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

27 May 2014, SSC Meeting at University of Toronto

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).

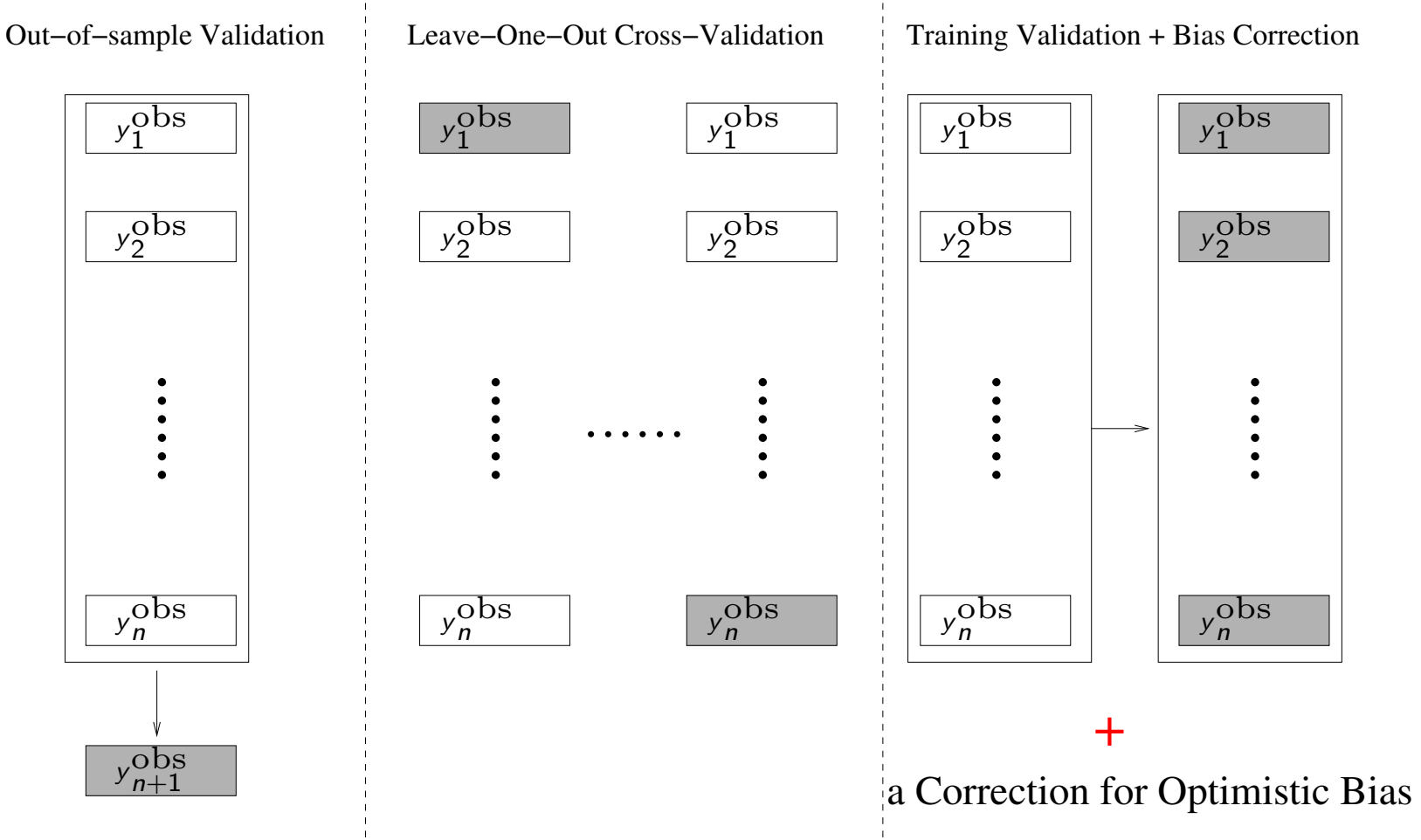
Outline

Section 1

Introduction of the Problem

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:

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

   Justified for models with non-identifiable parameters, but not yet for models with correlated units.

Section 2

Actual Bayesian Cross-validatory Evaluation

Bayesian Models with Unit-specific Latent Variables

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

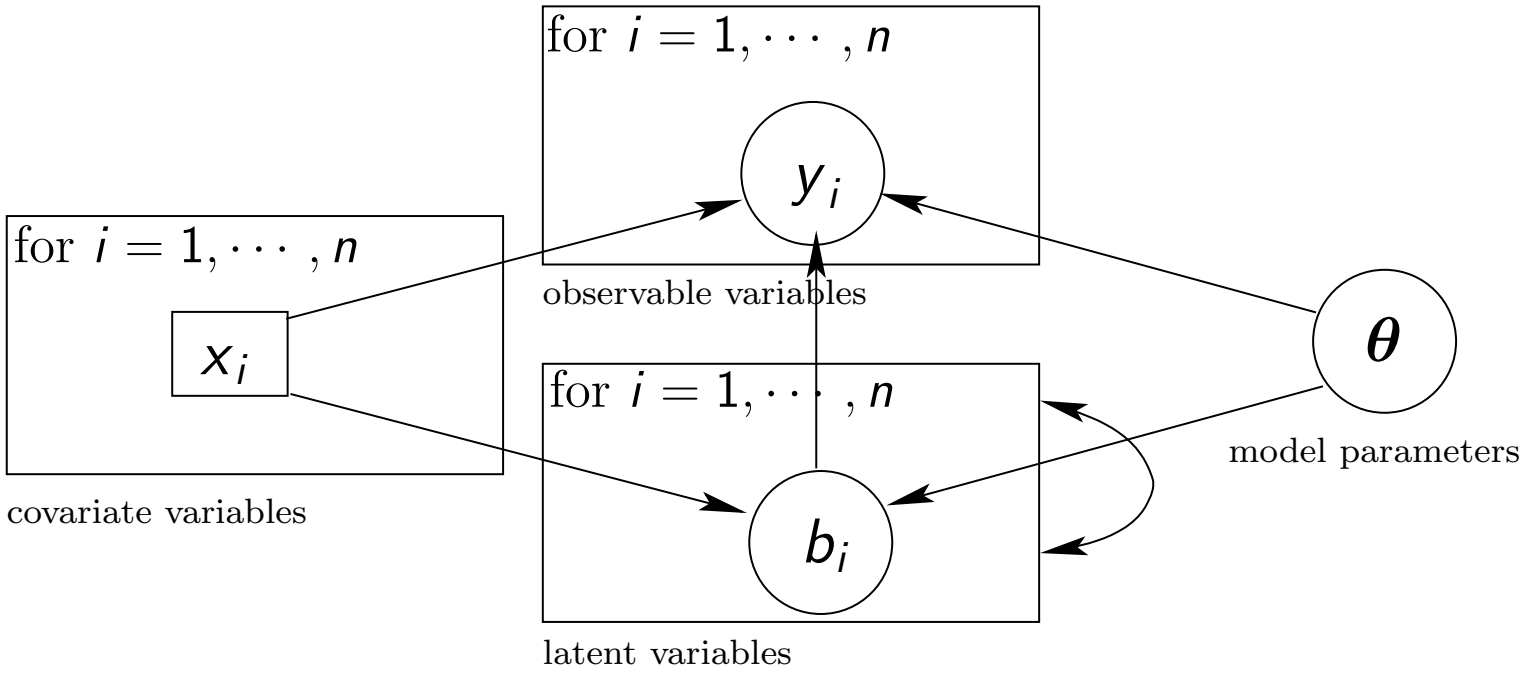


Figure 1: Graphical representation. The double arrows in the box for $b_{1:n}$ mean possible dependency between $b_{1:n}$. Note that the covariate x_i will be omitted in the conditions of densities for b_i and y_i throughout this paper for simplicity.

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**:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n} | y_{-i}^{\text{obs}}) = \prod_{j \neq i} P(y_j^{\text{obs}} | b_j, \boldsymbol{\theta}) P(b_{1:n} | \boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \qquad (5)$$

- If we drop b_i from samples of $(\boldsymbol{\theta}, b_{1:n}) \sim (5)$, we obtain samples of $(\boldsymbol{\theta}, b_{-i})$ from the **marginalized CV posterior**:

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

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

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n} | y_{-i}^{\text{obs}}) = P_{\text{post(-i), M}}(\boldsymbol{\theta}, b_{-i} | y_{-i}^{\text{obs}}) P(b_i | b_{-i}, \boldsymbol{\theta}) \qquad (7)$$

  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 (7) or (5):

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) = \int a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}}) d\boldsymbol{\theta} db_{1:n} \quad (8)$$

A Special Case: CV Information Criterion

Let a be the value of predictive density:

$$a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) = P(y_i^{\text{obs}} | \boldsymbol{\theta}, b_i). \tag{9}$$

Then

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) = P(y_i^{\text{obs}} | y_{-i}^{\text{obs}}) \tag{10}$$

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:

$$\text{CVIC} = -2 \sum_{i=1}^{n} \log(P(y_i^{\text{obs}} | y_{-i}^{\text{obs}})). \tag{11}$$

Section 3

Importance Sampling (IS) and WAIC Approximations

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 $\boldsymbol{\theta}$:

$$P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y_{1:n}^{\text{obs}}) = \prod_{j=1}^{n} P(y_j^{\text{obs}}|b_j, \boldsymbol{\theta}) P(b_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) / C_1, \tag{12}$$

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

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 (8), importance weighting method is based on the following equality for CV expected evaluation:

$$E_{\text{post(-i)}}(a(y^{\text{obs}}_i, \boldsymbol{\theta}, b_i)) = \frac{E_{\text{post}}\left[a(y^{\text{obs}}_i, \boldsymbol{\theta}, b_i) W^{\text{nIS}}_i(\boldsymbol{\theta}, b_{1:n})\right]}{E_{\text{post}}\left[W^{\text{nIS}}_i(\boldsymbol{\theta}, b_{1:n})\right]}, \tag{13}$$

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

$$W^{\text{nIS}}_i(\boldsymbol{\theta}, b_{1:n}) = \frac{P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y^{\text{obs}}_{-i})}{P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y^{\text{obs}}_{1:n})} \times \frac{C_2}{C_1} = \frac{1}{P(y^{\text{obs}}_i|\boldsymbol{\theta}, b_i)}. \tag{14}$$

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 $P(\boldsymbol{\theta}, b_{1:n}|y_{1:n}^{\text{obs}})$:

$$P(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{1}{E_{\text{post}}\left[1/P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)\right]}. \tag{15}$$

Based on (15), **nIS** estimates the CV posterior predictive density by:

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

The corresponding nIS estimate of CVIC using (16) is

$$\widehat{\text{CVIC}}^{\text{nIS}} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{nIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})) \tag{17}$$

Integrated Importance Sampling (iIS)

- nIS often does not work well, because:
  - MCMC sample of b_i from the full data posterior $P_{\text{post}}(\theta, b_{1:n}|y^{\text{obs}}_{1:n})$ fit y^{obs}_i so well because it receives information from y^{obs}_i.
  - In actual CV simulation, b_i does not get information from y^{obs}_i because it is omitted from the data.

  Therefore, $P(b_i|y^{\text{obs}}_{1:n})$ and $(b_i|y^{\text{obs}}_{-i})$ differ so much that the nIS estimate becomes inaccurate and unstable.
- Solution

  For each unit i, drop b_i *temporarily* from full data posterior sample, regenerate b_i without reference to y^{obs}_i, that is, from $P(b_i|b_{-i}, \theta)$.

iIS Formula

Using the standard importance weighting method, we will estimate (8) by

$$E_{\text{post(-i), M}}(A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i})) = \frac{E_{\text{post, M}}\left[A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i})\ W_i^{\text{iIS}}(\boldsymbol{\theta}, b_{-i})\right]}{E_{\text{post, M}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, b_{-i})\right]}, \quad (18)$$

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

$$W_i^{\text{iIS}}(\boldsymbol{\theta}, b_{-i}) = \frac{P_{\text{post(-i), M}}(\boldsymbol{\theta}, b_{-i}|y_{-i}^{\text{obs}})}{P_{\text{post, M}}(\boldsymbol{\theta}, b_{-i}|y_{-i}^{\text{obs}})} \times \frac{C_2}{C_1} = \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})}, \quad (19)$$

where $(A(y_i^{\text{obs}}, \boldsymbol{\theta}, b_{-i})$ and $P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})$ result from integrating 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 does not use y_i^{obs}, respectively.

A Special Case: 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

$$\hat{P}^{\text{iIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post, M}}\left[1/P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})\right]}. \tag{20}$$

Accordingly, iIS estimate of CVIC is

$$\widehat{\text{CVIC}}^{\text{iIS}} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{iIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})) \tag{21}$$

iWAIC for Latent Variables Models

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

$$\hat{P}^{\text{iWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{E_{\text{post}}(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i}))}{\exp\left\{V_{\text{post}}(\log(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})))\right\}}. \tag{22}$$

Accordingly, iWAIC for approximating CVIC is given by:

$$\text{iWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{iWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})). \tag{23}$$

Section 4

Real Data Examples

Subsection 1

Mixture Models

Galaxy Data

The data set is a numeric vector of velocities (km/sec) of 82 galaxies from 6 well-separated conic sections.

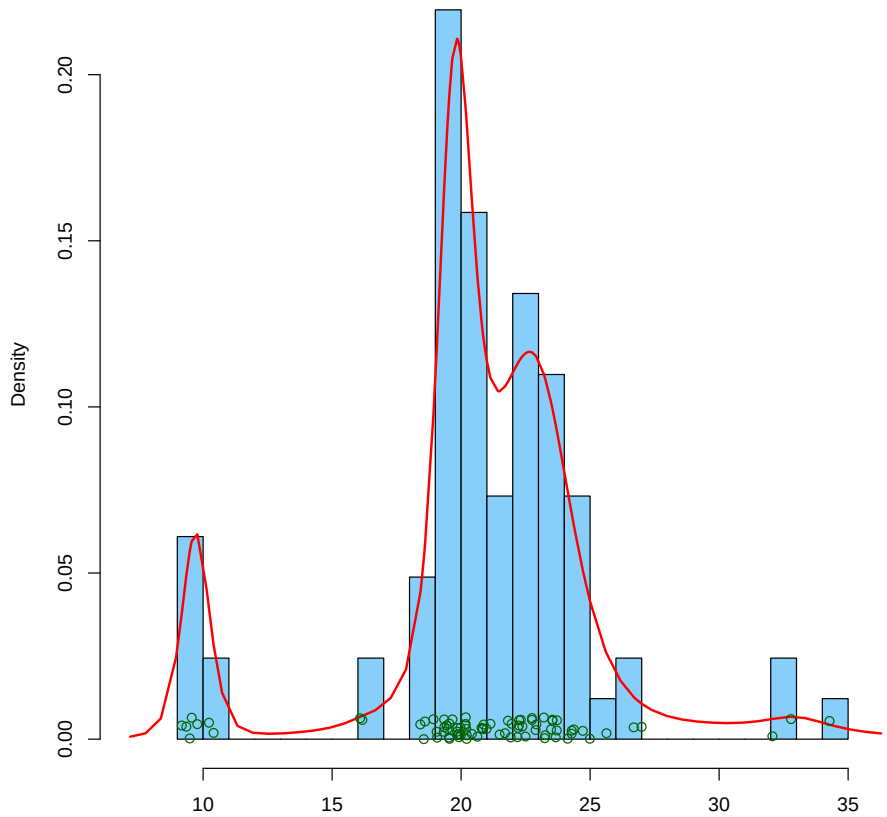


Figure 2: Histograms of Galaxy data and a estimated density curve.

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:

$$y_i|z_i = k, \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}_{1:K} \sim N(\mu_k, \sigma_k^2), \text{ for } i = 1, \ldots, n \quad (24)$$

$$z_i|p_{1:K} \sim \text{Category}(p_1, \ldots, p_K), \text{ for } i = 1, \ldots, n \quad (25)$$

$$\mu_k \sim N(20, 10^4), \text{ for } k = 1, \ldots, K \quad (26)$$

$$\sigma_k^2 \sim \text{Inverse-Gamma}(0.01, 0.01 \times 20), \text{ for } k = 1, \ldots, K \quad (27)$$

$$p_k \sim \text{Dirichlet}(1, \ldots, 1) \text{ for } k = 1, \ldots, K \quad (28)$$

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.

The Mixture Model is a Latent Variable Model

- 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 $(\boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}^2_{1:K}, p_{1:K})$.

Computing nIS, iIS, nWAIC, iWAIC in Mixture Models

- The non-integrated predictive density:

$$P(y_i^{\text{obs}}|z_i, \boldsymbol{\theta}) = \phi(y_i^{\text{obs}}|\mu_{z_i}, \sigma_{z_i}). \tag{29}$$

- The integrated predictive density:

$$P(y_i^{\text{obs}}|\boldsymbol{\theta}, z_{-i}) = P(y_i^{\text{obs}}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \phi(y_i^{\text{obs}}|\mu_k, \sigma_k) \tag{30}$$

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.

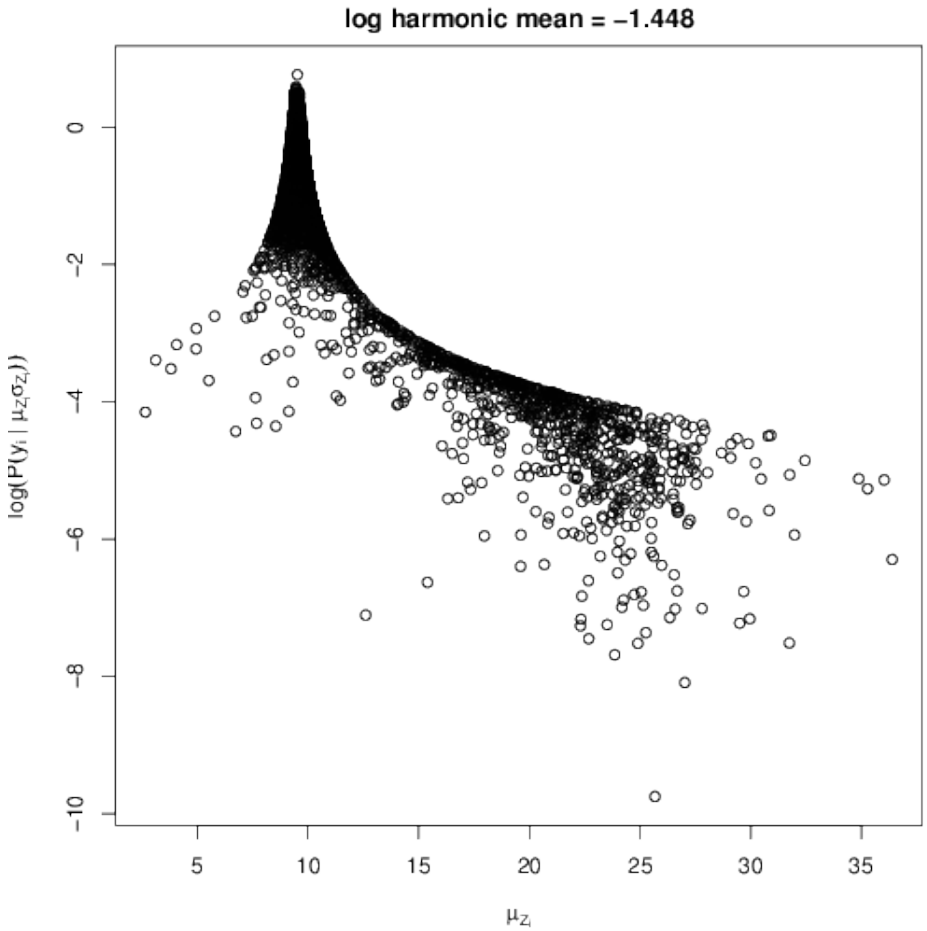
Comparison of 5 Information Criteria

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.

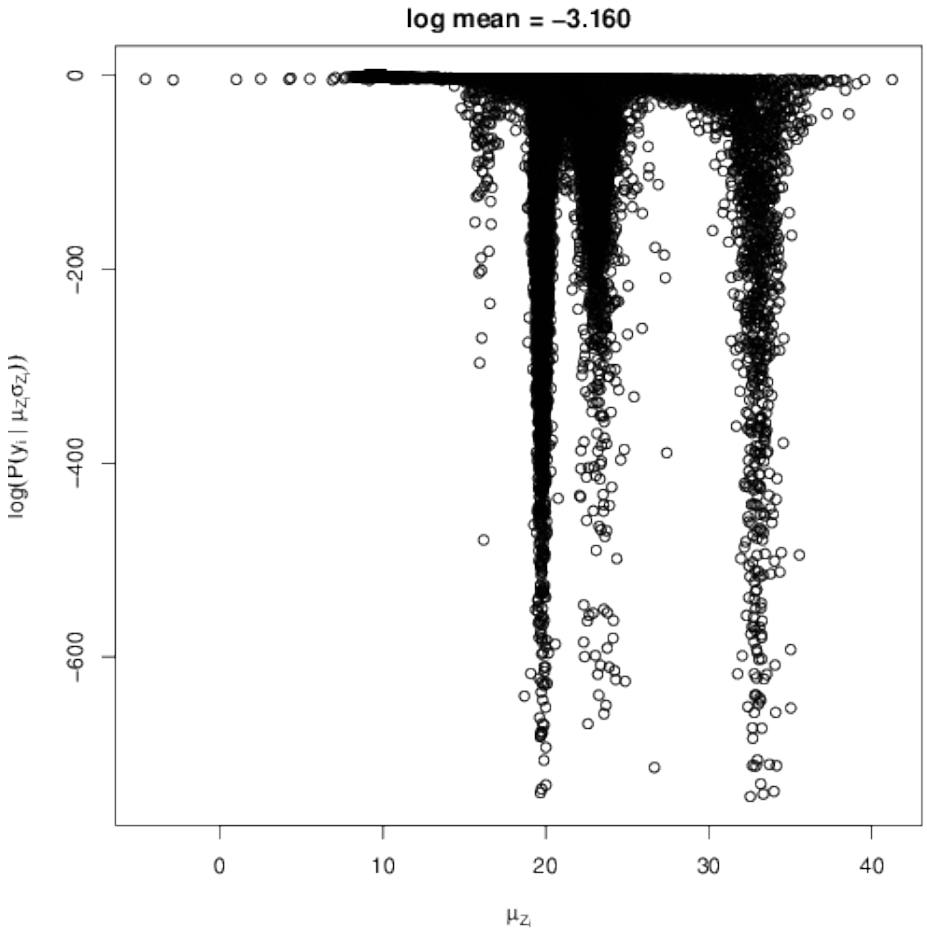
K	DIC	nWAIC	nIS	iWAIC	iIS	CVIC
2	445.38(1.64)	420.27(0.39)	425.63(3.45)	449.56(0.14)	449.62(0.17)	450.55
3	528.78(45.12)	384.94(9.94)	391.29(6.17)	437.23(4.70)	436.43(3.79)	427.46
4	774.85(31.58)	339.91(1.87)	363.55(5.32)	422.43(0.53)	422.76(0.54)	423.16
5	710.88(25.34)	328.19(0.29)	362.30(3.70)	421.02(0.09)	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

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.



(a) Full Data Posterior



(b) CV Posterior

Subsection 2

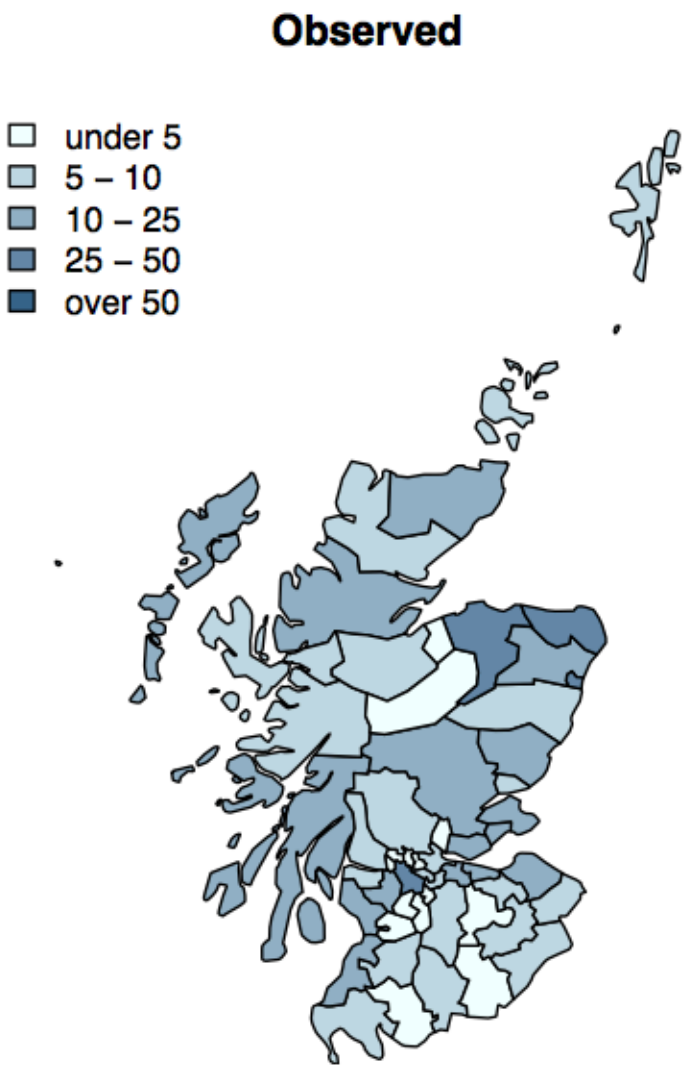
Correlated Random Spatial Effect Models

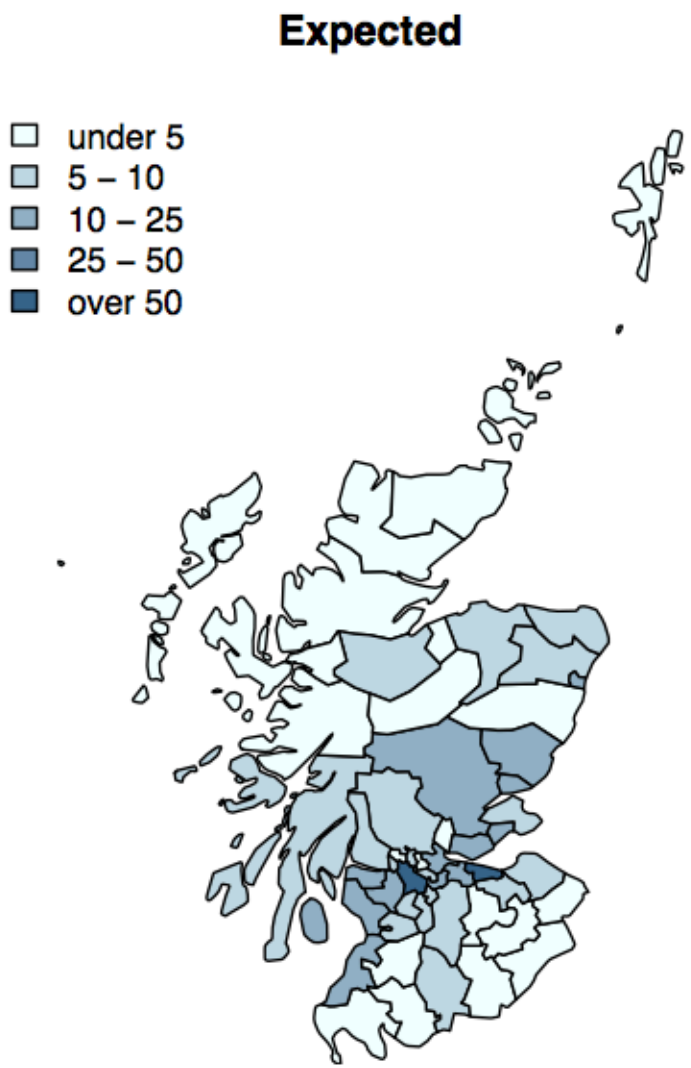
Scottish Lip Cancer Data I

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:

$$y_i | E_i, \lambda_i \sim \text{Poisson}(\lambda_i E_i), \tag{31}$$

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 $s = (s_1, \cdots, s_n)'$:

$$\text{model 1 (spatial+linear, full)} : s \sim N_n(\alpha + X\beta, \Phi\tau^2), \tag{32}$$
$$\text{model 2 (spatial)} : s \sim N_n(\alpha, \Phi\tau^2), \tag{33}$$
$$\text{model 3 (linear)} : s \sim N_n(\alpha + X\beta, I_n\tau^2), \tag{34}$$
$$\text{model 4 (exchangable)} : s \sim N_n(\alpha, I_n\tau^2), \tag{35}$$

where Φ specify spatial association between districts.

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).

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):

$$s_i | s_{-i}, \boldsymbol{\theta} \sim N(\alpha + x_i\beta + \phi \sum_{j \in N_i} (c_{ij}(s_j - \alpha - x_j\beta)), \tau^2 m_{ii}), \tag{36}$$

  where N_i is the set of neighbours of district i.

- Integrated predictive density:

$$P(y_i^{\text{obs}} \,|\, \boldsymbol{\theta}, s_{-i}) = \int \text{dpoisson}(y_i^{\text{obs}} | \lambda_i E_i) P(s_i \,|\, \boldsymbol{\theta}, s_{-i}) ds_i \tag{37}$$

  We generate 200 random numbers of s_i from the distribution (36), and then estimate the integral in (37).

Comparison of 5 Information Criteria

Table 2: 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)

Section 5

Conclusions and Future Work

Conclusions and Future Work

- The proposed iIS and iWAIC significantly reduce the bias of nIS and nWAIC in evaluating Bayesian models with unit-specific latent variables.
- 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 and in more complex models empirically.
- 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?