta)].
\]

This measures the impacts of the observation on the generalization with the evaluation function \( \nu \) and the quasi-posterior \( \pi(\theta \mid X) \).

Figure 4 shows how the measure \( M_{\nu,i} \) works in the example of Subsection 3.3. We consider the \( \ell_2 \) losses and the scaled \( \ell_1 \) losses, as \( \nu \). In both losses, the measures \( M_{\nu,i} \) divided by \( \sum_{i=1}^{n} M_{\nu,i} \) have peaks at \( i = 50 \), which implies that the measures \( M_{\nu,i} \) successfully detect the influential observation.

4.2. The infinitesimal jackknife approximation of leave-one-out cross validation. The infinitesimal jackknife is a general methodology that approximates algorithms requiring the re-fitting of models, such as cross validation and the bootstrap methods. In the recent machine learning literature, this methodology has been rekindled as a linear approximation of LOOCV; see Beirami et al. (2017); Koh and Liang (2017); Giordano et al. (2019); Rad and Maleki (2020).

In the literature on information criteria, this methodology has been used to show the asymptotic equivalence between information criteria and LOOCV (Stone, 1974). Our proof focuses on this equivalence result and derives the proposed method from the Bayesian LOOCV. The infinitesimal jackknife approximation of the LOOCV estimate requires the second order differentiation and its inverse calculation (c.f., Beirami et al., 2017; Koh and Liang, 2017). By presenting the proof in Subsection 2.2, we emphasise
that these calculi are avoidable in Bayesian framework. Note that, in assessing point estimates, Giordano (2017) also points out this computational merit of the Bayesian infinitesimal jackknife approximation.

5. CONCLUSION

We have proposed a novel, computationally low-cost method of estimating the Gibbs generalization errors and plugin generalization errors for arbitrary loss functions. We have demonstrated the usefulness of the proposed method in privacy-preserving learning. An important practical implication of this study is that the posterior covariance provides an easy-to-implement generalization error estimate for arbitrary loss functions, and can avoid the cumbersome refitting in LOOCV as well as the importance sampling technique that is sensitive to the presence of influential observations. Also, theoretical connections between WAIC, the Bayesian sensitivity analysis, and the infinitesimal jackknife approximation of Bayesian LOOCV are clarified by our proof for the asymptotic unbiasedness.

ACKNOWLEDGEMENT

The authors would like to thank Akifumi Okuno, Hironori Fujisawa, Yoshiyuki Ninomiya, and Yusaku Ohkubo for providing them with fruitful discussions. This work was supported by Japan Society for the Promotion of Science (JSPS) [Grant Nos. 19K20222, 21H05205, 21K12067], the Japan Science and Technology Agency’s Core Research for Evolutional Science and Technology (JST CREST) [Grant No. JPMJCR1763], and the MEXT Project for Seismology toward Research Innovation with Data of Earthquake (STAR-E) [Grant No. JPJ010217].

REFERENCES

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory (eds. B. Petrov and F. Csaki), 267–281.
[2] Ando, T. (2007) Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94, 443–458.
[3] Beirami, A., Razaviyayn, M., Shahrampour, S. and Tartokh, V. (2017) On optimal generalizability in parametric learning. In *Advances in Neural Information Processing Systems*, 3458–3468.
[4] Dua, D. and Graff, C. (2017) UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.
[5] Dwork, C. (2006) Differential privacy. In *Proceedings of the 33rd International conference on Automata, Languages and Programming, Part II*, 1–12.
[6] Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of American Statistical Association*, 70, 320–328.
[7] Gelfand, A., Dey, D. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics* (eds. J. Bernardo, J. Berger, A. Dawid and A. Smith), 147–167. Oxford University Press.
[8] Giordano, R. (2017) Stansensitivity. https://github.com/rgiordan/StanSensitivity
[9] Giordano, R., Broderick, T. and Jordan, M. (2018) Covariances, robustness, and variational bayes. *Journal of Machine Learning Research*, 19, 1–49. URL: http://jmlr.org/papers/v19/17-670.html
[10] Giordano, R., Stephenson, W., Liu, R., Jornan, M. and Broderick, T. (2019) A swiss army infinitesimal jackknife. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* 2019, 89.
[11] Iba, Y. and Yano, K. (2020) Posterior covariance information criterion. ArXiv:2106.13694.
[12] Koh, P. and Liang, P. (2017) Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 1885–1894.
[13] Konishi, S. and Kitagawa, G. (1996) Generalised information criteria in model selection. *Biometrika*, 83, 875–890.
[14] Millar, R. and Stewart, W. (2007) Assessment of locally influential observations in Bayesian models. *Bayesian Analysis*, 2, 365–384.
This appendix provides the detailed calculation used in Section 3.1. The calculation employs the following lemma.

**Lemma A.1** (Kumar, 1973). Let \( w \) be a random vector from \( N(0, I_d) \). For \( d \times d \) symmetric matrices \( B \) and \( C \), we have

\[
\mathbb{E}[(w^\top Bw)(w^\top Cw)] = 2\text{tr}(BC) + (\text{tr} B)(\text{tr} C).
\]
Step 1. Establishing (7): Let us begin by expanding $\text{Cov}_{\text{pos}}[s(X_i, \theta), \nu(X_i, \theta)]$. For $i = 1, \ldots, n$, we have

$$
\begin{align*}
\mathbb{E}_{\text{pos}}[(X_i - \theta)^\top A(X_i - \theta)(X_i - \theta)^\top (X_i - \theta)] & = \mathbb{E}_{\text{pos}} \left\{ (X_i - \hat{\theta})^\top A(X_i - \hat{\theta}) + 2(\hat{\theta} - \theta)^\top A(X_i - \hat{\theta}) + \hat{\theta} - \theta)^\top A(\hat{\theta} - \theta) \right\} \\
& = \bar{X}_i^\top A\bar{X}_i + \mathbb{E}_{\text{pos}}[2\bar{X}_i^\top A\bar{X}_i \tilde{\theta}^\top \tilde{\theta}] + \mathbb{E}_{\text{pos}}[\bar{X}_i^\top A\bar{X}_i \theta^\top \tilde{\theta}] + \mathbb{E}_{\text{pos}}[2\tilde{\theta}^\top A\bar{X}_i \tilde{\theta}] + \mathbb{E}_{\text{pos}}[2\tilde{\theta}^\top A\bar{X}_i \tilde{\theta}]
\end{align*}
$$

where $\bar{X}_i := X_i - \hat{\theta}$ and $\tilde{\theta} := \hat{\theta} - \theta$. Lemma A.1 gives

$$
\mathbb{E}_{\text{pos}}[(\tilde{\theta} - \theta)^\top (\tilde{\theta} - \theta) A(\hat{\theta} - \theta)] = \frac{2 + d}{(n\beta + 1/\tau)^2} \text{tr}(A)
$$

and thus we get

$$
\begin{align*}
\mathbb{E}_{\text{pos}}[(X_i - \theta)^\top (X_i - \theta)(X_i - \theta)^\top A(X_i - \theta)] & = \bar{X}_i^\top A\bar{X}_i + \frac{4 + d}{n\beta + 1/\tau} \bar{X}_i^\top A\bar{X}_i + \frac{\text{tr}(A)}{n\beta + 1/\tau} \bar{X}_i^\top A\bar{X}_i + \frac{2 + d}{(n\beta + 1/\tau)^2} \text{tr}(A)
\end{align*}
$$

Further, for $i = 1, \ldots, n$, we have

$$
\begin{align*}
\mathbb{E}_{\text{pos}}[(X_i - \theta)^\top A(X_i - \theta)] & = \left\{ \bar{X}_i^\top \bar{X}_i + \frac{d}{n\beta + 1/\tau} \right\} \left\{ \bar{X}_i^\top A\bar{X}_i + \frac{\text{tr}(A)}{n\beta + 1/\tau} \right\} \\
& = \bar{X}_i^\top \bar{X}_i A\bar{X}_i + \frac{\text{tr}(A)}{n\beta + 1/\tau} \bar{X}_i^\top \bar{X}_i + \frac{d}{n\beta + 1/\tau} \bar{X}_i^\top A\bar{X}_i + \frac{d}{n\beta + 1/\tau} \frac{\text{tr}(A)}{n\beta + 1/\tau}.
\end{align*}
$$

Combining (11) and (12) yields

$$
\text{Cov}_{\text{pos}}[\nu(X_i, \theta), s(X_i, \theta)] = -\frac{\beta}{2} \left\{ \frac{4}{n\beta + 1/\tau} \bar{X}_i^\top A\bar{X}_i + \frac{2\text{tr}(A)}{(n\beta + 1/\tau)^2} \right\},
$$

which implies (7).

Step 2. Establishing (8): Next, we take the expectation of $(1/n) \sum_{i=1}^n (X_i - \hat{\theta})^\top A(X_i - \hat{\theta})$. Let $a := (n\beta)(n\beta + 1)/(n\beta + 1)$. Then, we have

$$
\begin{align*}
\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\theta})^\top A(X_i - \hat{\theta}) \right] & = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \theta^*)^\top A(X_i - \theta^*) \right] + (1 - a)^2(\theta^*)^\top A\theta^* + (a^2 - 2a)\mathbb{E} \left[ (\bar{X} - \theta^*)^\top A(\bar{X} - \theta^*) \right] \\
& = \text{tr}(A) + (1 - a)^2(\theta^*)^\top A\theta^* + (a^2 - 2a) \frac{\text{tr}(A)}{n},
\end{align*}
$$

which implies (8).
Step 3. Establishing (10): Finally, we consider $\text{Cov}_{\text{pos}}[(X_i - \theta)^\top A(X_i - \theta), \theta^\top \theta]$. For $i = 1, \ldots, n$, we have

$$\begin{align*}
\text{E}_{\text{pos}}[(X_i - \theta)^\top A(X_i - \theta)\theta^\top \theta] &= \text{E}_{\text{pos}}\left\{ (X_i - \hat{\theta})^\top A(X_i - \hat{\theta}) + (\hat{\theta} - \theta)^\top A(\hat{\theta} - \theta) + 2(X_i - \hat{\theta})^\top A(\hat{\theta} - \theta) \right\} \\
&= \frac{1}{n\beta + 1/\tau} (X_i - \hat{\theta})^\top A(X_i - \hat{\theta}) + (X_i - \hat{\theta})^\top A(X_i - \hat{\theta})\hat{\theta}^\top \hat{\theta} + \frac{\text{tr}(A)}{n\beta + 1/\tau} \hat{\theta}^\top \hat{\theta} \\
&\quad + \frac{2 + d}{(n\beta + 1/\tau)^2} \text{tr}(A) + \frac{4}{n\beta + 1/\tau} (X_i - \hat{\theta})^\top A\hat{\theta}
\end{align*}$$

and we have

$$\begin{align*}
\text{E}_{\text{pos}}[(X_i - \theta)^\top A(X_i - \theta)]\text{E}_{\text{pos}}[\theta^\top \theta] &= \left\{ (X_i - \hat{\theta})^\top A(X_i - \hat{\theta}) + \frac{\text{tr}(A)}{n\beta + 1/\tau} \right\} \left\{ \hat{\theta}^\top \hat{\theta} + \frac{1}{n\beta + 1/\tau} \right\} \\
&= \frac{1}{n\beta + 1/\tau} (X_i - \hat{\theta})^\top A(X_i - \hat{\theta}) + (X_i - \hat{\theta})^\top A(X_i - \hat{\theta})\hat{\theta}^\top \hat{\theta} + \frac{\text{tr}(A)}{n\beta + 1/\tau} \hat{\theta}^\top \hat{\theta} + \frac{\text{tr}(A)}{(n\beta + 1/\tau)^2}.
\end{align*}$$

Combining these yields

$$\text{Cov}_{\text{pos}}[(X_i - \theta)^\top A(X_i - \theta), \theta^\top \theta] = -\frac{4}{n\beta + 1/\tau} (X_i - \hat{\theta})^\top A\hat{\theta} + \frac{1 + d}{(n\beta + 1/\tau)^2} \text{tr}(A).$$

This, together with the identity

$$\text{E}\left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\theta})^\top A\hat{\theta} \right] = a(1 - a) \text{E}\left[ \overline{X}^\top A\overline{X} \right] = a(1 - a) \{ (\theta^*)^\top A\theta^* + \text{tr}(A)/n \},$$

yields (10).

Appendix B. Proof for Lemma 2.6

This appendix provides the proof of Lemma 2.6. The Lebesgue dominated convergence theorem ensures the exchange of differentiation and integration in $(\partial^k/\partial w_i^k)\text{E}_{\text{pos}}^w[\nu(X_i, \theta)]$ under Condition C3. Then, considering

$$F(w) := \int \exp\left\{ \sum_{i=1}^{n} w_is(X_i, \theta) \right\} \pi(\theta) d\theta$$

gives

$$\begin{align*}
\frac{\partial}{\partial w_i} \text{E}_{\text{pos}}^w[\nu(X_i, \theta)] &= \frac{\int s(X_i, \theta) \nu(X_i, \theta)e^{\sum_{i=1}^{n} w_is(X_i, \theta)} \pi(\theta) d\theta F(w)}{F^2(w)} \\
&\quad - \frac{\int s(X_i, \theta)e^{\sum_{i=1}^{n} w_is(X_i, \theta)} \pi(\theta) d\theta F(w)}{F^2(w)} \\
&\quad \int \nu(X_i, \theta)e^{\sum_{i=1}^{n} w_is(X_i, \theta)} \pi(\theta) d\theta F(w) \\
&\quad - \frac{\int \nu(X_i, \theta)e^{\sum_{i=1}^{n} w_is(X_i, \theta)} \pi(\theta) d\theta F(w)}{F^2(w)},
\end{align*}$$

which implies

$$\frac{\partial}{\partial w_i} \text{E}_{\text{pos}}^w[\nu(X_i, \theta)] = \text{E}_{\text{pos}}^w \left\{ \nu(X_i, \theta) - \text{E}_{\text{pos}}^w[\nu(X_i, \theta)] \right\} \{ s(X_i, \theta) - \text{E}_{\text{pos}}^w[s(X_i, \theta)] \}.$$

14
Further, we have

\[
\frac{\partial}{\partial w_i} \left( \int s(X_i, \theta) \nu(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta F(w) \right) = \int s^2(X_i, \theta) \nu(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta F(w) - \int s(X_i, \theta) \nu(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta \int s(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta F^2(w) + \int s(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta \int \nu(X_i, \theta) s(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta d\theta F^2(w) - 2 \left( \int s(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta \right)^2 \int \nu(X_i, \theta) e^{\sum_{i=1}^{n} w_i s(X_i, \theta) \pi(\theta)} d\theta F(w)
\]

which implies

\[
\frac{\partial^2}{\partial w_i^2} \mathbb{E}_{\text{pos}}^{w}[\nu(X_i, \theta)] = \mathbb{E}_{\text{pos}}^{w} \left[ \nu(X_i, \theta) - \mathbb{E}_{\text{pos}}^{w}[\nu(X_i, \theta)] \right] \{s(X_i, \theta) - \mathbb{E}_{\text{pos}}^{w}[s(X_i, \theta)] \}^2
\]

and completes the proof. \(\square\)

REFERENCES FOR APPENDICES

[1] Kumar, A. (1973) Expectation of products of quadratic forms. *Sankha*, **35**, 359–362.