Cross-validatory Model Comparison and Divergent Regions Detection using iIS for Disease Mapping

Longhai Li

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

Presented on 21 January 2016
Seminar of Department of Statistics at the University of Manitoba

Acknowledgements

- Joint work with **Shi Qiu 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 Prof. Liqun Wang and Prof. Alexandre Leblanc for their warm hosting of my visit.

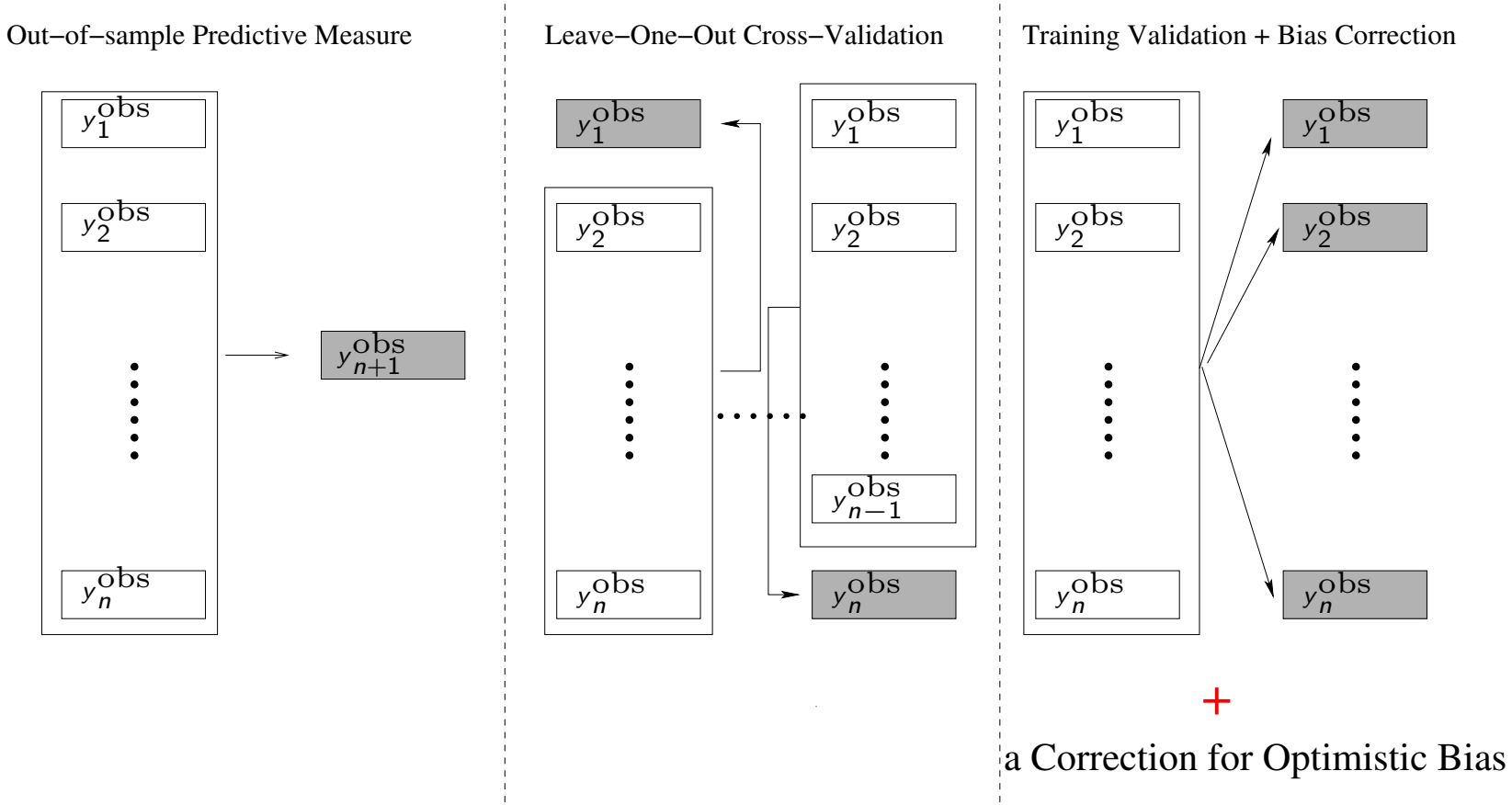
Outline

Section 1

An Introduction to Predictive Model Assessment Methods

Out-of-Sample Predictive Assessment

Predictive assessment is often used for model comparison, diagnostics, and outlier detection in practice.



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

Review of Various Information Criterion 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 (Spiegelhalter et al., 2002) was proposed:

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

The DIC is justified only for models with identifiable parameters.

Review of Various Information Criterion II

3. Widely Applicable Information Criterion (Watanabe,2009)

$$\text{WAIC} = -2\left(\sum_{i=1}^{n}\log(E_{\text{post}}(P(y_i^{\text{obs}}|\theta))) - p_{waic}\right) \quad (5)$$

$$p_{waic} = \sum_{i=1}^{n} V_{\text{post}}\left(\log(P(y_i^{\text{obs}}|\theta))\right) \quad (6)$$

The WAIC is justified for models with non-identifiable parameters.

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

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

$$\text{IS estimate of IC} = -2\sum_{i=1}^{n}\log(\widehat{P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})}^{\text{IS}}) \quad (8)$$

What Will We Propose?

We propose an improved importance sampling method (iIS) for approximating cross-validatory (CV) predictive assessment.

iIS is applicable to Bayesian models with correlated unit-specific latent variables, for example those models for spatial and temporal data.

Section 2

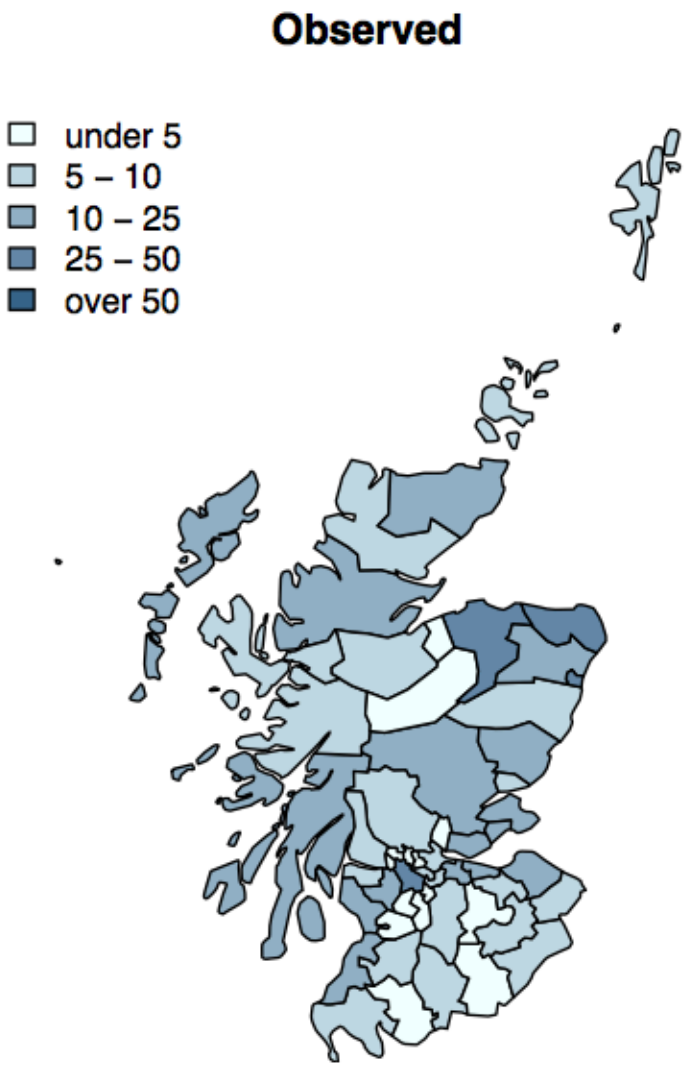
Disease Mapping Models

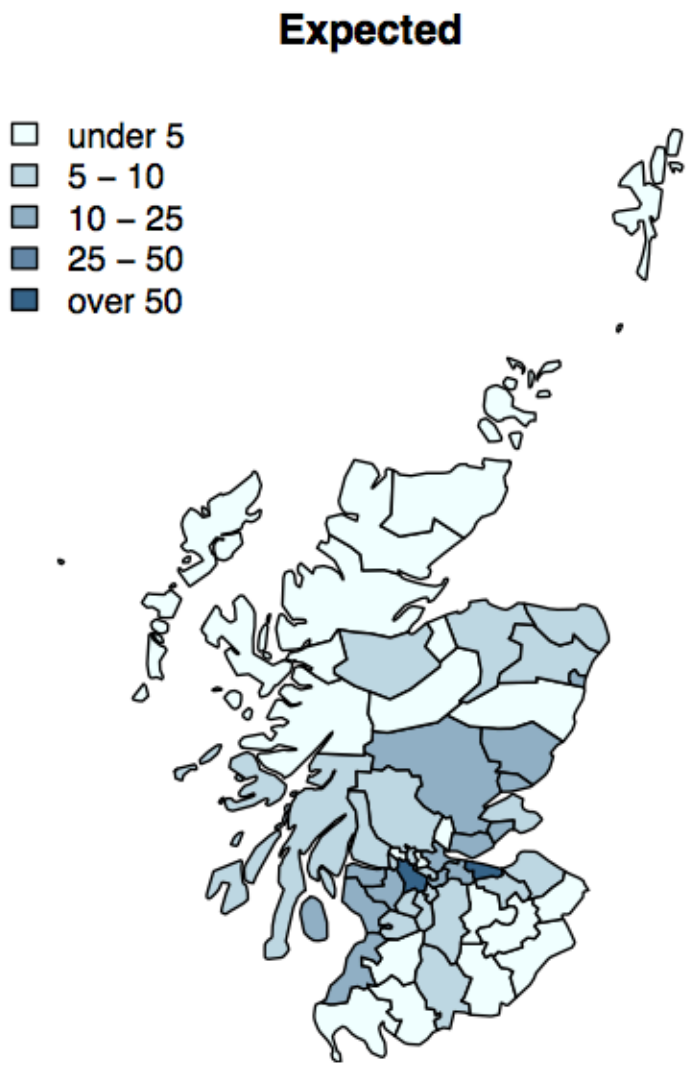
Scottish Lip Cancer Data I

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





A Subset of the Dataset

ID	District name	*Y*	*E*	*SMR*	*X*	Neighbours
1	Skye-Lochalsh	9	1.38	6.52	16	5,9,11,19
2	Banff-Buchan	39	8.66	4.50	16	7,10
3	Caithness	11	3.04	3.62	10	6,12
11	Western Isles	13	4.40	2.95	7	1,5,9,12
15	NE Fife	17	7.84	2.17	7	25,29,50
17	Badenoch	2	1.07	1.87	10	7,9,13,16,19,29
26	Dunfermline	15	12.49	1.20	1	25,29,42,43
38	Monklands	8	9.35	0.86	1	30,42,44,49,51,54
42	Falkirk	8	15.78	0.51	16	26,30,34,38,43,51
45	Edinburgh	19	50.72	0.37	1	28,30,33,56
49	Glasgow	28	88.66	0.32	0	38,40,41,44,47,48,52,53,54
50	Dundee	6	19.62	0.31	1	15,21,29
55	Annandale	0	4.16	0	16	18,20,24,27,56
56	Tweeddale	0	1.76	0	10	18,24,30,33,45,55

A Hierarchical Bayesian Spatial Model for y_i's

- **A model for the observed variables given latent variables**

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

where λ_i denotes the underlying relative risk for district i.

- **A model for latent log relative risks** $s_i = \log(\lambda_i)$

$$(s_1, \ldots, s_n)' \sim N_n(\alpha + X\beta, \Phi\tau^2)$$

where $\Phi = (I_n - \phi C)^{-1} M$ is a matrix modelling spatial dependency with *proper conditional auto-regressive* (CAR) method.

- **A model (prior) for parameters**

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

Section 3

Integrated Important Sampling (iIS) in General

Bayesian Models with Unit-specific Latent Variables

iIS can be applied to models described as follows:

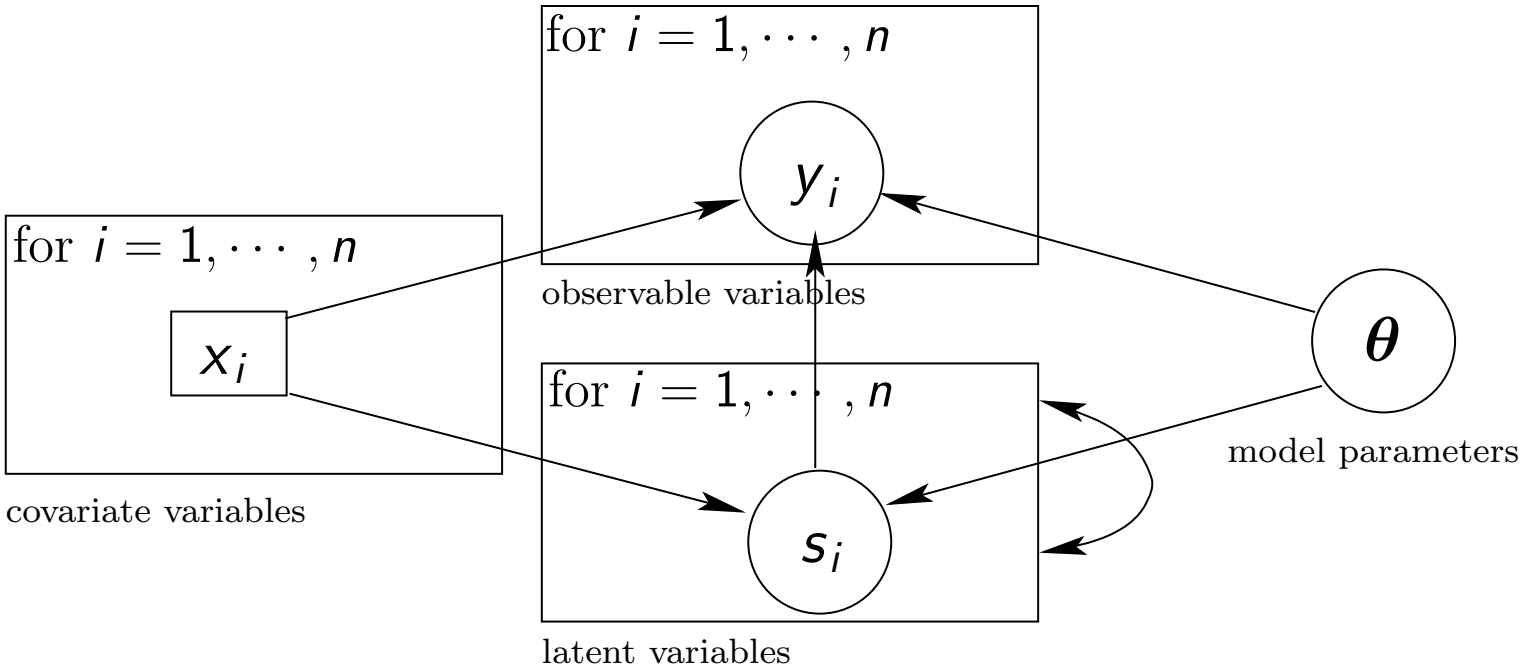


Figure 1: Graphical representation. The double arrows in the box for $s_{1:n}$ mean possible dependency between $s_{1:n}$. Note that the covariate x_i will be omitted in the conditions of densities for s_i and y_i throughout this presentation for simplicity.

Subsection 1

Leave-one-out cross-validatory (LOOCV) Assessment

Cross-validatory Predictive Assessment I

- Suppose we have specified a Bayesian model:
  - a density for y_i given s_i: $P(y_i|s_i, \boldsymbol{\theta})$,
  - a joint density for latent variables $s_{1:n}$: $P(s_{1:n}|\boldsymbol{\theta})$, and
  - a prior density for $\boldsymbol{\theta}$: $P(\boldsymbol{\theta})$.

- **CV posterior distribution** with y_i^{obs} removed from the data set:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, s_{1:n}|y_{-i}^{\text{obs}}) = \prod_{j \neq i} P(y_j^{\text{obs}}|s_j, \boldsymbol{\theta}) P(s_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \qquad (9)$$

Cross-validatory Predictive Assessment II

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

- **CV posterior predictive assessment** is defined as the expectation of the $a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$ with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, s_{1:n}|y_{-i}^{\text{obs}})$:

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)) = \int a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) P_{\text{post(-i)}}(\boldsymbol{\theta}, s_{1:n}|y_{-i}^{\text{obs}}) d\boldsymbol{\theta} ds_{1:n} \tag{10}$$

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

Subsection 2

Two Predictive Model Assessment Questions

Model Comparison with CV Information Criterion (CVIC)

- Using the likelihood of $(\boldsymbol{\theta}, s_i)$ given y_i^{obs} as an evaluation function:

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

- CV posterior predictive density at y_i^{obs}:

$$\begin{aligned} P(y_i^{\text{obs}} | y_{-i}^{\text{obs}}) &= E_{\text{post(-i)}}(P(y_i^{\text{obs}} | \boldsymbol{\theta}, s_i)) \\ &= \int P(y_i^{\text{obs}} | \boldsymbol{\theta}, s_i) P_{\text{post(-i)}}(\boldsymbol{\theta}, s_{1:n} | y_{-i}^{\text{obs}}) d\boldsymbol{\theta} ds_{1:n} \end{aligned}$$

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

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

Outlier Detection with CV Predictive p-value

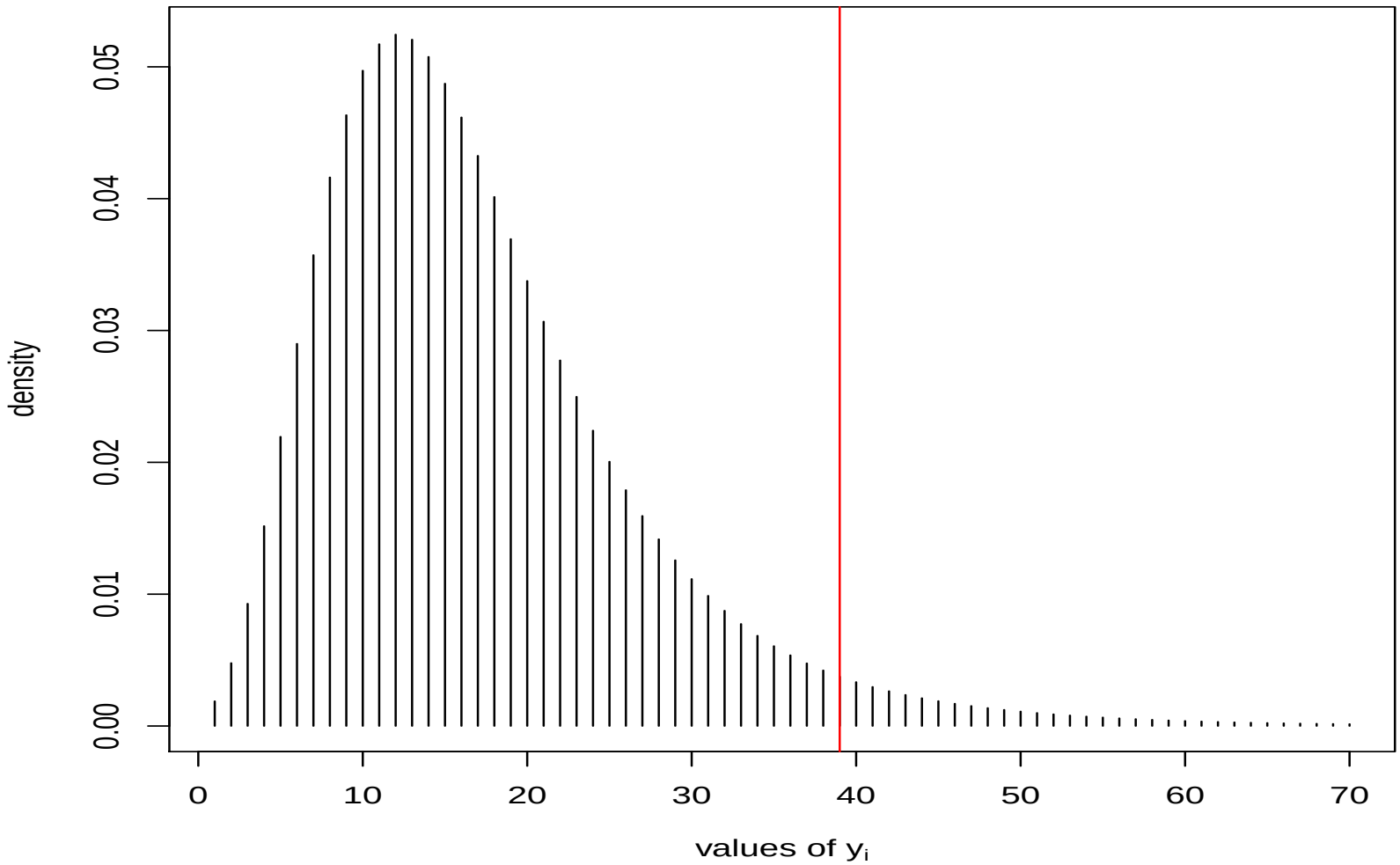
- Using a tail probability of $P(y_i|\boldsymbol{\theta}, s_i)$ as an evaluation function:

$$\begin{aligned} a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) &= \text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) \\ &= Pr(y_i > y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) + 0.5Pr(y_i = y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) \end{aligned}$$

- **CV predictive p-value** for detecting outliers:

$$\begin{aligned} \text{p-value}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) &= E_{\text{post(-i)}}(\text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)) \\ &= Pr(y_i > y_i^{\text{obs}}|y_{-i}^{\text{obs}}) + 0.5Pr(y_i = y_i^{\text{obs}}|y_{-i}^{\text{obs}}) \end{aligned}$$

CV Predictive Density and p-value



Problem with Actual LOOCV

We need to repeat this procedure for each $i = 1, \ldots, n$. Time consuming!.

We want to fit MCMC given the full data only once, then find the above integrals for all $i = 1, \ldots, n$.

Subsection 3

Non-integrated Importance Sampling (nIS)

Importance Weight for (θ, s_i)

- The full data posterior of $(s_{1:n}, \theta)$ given observations $y^{\text{obs}}_{1:n}$:

$$P_{\text{post}}(\theta, s_{1:n}|y^{\text{obs}}_{1:n}) = \prod_{j=1}^{n} P(y^{\text{obs}}_j|s_j, \theta)P(s_{1:n}|\theta)P(\theta)/C_1, \tag{11}$$

- The CV posterior of $(\theta, s_{1:n})$ given y^{obs}_{-i}:

$$P_{\text{post(-i)}}(\theta, s_{1:n}|y^{\text{obs}}_{-i}) = \prod_{j\neq i} P(y^{\text{obs}}_j|s_j, \theta)P(s_{1:n}|\theta)P(\theta)/C_2 \tag{12}$$

- Importance weight:

$$W^{\text{nIS}}_i(\theta, s_i) = \frac{P_{\text{post(-i)}}(\theta, s_{1:n}|y^{\text{obs}}_{-i})}{P_{\text{post}}(\theta, s_{1:n}|y^{\text{obs}}_{1:n})} \propto \frac{1}{P(y^{\text{obs}}_i|\theta, s_i)} \tag{13}$$

Importance Sampling Method

- Importance reweighing method:

$$E_{\text{post(-i)}}(a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)) = \frac{E_{\text{post}}\left[a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) W_i^{\text{nIS}}(\boldsymbol{\theta}, s_i)\right]}{E_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, s_i)\right]} \tag{14}$$

- **Direct Understanding**
  Samples of $(\boldsymbol{\theta}, s_i)$ that fit better y_i^{obs} should be given lower in validating y_i^{obs}, as a way to combat against the optimistic bias.

IS Estimate of CVIC

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

$$a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) W_i^{\text{nIS}}(\boldsymbol{\theta}, s_i) = 1.$$

- The CV posterior predictive density $P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})$:

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

- nIS estimate of CVIC 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 4

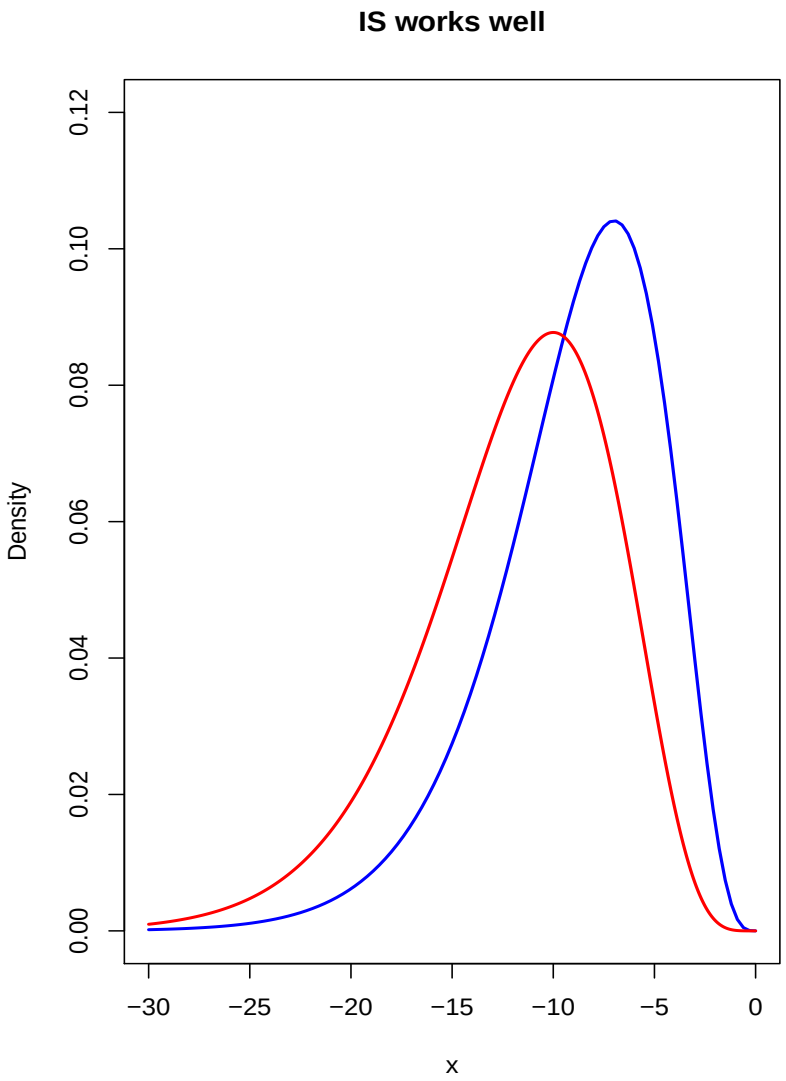
Integrated Importance Sampling (iIS)

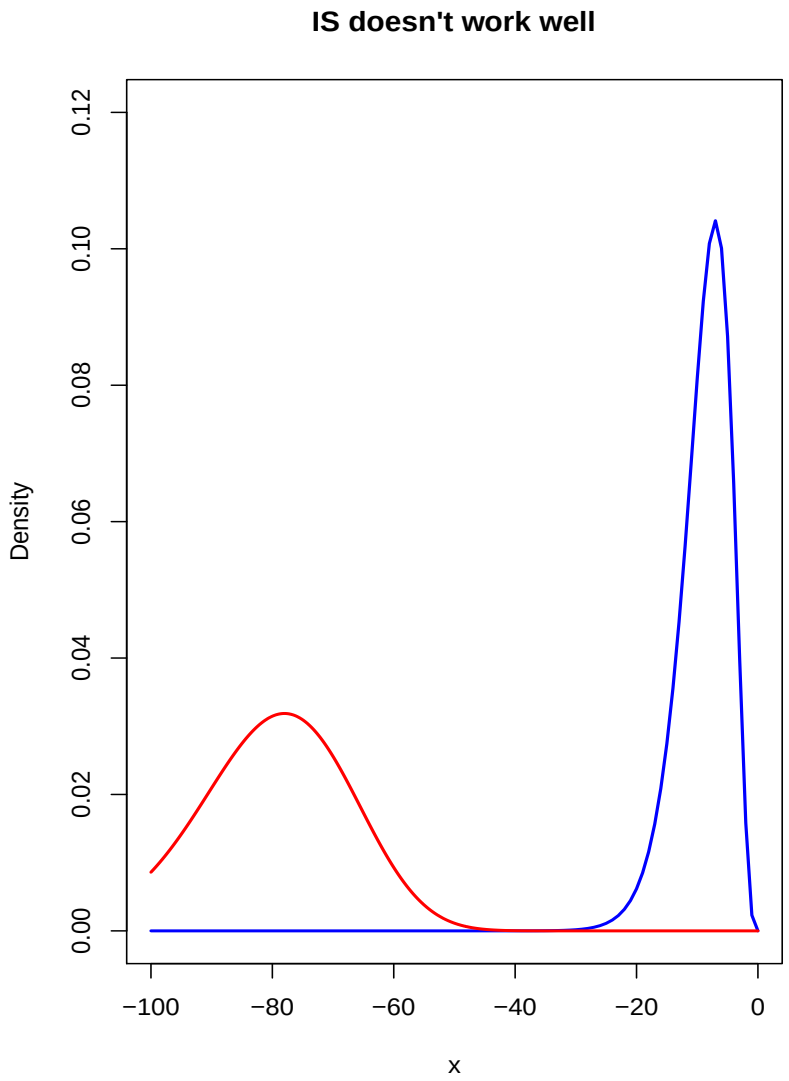
Problem of Non-integrated Importance Sampling (nIS)

- Unfortunately, nIS often does not work well. The full data posterior of $P(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)$ (as a function of $(\boldsymbol{\theta}, s_i)$) favors much larger values than the corresponding CV posterior, because s_i receives much information from y_i^{obs}. That is, s_i and $\boldsymbol{\theta}$ are bounded to the area giving high values of $P(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)$.
- $P(s_i, \boldsymbol{\theta}|y_{1:n}^{\text{obs}})$ and $P(s_i, \boldsymbol{\theta}|y_{-i}^{\text{obs}})$ may differ drastically.

Schematic Density of $x(\boldsymbol{\theta}, s_i) = \log P(y_i^{\text{obs}} | \boldsymbol{\theta}, s_i)$

Red Curve: CV posterior with y_i^{obs} removed, Blue Curve: full data posterior





Idea in Integrated Importance Sampling

- Drop s_i *temporarily* from full data posterior sample, regenerate s_i from $P(s_i|s_{-i}, \boldsymbol{\theta})$ as in actual CV,
- In other words, and apply importance sampling to find

$$\text{expectation w.r.t. } P_{\text{post(-i)}}(\boldsymbol{\theta}, s_{-i}|y^{\text{obs}}_{-i})$$

with

$$\text{expectation w.r.t. } P_{\text{post}}(\boldsymbol{\theta}, s_{-i}|y^{\text{obs}}_{1:n})$$

Expectation w.r.t. CV Posterior of (θ, s_{-i})

- **CV Posterior**

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

- **Expectation of a function of (θ, s_{-i})**

$$E_{\text{post(-i)}}(a(y^{\text{obs}}_i, \theta, s_i)) = \int\int A(y^{\text{obs}}_i, \theta, s_{-i})P_{\text{post(-i)}}(\theta, s_{-i}|y^{\text{obs}}_{-i})d\theta ds_{-i}$$

where,

$$\begin{aligned} A(y^{\text{obs}}_i, \theta, s_{-i}) &= \int a(y^{\text{obs}}_i, \theta, s_i)P(s_i|s_{-i}, \theta)ds_i, \\ P_{\text{post(-i)}}(\theta, s_{-i}|y^{\text{obs}}_{-i}) &= \prod_{j \neq i} P(y^{\text{obs}}_j|s_j, \theta)P(s_{-i}|\theta)P(\theta)/C_2 \end{aligned}$$

Full data posterior of (θ, s_{-i})

- **Full data posterior**

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

- **Marginalize** s_i

$$P_{\text{post}}(\theta, s_{-i}|y^{\text{obs}}_{1:n}) = \Big[\prod_{j\neq i} P(y^{\text{obs}}_j|s_j, \theta)P(s_{-i}|\theta)P(\theta)\Big] P(y^{\text{obs}}_i|\theta, s_{-i})/C_1,$$

where,

$$P(y^{\text{obs}}_i|\theta, s_{-i}) = \int P(y^{\text{obs}}_i|s_i, \theta)P(s_i|s_{-i}, \theta)ds_i.$$

Integrated Importance Sampling Weight and Formula

- Importance Weight for $(\boldsymbol{\theta}, s_{-i})$

$$W_i^{\text{iIS}}(\boldsymbol{\theta}, s_{-i}) = \frac{P_{\text{post(-i)}}(\boldsymbol{\theta}, s_{-i}|y_{-i}^{\text{obs}})}{P_{\text{post}}(\boldsymbol{\theta}, s_{-i}|y_{1:n}^{\text{obs}})} = \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, s_{-i})}. \quad (16)$$

- iIS formula

$$E_{\text{post(-i)}}(A(y_i^{\text{obs}}, \boldsymbol{\theta}, s_{-i})) = \frac{E_{\text{post}}\left[A(y_i^{\text{obs}}, \boldsymbol{\theta}, s_{-i})\ W_i^{\text{iIS}}(\boldsymbol{\theta}, s_{-i})\right]}{E_{\text{post}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, s_{-i})\right]}, \quad (17)$$

Difference of iIS and nIS

- Evaluation Function

$$a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) \Longrightarrow A(y_i^{\text{obs}}, \boldsymbol{\theta}, s_{-i}) = \int a(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) P(s_i | s_{-i}, \boldsymbol{\theta}) ds_i.$$

- Importance Weight

$$P(y_i^{\text{obs}} | \boldsymbol{\theta}, s_i) \Longrightarrow P(y_i^{\text{obs}} | \boldsymbol{\theta}, s_{-i}) = \int P(y_i^{\text{obs}} | s_i, \boldsymbol{\theta}) P(s_i | s_{-i}, \boldsymbol{\theta}) ds_i.$$

- Find these two quantities using Monte Carlo by generating s_i from $P(s_i | s_{-i}, \boldsymbol{\theta})$ or other methods.

A Special Case: iIS Estimate of CVIC

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

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

- iIS estimate of CVIC is

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

Section 4

Applications to Disease Mapping Models

Subsection 1

Model Comparison with Information Criterion

Four Models Considered for Scottish Lip Cancer Data

- Model for y_i:

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

  where λ_i denotes the underlying relative risk for district i.
- Let $s_i = \log(\lambda_i)$. Four different models for spatial effects $s = (s_1, \cdots, s_n)'$:
  - model 1 (spatial+linear, full) : $s \sim N_n(\alpha + X\beta, \Phi\tau^2)$
  - model 2 (spatial) : $s \sim N_n(\alpha, \Phi\tau^2)$
  - model 3 (linear) : $s \sim N_n(\alpha + X\beta, I_n\tau^2)$
  - model 4 (exchangable) : $s \sim N_n(\alpha, I_n\tau^2)$

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.

Implementation of DIC, nIS and nWAIC

- DIC
  apply DIC formula (2) with $(\boldsymbol{\theta}, s_{1:n})$

- nIS and nWAIC
  apply importance sampling (7) and WAIC formula (5) with

$$P(y_i^{\text{obs}}|s_i, \boldsymbol{\theta}) = \text{pois}(y_i^{\text{obs}}|\lambda_i E_i), \text{where } \lambda_i = \exp(s_i)$$

iIS and iWAIC

- Integrated predictive density:

$$P(y_i^{\text{obs}} \mid \boldsymbol{\theta}, s_{-i}) = \int \text{pois}(y_i^{\text{obs}} | \lambda_i E_i) P(s_i \mid \boldsymbol{\theta}, s_{-i}) ds_i \tag{19}$$

$$s_i | s_{-i}, \boldsymbol{\theta} \sim N(\alpha + x_i \beta + \phi \sum_{j \in N_i} (c_{ij}(s_j - \alpha - x_j \beta)), \tau^2 m_{ii}), \tag{20}$$

  where N_i is the set of neighbours of district i.
  We generate 200 random numbers of s_i from the distribution (20), and then estimate the integral in (19).

- iIS and iWAIC
  apply importance sampling (7) and WAIC formula (5) with $P(y_i^{\text{obs}} \mid \boldsymbol{\theta}, s_{-i})$

Comparison of 5 Information Criteria

Table 1: 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)

Subsection 2

Detecting Divergent Regions with CV Predictive p-value

A Hierarchical Bayesian Spatial Model for y_i's

- **A model for the observed variables given latent variables**

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

where λ_i denotes the underlying relative risk for district i.

- **A model for latent log relative risks** $s_i = \log(\lambda_i)$

$$(s_1, \ldots, s_n)' \sim N_n(\alpha + X\beta, \Phi\tau^2)$$

where $\Phi = (I_n - \phi C)^{-1} M$ is a matrix modelling spatial dependency.

- **A model (prior) for parameters**

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

p-value given $(\boldsymbol{\theta}, s_i)$ and CV Predictive p-value

- p-value given $(\boldsymbol{\theta}, s_i)$:

$$\begin{aligned} \text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) &= Pr(y_i > y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) + 0.5Pr(y_i = y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) \\ &= \sum_{y_i > y_i^{\text{obs}}} \text{pois}(y_i|\lambda_i E_i) + 0.5\text{pois}(y_i^{\text{obs}}|\lambda_i E_i) \quad (21) \end{aligned}$$

- **CV predictive p-value** for detecting outliers:

$$\text{p-value}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = E_{\text{post(-i)}}(\text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)) \quad (22)$$

Other Methods for Computing a Predictive p-value

- **Posterior Check**

$$\text{p-value}^{\text{Post.check}}(y_i^{\text{obs}}) = E_{\text{post}}(\text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i))$$

- **Ghosting Method** (Marshall and Spiegelhalter, 2007)

$$\text{p-value}^{\text{Ghost}}(y_i^{\text{obs}}) = E_{\text{ghost}}(\text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)), \text{where}$$

$$P_{\text{ghost}}(s_i, \boldsymbol{\theta}) = P_{\text{post}}(\boldsymbol{\theta}, s_{-i}|y_{1:n}^{\text{obs}}) \times P(s_i|s_{-i}, \boldsymbol{\theta})$$

- **Important Sampling** (Stern and Cressie, 2000)

$$\text{p-value}^{\text{nIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{E_{\text{post}}\left[\text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) W_i^{\text{nIS}}(\boldsymbol{\theta}, s_i)\right]}{E_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, s_i)\right]}, \text{where}$$

$$W_i^{\text{nIS}}(\boldsymbol{\theta}, s_i) = \frac{1}{\text{pois}(y_i^{\text{obs}}|\lambda_i E_i)}$$

iIS estimate of Predictive p-value

$$\text{p-value}^{\text{iIS}}(y_i^{\text{obs}}|y_{-i}^{\text{obs}}) = \frac{E_{\text{post}}\left[\text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_{-i}) W_i^{\text{iIS}}(\boldsymbol{\theta}, s_{-i})\right]}{E_{\text{post}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, s_{-i})\right]}, \text{ where}$$

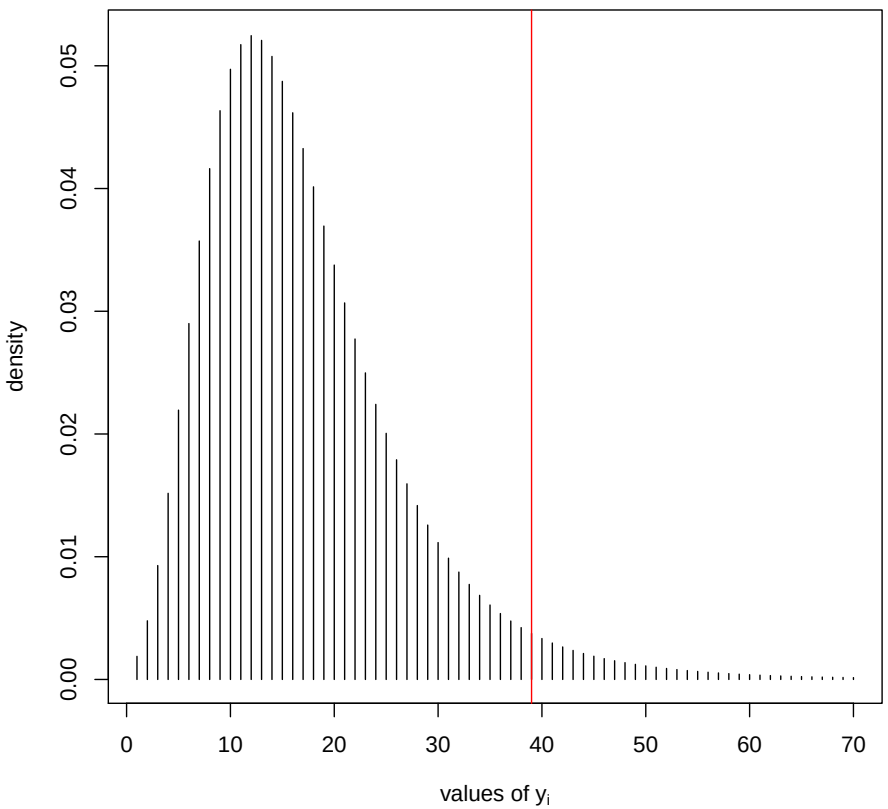
$$\begin{aligned}
\text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_{-i}) &= \int \text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) P(s_i \mid \boldsymbol{\theta}, s_{-i}) ds_i \\
P(y_i^{\text{obs}} \mid \boldsymbol{\theta}, s_{-i}) &= \int \text{pois}(y_i^{\text{obs}}|\lambda_i E_i) P(s_i \mid \boldsymbol{\theta}, s_{-i}) ds_i \\
W_i^{\text{iIS}}(\boldsymbol{\theta}, s_{-i}) &= 1/P(y_i^{\text{obs}} \mid \boldsymbol{\theta}, s_{-i})
\end{aligned}$$

Predictive p-values of Selected Districts

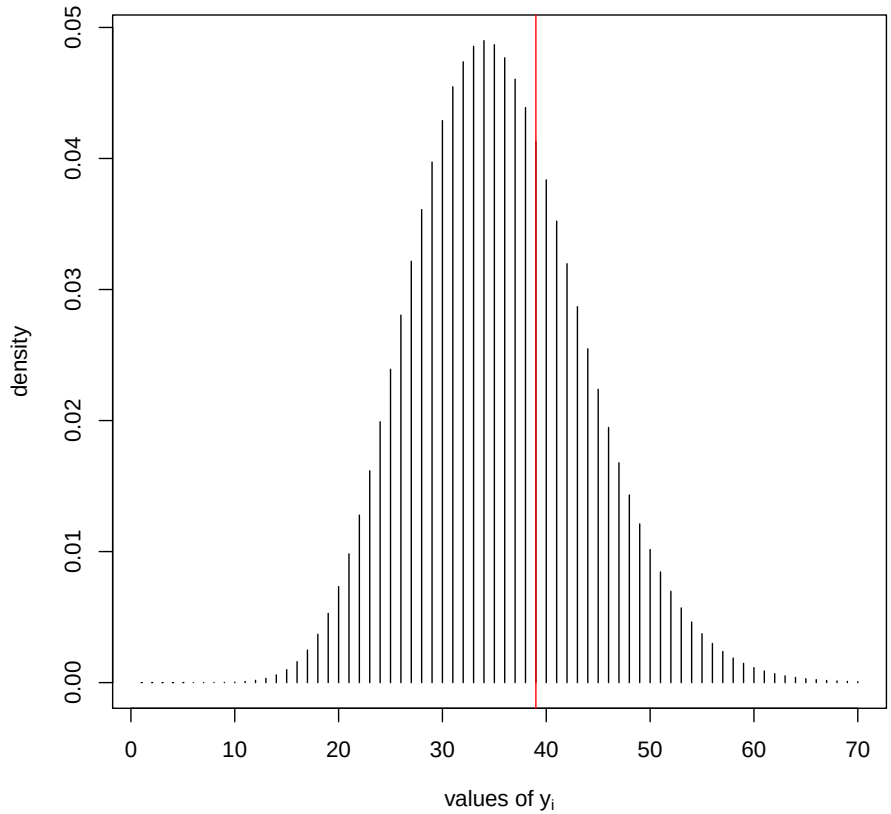
Table 2: The estimated predictive p-value(y_i^{obs}) for a selected subset of the 56 districts in the Scottish lip cancer data.

ID	CV	PCH	GHO	nIS	iIS
1	0.31	0.42	0.32	0.30	0.31
2	0.03	0.32	0.05	0.03	0.03
3	0.09	0.33	0.10	0.12	0.09
11	0.13	0.34	0.13	0.11	0.12
15	0.06	0.27	0.07	0.07	0.06
17	0.60	0.47	0.60	0.53	0.61
45	0.95	0.78	0.89	0.95	0.96
50	0.96	0.82	0.93	0.95	0.96
55	0.99	0.92	0.99	0.99	0.99
56	0.84	0.73	0.83	0.82	0.84

Figure 2: **Illustration of Optimistic Bias in Posterior Checking**

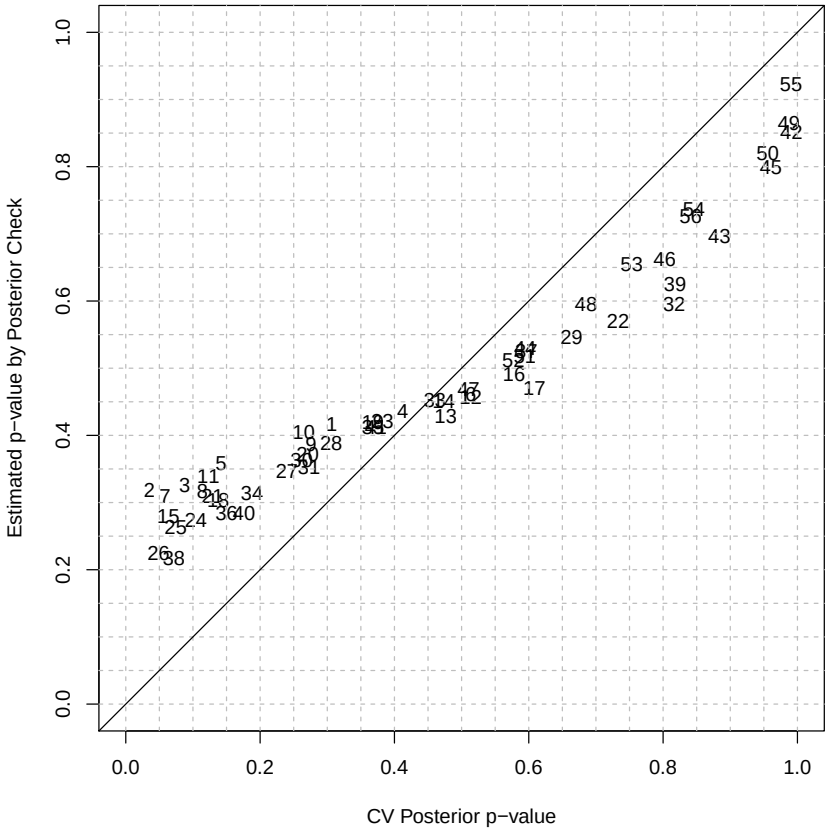


(a) CV predictive PMF of y_2 for district 2 (Banff-Buchan)

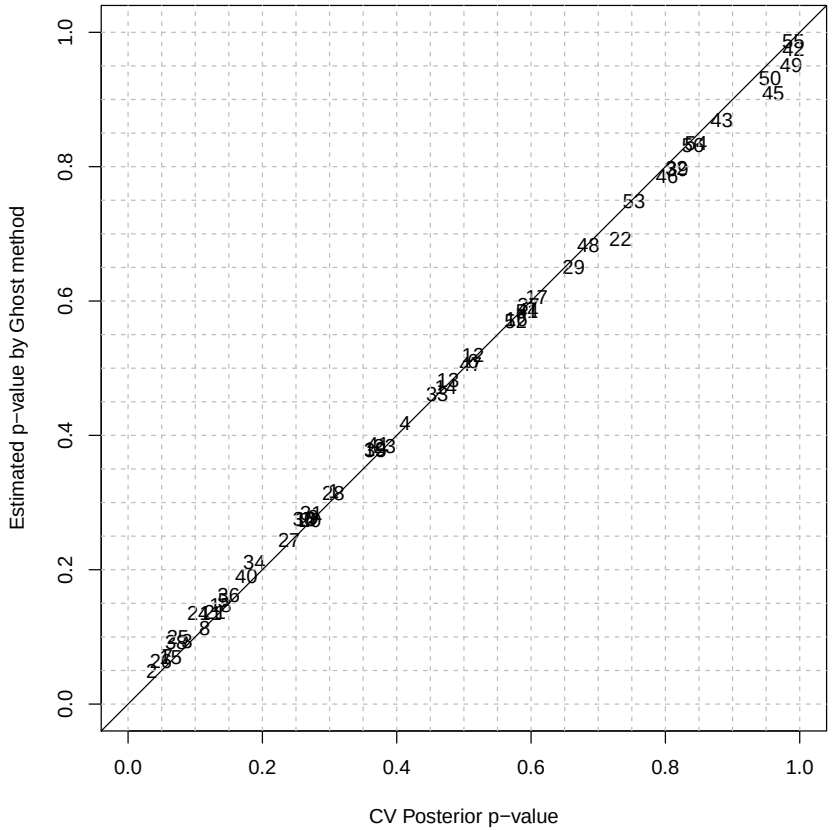


(b) Full data predictive PMF of y_2 for district 2 (Banff-Buchan)

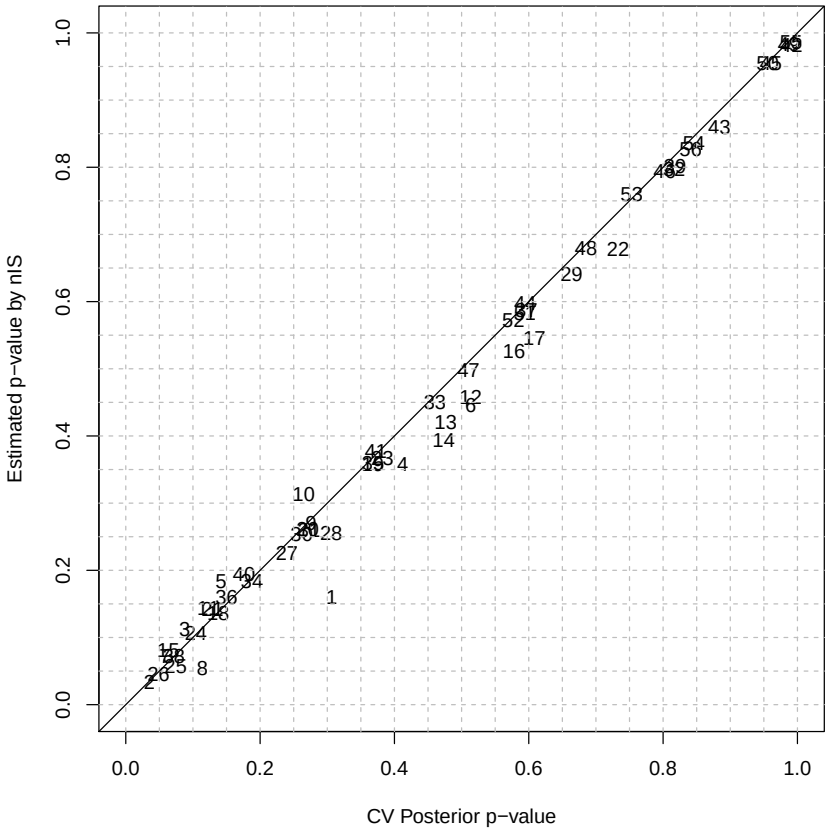
Figure 3: **Comparing estimated p-values with CV predictive p-values**



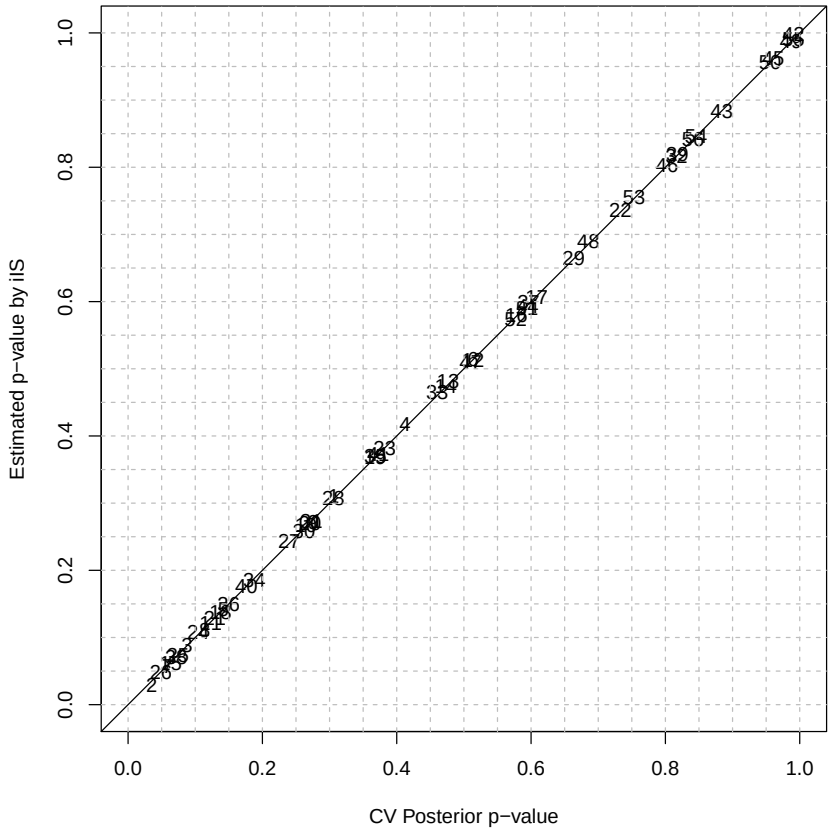
(a) Posterior checking



(b) Ghosting method



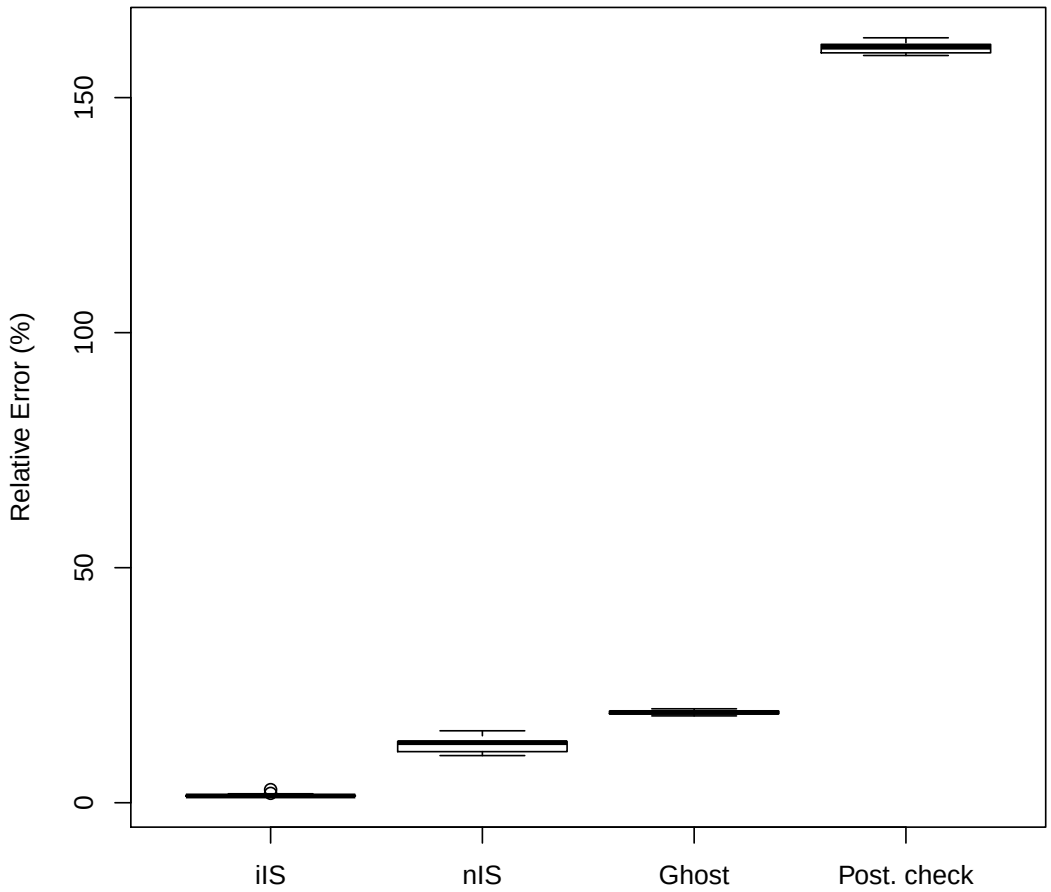
(c) Non-integrated IS (nIS)



(d) Integrated IS (iIS)

Box-plots of Relative Errors in the Estimated p-value

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



Computation Time

	CV	iIS	nIS	GHO	PCH
MCMC	1137.56	20.05	19.97	19.95	19.90
Computing p-value	0.99	143.65	1.25	84.06	1.12
Total	1138.55	163.70	21.22	104.00	21.01

Conclusions

- Naive application of IS to latent variables models by treating latent variables as parameters may give wrong results in predictive model assessment.
- The new proposed iIS significantly improve the accuracy of IS in assessing Bayesian models with unit-specific latent variables. In our studies, they gave results very close to what given by the actual cross-validation.

Directions for Future Work

- Investigation of iIS and ordinary IS in many other models with unit-specific latent variables, including factor models, hidden Markov models, stochastic volatility models, and other time series models.
- Use of CV predictive p-values to define "residuals" for model diagnostics, as alternatives to Pearson's and deviance residuals. The attractiveness is that CV predictive p-values are always *uniformly distributed* when the model is right for the dataset.
- Other methods to improve importance sampling in more general situation. A recent proposal by Vehtari and Gelman (2015): truncating large importance weight.
- Determine thresholds for CVIC.

References

- Vehtari, A. and Ojanen, J. (2012), "A survey of Bayesian predictive methods for model assessment, selection and comparison," Statistics Surveys, 6, 142-228.
- 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. (2014), "Understanding predictive information criteria for Bayesian models," Statistics and Computing, 24(6), 997-1016
- Li, L., Qiu, S.*, Zhang, B.*, and Feng, C.X. (2016+). Approximating Cross-validatory Predictive Evaluation in Bayesian Latent Variables Models with Integrated IS and WAIC, Statistics and Computing, online first.
- Vehtari, A., & Gelman, A. (2015). Pareto Smoothed Importance Sampling. arXiv:1507.02646 [stat].