Approximating Cross-validatory Predictive Evaluation in Bayesian Latent Variables Models with Integrated IS and WAIC

Longhai Li[*‡], Shi Qiu[*], Bei Zhang[*], and Cindy X. Feng[†]

17 April 2015

Abstract: Looking at predictive accuracy is a traditional method for comparing models. A natural method for approximating out-of-sample predictive accuracy is leave-one-out cross-validation (LOOCV) — we alternately hold out each case from a full data set and then train a Bayesian model using Markov chain Monte Carlo (MCMC) without the held-out case; at last we evaluate the posterior predictive distribution of all cases with their actual observations. However, actual LOOCV is time-consuming. This paper introduces two methods, namely iIS and iWAIC, for approximating LOOCV with only Markov chain samples simulated from a posterior based on a *full* data set. iIS and iWAIC aim at improving the approximations given by importance sampling (IS) and WAIC in Bayesian models with possibly correlated latent variables. In iIS and iWAIC, we first integrate the predictive density over the distribution of the latent variables associated with the held-out without reference to its observation, then apply IS and WAIC approximations to the integrated predictive density. We compare iIS and iWAIC with other approximation methods in three kinds of models: finite mixture models, models with correlated spatial effects, and a random effect logistic regression model. Our empirical results show that iIS and iWAIC give substantially better approximates than non-integrated IS and WAIC and other methods.

Key phrases: MCMC, cross-validation, posterior predictive check, predictive model assessment, DIC, WAIC, Bayesian latent variable models

[*] Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Rd, Saskatoon, SK, S7N5E6, Canada. E-mails: longhai@math.usask.ca, shq471@mail.usask.ca, bez733@mail.usask.ca.

[†] School of Public Health and Western College of Veterinary Medicine, University of Saskatchewan, Health Sciences Building, 107 Wiggins Road, Saskatoon, SK, S7N 5E5 Canada. Email: cindy.feng@usask.ca.

[‡] Correspondance author.

1 Introduction

Evaluating goodness of fit of models to a data set is of fundamental importance in statistics. The goodness-of-fit evaluation is necessary for many tasks, such as comparing competing models (which may be non-nested), testing hypotheses, and detecting outliers in a data set. To date, evaluating model goodness-of-fit remains a daunting problem for Bayesian statisticians. There have been a wide range of methods for this problem in addition to the classic significance test for parameters used to link a family of nested models. In particular, the Bayes factor (Kass and Raftery, 1995) based on marginal likelihood is widely used for comparing multiple Bayesian models. However, it is notorious that the marginal likelihood function can be arbitrarily small if the prior is sufficiently diffuse — a problem called the Jeffrey-Lindley paradox (Lindley, 1957; Robert, 2013) – therefore the Bayes factor cannot be used in models with uninformative or improper priors. Much research has been done to remedy this problem, for example the fractional Bayes factor (O'Hagan, 1995, 1997), the intrinsic Bayes factor (Berger and Pericchi, 1996), as well as other methods treating model selection as a decision problem with continuous loss functions, see Bernardo and Rueda (2002); Li and Yu (2012); Li et al. (2014), and the references therein. In addition, computing marginal likelihood is tremendously difficult for complex models, see discussion in Chib (1995); Raftery et al. (2006), and the references therein. Another traditional approach, often referred to as predictive model assessment, looks at the accuracy of competing models in predicting out-of-sample observations; this method is free of the Jeffrey-Lindley paradox. An extensive review of predictive model assessment methods is provided by Vehtari and Ojanen (2012).

Cross-validation (CV) is a natural way to approximate the out-of-sample predictive performance of a model. Throughout this paper, we will discuss only leave-one-out cross-validation (LOOCV); hence in what follows, CV means LOOCV. In CV, we hold out a unit from a full data set, fit/train a model using Markov chain Monte Carlo (MCMC) without the holdout, and then find a predictive distribution of what would be observed from the holdout. We repeat this procedure with each observation as a holdout. We can then compare the CV predictive distributions with the actual observations in terms of a chosen loss function. A widely used loss function is negative twice log predictive density of the actual observation. Predictive evaluations based on this loss are often called *information criteria* (IC) for historical reasons (Gelman et al., 2014). CV predictive evaluation can also be used to check whether the actual observation is an outlier by looking at tail probability of the predictive distribution (Marshall and Spiegelhalter, 2003, 2007). Actual Bayesian CV is time-consuming for complex models because it requires an MCMC simulation for each held-out unit. Alternative methods have been proposed to approximate out-of-sample or CV predictive evaluations using only MCMC samples drawn from the posterior based on the full data set. These methods aim to correct for optimistic bias in training (also called within-sample) predictive evaluation. Gelfand et al. (1992) introduce importance sampling (IS) method that weights MCMC samples using inverse training predictive density for each unit; this method is widely applicable to many loss functions. IS has been innovatively applied to many problems, such as off-policy reinforcement learning problems (Hachiya et al., 2008) and "inverse problems" (Bhattacharya and Haslett, 2007). However, many applications show that IS approximations have large biases and variance (Peruggia, 1997; Vehtari, 2001; Vehtari and Lampinen, 2002; Epifani et al., 2008).

There are also many other methods that focus on estimating out-of-sample information criterion by adjusting a version of the training predictive information criterion with a correction for optimistic bias (Spiegelhalter et al., 2002; Ando, 2007; Plummer, 2008; Gelman et al., 2014). In recent years, the deviance information criterion (DIC) of Spiegelhalter et al. (2002), which is readily available in WinBUGS, may be the most popular choice in Bayesian applications. However, a number of difficulties have been noted with DIC (and its variants), particularly in Bayesian models in which latent variables and model parameters are nonidentifiable from data — a typical example is mixture models; see Celeux et al. (2006), and

Plummer (2008), and many of the discussions following the paper by Spiegelhalter et al. (2002). Some authors have pointed out connections and discrepancies of DIC with out-of-sample information criterion [see Plummer (2008); Watanabe (2010a); Gelman et al. (2014)]. However, nowadays we often need to compare models with latent variables. DIC is typically implemented by treating latent variables as unknown parameters (otherwise DIC will be too hard to implement); however, this treatment lacks theoretical justification – for a detailed discussion see Li et al. (2012). Recently, Watanabe (2009, 2010b,c) proposes a newer criterion called WAIC (widely applicable information criterion), which has been evaluated in several simple models by Gelman et al. (2014). WAIC operates on predictive probability density of observed variables rather than on model parameters, hence it can be applied in singular statistical models (*i.e.* models with non-identifiable parameterization). Watanabe (2010a) has proved that WAIC is asymptotically equivalent to the CV information criterion as random variables of training data, and that on average of both training and evaluation (future) data, WAIC and CV information criterion are both asymptotically equivalent to the out-of-sample information criterion (Watanabe, 2009). However, WAIC is only justified for problems where the observed data is independently distributed with a population distribution.

In this article, we introduce two predictive evaluation methods based on IS and WAIC for use in Bayesian models with unit-specific and possibly correlated latent variables. IS and WAIC can be simply applied to the (non-integrated) predictive density of observed variables, which is conditional not only on the model parameters, but also on latent variables associated with a validation unit that is supposed to be left out in CV. However, the actual observations on the validation unit used in the full data posterior often bring more bias into the latent variables associated with the validation unit (perhaps more than into the model parameters) than IS or WAIC correction alone can eliminate. To eliminate the bias in the latent variables associated with the validation unit, one remedy is to temporarily discard the latent variables in the full data posterior sample and integrate the non-integrated predictive density with respect to the conditional distribution of the latent variables associated with the validation unit that is conditional only on the model parameters, *not on the actual observations*. This process will lead to an *integrated* predictive density. Using the same process we also obtain an integrated evaluation function. We then apply IS and WAIC formulae to the integrated predictive density and evaluation functions, which results in two predictive evaluation methods — Integrated Importance Sampling (iIS) and Integrated WAIC (iWAIC). The required integrals can be obtained analytically in some models, otherwise it can often be easily approximated using Monte Carlo methods or other numerical methods.

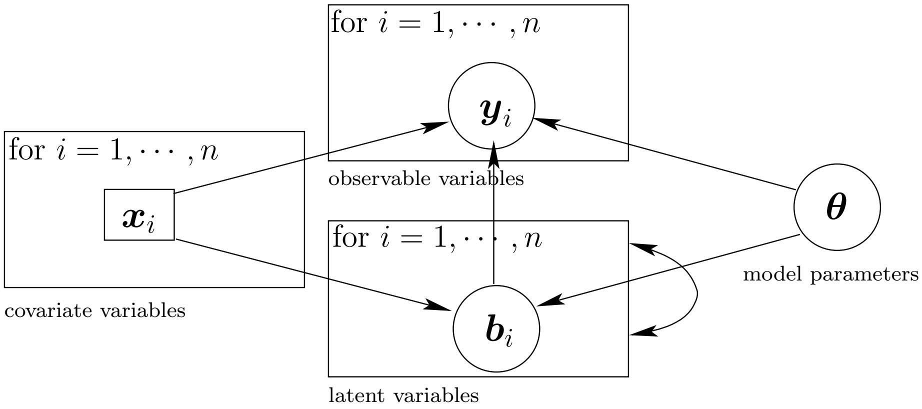
According to the software manual for the Matlab toolbox `GPstuff` (Vanhatalo et al., 2012, 2013), the method of iIS has been used for computing information criterion of Gaussian process latent variable models. For computing the information criterion, one uses only the integrated predictive density (see equation (21)) for which `GPstuff` uses the analytical method for Gaussian likelihoods and the numerical approximation method for non-Gaussian likelihoods. However, to the best of our knowledge, iIS has not been previously described formally with theoretical justification. This article describes iIS formally for general latent variable models as well as for general evaluation functions. For example, we will show that iIS can also be used for computing CV posterior p-values. This article also theoretically proves the equivalence of iIS and actual CV evaluation. The main contribution of this article is to use illustrative examples to demonstrate the necessity and benefit of integrating away the latent variables associated with the validation unit. For computing CV posterior p-values, iIS is also related to the ghosting method which was proposed by Marshall and Spiegelhalter (2007) and also discussed by Held et al. (2010). The ghosting method discards latent variables associated with the validation unit and re-generates them from the distribution without reference to the actual observations of the validation unit for obtaining Monte Carlo estimates of tail probabilities. However, the ghosting method does not use importance re-weighting to correct for the bias in model parameters; hence, the ghosting method can be deemed as a partial implementation of iIS.

The remainder of this article will be organized as follows. In Section 1, we describe a class of Bayesian models with unit-specific models that iIS and iWAIC can be applied to. In Section 2, we describe how to perform actual cross-validation evaluation, and give relevant posterior distributions. We will then describe iIS and iWAIC in general terms in Sections 4 and 5, respectively. In Section 6, we compare iIS and iWAIC to other approximation methods in three simple examples — a mixture modelling problem, a problem using random effect logistic models, and a disease mapping problem that uses spatially correlated random effects. Our empirical results show that iIS and iWAIC provide significantly closer approximates to actual CV evaluation results than ordinary IS and WAIC, as well as other methods. The article will be concluded in Section 7. In Appendices, we give a sketch of the working procedures of iIS and iWAIC.

2 Bayesian Models with Unit-specific Latent Variables

The new predictive evaluation methods that we will describe hereis for use in Bayesian models with unit-specific latent variables. Throughout this paper, we use bold-faced letters to denote vectors and matrices. Suppose we have n observations $\boldsymbol{y}_1^{\text{obs}}, \cdots, \boldsymbol{y}_n^{\text{obs}}$ on n observation units (*a.k.a.* cases, such as persons, locations, time points, or a combination of them). These observations are modelled as a realization of random variables $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$. In many problems, we introduce a latent variable (often a random vector, sometimes called random effects or missing data) $\boldsymbol{b}_i$ for each unit i from which $\boldsymbol{y}_i^{\text{obs}}$ is observed; we then model $\boldsymbol{y}_i$ and $\boldsymbol{b}_i$ with certain statistical distributions parametrized by $\boldsymbol{\theta}$. Conditional on $\boldsymbol{b}_i$ and $\boldsymbol{\theta}$ (often also on a covariate variable $\boldsymbol{x}_i$ that will be omitted in following equations for simplicity), we assume that $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ are statistically independent with probability density $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$, which we will call the **non-integrated predictive density** in this article. If we assume independence between $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n$ given $\boldsymbol{\theta}$, then the marginalized distributions of random variables $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ given $\boldsymbol{\theta}$ are also independent for each i, as is the case in mixture models. For modelling spatial and time series data, we often assume that the latent variables

Figure 1: Graphical representation of Bayesian latent variables models. The double arrows in the box for $\boldsymbol{b}_{1:n}$ mean possible dependency between $\boldsymbol{b}_{1:n}$. Note that the covariate $\boldsymbol{x}_i$ will be omitted in the conditions of densities for $\boldsymbol{b}_i$ and $\boldsymbol{y}_i$ throughout this paper for simplicity.



$\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n$ are dependent for modelling correlations between locations or time points (see an example in Section 6.3). In the following general discussion, we will assume that $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ are correlated. Figure 1 gives a graphical representation of the models described here.

Throughout this paper, we will use notation $\boldsymbol{a}_{1:n}$ to denote the collection of all $\boldsymbol{a}_j$: $\{\boldsymbol{a}_j | j = 1, \ldots, n\}$, and use $\boldsymbol{a}_{-i}$ to denote the collection of all $\boldsymbol{a}_j$ except $\boldsymbol{a}_i$: $\{\boldsymbol{a}_j | j = 1, \ldots, n, j \neq i\}$.

Suppose conditional on $\boldsymbol{\theta}$, we have specified a density for $\boldsymbol{y}_i$ given $\boldsymbol{b}_i$: $P(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\theta})$, a joint prior density for latent variables $\boldsymbol{b}_{1:n}$: $P(\boldsymbol{b}_{1:n} | \boldsymbol{\theta})$, and a prior density for $\boldsymbol{\theta}$: $P(\boldsymbol{\theta})$. The posterior of $(\boldsymbol{b}_{1:n}, \boldsymbol{\theta})$ given observations $\boldsymbol{y}_{1:n}^{\text{obs}}$ is proportional to the joint density of $\boldsymbol{y}_{1:n}^{\text{obs}}$, $\boldsymbol{b}_{1:n}$, and $\boldsymbol{\theta}$:

$$P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n} | \boldsymbol{y}_{1:n}^{\text{obs}}) = \prod_{j=1}^{n} P(\boldsymbol{y}_j^{\text{obs}} | \boldsymbol{b}_j, \boldsymbol{\theta}) P(\boldsymbol{b}_{1:n} | \boldsymbol{\theta}) P(\boldsymbol{\theta}) / C_1, \tag{1}$$

where C_1 is the normalizing constant involving only with $\boldsymbol{y}_{1:n}^{\text{obs}}$.

3 Actual Cross-validatory Predictive Evaluation

To do cross-validation, for each $i = 1, \ldots, n$, we omit observation $\boldsymbol{y}_i^{\text{obs}}$, and then draw MCMC samples from the **CV posterior distribution** of model parameters and latent variables

$P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\mathrm{obs}})$:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\mathrm{obs}}) = \prod_{j \neq i} P(\boldsymbol{y}_j^{\mathrm{obs}}|\boldsymbol{b}_j, \boldsymbol{\theta}) P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \tag{2}$$

where C_2 is the normalizing constant involving only with $\boldsymbol{y}_{-i}^{\mathrm{obs}}$. Note that, in equation (2), we assume that the structural information (*e.g.* spatial relationships between n locations) among $\boldsymbol{b}_{1:n}$ is not lost, rather only that the value of $\boldsymbol{y}_i^{\mathrm{obs}}$ is omitted. After we draw MCMC samples of $(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})$ from (2) and then drop $\boldsymbol{b}_i$, we obtain an MCMC sample of $(\boldsymbol{\theta}, \boldsymbol{b}_{-i})$ from the marginalized CV posterior $P(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\mathrm{obs}})$:

$$P_{\text{post(-i), M}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\mathrm{obs}}) = \prod_{j \neq i} P(\boldsymbol{y}_j^{\mathrm{obs}}|\boldsymbol{b}_j, \boldsymbol{\theta}) P(\boldsymbol{b}_{-i}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \tag{3}$$

where $P(\boldsymbol{b}_{-i}|\boldsymbol{\theta})$ is the marginalized prior density for $\boldsymbol{b}_{-i}$ induced from the specified joint prior for $\boldsymbol{b}_{1:n}$, *i.e.* $P(\boldsymbol{b}_{-i}|\boldsymbol{\theta}) = \int P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta}) d\boldsymbol{b}_i$. Using the conditional prior $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta}) = P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta})/P(\boldsymbol{b}_{-i}|\boldsymbol{\theta})$, we can write

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\mathrm{obs}}) = P_{\text{post(-i), M}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\mathrm{obs}}) P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta}). \tag{4}$$

From the above expression, we see that sampling from $P_{\text{post(-i)}}$ is equivalent to sampling from $P_{\text{post(-i), M}}$ and then generating $\boldsymbol{b}_i$ from the conditional prior $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta})$. Therefore, this method for performing cross-validation makes use of the assumed structure in $\boldsymbol{b}_{1:n}$ (such as neighbouring relationships between spatial units, see the example presented in Section 6.3) through $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta})$ in predicting $\boldsymbol{y}_i$ given $\boldsymbol{y}_{-i}^{\mathrm{obs}}$. This treatment indeed regards the structural information in $\boldsymbol{b}_{1:n}$ as a known fixed covariate. We feel that this treatment is reasonable because we are interested in comparing competing models for the conditional distribution of $\boldsymbol{y}_{1:n}$ given the structure between the n units, rather than the distribution of the structure itself. This is similar to how the cross-validation is done in linear models, for which we assume that the values of the covariates (explanatory variables) of the test case are known when we make prediction of the response of the test case.

The purpose of performing CV is to evaluate certain compatibility (or discrepancy) between the posterior $P(\boldsymbol{y}_i|\boldsymbol{y}_{-i}^{\text{obs}})$ and the actual observation $\boldsymbol{y}_i^{\text{obs}}$. We will specify an evaluation function $a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)$ that measures certain goodness-of-fit (or discrepancy) of the distribution $P(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{b}_i)$ to the actual observation $\boldsymbol{y}_i^{\text{obs}}$. The **CV posterior predictive evaluation** is defined as the expectation of the $a(\boldsymbol{y}_{1:n}^{\text{obs}}, ., .)$ with respect to $P_{\text{post(-i)}}$:

$$E_{\text{post(-i)}}(a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)) = \int a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}}) d\boldsymbol{\theta} d\boldsymbol{b}_{1:n}. \tag{5}$$

The expectation in (5) can be approximated by averaging $a(\boldsymbol{y}_i^{\text{obs}}, \cdot, \cdot)$ over MCMC samples of $(\boldsymbol{\theta}, \boldsymbol{b}_i)$ drawn from $P_{\text{post(-i)}}$.

The first example of a is the value of predictive density function $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$ at the actual observation $\boldsymbol{y}_i^{\text{obs}}$:

$$a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) = P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i). \tag{6}$$

The expectation of (6) with respect to $P_{\text{post(-i)}}$ is referred to here as the **CV posterior predictive density** $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$. The **CV information criterion** (CVIC) is defined as minus two times the sum of the CV posterior predictive densities over all validation units:

$$\text{CVIC} = -2\sum_{i=1}^{n} \log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})). \tag{7}$$

A smaller value of the CVIC indicates a better fit of a Bayesian model to a real data set. The second example of a in (5) is the p-value given the model parameters and latent variables for unit i (Marshall and Spiegelhalter, 2003, 2007):

$$a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) = Pr(\boldsymbol{y}_i > \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i) + 0.5Pr(\boldsymbol{y}_i = \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i), \tag{8}$$

where Pr means probability of a set, as we have used P as density; also $\boldsymbol{y}_i$ should be a scalar for such situations. The expectation of (8) with respect to $P_{\text{post(-i)}}$ gives the **CV posterior p-value**:

$$\text{CV posterior p-value } (\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = Pr(\boldsymbol{y}_i > \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) + 0.5Pr(\boldsymbol{y}_i = \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}), \tag{9}$$

which is a tail probability of the CV posterior predictive distribution with density $P(\boldsymbol{y}_i|\boldsymbol{y}_{-i}^{\text{obs}})$. The purpose of computing the CV posterior p-value is to check the discrepancy of the observation $\boldsymbol{y}_i^{\text{obs}}$ to the CV posterior predictive distribution of $\boldsymbol{y}_i$ that is conditional on other observations (*i.e.* $\boldsymbol{y}_{-i}^{\text{obs}}$). Both very large and very small values of posterior p-value indicate that $\boldsymbol{y}_i^{\text{obs}}$ may be an outlier (unusually small or large) compared to other observations.

Actual CV requires n Markov chain simulations (each of which may use multiple parallel chains), one for each validation unit. This is very time consuming, especially when the model is complex and n is fairly large. Therefore, we are interested in approximating the expectations in (5) for all validation units $i = 1, \ldots, n$ with samples of $(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})$ obtained from a single MCMC simulation based on the full data set; that is, with samples drawn from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ (referred to as the **full data posterior** hereafter). However, we cannot simply treat samples from the full data posterior as CV posteriors, because the inclusion of $\boldsymbol{y}_i^{\text{obs}}$ has introduced optimistic bias in validating $\boldsymbol{y}_i^{\text{obs}}$. This optimistic bias means that the "posterior predictive distribution" of $\boldsymbol{y}_i$ formed by averaging $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$ with respect to $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ fits $\boldsymbol{y}_i^{\text{obs}}$ better than the actual CV posterior predictive distribution of $\boldsymbol{y}_i$ that averages $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$ with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$. Therefore, we need to correct for this optimistic bias using a certain method to obtain an unbiased approximate/estimate of the actual CV posterior predictive evaluation. We will introduce two approximating methods in Sections 4 and 5, respectively.

4 Importance Sampling (IS) Approximation

4.1 Non-integrated Importance Sampling

Importance weighting (Gelfand et al., 1992) is a natural choice for approximating CV prediction evaluation based on the posterior given the full data set. For general and detailed discussions of importance sampling techniques, one can refer to Geweke (1989); Neal (1993); Gelman and Meng (1998); Liu (2001). If our samples are from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ but we are interested in estimating the mean of a with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ as in (5), the importance weighting method is based on the following equality for CV expected evaluation:

$$E_{\text{post(-i)}}(a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)) = \frac{E_{\text{post}}\left[a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})\right]}{E_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})\right]}, \tag{10}$$

where $E_{\text{post}}[\,]$ is expectation with respect to $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n} | \boldsymbol{y}_{1:n}^{\text{obs}})$, and

$$W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}) = \frac{P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n} | \boldsymbol{y}_{-i}^{\text{obs}})}{P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n} | \boldsymbol{y}_{1:n}^{\text{obs}})} \times \frac{C_2}{C_1} = \frac{1}{P(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{\theta}, \boldsymbol{b}_i)}. \tag{11}$$

Note that we can multiply any constant to the above importance weight since they will be canceled in the fraction of (10); also we use the superscript $^{\text{nIS}}$ denote application of importance sampling (shortened by **nIS**) to the **non-integrated predictive density**, in contrast to iIS to be given in next section. In words, importance sampling estimates the expected evaluation by finding Monte Carlo estimates of the two means in the fraction of (10) with only MCMC samples from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n} | \boldsymbol{y}_{1:n}^{\text{obs}})$. We can apply equation (10) to estimate means of any evaluation function a with respect to the CV posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{b}_i)$.

Particularly, in computing CVIC, the evaluation function $a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) = P(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{\theta}, \boldsymbol{b}_i)$ is the same as $1/W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})$ in equation (11). Therefore, the numerator of (10) is just 1 when applied to compute CVIC. Therefore, the CV posterior predictive density $P(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{y}_{-i}^{\text{obs}})$ is equal to the harmonic mean of the non-integrated predictive density $P(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{\theta}, \boldsymbol{b}_i)$ with respect to $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n} | \boldsymbol{y}_{1:n}^{\text{obs}})$:

$$P(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{y}_{-i}^{\text{obs}}) = \frac{1}{E_{\text{post}}\left[1/P(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{\theta}, \boldsymbol{b}_i)\right]}. \tag{12}$$

Based on the equality in (12), **nIS** estimates the CV posterior predictive density by:

$$\hat{P}^{\text{nIS}}(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{y}_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post}}\left[1/P(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{\theta}, \boldsymbol{b}_i)\right]}. \tag{13}$$

The corresponding nIS estimate of CVIC using (13) is $-2\sum_{i=1}^{n} \log(\hat{P}^{\text{nIS}}(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{y}_{-i}^{\text{obs}})$. Note that if there are no latent variables used in the model, there will be no $\boldsymbol{b}_i$ in (12) and (13).

4.2 Integrated Importance Sampling

In theory, the nIS estimate (10) is valid for almost all Bayesian models with latent variables as long as the integral itself exists and the supports of $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ and $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ are the same. However, in simulating MCMC from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$, the latent variable $\boldsymbol{b}_i$ is largely confined to regions that fit the observation $\boldsymbol{y}_i^{\text{obs}}$ well. Therefore, the distribution of $\boldsymbol{b}_i$ marginalized from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ may be highly biased to regions that fit the observation $\boldsymbol{y}_i^{\text{obs}}$ well, compared to the distribution of $\boldsymbol{b}_i$ marginalized from $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$, which can cover a much larger area. Although the supports of $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ and $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ are the same in theory, the effective support of $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ may be much smaller than that of $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$. We will illustrate this in the mixture model example with Figure 3; this results in the inaccuracy of nIS.

To improve nIS, we can re-generate $\boldsymbol{b}_i$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta})$ with the observation $\boldsymbol{y}_i^{\text{obs}}$ removed as the actual cross-validation simulation does; see equation (4). The formal formulation of such a re-generation procedure is given as follows. First we note that using equation (4), we can rewrite the expectation in (5) as

$$E_{\text{post(-i)}}(a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)) = E_{\text{post(-i), M}}(A(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_{-i})) \tag{14}$$

$$= \int\int A(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_{-i}) P(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}}) d\boldsymbol{\theta} d\boldsymbol{b}_{-i}, \tag{15}$$

where

$$A(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_{-i}) = \int a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta}) d\boldsymbol{b}_i. \tag{16}$$

We refer to (16) as an **integrated evaluation function**.

We will also discard $\boldsymbol{b}_i$ temporarily for validation unit i in MCMC samples from the full data posterior $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$. The marginalized full data posterior of $(\boldsymbol{\theta}, \boldsymbol{b}_{-i})$ is

$$P_{\text{post, M}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{1:n}^{\text{obs}}) = \prod_{j \neq i} P(\boldsymbol{y}_j^{\text{obs}}|\boldsymbol{b}_j, \boldsymbol{\theta}) P(\boldsymbol{b}_{-i}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \times \int P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{b}_i, \boldsymbol{\theta}) P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta}) d\boldsymbol{b}_i / C_1. \tag{17}$$

We refer to the second factor in (17) as an **integrated predictive density** because it integrates away $\boldsymbol{b}_i$ without reference to $\boldsymbol{y}_i^{\text{obs}}$. For ease in reference, it is explicitly given below:

$$P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i}) = \int P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{b}_i, \boldsymbol{\theta}) P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta}) d\boldsymbol{b}_i. \tag{18}$$

Using the standard importance weighting method, we will estimate (15) by

$$E_{\text{post(-i), M}}(A(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_{-i})) = \frac{E_{\text{post, M}}\left[A(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_{-i})\ W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i})\right]}{E_{\text{post, M}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i})\right]}, \tag{19}$$

where W_i^{iIS} is the integrated importance weight:

$$W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}) = \frac{P_{\text{post(-i), M}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}})}{P_{\text{post, M}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{1:n}^{\text{obs}})} \times \frac{C_2}{C_1} = \frac{1}{P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})}. \tag{20}$$

In particular, for estimating CVIC, $A \times W_i^{\text{iIS}} = 1$. Therefore, the iIS estimate of the CV posterior predictive density based on equality (19) is given by:

$$\hat{P}^{\text{iIS}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post, M}}\left[1/P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})\right]}. \tag{21}$$

Accordingly, the iIS estimate of CVIC using (21) is $-2\sum_{i=1}^{n} \log(\hat{P}^{\text{iIS}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}))$. The only difference from the nIS estimate in (13) is the replacement of the non-integrated predictive density $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)$ with the integrated predictive density $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})$. Note that we can also write the expectation $E_{\text{post, M}}(\ )$ in equations (19) and (21) as $E_{\text{post}}(\ )$, because we still find Monte Carlo estimates with samples of $(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})$ from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$, but without using $\boldsymbol{b}_i$.

The integration over $\boldsymbol{b}_i$ in equations (16) and (18) is the essential difference of iIS to nIS; for using iIS, we need to find their values. In some problems these values can be approximated with finite summation or calculated analytically. Otherwise, we will re-generate $\boldsymbol{b}_i$ given $(\boldsymbol{b}_{-i}, \boldsymbol{\theta})$ with no reference to $\boldsymbol{y}_i^{\text{obs}}$, which is often easy. Note that this re-generation needs to be done for each $i = 1, \ldots, n$. Sometimes, much computation can be shared by these n re-generating processes since they are all conditional on $\boldsymbol{\theta}$; see the example in Section 6.3.

5 WAIC Approximations

In this section, we describe a generalized WAIC method, iWAIC, for approximating the CV predictive density in Bayesian models with correlated latent variables.

We will first describe WAIC for models with no latent variables (or models after we integrate away latent variables that are independent for units given parameters). In such models, observed variables $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independently distributed with a probability distribution $P(\boldsymbol{y}|\boldsymbol{\theta})$ conditional on model parameters $\boldsymbol{\theta}$. After we obtain MCMC samples for $\boldsymbol{\theta}$ given observations $\boldsymbol{y}_1^{\text{obs}}, \ldots, \boldsymbol{y}_n^{\text{obs}}$, a version of WAIC (Watanabe, 2009, 2010b,c) is given by:

$$\text{WAIC} = -2\sum_{i=1}^{n} \left[\log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}))) \right], \tag{22}$$

where E_{post} and V_{post} stand for mean and variance over $\boldsymbol{\theta}$ with respect to $P(\boldsymbol{\theta}|\boldsymbol{y}_1^{\text{obs}}, \ldots, \boldsymbol{y}_n^{\text{obs}})$. By comparing the forms of WAIC and CVIC (7), we can think of the CV posterior predictive density in WAIC as being approximated by:

$$\hat{P}^{\text{WAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \exp\left\{ \log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}))) \right\}. \tag{23}$$

In words, WAIC corrects for the bias in the mean of the training predictive density of $\boldsymbol{y}_i^{\text{obs}}$ by dividing the exponential of the variance of the log predictive density of $\boldsymbol{y}_i^{\text{obs}}$ with respect to the posterior of $\boldsymbol{\theta}$ given the full data set. Watanabe (2010a) has proven that WAIC is asymptotically equivalent to CVIC when the observed variables are independently distributed and conditional on $\boldsymbol{\theta}$. Watanabe has also shown the asymptotic equivalence of the Taylor expansions of (23) and the harmonic mean (13) (without $\boldsymbol{b}_i$). From our research, we do see that (23) provides results very close to the CV posterior predictive density of each $\boldsymbol{y}_i^{\text{obs}}$. This way to look at WAIC also provides an approach to assess the statistical significance of differences of WAICs of different models by looking at differences in the means of log CV posterior predictive densities, which was advocated by Vehtari and Lampinen (2002) for CVIC itself.

For the models given in Section 2 with possibly correlated latent variables, a naive way to approximate the CVIC is to apply WAIC directly to the non-integrated predictive density of $\boldsymbol{y}_i^{\text{obs}}$ conditional on $\boldsymbol{\theta}$ and $\boldsymbol{b}_i$:

$$\hat{P}^{\text{nWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \exp\left\{\log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)))\right\}. \quad (24)$$

We will refer to (24) as the non-integrated WAIC (or nWAIC for short) method for approximating CV posterior predictive density. The corresponding information criterion based on (24) is:

$$\text{nWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{nWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})). \quad (25)$$

This way to apply WAIC indeed treats latent variables as model parameters. nWAIC is not justified by the theory for WAIC; however, practitioners may likely apply WAIC to Bayesian models with latent variable this way for the sake of convenience.

Our research (to be presented next) will show that nWAIC cannot correct for the bias in unit-specific latent variables entirely. We propose to apply the WAIC approximation to the integrated predictive density (18) in order to estimate the CV posterior predictive density:

$$\hat{P}^{\text{iWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \exp\left\{\log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i}))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})))\right\}. \quad (26)$$

Accordingly, iWAIC can be used to approximate CVIC by :

$$\text{iWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{iWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})). \quad (27)$$

In Section 4, we have theoretically shown the equivalence of iIS to CV predictive evaluation for models with correlated latent variables, which holds as long as the support of the full data posterior is not a subset of the CV posterior. However, we haven't proven any sort of equivalences of $\hat{P}^{\text{iWAIC}}$ and $\hat{P}^{\text{nWAIC}}$ to CVIC. The derivations of formulae for nWAIC and iWAIC for models with correlated latent variables are only heuristic, borrowing the asymptotic equivalence of the WAIC estimate (23) and CVIC expressed with the harmonic mean (IS) (12) (without $\boldsymbol{b}_i$) for models without latent variables, which is proved by Watanabe

(2010a).

6 Data Examples

6.1 Finite Mixture Models for Galaxy Data

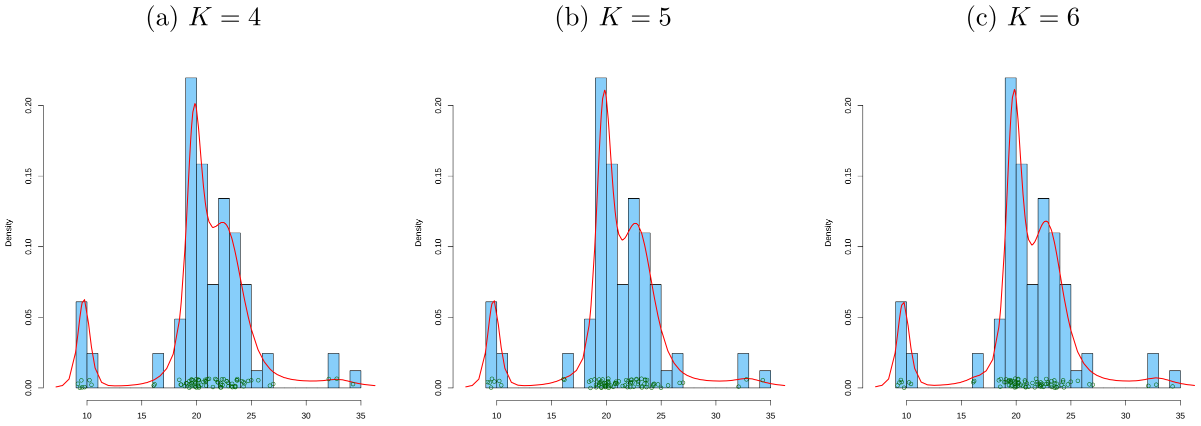
In this section we look at the performance of iIS and iWAIC in approximating the CVIC of finite mixture models fitted to `Galaxy` data (Postman et al., 1986; Roeder, 1990); this data is used very often to demonstrate mixture modelling methods. We obtained the data set from the R package `MASS`. The data set is a numeric vector of velocities (km/sec) of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. We applied mixture modelling to the velocities divided by 1000. A histogram of these 82 numbers is shown in each plot of Figure 2, which also shows three fitted density functions to be discussed later. Our purpose of computing CVIC for finite mixture models is to determine the number of mixture components K that can adequately capture the heterogeneity in the data without overfitting it. The finite mixture model that we used to fit Galaxy data is as follows:

$$
\begin{aligned}
y_i | z_i = k, \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}_{1:K} &\sim N(\mu_k, \sigma_k^2), \text{ for } i = 1, \ldots, n, & (28)\\
z_i | \boldsymbol{p}_{1:K} &\sim \text{Category}(p_1, \ldots, p_K), \text{ for } i = 1, \ldots, n, & (29)\\
\mu_k &\sim N(20, 10^4), \text{ for } k = 1, \ldots, K, & (30)\\
\sigma_k^2 &\sim \text{Inverse-Gamma}(0.01, 0.01 \times 20), \text{ for } k = 1, \ldots, K, & (31)\\
p_k &\sim \text{Dirichlet}(1, \ldots, 1) \text{ for } k = 1, \ldots, K. & (32)
\end{aligned}
$$

Here we set the prior mean of μ_k to 20 (which is the mean of the 82 numbers) and set the scale for Inverse Gamma prior for σ_k^2 to 20 (which is the variance of the 82 numbers).

This finite mixture model (equations (28) - (32)) falls in the class of models depicted by Figure 1: the observed variable is y_i, the model parameters denoted by $\boldsymbol{\theta}$ are $(\boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}^2_{1:K}, \boldsymbol{p}_{1:K})$, and the latent variable $\boldsymbol{b}_i$ is the mixture component indicator z_i. In this

Figure 2: Histograms of Galaxy data and three estimated density curves using MCMC samples from fitting finite mixture models with different numbers of components, $K = 4, 5, 6$ and the full data set.


(a) $K = 4$ (b) $K = 5$ (c) $K = 6$

model, the latent variables $z_1, \ldots, z_n$ are independent given the model parameter $\boldsymbol{\theta}$. It follows that $y_1, \ldots, y_n$ are independent given $\boldsymbol{\theta}$.

We used JAGS (Plummer, 2003) to run MCMC simulations for fitting the above model to `Galaxy` data with various choices of K. To avoid the problem in which MCMC may get stuck in a model with only one component, we followed the JAGS `eyes` example to restrict the MCMC to have at least a data point in each component. All MCMC simulations started with randomly generated $\boldsymbol{z}_{1:n}$ and ran 5 parallel chains, each doing 2000, 2000, and 100,000 iterations for adapting, burning, and sampling, respectively.

We ran 82 cross-validatory MCMC simulations with each of the 82 numbers removed (set to `NA` in JAGS). After each simulation, we computed the actual CV posterior predictive density $P(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$ using equation (5) with the evaluation function a set to $\phi(y_i^{\text{obs}}|\mu_{z_i}, \sigma^2_{z_i})$, where ϕ represents normal density. Using all 82 values of the CV posterior predictive densities, we can compute the CVIC using equation (7). The CVICs for different choices of K based on one simulation for each K are displayed in Table 1. We repeated computing CVICs quite a few times and the results were almost the same, with differences only in the 2nd decimal.

We then considered approximating CVIC using four different methods (nIS, nWAIC, iIS, iWAIC) from a single MCMC simulation that is based on all of the 82 numbers. The non-integrated predictive density for this model is $P(y_i^{\text{obs}}|z_i, \boldsymbol{\theta})$ as specified in (28); this is a normal density with mean μ_{z_i} and standard deviation σ_{z_i}, denoted by $\phi(y_i^{\text{obs}}|\mu_{z_i}, \sigma_{z_i})$. The values of $P(y_i^{\text{obs}}|z_i, \boldsymbol{\theta})$ computed with a collection of MCMC samples of $(z_i, \boldsymbol{\theta})$ are then used for computing nIS and nWAIC approximates of CV posterior predictive densities (with equations (13) and (24) respectively). We can then compute the nIS information criterion and the nWAIC by plugging the approximates of CV posterior predictive densities into (7). The integrated predictive density is $P(y_i^{\text{obs}}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \phi(y_i^{\text{obs}}|\mu_k, \sigma_k)$ (note that $\boldsymbol{z}_{-i}$ and y_i are independent given $\boldsymbol{\theta}$). We can then use $P(y_i^{\text{obs}}|\boldsymbol{\theta})$ for computing iIS and iWAIC approximates of CV posterior predictive densities (with equations (21) and (26) respectively), and corresponding information criterion values. In this example, iIS and iWAIC are just applications of IS and WAIC to mixture models with latent variables $\boldsymbol{z}_{1:n}$ integrated out.

Table 1: Comparisons of 5 information criteria for mixture models. The numbers are the averages of ICs from 100 independent MCMC simulations. The numbers in brackets indicates standard deviations.

K	DIC	nWAIC	nIS	iWAIC	iIS	CVIC
2	445.38(1.64)	420.27(0.39)	425.63(3.45)	449.56(0.14)	449.62(0.17)	450.55
3	528.78(45.12)	384.94(9.94)	391.29(6.17)	437.23(4.70)	436.43(3.79)	427.46
4	774.85(31.58)	339.91(1.87)	363.55(5.32)	422.43(0.53)	422.76(0.54)	423.16
5	710.88(25.34)	328.19(0.29)	362.30(3.70)	421.02(0.09)	421.41(0.10)	421.10
6	679.95(17.48)	323.62(1.33)	355.49(5.72)	420.97(0.27)	421.35(0.31)	421.34
7	675.27(18.57)	321.61(0.30)	364.41(4.49)	421.25(0.07)	421.64(0.12)	421.53

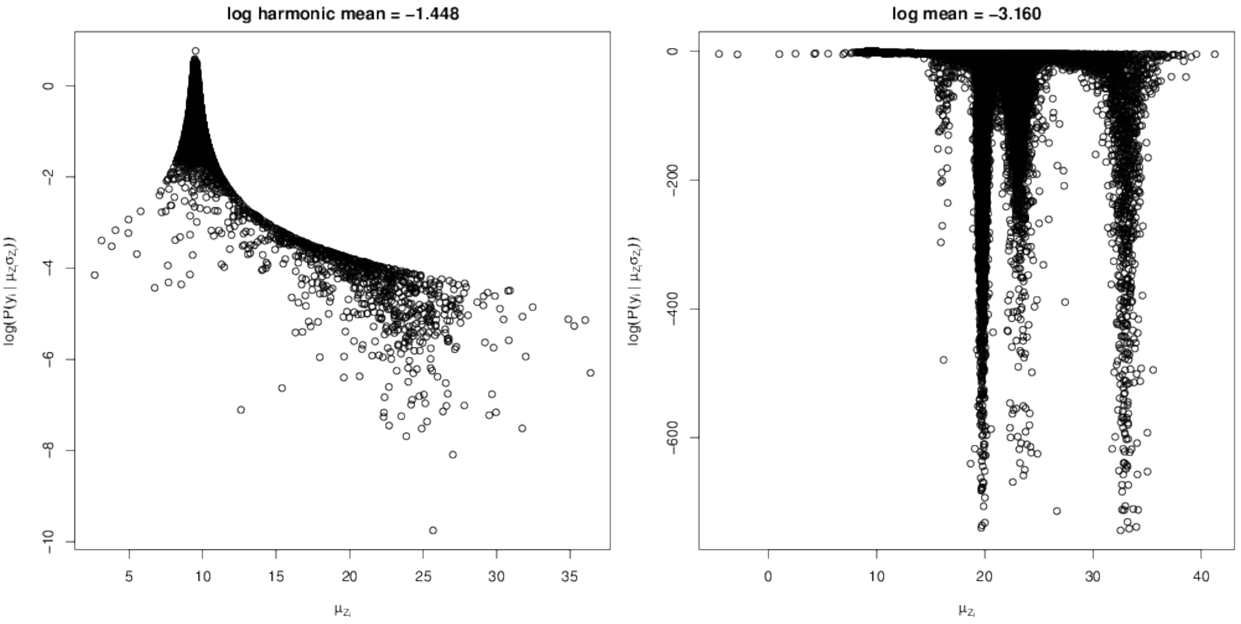
For each choice of K, we computed the above four criteria as well as DIC (using R package R2jags) for 100 independent MCMC simulations. Table 1 shows the means of these 100 information criterion values for each approximation method, with standard deviations shown in brackets. From the table, we see that the naive applications of WAIC and IS to non-integrated predictive densities $P(y_i^{\text{obs}}|z_i, \boldsymbol{\theta})$ do not work satisfactorily. They are both highly downward biased. Furthermore, nWAIC chooses over-complex models because nWAICs keep decreasing until $K = 7$, and nIS estimates of CVIC have very high variances. DICs for this example turn into a mess because the model parameters are non-identifiable. iIS and iWAIC provide significantly closer estimates of actual CVIC with much smaller standard deviations than other methods. These results show that using integrated predictive densities significantly improves accuracy of nIS and nWAIC. The results of iWAIC may not be surprising because here iWAIC is just an application of WAIC to the marginalized models with latent variables $\boldsymbol{z}_{1:n}$ integrated out in which the observed variables $y_1, \dots, y_n$ are independent given the model parameters. Watanabe (2010a) has proven the asymptotic equivalence of WAIC and CVIC in such models. iIS is also theoretically justified in Section 4.

Table 2: One-sided paired t-test p-values for comparing means of 82 log posterior predictive densities for Galaxy data given by mixture models with different number of mixture components, K.

pair of models	nWAIC	nIS	iWAIC	iIS	CVIC
K =3 vs $K = 2$	0.000	0.000	0.016	0.013	0.010
$K = 4$ vs $K = 3$	0.000	0.019	0.030	0.032	0.190
$K = 5$ vs $K = 4$	0.000	0.249	0.070	0.066	0.027
$K = 6$ vs $K = 5$	0.002	0.203	0.489	0.476	0.674
$K = 7$ vs $K = 6$	0.110	0.840	0.716	0.711	0.700

The CVIC is defined as being minus two times the sum of the log CV posterior predictive densities. Therefore, the statistical significance of the differences of two CVICs (or estimates) can be assessed by looking at the population mean differences of two groups of log CV posterior predictive densities (Vehtari, 2001; Vehtari and Lampinen, 2002). We conducted a one-sided paired t-test to determine whether a finite mixture model with K components provides a better fit (larger mean of CV posterior predictive densities) to `Galaxy` data than a mixture model with $K-1$ components. The p-values of the comparisons for $K = 3, \dots, 7$ for actual CV posterior predictive densities are given in Table 2 (column CVIC). We also conducted the same test for the log CV posterior predictive densities estimated by four different methods (nIS, iIS, nWAIC, iWAIC). Due to the variations in these estimates, we computed the p-values 1000 times by randomly drawing two simulation results from models with K and $K-1$ components. We then computed the mean of the 1000 p-values. Table 2 shows the results for all four different estimation methods. From the table, we see that iIS and iWAIC provides much closer p-values to those based on actual CV posterior predictive densities than nIS and nWAIC. These p-values indicate that mixture models with 5 components are adequate to capture the heterogeneity in Galaxy data, and 6-component mixture models do not provide a significantly better fit. These conclusions can be visualized by the density curves given by the fits for $K = 4, 5, 6$, where the curves with $K = 4$ and $K = 5$ are different, but the curves with $K = 5$ and $K = 6$ are almost the same.

Figure 3: Scatter-plot of non-integrated predictive densities against μ_{z_i}, given MCMC samples from the full data posterior (3a) and the actual CV posterior with the 3rd number removed (3b), when $K = 5$ components are used.


(a)

(b)

Last, we explain why naive applications of IS and WAIC to non-integrated predictive densities cannot provide good estimates of CV posterior predictive densities. Figure 3 shows scatter-plots of the log non-integrated predictive density, *i.e.* $\log(P(y_i^{\text{obs}}|z_i, \boldsymbol{\theta})) = \log(\phi(y_i^{\text{obs}}|\mu_{z_i}, \sigma_{z_i}))$, against μ_{z_i}, computed with each MCMC sample of $(z_i, \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}_{1:K})$ from the full data posterior (Figure (3a)) and the actual CV posterior with y_i^{obs} removed (Figure (3b)), where y_i^{obs} is the 3rd of the 82 numbers. In this figure we see great discrepancy between the posterior distribution of the non-integrated predictive density with and without y_i^{obs} included in MCMC simulations. When we simulate MCMC with the full data (y_i^{obs} included), most of the z_i visit components that fit y_i^{obs} well, with μ_{z_i} typically being close to 10. Thus, the non-integrated predictive densities are usually very large. When we simulate MCMC with y_i^{obs} removed, most of the z_i visit large components, with μ_{z_i} typically being in the interval from 10 to 35; such cases often do not fit y_i^{obs} well. The reason is that without the inclusion of y_i^{obs}, z_i will more likely take larger components. Thus, values of $P(y_i|\boldsymbol{\theta}, z_i)$ in the CV posterior are orders of magnitude smaller than those in the full data posterior. This indicates that the difference between the CV posterior and full data posterior of z_i is huge. Applying IS and WAIC to the non-integrated predictive densities alone is unable to correct for much of the bias due to the inclusion of y_i^{obs} in MCMC simulation. By averaging the non-integrated predictive density over regenerated z_i given $\boldsymbol{\theta}$ but not y_i^{obs}, we significantly reduce the optimistic bias in $P(y_i^{\text{obs}}|\boldsymbol{\theta}, z_i)$ due to inclusion of y_i^{obs}. This explains why iIS and iWAIC provide significantly closer estimates to CVIC than nIS and nWAIC.

6.2 A Simulation Study with Finite Mixture Models

In this section, we report a simulation study with the same mixture models that were described in Section 6.1. We simulated 100 data sets, each containing 200 data points y_i from a mixture distribution with $K = 4$ normal components: $(1/4)N(-7, 1) + (1/4)N(-2, 1) + (1/4)N(1, 1) + (1/4)N(7, 1)$. The kernel density of one of the data sets is shown in Figure 4. From this plot, we see that the middle two components may be hard to separate in some data sets.

We fitted each of the 100 data sets using the procedure described in Section 6.1 and then computed the information criterion (IC) using each of the five methods (nWAIC, nIS, iWAIC, iIS and DIC). Table 3 shows the IC values for two selected data sets. Table (4a) shows average IC values in the 100 data sets for each model indexed by K (row) and for each method for approximating CVIC (column). Table (4b) shows frequencies of selected models in the 100 data sets by looking at the minimum IC value computed with each of the five methods (column).

Figure 4: Kernel Density of a Simulated Data Set.

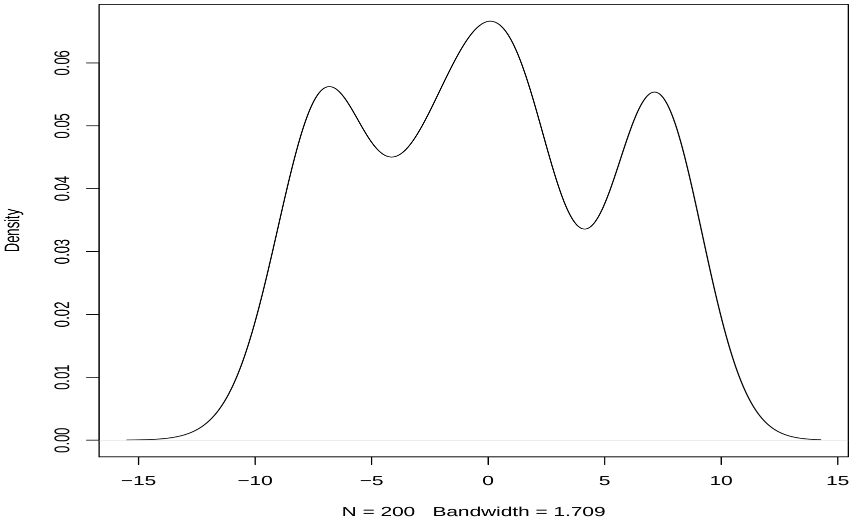


Table 3: Values of information criterion for two selected data sets. The bold-faced numbers show the smallest IC values for each method.

(a) iIS and iWAIC select $K = 4$

K	nWAIC	nIS	iWAIC	iIS	DIC
2	1022.95	1081.11	1163.96	1163.99	1143.19
3	690.39	855.85	1088.20	1088.26	**850.28**
4	642.82	782.94	**1083.16**	**1083.28**	1416.81
5	640.27	754.13	1084.29	1084.48	1351.18
6	638.25	756.82	1085.25	1085.51	1382.83
7	**637.12**	**727.84**	1086.46	1086.76	1479.03

(b) iIS and iWAIC do not select $K = 4$

K	nWAIC	nIS	iWAIC	iIS	DIC
2	1144.14	1122.07	1202.25	1199.90	1192.42
3	758.35	924.19	1101.48	1101.53	**1024.83**
4	706.30	830.63	1095.42	1095.54	1510.18
5	691.02	821.51	1094.84	1094.99	1561.51
6	679.71	800.38	**1094.64**	**1094.80**	1652.05
7	**673.63**	**794.05**	1094.69	1094.87	1740.80

Table 4: Model selection results for 100 data sets simulated from finite mixture models with $K = 4$. Table (4a) shows average IC values in the 100 data sets for each model with K components (row) and for different methods (column). Each column of Table (4b) shows frequencies of selected models indexed by K by looking at the minimum IC value in the 100 data sets.

(a) Average of IC values in 100 data sets

K	nIS	nWAIC	iIS	iWAIC	DIC
2	1112.48	1103.95	1181.60	1182.33	1248.97
3	922.88	751.58	1105.18	1105.11	990.51
4	827.06	682.62	1099.42	1099.26	1572.80
5	810.42	674.39	1099.18	1098.96	1562.05
6	801.24	669.57	1099.60	1099.31	1630.02
7	796.65	666.39	1100.09	1099.77	1700.12

(b) Frequencies of selected models

K	nIS	nWAIC	iIS	iWAIC	DIC
2	0	0	0	0	2
3	0	0	15	15	94
4	6	15	39	37	4
5	10	4	21	20	0
6	30	8	11	13	0
7	54	73	14	15	0

In both of the data sets shown in Table 3, nWAIC and nIS select the model with $K = 7$, which is more flexible than the true data generating model, which has $K = 4$. This is typical for the 100 data sets, which can be seen from Tables (4a) and (4b). DIC, on the other hand, almost always selects the model with $K = 3$ which is simpler than the true model ($K = 4$). These results are consistent with what we observe from the analysis of `Galaxy` data. For iWAIC and iIS, their IC values have a sharp decrease from $K = 2$ until $K = 4$, compared to the changes in the IC values after $K = 4$, which stabilize and have only very small variation. This small variation sometimes leads to wrong model selection results if one chooses the model with the smallest IC value, for example in the data set shown by Table (3b). Therefore, IC may not be sensitive enough in penalizing over-complex models. Overall, in this example, iWAIC and iIS outperform nWAIC, nIS and DIC in comparing models because of the improved estimation of CVIC. With IC values computed by iWAIC and iIS, the appropriate decision ($K = 4$) can be made for most data sets if one does not simply look only at which model has the smallest IC value, but rather by also looking at the change of IC values in all models considered.

In this example, IC does not penalize over-complex models sensitively. This insensitivity occurs because the posterior inference with MCMC itself is robust to over-complexity in models; that is, MCMC simulation can automatically adjust the model complexity. In this example, although we fit a mixture model with $K = 7$ components, some components have very small proportions in MCMC samples, which is effectively a simpler model. This has been long known as a good property for Bayesian inference; see extensive discussions by Neal (1995). However, this poses some difficulty when selecting models by looking at CVIC. We've noticed that recently Wang and Gelman (2014) have also discussed the insensitivity of CVIC, where they explain that CVIC itself is not sensitive in distinguishing models for binary data. Overall, determining a threshold for CVIC (even computed with actual cross-validation) for selecting models, particularly among models with slight difference, is still a problem, which demands further study. Looking at the population mean of log CV posterior predictive density may be an option, as we discuss in the `Galaxy` data analysis results. However, we feel that generally this may be too conservative because a model is better than another not because it can provides better predictive accuracy for "most" observations (resulting in a sharp change in population mean — CVIC), but rather it can provide better prediction only for a fraction of observations. Perhaps, we should look at the proportion of units whose predictions have been improved with a more complex model.

6.3 Random Spatial Effect Models for Scottish Lip Cancer Data

In this section, we investigated the performance of iIS and iWAIC in an analysis of Scottish lip cancer data which was used in Stern and Cressie (2000); Spiegelhalter et al. (2002); Plummer (2008). The data set was extracted from the paper by Stern and Cressie (2000). The data represents male lip cancer counts (over the period 1975 - 1980) in the $n = 56$ districts of Scotland. At each district i, the data include these fields: (1) the number of observed cases of lip cancer, y_i; (2) the number of expected cases, E_i, calculated based on standardization of "population at risk" across different age groups; (3) the percent of population employed in agriculture, fishing and forestry, x_i, used as a covariate; and (4) a list of the neighbouring regions.

The y_i, for $i = 1, \ldots, n$, is modelled as an independent Poisson random variable conditional on λ_i and E_i:

$$y_i | E_i, \lambda_i \sim \text{Poisson}(\lambda_i E_i), \tag{33}$$

where λ_i denotes the underlying relative risk for district i, and E_i stands for expected counts. Let $s_i = \log(\lambda_i)$ and $\boldsymbol{X} = (x_1, \ldots, x_n)'$. We consider four different models for the vector $\boldsymbol{s} = (s_1, \cdots, s_n)'$ conditional on $\boldsymbol{X}$ and neighbouring information between districts:

$$\text{spatial+linear (called } \textit{full} \text{ for short)} : \boldsymbol{s} \sim N_n(\alpha + \boldsymbol{X}\beta, \Phi\tau^2), \tag{34}$$

$$\text{spatial} : \boldsymbol{s} \sim N_n(\alpha, \Phi\tau^2), \tag{35}$$

$$\text{linear} : \boldsymbol{s} \sim N_n(\alpha + \boldsymbol{X}\beta, I_n\tau^2), \tag{36}$$

$$\text{exchangable} : \boldsymbol{s} \sim N_n(\alpha, I_n\tau^2), \tag{37}$$

where $\Phi = (I_n - \phi C)^{-1} M$ is a matrix for capturing the spatial correlations amongst the n districts, in which the elements of C are: $c_{ij} = (E_j/E_i)^{1/2}$ if areas i and j are neighbours, and 0 otherwise; the elements of M are: $m_{ii} = E_i^{-1}$ and $m_{ij} = 0$ if $i \neq j$. The multivariate normal distribution with Φ as its covariance matrix is called a **proper conditional autoregression (CAR) model**. Derived from the joint distribution in (34), the conditional distribution of $s_i | \boldsymbol{s}_{-i}, \alpha, \beta, \phi$ is:

$$s_i | \boldsymbol{s}_{-i}, \boldsymbol{\theta} \sim N(\alpha + x_i\beta + \phi \sum_{j \in N_i} (c_{ij}(s_j - \alpha - x_j\beta)), \tau^2 m_{ii}), \tag{38}$$

where N_i is the set of neighbours of district i. From (38), we see that ϕ controls the degree of spatial dependency of s_i on its neighbours. At a higher level, diffused priors are assigned to α, β, τ, and ϕ: $\alpha \sim N(0, 1000^2)$, $\beta \sim N(0, 1000^2)$, $\tau^2 \sim \text{Inv-Gamma}(0.5, 0.0005)$, $\phi \sim \text{Unif}(\phi_0, \phi_1)$, where (ϕ_0, ϕ_1) is the interval for ϕ such that Φ is positive-definite (see Stern and Cressie, 2000). In model (34), we consider both spatial and linear effects of x_i in modelling $\boldsymbol{s}$. One may also consider other models. Model (35) considers only spatial effects, model (36) considers only linear effects, and model (37) considers no spatial or linear effects. We are interested in comparing goodness-of-fits of the four models to the lip cancer data set so as to determine which model is the most appropriate for Scottish lip cancer data. CVIC is one criterion for measuring such a goodness-of-fit.

All the above four models belong to the class of Bayesian latent variable models depicted by Figure 1. The observable variable is y_i, the latent variable is s_i , and the model parameters $\boldsymbol{\theta}$ in model (34) are $(\alpha, \beta, \tau, \phi)$. A subset of these parameters are considered for other models depending on which parameters are used in the respective models. We used OpenBUGS through R package `R2OpenBUGS` to run MCMC simulations for fitting each of the above models to the Scottish lip cancer data. For each simulation, we ran two parrallel chains, each with 15000 iterations, and the first 5000 were discarded as burn-in.

For each model, we first ran actual 56 cross-validatory MCMC simulations with each of the 56 obervations removed (set to `NA` in OpenBUGS) and then computed the actual CV posterior predictive density $P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})$ using equation (5) with the evaluation function set to $\text{dpoisson}(y_i^{\text{obs}}|\lambda_i E_i)$, *i.e.* the Poisson probability mass function with parameter $\lambda_i E_i$. Then we computed the CVIC using equation (7). We computed the actual CVIC 10 times for each model although actual LOOCV gives very stable results. The averages and standard deviations of 10 CVICs for different models are displayed in Table 5. From this table, we see that the spatial+linear model is the best fit for the Scottish lip cancer data according to CVIC.

We then consider approximating the CVIC with four different methods (nIS, nWAIC, iIS, and iWAIC) from a single MCMC simulation based on all of the 56 observations. The non-integrated predictive density used in computing nIS and nWAIC with equations (13) and (24) is $\text{dpoisson}(y_i^{\text{obs}}|\lambda_i E_i)$, where $\lambda_i = \exp(s_i)$. Next, we describe how to compute iIS and iWAIC for model (34). The integrated predictive density (18) required by (21) and (26) is:

$$P(y_i^{\text{obs}} \,|\, \boldsymbol{\theta}, \boldsymbol{s}_{-i}) = \int \text{dpoisson}(y_i^{\text{obs}}|\lambda_i E_i) P(s_i \,|\, \boldsymbol{\theta}, \boldsymbol{s}_{-i}) ds_i, \tag{39}$$

where $P(s_i|\boldsymbol{\theta}, \boldsymbol{s}_{-i})$ is given by equation (38). Because there is no closed form for the integral (39), we use Monte Carlo methods to estimate it by generating 200 random numbers from $P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})$ (note, this is done for each retained MCMC sample of $(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})$ and each validation unit i, with s_i alternately discarded). Finally, based on computed values of $P(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i})$ for all MCMC samples, we can then compute iIS and iWAIC approximates of the CV posterior predictive densities (with equations (21) and (26) respectively) and then the corresponding iIS information criterion and iWAIC. iIS and iWAIC are computed similarly for models (35) - (37), with only a change of the conditional distribution (38) according to their joint prior distributions.

We repeated computing the values of the above four criteria as well as DIC for 100 independent MCMC simulations based on each model. The means of these 100 information criterion values for each method and each model are shown in Table 5, with standard deviations shown in brackets. We see from this table that iIS and iWAIC provide significantly closer approximates to actual CVIC than nIS, nWAIC and DIC; furthermore, the approximates by iWAIC and iIS are almost identical to actual CVIC. In contrast, DIC has large biases and variances when spatial effects are considered, and also the mean DIC of the spatial + linear model is bigger than the mean DIC of the model with spatial effects only. This suggests that if we randomly draw one MCMC simulation out of the 100 simulations based on each model, the probability that DIC does NOT pick up the spatial+linear model as the optimal model is high (56.6% if we assume the DICs are normally distributed). nWAIC and nIS also have large biases and variances. In particular, nWAIC nearly never chooses the spatial+linear model (with a probability close to 1 if nWAICs are normally distributed). nIS has a good chance (0.92 if the values are normally distributed) to choose the spatial+linear model. However, nIS is numerically unstable and has a fairly large variance, which has been well-known to many people (Spiegelhalter et al., 2002). In summary, the integration applied to latent variables associated with each validation unit substantially improves the estimates of CVIC given by nWAIC and nIS.

The good approximations of CVIC by iIS is not surprising, because our derivation in

Table 5: Comparisons of information criteria for Scottish lip cancer data. Except for CVIC, each table entry shows the average of 100 information criterion values computed from 100 independent MCMC simulations, and the standard deviation in bracket. For CVIC, the average and standard deviation are from 10 independent LOOCV evaluations.

	DIC	nWAIC	nIS	iWAIC	iIS	CVIC
spa.+lin.	269.43(12.30)	306.82(0.21)	335.54(1.27)	344.47(0.12)	345.21(0.19)	343.88(0.14)
spatial	266.79(10.15)	304.61(0.18)	338.77(1.85)	354.11(0.06)	356.06(0.37)	352.54(0.14)
linear	310.42(0.11)	306.94(0.21)	338.81(3.02)	350.48(0.05)	350.54(0.05)	349.48(0.11)
exch.	312.57(0.12)	306.74(0.17)	346.55(3.46)	368.01(0.03)	368.08(0.03)	366.61(0.00)

Section 4.2 has shown the equivalence in these models. It is surprising to note that the heuristic iWAIC also gives estimates very close to CVIC for model (34) and (35), which contain actually correlated random effects. Furthermore, note that iWAIC has smaller standard deviations and biases than iIS. Therefore, the equivalence of iWAIC to iIS (or CVIC) deserves more empirical and theoretical investigation in the future.

6.4 CV Posterior p-values in Logistic Regression for Seeds Data

We compare here different methods for computing posterior p-values for the purpose of identifying outliers in applying logistic regression with random effects to `Seeds` data, a classic example of WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/Vol1.pdf). We obtained the data set from the previous link. The example is taken from Table 3 of Crowder (1978). The study data contains the proportion of seeds that germinated on each of 21 plates arranged according to a 2 by 2 factorial layout by seed and type of root extract. For $i = 1, \ldots, 21$, let r_i be the number of germinated seeds in the ith plate, n_i be the total number of seeds in the ith plate, x_{i1} be the seed type (0/1), and x_{i2} be root extract (0/1). The conditional distribution of r_i given n_i, x_{i1} and x_{i2} are specified as follows:

$$r_i|n_i, p_i \sim \text{Binomial}(n_i, p_i), \tag{40}$$

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i, \tag{41}$$

$$b_i \sim N(0, \sigma^2), \tag{42}$$

and parameters $\alpha_0, \alpha_1, \alpha_2, \alpha_{12}$ are assigned with $N(0, 10^6)$ as prior, and σ^2 is assigned with Inverse-Gamma (0.001, 0.001) as prior. The above model is a member of Bayesian latent variable models depicted by Figure 1. The observable variable is r_i, the latent variable is b_i, the covariate variable vector is (n_i, x_{i1}, x_{i2}), and the model parameter vector $\boldsymbol{\theta}$ is $(\alpha_0, \alpha_1, \alpha_2, \alpha_{12})$. We used JAGS to run MCMC to fit the above model to the `Seeds` data. For each simulation, we ran 5 parallel chains, each running 1000 iterations for adapting, 2500 iterations for burning in, and 10000 iterations for sampling.

The p-value (given parameters and latent variables) defined by (8) for this example is the right tail probability of the Binomial distribution with n_i trials and success rate p_i:

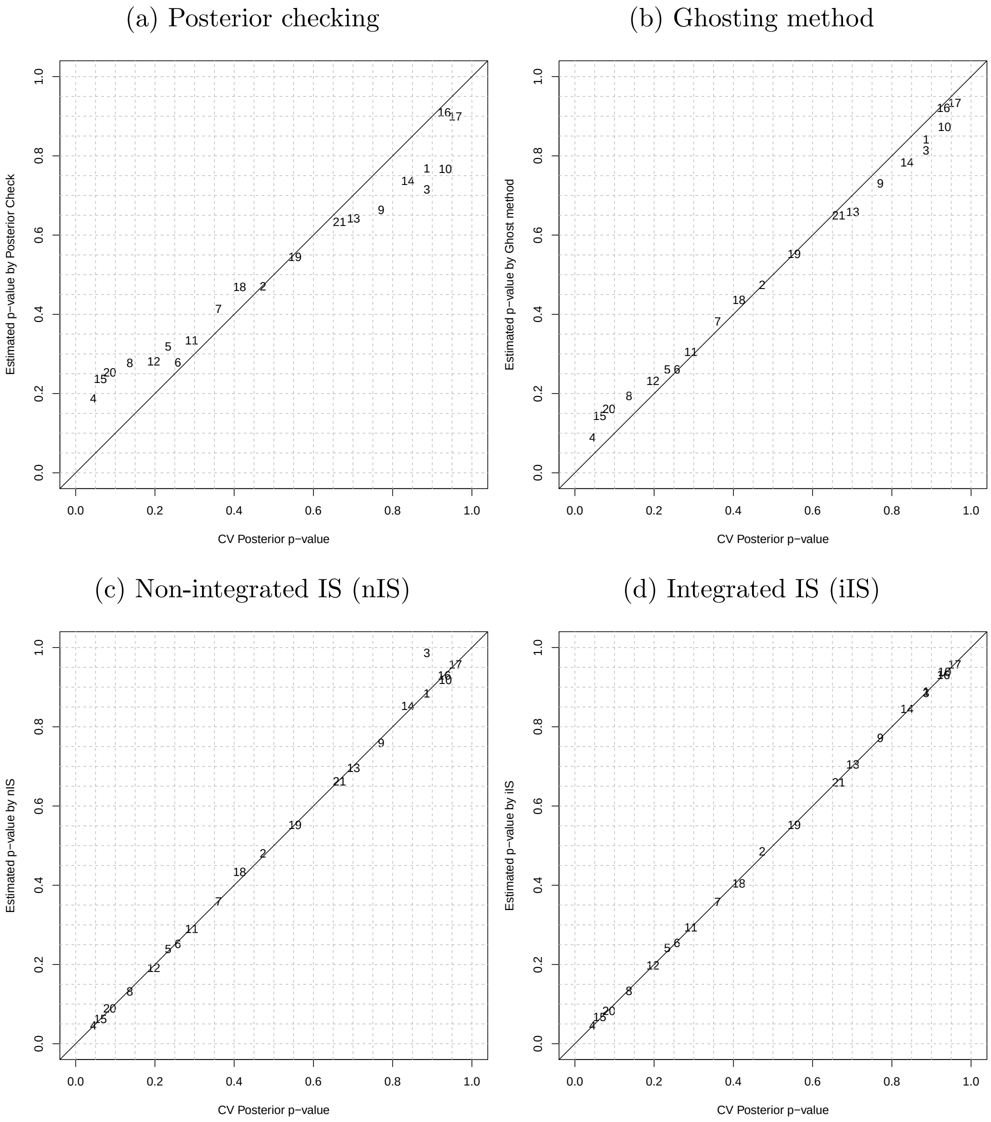
$$\text{p-value}(r_i^{\text{obs}}, \boldsymbol{\theta}, b_i) = 1 - \text{pbinom}(r_i^{\text{obs}}; n_i, p_i) + 0.5\,\text{dbinom}(r_i^{\text{obs}}; n_i, p_i), \tag{43}$$

where r_i^{obs} is the actual observation of r_i, and pbinom and dbinom denote the CDF and PMF of Binomial distribution. Very small or very large p-values indicate that the actual observed r_i^{obs} falls on the tails of (*i.e.* is unusual to) Binomial (n_i, p_i). The CV posterior p-value (Marshall and Spiegelhalter, 2003) for the observation r_i^{obs} is the mean of p-value$(r_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ with respect to the CV posterior distribution $P(\boldsymbol{\theta}, b_i|\boldsymbol{r}_{-i}^{\text{obs}})$. If we get a very small or very large CV posterior p-value for observation r_i^{obs}, it indicates that r_i^{obs} is unusual to the predictive distribution of r_i given $\boldsymbol{r}_{-i}^{\text{obs}}$. For this example, when the CV posterior p-value for r_i^{obs} is very small or very large, the germination rate (r_i^{obs}) of the ith plate is probably an outlier to the other plates. Marshall and Spiegelhalter (2007) showed that the CV posterior p-values are uniformly distributed on the interval $(0, 1)$. We ran actual CV MCMC simulations to find the CV posterior p-values for each of the 21 plates, and the results are displayed by the x-axis on the plots in Figure 5.

We compared four different methods for computing posterior p-values for identifying outliers with only a single MCMC simulation based on the full data set. One method is to apply the posterior checking idea of Gelman et al. (1996) without considering bias-correction; that is, to average each p-value$(r_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)$ with respect to the posterior of $(\boldsymbol{\theta}, b_i)$ given the full data set $\boldsymbol{r}_{1:21}^{\text{obs}}$. We will call this method *posterior checking*. Gelman et al. (1996) do not recommend this use of posterior checking because it uses the data in both model building and assessment. However, this method is convenient and so thus is perhaps used very often in practice; therefore, we include it in our comparison. To reduce the bias of including r_i^{obs} in model fitting, Marshall and Spiegelhalter (2003) propose the *ghosting method*: for each MCMC sample, one averages the p-value$(r_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ with respect to the conditional distribution of b_i given $\boldsymbol{\theta}$ (but without r_i^{obs}) to obtain the ghosting p-value (which can be done using Monte carlo methods by re-generating b_i given $\boldsymbol{\theta}$ with no reference to r_i^{obs}), then averages the ghosting p-values over all MCMC samples. The third method is the *non-integrated importance sampling* method (nIS) that averages the p-value $(r_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ after being weighted with the inverse of the probability density (mass) of r_i^{obs}: $1/\text{dbinom}(r_i^{\text{obs}}; n_i, p_i)$. The fourth method is *integrated importance sampling* (iIS). For each MCMC sample, we first average both the p-value $(r_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ and $\text{dbinom}(r_i^{\text{obs}}; n_i, p_i)$ with respect to $P(b_i|\boldsymbol{\theta})$ to find the integrated evaluation p-value (equation (47)) and the integrated predictive density (equation (18)) respectively, then compute the weighted average of the integrated p-values with the reversed integrated predictive density as weights over all MCMC samples using formula (19). We can see that the way to obtain the ghosting p-value is the same as finding the integrated p-value in (47) when $\boldsymbol{b}_{1:n}$ are independent given $\boldsymbol{\theta}$, but without using the reversed integrated density to correct for the optimistic bias in full data posterior of parameters. Therefore, the ghosting method can be viewed as a partial implementation of the iIS method presented here.

We calculated 21 posterior p-values using the four methods given a MCMC simulation based on the full data set, and repeated this calculation for 100 independent MCMC simulations. For computing integrated p-values and predictive densities as needed by nIS and the ghosting method, we generated 30 values of b_i from $N(0, \sigma^2)$ for each plate and each MCMC sample. Figure 5 shows the scatter-plots of four sets of estimated posterior p-values given by the four different methods against the actual CV posterior p-values from one MCMC

Figure 5: Scatterplots of estimated posterior p-values from an MCMC simulation against actual CV posterior p-values. The number for each point refers to the index of a plate.



(a) Posterior checking

(b) Ghosting method

(c) Non-integrated IS (nIS)

(d) Integrated IS (iIS)

simulation. From the figure, we see that the p-values given by posterior checking are more concentrated around 0.5 than the actual CV posterior p-values, and do not appear to be uniformly distributed (Gelman, 2013). This is because in computing each p-value, the observed value r_i^{obs} itself is included in model fitting, resulting in optimistic bias. The ghosting method reduces this bias, hence the estimated p-values are closer to the actual CV p-values and are more spread out over $(0, 1)$. However, for this example, the bias is still visible from Figure (5b). Both nIS and iIS give estimates that are very close to the actual values found by CV. However, nIS is less stable than iIS, and sometimes gives very poor estimates; for example the 3rd plate shown in Figure (5c).

To measure more precisely the accuracy of estimated p-values to the actual CV p-values, we use absolute relative error in percentage scale defined as

$$\text{RE} = (1/n) \sum_{i=1}^{n} \frac{|\hat{p}_i - p_i|}{\min(p_i,\ 1 - p_i)} \times 100, \tag{44}$$

where $\hat{\boldsymbol{p}}_{1:n}$ are estimates of $\boldsymbol{p}_{1:n}$. This measure greatly emphasizes the error between $\hat{p}_i$ and p_i when p_i is very small or very large, for which we demand more on absolute error than when p_i is close to 0.5. A similar measure (only using p_i in denominator) is used by Marshall and Spiegelhalter (2007). Here we modify the denominator because large p-values are important too. Table 6 shows the averages of REs over 100 independent simulations for each method. Clearly, we see that iIS is the best method among the four, and offers significant improvement over the ghosting and posterior checking methods.

Table 6: Comparisons of the averages of 100 absolute relative errors (in percentage) of estimated CV p-values from 100 independent MCMC simulations, for logistic regression example. The numbers in brackets indicate standard deviations.

iIS	nIS	Ghosting	Posterior checking
2.319(0.399)	5.234(1.083)	35.610(1.267)	93.887(3.854)

7 Conclusions and Discussions

In this article, we have introduced two methods (iIS and iWAIC) for approximating leave-one-out cross-validatory predictive evaluations for models with unit-specific and possibly correlated latent variables. The innovation in iIS and iWAIC is that we replace the non-integrated predictive density and evaluation functions by the integrated predictive density and evaluation functions. iIS is applicable to models with non-identifiable parametrization for which DIC may not be suitable; it is also applicable to models with correlated latent variables for which WAIC is not. The extent of applicability of iWAIC remains to be investigated. We have tested iIS and iWAIC in four examples, in which iIS and/or iWAIC provide almost identical approximates to what is given by actual leave-one-out cross-validation, whereas other methods show large discrepancies. In addition, we have found that iWAIC is slightly more stable than iIS.

Although our empirical results show that iIS and iWAIC provide better approximates of CVIC than DIC, we notice that the implementations of iIS and iWAIC are much more complicated, and requires users to have basic knowledge in statistics and scientific computing (for example a degree in statistics). For the moment, we do not know how to automate their applications as DIC, which can be embedded into a black-box MCMC sampler program. This is a direction for future research one can pursue.

Applicability of iWAIC to models with correlated latent variables still requires more empirical and theoretical investigations. The results of our empirical studies to the lip cancer data gives an example where iWAIC provides very close approximates to CVIC. However, we have to be cautious in the generalization of iWAIC to other models and problems. In the future, we will empirically test iWAIC in many other models using correlated latent variables, for example, the stochastic volatility models used for modelling financial time sequences (Jacquier et al., 2002; Gander and Stephens, 2007), multivariate spatial models (Feng and Dean, 2012), and many other models considering spatial and temporal correlations

(Waller et al., 1997). We will also investigate iWAIC theoretically, probably using the tools for singular statistical models developed by Watanabe (2009).

There is also much research work needed to generalize and extend iIS and iWAIC. We have only shown how to integrate latent variables away in the models where they are unit-specific to improve ordinary nIS and nWAIC. In many models, a latent variable is shared by many observations. It is still unclear to us how to improve nIS and nWAIC in such models. More ambitiously, we may wonder whether there is a method that requires little technical work but provides very good predictive evaluation for all Bayesian models.

The insensitivity of CVIC is another important problem that demands further research, as we discuss in Section 6.2. One may consider evaluation functions other than the log predictive density for capturing sensitively the difference among models. One may also consider other methods for comparing two sets of log predictive densities resulting from two competing models. However, we think that the method we present in this article for latent variable models may be generally useful for providing better approximation of cross-validatory quantities regardless of the choice of evaluation function.

Appendices

A Working procedure of iIS

1. Generate MCMC samples $\{(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{1:n}^{(s)})$; s= 1,. . . ,S$\}$ from $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$

2. For each $s = 1, \ldots, S$

   (a) For each $i = 1, \ldots, n$, generate $\{\boldsymbol{b}_i^{(s,r)}; r = 1, \ldots, R\}$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}^{(s)}, \boldsymbol{\theta}^{(s)})$, and estimate $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})$ by

$$\hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = (1/R)\sum_{r=1}^{R} P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}, \boldsymbol{b}_i^{(s,r)}). \tag{45}$$

Then, we can find the iIS weight:

$$W_i^{\mathrm{iIS}}(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = \frac{1}{\hat{P}(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})}. \tag{46}$$

(b) For each $i = 1, \ldots, n$, generate $\{\tilde{\boldsymbol{b}}_i^{(s,k)}; k = 1, \ldots, K\}$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}^{(s)}, \boldsymbol{\theta}^{(s)})$, and estimate the integrated evaluation function A by

$$A(\boldsymbol{y}_i^{\mathrm{obs}}, \boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = (1/K) \sum_{k=1}^{K} a\left(\boldsymbol{y}_i^{\mathrm{obs}}, \boldsymbol{\theta}^{(s)}, \tilde{\boldsymbol{b}}_i^{(s,k)}\right). \tag{47}$$

3. Estimate the expected evaluation function a with respect to $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\mathrm{obs}})$ by

$$\hat{E}_{\text{post(-i)}}^{\mathrm{iIS}}(a(\boldsymbol{y}_i^{\mathrm{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)) = \frac{(1/S)\sum_{s=1}^{S}\left[A(\boldsymbol{y}_i^{\mathrm{obs}}, \boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) W_i^{\mathrm{iIS}}(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})\right]}{(1/S)\sum_{s=1}^{S} W_i^{\mathrm{iIS}}(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})}. \tag{48}$$

Note that if we are only interested in computing CVIC, we don't need to do step 2(b); simply take the numerator in (48) to be 1 as warranted by theory.

B Working procedure of iWAIC

1. Generate MCMC samples $\{(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{1:n}^{(s)})$; s= 1,...,S$\}$ from $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\mathrm{obs}})$

2. For each $s = 1, \ldots, S$ and each $i = 1, \ldots, n$, generate $\{\boldsymbol{b}_i^{(s,r)}; r = 1, \ldots, R\}$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}^{(s)}, \boldsymbol{\theta}^{(s)})$, and estimate the integrated predictive density $P(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})$ by

$$\hat{P}(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = (1/R)\sum_{r=1}^{R} P(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}, \boldsymbol{b}_i^{(s,r)}). \tag{49}$$

3. Estimate the log CV posterior predictive density:

$$\log(\hat{P}(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{y}_{-i}^{\mathrm{obs}})) = \log((1/S)\sum_{s=1}^{S} \hat{P}(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})) - V_{s=1}^{S} \log(\hat{P}(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})), \tag{50}$$

where $V_{s=1}^{S} a^{(s)}$ stands for the sample variance of $(a^{(1)}, \ldots, a^{(S)})$.

4. Find iWAIC:

$$\mathrm{iWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}(\boldsymbol{y}_i^{\mathrm{obs}}|\boldsymbol{y}_{-i}^{\mathrm{obs}})). \tag{51}$$

References

Ando, T.: Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. Biometrika **94**(2), 443–458 (2007)

Berger, J.O., Pericchi, L.R.: The intrinsic bayes factor for model selection and prediction. Journal of the American Statistical Association **91**(433), 109–122 (1996). URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476668

Bernardo, J.M., Rueda, R.: Bayesian hypothesis testing: A reference approach. International Statistical Review **70**(3), 351–372 (2002). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2002.tb00175.x/abstract

Bhattacharya, S., Haslett, J.: Importance re-sampling MCMC for cross-validation in inverse problems. Bayesian Analysis **2**(2), 385–407 (2007)

Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M.: Deviance information criteria for missing data models. Bayesian Analysis **1**(4), 651–673 (2006)

Chib, S.: Marginal likelihood from the gibbs output. Journal of the American Statistical Association **90**(432), 1313–1321 (1995)

Crowder, M.J.: Beta-binomial anova for proportions. Applied Statistics **27**, 34–37 (1978)

Epifani, I., MacEachern, S.N., Peruggia, M.: Case-deletion importance sampling estimators: Central limit theorems and related results. Electronic Journal of Statistics **2**, 774–806 (2008)

Feng, C., Dean, C.: Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. Environmetrics **23**(6), 493–508 (2012)

Gander, M.P., Stephens, D.A.: Stochastic volatility modelling in continuous time with general marginal distributions: Inference, prediction and model selection. JSPI **137**(10), 3068–3081 (2007)

Gelfand, A.E., Dey, D.K., Chang, H.: Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In: Bayesian Statistics 4, pp. 147–167 (1992)

Gelman, A.: Understanding posterior p-values. unpublished online manuscript, available from Gelman's website. pp. 1–8 (2013)

Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for bayesian models. Statistics and Computing **24**(6), 997–1016 (2014). doi:10.1007/s11222-013-9416-2. URL http://link.springer.com/article/10.1007/s11222-013-9416-2

Gelman, A., Meng, X.: Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical Science **13**(2), 163–185 (1998)

Gelman, A., Meng, X., Stern, H.: Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica **6**(4), 733–760 (1996)

Geweke, J.: Bayesian inference in econometric models using monte carlo integration. Econometrica: Journal of the Econometric Society **57**, 1317–1339 (1989)

Hachiya, H., Akiyama, T., Sugiyama, M., Peters, J.: Adaptive importance sampling with automatic model selection in value function approximation. In: AAAI, pp. 1351–1356 (2008). URL http://www.aaai.org/Papers/AAAI/2008/AAAI08-214.pdf

Held, L., Schrdle, B., Rue, H.: Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In: Kneib, T., Tutz, G. (eds.) Statistical Modelling and Regression Structures, pp. 91–110. Physica-Verlag HD (2010). URL http://link.springer.com/chapter/10.1007/978-3-7908-2413-1_6

Jacquier, E., Polson, N.G., Rossi, P.E.: Bayesian analysis of stochastic volatility models. Journal of Business & Economic Statistics **20**(1), 69–87 (2002)

Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**(430), 773–795 (1995). doi:10.1080/01621459.1995.10476572. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

Li, Y., Yu, J.: Bayesian hypothesis testing in latent variable models. Journal of Econometrics **166**(2), 237–246 (2012). doi:10.1016/j.jeconom.2011.09.040. URL http://www.sciencedirect.com/science/article/pii/S0304407611002211

Li, Y., Zeng, T., Yu, J.: Robust deviance information criterion for latent variable models. SMU economics & statistics working paper series (2012). URL http://ink.library.smu.edu.sg/soe_research/1403/?utm_source=ink.library.smu.edu.sg%2Fsoe_research%2F1403&utm_medium=PDF&utm_campaign=PDFCoverPages

Li, Y., Zeng, T., Yu, J.: A new approach to bayesian hypothesis testing. Journal of Econometrics **178**(3), 602–612 (2014). URL http://www.sciencedirect.com/science/article/pii/S0304407613001991

Lindley, D.V.: A statistical paradox. Biometrika **44**(1/2), 187–187 (1957)

Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer-Verlag (2001)

Marshall, E.C., Spiegelhalter, D.J.: Approximate cross-validatory predictive checks in disease mapping models. Stat. Med. **22**(10), 1649–1660 (2003)

Marshall, E.C., Spiegelhalter, D.J.: Identifying outliers in bayesian hierarchical models: a simulation-based approach. Bayesian Analysis **2**(2), 409–444 (2007)

Neal, R.M.: Probabilistic inference using markov chain monte carlo methods. Tech. rep., Dept. of Computer Science, University of Toronto (1993)

Neal, R.M.: Bayesian learning for neural networks. Ph.D. thesis, University of Toronto (1995). URL http://www.db.toronto.edu/~radford/ftp/thesis.pdf

O'Hagan, A.: Fractional bayes factors for model comparison. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 99–138 (1995). URL http://www.jstor.org/stable/2346088

O'Hagan, A.: Properties of intrinsic and fractional bayes factors. Test **6**(1), 101–118 (1997). URL http://link.springer.com/article/10.1007/BF02564428

Peruggia, M.: On the variability of case-deletion importance sampling weights in the bayesian linear model. JASA **92**(437), 199–207 (1997)

Plummer, M.: Jags: A program for analysis of bayesian graphical models using gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March, pp. 20–22 (2003)

Plummer, M.: Penalized loss functions for bayesian model comparison. Biostatistics **9**(3), 523–539 (2008)

Postman, M., Huchra, J.P., Geller, M.J.: Probes of large-scale structure in the corona borealis region. The Astronomical Journal **92**, 1238–1247 (1986)

Raftery, A., Newton, M., Satagopan, J., Krivitsky, P.: Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series (2006). URL http://biostats.bepress.com/mskccbiostat/paper6

Robert, C.P.: On the Jeffreys-Lindley's paradox. ArXiv **1303.5973v**, 1–13 (2013)

Roeder, K.: Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. JASA **85**(411), 617–624 (1990)

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit. JRSSB **64**(4), 583–639 (2002)

Stern, H.S., Cressie, N.: Posterior predictive model checks for disease mapping models. Statistics in medicine **19**(17-18), 2377–2397 (2000). URL http://onlinelibrary.wiley.com/doi/10.1002/1097-0258(20000915/30)19:17/18%3C2377::AID-SIM576%3E3.0.CO;2-1/abstract

Vanhatalo, J., Riihimki, J., Hartikainen, J., Jylnki, P., Tolvanen, V., Vehtari, A.: Bayesian modeling with gaussian processes using the GPstuff toolbox. arXiv:1206.5754 [cs, stat] (2012). URL http://arxiv.org/abs/1206.5754

Vanhatalo, J., Riihimki, J., Hartikainen, J., Jylnki, P., Tolvanen, V., Vehtari, A.: GPstuff: bayesian modeling with gaussian processes. The Journal of Machine Learning Research **14**(1), 1175–1179 (2013). URL http://dl.acm.org/citation.cfm?id=2502617

Vehtari, A.: Bayesian model assessment and selection using expected utilities. Ph.D. thesis, HELSINKI UNIVERSITY OF TECHNOLOGY (2001)

Vehtari, A., Lampinen, J.: Bayesian model assessment and comparison using cross-validation predictive densities. Neural Comput. **14**(10), 2439–2468 (2002)

Vehtari, A., Ojanen, J.: A survey of bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys **6**, 142–228 (2012)

Waller, L.A., Carlin, B.P., Xia, H., Gelfand, A.E.: Hierarchical Spatio-Temporal mapping of disease rates. JASA **92**(438), 607–617 (1997)

Wang, W., Gelman, A.: Difficulty of selecting among multilevel models using predictive accuracy. Working paper (2014). URL http://www.stat.columbia.edu/~gelman/research/unpublished/xval.pdf

Watanabe, S.: Algebraic geometry and statistical learning theory. Cambridge University Press (2009)

Watanabe, S.: Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research **11**, 3571–3594 (2010a)

Watanabe, S.: Equations of states in singular statistical estimation. Neural Networks **23**(1), 20–34 (2010b)

Watanabe, S.: Equations of states in statistical learning for an unrealizable and regular case. IEICE transactions on fundamentals of electronics, communications and computer sciences **93**(3), 617–626 (2010c)