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

Presented on 12 December 2014 at
Department of Mathematics
Tongji University, Shanghai, China

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).

- Special thanks to Dr. Lingjun Zhou for the invitation and warm host for my stay in Tongji University.

Outline

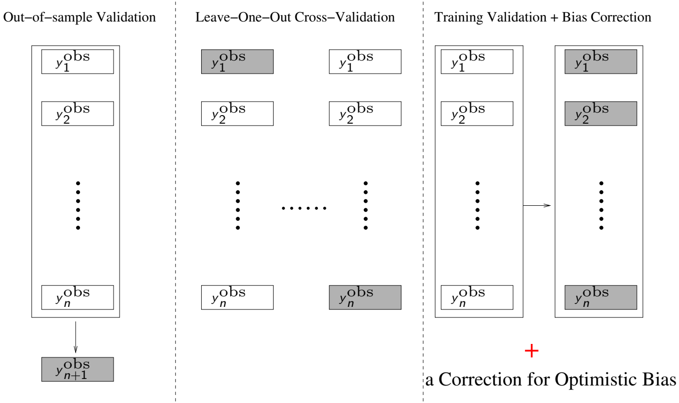
1. Introduction and Literature Review
2. Cross-validatory (CV) Posterior Predictive Evaluation
3. Importance Sampling (IS) Approximations
   - Non-integrated Importance Sampling (nIS)
   - Integrated Importance Sampling (iIS)
4. WAIC Approximations
5. Data Examples
   - Mixture Models for Galaxy Data
   - A Simulation Study with Mixture Models
   - Models with Correlated Spatial Effects
6. Conclusions and Future Work
7. References

Section 1

Introduction and Literature Review

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:



Optimistic bias = Training (within-sample) validation - Out-of-sample validation

Reviews of Bias-corrected Training Validation I

1. Akaike information criterion (Akaike, 1973) for classic statistics

$$\text{AIC} = -2\left(\log P(y^{\text{obs}}|\hat{\theta}_{\text{MLE}}) - p\right) \tag{1}$$

2. For Bayesian statistics, DIC (proposed by Spiegelhalter et al., 2002) was proposed:

$$\begin{aligned} \text{DIC} &= -2\left(\log P(y^{\text{obs}}|\hat{\theta}) - p_{\text{DIC}}\right), \text{where}, & (2) \\ \hat{\theta} &= E_{\text{post}}(\theta|\text{data}) & (3) \\ p_{\text{DIC}} &= 2[\log P(y^{\text{obs}}|\hat{\theta}) - E_{\text{post}}\left(\log(P(y^{\text{obs}}|\theta))\right)] & (4) \end{aligned}$$

AIC and DIC are only justified only for models with identifiable parameters.

Reviews of Bias-corrected Training Validation II

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

$$P(\widehat{y_i^{\text{obs}}|y_{-i}^{\text{obs}}}) = \frac{E_{\text{post}}(P(y_i^{\text{obs}}|\theta))}{\exp\left\{V_{\text{post}}\left(\log(P(y_i^{\text{obs}}|\theta))\right)\right\}} \tag{5}$$

$$\text{WAIC} = -2\log(P(\widehat{y_i^{\text{obs}}|y_{-i}^{\text{obs}}})) \tag{6}$$

   WAIC is justified for models with non-identifiable parameters (therefore widely applicable), but currently for only independent samples.

4. Importance Sampling or harmonic mean estimates (proposed by Gelfand et al. (1992)). For each unit:

$$P(\widehat{y_i^{\text{obs}}|y_{-i}^{\text{obs}}}) = \frac{1}{E_{\text{post}}\left(1/P(y_i^{\text{obs}}|\theta)\right)} \tag{7}$$

$$\text{IS estimate of IC} = -2\log(P(\widehat{y_i^{\text{obs}}|y_{-i}^{\text{obs}}})) \tag{8}$$

What Will We Propose?

We propose two improved methods (namely iIS, and iWAIC) inspired by importance sampling formulae for approximating cross-validatory (CV) predictive evaluation.

Our Goal:

To improve predictive model evaluation in Bayesian models with correlated unit-specific latent variables, for example those models for spatial and temporal data.

Section 2

Cross-validatory (CV) Posterior Predictive Evaluation

Bayesian Models with Unit-specific Latent Variables

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

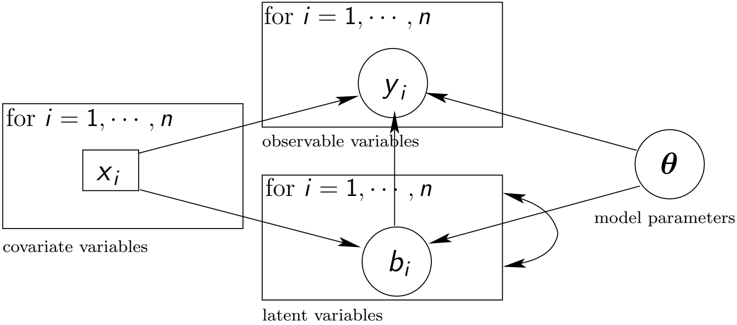
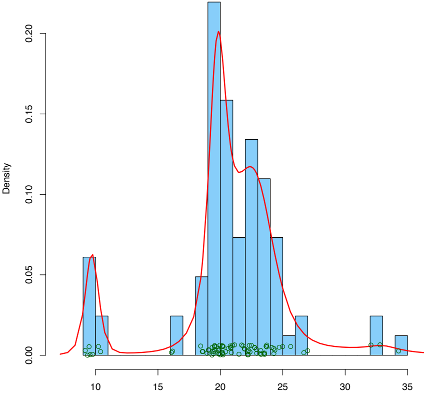


Figure 1: Graphical representation. The double arrows in the box for $b_{1:n}$ mean possible dependency between $b_{1:n}$. Note that the covariate x_i will be omitted in the conditions of densities for b_i and y_i throughout this paper for simplicity.

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.



Mixture Models with a Fixed Number, K, of Components

- Considering the heterogeneity, we model the Galaxy data with mixture models:

$$\begin{aligned} y_i | z_i = k, \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}^2_{1:K} \sim N(\mu_k, \sigma^2_k),\ \text{for } i = 1, \ldots, n \qquad (9) \\ z_i | p_{1:K} &\sim \text{Category}(p_1, \ldots, p_K),\ \text{for } i = 1, \ldots, n \qquad (10) \\ \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}^2_{1:K}, p_{1:K} &\sim \text{certain prior} \qquad (11) \end{aligned}$$

- The finite mixture model is an example of the models with unit-specific latent variables:
  - the observed variable is y_i,
  - the mixture component indicator z_i is the unit-specific latent variable
  - the model parameters $\boldsymbol{\theta}$ is $(\boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}^2_{1:K}, p_{1:K})$.
- We are interested in determining $K = ?$.

Generals of CV Posterior Predictive Evaluation: I

- 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 $\boldsymbol{\theta}$: $P(\boldsymbol{\theta})$.
- 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**:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}}) = \prod_{j \neq i} P(y_j^{\text{obs}}|b_j, \boldsymbol{\theta}) P(b_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \qquad (12)$$

- Based on the above posterior, we can form a posterior predictive distribution for y_i:

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

Then we can compare the predictive distribution with y_i^{obs}.

Generals of CV Posterior Predictive Evaluation: II

- 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 $P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}})$:

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

  We could use MCMC to draw samples of $(\boldsymbol{\theta}, b_{1:n})$ from CV posterior, and then use the samples to approximate the above integral.

- Important: We need to repeat this procedure for each $i = 1, \dots, n$. Time consuming!. We want to fit MCMC given the full data only once, then find the above integrals for all $i = 1, \dots, n$.

A Special Case: CV Information Criterion (CVIC)

- Evaluation function:

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

- CV posterior predictive density:

$$\begin{aligned} E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) &= \int P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i) {\color{red} P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}})} d\boldsymbol{\theta} d b_{1:n} \\ &= P(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) \end{aligned}$$

- **CV information criterion** (CVIC) for comparing Bayesian models is:

$$\text{CVIC} = -2 \sum_{i=1}^{n} \log(P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})). \tag{13}$$

Section 3

Importance Sampling (IS) Approximations

Subsection 1

Non-integrated Importance Sampling (nIS)

Posterior Distribution Conditional on Full Data

The posterior of $(b_{1:n}, \boldsymbol{\theta})$ given observations $y^{\text{obs}}_{1:n}$ is proportional to the joint density of $y^{\text{obs}}_{1:n}$, $b_{1:n}$, and $\boldsymbol{\theta}$:

$$P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y^{\text{obs}}_{1:n}) = \prod_{j=1}^{n} P(y^{\text{obs}}_j|b_j, \boldsymbol{\theta}) P(b_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) / C_1, \tag{14}$$

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

The ratio between CV posterior and the full data posterior is:

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

Importance Sampling Method

- Our samples $(\boldsymbol{\theta}, b_i) \sim P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y_{1:n}^{\text{obs}})$, but we are interested in estimating the mean of a function w.r.t. $P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}})$.
- Importance Reweighing method:

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)) = \frac{E_{\text{post}}\left[a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) W_i^{\text{nIS}}(\boldsymbol{\theta}, b_{1:n})\right]}{E_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, b_{1:n})\right]}, \text{ where,} \tag{15}$$

$$W_i^{\text{nIS}}(\boldsymbol{\theta}, b_{1:n}) = \frac{P_{\text{post(-i)}}(\boldsymbol{\theta}, b_{1:n}|y_{-i}^{\text{obs}})}{P_{\text{post}}(\boldsymbol{\theta}, b_{1:n}|y_{1:n}^{\text{obs}})} = \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)} \times \frac{C_1}{C_2}. \tag{16}$$

- **Intuition of this formula**: Full data posterior sample $(\boldsymbol{\theta}, b_i)$ that fit better y_i^{obs} should be considered less in validating y_i^{obs}, as a way to correct for the optimistic bias.

IS Estimate of CVIC

- In CVIC, $a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) = P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)$, therefore, in the numerator,

$$a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i) W_i^{\text{nIS}}(\boldsymbol{\theta}, b_{1:n}) = \frac{C_1}{C_2}$$

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

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

- nIS (non-integrated IS) estimate of $P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})$:

$$\hat{P}^{\text{nIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post}}\left[1/P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_i)\right]}. \qquad (18)$$

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

Subsection 2

Integrated Importance Sampling (iIS)

Integrated Importance Sampling (iIS)

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

  Therefore, $P(b_i|y_{1:n}^{\text{obs}})$ and $(b_i|y_{-i}^{\text{obs}})$ differ so much that the nIS estimate becomes inaccurate and unstable.
- **Our solution:**
  For each unit i, drop b_i *temporarily* from full data posterior sample, regenerate b_i without reference to y_i^{obs}, that is, from $P(b_i|b_{-i}, \boldsymbol{\theta})$.

Subsection 2

Integrated Importance Sampling (iIS)

Derivation of iIS Formula: I

1. Integrated Evaluation Function
   Rewrite the expectation in (8) as

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

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

where,

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

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 does not refer to y_i^{obs}.

Derivation of iIS Formula: II

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

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

where,

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

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

Derivation of iIS Formula: 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:

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

Summary for iIS

- Replace $a(y_i^{\text{obs}}, \boldsymbol{\theta}, b_i)$ with

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

- Replace $P(y_i^{\text{obs}} | b_i, \boldsymbol{\theta})$ with

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

iIS Estimate for CVIC

The iIS estimate for $P(y_i^{\text{obs}} | y_{-i}^{\text{obs}})$ is

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

Accordingly, iIS estimate of CVIC is

$$\widehat{\text{CVIC}}^{\text{iIS}} = -2 \sum_{i=1}^{n} \log(\hat{P}^{\text{iIS}}(y_i^{\text{obs}} | y_{-i}^{\text{obs}})) \tag{26}$$

Section 4

WAIC Approximations

WAIC for Models without Latent Variables

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

$$\mathrm{WAIC} = -2\sum_{i=1}^{n}\Big[\log(E_{\mathrm{post}}(P(y_i^{\mathrm{obs}}|\boldsymbol{\theta}))) - V_{\mathrm{post}}(\log(P(y_i^{\mathrm{obs}}|\boldsymbol{\theta})))\Big], \qquad (27)$$

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

$$\hat{P}^{\mathrm{WAIC}}(y_i^{\mathrm{obs}}|y_{-i}^{\mathrm{obs}}) = \frac{E_{\mathrm{post}}(P(y_i^{\mathrm{obs}}|\boldsymbol{\theta}))}{\exp\left\{V_{\mathrm{post}}(\log(P(y_i^{\mathrm{obs}}|\boldsymbol{\theta})))\right\}}. \qquad (28)$$

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:

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

We will refer to (29) as non-integrated WAIC (or nWAIC for short) method for approximating CV posterior predictive density. The corresponding information criterion based on (29) is:

$$\text{nWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{nWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})). \tag{30}$$

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:

$$\hat{P}^{\text{iWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{E_{\text{post}}(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i}))}{\exp\left\{V_{\text{post}}(\log(P(y_i^{\text{obs}}|\boldsymbol{\theta}, b_{-i})))\right\}}. \tag{31}$$

Accordingly, iWAIC for approximating CVIC is given by :

$$\text{iWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{iWAIC}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}})). \tag{32}$$

Section 5

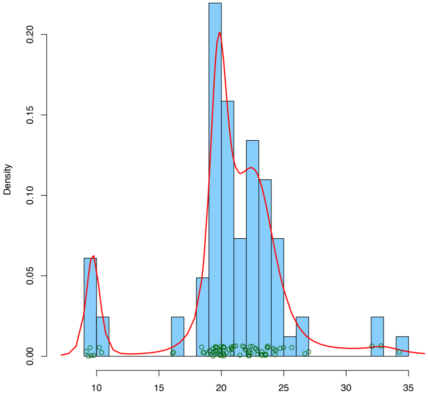
Data Examples

Subsection 1

Mixture Models for Galaxy Data

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.



Mixture Models with a Fixed Number, K, of Components

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

$$y_i | z_i = k, \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}_{1:K} \sim N(\mu_k, \sigma_k^2), \text{ for } i = 1, \ldots, n \tag{33}$$

$$\begin{aligned} z_i | p_{1:K} &\sim \text{Category}(p_1, \ldots, p_K), \text{ for } i = 1, \ldots, n & (34) \\ \mu_k &\sim N(20, 10^4), \text{ for } k = 1, \ldots, K & (35) \\ \sigma_k^2 &\sim \text{Inverse-Gamma}(0.01, 0.01 \times 20), \text{ for } k = 1, \ldots, K & (36) \\ p_k &\sim \text{Dirichlet}(1, \ldots, 1) \text{ for } k = 1, \ldots, K & (37) \end{aligned}$$

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.

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

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.

Non-integrated and integrated predictive density

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

- The non-integrated predictive density:

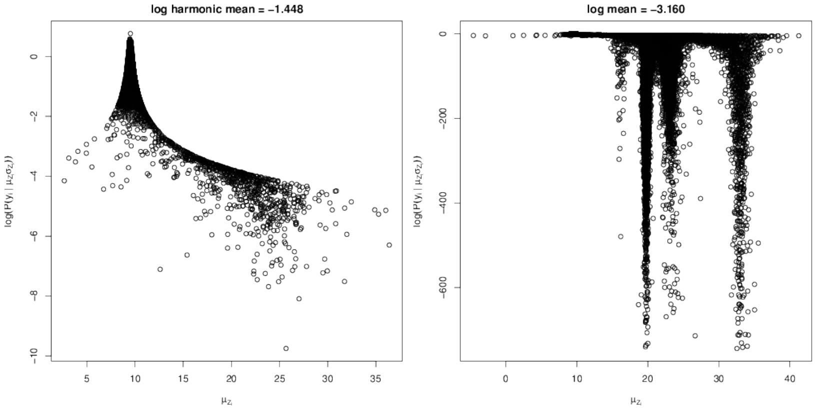
$$P(y_i^{\text{obs}} | z_i, \boldsymbol{\theta}) = \phi(y_i^{\text{obs}} | \mu_{z_i}, \sigma_{z_i})$$

- The integrated predictive density:

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

The Need of Using Integrated Predictive Density

Figure 2: 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) Full Data Posterior

(b) CV Posterior

Comparison of 5 Information Criteria

Table 1: Comparison of 5 information criteria for mixture models applied to Galaxy data. 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

An important note: CVIC (as well as iIS and iWAIC) is not sensitive in penalizing complex model because Bayesian methods can adjust model complexity automatically. A question that is interesting to address in the future!

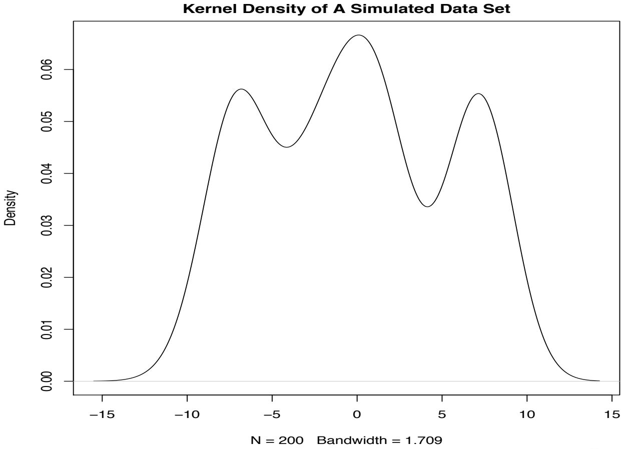
Subsection 2

A Simulation Study with Mixture Models

Data Generating Model

I simulated 100 data sets, each containing 200 data points y_i from the following mixture model with $K = 4$ components:

$$(1/4)N(-7,1) + (1/4)N(-2,1) + (1/4)N(1,1) + (1/4)N(7,1)$$



Details of Simulation Studies

- Models and MCMC simulations are the same as for Galaxy data
- For each of the 100 data sets, we simulate an MCMC to fit finite mixture models with $K = 2, \ldots, 7$ components to the full data, then we compare them using 5 information criteria.

Average of Information Criteria in 100 Data sets

Table 2: Each number in the table is average of IC values in 100 replicates of data sets, for a model with certain number of components, K, and given a certain criterion (column).

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

Frequency of Selected Models

Table 3: Frequency of models with different K being selected by looking at the minimum information criterion value based on 5 information criteria. True model is $K = 4$.

K	IS	WAIC	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
total	100	100	100	100	100

Subsection 3

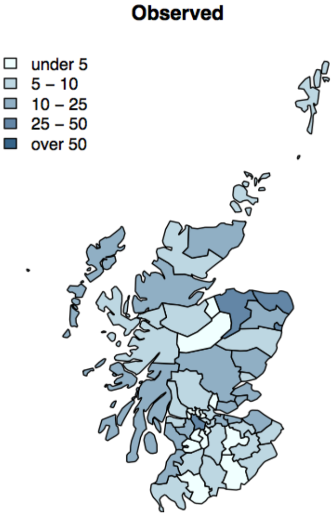
Models with Correlated Spatial Effects

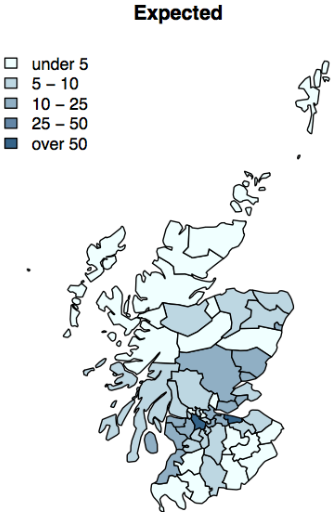
Scottish Lip Cancer Data I

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


Observed


Expected

Four Models Considered: I

The y_i is modelled as a Poisson random variable:

$$y_i|E_i, \lambda_i \sim \text{Poisson}(\lambda_i E_i), \tag{38}$$

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

$$\text{model 1 (spatial+linear, full)} : s \sim N_n(\alpha + X\beta, \Phi\tau^2), \tag{39}$$

$$\text{model 2 (spatial)} : s \sim N_n(\alpha, \Phi\tau^2), \tag{40}$$

$$\text{model 3 (linear)} : s \sim N_n(\alpha + X\beta, I_n\tau^2), \tag{41}$$

$$\text{model 4 (exchangable)} : s \sim N_n(\alpha, I_n\tau^2), \tag{42}$$

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. This model is called **proper conditional auto regression (CAR) model**.

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

$$\tau^2 \sim \text{Inv-Gamma}(0.5, 0.0005) \tag{43}$$

$$\beta \sim N(0, 1000^2) \tag{44}$$

$$\phi \sim \text{Unif}(\phi_0, \phi_1), \tag{45}$$

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

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

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.

Non-integrated and integrated predictive density

For each unit i, and for each MCMC sample of $(s_1, \ldots, s_n, \boldsymbol{\theta})$:

- Non-integrated predictive density

$$P(y_i^{\text{obs}}|s_i, \boldsymbol{\theta}) = \text{dpoisson}(y_i^{\text{obs}}|\lambda_i E_i) \tag{46}$$

- Conditional distribution

$$P(s_i|s_{-i}, \boldsymbol{\theta}) \sim N(\alpha + x_i\beta + \phi \sum_{j \in N_i} (c_{ij}(s_j - \alpha - x_j\beta)), \tau^2 m_{ii}), \tag{47}$$

  where N_i is the set of neighbours of district i.

- Integrated predictive density:

$$P(y_i^{\text{obs}} \mid \boldsymbol{\theta}, s_{-i}) = \int \text{dpoisson}(y_i^{\text{obs}}|\lambda_i E_i) P(s_i \mid \boldsymbol{\theta}, s_{-i}) ds_i \tag{48}$$

  We generate 200 random numbers of s_i from the distribution (47), and then estimate the integral in (48).

Comparison of 5 Information Criteria

Table 4: 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)

Conclusion

- Naive use of IS and WAIC to latent variables models by treating latent variables as parameters may give wrong results in model comparison, hence unreliable.
- 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.

Future Work I

- Demonstration of iIS and iWAIC in many other models used in many applications. A undergraduate student, Zhouji Zheng, has considered stochastic volatility models for financial time series data. Shi Qiu has done further research in disease mapping problems. The comparison results are very favourable to iIS and iWAIC. We will consider many other models: hidden Markov models? Structural equation models? Factor Analysis Models? and more ... Suitable for Master-level training.
- Looking at Log density function (or deviance) has many limitations (scale-variant, hard to assess practical significance). Use of iIS with other choice of evaluation function, for example, the tail probability of posterior predictive distribution — **posterior predictive p-value**. Such p-values can be used to
  - determine whether an observation is an outlier or not;
  - check the goodness of a model to a data set (not comprising a set of models!);

Future Work II

    - compare models too.
- How to define such posterior predictive p-values appropriately is also an active research area! Very thought-provoking area.
- iWAIC works very well in the spatial random effect models. The result is surprising and encouraging. One may consider investigating the theoretical validity of iWAIC, which we have not done. Suitable for a PhD in mathematics. It will require good knowledge in **algebraic geometry results about singularity**. See Watanabe (2009).
- 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 approximation for 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`.