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

Longhai Li

Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, SK, CANADA

3 April 2014

Acknowledgements

- Joint work with **Shi Qiu, Bei Zhang and Cindy X. Feng**.

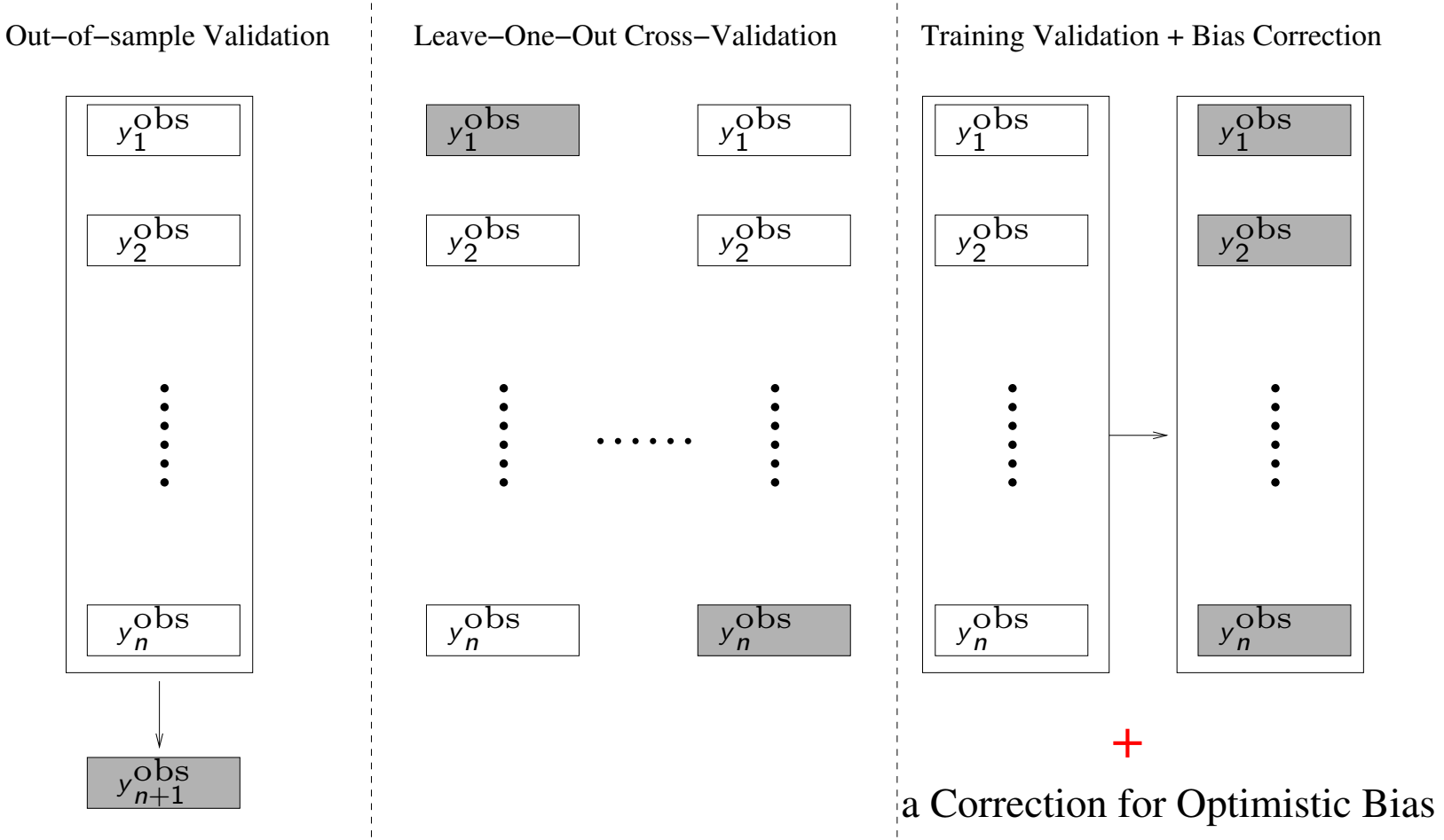
- The work was supported by grants from Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Foundation for Innovation (CFI).

- Thank Dr. Yao and Dr. Du for their warm hosting of my visit to Kansas State University.

Outline

Approximations for Out-of-Sample Predictive Evaluation

Predictive evaluation is often used for model comparison, diagnostics, and detecting outliers in practice. There are three ways for this with their own advantages and limitations:



Reviews of Bias-corrected Training Validation

1. AIC, DIC and others (eg., Spiegelhalter et al. (2002), Celeux et al. (2006), Plummer (2008), and Ando (2007)). Particularly,

$$\text{DIC} = -2\left(\log P(y^{\text{obs}}|\hat{\theta}) - p_{\text{DIC}}\right), \text{where,} \quad (1)$$

$$p_{\text{DIC}} = 2[\log P(y^{\text{obs}}|\hat{\theta}) - E_{\text{post}}\left(\log(P(y^{\text{obs}}|\theta))\right)] \quad (2)$$

   Good for models with identifiable parameters.

2. Importance Sampling (eg. Gelfand et al. (1992)). For each unit:

$$P(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = 1/E_{\text{post}}\left(1/P(y_i^{\text{obs}}|\theta)\right) \quad (3)$$

3. Widely Applicable Information Criterion (WAIC, proposed by Watanabe (2009)). For each unit:

$$\widehat{P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})} = E_{\text{post}}(P(y_i^{\text{obs}}|\theta))/\exp\left[V_{\text{post}}\left(\log(P(y_i^{\text{obs}}|\theta))\right)\right] \quad (4)$$

   Applicable to models with non-identifiable parameters, but not to models with correlated units.

What Will We Propose?

Two improved methods (namely iIS, and iWAIC) inspired by importance sampling formulae for Bayesian models with unit-specific latent variables that may be correlated.

Bayesian Models with Unit-specific Latent Variables

The two methods to be proposed aim at improving IS and WAIC evaluation for such models:

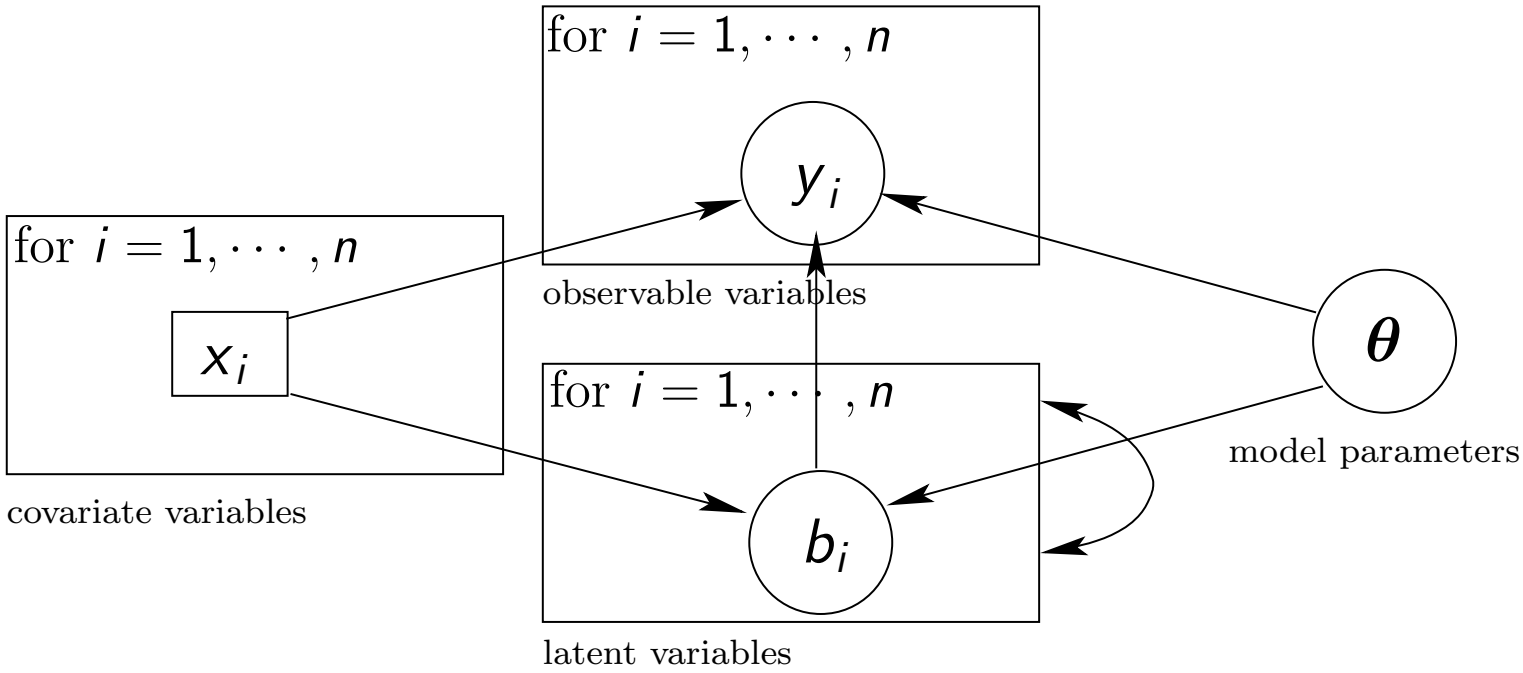


Figure 1: Graphical representation. The double arrows in the box for $b_{1:n}$ mean possible dependency between $b_{1:n}$. Note that the covariate x_i will be omitted in the conditions of densities for b_i and y_i throughout this paper for simplicity.

Posterior Distribution Given Full Data

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

$$P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y_{1:n}^{\text{obs}}) = \prod_{j=1}^{n} P(y_j^{\text{obs}}|b_j, \boldsymbol{\theta}) P(b_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) / C_1, \tag{5}$$

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

CV Posterior Distributions

- To do cross-validation, for each $i = 1, \ldots, n$, we omit observation y_i^{obs}, and then draw MCMC samples from **CV posterior distribution**:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}}) = \prod_{j \neq i} P(y_j^{\text{obs}}|b_j, \boldsymbol{\theta}) P(b_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \qquad (6)$$

- If we drop b_i from samples of $(\boldsymbol{\theta}, b_{1:n}) \sim (6)$, we obtain samples of $(\boldsymbol{\theta}, b_{-i})$ from the **marginalized CV posterior**:

$$P_{\text{post(-i), M}}(\boldsymbol{\theta}, b_{-i}|y_{-i}^{\text{obs}}) = \prod_{j \neq i} P(y_j^{\text{obs}}|b_j, \boldsymbol{\theta}) P(b_{-i}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \qquad (7)$$

  where $P(b_{-i}|\boldsymbol{\theta}) = \int P(b_{1:n}|\boldsymbol{\theta}) db_i$.
- It is useful to note that

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}}) = P_{\text{post(-i), M}}(\boldsymbol{\theta}, b_{-i}|y_{-i}^{\text{obs}}) P(b_i|b_{-i}, \boldsymbol{\theta}) \qquad (8)$$

  Sampling $P_{\text{post(-i)}}$= sampling $P_{\text{post(-i), M}}$ + drawing $b_i \sim P(b_i|b_{-i}, \boldsymbol{\theta})$.

CV Posterior Predictive Evaluation: General

Suppose we specify an evaluation function $a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ that measures certain goodness-of-fit (or discrepancy) of the distribution $P(y_i|\boldsymbol{\theta}, b_i)$ to the actual observation y_i^{obs} .

CV posterior predictive evaluation is defined as the expectation of the $a(y_{1:n}^{\text{obs}}, ., .)$ with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}})$ given in equations (8) or (6):

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) = \int a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}}) d\boldsymbol{\theta} db_{1:n} \quad (9)$$

CV Posterior Predictive Evaluation: Two Specific Cases

1. Let a be the value of predictive density:

$$a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) = P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i). \tag{10}$$

Then

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) = P(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) \tag{11}$$

We call it **CV posterior predictive density** for the held-out unit y_i^{obs}. **CV information criterion** (CVIC) for evaluating a Bayesian model is:

$$\text{CVIC} = -2\sum_{i=1}^{n} \log(P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})). \tag{12}$$

2. Let a be a tail probability:

$$a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) = Pr(y_i > y_i^{\text{obs}}|\boldsymbol{\theta}, b_i) + 0.5Pr(y_i = y_i^{\text{obs}}|\boldsymbol{\theta}, b_i), \tag{13}$$

Then,
$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) = Pr(y_i > y_i^{\text{obs}}|y_{-i}^{\text{obs}}) + 0.5Pr(y_i = y_i^{\text{obs}}|y_{-i}^{\text{obs}})$. We call it **CV posterior p-value**.

Non-integrated IS (nIS) Approximation: General

If our samples are from $P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y^{\text{obs}}_{1:n})$, but we are interested in estimating the mean of a with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y^{\text{obs}}_{-i})$ as in (9), importance weighting method is based on the following equality for CV expected evaluation:

$$E_{\text{post(-i)}}(a(y^{\text{obs}}_i, \boldsymbol{\theta}, b_i)) = \frac{E_{\text{post}}\left[a(y^{\text{obs}}_i, \boldsymbol{\theta}, b_i) W^{\text{nIS}}_i(\boldsymbol{\theta}, b_{1:n})\right]}{E_{\text{post}}\left[W^{\text{nIS}}_i(\boldsymbol{\theta}, b_{1:n})\right]}, \tag{14}$$

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

$$W^{\text{nIS}}_i(\boldsymbol{\theta}, b_{1:n}) = \frac{P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y^{\text{obs}}_{-i})}{P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y^{\text{obs}}_{1:n})} \times \frac{C_2}{C_1} = \frac{1}{P(y^{\text{obs}}_i|\boldsymbol{\theta}, b_i)}. \tag{15}$$

Gelfand et al. (1992) may be the first to propose this method.

nIS Estimate of CVIC

To estimate CVIC, we set $a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) = P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)$, the CV posterior predictive density $P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})$ is equal to harmonic mean of the non-integrated predictive density $P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)$ with respect to $P(\boldsymbol{\theta}, b_{1:n}|y_{1:n}^{\text{obs}})$:

$$P(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{1}{E_{\text{post}}\left[1/P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)\right]}. \tag{16}$$

Based on (16), **nIS** estimates the CV posterior predictive density by:

$$\hat{P}^{\text{nIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post}}\left[1/P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)\right]}. \tag{17}$$

The corresponding nIS estimate of CVIC using (17) is

$$\widehat{\text{CVIC}}^{\text{nIS}} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{nIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})) \tag{18}$$

However, nIS often doesn't work well because b_i fits y_i^{obs} too well.

Theory for Integrated Importance Sampling (iIS): I

1. Integrated Evaluation Function
   Rewrite the expectation in (9) as

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) = E_{\text{post(-i), M}}(A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i})) \tag{19}$$

$$= \int\int A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i}) P(\boldsymbol{\theta}, b_{-i} | y_{-i}^{\text{obs}}) d\boldsymbol{\theta} db_{-i} \tag{20}$$

   where,

$$A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i}) = \int a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) P(b_i | b_{-i}, \boldsymbol{\theta}) db_i. \tag{21}$$

   Note: In (21), we integrate $a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ with respect to $P(b_i | b_{-i}, \boldsymbol{\theta})$, which is unconditional on y_i^{obs}.

Theory for Integrated Importance Sampling (iIS): II

2. Integrated Predictive Density
   The *full data* posterior of $(\boldsymbol{\theta}, b_{-i})$ is

$$P_{\text{post, M}}(\boldsymbol{\theta}, b_{-i}|y_{-i}^{\text{obs}}) = \Big[\prod_{j\neq i} P(y_j^{\text{obs}}|b_j, \boldsymbol{\theta})P(b_{-i}|\boldsymbol{\theta})P(\boldsymbol{\theta})\Big] P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})/C_1, \tag{22}$$

   where,

$$P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i}) = \int P(y_i^{\text{obs}}|b_i, \boldsymbol{\theta})P(b_i|b_{-i}, \boldsymbol{\theta})db_i. \tag{23}$$

   We will call (23) **integrated predictive density**, because it integrates away b_i without reference to y_i^{obs}.

Theory for Integrated Importance Sampling (iIS): III

3. Integrated Importance Sampling Formula
   Using the standard importance weighting method, we will estimate (20) by

$$E_{\text{post(-i), M}}(A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i})) = \frac{E_{\text{post, M}}\left[A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i})\ W_i^{\text{iIS}}(\boldsymbol{\theta}, b_{-i})\right]}{E_{\text{post, M}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, b_{-i})\right]}, \tag{24}$$

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

$$W_i^{\text{iIS}}(\boldsymbol{\theta}, b_{-i}) = \frac{P_{\text{post(-i), M}}(\boldsymbol{\theta}, b_{-i}|y_{-i}^{\text{obs}})}{P_{\text{post, M}}(\boldsymbol{\theta}, b_{-i}|y_{-i}^{\text{obs}})} \times \frac{C_2}{C_1} = \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})}. \tag{25}$$

In summary, in iIS, we integrate the evaluation function $a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ and $P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)$ over b_i drawn from $P(b_i|b_{-i}, \boldsymbol{\theta})$, which is unconditional on y_i^{obs}, to find integrated evaluation and predictive density functions.

iIS Estimate for CVIC

In the special case of estimating CVIC, the evaluation function a is just the predictive density $P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)$, therefore, A is just reciprocal of W_i^{iIS}. Therefore, the iIS estimate for $P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})$ is

$$\hat{P}^{\text{iIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post, M}}\left[1/P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})\right]}. \tag{26}$$

Accordingly, iIS estimate of CVIC is

$$\widehat{\text{CVIC}}^{\text{iIS}} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{iIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})) \tag{27}$$

WAIC for Models without Latent Variables

Watanabe (2009) defines a version of WAIC for models without latent variables as follows:

$$\text{WAIC} = -2 \sum_{i=1}^{n} \left[\log(E_{\text{post}}(P(y_i^{\text{obs}}|\boldsymbol{\theta}))) - V_{\text{post}}(\log(P(y_i^{\text{obs}}|\boldsymbol{\theta}))) \right], \quad (28)$$

where E_{post} and V_{post} stand for mean and variance over $\boldsymbol{\theta}$ with respect to $P(\boldsymbol{\theta}|y_1^{\text{obs}}, \ldots, y_n^{\text{obs}})$. By comparing the forms of WAIC and CVIC, we can think of that in WAIC, the CV posterior predictive density is estimated by:

$$\hat{P}^{\text{WAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \exp\left\{ \log(E_{\text{post}}(P(y_i^{\text{obs}}|\boldsymbol{\theta}))) - V_{\text{post}}(\log(P(y_i^{\text{obs}}|\boldsymbol{\theta}))) \right\}. \quad (29)$$

nWAIC for Latent Variables Models

For the models with possibly correlated latent variables, a naive way to approximate CVIC is to apply WAIC directly to the non-integrated predictive density of y_i^{obs} conditional on $\boldsymbol{\theta}$ and b_i:

$$\hat{P}^{\text{nWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \exp\left\{ \log(E_{\text{post}}(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i))) - V_{\text{post}}(\log(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)))\right\}. \tag{30}$$

We will refer to (30) as non-integrated WAIC (or nWAIC for short) method for approximating CV posterior predictive density. The corresponding information criterion based on (30) is:

$$\text{nWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{nWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})). \tag{31}$$

iWAIC for Latent Variables Models

Using heuristics, we propose to apply WAIC approximation to the integrated predictive density (23) to estimate the CV posterior predictive density:

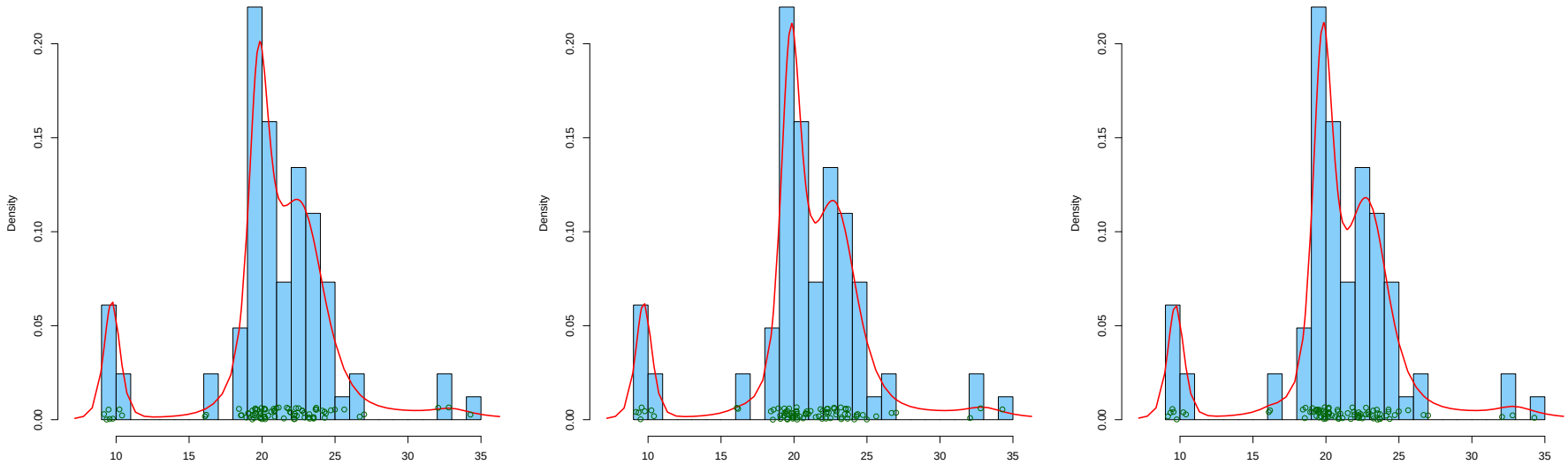
$$\hat{P}^{\text{iWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \exp\left\{ \log(E_{\text{post}}(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i}))) - V_{\text{post}}(\log(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})))\right\}. \tag{32}$$

Accordingly, iWAIC for approximating CVIC is given by :

$$\text{iWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{iWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})). \tag{33}$$

Galaxy Data

We obtained the data set from 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.



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

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.

Mixture Models with a Fixed Number, K, of Components

We applied mixture models to fit the 82 numbers. The finite mixture model that we used to fit Galaxy data is as follows:

$$y_i|z_i = k, \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}_{1:K} \sim N(\mu_k, \sigma_k^2), \text{ for } i = 1, \ldots, n \quad (34)$$

$$z_i|p_{1:K} \sim \text{Category}(p_1, \ldots, p_K), \text{ for } i = 1, \ldots, n \quad (35)$$

$$\mu_k \sim N(20, 10^4), \text{ for } k = 1, \ldots, K \quad (36)$$

$$\sigma_k^2 \sim \text{Inverse-Gamma}(0.01, 0.01 \times 20), \text{ for } k = 1, \ldots, K \quad (37)$$

$$p_k \sim \text{Dirichlet}(1, \ldots, 1) \text{ for } k = 1, \ldots, K \quad (38)$$

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.

Our purpose of computing CVIC for finite mixture models is to determine the numbers of mixture components, K, that can adequately capture the heterogeneity in a data but don't overfit the data.

How Did We Run MCMC?

We used JAGS to run MCMC simulations for fitting the above model to `Galaxy` data with various choice of K. To avoid the problem that MCMC may get stuck in a model with only one component, we followed JAGS `eyes` example to restrict the MCMC to have at least a data point in each component.

All MCMC simulations started with a randomly generated $z_{1:n}$, and ran 5 parallel chains, each doing 2000, 2000, and 100,000 iterations for adapting, burning, and sampling, respectively.

The Mixture Model is a Latent Variable Model

The finite mixture model (equations (35) - (38)) falls in the class of models depicted by Figure 1:

- the observed variable is y_i,
- the latent variable b_i is the mixture component indicator z_i, and
- the model parameters $\boldsymbol{\theta}$ is $(\boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}^2_{1:K}, p_{1:K})$.

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

Computing nIS, iIS, nWAIC, iWAIC in Mixture Models

For each MCMC sample of $(\boldsymbol{\theta}, z_1, \ldots, z_n)$ and each unit i, we compute

- The non-integrated predictive density: $P(y_i^{\text{obs}}|z_i, \boldsymbol{\theta}) = \phi(y_i^{\text{obs}}|\mu_{z_i}, \sigma_{z_i})$.
- The integrated predictive density:

$$P(y_i^{\text{obs}}|\boldsymbol{\theta}, z_{-i}) = P(y_i^{\text{obs}}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \phi(y_i^{\text{obs}}|\mu_k, \sigma_k) \tag{39}$$

  Notes: 1) z_{-i} and y_i are independent given $\boldsymbol{\theta}$. 2) the large component, not the component close to y_i^{obs}, dominates (39)

Then we can compute nIS, iIS, nWAIC and iWAIC.

We see that, to compute iIS and iWAIC, we just apply IS and WAIC to the marginalized models with $z_{1:n}$ integrated out, although $z_{1:n}$ are included in MCMC simulations.

Comparison of 5 Information Criteria

Table 1: Comparison 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

Comparison of Statistical Significance

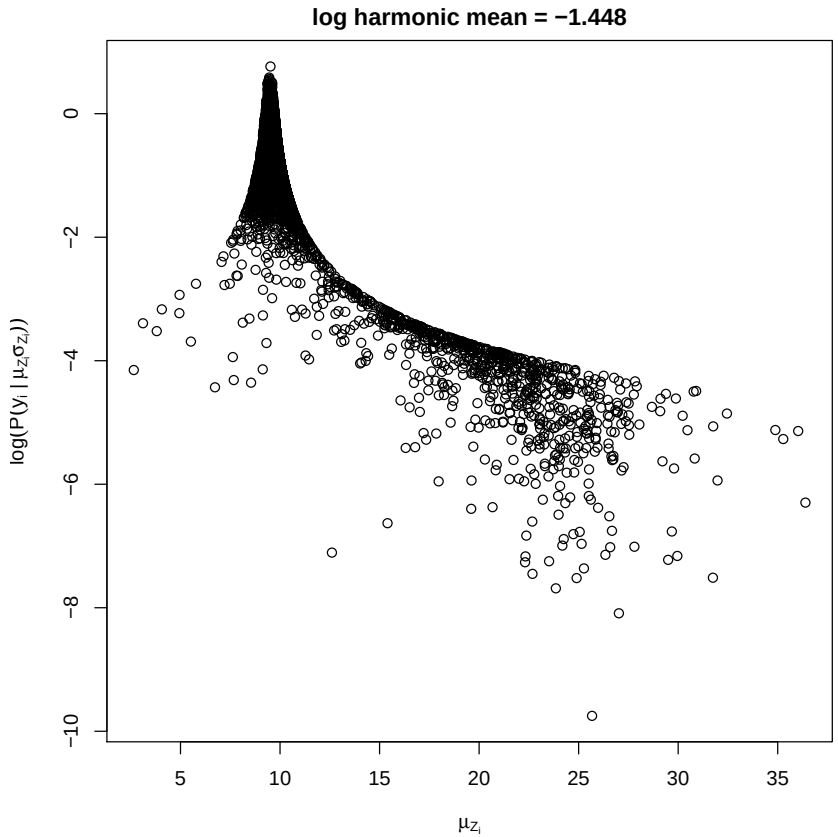
CVIC is the sum of minus twice of log CV posterior predictive densities. Therefore, the statistical significance of the differences of two CVICs (or estimates) can be accessed by looking at the population mean differences of two groups of log CV posterior predictive densities (or their estimates).

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.

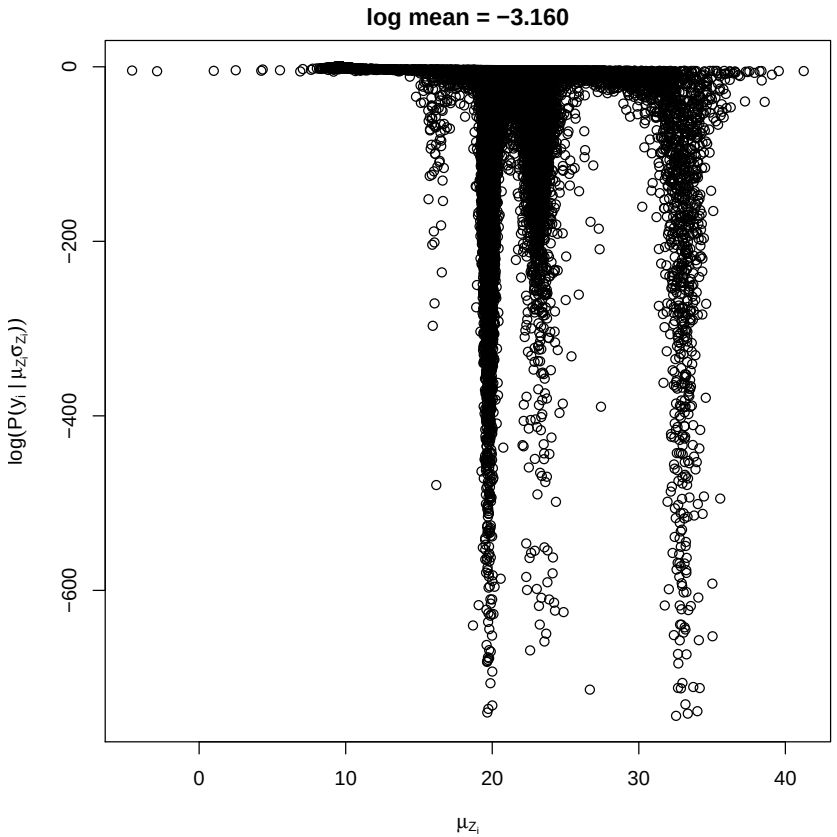
pair of models	nWAIC	nIS	iWAIC	iIS	CVIC
K =3 vs $K = 2$	0.000	0.000	0.016	0.013	0.010
$K = 4$ vs $K = 3$	0.000	0.019	0.030	0.032	0.190
$K = 5$ vs $K = 4$	0.000	0.249	0.070	0.066	0.027
$K = 6$ vs $K = 5$	0.002	0.203	0.489	0.476	0.674
$K = 7$ vs $K = 6$	0.110	0.840	0.716	0.711	0.700

Visualize the Need of Integrating z_i

Figure 3: Scatter-plot of non-integrated predictive densities against μ_{z_i}, given MCMC samples from the full data posterior (4a) and the actual CV posterior with the 3rd number removed (4b), when $K = 5$ components are used.



(a)



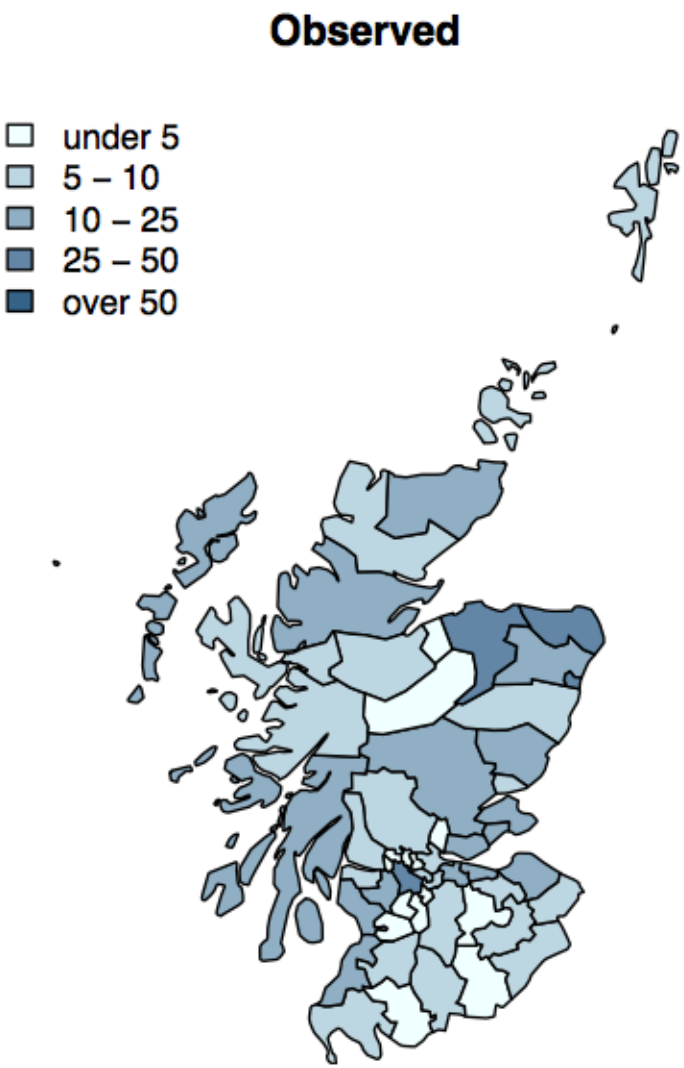
(b)

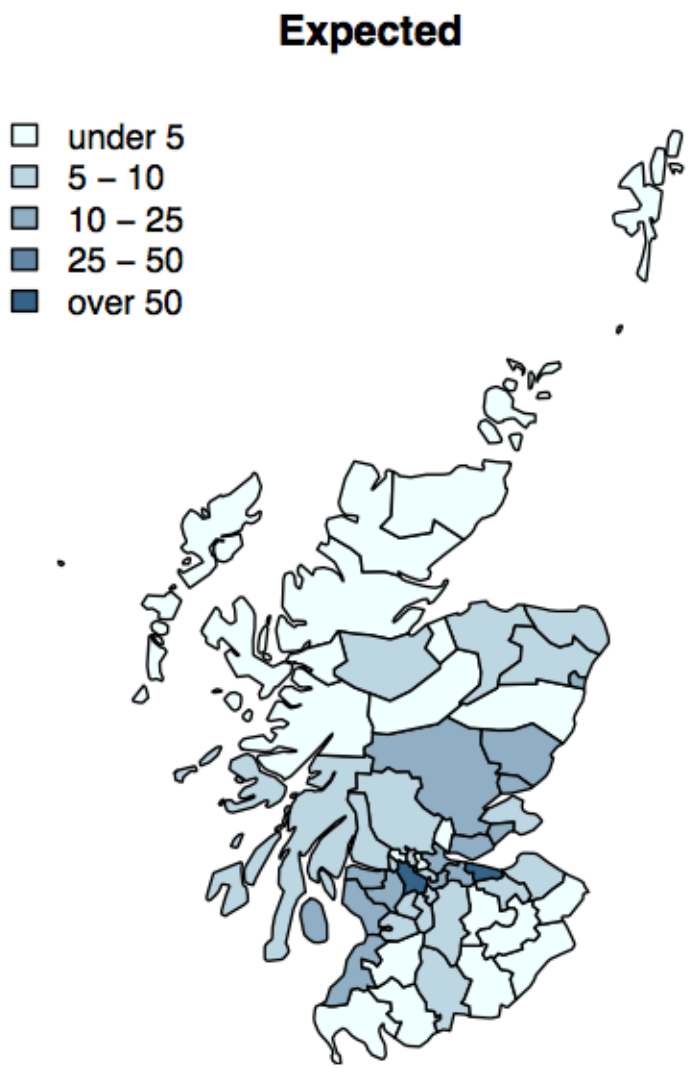
Scottish Lip Cancer Data I

The data represents male lip cancer counts (over the period 1975 - 1980) in the $n = 56$ districts of Scotland. The data includes these columns:

- the number of observed cases of lip cancer, y_i;
- the number of expected cases, E_i, which are based on age effects, and are proportional to a "population at risk" after such effects have been taken into account;
- the percent of population employed in agriculture, fishing and forestry, x_i, used as a covariate; and
- a list of the neighbouring regions.

Scottish Lip Cancer Data II





Four Models Considered: I

The y_i is modelled as a Poisson random variable:

$$y_i | E_i, \lambda_i \sim \text{Poisson}(\lambda_i E_i), \tag{40}$$

where λ_i denotes the underlying relative risk for district i.

Let $s_i = \log(\lambda_i)$. We consider four different models for the vector $s = (s_1, \cdots, s_n)'$:

$$\text{model 1 (spatial+linear, full)} : s \sim N_n(\alpha + X\beta, \Phi\tau^2), \tag{41}$$
$$\text{model 2 (spatial)} : s \sim N_n(\alpha, \Phi\tau^2), \tag{42}$$
$$\text{model 3 (linear)} : s \sim N_n(\alpha + X\beta, I_n\tau^2), \tag{43}$$
$$\text{model 4 (exchangable)} : s \sim N_n(\alpha, I_n\tau^2), \tag{44}$$

where Φ specify spatial association between districts, with details follow.

Four Models Considered: II

$\Phi = (I_n - \phi C)^{-1} M$ is a matrix modelling spatial dependency, in which, $c_{ij} = (E_j/E_i)^{1/2}$ if areas i and j are neighbours, equals to 0 otherwise, $m_{ii} = E_i^{-1}$ and $m_{ij} = 0$ if $i \neq j$. This model is called **proper conditional auto regression (CAR) model**. Looking at the conditional distribution of $s_i|s_{-i}, \alpha, \beta, \phi$ (48) may help understand this distribution.

At a higher level, we assign $\beta, \tau, and \phi$ with very diffuse prior:

$$\begin{aligned} \tau^2 &\sim \text{Inv-Gamma}(0.5, 0.0005) && (45) \\ \beta &\sim N(0, 1000^2) && (46) \\ \phi &\sim \text{Unif}(\phi_0, \phi_1), && (47) \end{aligned}$$

where (ϕ_0, ϕ_1) is the interval for ϕ such that Φ is positive-definite.

How Did we Run MCMC?

We used OpenBUGS through R package `R2OpenBUGS` to run MCMC simulations for fitting the above four models to lip cancer data. For each simulation, we ran two parrallel chains, each for 15000 iterations, and the first 5000 were discarded as burning.

For replicating computing information criterion (with each method), we ran 100 independent simulations as above by randomizing initial $\boldsymbol{\theta}$ and randomizing bugs random seed for OpenBUGS.

The Poisson Model is a Latent Variable Model

- the observed variable is y_i,
- the latent variable b_i is s_i (or λ_i)
- the model parameters $\boldsymbol{\theta}$ is (τ, β, ϕ).

In models 1 and 2, the latent variables $s_1, \ldots, s_n$ are *dependent* given the model parameter $\boldsymbol{\theta}$.

Computing iIS and iWAIC in Model 1

For each unit i, and for each MCMC sample of $(s, \boldsymbol{\theta})$:

- Conditional distribution (proper auto regression):

$$s_i | s_{-i}, \boldsymbol{\theta} \sim N(\alpha + x_i\beta + \phi \sum_{j \in N_i} (c_{ij}(s_j - \alpha - x_j\beta)), \tau^2 m_{ii}), \tag{48}$$

  where N_i is the set of neighbours of district i.

- Integrated predictive density:

$$P(y_i^{\text{obs}} \,|\, \boldsymbol{\theta}, s_{-i}) = \int \text{dpoisson}(y_i^{\text{obs}} | \lambda_i E_i) P(s_i \,|\, \boldsymbol{\theta}, s_{-i}) ds_i \tag{49}$$

  We generate 200 random numbers of s_i from the distribution (48), and then estimate the integral in (49).

Then we can compute iIS and iWAIC with the integrated predictive density.

Comparison of 5 Information Criteria

Table 3: Comparisons of information criteria for lip cancer data. Each table entry shows the average of 100 information criteria computed from 100 independent MCMC simulations, and the standard deviation in bracket.

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

Seeds Data

The example is taken from Table 3 of Crowder (1978). The study concerns about 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).

Logistic Regression Models with Random Effects

The conditional distribution of r_i given n_i, x_{i1} and x_{i2} are specified as follows:

$$
\begin{aligned}
r_i | n_i, p_i &\sim \text{Binomial}(n_i, p_i) && (50)\\
\text{logit}(p_i) &= \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i && (51)\\
b_i &\sim N(0, \sigma^2), && (52)
\end{aligned}
$$

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.

CV Posterior p-value for Outlier Detection

The p-value (given parameters and latent variable) defined by (13) for this example is the right tail probability of Binomial distribution with number of trials n_i and success rate p_i:

$$\text{p-value}(r_i^{\text{obs}}, \boldsymbol{\theta}, b_i) = 1 - \text{pbinom}(r_i^{\text{obs}}; n_i, p_i) + 0.5\,\text{dbinom}(r_i^{\text{obs}}; n_i, p_i), \quad (53)$$

where r_i^{obs} is the actual observation of r_i, and pbinom and dbinom denote CDF and PMF of Binomial distribution.

CV posterior p-value for 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 $P(\boldsymbol{\theta}, b_i | r_{-i}^{\text{obs}})$.

Four Methods for Approximating CV Posterior p-value: I

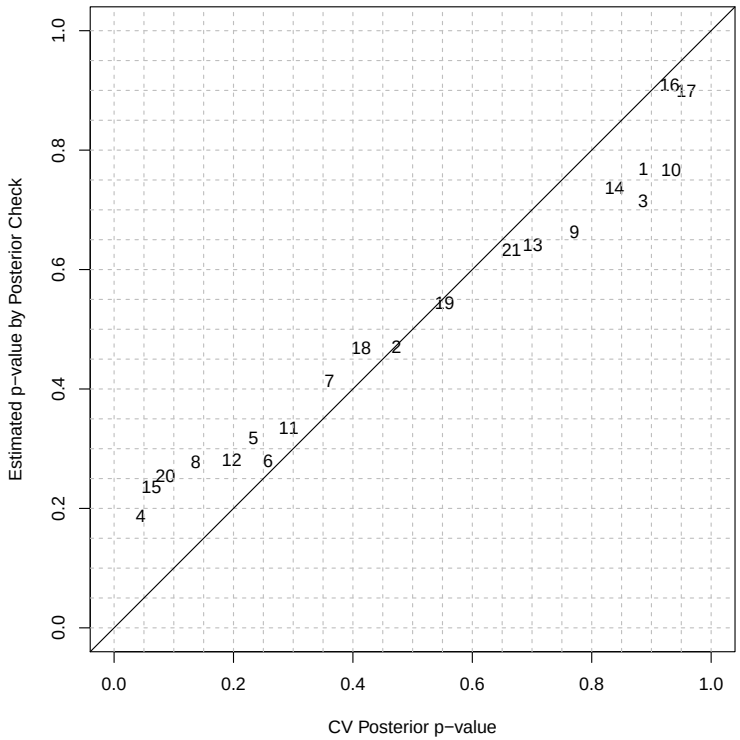
- Posterior check (Gelman et al. 1996)
  Average each p-value($r_i^{\text{obs}}, \boldsymbol{\theta}, b_i$) with respect to the posterior of ($\boldsymbol{\theta}, b_i$) given the full data set $r_{1:21}^{\text{obs}}$.
- Ghosting method (Marshall and Spiegelhater, 2003)
  For each MCMC sample, one averages 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 ghosting p-value, then averages the ghosting p-values over all MCMC samples.
- nIS:
  Average p-value($r_i^{\text{obs}}, \boldsymbol{\theta}, b_i$) after being weighted with the inverse of probability density (mass) of r_i^{obs}: $1/\text{dbinom}(r_i^{\text{obs}}; n_i, p_i)$

Four Methods for Approximating CV Posterior p-value: II

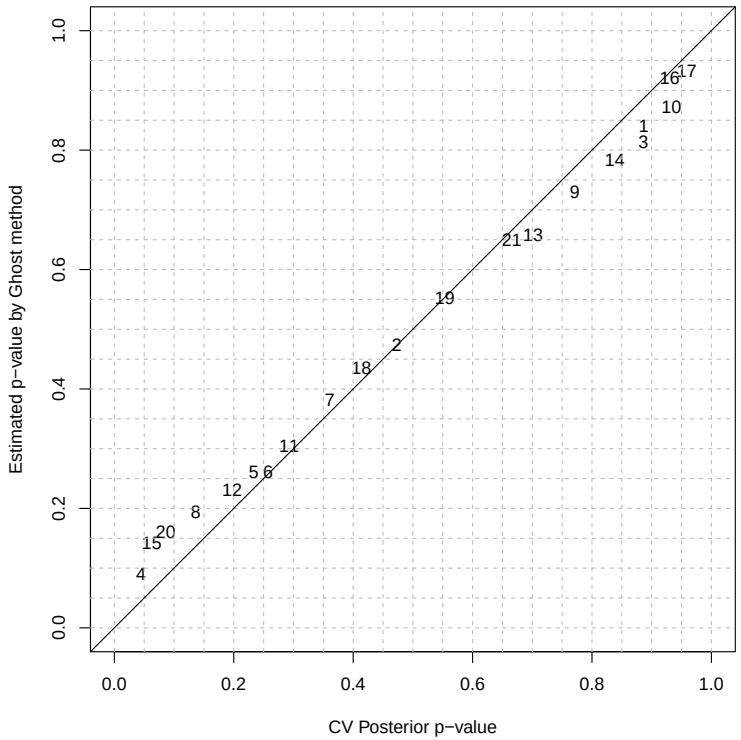
- iIS:
  For each MCMC sample, we first average each of p-value$(r_i^{\text{obs}}, \theta, b_i)$ and dbinom$(r_i^{\text{obs}}; n_i, p_i)$ **over 30** b_i **randomly generated from** $b_i|\theta$ to find the integrated p-value and the integrated predictive density respectively.
  Then compute the weighted average of the integrated p-values with the reversed integrated predictive density as weights over all MCMC samples using formula (24).

Comparison of Estimated CV Posterior p-value I

Figure 4: Scatterplots of estimated posterior p-values from an MCMC simulation against actual CV posterior p-values. The number for points show indices of plates

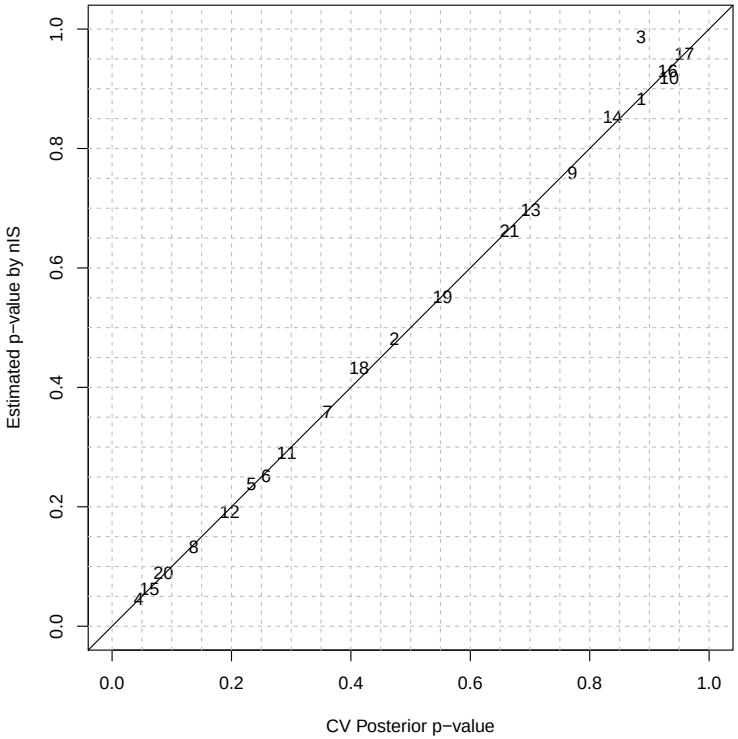


(a) Posterior checking

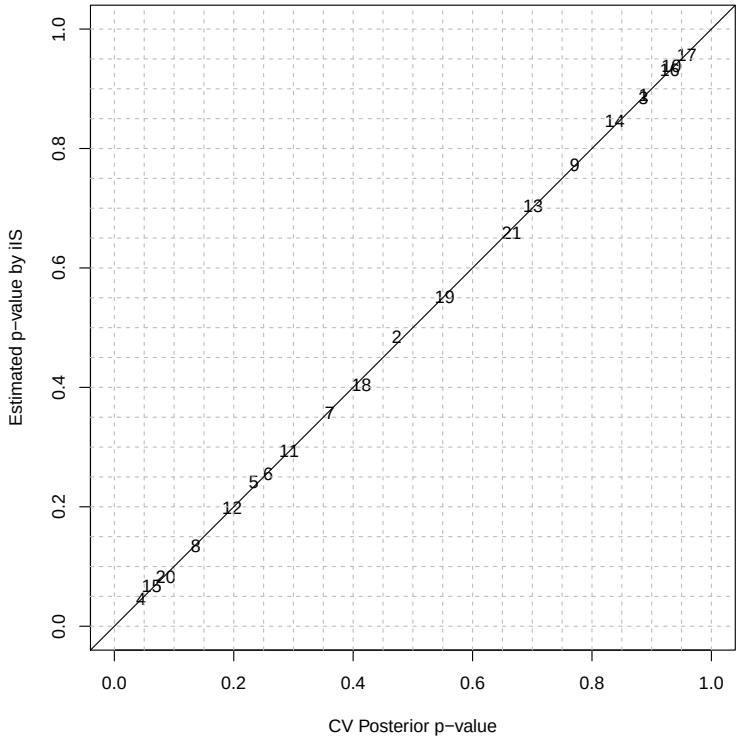


(b) Ghosting method

Comparison of Estimated CV Posterior p-value II



(c) Non-integrated IS (nIS)



(d) Integrated IS (iIS)

Replication Study

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

$$\text{RE} = (1/n)\sum_{i=1}^{n}\frac{|\hat{p}_i - p_i|}{\min(p_i,\ 1-p_i)} \times 100, \tag{54}$$

where $\hat{p}_{1:n}$ are estimates of $p_{1:n}$. This measure emphasizes greatly on 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.

Table 4: 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)

Conclusions and Future Work

- The new proposed iIS and iWAIC significantly reduce the bias of nIS and nWAIC in evaluating Bayesian models with unit-specific latent variables. In our studies, they gave results very close to what given by the actual cross-validation.
- iWAIC works very well in the spatial random effect models. The result is surprising and encouraging. One may consider investigating the validity of iWAIC theoretically.
- iIS and iWAIC are limited to Bayesian model with *unit-specific* latent variables. In many models, a latent variable is shared by multiple units. How to improve IS and WAIC for such models?
- Advancing applications to other models. An interesting model is auto logistic regression model for spatial data, where the model is defined with conditional distribution only. I will investigate the applicability of iIS, iWAIC or cross-validation itself for comparison and diagnostics of such models.

References

To read more of this topic, the following is a short list of references:

- Vehtari, A. and Ojanen, J. (2012), "A survey of Bayesian predictive methods for model assessment, selection and comparison," Statistics Surveys, 6, 142-228.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit," JRSSB, 64, 583-639.
- Watanabe, S. (2009), "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," Journal of Machine Learning Research, 11, 3571-3594.
- Gelman, A., Hwang, J., and Vehtari, A. (2013), "Understanding predictive information criteria for Bayesian models," unpublished online manuscript, available from Gelman's website.

- The paper with more details about this talk can be found from: `http://math.usask.ca/~longhai/doc`.