Estimating Cross-validatory Predictive p-values with Integrated Importance Sampling for Disease Mapping Models

Longhai Li

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

Presented on 11 August 2019
The 4th ICSA-Canada-Chapter Symposium
Queen's University

Acknowledgements

- Thanks to my co-authors: **Cindy X. Feng and Shi Qiu**.
- The work was supported by grants from Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Foundation for Innovation (CFI).

Outline

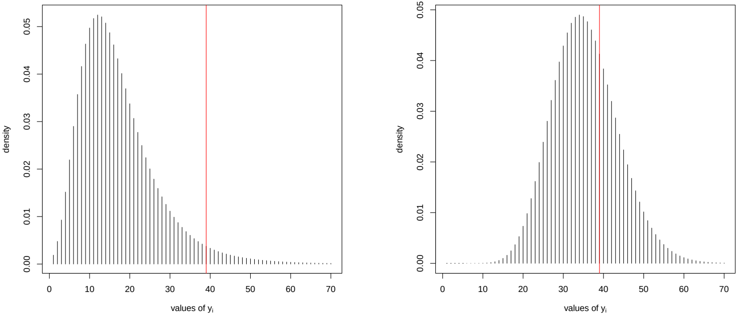
1. Predictive *p*-value
2. A Disease Mapping Model
3. Methods for Computing Predictive *p*-values
    - Posterior Predictive Checking VS LOOCV Checking
    - Non-integrated Importance Sampling
    - Ghosting Method
    - Integrated (Marginalized) Importance Sampling
4. Numerical Comparisons with Two Real Datasets
    - Lip Cancer Data in Scottland
    - Larynx Cancer Data in Germany
5. Conclusions and Future Work
6. References

Section 1

Predictive *p*-value

Predictive *p*-value

- Predictive *p*-value is the tail probability of a predictive distribution:



- Predictive *p*-value can be used to
  - check model
  - identify "outliers" (or divergent regions in disease mapping problems)

Section 2

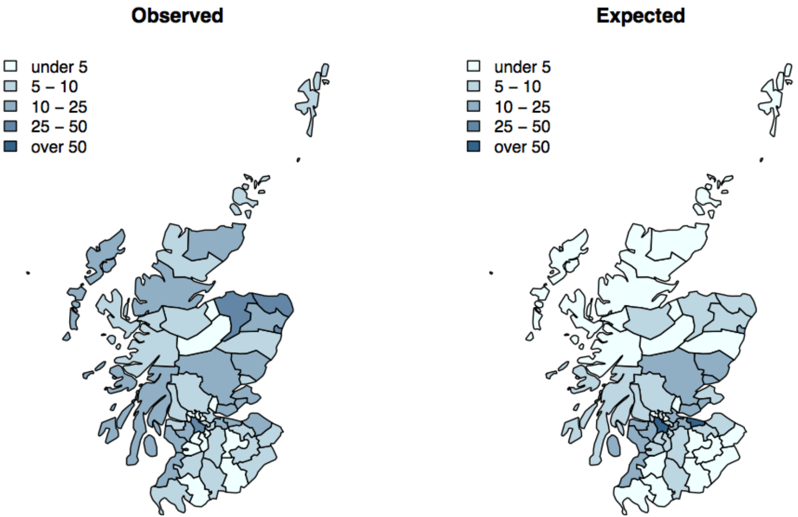
A Disease Mapping Model

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** $s_i = \log(\lambda_i)$

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

where Φ 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

Methods for Computing Predictive *p*-values

Subsection 1

Posterior Predictive Checking VS LOOCV Checking

Posterior Predictive Checking

- The **full data posterior** density of $(\boldsymbol{s}_{1:n}, \boldsymbol{\theta})$ given observations $\boldsymbol{y}_{1:n}^{\text{obs}}$ is given by:

$$P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n} | \boldsymbol{y}_{1:n}^{\text{obs}}) = \prod_{j=1}^{n} P_y(y_j^{\text{obs}} | s_j, \boldsymbol{\theta}) P_s(\boldsymbol{s}_{1:n} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \,/\, C_1. \quad (1)$$

- Predictive p-value for y_i given parameters and latent variable:

$$p\text{-value}(y_i^{\text{obs}} | \boldsymbol{\theta}, s_i) = Pr(y_i > y_i^{\text{obs}} | \boldsymbol{\theta}, s_i) + 0.5 Pr(y_i = y_i^{\text{obs}} | \boldsymbol{\theta}, s_i). \quad (2)$$

- The posterior predictive p-value:

$$p\text{-value}^{\text{Post}}(y_i^{\text{obs}}) = E_{\text{post}}(p\text{-value}(y_i^{\text{obs}} | \boldsymbol{\theta}, s_i)), \quad (3)$$

- Given MCMC samples $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{s}_{1:n}^{(t)}); t = 1, \ldots, T\}$ from the full data posterior (1), posterior predictive p-value (3) is computed as follows:

$$\widehat{p\text{-value}}^{\text{Post}}(y_i^{\text{obs}}) = \frac{\sum_t^T p\text{-value}(y_i^{\text{obs}} | \boldsymbol{\theta}^{(t)}, s_i^{(t)})}{T}. \quad (4)$$

Optimistic Bias (Conservatism) in Posterior Checking

- Posterior predictive checking uses the dataset twice: y_i^{obs} is used to obtain the posterior predictive distribution of y_i, and is also used to test itself. Generally, y_i^{obs} appears better predictable by the model.
- The posterior predictive p-values are concentrated around 0.5 rather than uniformly distributed on the interval (0,1).
- Failure in identifying “outliers”

LOOCV Predictive p-value (Gold Standard)

- LOOCV distribution $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ is given as follows:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}}) = \prod_{j=1,\ldots,i-1,i+1,\ldots,n} P_y(y_j^{\text{obs}}|s_j, \boldsymbol{\theta}) P_s(\boldsymbol{s}_{1:n}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \,/\, C_2. \tag{5}$$

- The LOOCV predictive p-value for y_i^{obs}

$$p\text{-value}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = E_{\text{post(-i)}}(p\text{-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)). \tag{6}$$

- Given MCMC samples $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{s}_{1:n}^{(t)}); t = 1, \ldots, T\}$ from the LOOCV posterior (5), the LOOCV predictive p-value is computed as follows:

$$\widehat{p\text{-value}}^{\text{CV}}(y_i^{\text{obs}}) = \frac{\sum_t^T p\text{-value}(y_i^{\text{obs}}|\boldsymbol{\theta}^{(t)}, s_i^{(t)})}{T}. \tag{7}$$

- MCMC sampling needs to be redone for each y_i^{obs}.

Subsection 2

Non-integrated Importance Sampling

Review of Importance Sampling

- Our goal is to find the expectation of a function $a(X)$ when X has a probability density proportional to $f(x)$, denoted by

$$E_f\left(a(X)\right)$$

- Instead of drawing samples from $f(x)$, we draw samples from an approximate distribution with a probability density $g(x)$.
- Importance weight:

$$W(x) = \frac{f(x)}{g(x)}$$

- *Importance reweighting formula*

$$E_f\left(a(X)\right) \quad = \quad \frac{E_g\left(a(X)W(X)\right)}{E_g\left(W(X)\right)}. \tag{8}$$

- The intuitive explanation of the *importance reweighting formula* (8) is that samples that are more compatible with the target distribution f will be assigned more weight (and vice versa).

Non-integrated Importance Sampling

- We can estimate expectations with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}|\boldsymbol{y}^{\text{obs}}_{-i})$ in (5) by reweighting samples from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}|\boldsymbol{y}^{\text{obs}}_{1:n})$ (1):

$$p\text{-value}(y_i^{\text{obs}}|\boldsymbol{y}^{\text{obs}}_{-i}) = \frac{E_{\text{post}}\left[p\text{-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)\, W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_i)\right]}{E_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_i)\right]}, \qquad (9)$$

- $W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_i)$ is the ratio:

$$W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_i) = \frac{P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}|\boldsymbol{y}^{\text{obs}}_{-i})}{P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}|\boldsymbol{y}^{\text{obs}}_{1:n})} \times \frac{C_2}{C_1} = \frac{1}{P_y(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)}. \qquad (10)$$

- Given MCMC samples $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{s}^{(t)}_{1:n}); t = 1, \ldots, T\}$ from the full data posterior (1), the nIS predictive p-value (9) is computed as follows:

$$\widehat{p\text{-value}}^{\text{nIS}}(y_i^{\text{obs}}) = \frac{\sum_t^T \left[p\text{-value}(y_i^{\text{obs}}|\boldsymbol{\theta}^{(t)}, s_i^{(t)})\ W_i^{\text{nIS}}(\boldsymbol{\theta}^{(t)}, \boldsymbol{s}_i^{(t)})\right] \Big/ T}{\sum_t^T W_i^{\text{nIS}}(\boldsymbol{\theta}^{(t)}, \boldsymbol{s}_i^{(t)}) \Big/ T}. \qquad (11)$$

Problems in Importance Sampling

- Most of the MCMC samples of $\boldsymbol{s}_i$ are largely bound to regions that fit the observation y_i^{obs} well.
- The estimate (11) is dominated by a single or a few very incompatible MCMC samples.
- The "effective" sample size in (11) may be very small.
- This leads to the notorious instability problem of importance sampling.

Subsection 3

Ghosting Method

Ghosting Method

- Marshall and Spiegelhalter (2007) propose that the $\boldsymbol{s}_i$ is replaced with a re-generated $\boldsymbol{s}_i$ without reference to y_i^{obs}. Technically, draw samples from the "ghosting" distribution of $(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})$:

$$P_{\text{ghost}}(\boldsymbol{s}_{1:n}, \boldsymbol{\theta}) = P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i} | \boldsymbol{y}_{1:n}^{\text{obs}}) \times P(\boldsymbol{s}_i | \boldsymbol{s}_{-i}, \boldsymbol{\theta}), \tag{12}$$

- The "ghosting" predictive p-value

$$p\text{-value}^{\text{ghost}}(y_i^{\text{obs}}) = E_{\text{ghost}}\left(p\text{-value}(y_i^{\text{obs}} | \boldsymbol{\theta}, s_i)\right), \tag{13}$$

- Given "ghosting" samples $\{(\boldsymbol{\theta}^{(t)}, \tilde{\boldsymbol{s}}_i^{(t)}); t = 1, \ldots, T\}$, the ghosting predictive p-value is computed as follows:

$$\widehat{p\text{-value}}^{\text{ghost}}(y_i^{\text{obs}}) = \frac{\sum_t^T p\text{-value}(y_i^{\text{obs}} | \boldsymbol{\theta}^{(t)}, \tilde{\boldsymbol{s}}_i^{(t)})}{T}. \tag{14}$$

Subsection 4

Integrated (Marginalized) Importance Sampling

Intuition in Integrated Importance Sampling (iIS)

- The "ghosting" predictive p-value is not equivalent to the LOOCV p-value in theory.
- Particularly, the MCMC sample of $(\boldsymbol{\theta}^{(t)}, \boldsymbol{s}_{-i}^{(t)})$ still contain information of y_i^{obs}.
- $P_y(y_i^{\text{obs}}|\boldsymbol{s}_i, \boldsymbol{\theta})$ is integrated out with respect to the distribution of $\boldsymbol{s}_i$ given $\boldsymbol{\theta}$ without reference to the actual observation y_i^{obs}.
- The **integrated predictive density** of y_i^{obs}:

$$P(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i}) = \int P_y(y_i^{\text{obs}}|s_i, \boldsymbol{\theta}) P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta}) d\ s_i. \tag{15}$$

- We use the **integrated predictive density** of y_i^{obs} to correct for the bias in $\boldsymbol{\theta}, \boldsymbol{s}_{-i}$ due to inclusion of y_i^{obs} in full data posterior.

iIS Predictive p-value

- For each MCMC sample, we first generate two sets of new $\boldsymbol{s}_i$ from $P(\boldsymbol{s}_i|\boldsymbol{s}_{-i}^{(t)}, \boldsymbol{\theta}^{(t)})$, denoted by $\{\tilde{\boldsymbol{s}}_i^{(A,k)}; k = 1, \dots, R\}$ and $\{\tilde{\boldsymbol{s}}_i^{(W,k)}; k = 1, \dots, R\}$ respectively.
- Computing integrated p-value and importance weight:

$$\widehat{A}_i^{(t)} = \frac{\sum_{k=1}^{R} p\text{-value}(y_i^{\text{obs}}|\boldsymbol{\theta}^{(t)}, \tilde{s}_i^{(A,k)})}{R} \tag{16}$$

$$\widehat{W}_i^{(t)} = 1 \Big/ \frac{\sum_{k=1}^{R} P_y(y_i^{\text{obs}}|\boldsymbol{\theta}^{(t)}, \tilde{s}_i^{(W,k)})}{R}. \tag{17}$$

- The LOOCV p-value is then estimated as follows:

$$\widehat{p\text{-value}}^{\text{iIS}}(y_i^{\text{obs}}) = \frac{\sum_{t=1}^{T} \widehat{A}_i^{(t)} \widehat{W}_i^{(t)} \Big/ T}{\sum_{t=1}^{T} \widehat{W}_i^{(t)} \Big/ T}. \tag{18}$$

- It can be shown that this estimate asymptotically equals to the true LOOCV predictive p-value when $T \to \infty$

Section 4

Numerical Comparisons with Two Real Datasets

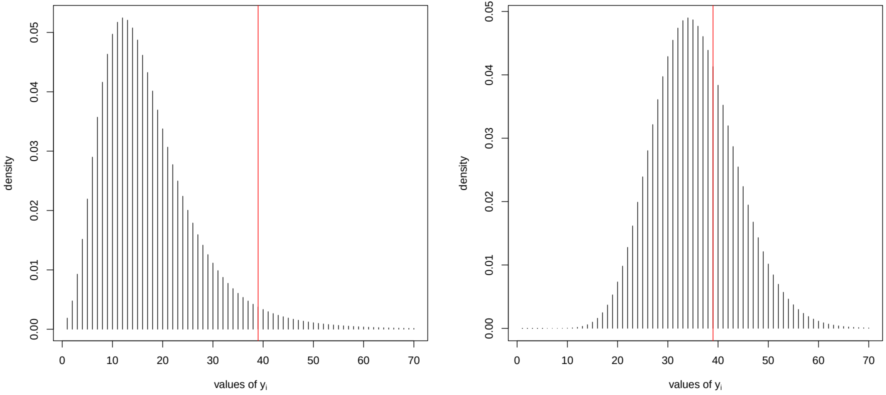
Subsection 1

Lip Cancer Data in Scottland

Estimated predictive *p*-values(y_i^{obs})

ID	LOOCV	PCH	GHO	nIS	iIS	ID	LOOCV	PCH	GHO	nIS	iIS
1	0.308	0.417	0.310	0.319	0.307	29	0.667	0.547	0.651	0.631	0.664
2	0.033	**0.320**	**0.050**	**0.074**	0.030	30	0.260	0.367	0.278	0.263	0.262
3	0.090	0.325	0.096	0.089	0.090	31	0.275	0.359	0.283	0.262	0.274
4	0.418	0.437	0.423	0.430	0.417	32	0.816	0.601	0.799	0.768	0.818
5	0.139	0.357	0.155	0.159	0.140	33	0.469	0.455	0.467	0.466	0.463
6	0.512	0.463	0.512	0.458	0.514	34	0.188	0.317	0.211	0.189	0.190
7	0.060	0.312	0.072	**0.041**	0.058	35	0.370	0.414	0.372	0.364	0.370
8	0.113	0.313	0.114	0.112	0.112	36	0.151	0.284	0.162	0.154	0.149
9	0.267	0.386	0.281	0.261	0.271	37	0.596	0.524	0.590	0.598	0.601
10	0.269	0.405	0.279	0.300	0.267	38	0.071	0.221	0.092	0.076	0.073
11	0.127	0.334	0.137	0.138	0.122	39	0.820	0.627	0.794	0.804	0.821
12	0.514	0.458	0.518	0.445	0.515	40	0.182	0.285	0.192	0.181	0.178
13	0.484	0.433	0.485	0.412	0.479	41	0.376	0.413	0.384	0.375	0.376
14	0.474	0.455	0.472	0.451	0.477	42	0.991	**0.853**	0.977	0.987	0.992
15	0.061	0.280	0.070	0.056	0.062	43	0.880	0.699	0.872	0.866	0.883
16	0.578	0.496	0.571	0.540	0.578	44	0.599	0.532	0.585	0.588	0.593
17	0.609	0.473	0.602	0.536	0.606	45	0.962	**0.798**	**0.904**	0.973	0.971
18	0.138	0.303	0.146	0.144	0.136	46	0.802	0.664	0.788	0.807	0.802
19	0.369	0.422	0.378	0.373	0.366	47	0.510	0.470	0.506	0.506	0.511
20	0.271	0.366	0.277	0.245	0.271	48	0.687	0.598	0.684	0.692	0.688
21	0.133	0.309	0.139	0.127	0.129	49	0.987	**0.865**	**0.949**	0.983	0.987
22	0.734	0.572	0.695	0.700	0.744	50	0.954	**0.819**	**0.930**	0.951	0.955
23	0.382	0.427	0.390	0.381	0.384	51	0.590	0.519	0.586	0.581	0.591
24	0.106	0.278	0.140	0.118	0.109	52	0.574	0.512	0.571	0.576	0.575
25	0.075	0.259	0.093	0.079	0.073	53	0.757	0.657	0.748	0.750	0.757
26	0.049	**0.224**	**0.061**	**0.052**	0.048	54	0.847	0.739	0.837	0.841	0.847
27	0.244	0.348	0.250	0.248	0.244	55	0.990	**0.923**	0.987	0.990	0.991
28	0.305	0.383	0.315	0.302	0.308	56	0.841	0.728	0.833	0.826	0.842

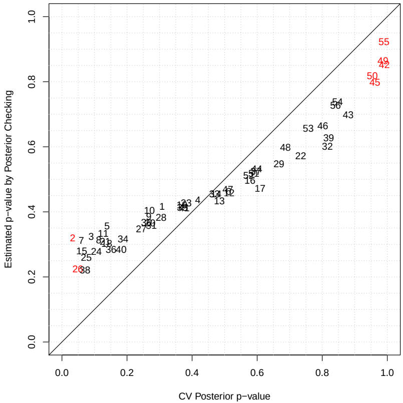
Illustration of Optimistic Bias in Posterior Checking



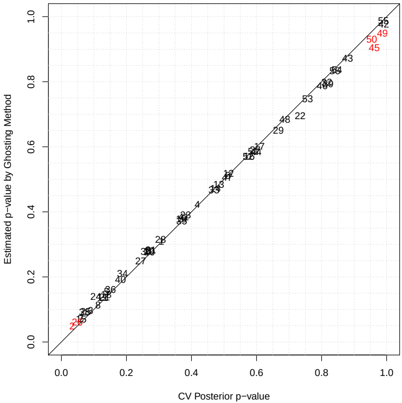
(a) CV predictive PMF of y_2

(b) Posterior Checking PMF of y_2

Comparing Estimated *p*-values with LOOCV *p*-values I

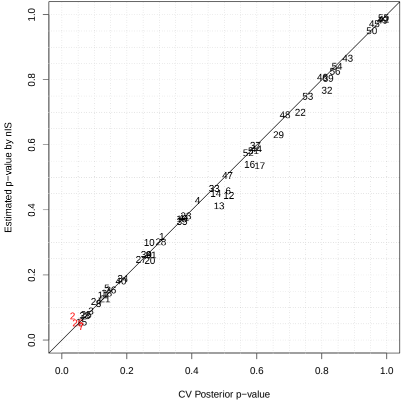


(c) Posterior checking

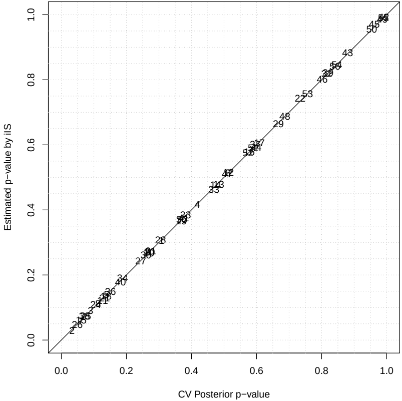


(d) Ghosting method

Comparing Estimated *p*-values with LOOCV *p*-values II



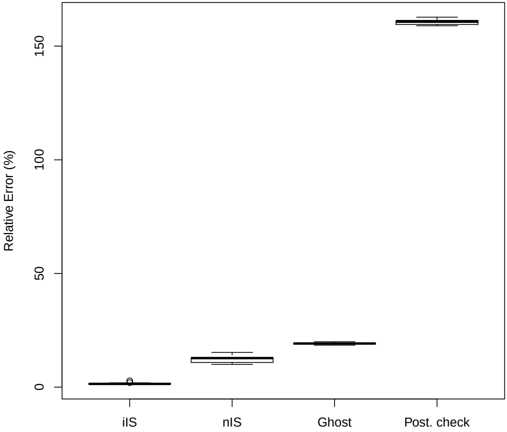
(e) Non-integrated IS (nIS)



(f) 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

Table 1: Comparison of computation time (in seconds). (Abbreviations: LOOCV: actual cross validation, PCH: posterior predictive checking, GHO: Ghosting, nIS: naive importance sampling and iIS: integrated importance sampling).

	LOOCV	PCH	nIS	GHO	iIS
MCMC fitting	1138	20	20	20	20
Computing *p*-values	1	1	1	84	144
Total	1139	21	21	104	164

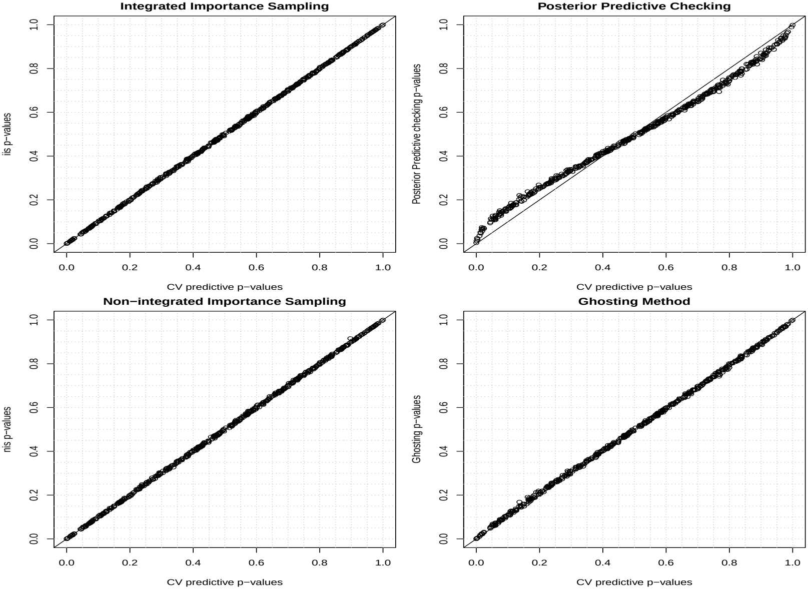
Subsection 2

Larynx Cancer Data in Germany

Dataset Information

- $N = 544$ districts.
- y_i: number of larynx cancer mortality counts
- x_i: level of smoking consumption

Comparing Estimated *p*-values with LOOCV *p*-values



Comparing Mis-categorization

Table 2: Contingency table of districts categorized by cutting predictive *p*-values with 0.1 and 0.9 in German larynx cancer example. The bolded numbers show the mis-categorized districts compared to CV.

	Posterior predictive checking			Ghosting method			nIS			iIS		
CV	[0,0.1)	[0.1,0.9)	[0.9,1]	[0,0.1)	[0.1,0.9)	[0.9,1]	[0,0.1)	[0.1,0.9)	[0.9,1]	[0,0.1)	[0.1,0.9)	[0.9,1]
[0, 0.1)	16	**31**	0	42	**5**	0	47	0	0	47	0	0
[0.1, 0.9)	0	455	0	0	455	0	0	454	**1**	0	455	0
[0.9, 1]	0	**21**	21	0	**3**	39	0	0	42	0	0	42

Computation Time

Table 3: Comparison of computation time in larynx cancer example.

	LOOCV	PCH	nIS	GHO	iIS
MCMC fitting (seconds)	4.8×10^6	7816	7816	7816	7816
Computing *p*-values (seconds)	2	2	2	284	522
Total (hours)	1333	2.17	2.17	2.25	2.32
Total (relative to CV)	1	0.162%	0.162%	0.168%	0.173%

Conclusion and Discussion

- Non-integrated IS by treating latent variables as parameters may give wrong results in predictive model assessment.
- The new proposed iIS can 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.
- The iIS method can be applied to many other models with correlated or independent random effects provided that the random effect is specific to each test observation or unit, for example the zero-inflated models, which is a special case of mixture models.
- iIS and Ghosting method are not yet applicable in models with complex structure in latent variables. For such models, non-integrated importance sampling is still valid. There is an improved importance sampling that is more widely applicable: **"Pareto smoothed importance sampling"** (Vehtari and Gelman, 2015).

References

- Li, L., Feng, C.X., Qiu, S.* (2017), Estimating Cross-validatory Predictive *p*-values with Integrated Importance Sampling for Disease Mapping Models. Statistics in Medicine, Volume 36, Issue 14, Pages 2220-2236.
- 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, Volume 26, Issue 4, pp 881-897.
- Vehtari, A., & Gelman, A. (2015). Pareto Smoothed Importance Sampling. arXiv:1507.02646 [stat].
- Vehtari A, Gelman A, Gabry J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. Sep 1;27(5):1413-32.