

Robust leave-one-out cross-validation for high-dimensional Bayesian models

Luca Silva and Giacomo Zanella

September 20, 2022

Abstract

Leave-one-out cross-validation (LOO-CV) is a popular method for estimating out-of-sample predictive accuracy. However, computing LOO-CV criteria can be computationally expensive due to the need to fit the model multiple times. In the Bayesian context, importance sampling provides a possible solution but classical approaches can easily produce estimators whose variance is infinite, making them potentially unreliable. Here we propose and analyze a novel mixture estimator to compute Bayesian LOO-CV criteria. Our method retains the simplicity and computational convenience of classical approaches, while guaranteeing finite variance of the resulting estimators. Both theoretical and numerical results are provided to illustrate the improved robustness and efficiency. The computational benefits are particularly significant in high-dimensional problems, allowing to perform Bayesian LOO-CV for a broader range of models. The proposed methodology is easily implementable in standard probabilistic programming software and has a computational cost roughly equivalent to fitting the original model once.

Keywords: Leave-One-Out Cross-Validation, Importance Sampling, Model Evaluation, Bayesian Analysis, Markov Chain Monte Carlo.

1 Introduction

Consider a Bayesian model with conditionally independent observations $y = (y_1, \dots, y_n)$ given a set of parameters θ , and denote the resulting joint distribution of θ and y as

$$p(\theta, y) = p(\theta) \prod_{i=1}^n p(y_i | \theta). \quad (1)$$

Given some observed data y , the model yields a posterior distribution over the unknown parameters, $p(\theta | y)$, and a posterior predictive distribution at a new point y_{new} given by

$$p(y_{new} | y) = \int p(y_{new} | \theta) p(\theta | y) d\theta.$$

In many contexts, one is interested in quantifying the out-of-sample performances of such predictive distribution, for example to optimize some tuning hyper-parameter or to compare different models. Assuming the existence of a true data-generating process p^* , a

gold-standard measure of predictive performances is the expected log predictive density (ELPD) defined as

$$\text{ELPD} = \int \log p(y_{\text{new}}|y) p^*(y_{\text{new}}) dy_{\text{new}}. \quad (2)$$

While alternative scoring rules for predictive distributions could be used, here we focus on the logarithmic one since it has both strong decision theoretic justification and is the one most commonly used in practice, see Section 1.1 for more details.

The true data generating distribution p^* is unknown in practice, and the ELPD is typically approximated through cross-validation. In particular, leave-one-out cross-validation (LOO-CV) leads to an estimator of (n times) the ELPD defined as

$$\psi := \sum_{i=1}^n \log p(y_i|y_{-i}) = \sum_{i=1}^n \log \left(\int p(y_i|\theta) p(\theta|y_{-i}) d\theta \right), \quad (3)$$

where $y_{-i} = (y_j)_{j \neq i}$. The LOO-CV estimator ψ has appealing statistical properties [Vehtari and Ojanen, 2012], such as smaller bias compared to k -fold CV with small k , especially in high-dimensional contexts [Rad et al., 2020]. However, naive approaches to compute ψ require to fit the original model n times, one for each LOO dataset y_{-i} , thus being computationally infeasible. When using Monte Carlo methods to perform computations, a classical solution is to draw samples from $p(\theta|y)$ only once and then resort to importance sampling to approximate each LOO posterior $p(\theta|y_{-i})$ [Gelfand et al., 1992]. However, as previously noted in the literature, the resulting estimators of $\{p(y_i|y_{-i})\}_{i=1}^n$ are often unreliable and can easily have infinite variance [Peruggia, 1997, Epifani et al., 2008]. Here we propose novel estimators of $\{p(y_i|y_{-i})\}_{i=1}^n$, based on a mixture representation of leave-one-out posteriors. Unlike standard estimators in the literature, our method provides guarantees on the finiteness of the estimator’s variance and performs dramatically better in high-dimensional problems, where LOO-CV is particularly appealing and standard competitors break down (see e.g. results in Sections 3 and 4). Crucially, our methodology requires only a single sampling procedure and it can be trivially implemented in probabilistic programming languages, thus preserving the practicality and limited computational cost of previously proposed and widely used solutions [Gelfand et al., 1992, Vehtari et al., 2017], while offering drastically improved robustness to high-dimensional scenarios.

More generally, our work support recent evidence, both in the Bayesian and frequentist literature [Beirami et al., 2017, Rad et al., 2020, Giordano et al., 2019, Paananen et al., 2021], that LOO-CV criteria can be reliably approximated with a computational cost comparable to the one of a single model fit, thus being not only statistically appealing but also computationally practical. In this sense LOO-CV can be computationally cheaper than k -fold CV by a factor of k , since the latter requires fitting k separate models and is not easily amenable to the same importance sampling tricks as LOO-CV. Such k -times speed-up can be crucial in the context of Bayesian computation with Monte Carlo methods where each model fitting can be expensive. In this context, our work contributes to prevent one of the main factor limiting the applicability of Bayesian LOO-CV, i.e. the potential instability of classical estimators of $\{p(y_i|y_{-i})\}_{i=1}^n$.

The article is organized as follows: after briefly reviewing relevant literature in Section 1.1, we describe our proposed methodology and compare to the classical one in Section 2. Section 3 provides some theoretical analysis of the resulting estimators, including a proof of finite variance and a comparison to standard methods in high-dimensional regression contexts. Section 4 provides numerical results that support the theoretical findings and illustrate the improved robustness both to the presence of model misspecification and to high-dimensionality of the parameter space. Finally, Section 5 discusses simple extensions of our methodology (e.g. different scoring rules or non conditionally-independent models). For notational brevity, throughout the paper we use the same letter p to denote appropriate joint, marginal and conditional distributions of the model for θ , y and y_{new} , as done in (1)-(3). Similarly, we leave the dependence of $p(y_i|\theta)$ and $p(y_{new}|\theta)$ on additional covariates or other variables implicit in the notation.

1.1 Predictive criteria for Bayesian model validation and selection

Motivations to estimate out-of-sample predictive measures such as ELPD in (2) include hyper-parameters tuning and model validation, selection and averaging. In such contexts, criteria based on predictive distributions have complementary roles compared to ones based directly on posterior distributions, such as Bayes factors and classical Bayesian model averaging [Hoeting et al., 1999]. Using Box [1980] words, posterior distributions provide a basis for “estimation of parameters conditional on the adequacy of the entertained model” while predictive distributions enable “criticism of the entertained model in the light of current data”. Practical advantages of predictive-based criteria include being more directly comparable across different models (including non-nested ones), and being typically less sensitive to prior specifications compared to Bayes factors, including vague priors as in e.g. Bartlett’s paradox in Bayesian model selection Bartlett, 1957, Lindley, 1957, Liang et al., 2008. The literature on the topic is vast: see for example Gelfand and Dey [1994], Vehtari and Ojanen [2012] and reference therein for an overview and some arguments in favour of Bayesian predictive measures and cross-validation criteria, and Fong and Holmes [2020] for a recent theoretical comparison between cross-validation and marginal likelihood techniques. While there are various scoring functions to evaluate predictive distributions [Gneiting, 2011], here we focus on the logarithmic one, which is the unique local and proper scoring rule [Bernardo, 1979] and the most widely used in practice. See for example Gelman et al. [2014] for arguments in favour of using the ELPD metric in (2) and its LOO-CV estimator in (3). Beyond computing ELPD estimates as in (3), the LOO predictive probabilities $\{p(y_i|y_{-i})\}_{i=1}^n$ are also of interest in themselves, as they allow to implement methodologies aiming at optimizing predictive performances such as Bayesian stacking (see e.g. Yao et al. [2018] and references therein) or at identify discording observations (see e.g. the notion of conditional predictive ordinate Pettit, 1990) which then leads to model improvements and refinements. See also Weiss and Cho [1998], Epifani et al. [2008] and references therein for related discussion.

2 Computing Bayesian leave-one-out cross validation

In this paper we focus on Monte Carlo methodologies to compute the LOO predictive probabilities $\{p(y_i|y_{-i})\}_{i=1}^n$. Depending on the context, these may be themselves the quantities of interest, or an intermediate step to compute LOO-CV criteria such as ψ defined in (3). In the latter case an estimate of ψ is simply obtained by plugging-in the estimates of $p(y_i|y_{-i})$ in (3).

The first, somehow brute-force, approach to this computation would be to fit n times the model separately. Recalling that $p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta$, one could draw S Monte Carlo samples from each LOO posterior $p(\theta|y_{-i})$, using e.g. n separate MCMC runs, and then estimate $p(y_i|y_{-i})$ with the resulting sample average of $p(y_i|\theta)$. We denote the resulting estimators of $\mu_i := p(y_i|y_{-i})$ as

$$\hat{\mu}_i^{(loo)} = S^{-1} \sum_{s=1}^S p(y_i|\theta_s) \quad (4)$$

where $\theta_1, \theta_2, \dots, \theta_S \sim p(\theta|y_{-i})$. Assuming the computational cost of each Monte Carlo sample to grow linearly with n , this would require $\Theta(Sn)$ samples and $\Theta(Sn^2)$ computational cost in total, which is typically unfeasible.

A potential solution proposed in [Gelfand et al., 1992] is to instead draw only one set of samples from the full-data posterior, and then use importance sampling to approximate expectations with respect to the n different LOO posteriors. This leads to unnormalized importance weights between the i -th LOO posterior and the full posterior equal to

$$w_i^{(post)}(\theta) = p(y_i|\theta)^{-1} \propto \frac{p(\theta|y_{-i})}{p(\theta|y)}.$$

The corresponding self-normalized importance sampling estimator of $p(y_i|y_{-i})$ is

$$\hat{\mu}_i^{(post)} = \frac{\sum_{s=1}^S p(y_i|\theta_s) w_i^{(post)}(\theta_s)}{\sum_{s=1}^S w_i^{(post)}(\theta_s)} = \left(S^{-1} \sum_{s=1}^S p(y_i|\theta_s)^{-1} \right)^{-1} \quad (5)$$

where $\theta_1, \theta_2, \dots, \theta_S \sim p(\theta|y)$. This procedure is practically appealing because it only requires one sampling routine and has $\Theta(Sn)$ total cost, including the computation of the n estimators $\{\hat{\mu}_i^{(post)}\}_{i=1}^n$, each of which can be obtained at $\Theta(S)$ cost given the samples $\{\theta_s\}_{s=1}^S$ using definition (5). The drawback is that the resulting importance sampling estimators can be unstable and even have infinite variance. In such cases the estimators are still consistent, i.e. $\lim_{S \rightarrow \infty} \hat{\mu}_i^{(post)} = p(y_i|y_{-i})$ almost surely, but the central limit theorem and the $S^{-1/2}$ rate of convergence may not hold [Epifani et al., 2008]. These issues are not surprising if one realizes that (5) is a variation of the classical harmonic-mean estimator [Newton and Raftery, 1994], which has well-known stability issues. This has motivated proposals in the literature to improve the stability of LOO-CV estimators as well as to diagnose their potential failure. A notable example that we compare with in simulations later on is the Pareto-smoothed importance sampling methodology of [Vehtari et al., 2017]

implemented in the popular LOO R package [Vehtari et al., 2020]. See also Alqallaf and Gustafson [2001], Bornn et al. [2010], Rischard et al. [2018], Paananen et al. [2021] for other work in the area, and Section 4.3.1 for comparison with some of those.

2.1 Mixture estimators

Here we propose a different set of estimators for $\{p(y_i|y_{-i})\}_{i=1}^n$ with drastically improved robustness to high-dimensionality, which we achieve expressing the problem in terms of mixtures rather than harmonic mean identities. We introduce a component indicator I , formally a random variable on $\{1, \dots, n\}$, and define a joint distribution for θ and I as

$$q_{mix}(\theta, I) = \frac{p(\theta)p(y_{-I}|\theta)}{\sum_{j=1}^n p(y_{-j})} \quad (\theta, I) \in \Theta \times \{1, \dots, n\}. \quad (6)$$

Here q_{mix} is defined so that $q_{mix}(\theta|I=i) = p(\theta|y_{-i})$ and thus $p(y_i|y_{-i})$ can be written as the following conditional expectation

$$p(y_i|y_{-i}) = \mathbb{E}_{(\theta, I) \sim q_{mix}} [p(y_i|\theta)|I=i]. \quad (7)$$

This representation leads to our proposed set of estimators, which are obtained through the following steps:

- (i) draw S samples $\theta_1, \theta_2, \dots, \theta_S$ from $q_{mix}(\theta)$, where

$$q_{mix}(\theta) = Z^{-1} \sum_{j=1}^n p(\theta)p(y_{-j}|\theta) \propto p(\theta|y) \left(\sum_{j=1}^n p(y_j|\theta)^{-1} \right), \quad (8)$$

is the marginal distribution of θ under the joint $q_{mix}(\theta, I)$ and $Z = \sum_{j=1}^n p(y_{-j})$. Sampling from (8) can be done using standard MCMC algorithms, as discussed below and in Appendix A;

- (ii) for each $i \in \{1, \dots, n\}$, obtain weighted samples from $p(\theta|y_{-i})$ assigning to each sample in $\{\theta_s\}_{s=1, \dots, S}$ the weight

$$w_i^{(mix)}(\theta) = q_{mix}(I=i|\theta) = \frac{p(y_i|\theta)^{-1}}{\sum_{j=1}^n p(y_j|\theta)^{-1}},$$

which is the conditional probability of $I=i$ given θ under the joint distribution $q_{mix}(\theta, I)$;

- (iii) for each $i \in \{1, \dots, n\}$, estimate $p(y_i|y_{-i})$ with

$$\hat{\mu}_i^{(mix)} = \frac{\sum_{s=1}^S p(y_i|\theta_s) w_i^{(mix)}(\theta_s)}{\sum_{s=1}^S w_i^{(mix)}(\theta_s)}. \quad (9)$$

The estimator in (9) can also be interpreted as a self-normalized importance sampling estimator with importance distribution $q_{mix}(\theta)$ and target distribution $p(\theta|y_{-i})$, so that $w_i^{(mix)}(\theta)$ are unnormalized importance weights between target and importance distribution. We often use this formulation when proving theoretical results in Section 3.

The proposed estimators $\{\hat{\mu}_i^{(mix)}\}_{i=1}^p$ retain the simplicity and computational practicality of the classical ones in (5). In fact a single sampling routine is required, this time from $q_{mix}(\theta)$, and the total computational cost to obtain the n estimators $\{\hat{\mu}_i^{(mix)}\}_{i=1}^n$ is still $\Theta(Sn)$. The latter follows from two crucial remarks. First, evaluating $q_{mix}(\theta)$ up to normalizing constant requires $\Theta(n)$ cost using the last expression in (8), see also (21) in Appendix A.1. Note that a naive use of the first expression in (8) would instead incur in a $\Theta(n^2)$ cost. Second, computing $\{\hat{\mu}_i^{(mix)}\}_{i=1}^n$ in (9) requires first an $\Theta(Sn)$ computation common to all i 's, namely the computation of $\{\sum_{j=1}^n p(y_j|\theta_s)^{-1}\}_{s=1}^S$ and, given the latter, each $\hat{\mu}_i^{(mix)}$ can be computed at $\Theta(S)$ cost. See Appendix A.2 for more details.

Also, evaluating gradients of the log of the mixture distribution, $\nabla \log q_{mix}(\theta)$, involves an $\Theta(n)$ cost, analogously to gradients of the standard log-posterior $\nabla \log p(\theta|y)$, and the whole routine is trivial to implement in probabilistic programming languages that rely on gradient-based MCMC, such as STAN [Stan Development Team, 2020]. In our numerical experiments, sampling from $p(\theta|y)$ and $q_{mix}(\theta)$ with STAN required a comparable amount of time, with only a slight overhead for q_{mix} . See Appendix A for more details on efficient and numerically stable implementation of the sampling procedure.

Remark 1 (Mixture interpretation). *The distribution q_{mix} can be interpreted as a mixture of LOO posteriors writing $q_{mix}(\theta) = \sum_{i=1}^n \pi_i p(\theta|y_{-i})$ with mixture probabilities $\pi_i = Z^{-1}p(y_{-i})$ satisfying $\sum_i \pi_i = 1$ and $\pi_i \geq 0$. Rewriting $\pi_i = \tilde{Z}^{-1}p(y_i|y_{-i})^{-1}$, with $\tilde{Z} = \sum_j p(y_j|y_{-j})^{-1}$, we can express the quantity of interest as $p(y_i|y_{-i}) = \tilde{Z}^{-1}/\pi_i$. Indeed, the denominator in (9) times S^{-1} is a consistent estimator of π_i while the numerator times S^{-1} is a consistent estimator of \tilde{Z}^{-1} . Thus, the algorithm is effectively estimating the probability π_i of each component in the mixture representation and using that to estimate $p(y_i|y_{-i})$. This is arguably where the improvement in performances of $\hat{\mu}_i^{(mix)}$ compared to $\hat{\mu}_i^{(post)}$ comes from, since mixture probabilities are typically easier to estimate than harmonic means. For example the weights $w_i^{(mix)}(\theta)$ are by construction upper bounded by 1, being conditional probabilities, which is a desirable feature to improve robustness of importance sampling estimators.*

The idea of using mixtures to derive estimators with improved stability underlies various methodologies in the Monte Carlo literature, such as Bridge Sampling and variations [Bennett, 1976, Geyer, 1991, Meng and Wong, 1996, Shirts and Chodera, 2008]. In this sense, one can think at our proposed methodology as an effective and practical way to extend these techniques to LOO-CV computation contexts while preserving a $\Theta(Sn)$ total computational cost.

Remark 2 (Choice of mixture probabilities). *Note that the mixture probabilities π_i involve the intractable terms $p(y_{-i})$ that are typically not available in closed form. However, these terms cancel with the denominator of $p(\theta|y_{-i}) = p(\theta)p(y_{-i}|\theta)/p(y_{-i})$, making $q_{mix}(\theta)$ computable up to a single intractable normalizing constant Z as in (8) and thus amenable to standard sampling algorithms. Also, since $\pi_i \propto p(y_i|y_{-i})^{-1}$, q_{mix} naturally puts more weight on mixture components with smaller $p(y_i|y_{-i})$. This is desirable since small values of $p(y_i|y_{-i})$ are typically harder to estimate and contribute more to $\psi = \sum_{i=1}^n \log p(y_i|y_{-i})$.*

In Sections 3.1 and 5 we discuss extensions to mixture constructions with general choices of mixture probabilities.

3 Analysis of the proposed estimator

In this section we provide a theoretical analysis of the proposed estimators $\{\hat{\mu}_i^{(mix)}\}_{i=1}^n$ with particular focus on comparing them with the classical ones $\{\hat{\mu}_i^{(post)}\}_{i=1}^n$. The efficiency of the different estimators, both classical and novel ones, is measured in terms of their relative asymptotic variances, defined as

$$AV_i^{(post)} = \lim_{S \rightarrow \infty} S \text{var}(\hat{\mu}_i^{(post)} / \mu_i) \quad \text{and} \quad AV_i^{(mix)} = \lim_{S \rightarrow \infty} S \text{var}(\hat{\mu}_i^{(mix)} / \mu_i), \quad (10)$$

where $\mu_i = p(y_i | y_{-i})$ as before. By the delta method we also have

$$AV_i^{(post)} = \lim_{S \rightarrow \infty} S \text{var}(\log(\hat{\mu}_i^{(post)})) \quad \text{and} \quad AV_i^{(mix)} = \lim_{S \rightarrow \infty} S \text{var}(\log(\hat{\mu}_i^{(mix)})),$$

meaning that the above terms can also be interpreted as the asymptotic variances of the plug-in estimators on the log-scale, $\log(\hat{\mu}_i^{(post)})$ and $\log(\hat{\mu}_i^{(mix)})$. Thus $\{AV_i^{(post)}\}_{i=1}^n$ and $\{AV_i^{(mix)}\}_{i=1}^n$ are a natural measure of performance when the quantities of interest are $\{\log(p(y_i | y_{-i}))\}_{i=1}^n$ or ψ in (3).

Note that the asymptotic variances in (10) refer to the case when $(\theta_s)_{s=1}^S$ in (5) and (9) are i.i.d. samples from, respectively, $p(\theta | y)$ and $q_{mix}(\theta)$. In practice, one is rarely able to draw i.i.d. samples from such distributions and instead typically relies on MCMC schemes, leading to correlated samples. In such cases the asymptotic variances of the actual estimators used in practice can be decomposed as the product of an importance sampling contribution times an MCMC contribution, namely as the product of the asymptotic variances in (10) times an MCMC integrated autocorrelation time, see e.g. Lemma 1 of Zanella and Roberts [2019]. Thus, while formally referring to the i.i.d. case, the asymptotic variances in (10) are directly relevant also to the case of MCMC sampling.

3.1 Finiteness of asymptotic variances

As mentioned above, a serious drawback of the classical estimator is that its variance $AV_i^{(post)}$ can be very large, even infinite. Indeed Peruggia [1997], Epifani et al. [2008] provide various examples, even simple ones, where $AV_i^{(post)}$ is infinite. Our first key theoretical result states that, on the contrary, the proposed mixture estimators lead finite asymptotic variances under minimal technical assumptions. In particular, we will only require that

$$p(y_i | y_{-i}) > 0 \quad \text{and} \quad \int_{\Theta} p(y_i | \theta) p(\theta | y) d\theta < \infty \quad \text{for all } i = 1, \dots, n. \quad (A1)$$

The above assumptions require the quantity of interest $p(y_i | y_{-i})$ to be non-zero, otherwise $\log p(y_i | y_{-i})$ would not be well defined, and the predictive distribution $p(y_{new} | y)$ based on the full data to be finite at $y_{new} = y_i$ for each i . These are minimal assumptions that are

typically satisfied for any model where LOO-CV quantities are of interest. Given these, we can state the following result.

Theorem 3.1.1. *Under (A1) we have that $AV_i^{(mix)} < \infty$ for all $i = 1, \dots, n$.*

Theorem 3.1.1 holds also in the more general case where q_{mix} in (8) is replaced by a weighted version $q_{mix}^{(\alpha)}(\theta) = Z_{\alpha}^{-1} \sum_{i=1}^n \alpha_i p(y_{-i}|\theta)p(\theta)$ where $Z_{\alpha} = \sum_{i=1}^n \alpha_i p(y_{-i})$ and $\alpha = (\alpha_i)_{i=1}^n$ are arbitrary weights satisfying $\alpha_i \in (0, \infty)$ for all i . In the supplement we prove the result in such more general version. See also Remark 2 and Section 5 for more details on the practical relevance of the more general weighted mixture $q_{mix}^{(\alpha)}$.

Theorem 3.1.1 highlights a first sharp distinction between the classical and mixture estimators. In fact, the difference between having finite or infinite asymptotic variance has major practical consequences, such as guaranteeing that mixture estimators will converge at the usual $S^{-1/2}$ rate, while classical estimators $\hat{\mu}_i^{(post)}$ may not. This is indeed observed in practice even for simple models, see e.g. Figure 3 in Section 4 below, and implies that in those situations the improvement in efficiency between $\hat{\mu}_i^{(mix)}$ and $\hat{\mu}_i^{(post)}$ increases to infinity as $S \rightarrow \infty$.

3.2 High-dimensional regression models

In this section we provide a more refined analysis of the behavior of $AV_i^{(post)}$ and $AV_i^{(mix)}$, focusing on high-dimensional regression models, first considering the linear case and then a more general regression context. Our results suggest that the classical estimator is highly sensitive to high-dimensionality and in particular it deteriorates as the ratio p/n increases, while the mixture estimator exhibits drastically improved robustness.

3.2.1 Connection to Bayesian leverage and the impact of high-dimensionality

Consider the regression model

$$\begin{aligned} y_i|\theta &\sim N(x_i^T \theta, \sigma^2) & i = 1, \dots, n \\ \theta &\sim N(\theta_0, \Sigma), \end{aligned} \tag{11}$$

where x_i and θ indicate $p \times 1$ matrices of, respectively, covariates and parameters. We assume the noise level σ^2 and the prior mean and covariance, θ_0 and Σ , to be fixed and known. For the linear model in (11), the finiteness of $AV_i^{(post)}$ is elegantly related to the notion of Bayesian leverage. Denoting by X the $n \times p$ matrix of covariates, define the *Bayesian hat matrix*, or *Ridge hat matrix*, as

$$H = X(X^T X + \sigma^2 \Sigma^{-1})^{-1} X^T, \tag{12}$$

which collapses to the standard (frequentist) hat matrix in the flat prior case, i.e. when $\Sigma^{-1} = 0$. The diagonal element H_{ii} represent the Bayesian leverage of the i -th observation. Thus, a higher value of H_{ii} indicates a higher discrepancy between the full posterior $p(\theta|y)$ and the LOO posterior $p(\theta|y_{-i})$, which in turn implies that the importance sampling estimator in (5) can have poor behavior. The theorem below makes the connection precise.

The connection between leverages and the finiteness of $AV_i^{(post)}$ was previously studied in Peruggia [1997]. The following result extends results therein, allowing for $p > n$ and using the notion of Bayesian leverage, rather than the frequentist one (which corresponds to $\Sigma^{-1} = 0$).

Theorem 3.2.1. *Under (11), for each $i \in \{1, \dots, n\}$, we have $AV_i^{(post)} < \infty$ if and only if $H_{ii} < 0.5$.*

The connection to Bayesian leverages provides useful insight in the behavior of the classical estimator in (5) and in particular on its dependence with respect to the dimensionality of θ and the amount of prior shrinkage.

Consider first the case of flat improper prior for θ , corresponding to $p < n$ and $\Sigma^{-1} = 0$. In such case H is the standard (frequentist) hat matrix and its trace satisfies $\sum_{i=1}^n H_{ii} = rk(X)$, where $rk(X)$ denotes the rank of X . For linearly independent predictors we have $rk(X) = p$, which implies that $H_{ii} \geq p/n$ for at least one i . Thus, by Theorem 3.2.1, as soon as $p \geq n/2$ some $AV_i^{(post)}$ must be infinite. When the entries of X are random variables (r.v.s) with complex Gaussian distributions, it holds $H_{ii} \sim Beta(p, n - p)$, see Appendix A of Chave and Thomson [2003]. This provides a more refined description of the distribution of leverages under a random design assumption and further highlights the key role of the ratio p/n , since in that case $E[H_{ii}] = p/n$. The same will hold by symmetry for any random designs with $rk(X) = p$ almost surely and distribution of X exchangeable over rows. This is consistent with our numerical experiments, where the performance of the classical estimators quickly degrade as p increases and degenerate when p is of the same order as n .

More generally, when $\Sigma = \nu^2 \mathbb{I}_p$, with \mathbb{I}_p being the $p \times p$ identity matrix, each H_{ii} is a strictly decreasing function of the so-called ridge regularization parameter $\lambda = \sigma^2 \nu^{-2}$ and the trace of H satisfies $\sum_{i=1}^n H_{ii} = \sum_{j=1}^{rk(X)} \frac{d_j^2}{d_j^2 + \lambda}$, where $(d_j)_{j=1}^{rk(X)}$ are the singular values of X [Walker and Birch, 1988]. Thus, increasing the amount of prior regularization lowers the values of the Bayesian leverages, increasing the chances of having $AV_i^{(post)} < \infty$ for all i . This is consistent with the intuition that stronger shrinkage and regularization decreases the sensitivity of the posterior to each single observation, making LOO-CV calculations potentially easier. Nonetheless, as illustrated in Figure 1 we see below, even under strong prior shrinkage the leverages H_{ii} can be large when p/n is large, leading to instability of the classical estimator.

3.2.2 Behavior of the classical and mixture estimators in large p regimes

We now provide a high-dimensional asymptotic analysis of $AV_i^{(post)}$ and $AV_i^{(mix)}$ under random design assumptions. Specifically, we assume that

$$(X_{ij})_{i,j \geq 1} \text{ are independent r.v.s with } E[X_{ij}] = 0, \text{ Var}(X_{ij}) = \tau^2 < \infty \text{ and } E[X_{ij}^4] \leq c_x \quad (\text{A2})$$

for some $c_x < \infty$. The assumption of zero mean and constant variance is realistic in settings where the regressors are standardized. While the assumption of independence is

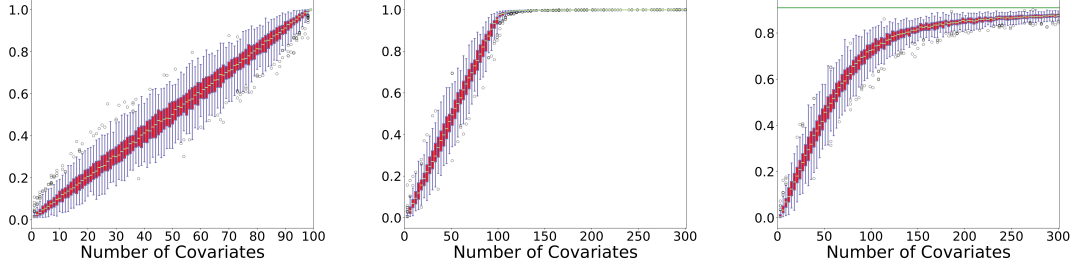


Figure 1: Distribution of the leverages H_{ii} as a function of p for $n = 100$ and $X_{ij} \stackrel{iid}{\sim} N(0, 1)$. Left: $\Sigma^{-1} = 0$, center: $\Sigma = 10\mathbb{I}_p$, right: $\Sigma = p^{-1}10\mathbb{I}_p$.

potentially restrictive, it allows to derive more intuitive and explicit results. We expect our conclusions to hold well beyond such assumption but we leave such extensions to future work. For example, a direct and relatively immediate extension would be to consider cases of weak dependence among predictors, such as Assumption 3 in Fasano et al. [2022].

We consider settings where p can be large. In such cases, it may be appropriate to assume the prior covariance of θ to vary with p . An interesting and natural setting is to take $\Sigma = \nu_p^2 \mathbb{I}_p$ with $\nu_p^2 = c/p$ for some fixed $c > 0$, which induces a prior variance of the linear predictors $\text{var}(x_i^T \theta) = c(p^{-1} \sum_{j=1}^p X_{ij}^2)$ that is approximately constant w.r.t. p and converges to the non-degenerate value $c\tau^2 \in (0, \infty)$ as $p \rightarrow \infty$ under (A2). Other regimes considered in the literature are ones where ν_p^2 is constant or where it scales as $\Theta(n/p)$. The following proposition characterizes the behaviour of H_{ii} when $p \rightarrow \infty$ for all such cases, which can be obtained with different choices of c .

Proposition 3.2.2. *Assume (11) and (A2), with $\Sigma = \nu_p^2 \mathbb{I}_p$ and $\lim_{p \rightarrow \infty} p\nu_p^2 = c \in [0, \infty]$. For each $i \in \{1, \dots, n\}$, we have*

$$H_{ii} \rightarrow \frac{c\tau^2}{\sigma^2 + c\tau^2} \quad \text{almost surely as } p \rightarrow \infty. \quad (13)$$

In the above convergence n is fixed while $p \rightarrow \infty$, and $\frac{c\tau^2}{\sigma^2 + c\tau^2} = 1$ when $c = \infty$. It follows that $AV_i^{(post)} = \infty$ almost surely for large enough p if $c\tau^2 > \sigma^2$, while $\limsup_{p \rightarrow \infty} AV_i^{(post)} < \infty$ almost surely if $c\tau^2 < \sigma^2$.

The statement about $AV_i^{(post)}$ being eventually infinite for a large enough p when $c\tau^2 > \sigma^2$ is a direct consequence of (13) and Theorem 3.2.1. This is coherent with the numerical simulations of Section 4, where the classical estimator eventually breaks down as p/n increases. The condition $c\tau^2 > \sigma^2$ is satisfied for most common prior specifications. It is obviously satisfied when ν is constant since $c = \infty$ there. Under stronger prior shrinkage where $\nu_p^2 = c/p$ with $c < \infty$, one typically sets c to some value that is significantly larger than the noise variance σ^2 , to avoid overly informative priors for the linear predictors $x_i^T \theta$, and thus $c\tau^2 > \sigma^2$ will typically hold also there. Finally, the condition $c\tau^2 > \sigma^2$ can be

directly interpreted as a comparison between prior and likelihood information, in particular as requiring the latter to be stronger than the former.

Taking the limit for $p \rightarrow \infty$ when n is fixed mimics a regime where p is large compared to n . As shown in the simulations of Section 4, such regime is highly challenging for Monte Carlo methods performing LOO-CV computations, the intuition being that the discrepancy among LOO posteriors is maximal in such regime. Large- p -small- n regimes are also interesting to consider since LOO-CV methods are particularly appealing there, due to the potentially large bias that k -fold CV methods incur in estimating ELPD in such contexts (see e.g. Rad et al. [2020] and references therein).

We now study the behaviour of $AV_i^{(mix)}$ in settings similar to Proposition 3.2.2. We first consider the case where $c < \infty$.

Theorem 3.2.3. *Assume (11) and (A2), with $\Sigma = \nu_p^2 \mathbb{I}_p$ and $\lim_{p \rightarrow \infty} p\nu_p^2 = c \in [0, \infty)$. Then we have $\limsup_{p \rightarrow \infty} AV_i^{(mix)} < \infty$ almost surely for every $i \in \{1, \dots, n\}$.*

Compared to Theorem 3.1.1, which guarantees that $AV_i^{(mix)} < \infty$ for every fixed dataset and thus for every p , Theorem 3.2.3 proves the stronger statement that each $AV_i^{(mix)}$ is also uniformly bounded with respect to p , suggesting that mixture estimators are remarkably robust to high-dimensionality of the parameter space.

3.2.3 More general regression models

We now extend some of the results derived above for the Gaussian model (11) to more general regression contexts. The results suggest that the improved robustness of $\hat{\mu}_i^{(mix)}$ compared to $\hat{\mu}_i^{(post)}$, especially in high-dimensions, is not specific to Gaussian likelihoods but rather it holds more generally. We consider regression models with general likelihood and Gaussian prior, where

$$\begin{aligned} \theta &\sim N(\theta_0, \Sigma), \\ p(y|\theta) &= \prod_{i=1}^n g(y_i|\eta_i), \quad \text{where } \eta_i = x_i^T \theta \text{ for } i = 1, \dots, n, \end{aligned} \tag{14}$$

and $g(\cdot|\cdot) : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty)$ is a generic likelihood function. The above formulation includes generalized linear models (GLM's) with Gaussian prior. Throughout, we assume the likelihood to be upper bounded, i.e. $\sup_{\eta_i} g(y_i|\eta_i) < \infty$ for any fixed $y_i \in \mathbb{R}$. The latter is arguably a mild assumption that is typically satisfied in practice.

Theorem 3.2.4. *Assume (14) and (A2), with $\Sigma = \nu_p^2 \mathbb{I}_p$ and $\lim_{p \rightarrow \infty} p\nu_p^2 = c \in [0, \infty)$. Then we have that almost surely, for each $i \in \{1, \dots, n\}$:*

- (a) $\limsup_{p \rightarrow \infty} AV_i^{(mix)} < \infty$
- (b) $\limsup_{p \rightarrow \infty} AV_i^{(post)} < \infty$ if

$$\int \exp(-\delta \eta_i^2) g(y_i|\eta_i)^{-1} d\eta_i < \infty, \tag{15}$$

for some $\delta < (2c\tau^2)^{-1}$, while $AV_i^{(post)} = \infty$ for large enough p if the integral in (15) is equal to infinity for some $\delta > (2c\tau^2)^{-1}$.

Theorem 3.2.4 extends the results of Section 3.2.2 to generic likelihoods. Namely $AV_i^{(mix)}$ is shown to remain bounded away from infinity as p grows, while $AV_i^{(post)}$ is shown to become eventually equal to ∞ when (15) does not hold, i.e. provided the likelihood function has light enough tails. In the Gaussian likelihood case, (15) coincides with requiring $\sigma^2 > c\tau^2$, which directly relates to Proposition 3.2.2 and the discussion thereafter. Condition (15) also relates to the study of $AV_i^{(post)}$ under thick-tail or light-tail priors in Epifani et al. [2008], although there the opposite scenario is considered where the likelihood is Gaussian and the prior is general and no asymptotic regime is considered.

Finally, we consider the case where the prior variance of the linear predictors diverges with p , i.e. $\lim_{p \rightarrow \infty} p\nu_p^2 = \infty$. This happens for example when ν_p^2 remains constant as $p \rightarrow \infty$. In this case $AV_i^{(mix)}$ can also diverge as $p \rightarrow \infty$, depending on the tail behavior of the likelihood function. The underlying reason is that in such cases the LOO predictive probabilities $p(y_i|y_{-i})$ go to 0 as $p \rightarrow \infty$ and even the asymptotic variance of the LOO estimators $\hat{\mu}_i^{(loo)}$, which we regard as the gold-standard but computationally expensive approach, diverge. We denote $AV_i^{(loo)} = \lim_{S \rightarrow \infty} S \text{var}(\hat{\mu}_i^{(loo)}/\mu_i)$ in the next theorem.

Theorem 3.2.5. Assume (14) and (A2), with $\Sigma = \nu_p^2 \mathbb{I}_p$ and $\lim_{p \rightarrow \infty} p\nu_p^2 = \infty$. Then we have:

- (a) if $\int g(y_i|\eta_i)d\eta_i < \infty$ for $i = 1, \dots, n$ then $\lim_{p \rightarrow \infty} AV_i^{(loo)} = \lim_{p \rightarrow \infty} AV_i^{(mix)} = \infty$ almost surely for $i = 1, \dots, n$;
- (b) if

$$\lim_{\eta_i \rightarrow \infty} g(y_i|\eta_i) + \lim_{\eta_i \rightarrow -\infty} g(y_i|\eta_i) \in (0, \infty) \quad \text{for } i = 1, \dots, n \quad (16)$$

then $\limsup_{p \rightarrow \infty} AV_i^{(mix)} < \infty$ and $\limsup_{p \rightarrow \infty} AV_i^{(loo)} < \infty$ almost surely as $p \rightarrow \infty$ for $i = 1, \dots, n$. If (16) holds and $\lim_{\eta_i \rightarrow \infty} g(y_i|\eta_i) = 0$ or $\lim_{\eta_i \rightarrow -\infty} g(y_i|\eta_i) = 0$ for $i = 1, \dots, n$, then $\lim_{p \rightarrow \infty} AV_i^{(post)} = \infty$.

Theorem 3.2.5 shows that, when $\lim_{p \rightarrow \infty} p\nu_p^2 = \infty$, the asymptotic behaviour of $AV_i^{(mix)}$, as well as $AV_i^{(loo)}$, depends on the type of likelihood in the model. For integrable likelihoods, i.e. ones satisfying $\int g(y_i|\eta_i)d\eta_i < \infty$ such as for Gaussian, Poisson, etc., the performances of all estimators under consideration (including the mixture and the gold-standard but expensive LOO ones) deteriorate as $p \rightarrow \infty$, see case (a) of Theorem 3.2.5. As mentioned above, the deterioration of performances of the mixture and LOO estimators in this case is related to the target probabilities $p(y_i|y_{-i})$ going to 0 as $p \rightarrow \infty$. Instead, for non-integrable likelihoods such as the logistic one, which falls into case (b) of Theorem 3.2.5, we have that $\lim_{p \rightarrow \infty} AV_i^{(post)} = \infty$ while $\limsup_{p \rightarrow \infty} AV_i^{(mix)} < \infty$.

4 Numerical simulations and real data examples

In this section we provide extensive numerical simulations, both on synthetic and real data, to compare the efficiency of the classical and mixture estimators. We also include

the Pareto-smoothed importance sampling (PSIS) estimator of Vehtari et al. [2017] in the comparison, which is the default methodology implemented in the popular *loo* R package [Vehtari et al., 2020] for Bayesian LOO-CV calculations. PSIS estimators are a modification of classical posterior estimators $\hat{\mu}_i^{(post)}$, where the importance weights are regularised to alleviate potential instability due to heavy-tail weight distributions.

We test the above estimators in challenging cases where the difficulty in computing the LOO predictive probabilities $\{p(y_i|y_{-i})\}_{i=1}^n$ arises mostly from two sources: (a) high-dimensionality of the parameter space and (b) presence of observations that are not well fit by the model or more generally presence of model misspecification. We test (a) mainly by considering large p scenarios and (b) by considering real datasets with (either known or potential) observations not well fit by the model. The results suggest that the mixture estimator dominates the classical and PSIS ones and, in line with the theoretical results of Section 3, that the magnitude of the improvement increases with the dimensionality of problem, while also being potentially large for low dimensional problems with highly influential observations. In Section 4.3.1 we consider also comparisons to the methodologies in Alqallaf and Gustafson [2001] and Bornn et al. [2010].

4.1 High-dimensional linear regression

We start by considering high-dimensional linear regression models where the quantities of interest $\{p(y_i|y_{-i})\}_{i=1}^n$ can be computed in closed form and the different estimators can be compared in terms of the induced mean squared errors (MSE) for a variety of setting.

4.1.1 Dependence of the estimators efficiency on n and p

First we explore how the performances of the different estimators depend on the number of data points n and parameters p . We consider the model in (11), with $\sigma^2 = 1$ and two prior specifications, one where $\Sigma = 10\mathbb{I}_p$ and one where $\Sigma = 100/p\mathbb{I}_p$. We take $n \in \{50, 100, 150\}$ and for every such value we vary p/n ranging from 0.1 to 3. For every resulting (n, p) pair we generate 1000 synthetic datasets, simulating the design matrix X with i.i.d. standard normal entries and the data y from the corresponding model likelihood in (11). For each generated dataset, we compute the exact values of $\{p(y_i|y_{-i})\}_{i=1}^n$, as well as the corresponding classical, mixture and PSIS estimators based on $S = 2 \times 10^3$ i.i.d. samples from either $p(\theta|y)$ or $q_{mix}(\theta)$. We compute the PSIS estimator using the PYTHON code available at <https://github.com/avehtari/PSIS>. We then compute the MSE of the estimators on the log scale, e.g. $\mathbb{E}[(\log(\hat{\mu}_i^{(post)}) - \log(\mu_i))^2]$ for the classical estimator. For each (n, p) pair we report the average MSE, averaging both over datasets and over $i = 1, \dots, n$. The large number of repeated datasets for each (n, p) pair was needed to ensure stable MSE estimates.

The results are reported in Figure 2. In these simulations, the PSIS estimators mildly improve over the classical ones for small-to-moderate ratios p/n but overall the two perform similarly. For example, the MSE of PSIS is never smaller than the one of posterior by more than a factor of 2, with largest reduction in MSE being roughly of 40% for values of $p/n \approx 0.35$. The mixture estimator outperforms the posterior and PSIS ones in all

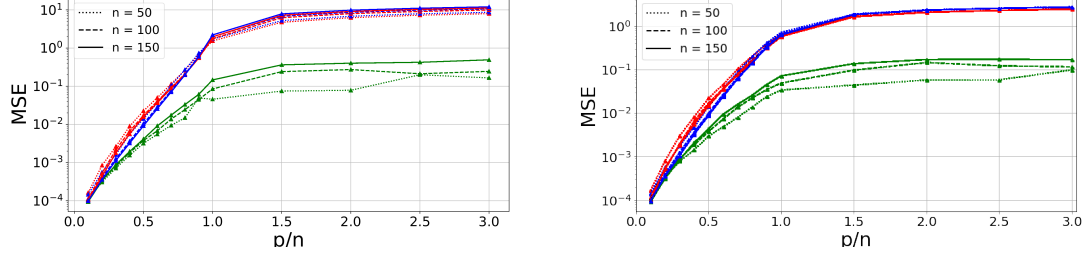


Figure 2: Mean squared error (MSE) for the posterior, PSIS and mixture estimators of $\{\log p(y_i|y_{-i})\}_{i=1}^n$ for the high-dimensional linear regression experiments of Section 4.1.1. Left: prior variance $\sigma_p^2 = 10$. Right: prior variance $\sigma_p^2 = 100/p$.

settings, with improvements that become larger as the ratio p/n increases, reaching up to one-to-two orders of magnitude reduction in MSE. In these high-dimensional regimes the classical and PSIS estimators break down (note the log-scale) while the mixture estimator remains reliable with a moderate MSE. This is in agreement with the theory in Section 3, which implies that $AV_i^{(post)}$ becomes infinite for p sufficiently large, while $AV_i^{(mix)}$ is always finite and also is uniformly bounded with respect to p when $\Sigma = c/p\mathbb{I}_p$ for some finite c . All methods perform better when the prior is more informative, i.e. when $\Sigma = 100/p\mathbb{I}_p$ compared to $\Sigma = 10\mathbb{I}_p$, which is again in accordance with the theoretical results of Section 3.

4.1.2 Infinite asymptotic variance and failure of standard rate of convergence

Next we explore more directly the impact of having a finite versus infinite asymptotic variance. This is better understood in terms of dependence of the MSE on the number of samples S , rather than fixing S and varying, e.g., n or p . In fact, when $AV_i^{(post)} = \infty$ the MSE of the estimator $\log(\hat{\mu}_i^{(post)})$ will decay at a rate slower than the classical $\mathcal{O}(S^{-1})$ Monte Carlo rate as $S \rightarrow \infty$. Figure 3 illustrate such phenomenon. The model setting and MSE computation is analogous to Figure 2, but now we vary S while fixing $p = n = 100$ and $\Sigma = 10\mathbb{I}_p$. In this setting the results of Section 3 suggest that $AV_i^{(post)} = \infty$ while $AV_i^{(mix)} < \infty$. In Figure 3 we see the MSE of the classical and PSIS estimators decaying approximately at a rate $\mathcal{O}(S^{-0.1})$ while the MSE of the mixture estimators follows the theoretical $\mathcal{O}(S^{-1})$ rate. In practice, this means that in such scenarios, despite being consistent as $S \rightarrow \infty$, the classical and PSIS estimators will require an extremely large number of samples to make the MSE small.

4.1.3 Real data, misspecification and non-conjugate priors

We now move to study how our estimator performs in a regression setting on a real dataset. We consider the *Bladder cancer* data available in the Gene Expression Omnibus (GEO) repository at <https://www.ncbi.nlm.nih.gov/gds>, with accession number *GSE31684*. The full dataset has 93 observations, and for every observation, we have 54680 covariates,

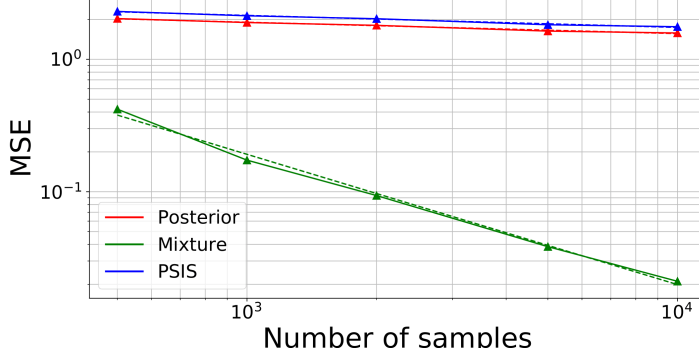


Figure 3: MSE (solid lines) as a function of the number of samples S for the different estimators, see Section 4.1.2 for details. The dashed lines have slopes of -0.984 for mixture, -0.086 for posterior, -0.091 for PSIS. Slopes far from -1 are indicative of infinite asymptotic variance and of a MSE decaying at a rate slower than the standard $\mathcal{O}(S^{-1})$ Monte Carlo one.

most of which are gene expressions of the patients. We derive different sub-datasets with varying p/n ratios by taking the first p covariates of the original dataset for $p \in \{\frac{n}{2}, n, 2n, 3n, 4n, 5n\}$. For each of the resulting six datasets, we standardize covariates and response variable to have zero mean and unit variance before fitting the model. First, we employ the usual Bayesian linear regression model with conjugate prior

$$y|X, \sigma^2 \sim N(X\theta, \sigma^2 \mathbb{I}_n) \quad \text{and} \quad \theta|\sigma^2 \sim N(\theta_0, \sigma^2 \Sigma),$$

with $\theta_0 = 0$ and $\Sigma = 100/p\mathbb{I}_p$, and set $\sigma^2 = \operatorname{argmax}_{\sigma} p(y|\sigma^2)$ in an empirical Bayes fashion. The latter operation was not needed for synthetic data since there we could simply set σ to the true data-generating value. Note that the value of σ^2 will also influence the prior variance for θ as indicated in the above model specification.

We compute estimators based on $S = 2 \times 10^4$ i.i.d. samples from either the posterior and the mixture. Figure 4 shows the resulting MSE, averaged over 100 independent repetitions. We can see that, in this real data example, the MSE values are significantly larger than the ones for simulated data with similar dimensionality and data size (see e.g. Figure 2), suggesting that real data and potential model misspecification make LOO-CV computations harder. Table 1 provides more detailed information, including averages and maximum MSE with respect to data points indices $i = 1, \dots, n$, as well as percentages of data points with large Pareto shape parameter k computed with the LOO R package [Vehtari et al., 2020] which are commonly used to diagnose instability of the classical estimators. We did not calculate such shape parameters for the mixture estimators since, by construction, their weights are upper bounded by 1 and hence the tail shape parameter is not well defined.

Finally, we consider non-conjugate priors, namely independent Laplace, or double-Exponential, priors for $\theta_1, \dots, \theta_p$ with mean parameter equal to 0 and scale parameter

num. of covariates	Estimator	MSE (average over $i = 1, \dots, n$)	MSE (max over $i = 1, \dots, n$)	% of k -shape $> .7$
p=n/2	Mixture	1.1e-03	7.0e-03	-
	Posterior	1.5e-01	5.4e-01	24%
	PSIS	1.7e-01	2.4e-01	-
p=n	Mixture	2.9e-01	1.5e+00	-
	Posterior	2.8e+00	6.1e+00	86%
	PSIS	3.1e+00	4.1e+00	-
p=2n	Mixture	7.9e-02	4.0e-01	-
	Posterior	2.6e+00	5.9e+01	99%
	PSIS	2.9e+00	3.7e+00	-
p=5n	Mixture	2.9e-02	1.2e-01	-
	Posterior	2.1e+00	4.9e+00	99%
	PSIS	2.4e+00	3.0e+00	-

Table 1: Mean squared error (MSE) of the different estimators for sub-datasets of increasing dimensionality extracted from the Bladder dataset. The % of k -shape $> .7$ refers to the diagnostic produced by the *loo* R package [Vehtari et al., 2020] to indicate unreliable estimates provided by the posterior estimators.

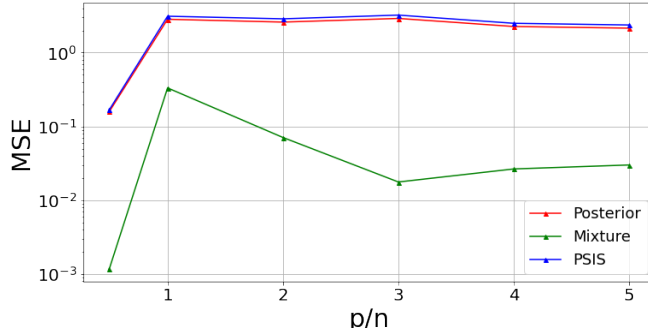


Figure 4: Average MSE on different sub-datasets of the Bladder Cancer data.

equal to $b = \sqrt{50/p}$, so to have prior variance for each coefficient equal to $100/p$. We keep a Gaussian likelihood, $y|X, \sigma^2 \sim N(X\theta, \sigma^2 \mathbb{I}_n)$, treating the noise parameter σ as unknown and assigning a *InvGamma*(4, 6) prior to it. We consider the subset of the Bladder data with $p = 2n$. Non-conjugate high-dimensional problems are challenging for Bayesian LOO-CV computations based on importance sampling and indeed most examples considered in the literature are of low or moderate dimensionality, with exceptions including [Lamnissos et al., 2012, Paananen et al., 2021]. Since the model is not conjugate the true values are not available and thus we computed an accurate approximation to those that we use as benchmark, using leave-one-out estimators based on long MCMC runs (namely using 10

num. of covariates	Estimator	MSE (average over $i = 1, \dots, n$)	MSE (max over $i = 1, \dots, n$)	% of k - shape $> .7$
p=n/2 (Laplace prior)	Mixture	3.0e-02	2.7e-01	-
	Posterior	5.6e-01	3.6e+00	86%
	PSIS	6.1e-01	2.9e+00	-

Table 2: Mean squared error (MSE) of the different estimators for a sub-dataset of the Bladder dataset with non-conjugate Laplace prior and unknown noise level.

chains with 8×10^3 samples each, resulting in 4×10^4 total samples after discarding the first half as burn-in). To ensure high quality of the samples both from the posterior and the mixture we set the STAN control values to $adapt_delta = 0.99$ and $max_treedepth = 15$ respectively. We then compute 25 independent replications of the posterior and mixture estimators based on the default STAN value of $S = 4 \times 10^3$ and report the resulting MSE in Table 2. In this example the mixture estimators provide roughly a 20 times reduction in MSE compared to the posterior ones.

4.2 Examples from the Bayesian LOO-CV literature

4.2.1 Leukaemia survival dataset

We now consider the leukaemia dataset, which is a standard example in the literature on Bayesian LOO-CV computation [Epifani et al., 2008, Vehtari et al., 2017, Rischard et al., 2018]. The dataset is used to estimate the survival distribution for leukaemia patients. The response variable is survival time (from diagnosis), and the two explanatory variables are white blood cell count at diagnosis (WBC) and the outcome of a test related to white blood cell characteristics Cook and Weisberg [1982]. Following previous analysis in the literature, we dichotomize survival times to indicate survival past 50 weeks, and we discard three repeated observation. The resulting dataset has $n = 30$ binary responses, $p = 3$ regressors including the intercept and is available at <https://github.com/luchinoprince/MixtureIS/>. We fit a Bayesian logistic regression model, meaning that each response $y_i \in \{0, 1\}$ is modelled as a Bernoulli random variables taking value 1 with success probability $(1 + \exp(x_i^T \theta))^{-1} \exp(x_i^T \theta)$, where x_i is a vector of covariates. We assume independent Laplace, or double-Exponential, priors for $\theta_1, \dots, \theta_p$ with mean parameter equal to 0 and scale parameter equal to $b = \sqrt{50/p}$, so to have prior variance for each coefficient equal to $100/p$.

This dataset is challenging for LOO-CV calculations due to the presence of a highly-influential observation, a patient with a high WBC and a survival time of more than 50 weeks, here corresponding to $i = 15$. In particular, [Epifani et al., 2008] show that for this dataset $AV_{15}^{(post)} = \infty$, while we know by Theorem 3.1.1 that $AV_{15}^{(mix)} < \infty$.

The values of $\{p(y_i|y_{-i})\}_{i=1}^n$ are not available analytically, and we compute an accurate approximations of $\{p(y_i|y_{-i})\}_{i=1}^n$ running a separate long MCMC chain to sample from $p(\theta|y_{-i})$, for each $i = 1, \dots, n$, with 10^6 iterations and first half discarded as burn in. We

treat such estimates as ground truth values, since their Monte Carlo error is negligible compared to the ones of the other estimators involved in this analysis. We then run 100 independent MCMC chains sampling from $p(\theta|y)$ and from $q_{mix}(\theta)$, of length 2×10^4 iterations each with the first half discarded as burn-in, and use the resulting samples to compute 100 i.i.d. replicates of the classical, mixture and PSIS estimators. All MCMC runs were obtained with the STAN interface in PYTHON, see e.g. <https://pystan.readthedocs.io/>, using default settings, see Appendix A for detail on how to sample from q_{mix} with STAN. No convergence or mixing issues were found using standard diagnostics.

Figure 5 reports the results displaying, for each $i = 1, \dots, n$, a box-plot of the differences between the log probability $\log(p(y_i|y_{-i}))$ and its 100 estimates. As we can see, the classical and PSIS estimators struggle to recover the true value of $\log(p(y_i|y_{-i}))$ for $i = 15$ providing highly biased estimates, which is in line with the results of Epifani et al. [2008], Vehtari et al. [2017]. On the contrary, the mixture estimator has a drastically smaller error and is centred around the correct value. All methods are able to accurately recover the ground truth values for the other values of i .

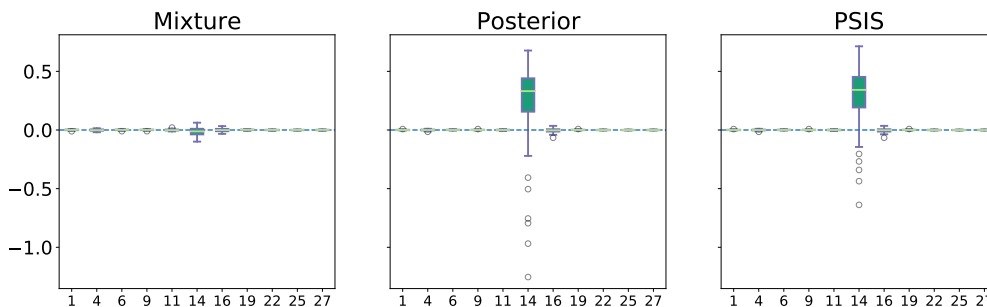


Figure 5: Errors in estimating $\{\log p(y_i|y_{-i})\}_{i=1}^n$ for the Leukaemia dataset.

4.2.2 Stack Loss dataset

We now consider a second dataset previously analysed in the Bayesian LOO-CV literature, namely the *Stack Loss* dataset. We follow Peruggia [1997] and Vehtari et al. [2017, Section 4.3], obtaining a linear regression model with $n = 21$ observations and $p = 3$ regressors. For this example Peruggia [1997] shows $AV_i^{(post)} = \infty$ for $i = 21$. Figure 6 displays the root mean squared error (RMSE) in estimating $\log(p(y_i|y_{-i}))$ for the problematic observation, $i = 21$, as well as a more ordinary observation, $i = 1$. We fit the model with different values of σ^2 , varying them over a grid centred on the maximum marginal likelihood estimator, in order to explore sensitivity to the likelihood strength. In this example PSIS improves over the posterior estimator for both $i = 1$ and $i = 21$. For both posterior and PSIS, the RMSE for $i = 21$ is an order of magnitude larger than the one for $i = 1$, while for the mixture they are of comparable order. As a result, the mixture estimator provides a major improvement for $i = 21$, while it performs comparably for $i = 1$ (slightly better or worse than posterior and PSIS depending on the value of σ^2). This relates to the fact that

mixture estimators implicitly focus more computational effort on smaller and harder to estimate values of $p(y_i|y_{-i})$ (see e.g. Remark 2), and thus they are particularly useful for those.

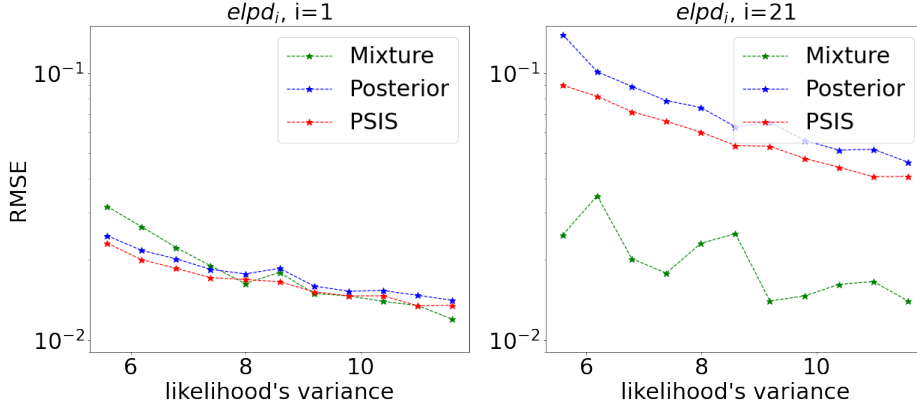


Figure 6: Root mean squared error (RMSE) in estimating $\log(p(y_i|y_{-i}))$ for the Stack Loss data for $i = 1$ (left) and $i = 21$ (right). The x-axis reports the value of σ^2 .

4.3 High-dimensional binary regression

We now consider three high-dimensional binary regression examples. We consider three real datasets from the UCI machine learning repository at <https://archive.ics.uci.edu/>, namely the *Arrhythmia*, *Voice* and *Parkinson* ones, which cover different n/p ratios. Preprocessing of the data included removal of covariates that were almost equal for all individuals, which created stability problems to the HMC algorithm implemented in STAN especially for the *Arrhythmia* dataset, and normalisation of all remaining covariates to have zero mean and unit variance. The values of (n, p) for the three datasets in their final format, which can be found at <https://github.com/luchinoprince/MixtureIS>, are (452, 208) for *Arrhythmia*, (756, 755) for *Parkinson* and (126, 312) for *Voice*.

For each dataset we ran four MCMC chains for 2×10^3 iterations each, removing the first half as burn-in, leaving us with $S = 4 \times 10^3$ samples from both the posterior and the mixture distributions, which were used to compute the classical, mixture, and PSIS estimators. STAN with defaults setting was used and no convergence or mixing issues were detected with standard diagnostics. Table 3 summarizes the resulting MSE of the estimators relative to the ground truth values, averaging over 10 independent repetitions for each combination of dataset and method.

For the *Arrhythmia* and *Voice* datasets we obtained accurate estimates (which we treat as ground truth values) for $\{\log(p(y_i|y_{-i}))\}_{i=1}^n$ by drawing 5×10^4 samples from each of the n LOO posteriors separately as done in Section 4.2.1. For the *Parkinson* dataset, the above procedure would have been computationally unfeasible and we instead obtained ground truth values for $\{\log(p(y_i|y_{-i}))\}_{i=1}^n$ running a long chain sampling from q_{mix} and then computing the mixture estimators based on 10^6 samples. Standard diagnostics suggested

Dataset	Estimator	MSE (mean over $i = 1, \dots, n$)	MSE (max over $i = 1, \dots, n$)	% of k - shape $> .7$
Arrhythmia n=452 p=208	Mixture	4.4e-03	3.9e-01	-
	Bronze	8.0e-03	1.2e+00	23%
	Posterior	9.3e-03	1.1e+00	25%
	PSIS	6.4e-03	8.7e-01	-
Parkinson n=756 p=755	Mixture	3.6e-03	3.3e-01	-
	Bronze	8.7e-03	1.2e+00	49%
	Posterior	1.0e-02	2.0e+00	53%
	PSIS	6.0e-03	5.0e-01	-
Voice n=126 p=312	Mixture	2.3e-03	6.6e-02	-
	Bronze	2.4e-02	1.1e+00	54%
	Posterior	1.8e-02	9.7e-01	42%
	PSIS	2.0e-02	1.0e+00	-

Table 3: Mean squared errors in estimating $\{\log(p(y_i|y_{-i}))\}_{i=1}^n$ for three high-dimensional binary regression datasets. Mean and quantiles are intended over $i \in \{1, \dots, n\}$ for a single run of each method. See Section 4.3.1 for definition of the bronze estimator.

that the Monte Carlo error for these estimates was at least one order of magnitude smaller than the one of the other estimates under consideration.

In Table 3, the mixture estimator always performs significantly better than both the classical and PSIS estimators, see below for discussion on the bronze estimator also reported in Table 3. Figure 7 displays the evolution of the classical and mixture estimators for the 20 data points with largest absolute value of $\log(p(y_i|y_{-i}))$. Some classical estimators exhibit very large jumps even at high number of iterations, which is a typical pathological behaviour of estimators with infinite or excessively large variance. The mixture estimators, despite having some jumps in a few cases, display a much more stable evolution and convergence.

4.3.1 Comparison to additional alternative computational methodologies

In this section we provide a brief comparison with other alternative methodologies from the Bayesian LOO-CV computation literature, using the three datasets of Table 3. We consider the gold, silver and bronze estimators proposed in [Alqallaf and Gustafson, 2001] and the Sequential Monte Carlo (SMC) approach of [Bornn et al., 2010].

The *bronze* estimator of [Alqallaf and Gustafson, 2001] is the easiest to compare with. In our framework, such methodology estimates $\{p(y_i|y_{-i})\}_{i=1}^n$ performing self-normalized importance sampling with importance distribution given by the following tempered posterior

$$q_B(\theta) \propto \left(\prod_{i=1}^n p(y_i|\theta) \right)^{\frac{n-1}{n}} p(\theta). \quad (17)$$

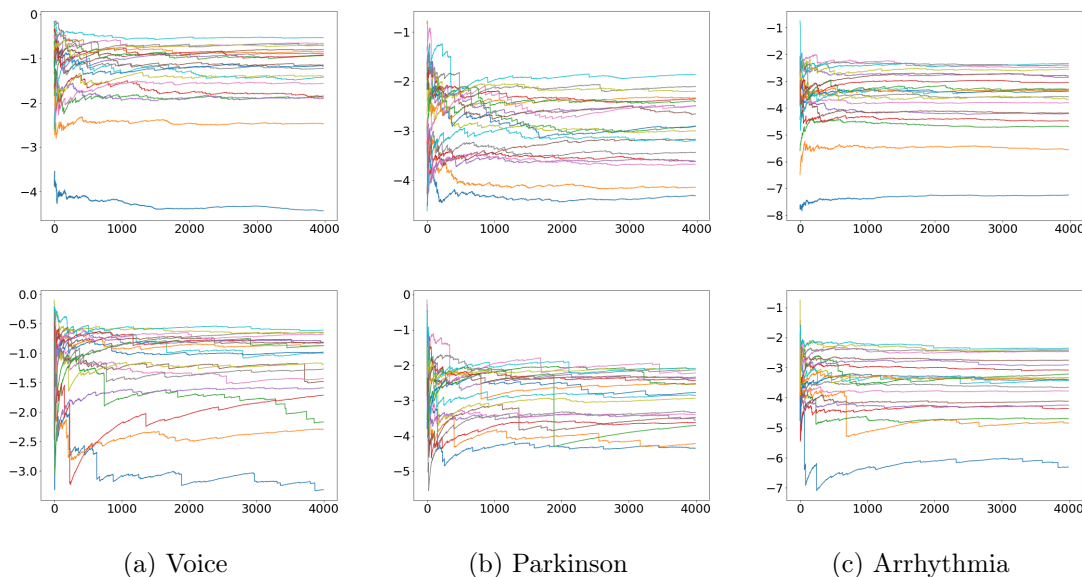


Figure 7: Evolution of the mixture (first row) and classical (second row) estimators, with number of samples on the x -axis, for three datasets (one per column). The traceplots of the estimators corresponding to the 20 data points with largest absolute value of $\log(p(y_i|y_{-i}))$ are displayed.

This procedure has a computational cost comparable to the posterior, mixture and PSIS ones for the same number of samples. We thus test it on the examples in Table 3 using the same number of samples as well as STAN settings. The resulting MSE, which are reported in Table 3, are closer to the ones of the posterior and PSIS estimators rather than the mixture ones.

The SMC methodology of [Bornn et al., 2010], when applied to our context, coincides with running n SMC routines, one for each target value $p(y_i|y_{-i})$, initialized from the same samples drawn from the posterior $p(\theta|y)$. When an adaptive SMC approach is employed, this procedure ends up performing pure importance sampling (with the posterior as importance distribution) for data points inducing well behaved importance weights (e.g. ones with ESS above a given threshold) while performing a genuine SMC routine involving resampling and MCMC moves for the other values. While the resulting estimators are often guaranteed to have finite variance (see Bornn et al., 2010), the total computational cost can be quadratic in n if a considerable proportion of data points requires non-trivial SMC routines. We thus test how many data points require non-trivial SMC routines for the high-dimensional binary regression examples above. The results suggest that approximately 40% for the *Voice* Dataset, 22% for the *Parkinson* dataset and 64% for the *Arrhythmia* dataset. Such percentages were calculated by looking at the effective sample size (ESS) of the weights of the posterior, and assessing how many were under the threshold of $1/2$, which is a default value commonly used in the literature [Chopin and Papaspiliopoulos, 2020]. Such high percentages suggest that SMC, at least in the above

version, is not particularly suited to such a *cross-sectional* estimation procedure, since running $\Theta(n)$ separate SMC routines makes it computationally too demanding, while it can be very appealing in *longitudinal* scenarios, such as hyper-parameter tuning, see e.g. Bornn et al. [2010].

Finally, we consider the *gold* and *silver* of [Alqallaf and Gustafson, 2001]. These allow to obtain only an estimation of the whole LOO-CV sum $\psi = \sum_{i=1}^n \log(p(y_i|y_{-i}))$ in (3), as opposed to the n terms $\{p(y_i|y_{-i})\}_{i=1}^n$. In particular, the *gold* estimator of ψ is defined as

$$\hat{\psi}_{gold} = \frac{n}{K} \sum_{i \in I} \log(p(y_i|y_{-i})), \quad (18)$$

where K is a fixed integer in $\{1, \dots, n\}$ and I is a collection of K indices uniformly sampled without replacement from $\{1, 2, \dots, n\}$. The gold estimator is not computable in practice since we do not know the exact values of $p(y_i|y_{-i})$. A practical approach is given by the so-called silver estimator, which is defined as

$$\hat{\psi}_{silv} = \frac{n}{K} \sum_{i \in I} \log(\hat{\mu}_i^{(loo)}), \quad (19)$$

with K and I defined as for the gold estimator and $\hat{\mu}_i^{(loo)}$ as in (4). We compare the silver estimator with the estimator of ψ obtained from the mixture estimators by plug-in, i.e. $\hat{\psi}_{mix} = \sum_{i=1}^n \log(\hat{\mu}_i^{(mix)})$. To ensure comparability, we fix the total computational resources to 2×10^4 samples (including burn-in ones) both for the silver and mixture estimators. Thus, for a given value of K , each chain used to compute a single $\hat{\mu}_i^{(loo)}$ has a total of $2 \times 10^4/K$ samples. Figure 8 shows the errors in estimating ψ obtained with $\hat{\psi}_{silv}$ for different values of K and with $\hat{\psi}_{mix}$. We can see that, for small values of K , $\hat{\psi}_{silv}$ has a large variance due to the variability in the choice of the subset I . On the contrary, as K increases the bias of each estimator $\hat{\mu}_i^{(loo)}$ increases, since these are self-normalized importance sampling estimators based on $2 \times 10^4/K$ samples, which became too few samples as K increases (in the extreme case of $K = 721$ for the Parkinson data one has $2 \times 10^4/K \approx 28$ samples for every estimator). As a result, regardless of the value of K , $\hat{\psi}_{silv}$ has a much larger estimation error (note the log-scale on the y axis) than $\hat{\psi}_{mix}$ with the same number of total samples. Note that for the *Voice* dataset, given the small values of n and the large number of total samples, the performances of the silver estimator are monotonically increasing with K and the optimal value is $K = n$, which makes the silver estimator coincide with the brute force approach discussed in Section 2.

5 Extensions

The proposed mixture estimator can be extended in various directions.

First, one could extend the mixture estimators to compute LOO-CV criteria for general scoring rules beyond the logarithmic one, see e.g. [Bernardo, 1979, Vehtari and Ojanen, 2012]. In such case one would be interested in LOO-CV estimators of quantities such as $\mathbb{E}_{y_{new} \sim p^*}[S(y_{new}, p(\cdot|y))]$ where S is a scoring rule and $p(\cdot|y)$ is the predictive distribution

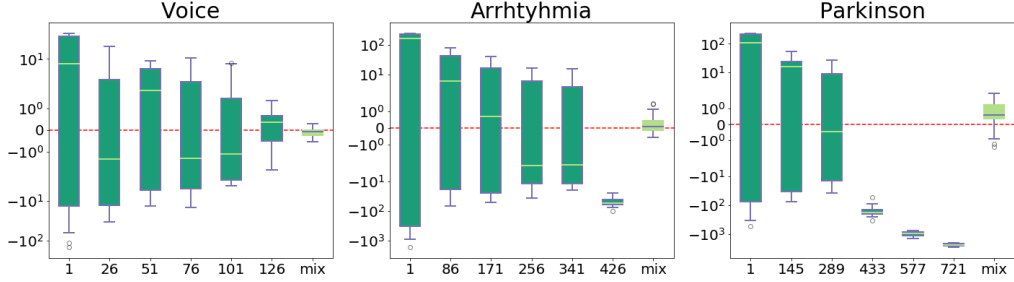


Figure 8: Errors in estimating ψ for $\hat{\psi}_{silv}$, with different values of K on the x -axis and $\hat{\psi}_{mix}$. Boxplots are based on 25 independent repetitions for each estimator.

of y_{new} given the observed data y . The main difference in terms of computational methodology that may arise is the need for another layer of integration if the scoring rule is not *local* [Bernardo, 1979], but instead defined itself as an integral.

Another important extension is to models with non conditionally independent observations, i.e. where the equality in (1) is not satisfied. There, the mixture distribution can be written as

$$q_{mix}(\theta) = Z^{-1} \sum_{i=1}^n p(\theta) p(y_{-i}|\theta) \propto p(\theta|y) \left(\sum_{i=1}^n p(y_i|\theta, y_{-i})^{-1} \right),$$

but $p(y_i|\theta, y_{-i}) \neq p(y_i|\theta)$ in general and thus the last equality in (8) does not hold. One should then replace $p(y_i|\theta)^{-1}$ with $p(y_i|\theta, y_{-i})^{-1}$ throughout for both the posterior and mixture estimators, e.g. in (5), (8) and (9). In such contexts, the mixture estimators remain appealing *provided* one can compute the n predictive likelihood terms $\{p(y_i|\theta, y_{-i})\}_{i=1}^n$ for a given θ at $\Theta(n)$ total computational cost. This will be the case when, after the computation of $p(y|\theta)$, one can compute $p(y_{-i}|\theta)$ for a given i at $\Theta(1)$ additional cost, e.g. using rank-one updates in regression-type models. If instead computing each $p(y_i|\theta, y_{-i})$ term can only be done at $\Theta(n)$ cost separately for each i , then computing $\{p(y_i|\theta, y_{-i})\}_{i=1}^n$ has $\Theta(n^2)$ total cost and both the mixture and posterior estimators are likely to be impractical and no better than the brute force approach discussed at the beginning of Section 2.

Finally, another interesting direction to explore in future work is the extension of the proposed mixture estimator to leave- p -out contexts for $p > 1$. A naive application of the mixture methodology, however, where the mixture is defined as $q_{mix}(\theta) \propto \sum_A p(\theta) p(y_{-A}|\theta)$ where A runs over subsets of $\{1, \dots, n\}$ of size p , would incur a p -choose- n cost per iteration, thus being impractical. Nonetheless, we expect such cost to be avoidable using, for example, appropriate unbiased likelihood estimators in conjunction with pseudo-marginal MCMC algorithms. We leave such extensions to future work.

5.1 Algorithmic variations

As mentioned in Section 3.1, the mixture distribution q_{mix} could be replaced by a more general, weighted version $q_{mix}^{(\alpha)}(\theta) = Z_{\alpha}^{-1} \sum_{i=1}^n \alpha_i p(y_{-i}|\theta) p(\theta)$ with $\alpha = (\alpha_1, \dots, \alpha_n) \in$

$(0, \infty)^n$ being arbitrary weights. In such case Theorem 3.1.1 would still hold, as shown in its proof. In practice, such weighted version directly affects the value of the mixture weight components (π_1, \dots, π_n) that in general satisfy $\pi_i \propto \alpha_i p(y_i|y_{-i})^{-1}$ for $i = \dots, n$, see also Remark 1. Since larger values of π_i tend to lead to estimators of $p(y_i|y_{-i})$ with smaller variance, it follows that increasing π_i corresponds to putting more computational effort in estimating $p(y_i|y_{-i})$ relative to other $p(y_j|y_{-j})$ for $j \neq i$. Thus, having direct control on π_i might be useful to, e.g., design adaptive versions of the algorithm that adapt the weights α on the fly to put more effort on more important or harder to estimate values of $p(y_i|y_{-i})$. In the default version, $\alpha_i = 1$ and $\pi_i \propto p(y_i|y_{-i})^{-1}$ for $i = 1, \dots, n$. As discussed in Remark 2, this is a reasonable default choice that gives more weight to data points y_i with larger values of $|\log p(y_i|y_{-i})|$, which are typically more important (e.g. contribute more to LOO-CV) and harder to estimate. However, $\pi_i \propto p(y_i|y_{-i})^{-1}$ may not be the optimal choice in general, and thus weighted versions $q_{mix}^{(\alpha)}$ might be useful to increase robustness of the proposed estimating procedure to, e.g. overly large values of π_i . In our preliminary exploration, different values of α did not lead to major improvements compared to the default version, which is why we only presented results for that version in this paper. However, we do not exclude that more complex or extreme examples may benefit from tuning of α .

As discussed in Remark 1, the estimators $\{\hat{\mu}_i^{(mix)}\}_{i=1}^n$ effectively estimate the mixture weights $\{\pi_i\}_{i=1}^n$ and the normalizing constant \tilde{Z} and then compute $p(y_i|y_{-i}) = \tilde{Z}^{-1} \pi_i^{-1}$. One might consider more advanced methodologies, e.g. Bridge Sampling [Bennett, 1976, Meng and Wong, 1996], to estimate the normalizing constant \tilde{Z} between $q_{mix}(\theta)$ and $p(\theta|y)$, but we expect this to lead to minimal improvements. In fact, the largest relative variance in all our experiments was given by the estimators of π_i , i.e. the numerators in (9), and thus employing a better estimator of \tilde{Z} would only provide minimal improvements. In this sense, we found that the key and main task required to estimate $\{p(y_i|y_{-i})\}_{i=1}^n$ is estimating the mixture weights $\{\pi_i\}_{i=1}^n$, while estimating \tilde{Z} was significantly easier in all examples we considered.

6 Discussion

We proposed a novel estimator for Bayesian LOO-CV estimator that retains appealing features of classical estimators, such as simplicity of implementation and $\Theta(Sn)$ total cost, while significantly improving robustness to high-dimensionality. We expect our proposed computational methodology to be most useful when the number of parameters is of comparable order, or even larger, than the number of data points. Interestingly, these are regimes where LOO-CV exhibits considerably smaller bias in estimating the ELPD compared to other cross-validation strategies such as k -fold, see e.g. Rad et al. [2020] and references therein. Our work supports the idea that Bayesian LOO-CV computations can be efficiently accomplished with Monte Carlo methods, requiring a computational effort comparable to fitting the model once. This seems to be a computational advantages compared to, e.g., marginal likelihood or Bayes Factors approximation, which is typically a much harder task.

Directions for future research include characterizing how easy or hard it is to sample from the mixture q_{mix} compared to $p(\theta|y)$, which would provide a more complete theoretical picture on the comparison between the efficiency of classical and mixture estimators; and extending the asymptotic analysis of Section 3.2 to cases where both n and p diverge simultaneously.

References

- Fatemah Alqallaf and Paul Gustafson. On cross-validation of bayesian models. *Canadian Journal of Statistics*, 29(2):333–340, 2001.
- Maurice S Bartlett. A comment on D.V. Lindley statistical paradox. *Biometrika*, 44(1-2): 533–534, 1957.
- Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, and Vahid Tarokh. On optimal generalizability in parametric learning. *Advances in Neural Information Processing Systems*, 30, 2017.
- Charles H Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.
- José Bernardo. Expected information as expected utility. 7(3):686–690, 1979. URL www.jstor.org/stable/2958753.
- Luke Bornn, Arnaud Doucet, and Raphael Gottardo. An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1): 47–64, 2010.
- George E. P. Box. Sampling and bayes’ inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383–430, 1980. URL <http://www.jstor.org/stable/2982063>.
- Alan D Chave and David J Thomson. A bounded influence regression estimator based on the statistics of the hat matrix. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(3):307–322, 2003.
- Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer, 2020.
- R. Dennis. Cook and Sanford Weisberg. *Residuals and influence in regression*. Chapman and Hall, New York, 1982.
- Ilenia Epifani, Steven N. MacEachern, and Mario Peruggia. Case-deletion importance sampling estimators: Central limit theorems and related results. *Electron. J. Statist.*, 2: 774–806, 2008. doi: 10.1214/08-EJS259. URL <https://doi.org/10.1214/08-EJS259>.
- Augusto Fasano, Daniele Durante, and Giacomo Zanella. Scalable and accurate variational bayes for high-dimensional binary regression models. *Biometrika*, In press, 2022.

- Edwin Fong and Chris Holmes. On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496, 2020.
- Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.
- Alan E. Gelfand, Dipak K. Dey, and Hong Chang. Model determination using predictive distributions with implementation via sampling-based-methods (with discussion). In *Bayesian Statistics 4*. University Press, 1992.
- Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.
- Charles J Geyer. Estimating normalizing constants and reweighting mixtures in markov chain monte carlo. 1991.
- Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019.
- Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 06 2011. URL <https://doi.org/10.1198/jasa.2011.r10138>.
- Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401, 1999.
- Demetris Lamnisis, Jim E Griffin, and Mark FJ Steel. Cross-validation prior choice in Bayesian probit regression with many covariates. *Statistics and Computing*, 22(2):359–373, 2012.
- Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.
- Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.
- Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.
- Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.
- Michael A Newton and Adrian E Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.
- Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

- Topi Paananen, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):1–19, 2021.
- Mario Peruggia. On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association*, 92(437):199–207, 1997.
- LI Pettit. The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):175–184, 1990.
- Kamiar Rahnema Rad, Arian Maleki, et al. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996, 2020.
- Maxime Rischard, Pierre E Jacob, and Natesh Pillai. Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation. *arXiv preprint arXiv:1810.01382*, 2018.
- Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12):124105, 2008.
- Stan Development Team. RStan: the R interface to Stan, 2020. URL <http://mc-stan.org/>.
- Aki Vehtari and Janne Ojanen. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.
- Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017. URL <https://doi.org/10.1007/s11222-016-9696-4>.
- Aki Vehtari, Jonah Gabry, Mans Magnusson, Yuling Yao, Paul-Christian Bürkner, Topi Paananen, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models, 2020. URL <https://mc-stan.org/loo/>. R package version 2.4.1.
- Esteban Walker and Jeffrey B Birch. Influence measures in ridge regression. *Technometrics*, 30(2):221–227, 1988.
- Robert E Weiss and Meehyung Cho. Bayesian marginal influence assessment. *Journal of statistical planning and inference*, 71(1-2):163–177, 1998.
- Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007, 2018.
- Giacomo Zanella and Gareth Roberts. Scalable importance tempering and bayesian variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):489–517, 2019.

A Implementation details

A.1 Sampling from the proposed mixture with MCMC

For models with conditionally independent data as in (1) the log posterior is typically computed as the sum of log prior and log likelihood contributions as follows

$$\log p(\theta|y) = \log p(\theta) + \sum_{i=1}^n \log p(y_i|\theta) + \text{const}, \quad (20)$$

where *const* denotes terms that do not depend on θ . For $q_{mix}(\theta)$ defined in (8) we have the same expression plus an additional term that can be written as follows to ensure numerical stability

$$\log q_{mix}(\theta) = \log p(\theta) + \sum_{i=1}^n \log p(y_i|\theta) + LSE(\{-\log p(y_i|\theta)\}_{i=1}^n) + \text{const}, \quad (21)$$

where *LSE* denotes the usual *LogSumExp* function defined as $LSE(\mathbf{x}) = \log(\sum_{i=1}^n \exp(x_i))$ for $\mathbf{x} = \{x_i\}_{i=1}^n \in \mathbb{R}^n$. The expression in (21) is trivial to compute whenever the log-prior and log-likelihoods are computable and requires $\Theta(n)$ operations per evaluation, exactly as $\log p(\theta|y)$. In other words, $q_{mix}(\theta)$ can be computed up to normalizing constant whenever the original posterior $p(\theta|y)$ can. We further note that, while computing $\log q_{mix}(\theta)$ in (21) may appear to require roughly twice as many computations as $\log p(\theta|y)$ in (20), as one needs to compute both the sum and the *LSE* quantities, for most models the cost of computing the n likelihood terms $\{\log p(y_i|\theta)\}_{i=1}^n$ dominates the cost of computing their sum or the *LSE* function, e.g. a $\Theta(np)$ cost for the former versus a $\Theta(n)$ cost for the latter for a regression model with n data points and p covariates. Thus in such cases computing $\log q_{mix}(\theta)$ and $\log p(\theta|y)$ have roughly the same cost.

The expression in (21) is also trivial to differentiate, allowing to compute the gradient $\nabla \log q_{mix}(\theta)$, and is amenable to standard probabilistic programming software based automatic differentiation. For example, for the logistic regression model used in Sections 4.2.1 and 4.3 the STAN code to define the posterior $p(\theta|y)$ is given by

```
data {
  int <lower=0> n;
  int <lower=0> k;
  int <lower=0, upper=1> y[n];
  matrix [n,p] X;
  real <lower=0.0> prior_scale;
}
parameters {
  vector[p] beta;
}
model{
  vector[n] means=X*beta;
```

```

    target += double_exponential_lpdf(beta | 0, prior_scale);
    target += bernoulli_logit_lpmf(y | means);
}

```

while to define the mixture distribution $q_{mix}(\theta)$ one should replace the model section with

```

model{
  vector[n] means=X*beta;
  vector[n] log_lik;
  for (index in 1:n){
    log_lik[index]= bernoulli_logit_lpmf(y[index] | means[index]);
  }
  target += double_exponential_lpdf(beta | 0, prior_scale);
  target += sum(log_lik);
  target += log_sum_exp(-log_lik);
}

```

See also the github repository accompanying the paper, <https://github.com/luchinoprince/MixtureIS>, for more details and examples of software implementations.

A.2 Efficient computation of the mixture estimators

Given S samples $\{\theta_s\}_{s=1}^S$ from $q_{mix}(\theta)$, the n estimators $\{\hat{\mu}_i^{(mix)}\}_{i=1}^n$ defined in (9) can be computed at $\Theta(nS)$ total cost in a numerically stable way as follows:

- (i) compute the $n \times S$ matrix of log-likelihood terms $\{\ell_{is}\}_{i,s}$, where $\ell_{is} = \log p(y_i|\theta_s)$ for $i = 1, \dots, n$ and $s = 1, \dots, S$;
- (ii) compute the $n \times S$ matrix of log-weights $\{\tilde{w}_{is}\}_{i,s}$ defined as $\tilde{w}_{is} = \log(w_i^{(mix)}(\theta_s))$, using the equality $\tilde{w}_{is} = -\ell_{is} - \tilde{z}_s$ for $i = 1, \dots, n$ and $s = 1, \dots, S$, where $\tilde{z}_s = LSE(\{-\ell_{is}\}_{i=1}^n)$ for $s = 1, \dots, S$;
- (iii) compute the log-estimators exploiting the equality $\log \hat{\mu}_i^{(mix)} = \tilde{z} - LSE(\{\tilde{w}_{is}\}_{s=1}^S)$ for $i = 1, \dots, n$ where $\tilde{z} = LSE(\{-z_s\}_{s=1}^S)$.

The above operations (i)-(iii) require $\Theta(nS)$ computational cost. In terms of memory requirements, the simplest implementation of the above operations, which creates the $n \times S$ matrices $\{\ell_{is}\}_{i,s}$ and $\{\tilde{w}_{is}\}_{i,s}$, require $\Theta(nS)$ storage, but this can be easily reduced to $\Theta(n)$ storage, if required, by storing only one column at a time.

B Proofs

Proof of Theorem 3.1.1. A standard application of the delta method for the derivation of the relative asymptotic variance of self-normalized importance sampling estimators, see

e.g. Liu [2001, eq.(2.7)] or Owen [2013, eq.(9.8)], applied to $\hat{\mu}_i^{(mix)}$ leads to

$$\begin{aligned} AV_i^{(mix)} &= \lim_{S \rightarrow \infty} S \text{var} \left(\frac{\hat{\mu}_i^{(mix)}}{\mu_i} \right) = \int \left(\frac{p(\theta|y_{-i})}{q_{mix}^{(\alpha)}(\theta)} \right)^2 \left(\frac{p(y_i|\theta)}{\mu_i} - 1 \right)^2 q_{mix}^{(\alpha)}(\theta) d\theta \\ &= \int \frac{p(\theta|y)^2}{q_{mix}^{(\alpha)}(\theta)} d\theta - 2 \int \frac{p(\theta|y)p(\theta|y_{-i})}{q_{mix}^{(\alpha)}(\theta)} d\theta + \int \frac{p(\theta|y_{-i})^2}{q_{mix}^{(\alpha)}(\theta)} d\theta, \end{aligned} \quad (22)$$

where in the last equality we re-arranged terms and used $\mu_i^{-1}p(\theta|y_{-i})p(y_i|\theta) = p(\theta|y)$. Writing $q_{mix}^{(\alpha)}(\theta) = \sum_{j=1}^n \pi_j p(\theta|y_{-j})$ with $\pi_j = Z_{\alpha}^{-1} \alpha_j p(y_{-j})$ and upper bounding the negative terms in (22) by 0, we have

$$AV_i^{(mix)} \leq \int \frac{p(\theta|y)^2}{\sum_{j=1}^n \pi_j p(\theta|y_{-j})} d\theta + \int \frac{p(\theta|y_{-i})^2}{\sum_{j=1}^n \pi_j p(\theta|y_{-j})} d\theta. \quad (23)$$

From $\sum_{j=1}^n \pi_j p(\theta|y_{-j}) \geq \pi_i p(\theta|y_{-i})$ it follows

$$\int \frac{p(\theta|y_{-i})^2}{\sum_{j=1}^n \pi_j p(\theta|y_{-j})} d\theta \leq \int \frac{p(\theta|y_{-i})^2}{\pi_i p(\theta|y_{-i})} d\theta = \pi_i^{-1}$$

and

$$\int \frac{p(\theta|y)^2}{\sum_{j=1}^n \pi_j p(\theta|y_{-j})} d\theta \leq \pi_i^{-1} \int \frac{p(\theta|y)^2}{p(\theta|y_{-i})} d\theta = \pi_i^{-1} p(y_i|y_{-i})^{-1} \int p(y_i|\theta) p(\theta|y) d\theta,$$

where in the last equality we also used $p(\theta|y_{-i})^{-1}p(\theta|y) = p(y_i|y_{-i})^{-1}p(y_i|\theta)$. Combining the above with (23) we obtain

$$AV_i^{(mix)} \leq \pi_i^{-1} \left(1 + p(y_i|y_{-i})^{-1} \int p(y_i|\theta) p(\theta|y) d\theta \right). \quad (24)$$

The latter upper bound is finite by (A1) and the fact that $\alpha_i > 0$ implies $\pi_i > 0$. \square

B.1 Proof of Theorem 3.2.1

Lemma B.1.1. *For any $h \in (0, 1)$, the matrix $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]$ is singular if and only if $H_{ii} = h$, with H as in (12). If $H_{ii} \neq h$ then*

$$\begin{aligned} [X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]^{-1} &= \\ (X^T X)^{-1} + \frac{h^{-1}}{1 - h^{-1} H_{ii}} \cdot (X^T X + \sigma^2 \Sigma^{-1})^{-1} x_i x_i^T (X^T X + \sigma^2 \Sigma^{-1})^{-1}. \end{aligned} \quad (25)$$

Proof. Assume first that $H_{ii} = h$, then $x_i \neq 0$ (the zero vector has leverage zero). Multiplying $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]$ by the non-zero vector $(X^T X + \sigma^2 \Sigma^{-1})^{-1} x_i$ yields $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T](X^T X + \sigma^2 \Sigma^{-1})^{-1} x_i = x_i + h^{-1} H_{ii} x_i = 0$. Hence we have proved that in this case $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]$ is singular. We now verify that (25)

is the inverse of $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]$ when $H_{ii} \neq h$. Multiplying the two matrices we get:

$$\begin{aligned} \mathbb{I} + \frac{h^{-1}}{1 - h^{-1} H_{ii}} x_i x_i^T (X^T X + \sigma^2 \cdot \Sigma^{-1})^{-1} - h^{-1} x_i x_i^T (X^T X + \sigma^2 \cdot \Sigma^{-1})^{-1} \\ + \frac{h^{-2} H_{ii}}{1 - h^{-1} H_{ii}} x_i x_i^T (X^T X + \sigma^2 \cdot \Sigma^{-1})^{-1} \\ = \mathbb{I} + x_i x_i^T (X^T X + \sigma^2 \cdot \Sigma^{-1})^{-1} \left(\frac{h^{-1}}{1 - h^{-1} H_{ii}} + h^{-1} - \frac{h^{-2} H_{ii}}{1 - h^{-1} H_{ii}} \right) = \mathbb{I}. \end{aligned}$$

□

Lemma B.1.2. *Let Σ be a positive definite $p \times p$ matrix, X a $n \times p$ matrix, $\sigma > 0$ and $M = X^T X - h^{-1} x_i x_i^T + \sigma^2 \Sigma^{-1}$ with $i \in \{1, \dots, n\}$ and $h \in (0, 1)$. Then M is positive definite if and only if $H_{ii} < h$, with H as in (12).*

Proof. Assume that $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]$ is positive definite. If $x_i = 0$, then $H_{ii} = 0 < h$. If $x_i \neq 0$, then $(X^T X + \sigma^2 \Sigma^{-1})^{-1} x_i$ is a non-zero vector and we must have, by positive definiteness, $0 < x_i^T (X^T X + \sigma^2 \Sigma^{-1})^{-1} [X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T] (X^T X + \sigma^2 \Sigma^{-1})^{-1} x_i = H_{ii}(1 - h^{-1} H_{ii})$. This implies that $H_{ii} < h$.

Conversely, suppose that $H_{ii} < h$. Then $h^{-1}/(1 - h^{-1} H_{ii}) > 0$, and (25) shows that $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]^{-1}$ can be written as the sum of a positive definite matrix and a positive semi-definite one. As such, it is positive definite, and $[X^T X + \sigma^2 \cdot \Sigma^{-1} - h^{-1} x_i x_i^T]$ must be positive definite as well. □

Proof of Theorem 3.2.1. The relative asymptotic variance of the classical estimator $\hat{\mu}_i^{(post)}$ can be derived in analogous way to the derivation in (22), with the importance distribution $q_{mix}^{(\alpha)}(\theta)$ replaced by the posterior $p(\theta|y)$. After simplifications, this leads to

$$AV_i^{(post)} = \int \left(\frac{p(\theta|y-i)}{p(\theta|y)} \right)^2 p(\theta|y) d\theta - 1. \quad (26)$$

By (26) and (11) we have

$$AV_i^{(post)} + 1 = c_1 \int \exp \left\{ \frac{2(y_i - x_i^T \theta)^2}{2\sigma^2} - \frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2} - \frac{(\theta - \theta_0)^T \Sigma^{-1} (\theta - \theta_0)}{2} \right\} d\theta, \quad (27)$$

where c_1 is a constant independent of θ . Grouping together quadratic and linear terms in θ we obtain

$$AV_i^{(post)} = c_2 \int \exp \left\{ -\theta^T M \theta + \theta^T v \right\} d\theta, \quad M = (2\sigma^2)^{-1} [X^T X - 2x_i x_i^T + \sigma^2 \Sigma^{-1}] \quad (28)$$

where v is a p -dimensional vector and c_2 is a non-zero scalar, and both are independent of θ . It follows that $AV_i^{(post)}$ is finite if and only if M is positive definite. The thesis follows by Lemma B.1.2 with $h = 0.5$. □

B.2 Proof of Proposition 3.2.2 and Theorems 3.2.3, 3.2.4 and 3.2.5

Proof of Proposition 3.2.2. Denoting $\lambda_p = \nu_p^{-2}\sigma^2$ and applying Woodbury matrix identity we have

$$\begin{aligned} H &= X(X^T X + \lambda_p \mathbb{I}_p)^{-1} X^T, \\ &= X(\lambda_p^{-1} \mathbb{I}_p - \lambda_p^{-1} X^T (\mathbb{I}_n + \lambda_p^{-1} X X^T)^{-1} \lambda_p^{-1} X) X^T \\ &= \lambda_p^{-1} X X^T - \lambda_p^{-1} X X^T (\mathbb{I}_n + \lambda_p^{-1} X X^T)^{-1} \lambda_p^{-1} X X^T. \end{aligned}$$

Since $\lim_{p \rightarrow \infty} p \lambda_p^{-1} = \sigma^{-2} c$, by Kolmogorov's criterion of SLLN and the random design assumption (A2) on X , we have $p^{-1} X X^T \rightarrow \tau^2 \mathbb{I}_n$ almost surely element-wise as $p \rightarrow \infty$, see e.g. Fasano et al. [2022, Lemma 1] for a more detailed proof of the latter statement. It follows that $\lambda_p^{-1} X X^T \rightarrow \frac{c\tau^2}{\sigma^2} \mathbb{I}_n$ almost surely element-wise as $p \rightarrow \infty$ and H converges in the same way to

$$\frac{c\tau^2}{\sigma^2} \mathbb{I}_n - \frac{c\tau^2}{\sigma^2} \mathbb{I}_n (\mathbb{I}_n + \frac{c\tau^2}{\sigma^2} \mathbb{I}_n)^{-1} \frac{c\tau^2}{\sigma^2} \mathbb{I}_n = \frac{c\tau^2}{\sigma^2 + c\tau^2} \mathbb{I}_n,$$

which implies the desired convergence of H_{ii} . The statement about $AV_i^{(post)}$ follows by combining the above result with Theorem 3.2.1. \square

Proof of Theorem 3.2.3. The statement follows from part (a) of Theorem 3.2.4, since (11) is a special case of (14). \square

Proof of Theorem 3.2.4. First we prove part (a). Using $\pi_i^{-1} = \left(\sum_{j=1}^n p(y_j | y_{-j})^{-1} \right) p(y_i | y_{-i})$ we can re-write the upper bound in (24) as

$$AV_i^{(mix)} \leq \left(\sum_{j=1}^n p(y_j | y_{-j})^{-1} \right) \left(p(y_i | y_{-i}) + \int p(y_i | \theta) p(\theta | y) d\theta \right). \quad (29)$$

By the subadditivity and submultiplicativity of \limsup , and monotonicity of $t \mapsto t^{-1}$ on $(0, \infty)$, it follows

$$\limsup_{p \rightarrow \infty} AV_i^{(mix)} \leq \left(\sum_{j=1}^n (\liminf_{p \rightarrow \infty} p(y_j | y_{-j}))^{-1} \right) \left(\limsup_{p \rightarrow \infty} p(y_i | y_{-i}) + \limsup_{p \rightarrow \infty} \int p(y_i | \theta) p(\theta | y) d\theta \right). \quad (30)$$

We now prove that all terms on the right-hand side are finite. We have $p(y_i | y_{-i}) = p(y) / p(y_{-i})$ where, by (14),

$$p(y) = \int \prod_{j=1}^n g(y_j | \eta_j) p(\eta) d\eta \quad \text{and} \quad p(y_{-i}) = \int \prod_{j \neq i}^n g(y_j | \eta_j) p(\eta) d\eta$$

where $p(\eta) = N(\eta; 0, A_p)$ with $A_p = \nu_p^2 X X^T$ is the prior distribution on $\eta = (\eta_1, \dots, \eta_n)$ induced by the prior on $(\theta_1, \dots, \theta_p)$ and the linear transformation $\eta = X\theta$. As shown

in the proof of Proposition 3.2.2, we have $p^{-1}XX^T \rightarrow \tau^2\mathbb{I}_n$ almost surely element-wise as $p \rightarrow \infty$, and thus also $A_p = \nu_p^2 XX^T = p\nu_p^2(p^{-1}XX^T) \rightarrow c\tau^2\mathbb{I}_n$, which implies that $p(\eta) \rightarrow N(\eta; 0, c\tau^2\mathbb{I}_n)$ almost surely as $p \rightarrow \infty$, where the convergence is point-wise in $\eta \in \mathbb{R}^n$. Also, since $A_p \rightarrow c\tau^2\mathbb{I}_n$ as $p \rightarrow \infty$, we have that, almost surely for large enough p , A_p is invertible, its determinant satisfies $|A_p| > (c\tau^2/2)^n$ and $(A_p^{-1} - (2c\tau^2)^{-1}\mathbb{I}_n)$ is positive definite. These observations imply that, almost surely, for large enough p allow we have

$$p(\eta) < (\pi c\tau^2)^{-n/2} \exp(-(4c\tau^2)^{-1}\|\eta\|^2),$$

for every $\eta \in \mathbb{R}^n$. Combining the above bound with the boundedness of the likelihood, we can apply the dominated convergence theorem and deduce that

$$p(y) \rightarrow \int_{\mathbb{R}^n} \prod_{j=1}^n g(y_j|\eta_j) N(\eta_j; 0, c\tau^2\mathbb{I}_n) d\eta = \prod_{j=1}^n \int_{\mathbb{R}} g(y_j|\eta_j) N(\eta_j; 0, c\tau^2) d\eta_j \in (0, \infty)$$

almost surely as $p \rightarrow \infty$. Applying the same argument to $p(y_{-i})$ we obtain

$$p(y_i|y_{-i}) = \frac{p(y)}{p(y_{-i})} \rightarrow \frac{\prod_{j=1}^n \int_{\mathbb{R}} g(y_j|\eta_j) N(\eta_j; 0, c\tau^2) d\eta_j}{\prod_{j \neq i} \int_{\mathbb{R}} g(y_j|\eta_j) N(\eta_j; 0, c\tau^2) d\eta_j} = \int_{\mathbb{R}} g(y_i|\eta_i) N(\eta_i; 0, c\tau^2) d\eta_i \in (0, \infty), \quad (31)$$

meaning that $\limsup_{p \rightarrow \infty} p(y_i|y_{-i}) < \infty$ and $(\liminf_{p \rightarrow \infty} p(y_i|y_{-i}))^{-1} < \infty$ almost surely. By the same argument we also have $(\liminf_{p \rightarrow \infty} p(y_j|y_{-j}))^{-1} < \infty$ for $j = 1, \dots, n$. Finally, by (14) and Bayes Theorem, we can write

$$\int p(y_i|\theta) p(\theta|y) d\theta = \frac{\int p(y_i|\theta) p(y|\theta) p(\theta) d\theta}{\int p(y|\theta) p(\theta) d\theta} = \frac{\int_{\mathbb{R}^n} g(y_i|\eta_i)^2 \prod_{j \neq i} g(y_j|\eta_j) p(\eta) d\eta}{\int_{\mathbb{R}^n} \prod_{j=1}^n g(y_j|\eta_j) p(\eta) d\eta}$$

and applying dominated convergence arguments analogous to above we obtain

$$\int p(y_i|\theta) p(\theta|y) d\theta \rightarrow \frac{\int_{\mathbb{R}} g(y_i|\eta_i)^2 N(\eta_i; 0, c) d\eta_i}{\int_{\mathbb{R}} g(y_i|\eta_i) N(\eta_i; 0, c) d\eta_i} \in (0, \infty),$$

which implies that $\limsup_{p \rightarrow \infty} \int p(y_i|\theta) p(\theta|y) d\theta$. Combining the above bounds with (30) we deduce $\limsup_{p \rightarrow \infty} AV_i^{(mix)} < \infty$.

Consider now part (b) and assume $\int_{\mathbb{R}} g(y_i|\eta_i)^{-1} \exp(-\delta\eta_i^2) d\eta_i < \infty$ for some $\delta < (2c\tau^2)^{-1}$. By (26)

$$AV_i^{(post)} + 1 = \int \left(\frac{p(\theta|y_{-i})}{p(\theta|y)} \right)^2 p(\theta|y) d\theta = \frac{p(y_i|y_{-i})}{p(y_{-i})} \int \frac{p(y_{-i}|\theta)^2}{p(y|\theta)} p(\theta) d\theta.$$

By (31) we have $\lim_{p \rightarrow \infty} \frac{p(y_i|y_{-i})}{p(y_{-i})} = a$ for some $a \in (0, \infty)$. Thus

$$\limsup_{p \rightarrow \infty} AV_i^{(post)} + 1 = a \limsup_{p \rightarrow \infty} \int \frac{p(y_{-i}|\theta)^2}{p(y|\theta)} p(\theta) d\theta.$$

By (14)

$$\int \frac{p(y_{-i}|\theta)^2}{p(y|\theta)} p(\theta) d\theta = \int_{\mathbb{R}^n} \frac{\prod_{j \neq i} g(y_j|\eta_j)}{g(y_i|\eta_i)} p(\eta) d\eta \leq \left(\prod_{j \neq i} \sup_{\eta_j} g(y_j|\eta_j) \right) \int_{\mathbb{R}} g(y_i|\eta_i)^{-1} p(\eta_i) d\eta_i$$

with $p(\eta) = N(\eta; 0, A_p)$ as above and $p(\eta_i) = N(\eta_i; 0, a_p^{(i)})$ where $a_p^{(i)}$ is the i -th diagonal term of A_p . By $A_p \rightarrow c\tau^2 \mathbb{I}_n$ almost surely, we have $a_p^{(i)} \rightarrow c\tau^2$ and thus $(2a_p^{(i)})^{-1} > \delta$ eventually as $p \rightarrow \infty$ since $\delta < (2c\tau^2)^{-1}$. It follows

$$\begin{aligned} \limsup_{p \rightarrow \infty} \int_{\mathbb{R}} g(y_i|\eta_i)^{-1} p(\eta_i) d\eta_i &= (2\pi c\tau^2)^{-1/2} \limsup_{p \rightarrow \infty} \int_{\mathbb{R}} g(y_i|\eta_i)^{-1} \exp(-(2a_p^{(i)})^{-1} \eta_i^2) d\eta_i \\ &\leq (2\pi c\tau^2)^{-1/2} \limsup_{p \rightarrow \infty} \int_{\mathbb{R}} g(y_i|\eta_i)^{-1} \exp(-\delta \eta_i^2) d\eta_i < \infty. \end{aligned}$$

Combining the above inequalities we obtain $\limsup_{p \rightarrow \infty} AV_i^{(post)} < \infty$ as desired.

Finally, consider part (b) and assume $\int_{\mathbb{R}} g(y_i|\eta_i)^{-1} \exp(-\delta \eta_i^2) d\eta_i = \infty$ for some $\delta > (2c\tau^2)^{-1}$. In this case, using that $A_p \rightarrow c\tau^2 \mathbb{I}_n$ as $p \rightarrow \infty$ we have

$$\begin{aligned} \limsup_{p \rightarrow \infty} \frac{\prod_{j \neq i} g(y_j|\eta_j)}{g(y_i|\eta_i)} p(\eta) d\eta &= (2\pi c\tau^2)^{-n/2} \limsup_{p \rightarrow \infty} \int_{\mathbb{R}^n} \frac{\prod_{j \neq i} g(y_j|\eta_j)}{g(y_i|\eta_i)} \exp(-\eta^T A_p \eta) d\eta \\ &\geq (2\pi c\tau^2)^{-n/2} \int_{\mathbb{R}^n} \frac{\prod_{j \neq i} g(y_j|\eta_j)}{g(y_i|\eta_i)} \exp(-\delta \|\eta\|^2) d\eta \\ &= (2\pi c\tau^2)^{-n/2} \left(\prod_{j \neq i} \int_{\mathbb{R}} g(y_j|\eta_j) \exp(-\delta \eta_j^2) d\eta_j \right) \int_{\mathbb{R}} g(y_i|\eta_i)^{-1} \exp(-\delta \eta_i^2) d\eta_i = \infty, \end{aligned}$$

where we used the fact that $\delta \mathbb{I}_n - A_p$ is eventually positive definite as $p \rightarrow \infty$ since $\delta > (2c\tau^2)^{-1}$. \square

Proof of Theorem 3.2.5. Part (a). We first show that $AV_i^{(loo)}$ diverges as $p \rightarrow \infty$. A derivation analogous to (22), with the importance distribution $q_{mix}^{(\alpha)}(\theta)$ replaced by the LOO posterior $p(\theta|y_{-i})$ and some simple algebraic simplifications, leads to

$$AV_i^{(loo)} = \int \frac{p(\theta|y)^2}{p(\theta|y_{-i})} d\theta - 1 = \frac{p(y_{-i})}{p(y)^2} \int p(y_i|\theta) p(y|\theta) p(\theta) d\theta - 1,$$

where we also used the conditional independence assumption $p(y|\theta) = \prod_{i=1}^n p(y_i|\theta)$. Combining the above with (14) we have

$$AV_i^{(loo)} + 1 = \frac{\left(\int \prod_{j \neq i} g(y_j|\eta_j) p(\eta_{-i}) d\eta_{-i} \right) \left(\int h_i(\eta) p(\eta) d\eta \right)}{\left(\int \prod_{j=1}^n g(y_j|\eta_j) p(\eta) d\eta \right)^2}, \quad (32)$$

where $p(\eta_{-i})$ and $p(\eta)$ denote the prior distributions of η_{-i} and η under (14), and $h_i(\eta) = g(y_i|\eta_i) \prod_{j=1}^n g(y_j|\eta_j)$. Since $p(\eta) = N(\eta; 0, A_p)$ with $A_p = \nu_p^2 XX^T$ and $p(\eta_{-i}) = N(\eta_{-i}; 0, A_p^{(i)})$ with $A_p^{(i)} = \nu_p^2 X_{-i} X_{-i}^T$, we can rewrite $AV_i^{(loo)} + 1$ as

$$\sqrt{\frac{2\pi p\nu_p^2 \left| \frac{1}{p} XX^T \right|}{\left| \frac{1}{p} X_{-i} X_{-i}^T \right|}} \frac{\left(\int \prod_{j \neq i} g(y_j|\eta_j) K_{-i}(\eta_{-i}) d\eta_{-i} \right) \left(\int h_i(\eta) K(\eta) d\eta \right)}{\left(\int \prod_{j=1}^n g(y_j|\eta_j) K(\eta) d\eta \right)^2}, \quad (33)$$

where $K(\eta) = \exp(-\eta^T (2\nu_p^2 XX^T)^{-1} \eta)$ and $K_{-i}(\eta_{-i}) = \exp(-\eta_{-i}^T (2\nu_p^2 X_{-i} X_{-i}^T)^{-1} \eta_{-i})$.

We now analyze the limiting behaviour of each term in (33). First we have $\lim_{p \rightarrow \infty} \frac{\left| \frac{1}{p} XX^T \right|}{\left| \frac{1}{p} X_{-i} X_{-i}^T \right|} = \tau^2$ since $p^{-1} XX^T \rightarrow \tau^2 \mathbb{I}_n$ and $p^{-1} X_{-i} X_{-i}^T \rightarrow \tau^2 \mathbb{I}_{n-1}$ almost surely, as shown in the proof of Proposition 3.2.2, and the determinant is a continuous function. Note that the latter convergences also imply that XX^T and $X_{-i} X_{-i}^T$ are almost surely eventually invertible as $p \rightarrow \infty$ so that K and K_{-i} are well defined. Then $K(\eta) \leq 1$ implies

$$\int \prod_{j=1}^n g(y_j|\eta_j) K(\eta) d\eta \leq I_y < \infty,$$

where $I_y = \int \prod_{j=1}^n g(y_j|\eta_j) d\eta = \prod_{j=1}^n \int g(y_j|\eta_j) d\eta_j$ is a positive and finite constant by the assumptions in part (a). Also, $K(\eta) = \exp\left(-\frac{1}{2\nu_p^2} \eta^T (p^{-1} XX^T)^{-1} \eta\right)$ combined with $p^{-1} XX^T \rightarrow \tau^2 \mathbb{I}_n$ and $p\nu_p^2 \rightarrow \infty$ implies that $K(\eta) \rightarrow \exp(0) = 1$ for every $\eta \in \mathbb{R}^n$ almost surely as $p \rightarrow \infty$, and similarly also $K_{-i}(\eta_{-i}) \rightarrow \exp(0) = 1$ for every $\eta_{-i} \in \mathbb{R}^{n-1}$. It follows by Fatou's lemma that

$$\liminf_{p \rightarrow \infty} \int \prod_{j \neq i} g(y_j|\eta_j) K_{-i}(\eta_{-i}) d\eta_{-i} \geq I_{y_{-i}} \quad \text{and} \quad \liminf_{p \rightarrow \infty} \int h_i(\eta) K(\eta) d\eta_{-i} \geq I_{\tilde{y}},$$

where $I_{\tilde{y}} = \int h_i(\eta) d\eta$ and $I_{y_{-i}} = \prod_{j \neq i} \int g(y_j|\eta_j) d\eta_j$ are positive and finite constants by the assumptions in part (a). Combining the above results with (33), the submultiplicativity of the \liminf and $p\nu_p^2 \rightarrow \infty$, we get

$$\liminf_{p \rightarrow \infty} AV_i^{(loo)} + 1 \geq \sqrt{2\pi\tau^2} \frac{I_{y_{-i}} I_{\tilde{y}}}{I_y^2} \liminf_{p \rightarrow \infty} \sqrt{p\nu_p^2} = \infty,$$

as desired.

We now prove that also $AV_i^{(mix)}$ diverges as $p \rightarrow \infty$ under the assumptions of part (a). By (22) and $\frac{p(\theta|y_{-i})}{q_{mix}(\theta)} \leq \pi_i^{-1}$ we have

$$AV_i^{(mix)} \geq \int \frac{p(\theta|y)^2}{q_{mix}(\theta)} d\theta - 2 \int \frac{p(\theta|y_{-i})}{q_{mix}(\theta)} p(\theta|y) d\theta \geq \int \frac{p(\theta|y)^2}{q_{mix}(\theta)} d\theta - 2\pi_i^{-1},$$

which implies

$$\liminf_{p \rightarrow \infty} AV_i^{(mix)} \geq \liminf_{p \rightarrow \infty} \int \frac{p(\theta|y)^2}{q_{mix}(\theta)} d\theta - \frac{2}{\liminf_{p \rightarrow \infty} \pi_i}.$$

We now prove that $\int \frac{p(\theta|y)^2}{q_{mix}(\theta)} d\theta$ diverges with p and that $\liminf_{p \rightarrow \infty} \pi_i > 0$ for every i , thus deducing $\lim_{p \rightarrow \infty} AV_i^{(mix)} = \infty$ from the inequality above. First, by (14) we have

$$\int \frac{p(\theta|y)^2}{q_{mix}(\theta)} d\theta = \frac{\sum_{j=1}^n p(y_{-j})}{p(y)^2} \int \frac{\prod_{i=1}^n g(y_i|\eta_i)^2}{\sum_{k=1}^n \prod_{i \neq k} g(y_i|\eta_i)} p(\eta) d\eta = \sum_{j=1}^n \frac{p(y_{-j}) \int h(\eta) p(\eta) d\eta}{p(y)^2}$$

with $h(\eta) = (\sum_{k=1}^n g(y_k|\eta_k)^{-1})^{-1} \prod_{i=1}^n g(y_i|\eta_i)$. Then, using $p(y_{-j}) = \int \prod_{i \neq j} g(y_i|\eta_i) p(\eta_{-j}) d\eta_{-j}$ with $p(\eta_{-j}) = (2\pi\nu_p^2 |X_{-j} X_{-j}^T|)^{-(n-1)/2} K_{-j}(\eta_{-j})$ as defined above, we have

$$\int \frac{p(\theta|y)^2}{q_{mix}(\theta)} d\theta = \sum_{j=1}^n \sqrt{2\pi\nu_p^2 \frac{|\frac{1}{p} X X^T|}{|\frac{1}{p} X_{-j} X_{-j}^T|}} \frac{\left(\int \prod_{k \neq j} g(y_k|\eta_k) K_{-j}(\eta_{-j}) d\eta_{-j} \right) \left(\int h(\eta) K(\eta) d\eta \right)}{\left(\int \prod_{j=1}^n g(y_j|\eta_j) K(\eta) d\eta \right)^2}.$$

Proceeding as done above for $AV_i^{(loo)} + 1$, exploiting the almost sure point-wise convergences $K(\eta) \rightarrow 1$ and $K_{-j}(\eta_{-j}) \rightarrow 1$, one can derive

$$\liminf_{p \rightarrow \infty} \int \frac{p(\theta|y)^2}{q_{mix}(\theta)} d\theta \geq \sum_{j=1}^n \sqrt{2\pi\tau^2} \frac{I_{y_{-j}} I_{mix}}{I_y^2} \liminf_{p \rightarrow \infty} \sqrt{p\nu_p^2} = \infty,$$

where $I_{mix} = \int h(\eta) p(\eta) d\eta$ is a positive constant.

We now prove that $\liminf_{p \rightarrow \infty} \pi_i > 0$ for every i . From $\pi_i = \frac{p(y_{-i})}{\sum_{j=1}^n p(y_{-i})} = (1 + \sum_{j \neq i} \frac{p(y_{-j})}{p(y_{-i})})^{-1}$ it follows that

$$\liminf_{p \rightarrow \infty} \pi_i = \left(1 + \limsup_{p \rightarrow \infty} \sum_{j \neq i} \frac{p(y_{-j})}{p(y_{-i})} \right)^{-1} \geq \left(1 + \sum_{j \neq i} \limsup_{p \rightarrow \infty} \frac{p(y_{-j})}{p(y_{-i})} \right)^{-1}.$$

Then we write for every $j \neq i$

$$\frac{p(y_{-j})}{p(y_{-i})} = \left(\frac{|p^{-1} X_{-i} X_{-i}^T|}{|p^{-1} X_{-j} X_{-j}^T|} \right)^{1/2} \frac{\int \prod_{k \neq j} g(y_k|\eta_k) K_{-j}(\eta_{-j}) d\eta_{-j}}{\int \prod_{k \neq i} g(y_k|\eta_k) K_{-i}(\eta_{-i}) d\eta_{-i}},$$

which, using $p^{-1} X_{-i} X_{-i}^T \rightarrow \tau^2 \mathbb{I}_{n-1}$, $p^{-1} X_{-j} X_{-j}^T \rightarrow \tau^2 \mathbb{I}_{n-1}$, $K_{-j}(\eta_{-j}) \leq 1$ and $K_{-i}(\eta_{-i}) \rightarrow 1$, similarly to before, implies that $\limsup_{p \rightarrow \infty} \frac{p(y_{-j})}{p(y_{-i})} \leq \frac{I_{y_{-j}}}{I_{y_{-i}}} < \infty$.

Part (b). We start by proving $\limsup_{p \rightarrow \infty} AV_i^{(loo)} < \infty$. By (32) we can deduce

$$\limsup_{p \rightarrow \infty} AV_i^{(loo)} + 1 \leq B^2 \left(\liminf_{p \rightarrow \infty} \int \prod_{j=1}^n g(y_j|\eta_j) p(\eta) d\eta \right)^{-2} \quad (34)$$

where $B = \sup_{\eta \in \mathbb{R}^n} \prod_{i=1}^n p(y_i|\eta_i)$ is a finite constant by the assumption of upper bounded likelihood. Since $p(\eta) = N(\eta; 0, \nu_p^2 X X^T) \rightarrow 0$ almost surely for every $\eta \in \mathbb{R}^n$ as $p \rightarrow \infty$,

it is convenient to define the change of variables $\gamma = (p\nu_p^2)^{-1/2}\eta$ and re-write the integral above as

$$\int_{\mathbb{R}^n} \prod_{j=1}^n g(y_j|\eta_j) N(\eta; 0, \nu_p^2 X X^T) d\eta = \int_{\mathbb{R}^n} \prod_{j=1}^n g(y_j|\sqrt{p\nu_p^2}\gamma_j) N(\gamma; 0, p^{-1} X X^T) d\gamma. \quad (35)$$

Defining $a_i = \lim_{\eta_i \rightarrow -\infty} g(y_i|\eta_i)$ and $b_i = \lim_{\eta_i \rightarrow \infty} g(y_i|\eta_i)$, we have $\lim_{p \rightarrow \infty} g(y_j|\sqrt{p\nu_p^2}\gamma_j) = (a_i(1 - \text{sgn}(\gamma_i)) + b_i \text{sgn}(\gamma_i))$ for every i and every $\gamma_i \neq 0 \in \mathbb{R}$ and $\lim_{p \rightarrow \infty} N(\gamma; 0, p^{-1} X X^T) = N(\gamma; 0, \tau^2 \mathbb{I}_n)$ for every $\gamma \in \mathbb{R}^n$ almost surely as $p \rightarrow \infty$. Thus, by Fatou's lemma we have

$$\begin{aligned} \liminf_{p \rightarrow \infty} \int_{\mathbb{R}^n} \prod_{j=1}^n g(y_j|\sqrt{p\nu_p^2}\gamma_j) N(\gamma; 0, p^{-1} X X^T) d\gamma &\geq \\ \int_{\mathbb{R}^n} \prod_{j=1}^n (a_i(1 - \text{sgn}(\gamma_i)) + b_i \text{sgn}(\gamma_i)) N(\gamma; 0, \tau^2 \mathbb{I}_n) d\gamma &= \prod_{j=1}^n \left(\frac{a_i}{2} + \frac{b_i}{2} \right) > 0. \end{aligned} \quad (36)$$

The latter product is a positive constant by the assumption $a_i + b_i > 0$ for any i . Combining (36) and (34) we obtain $\limsup_{p \rightarrow \infty} AV_i^{(loo)} < \infty$ as desired.

We now prove $\limsup_{p \rightarrow \infty} AV_i^{(mix)} < \infty$. Equation (36) states that $\liminf_{p \rightarrow \infty} p(y) > 0$. An analogous derivation can be used to prove that $\liminf_{p \rightarrow \infty} p(y_{-j}) > 0$ for every $j = 1, \dots, n$. Combining the latter with $\limsup_{p \rightarrow \infty} p(y_{-j}) \leq B_{-j} < \infty$ for every $j = 1, \dots, n$, with $B_{-j} = \sup_{\eta \in \mathbb{R}^n} \prod_{i \neq j} p(y_i|\eta_i) < \infty$, we obtain that $\liminf_{p \rightarrow \infty} \pi_i \geq \frac{\liminf_{p \rightarrow \infty} p(y_{-i})}{\sum_{j=1}^n \limsup_{p \rightarrow \infty} p(y_{-i})} > 0$. One can then deduce

$$\limsup_{p \rightarrow \infty} AV_i^{(mix)} < (\limsup_{p \rightarrow \infty} \pi_i^{-1}) (\limsup_{p \rightarrow \infty} AV_i^{(loo)}) < \infty$$

as desired.

To conclude, we prove $\lim_{p \rightarrow \infty} AV_i^{(post)} = \infty$. By (14) and (26)

$$AV_i^{(post)} + 1 = \frac{p(y)}{p(y_{-i})^2} \int_{\mathbb{R}^n} \frac{\prod_{k \neq i} g(y_k|\eta_k)}{g(y_i|\eta_i)} p(\eta) d\eta.$$

Thus

$$\liminf_{p \rightarrow \infty} AV_i^{(post)} + 1 \geq \frac{\liminf_{p \rightarrow \infty} p(y)}{B_{-i}^2} \liminf_{p \rightarrow \infty} \int_{\mathbb{R}^n} \frac{\prod_{k \neq i} g(y_k|\eta_k)}{g(y_i|\eta_i)} p(\eta) d\eta,$$

where $B_{-i} < \infty$ and $\liminf_{p \rightarrow \infty} p(y) > 0$ as shown above. Using the same change of variable of (35) and proceeding as in (36) we obtain

$$\liminf_{p \rightarrow \infty} \int_{\mathbb{R}^n} \frac{\prod_{k \neq i} g(y_k|\eta_k)}{g(y_i|\eta_i)} p(\eta) d\eta \geq \left(\frac{1}{2a_i} + \frac{1}{2b_i} \right) \prod_{j \neq i} \left(\frac{a_j}{2} + \frac{b_j}{2} \right) = \infty$$

where the latter equality follows from the assumptions that $a_i b_i = 0$ and $(a_i + b_i) \in (0, \infty)$. It follows that $\liminf_{p \rightarrow \infty} AV_i^{(post)} = \infty$ almost surely, and thus also $\lim_{p \rightarrow \infty} AV_i^{(post)} = \infty$ almost surely as desired. \square