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

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the Department of Mathematics and Statistics

University of Saskatchewan

Saskatoon

By

Shi Qiu

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
Room 142 McLean Hall
106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

The well-documented problems associated with mapping raw rates of disease have resulted in an increased use of Bayesian hierarchical models to produce maps of "smoothed" estimates of disease rates. Two statistical problems arise in using Bayesian hierarchical models for disease mapping. The first problem is in comparing goodness of fit of various models, which can be used to test different hypotheses. The second problem is in identifying outliers/divergent regions with unusually high or low residual risk of disease, or those whose disease rates are not well fitted. The results of outlier detection may generate further hypotheses as to what additional covariates might be necessary for explaining the disease. Leave-one-out cross-validatory (LOOCV) model assessment has been used for these two problems. However, actual LOOCV is time-consuming. This thesis introduces two methods, namely iIS and iWAIC, for approximating LOOCV, using only Markov chain samples simulated from a posterior distribution based on a full data set. In iIS and iWAIC, we first integrate the latent variables without reference to holdout observation, then apply IS and WAIC approximations to the integrated predictive density and evaluation function. We apply iIS and iWAIC to two real data sets. Our empirical results show that iIS and iWAIC can provide significantly better estimation of LOOCV model assessment than existing methods including DIC, Importance Sampling, WAIC, posterior checking and Ghosting methods.

Acknowledgements

I cannot express enough thanks to my supervisor, Dr. Longhai Li, for his continued encouragement and mentorship. He encouraged me to not only grow as a Master student, but also as an instructor and a professional researcher. I was given the opportunity to develop my own individuality and self-sufficiency by his mentorship. This thesis could not have been written without his invaluable advice and patient guidance. I offer my sincere appreciation for the learning opportunities provided by Dr. Longhai Li.

I would also like to thank my co-supervisor, Dr. Cindy X. Feng, for her assistance and support in the topic of disease mapping, and extend my appreciation to a member of my committee: Dr. M.G. Bickis, for his support and suggestions. I thank the Department of Mathematics and Statistics for the funding provided through my Master training.

I am grateful to all of the professors, graduate students, and staff in the Department of Mathematics and Statistics, especially Dr. Chris Soteros, Dr. Juxin Liu, Ph.D student Lai Jiang who have provided me advice and help with my thesis.

Finally, I sincerely thank my parents, friends and girlfriend, who have loved me and supported me.

May God bless all of you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Disease mapping

Mapping of disease incidence mortality rates is of primary importance in many epidemiological studies. The use of crude rates to estimate rare disease risks in small areas, such as health units, census areas or administrative zones, is problematic since it does not account for the high variability of population sizes over the different regions, nor the spatial patterns of the regions under study. For this reason, interpretation of the spatial distribution of disease based on crude estimates is often misleading. Alternatively, Bayesian inference is widely used to produce stabilized risk maps by borrowing information from neighbourhoods across the map. Early developments of disease mapping methodology included the use of empirical Bayes (EB) techniques to estimate parameters, and a plug-in approximation of these for posterior inference, which yielded unbiased estimates of the relative risks. However, the variance of these estimates were underestimated, since the EB approach doses not account for the uncertainty arising from estimating hyperparameters. In recent years, fully Bayesian approaches have gained prominence. Inference is based on Markov chain Monte Carlo (MCMC) algorithms (Congdon, 2006) [1]. Bayesian methods for disease mapping are often termed as hierarchical spatial models. In this scenario, the goodness of fit of various models becomes salient for describing disease mapping. Therefore, model comparison methods must be carefully considered. For example, transmission of many vector borne diseases have multiple factor vectors related to environmental conditions. However, multiple potential models emerge considering these factors as hypothesis effects. There rises urgency in comparing these available models in goodness of fit. Jeefoo et al. (2010) [10] studied diffusion patterns of Dengue in Thailand and examined the spatial-temporal diffusion pattern. They affirm that outbreaks of Dengue are attributed to various factors, such as climate, breeding site density probability, urbanization and human population movements. These findings indicate that models for Dengue diffusion patterns can be constructed from multiple effects. In this thesis, we are interested in methods of comparing different models for spatial data to answer hypothesis-testing questions on whether effects(including spatial effect itself) are significant or not. In Chapter 3, we use two typical data to apply our methods for model comparison.

Disease mapping can also provide monitoring in public health surveillance by identifying the outliers/divergent regions of infectious and chronic diseases. Detection of disease divergent regions for public health surveillance is greatly helpful in improving the explaining of properties of a disease. For example, in one of the earliest studies of disease mapping in 1854, Dr. John Snow labelled on a dot-map residential addresses of people who died from cholera in London. He identified the Broad Street pump as the source of an intense cholera outbreak (McLeod, 2000) [15]. In recent years, a study found that chronic exposure to solar radiation might be a major risk factor in the development of lip cancer. This finding was supported by geographic residence of farmers, fishers and outdoor labours, compared to a global map of lip cancer incidences (Moore et al., 1999) [16]. The wide use of Bayesian methods for disease mapping has driven largely the improvement of outlier/divergent regions detection techniques. These techniques are highly demanded due to their better accuracy and high efficiency. For these practical considerations, model comparison and divergent region detection are two key tasks of model evaluation that need to be analysed carefully with reliable and precise methods.

1.2 Review of Model Comparison and Outlier Detection Methods

Spatial Bayesian models can be evaluated in several ways. One of the most popular ways to evaluate a model is to evaluate its predictive accuracy. Several measures are available to estimate the expected predictive accuracy without out-of-sample data. We list several reasonable-seeming predictive accuracy measures in three categories.

- **Within-sample predictive accuracy**

  A natural estimate of predictive density for out-of-sample data is the predictive density for observed sample data, which typically uses posterior predictive density in model evaluation. This is simple to understand, but is, in general, an overestimate of expected predictive density for out-of-sample data (optimistic biased) because it is evaluated on the data in which the model was fit.

- **Adjusted within-sample predictive accuracy**

  Since using posterior predictive density is a biased estimate of expected predictive density, the next logical step is to correct that bias. For instance, AIC, DIC, WAIC and other methods aim to correct optimistic bias by starting with posterior predictive density and then subtracting a correction for the number of parameters, or the effective number of parameters. These adjustments can give reasonable answers in many cases, but only in best scenarios.

- **Cross-validation**

  Cross-validation is a natural way to approximate out-of-sample predictive performance of a model. One must fit the model to training data set and then evaluate predictive accuracy on a holdout data set, where training data set and holdout data are departed from observed sample data set; then one must repeat this procedure with each holdout data set and summarize all subset predictive accuracy. Cross-validatory evaluation avoids the problem of optimistic bias. However, cross-validation can be computationally expensive: to get a stable estimate typically requires many data partitions and fits.

1.2.1 Model Comparison Methods

Suppose we have a simple Bayesian model without latent variables where $\boldsymbol{y}_{1:n}^{\text{obs}} = (y_1^{\text{obs}}, \ldots, y_n^{\text{obs}})$ denotes observation data and $\boldsymbol{y} = (y_1, \ldots, y_n)$ models independent random variables given parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$; thus probability density of $\boldsymbol{y}$ is $P(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{\theta})$. The posterior of $\boldsymbol{\theta}$, given full observations data $\boldsymbol{y}^{\text{obs}}$ is:

$$P_{\text{post}}(\boldsymbol{\theta}) = \prod_{j=1}^{n} P(y_j^{\text{obs}}|\theta)P(\boldsymbol{\theta})/C, \tag{1.1}$$

where C is normalizing constant. For historical reasons, model comparison methods usually treat predictive accuracy as information criteria(IC) which are typically defined based on the deviance of log predictive density of the observed data, multiplied by -2; that is, $-2\log P(\boldsymbol{y}^{\text{obs}}|\boldsymbol{\theta})$.

- **Leave-one-out cross-validation(LOO-CV)**

  In Bayesian cross-validation, the sample data are repeatedly partitioned into n, in which each holdout represents a single test case. Then the model is fit to the training set, with this fit to evaluate the predictive density of the holdout data. Assuming the holdout data observation is y_i^{obs} and training set is $\boldsymbol{y}_{-i}^{\text{obs}} = (y_1^{\text{obs}}, \ldots, y_{i-1}^{\text{obs}}, y_{i+1}^{\text{obs}}, \ldots, y_n^{\text{obs}})$, the posterior distribution for cross-validation is expressed as:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}) = \prod_{j \neq i} P(y_j^{\text{obs}}|\theta_j)P(\theta)/C_2 \tag{1.2}$$

  where C_2 is the normalizing constant involving only with $\boldsymbol{y}_{-i}$. We give LOO-CV information criterion based on $P(y_i^{\text{obs}}|\boldsymbol{\theta})$ with respect to the posterior distribution $P_{\text{post(-i)}}(\boldsymbol{\theta})$:

$$\text{LOO-CVIC} = -2\sum_{i=1}^{n} \log \int P(y_i^{\text{obs}}|\boldsymbol{\theta})P_{\text{post(-i)}}(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{1.3}$$

  which can also be expressed as

$$\text{LOO-CVIC} = -2\sum_{i=1}^{n} \log E_{\text{post(-i)}}[P(y_i^{\text{obs}}|\boldsymbol{\theta})], \tag{1.4}$$

  where $E_{\text{post(-i)}}[\ ]$ is integration over the posterior distribution $P_{\text{post(-i)}}(\boldsymbol{\theta})$. Hence, we notice that LOO-CV requires n times of the model fits, which requires much computing.

- **Deviance information criterion(DIC)**

DIC (Spiegelhalter et al., 2002) [19], is an information criterion of adjusted within-sample predictive accuracy, taking the formula:

$$\mathrm{DIC} = -2\big(\log P(\boldsymbol{y}_{1:n}^{\mathrm{obs}}|\boldsymbol{\theta}^{Bayes}) - p_{DIC}\big), \tag{1.5}$$

where $\boldsymbol{\theta}^{Bayes}$ is the posterior mean of $\boldsymbol{\theta}$ which is plugged into the predictive density function. p_{DIC} is the effective number of parameters for adjusting predictive accuracy, which is defined as

$$p_{DIC} = 2(\log P(\boldsymbol{y}_{1:n}^{\mathrm{obs}}|\boldsymbol{\theta}^{Bayes}) - E_{\mathrm{post}}(\log P(\boldsymbol{y}_{1:n}^{\mathrm{obs}}|\boldsymbol{\theta}))), \tag{1.6}$$

where the expectation in the second term E_{post} is an integral of $\log p(\boldsymbol{y}^{\mathrm{obs}}|\theta)$ over full posterior distribution $P_{\mathrm{post}}(\boldsymbol{\theta})$, that is:

$$E_{\mathrm{post}}[\log P(\boldsymbol{y}_{1:n}^{\mathrm{obs}}|\boldsymbol{\theta})] = \int \log P(\boldsymbol{y}_{1:n}^{\mathrm{obs}}|\boldsymbol{\theta}) P_{\mathrm{post}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{1.7}$$

The posterior mean of $\boldsymbol{\theta}$ will produce the maximum log predictive density when it happens to be same as the mode, and negative p_{DIC} can be produced if posterior mean is far from the mode. In addition, DIC become the most popular choice in Bayesian model comparisons since it is implemented in WinBUGS.

- **Importance sampling(IS)**

Importance sampling adjust predictive accuracy by importance weighting technique (Gelfand et al., 1992) [3], approximating CV prediction evaluation. Importance sampling information criterion calculates posterior predictive density with importance weighting point-wisely when averaging $\boldsymbol{\theta}$ over its posterior distribution:

$$\begin{aligned} \text{IS-IC} &= -2\sum_{i=n}^{n} \log \frac{E_{\mathrm{post}}[P(y_i^{\mathrm{obs}}|\boldsymbol{\theta})W_i]}{E_{\mathrm{post}}(W_i)} && (1.8) \\ &= -2\sum_{i=n}^{n} \log \frac{1}{E_{\mathrm{post}}[1/P(y_i^{\mathrm{obs}}|\boldsymbol{\theta})]} && (1.9) \end{aligned}$$

where, E_{post} is an integral to full data posterior mentioned in (1.1). W_i is importance weighting for random variable y_i. This case is that:

$$W_i = \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta})}. \tag{1.10}$$

- **Widely applicable information criterion(WAIC)**

WAIC stands for the approximate cross-validation approach for estimating the out-of-sample information criterion, starting with the computed pointwise posterior predictive density, then adding a correction for effective number of parameters to adjust for the bias from using the data twice:

$$\text{WAIC} = -2\sum_{i=1}^{n} \log E[P(y_i^{\text{obs}}|\boldsymbol{\theta})] + 2p_{WAIC}, \tag{1.11}$$

where p_{WAIC} represents a adjustment for effective number of parameters. There are two adjustments proposed in the literature:

$$\begin{aligned} p_{WAIC1} &= 2\sum_{i=1}^{n}\{\log(E_{\text{post}}[P(y_i^{\text{obs}}|\boldsymbol{\theta})] - E_{\text{post}}[\log P(y_i^{\text{obs}}|\boldsymbol{\theta})]\}, \text{ or} && (1.12)\\ p_{WAIC2} &= \sum_{i=1}^{n} V_{\text{post}}[\log P(y_i^{\text{obs}}|\boldsymbol{\theta})], && (1.13) \end{aligned}$$

where V_{post} stands for variance over $\boldsymbol{\theta}$ with respect to $P_{\text{post}}(\boldsymbol{\theta})$. Watanabe (2010) [26] has proven that WAIC is equivalent to LOO-CVIC asymptotically as random variables of training data. However, WAIC is only justified for problems where observed data are independently distributed with a population distribution. In this thesis, we will handle a problem where $y_1, \ldots, y_n$ are not independent given $\boldsymbol{\theta}$.

1.2.2 Divergent Region Detection Methods

The divergent region for disease mapping is usually termed as outlier in Bayesian spatial modelling. We have noticed that some of the methods above are based on adjustments of pointwise predictive density. Outlier detection is also an object related to pointwise prediction used for checking for whether the observed data point is at the extreme tails of the predictive distributions. Hence, we define a p-value as tail probability, that is p-value$(y_i^{\text{obs}}) = Pr(y_i \geq y_i^{\text{obs}}|\boldsymbol{\theta})$. We can use CV as an outlier detection method by looking at the predictive density of y_i^{obs} and all predictive densities of y_i that are greater than the observation y_i^{obs} value.

- **Leave-one-out cross-validation(LOO-CV)**:

  LOO-CV posterior predictive p-value is:

$$\text{p-value}^{\text{CV}}(y_i) = E_{\text{post(-i)}}[Pr(y_i \geq y_i^{\text{obs}}|\boldsymbol{\theta})] \tag{1.14}$$

- **Importance Sampling(IS)**:

  Importance sampling as mentioned above, can check the p-value by:

$$\text{p-value}^{\text{IS}}(y_i) = \frac{E_{\text{post}}[Pr(y \geq y_i^{\text{obs}}|\boldsymbol{\theta})W_i]}{E_{\text{post}}(W_i)}, \text{ where} \tag{1.15}$$

$$W_i = \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta})}. \tag{1.16}$$

- **Posterior checking**:

  Gelman et al. (1996) [5] provide a method which applies posterior predictive assessment in evaluating Bayesian models. We will call this method *posterior checking*, and use the subscript $^{\text{Post.check}}$ to denote application of posterior checking. One of the simplest formulas for estimating the p-value of y_i is:

$$\text{p-value}^{\text{Post.check}}(y_i) = E_{\text{post}}[Pr(y \geq y_i^{\text{obs}}|\boldsymbol{\theta})], \tag{1.17}$$

  which uses posterior checking for outlier detection. Gelman et al. (1996) [5] do not recommend this use of posterior checking because it involves double-use of the data, which leads to optimistic bias. However, due to convenience, this is often used in practice.

- **Ghosting**:

  Ghosting is a method mixing the CV and posterior checking approaches by Marshall and Spiegelhalter (2003) [13]. This method is applied to the models with latent variable

b_i. We use $^{\text{Ghost}}$ to denote the application of the Ghosting method. The formula is:

$$\text{p-value}(y_i)^{\text{Ghost}} = \int P(y_i \geq y_i^{\text{obs}} | \boldsymbol{\theta}, b_i) P(b_i | \boldsymbol{b}_{-i}, \boldsymbol{\theta}) P(\boldsymbol{b}_{-i}, \boldsymbol{\theta} | \boldsymbol{y}_{1:n}^{\text{obs}}) d\boldsymbol{\theta} db_i \tag{1.18}$$

where y_i^{obs} in the right side of equation is one of the observations in $\boldsymbol{y}_{1:n}^{\text{obs}}$. Ghosting avoids double use y_i directly and corrects optimistic bias to some extent.

1.3 Contribution of this thesis

LOO-CV is a natural way to approximate out-of-sample predictive evaluation in Bayesian spatial models and is recognized as a golden standard for other existing methods which aim to correct for optimistic bias. Moreover, the shortage of LOO-CV in expensive computing becomes the motivation for us to improve approximating CV methods. In the following context, we will use CV to represent LOO-CV, as this is the only method we will discuss.

In this thesis, we introduce two methods based on IS and WAIC for use in Bayesian spatial models. IS and WAIC can be simply applied to the predictive density of observed data. This data is modelled conditional not only on model parameters, but also on latent variables. However, actual validation observation units often bring optimistic bias into their latent variables in IS and WAIC methods. One remedy to eliminate the bias in the latent variables associated with the validatory units is to temporarily discard the latent variables in the full data posterior sample. One must integrate away the latent variables with respect to the conditional distribution of the latent variables associated with the validatory units conditional on only the model parameters *but not the actual observations.* This integration will lead to an *integrated* predictive density and *integrated* evaluation function, which result in two predictive evaluation methods: Integrated Importance Sampling (iIS) and Integrated WAIC (iWAIC). The required integrals can be obtained analytically in some models using Monte Carlo methods or other numerical methods.

Vehtari et al. (2001) [24] and Vanhatalo et al., 2012, 2013 [21, 22] have used iIS for computing information criterion; they have provided a special, but very important case of predictive evaluation, in Gaussian process latent variable models in their matlab toolbox

GPstuff. This is documented by the manual for *GPstuff*, but their technical report (Vanhatalo et al., 2012) [21] did not discuss the details of iIS. This thesis gives iIS a detailed discussion. In addition, we provide a formula for iIS that is applicable to general evaluation function; in particular, our formula can also be used for computing CV posterior p-value. Addition, we have also proven the equivalence of iIS and CV. The main contribution of this thesis is in illustrating the necessity of incorporating iIS and iWAIC in approximating CV. In computing CV posterior p-value, iIS is also related to Ghosting method, which was proposed by Marshall and Spiegelhalter (2007) [14] and discussed by Held et al. (2010) [9]. Ghosting method does not use importance re-weighting to correct the bias in model parameters; hence, Ghosting method can be deemed as a partial implementation of iIS.

This thesis will be organized as follows: we discuss in Chapter 2, a class of Bayesian models with unit-specific models to which iIS and iWAIC can be applied, how to perform actual cross-validation evaluation, and give relevant posterior distribution. We will then describe iIS and iWAIC in general terms. This Chapter is almost taken verbatim from Section 2 to 5 of the paper by Li et al. (2014) [11]. I am a co-author of this paper, and I have contributed to these sections partially. In Chapter 3, we compare iIS and iWAIC to other information criteria approximation methods for model comparison with two disease mapping data sets. One set of data concerns the prevalence of suicide in London, while the other deals with lip cancer in Scotland. The results of Scottish lip cancer data has been reported in Li et al. (2014) [11] to which I contributed partially; the results of London suicide data (Section 3.1) are originally reported here. In Chapter 4, we compare iIS with other methods in the problem of detecting divergent regions for disease mapping data: the first approach uses p-value based on predictive distribution of y_i, while the second predictive distribution of relative risk λ_i. The results of Chapter 4 are original in this thesis. Our empirical results show that iIS and iWAIC provide significantly closer approximating to actual CV evaluation results than ordinary IS and WAIC, as well as other methods. Chapter 5 will conclude this thesis by summarizing our findings and discussing advancements for the future. In Appendices, we give a sketch of the working procedures of iIS and iWAIC, long tables and R code.

Chapter 2

Integrated IS and WAIC [1]

2.1 Bayesian Models with Unit-specific Latent Variables

The new predictive evaluation methods that we will describe are for use in Bayesian models with unit-specific latent variables. Throughout this thesis, we use bold-faced letters to denote vectors and matrices, and supposing we have n observations $\boldsymbol{y}_1^{\text{obs}}, \cdots, \boldsymbol{y}_n^{\text{obs}}$ on n observation units (e.g. cases, such as persons, locations, time points, or a combination of them), we model them as a realization of random variables $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$. In many problems, we introduce a latent variable (often random vector, sometimes called random effects, missing data), $\boldsymbol{b}_i$ for each unit i from which $\boldsymbol{y}_i^{\text{obs}}$ is observed; then we model $\boldsymbol{y}_i$ and $\boldsymbol{b}_i$ with certain statistical distributions parametrized by $\boldsymbol{\theta}$. Conditional on $\boldsymbol{b}_i$ and $\boldsymbol{\theta}$ (often also on a covariate variable $\boldsymbol{x}_i$ that will be omitted in the following equations for simplicity), we assume that $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ are statistically independent, with probability density $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$, which we will call *non-integrated predictive density* in this thesis. If we assume independence between $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n$ given $\boldsymbol{\theta}$, then the marginalized distributions of random variables $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ are also independent for each i (e.g. in mixture models). In modelling spatial and time series data, we often assume that the latent variables $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n$ are dependent in modelling correlations between locations or time points. In the following general discussion, we will assume that $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ are correlated. Figure 2.1 gives a graphical representation of the models described here.

Throughout this thesis, we will use notation $\boldsymbol{a}_{1:n}$ to denote the collection of all $\boldsymbol{a}_j$: $\{\boldsymbol{a}_j | j =$

[1] This chapter is part of the co-authored paper "Li, L., Qiu, S., Zhang, B., and Feng, C.X. (2014). Approximating Cross-validatory Predictive Evaluation in Bayesian Latent Variables Models with Integrated IS and WAIC. Available from http://arxiv.org/abs/1404.2918 "

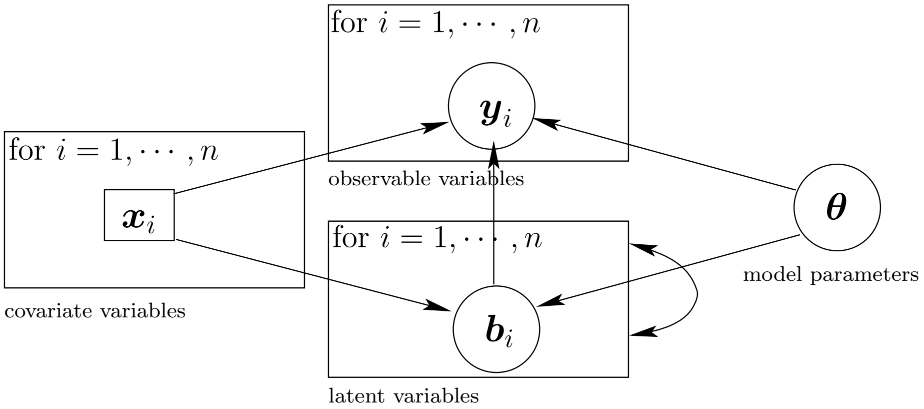


Figure 2.1: Graphical representation of Bayesian latent variables models. The double arrows in the box for $\boldsymbol{b}_{1:n}$ mean possible dependency between $\boldsymbol{b}_{1:n}$. Note that the covariate $\boldsymbol{x}_i$ will be omitted in the conditions of densities for $\boldsymbol{b}_i$ and $\boldsymbol{y}_i$ throughout this thesis for simplicity.

$1, \ldots, n\}$, and use $\boldsymbol{a}_{-i}$ to denote the collection of all $\boldsymbol{a}_j$ except $\boldsymbol{a}_i$: $\{\boldsymbol{a}_j | j = 1, \ldots, n, j \neq i\}$. Conditional on $\boldsymbol{\theta}$, we have specified a density for $\boldsymbol{y}_i$ given $\boldsymbol{b}_i$: $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$, a joint prior density for latent variables $\boldsymbol{b}_{1:n}$: $P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta})$, and a prior density for $\boldsymbol{\theta}$: $P(\boldsymbol{\theta})$. The posterior of $(\boldsymbol{b}_{1:n}, \boldsymbol{\theta})$ given observations $\boldsymbol{y}^{\text{obs}}_{1:n}$ is proportional to the joint density of $\boldsymbol{y}^{\text{obs}}_{1:n}$, $\boldsymbol{b}_{1:n}$, and $\boldsymbol{\theta}$:

$$P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}^{\text{obs}}_{1:n}) = \prod_{j=1}^{n} P(\boldsymbol{y}^{\text{obs}}_j|\boldsymbol{b}_j, \boldsymbol{\theta}) P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) / C_1, \tag{2.1}$$

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

2.2 Cross-validatory Predictive Evaluation

To do cross-validation, for each $i = 1, \ldots, n$, we omit observation $\boldsymbol{y}^{\text{obs}}_i$, and then draw MCMC samples from *CV posterior distribution* of model parameter and latent variables $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}^{\text{obs}}_{-i})$:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}^{\text{obs}}_{-i}) = \prod_{j \neq i} P(\boldsymbol{y}^{\text{obs}}_j|\boldsymbol{b}_j, \boldsymbol{\theta}) P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \tag{2.2}$$

where C_2 is the normalizing constant involving only with $\boldsymbol{y}^{\text{obs}}_{-i}$. Note that in equation (2.2), we assume that the possible structure's information (e.g. spatial relationships between n

locations) among $\boldsymbol{b}_{1:n}$ are not lost, as only the value of $\boldsymbol{y}_i^{\text{obs}}$ is omitted. After we draw MCMC samples of $(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})$ from (2.2), and then drop $\boldsymbol{b}_i$, we obtain MCMC sample of $(\boldsymbol{\theta}, \boldsymbol{b}_{-i})$ from the marginalized CV posterior $P(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}})$:

$$P_{\text{post(-i), M}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}}) = \prod_{j\neq i} P(\boldsymbol{y}_j^{\text{obs}}|\boldsymbol{b}_j, \boldsymbol{\theta}) P(\boldsymbol{b}_{-i}|\boldsymbol{\theta}) P(\boldsymbol{\theta}) \,/\, C_2, \tag{2.3}$$

where $P(\boldsymbol{b}_{-i}|\boldsymbol{\theta})$ is the marginalized prior density for $\boldsymbol{b}_{-i}$ generated from the specified joint prior for $\boldsymbol{b}_{1:n}$, i.e., $P(\boldsymbol{b}_{-i}|\boldsymbol{\theta}) = \int P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta}) d\boldsymbol{b}_i$. Using conditional prior

$$P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta}) = P(\boldsymbol{b}_{1:n}|\boldsymbol{\theta})/P(\boldsymbol{b}_{-i}|\boldsymbol{\theta}), \tag{2.4}$$

we can say:

$$P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}}) = P_{\text{post(-i), M}}(\boldsymbol{\theta}, \boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}}) P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta}). \tag{2.5}$$

From the above expression, we see that sampling from $P_{\text{post(-i)}}$ is equivalent to sampling from $P_{\text{post(-i), M}}$, and the conditional prior $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta})$. Therefore, this method of performing cross-validation makes use of the assumed structure in $\boldsymbol{b}_{1:n}$ (such as neighbouring relationships between spatial units) through $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\theta})$, in predicting $\boldsymbol{y}_i$ given $\boldsymbol{y}_{-i}^{\text{obs}}$. This treatment indeed regards the structure information in $\boldsymbol{b}_{1:n}$ as fixed covariate and being known. We feel that this treatment is reasonable because we are interested in comparing competing models for the conditional distribution of $\boldsymbol{y}_{1:n}$ given the structure between the n units, rather than the distribution of the structure itself. This is similar to how cross-validation is done in linear models, for which we assume that the values of the covariates (explanatory variables) of the test case are known when making a prediction of the test case response.

The purpose of performing CV is to evaluate certain compatibility (or discrepancy) between the posterior $P(\boldsymbol{y}_i|\boldsymbol{y}_{-i}^{\text{obs}})$ and the actual observation $\boldsymbol{y}_i^{\text{obs}}$. We will specify an evaluation function $a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)$ that measures certain goodness-of-fit (or discrepancy) of the distribution $P(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{b}_i)$ to the actual observation $\boldsymbol{y}_i^{\text{obs}}$. *CV posterior predictive evaluation* is defined as the expectation of the $a(\boldsymbol{y}_{1:n}^{\text{obs}}, ., .)$ with respect to $P_{\text{post(-i)}}$:

$$E_{\text{post(-i)}}(a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)) = \int a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}}) d\boldsymbol{\theta} d\boldsymbol{b}_{1:n}. \tag{2.6}$$

The expectation in (2.6) can be approximated by averaging $a(\boldsymbol{y}_i^{\text{obs}}, \cdot, \cdot)$ over MCMC samples of $(\boldsymbol{\theta}, \boldsymbol{b}_i)$ drawn from $P_{\text{post(-i)}}$.

One example of a is the value of predictive density function $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$ at the actual observation $\boldsymbol{y}_i^{\text{obs}}$:

$$a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) = P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i). \tag{2.7}$$

The expectation of (2.7) with respect to $P_{\text{post(-i)}}$ is *CV posterior predictive density* $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$. *CV information criterion* (CVIC) is defined as the sum of minus twice the CV posterior predictive densities over all validation units:

$$\text{CVIC} = -2\sum_{i=1}^{n} \log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})). \tag{2.8}$$

A smaller value of CVIC indicates a better fit of a Bayesian model to a real data set. A second example is to set a in (2.6) as the p-value given model parameter and latent variable for unit i: p-value $Pr(\boldsymbol{y}_i \geq \boldsymbol{y}_i^{\text{obs}})$ (Marshall and Spiegelhalter, 2003, 2007) [13, 14]; for discrete $\boldsymbol{y}$, we use:

$$a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) = Pr(\boldsymbol{y}_i > \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i) + 0.5Pr(\boldsymbol{y}_i = \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i), \tag{2.9}$$

where Pr means probability of a set, as we have used P as density; also $\boldsymbol{y}_i$ should be a scalar for such situations. The expectation of (2.9) with respect to $P_{\text{post(-i)}}$ gives *CV posterior p-value*:

$$\text{CV posterior p-value } (\boldsymbol{y}_i^{\text{obs}}) = Pr(\boldsymbol{y}_i > \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) + 0.5Pr(\boldsymbol{y}_i = \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}), \tag{2.10}$$

which is a tail probability of CV posterior predictive distribution with density $P(\boldsymbol{y}_i|\boldsymbol{y}_{-i}^{\text{obs}})$. The purpose of computing CV posterior p-value is to check the discrepancy of the observation $\boldsymbol{y}_i^{\text{obs}}$ to the CV posterior predictive distribution of $\boldsymbol{y}_i$ that is conditional on other observations $\boldsymbol{y}_{-i}^{\text{obs}}$. Both very large and very small values of posterior p-value indicate that $\boldsymbol{y}_i^{\text{obs}}$ may be an outlier (unusually small or large) compared to other observations.

Actual CV requires n of Markov chain simulations (each may use multiple parallel chains), one for each validation unit. This is very time consuming, especially when the model is complex and n is fairly large. Therefore, we are interested in approximating the expectations in (2.6) for all validation units where $i = 1, \dots, n$ with samples of $(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})$ obtained with a single MCMC simulation based on the full data set; that is, with samples drawn from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$, called *full data posterior* for short hereafter. However, we cannot simply treat samples from the full data posterior as CV posteriors, because the inclusion of $\boldsymbol{y}_i^{\text{obs}}$ has introduced optimistic bias in validating $\boldsymbol{y}_i^{\text{obs}}$. The optimistic bias means that the posterior predictive distribution of $\boldsymbol{y}_i$ formed by averaging $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$ with respect to $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ fits $\boldsymbol{y}_i^{\text{obs}}$ better than the actual CV posterior predictive distribution of $\boldsymbol{y}_i$ that averages $P(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$ with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$. Therefore, we need to correct for the optimistic bias with a certain method to obtain an unbiased approximate/estimate of actual CV posterior predictive evaluation. We will introduce two new approximating methods in Section 2.3 and 2.4, respectively.

2.3 Importance Sampling (IS) Approximation

2.3.1 Non-integrated Importance Sampling

Importance weighting (Gelfand et al., 1992) [3] is a natural choice for approximating CV prediction evaluation based on the posterior, given the full data set. For general and detailed discussion of importance sampling techniques, one can refer to Geweke, Neal, Gelman and Meng, Liu (1989, 1993, 1998, 2001) [8, 17, 4, 12]. If our samples are from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$, but we are interested in estimating the mean of a with respect to $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ as in (2.6), importance weighting method is based on the following equality for CV expected evaluation:

$$E_{\text{post(-i)}}(a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)) = \frac{E_{\text{post}}\left[a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i) W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})\right]}{E_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})\right]}, \tag{2.11}$$

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

$$W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}) = \frac{P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})}{P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})} \times \frac{C_2}{C_1} = \frac{1}{P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)}. \tag{2.12}$$

Note that we can multiply any constant to the above importance weight since it will be cancelled at the fraction of (2.11); also we use superscript $^{\text{nIS}}$ to denote application of importance sampling (shortened to *nIS*) to the *non-integrated predictive density*, in contrast to iIS to be given in next section. In words, importance sampling estimates the expected evaluation by finding Monte Carlo estimates of the two means in the fraction of (2.11) with only MCMC samples from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$. We can apply equation (2.11) to estimate means of any evaluation function a with respect to the CV posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{b}_i)$.

Particularly, in computing CVIC, the evaluation function $a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)$ is equal to $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)$ which is the same as $1/W_i^{\text{nIS}}$ in equation (2.12). Therefore, the numerator of (2.11) is just 1 when applied to compute CVIC. Hence, the CV posterior predictive density $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$ is equal to the harmonic mean of the non-integrated predictive density $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)$ with respect to $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$:

$$P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \frac{1}{E_{\text{post}}\left[1/P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)\right]}. \tag{2.13}$$

Based on the equality (2.13), **nIS** estimates the CV posterior predictive density by:

$$\hat{P}^{\text{nIS}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post}}\left[1/P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)\right]}. \tag{2.14}$$

The corresponding nIS estimate of CVIC using (2.14) is $-2\sum_{i=1}^{n}\log(\hat{P}^{\text{nIS}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$. Note that, if there are not latent variables used for a model, there will be no $\boldsymbol{b}_i$ in (2.13) and (2.14).

2.3.2 Integrated Importance Sampling

In theory, the nIS estimate (2.11) is valid for almost all Bayesian models with latent variables as long as the integral itself exists and the supports of $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ and $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ are the same. However, in simulating MCMC from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$, the latent variable $\boldsymbol{b}_i$ is largely confined to regions that fit the observation $\boldsymbol{y}_i^{\text{obs}}$ well. Therefore, the distribution of $\boldsymbol{b}_i$ marginalized from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ may be highly biased to regions that fit the observation $\boldsymbol{y}_i^{\text{obs}}$ well, compared to the distribution of $\boldsymbol{b}_i$ marginalized from $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$, which can cover a much larger area. Therefore, although the supports of $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ and $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ are the same in theory, the effective support of $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$ may be much smaller than that of $P_{\text{post(-i)}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$. This results in the inaccuracy of nIS.

To improve nIS, we can re-generate $\boldsymbol{b}_i$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i},\boldsymbol{\theta})$, with the observation $\boldsymbol{y}_i^{\text{obs}}$ removed, as the actual cross-validation simulation does; see equation (2.5). The formal formulation of such re-generation procedure is given as follows. First we note that using equation (2.5), we can rewrite the expectation in (2.6) as:

$$E_{\text{post(-i)}}(a(\boldsymbol{y}_i^{\text{obs}},\boldsymbol{\theta},\boldsymbol{b}_i)) = E_{\text{post(-i), M}}(A(\boldsymbol{y}_i^{\text{obs}},\boldsymbol{\theta},\boldsymbol{b}_{-i})) \tag{2.15}$$

$$= \int\int A(\boldsymbol{y}_i^{\text{obs}},\boldsymbol{\theta},\boldsymbol{b}_{-i})P(\boldsymbol{\theta},\boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}})d\boldsymbol{\theta}d\boldsymbol{b}_{-i} \tag{2.16}$$

where,

$$A(\boldsymbol{y}_i^{\text{obs}},\boldsymbol{\theta},\boldsymbol{b}_{-i}) = \int a(\boldsymbol{y}_i^{\text{obs}},\boldsymbol{\theta},\boldsymbol{b}_i)P(\boldsymbol{b}_i|\boldsymbol{b}_{-i},\boldsymbol{\theta})d\boldsymbol{b}_i. \tag{2.17}$$

We will call (2.17) an *integrated evaluation function.*

We will also discard $\boldsymbol{b}_i$ temporarily for validation unit i in MCMC samples from the full data posterior $P_{\text{post}}(\boldsymbol{\theta},\boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$. The marginalized full data posterior of $(\boldsymbol{\theta},\boldsymbol{b}_{-i})$ is

$$P_{\text{post, M}}(\boldsymbol{\theta},\boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}}) = \prod_{j\neq i} P(\boldsymbol{y}_j^{\text{obs}}|\boldsymbol{b}_j,\boldsymbol{\theta})P(\boldsymbol{b}_{-i}|\boldsymbol{\theta})P(\boldsymbol{\theta})\times\int P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{b}_i,\boldsymbol{\theta})P(\boldsymbol{b}_i|\boldsymbol{b}_{-i},\boldsymbol{\theta})d\boldsymbol{b}_i/C_1. \tag{2.18}$$

We will call the second factor in (2.18) *integrated predictive density*, because it integrates away $\boldsymbol{b}_i$ without reference to $\boldsymbol{y}_i^{\text{obs}}$. For ease in reference, it is explicitly given below:

$$P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta},\boldsymbol{b}_{-i}) = \int P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{b}_i,\boldsymbol{\theta})P(\boldsymbol{b}_i|\boldsymbol{b}_{-i},\boldsymbol{\theta})d\boldsymbol{b}_i. \tag{2.19}$$

Using the standard importance weighting method, we will estimate (2.16) by:

$$E_{\text{post(-i), M}}(A(\boldsymbol{y}_i^{\text{obs}},\boldsymbol{\theta},\boldsymbol{b}_{-i})) = \frac{E_{\text{post, M}}\left[A(\boldsymbol{y}_i^{\text{obs}},\boldsymbol{\theta},\boldsymbol{b}_{-i})\ W_i^{\text{iIS}}(\boldsymbol{\theta},\boldsymbol{b}_{-i})\right]}{E_{\text{post, M}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta},\boldsymbol{b}_{-i})\right]}, \tag{2.20}$$

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

$$W_i^{\text{iIS}}(\boldsymbol{\theta},\boldsymbol{b}_{-i}) = \frac{P_{\text{post(-i), M}}(\boldsymbol{\theta},\boldsymbol{b}_{-i}|\boldsymbol{y}_{-i}^{\text{obs}})}{P_{\text{post, M}}(\boldsymbol{\theta},\boldsymbol{b}_{-i}|\boldsymbol{y}_{1:n}^{\text{obs}})}\times\frac{C_2}{C_1} = \frac{1}{P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta},\boldsymbol{b}_{-i})}, \tag{2.21}$$

for estimating CVIC, $A\times W_i^{\text{iIS}} = 1$ in particular. Therefore, the iIS estimate of the CV posterior predictive density based on equality (2.20) is given by:

$$\hat{P}^{\text{iIS}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \frac{1}{\hat{E}_{\text{post, M}}\big[1/P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})\big]}. \tag{2.22}$$

Accordingly, iIS estimate of CVIC using (2.22) is $-2\sum_{i=1}^n \log(\hat{P}^{\text{iIS}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}))$. The only difference from nIS of estimate (2.14) is in the replacement of non-integrated predictive density $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)$ by integrated predictive density $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})$. Note that we can also write the expectation $E_{\text{post, M}}(\;)$ in equations (2.20) and (2.22) as $E_{\text{post}}(\;)$ because we still find Monte Carlo estimates with samples of $(\boldsymbol{\theta}, \boldsymbol{b}_{1:n})$ from $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$, but without using $\boldsymbol{b}_i$.

The integration over $\boldsymbol{b}_i$ in equations (2.17) and (2.19) is the essential difference between iIS and nIS. In order to use iIS, we need to find $\boldsymbol{b}_i$. In some problems, $\boldsymbol{b}_i$ can be approximated with finite summation, or calculated analytically. Otherwise, we will re-generate $\boldsymbol{b}_i$ given $(\boldsymbol{b}_{-i}, \boldsymbol{\theta})$ with no reference to $\boldsymbol{y}_i^{\text{obs}}$, which is often easy. Note that this re-generation needs to be done for each $i = 1, \ldots, n$. Sometimes much computation can be shared by these n re-generating processes since they are all conditional on $\boldsymbol{\theta}$; see the example in Chapter 3.

2.4 WAIC Approximations

In this section, we describe a generalized WAIC method, iWAIC, for approximating CV predictive density in Bayesian models with correlated latent variables.

We will first describe WAIC for models with no latent variables (or models after we integrate away latent variables that are independent for units given parameters). In such models, observed variables $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independently distributed with a probability distribution $P(\boldsymbol{y}|\boldsymbol{\theta})$ conditional on model parameters $\boldsymbol{\theta}$. After we obtain MCMC samples for $\boldsymbol{\theta}$ given observations $\boldsymbol{y}_1^{\text{obs}}, \ldots, \boldsymbol{y}_n^{\text{obs}}$, a version of WAIC (Watanabe, 2009, 2010, 2010) [25, 27, 28] is given by:

$$\text{WAIC} = -2\sum_{i=1}^n \big[\log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta})))\big], \tag{2.23}$$

where E_{post} and V_{post} stand for mean and variance over $\boldsymbol{\theta}$ with respect to $P(\boldsymbol{\theta}|\boldsymbol{y}_1^{\text{obs}}, \ldots, \boldsymbol{y}_n^{\text{obs}})$.

By comparing the forms of WAIC and CVIC (2.8), we can see the CV posterior predictive density approximated by:

$$\hat{P}^{\text{WAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \exp\Big\{\log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta})))\Big\}. \tag{2.24}$$

In words, WAIC corrects the bias in mean of training predictive density of $\boldsymbol{y}_i^{\text{obs}}$ by dividing exponential of variance of log predictive density of $\boldsymbol{y}_i^{\text{obs}}$ with respect to the posterior of $\boldsymbol{\theta}$, given the full data set. Watanabe (2010) [26] has proven that WAIC is asymptotically equivalent to CVIC when observed variables are independently distributed conditional on $\boldsymbol{\theta}$. He has shown the asymptotic equivalence of Taylor expansions of (2.24) and harmonic mean (2.14) (without $\boldsymbol{b}_i$). From our research, we do see that (2.24) provides results very close to CV posterior predictive density of each $\boldsymbol{y}_i^{\text{obs}}$. This perspective of WAIC also provides the approach to assess statistical significance of differences of WAICs of different models by looking at differences in means of log CV posterior predictive densities, which was advocated by Vehtari and Lampinen (2002) [23] for CVIC itself.

For the models given in Section 2.1 with possibly correlated latent variables, a naive way to approximate CVIC is to apply WAIC directly to the non-integrated predictive density of $\boldsymbol{y}_i^{\text{obs}}$ conditional on $\boldsymbol{\theta}$ and $\boldsymbol{b}_i$:

$$\hat{P}^{\text{nWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \exp\Big\{\log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_i)))\Big\}. \tag{2.25}$$

We will refer to (2.25) as non-integrated WAIC (or nWAIC for short) method for approximating CV posterior predictive density. The corresponding information criterion based on (2.25) is:

$$\text{nWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{nWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})). \tag{2.26}$$

This way to apply WAIC indeed treats latent variables as model parameters. nWAIC is not justified by the theory for WAIC. However, practitioners may likely apply WAIC to Bayesian models with latent variables this way for the sake of convenience.

Our research (to be presented next) will show that nWAIC cannot correct the bias in unit-specific latent variables entirely. We propose applying WAIC approximation to the integrated predictive density (2.19) in order to estimate the CV posterior predictive density:

$$\hat{P}^{\text{iWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = \exp\left\{\log(E_{\text{post}}(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i}))) - V_{\text{post}}(\log(P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})))\right\}. \tag{2.27}$$

Accordingly, iWAIC for approximating CVIC is given by :

$$\text{iWAIC} = -2\sum_{i=1}^{n} \log(\hat{P}^{\text{iWAIC}}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})). \tag{2.28}$$

In Section 2.3, we have theoretically shown the equivalence of iIS to CV predictive evaluation for models with correlated latent variables, which holds as long as the support of full data posterior is not a subset of the CV posterior. However, we have not proven any sort of equivalences of $\hat{P}^{\text{iWAIC}}$ and $\hat{P}^{\text{nWAIC}}$ to CVIC. The derivations of formulae for nWAIC and iWAIC for models with correlated latent variables are only heuristic, borrowing the asymptotic equivalence of WAIC estimate (2.24) and CVIC expressed with harmonic mean (IS) (2.13) (without $\boldsymbol{b}_i$) for models without latent variables, which is proven by Watanabe (2010) [26].

Chapter 3

Emprical Results on Model Comparison

3.1 London Suicide Data

In this section, we will investigate the performance of iIS and iWAIC in an analysis of London boroughs suicide data. The London suicide data is based on registered mortality under International Classification of Diseases (ICD) classes 950-959 and 980-989. This data set was downloaded from the link `http://webspace.qmul.ac.uk/pcongdon/BSM2.zip`, then opened with *OpenBUGS*. This data set is used in example 9.8 of the textbook by Congdon (2007) [2] for demonstrating spatial effects models. The data set is shown in Table 3.1. This example discusses the suicide mortality of 32 London boroughs from 1989-1993. The map of 32 London boroughs is shown in Figure 3.1. For $i = 1, \dots, n$ and $n = 32$, we denote random variables y_i as observed suicide counts. Expected suicide counts in the ith borough are E_i (derived using demographic methods). We also denote that y_i^{obs} is actual observed counts. The centroid of the ith borough was recorded as vector $\boldsymbol{x}_i = (x_{i1}, x_{i2})$, each with a unit of 10km as a reference point. The Euclidean distance d_{ij} between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is used to measure the distance between the ith and the jth borough. We also note the standardized morbidity ratio($SMR_i \equiv y_i/E_i$).

Table 3.1: London boroughs suicide mortality(male and female suicides combined over 1989-1993)

ID	Boroughs	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{Y}$	$\boldsymbol{E}$	$\boldsymbol{SMR}$
1	Barking and Dagenham	547.80	185.10	75	80.70	0.93
2	Barnet	524.30	191.70	145	169.80	0.85
3	Bexley	548.40	175.70	99	123.20	0.80

Table 3.1 – *Continued from previous page*

ID	Boroughs	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{Y}$	$\boldsymbol{E}$	***SMR***
4	Brent	520.70	185.50	168	139.50	1.20
5	Bromley	541.80	167.60	152	169.10	0.90
6	Camden	527.90	184.30	173	107.20	1.61
7	Croydon	533.30	165.10	152	179.80	0.85
8	Ealing	515.90	181.40	169	160.40	1.05
9	Enfield	533.10	195.30	130	147.50	0.88
10	Greenwich	542.80	176.80	117	116.80	1.00
11	Hackney	534.20	185.50	124	102.80	1.21
12	Hammersmith and Fulham	523.80	178.50	119	91.80	1.30
13	Haringey	531.50	189.60	134	119.60	1.12
14	Harrow	515.00	189.50	90	114.80	0.78
15	Havering	553.10	188.20	98	131.10	0.75
16	Hillingdon	508.60	183.80	89	136.10	0.65
17	Hounslow	514.00	175.80	128	116.60	1.10
18	Islington	531.10	185.10	145	98.50	1.47
19	Kensington and Chelsea	525.60	179.50	130	88.80	1.46
20	Kingston upon Thames	519.40	167.50	69	79.80	0.86
21	Lambeth	530.80	174.60	246	144.90	1.70
22	Lewisham	537.50	174.00	166	134.70	1.23
23	Merton	525.80	169.30	95	98.90	0.96
24	Newham	541.20	183.60	135	118.60	1.14
25	Redbridge	543.80	188.90	98	130.60	0.75
26	Richmond upon Thames	517.00	173.40	97	96.10	1.01
27	Southwark	526.60	164.50	202	127.10	1.59
28	Sutton	533.60	177.10	75	97.70	0.77
29	TowerHamlets	536.10	181.80	100	88.50	1.13
30	Waltham Forest	526.40	173.90	100	121.40	0.82
31	Wandsworth	527.20	181.10	153	156.80	0.98

Table 3.1 – *Continued from previous page*

ID	Boroughs	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{Y}$	$\boldsymbol{E}$	***SMR***
32	Westminster	537.90	189.60	194	114.00	1.70

We model $y_i|E_i, \mu_i \sim \text{Poisson}(\mu_i E_i)$, where μ_i denotes the underlying relative risk of London suicide for borough i. We consider four different models for the log relative risk $\log(\mu_i)$ as follows:

$$\text{model 1} \quad (\text{spatial+exchangeable}) : \log(\mu_i) = \alpha + s_i + u_i \tag{3.1}$$

$$\text{model 2} \quad (\text{spatial}) : \log(\mu_i) = \alpha + s_i \tag{3.2}$$

$$\text{model 3} \quad (\text{exchangeable}) : \log(\mu_i) = \alpha + u_i \tag{3.3}$$

$$\text{model 4} \quad (\text{pooled}) : \log(\mu_i) = \alpha \tag{3.4}$$

where α is an intercept for modelling pooled effect for all boroughs, and u_i and s_i are exchangeable (independent) and spatially correlated random effects on ith borough, respectively. According to (Congdon, 2007) [2], we assign priors to α, u_i and s_i with the following hierarchy:

$$\alpha \sim N(0, 1000) \tag{3.5}$$

$$u_i \sim N(0, \tau^2) \tag{3.6}$$

$$s_1, \ldots, s_n|\Sigma \sim N_n(\mathbf{0}, \Sigma) \tag{3.7}$$

$$\Sigma = \sigma^2 \boldsymbol{R}, \text{ with } r_{ij} = \exp[-(\phi d_{ij})^\delta] \tag{3.8}$$

$$1/\sigma^2, 1/\tau^2 \sim \text{Gamma}(1, 0.001) \tag{3.9}$$

$$\phi \sim \text{Uniform}([0.1, 5]) \tag{3.10}$$

$$\delta \sim \text{Uniform}([0, 2]) \tag{3.11}$$

All the above four models belong to the class of Bayesian latent variable models depicted by Figure 2.1. The observable variable is y_i, and the latent variable $\boldsymbol{b}_i$ is u_i and s_i (for

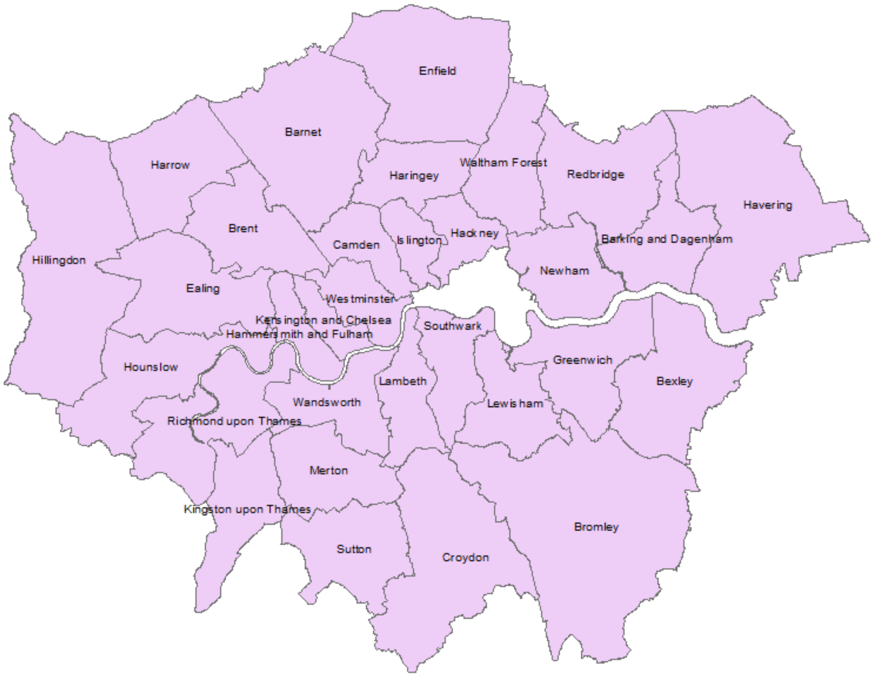


Figure 3.1: Map of London boroughs. This data of boundary files for this map was generated from the Office for National Statistics (ONS) 2011 Census. For information on the licensing of this data see http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/licences/index.html

model 1), s_i (for model 2), u_i (for model 3), or none (for model 4). Model parameter $\boldsymbol{\theta}$ is $(\alpha, \tau, \sigma, \phi, \delta)$ for model 1 or a subset of it for other models.

We used Winbugs to fit the above 4 models to London suicide data. For each model, we ran MCMC simulations with two parallel chains, each with 10000 iterations, from which we discarded the first 5000 iterations as burning. We then used MCMC simulations for each model to calculate the posterior inference of parameters. The results are summarized in Table 3.2.

For each model, we ran 32 actual cross-validatory MCMC simulations with each of the 32 observations removed (set y_i to NA in WinBugs). We then computed actual CV posterior predictive density $P(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$ using equation (2.6) with evaluation function set to dpoisson$(y_i^{\text{obs}}|\mu_i E_i)$ — Poisson probability mass function having the parameter $\mu_i E_i$. We computed CVIC for different models displayed in Table 3.3.

We then considered approximating CVIC with four different methods (nIS, nWAIC, iIS, and iWAIC) from a single MCMC simulation based on all of the 32 samples for each model. The dpoisson$(y_i^{\text{obs}}|\mu_i E_i)$ is non-integrated predictive density used in computing nIS

Table 3.2: Posterior inference parameters for the four models of London suicide data

Model	Parameters	Posterior distribution			
		Mean	2.5%	Median	97.5%
model 1 (spatial+exchangeable)	α	0.01	-0.25	0.03	0.16
	δ	0.94	0.05	0.86	1.96
	ϕ	2.74	0.29	2.84	4.90
	σ^2	0.025	0.00	0.00	0.1
	τ^2	0.04	0.00	0.04	0.09
model 2 (spatial)	α	-0.02	-0.30	-0.00	0.18
	δ	0.82	0.05	0.69	1.94
	ϕ	3.31	0.49	3.57	4.93
	σ^2	0.07	0.04	0.07	0.13
model 3 (exchangeable)	α	0.04	-0.05	0.04	0.14
	τ^2	0.06	0.03	0.06	0.11
model 4 (pooled)	α	0.07	0.03	0.07	0.10

and nWAIC with equations (2.14) and (2.25), where μ_i is computed with latent variables and model parameters used in respective models. We will now describe how to compute iIS and iWAIC for model 1 (3.2). The integrated predictive density (2.19) used in equations (2.22) and (2.27) is:

$$P(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i}, \boldsymbol{u}_{-i}) = \int\int \text{dpoisson}(y_i^{\text{obs}}|\mu_i E_i) P(s_i, u_i|\boldsymbol{\theta}, \boldsymbol{s}_{-i}, \boldsymbol{u}_{-i}) ds_i du_i. \tag{3.12}$$

where the second factor in the integral can be written as $P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})P(u_i|\boldsymbol{u}_{-i}, \boldsymbol{\theta})$ because the spatial effects s_i and the random effects u_i are assumed independent. We do not have closed-form solution for this integral. We therefore need to use Monte Carlo method to estimate (3.12) by generating random numbers from $P(u_i|\boldsymbol{u}_{-i}, \boldsymbol{\theta})$ and $P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})$ for each MCMC sample of $(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}, \boldsymbol{u}_{1:n})$ and each validation unit i. The $P(u_i|\boldsymbol{u}_{-i}, \boldsymbol{\theta})$ is $N(0, \tau^2)$ because $\boldsymbol{u}_{1:n}$ are independent given $\boldsymbol{\theta}$. The spatial effects $\boldsymbol{s}_{1:n}$ given Σ are distributed with $N_n(\mathbf{0}, \Sigma)$. Let $\boldsymbol{B} = \Sigma^{-1}$. By standard formula for conditional normal distribution, the $P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})$ is $N(-B^{-i}_{i,i}\boldsymbol{B}_{i,-i}\boldsymbol{s}_{-i}, B^{-1}_{ii})$, where $\boldsymbol{B}_{i,j}$ stands for the matrix containing only rows i and columns j of $\boldsymbol{B}$. By using this formula, we must invert the covariance matrix Σ only once, then $\boldsymbol{B}$ can be used for each i. In this computing, we generate 200 random numbers from each of the two conditional distributions for approximating the integral in (3.12). Finally, based on computed values of $P(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i}, \boldsymbol{u}_{-i})$ for all MCMC samples, we then compute iIS and iWAIC approximates of CV posterior predictive densities (with equations (2.22) and (2.27) respectively) or corresponding iIS-IC and iWAIC. For computing iIS and iWAIC in model 2 and model 3, we need only to integrate dpoisson($y_i^{\text{obs}}|\mu_i E_i$) with respect to $P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})$ and $P(u_i|\boldsymbol{u}_{-i}, \boldsymbol{\theta})$, respectively. For model 4, iIS and nIS are the same, as are iWAIC and nWAIC, since there is no latent variable.

We repeated computing nIS-IC, WAIC, iIS-IC, and iWAIC, as well as DIC for 100 independent MCMC simulations, each with 2 parallel chains. The means of these 100 information criteria for each method and each model are shown in Table 3.3, with standard deviations shown in brackets. From Table 3.3, we see that compared to nIS, nWAIC and DIC, iIS and iWAIC provide significantly closer approximates to the actual CVIC. For iWAIC, the approximates are almost identical to actual CVIC. This shows that the integration applied to latent variables associated with the validation unit indeed helps in correcting optimistic bias. The comparable results of iIS information criteria and CVIC may not be surprising since our derivation in Chapter 2 has shown their equivalence. However, it is surprising to see that the heuristic iWAIC also gives estimates very close to CVIC for model 1, as we have no theory to support this result. In addition, iWAIC has smaller standard deviations and smaller bias than iIS.

On the other hand, we see that although DIC, nIS and nWAIC are much downward biased, their model selection results (the ordering of these four models in terms of information criterion) are still correct. However, we believe that this may not generalize to more complex models. Finally, we notice that our model evaluation results (using actual CV) for the London suicide data indicate that there may not spatial effects. We can observe from Table 3.2 that the 95% credible intervals of parameter τ^2 and σ^2 in model 1 from posterior distribution are $(0.00, 0.1)$ and $(0.00, 0.09)$, respectively. We find that credible intervals of τ^2 are very close to zero, which means spatial effect can be ignored.

We also explored the performance of this experiment in Openbugs. We repeated the above processes by transferring identical data into Openbugs. We fitted the 4 models with Openbugs through R package `R2OpenBUGS`, then ran MCMC simulation with two parallel chains, 10000 iterations for each chain, burning the first 5000. The comparison results are shown in Table 3.4. When we compare Table 3.4 with Table 3.3, we can see that the results in iWAIC and

Table 3.3: Comparison results of cross-validation and other approximating methods using WinBugs in four models of London suicide mortality data

	CV	DIC	iWAIC	iIS	nWAIC	nIS
model 1	318.64(0.40)	273.99(0.16)	320.61(0.28)	326.86(4.12)	268.05(0.31)	297.99(4.99)
model 2	319.05(0.23)	274.21(0.16)	321.29(0.12)	327.41(3.79)	268.42(0.29)	298.38(4.87)
model 3	317.45(0.20)	273.59(0.15)	318.84(0.11)	319.09(0.19)	267.26(0.24)	295.48(4.42)
model 4	521.21(0.06)	511.77(0.03)	521.12(0.14)	521.13(0.16)	521.12(0.14)	521.13(0.16)

iIS are almost the same. However, the DIC in model 1 and 2 are extraordinarily different from DIC given by WinBUGS. This happened with negative value and large variance.

Table 3.4: Comparison results of cross-validation and other approximating methods using OpenBugs in four models of London suicide mortality data.

	CV	DIC	iWAIC	iIS	nWAIC	nIS
model 1	318.42(0.41)	167.25(93.39)	319.93(0.23)	323.64(2.48)	267.36(0.15)	295.83(2.97)
model 2	318.66(0.17)	135.37(276.90)	320.79(0.11)	326.36(2.83)	267.29(0.25)	293.42(3.57)
model 3	317.61(0.23)	273.60(0.13)	318.88(0.10)	319.17(0.19)	267.22(0.20)	294.99(4.84)
model 4	565.54(0.07)	511.79(0.03)	521.21(0.16)	521.20(0.16)	521.21(0.16)	521.20(0.16)

3.2 Scottish Lip Cancer Data

In this section, we apply iIS and iWAIC to another data set, Scottish lip cancer data, which was used in Stern and Cressie, Spiegelhalter et al., Plummer (2000, 2002, 2008) [20, 19, 18] and extracted from Stern and Cressie (2000) [20]. The data represents male lip cancer counts (over the period of 1975-1980) in the $n = 56$ districts of Scotland. At each district i, the data include these fields: (1) identity number of each district i; (2) name of each district; (3) number of observed cases of lip cancer, y_i; (4) number of expected cases, E_i, calculated based on standardization of "population at risk" across different age groups; (5) standardized morbidity ratio(SMR_i) for the ith districts, $SMR_i \equiv y_i/E_i$; (6) percent of population employed in agriculture, fishing and forestry, x_i, used as a covariate; and (7) group of IDs of neighbouring ith district. Figure 3.2 shows "ID" and "District name" on Scotland map, where we can verify the neighbouring districts graphically.

Table 3.5: Scotland lip cancer data

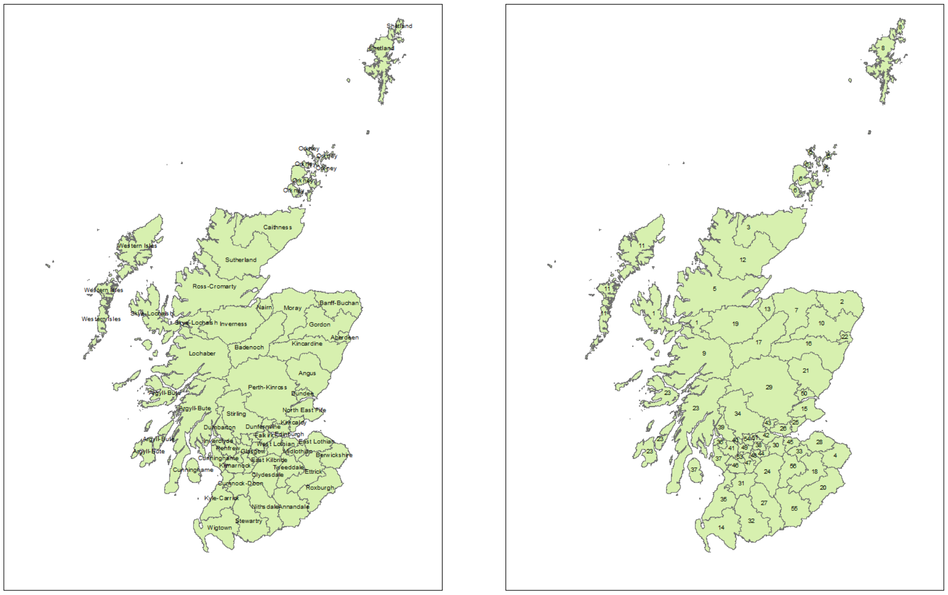
ID	District name	$\boldsymbol{Y}$	$\boldsymbol{E}$	$\boldsymbol{SMR}$	$\boldsymbol{X}$	Neighbours
1	Skye-Lochalsh	9	1.38	6.52	16	5,9,11,19
2	Banff-Buchan	39	8.66	4.50	16	7,10
3	Caithness	11	3.04	3.62	10	6,12
4	Berwickshire	9	2.53	3.56	24	18,20,28
5	Ross-Cromarty	15	4.26	3.52	10	1,11,12,13,19
6	Orkney	8	2.40	3.33	24	3,8
7	Moray	26	8.11	3.21	10	2,10,13,16,17
8	Shetland	7	2.30	3.04	7	6
9	Lochaber	6	1.98	3.03	7	1,11,17,19,23,29
10	Gorden	20	6.63	3.02	16	2,7,16,22
11	Western Isles	13	4.40	2.95	7	1,5,9,12
12	Sutherland	5	1.79	2.79	16	3,5,11
13	Nairn	3	1.08	2.78	10	5,7,17,19
14	Wigtown	8	3.31	2.42	24	31,32,35
15	NE Fife	17	7.84	2.17	7	25,29,50
16	Kincardine	9	4.55	1.98	16	7,10,17,21,22,29
17	Badenoch	2	1.07	1.87	10	7,9,13,16,19,29
18	Ettrick	7	4.18	1.67	7	4,20,28,33,55,56
19	Inverness	9	5.53	1.63	7	1,5,9,13,17
20	Roxburgh	7	4.44	1.58	10	4,18,55
21	Angus	16	10.46	1.53	7	16,29,50
22	Aberdeen	31	22.67	1.37	16	10,16
23	Argyll-Bute	11	8.77	1.25	10	9,29,34,36,37,39
24	Clydesdale	7	5.62	1.25	7	27,30,31,44,47,48,55,56
25	Kirkcaldy	19	15.47	1.23	1	15,26,29
26	Dunfermline	15	12.49	1.20	1	25,29,42,43
27	Nithsdale	7	6.04	1.16	7	24,31,32,55

Table 3.5 – *Continued from previous page*

ID	District name	$\boldsymbol{Y}$	$\boldsymbol{E}$	$\boldsymbol{SMR}$	$\boldsymbol{X}$	Neighbours
28	East-Lothian	10	8.96	1.12	7	4,18,33,45
29	Perth-Kinross	16	14.37	1.11	10	9,15,16,17,21,23,25,26,34,43,50
30	West Lothian	11	10.20	1.08	10	24,38,42,44,45,56
31	Cumnock-Doon	5	4.75	1.05	7	14,24,27,32,35,46,47
32	Stewartry	3	2.88	1.04	24	14,27,31,35
33	Midlothian	7	7.03	1.00	10	18,28,45,56
34	Stirling	8	8.53	0.94	7	23,29,39,40,42,43,51,52,54
35	Kyle-Carrick	11	12.32	0.89	7	14,31,32,37,46
36	Inverclyde	9	10.10	0.89	0	23,37,39,41
37	Cunninghame	11	12.68	0.87	10	23,35,36,41,46
38	Monklands	8	9.35	0.86	1	30,42,44,49,51,54
39	Dumbarton	6	7.20	0.83	16	23,34,36,40,41
40	Clydebank	4	5.27	0.76	0	34,39,41,49,52
41	Renfrew	10	18.76	0.53	1	36,37,39,40,46,49,53
42	Falkirk	8	15.78	0.51	16	26,30,34,38,43,51
43	Clackmannan	2	4.32	0.46	16	26,29,34,42
44	Motherwell	6	14.63	0.41	0	24,30,38,48,49
45	Edinburgh	19	50.72	0.37	1	28,30,33,56
46	Kilmarnock	3	8.20	0.37	7	31,35,37,41,47,53
47	East Kilbride	2	5.59	0.36	1	24,31,46,48,49,53
48	Hamilton	3	9.34	0.32	1	24,44,47,49
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
51	Cumbernauld	1	3.44	0.29	1	34,38,42,54
52	Bearsden	1	3.62	0.28	0	34,40,49,54
53	Eastwood	1	5.74	0.17	1	41,46,47,49
54	Strathkelvin	1	7.03	0.14	1	34,38,49,51,52
55	Annandale	0	4.16	0	16	18,20,24,27,56

Table 3.5 – *Continued from previous page*

ID	District name	$\boldsymbol{Y}$	$\boldsymbol{E}$	$\boldsymbol{SMR}$	$\boldsymbol{X}$	Neighbours
56	Tweeddale	0	1.76	0	10	18,24,30,33,45,55



(a) Map of Scotland with district names

(b) Map of Scotland with district ID

Figure 3.2: Maps of Scotland with district names and ID. Note: the GIS boundary files for ESRI and map was from from Local Government Boundary Commission for Scotland. For more information see the link http://www.lgbc-scotland.gov.uk/maps/datafiles/index_1995_on.asp

The y_i, for $i = 1, \ldots, n$, is modelled as an independent Poisson random variable conditional on λ_i and E_i:

$$y_i | E_i, \lambda_i \sim \text{Poisson}(\lambda_i E_i), \tag{3.13}$$

where λ_i denotes the underlying relative risk for district i, and E_i stands for expected counts. Let $s_i = \log(\lambda_i)$. We consider four different models for the vector $\boldsymbol{s} = (s_1, \cdots, s_n)$, conditional on $\boldsymbol{X} = (x_1, \ldots, x_n)'$ and neighbouring information between districts:

$$\text{spatial+linear (called } \textit{full} \text{ for short)} : \boldsymbol{s} \sim N_n(\alpha + \boldsymbol{X}\beta, \Phi\tau^2), \tag{3.14}$$

$$\text{spatial} : \boldsymbol{s} \sim N_n(\alpha, \Phi\tau^2), \tag{3.15}$$

$$\text{linear} : \boldsymbol{s} \sim N_n(\alpha + \boldsymbol{X}\beta, I_n\tau^2), \tag{3.16}$$

$$\text{exchangable} : \boldsymbol{s} \sim N_n(\alpha, I_n\tau^2), \tag{3.17}$$

In (3.14) and (3.15) above, $\Phi = (I_n - \phi C)^{-1}M$ is a matrix for capturing the spatial correlations amongst the n districts, in which the elements of C are: $c_{ij} = (E_j/E_i)^{1/2}$ if areas i and j are neighbours, and $c_{ij} = 0$ if otherwise; the elements of M are: $m_{ii} = E_i^{-1}$ and $m_{ij} = 0$ if $i \neq j$; ϕ is a parameter measuring spatial dependence; Φ can be expressed as $M^{1/2}(I - \phi M^{-1/2}CM^{1/2})^{-1}M^{1/2}$. For positive definite Φ, the range of ϕ, (ϕ_{min}, ϕ_{max}) is inverse of smallest and largest eigenvalues of $M^{-1/2}CM^{1/2}$(Stern and Cressie, 2000) [20]. The multivariate normal distributions with Φ as covariance matrix are called *proper conditional auto-regression (CAR) model*. In (3.14) and (3.16), β is a model parameter controlling linear effects of covariate $\boldsymbol{X}$ to logarithm of relative risk $\boldsymbol{s}$. In all four of these models, α represents a constant variable to standardized $\boldsymbol{s}$ and τ^2 represents a constant variance amongst n districts. Derived from the joint distribution in (3.14), the conditional distribution of $s_i|\boldsymbol{s}_{-i}, \alpha, \beta, \phi$ is:

$$s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta} \sim N(\alpha + x_i\beta + \phi \sum_{j \in N_i}(c_{ij}(s_j - \alpha - x_j\beta)), \tau^2 m_{ii}), \tag{3.18}$$

where N_i is the set of neighbours of district i. At a higher level, diffused priors are assigned to α, β, τ, and ϕ: $\alpha \sim N(0, 1000^2)$, $\beta \sim N(0, 1000^2)$, $\tau^2 \sim \text{Inv-Gamma}(0.5, 0.0005)$, $\phi \sim \text{Unif}(\phi_{min}, \phi_{max})$. In model (3.2), we consider both spatial and linear effects of x_i in modelling $\boldsymbol{s}$, whereas model (3.15) considers only spatial effect; model (3.16) considers only linear effect; and model (3.17) considers neither spatial nor linear effect. We are interested in comparing the goodness-of-fit of the four models to lip cancer data set, so as to determine which model is the most appropriate. CVIC is one criterion for measuring goodness-of-fit. All the above four models belong to the class of Bayesian latent variable models depicted by Figure 2.1. The observable variable is y_i, the latent variable is s_i, and the model parameters $\boldsymbol{\theta}$ in model

Table 3.6: Parameter summary fo posterior inferences of fitting the four models to the full lip cancer data

Model	Parameters	Posterior distribution			
		Mean	2.5%	Median	97.5%
Full(spatial+linear)	α	-0.57	-0.89	-0.57	-0.23
	β	6.31	3.57	6.30	9.17
	ϕ	0.14	0.02	0.15	0.17
	τ^2	2.00	0.99	1.91	3.57
Spatial	α	-0.21	-0.52	-0.20	0.11
	ϕ	0.16	0.11	0.16	0.17
	τ^2	3.14	1.77	3.01	5.25
Linear	α	-0.49	-0.82	-0.49	-0.18
	β	6.83	3.96	6.82	9.74
	τ^2	0.36	0.20	0.36	0.62
Exchangeable	α	0.08	-0.16	0.08	0.31
	τ^2	0.61	0.36	0.59	0.97

(3.14) are $(\alpha, \beta, \tau, \phi)$, with a subset for other models depending on which are used in respective models.

We used OpenBUGS through R package R2OpenBUGS to run MCMC simulations for fitting each of the above models to Scottish lip cancer data. For each simulation, we ran two parallel chains, each with 15000 iterations, with the first 5000 discarded as burn-in. We used one actual MCMC simulation for each model to compute the posterior inference of parameters. The results of posterior inference based on the full data are summarized in Table 3.6.

For each model, we first ran 56 actual cross-validatory MCMC simulations with each of the 56 observations removed (set y_i^{obs} to `NA` in OpenBUGS), and then computed actual CV posterior predictive density $P(y_i^{\text{obs}}|y_{-i}^{\text{obs}})$ using equation (2.6) with evaluation function set to dpoisson$(y_i^{\text{obs}}|\lambda_i E_i)$ — Poisson probability mass function with parameter $\lambda_i E_i$. Because E_i plays an important role in computing actual CV posterior predictive density, we take a brief digression here to explain it. Because $\boldsymbol{E}$ is internally standardized in the data set, E_i is determined by y_i. It is necessary therefore to recalculate expected counts E_{-i}. When y_i is removed, change E_{-i} to $c_i E_{-i}$, where $c_i = \dfrac{\sum_{j\neq i}^n y_j}{\sum_{j\neq i}^n E_j}$. In this data set, we found only a tiny change in the values of E_{-i}.

We computed CVIC using equation (2.8). We computed actual CVIC 10 times for each model although actual LOOCV gives very stable results. The averages and standard deviations of 10 CVICs for different models are displayed in Table 3.7. From this table, we see that the full model is optimal for the Scottish lip cancer data according to CVIC.

We then consider approximating CVIC with four different methods (nIS, nWAIC, iIS, and iWAIC) from a single MCMC simulation based on all of the 56 observations. The non-integrated predictive density used in computing nIS and nWAIC with equations (2.14) and (2.25) is dpoisson($y_i^{\text{obs}}|\lambda_i E_i$), where $\lambda_i = \exp(s_i)$. Next, we describe how to compute iIS and iWAIC for model (3.14). The integrated predictive density (2.19) required by (2.22) and (2.27) is:

$$P(y_i^{\text{obs}} \,|\, \boldsymbol{\theta}, \boldsymbol{s}_{-i}) = \int \text{dpoisson}(y_i^{\text{obs}}|\lambda_i E_i) P(s_i \,|\, \boldsymbol{\theta}, \boldsymbol{s}_{-i}) ds_i, \tag{3.19}$$

where $P(s_i|\boldsymbol{\theta}, \boldsymbol{s}_{-i})$ is given by equation (3.18). Because there is no closed form for the integral (3.19), we use Monte Carlo method to estimate it by generating 200 random numbers from $P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})$ (note that this is done for each retained MCMC sample of $(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})$ and each validation unit i, with s_i alternately discarded). Finally, based on computed values of $P(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i})$ for all MCMC samples, we can compute iIS and iWAIC approximates of CV posterior predictive densities (with equations (2.22) and (2.27), respectively), corresponding iIS information criterion and iWAIC. iIS and iWAIC are computed similarly for models (3.15) - (3.17), with only a change in the conditional distribution (3.18) according to their joint prior distributions.

We repeated computing the values of the above four criteria as well as DIC for 100 independent MCMC simulations based on each model. The means of these 100 information criterion values for each method and each model are shown in Table 3.7, with standard deviations shown in brackets. We see that CVIC chooses the full model; for confirming the CVIC's choice, we can verify from Table 3.6 that the 95% credible intervals of parameter β and ϕ in full model from posterior distribution, $(3.57, 9.17)$ and $(0.02, 0.17)$, are not including zero, which means that linear and spatial effect can not be ignored.

We notice iIS and iWAIC provide significantly closer approximates to actual CVIC than nIS, nWAIC and DIC. Furthermore, iWAIC and iIS are almost identical to actual CVIC. In contrast, DIC has large biases and variances when spatial effects are considered, and also

Table 3.7: Comparisons of information criteria for lip cancer data. Except CVIC, each table entry shows the average of 100 information criterion values computed from 100 independent MCMC simulations, and the standard deviation in bracket. For CVIC, the average and standard deviation are from 10 independent LOOCV evaluations.

	CV	DIC	iWAIC	iIS	nWAIC	nIS
full	343.93(0.12)	269.43(12.30)	344.47(0.12)	345.21(0.19)	306.82(0.21)	335.54(1.27)
spatial	352.54(0.11)	266.79(10.15)	354.11(0.06)	356.06(0.37)	304.61(0.18)	338.77(1.85)
linear	349.46(0.10)	310.42(0.11)	350.48(0.05)	350.54(0.05)	306.94(0.21)	338.81(3.02)
exch.	366.59(0.00)	312.57(0.12)	368.01(0.03)	368.08(0.03)	306.74(0.17)	346.55(3.46)

the mean DIC of full model is bigger than the mean DIC of the model with spatial effects only. This suggests that if we randomly draw one MCMC simulation out of the 100 based on each model, the probability that DIC does *not* pick up the full model as the optimal model is high (56.6% if we assume the DICs are normally distributed). nWAIC and nIS also have large biases and variances. In particular, nWAIC nearly never chooses the full model (with a probability close to 1 if nWAICs are normally distributed). nIS has a good chance to choose the spatial+linear model (0.92 if the values are normally distribute). However, nIS is numerically unstable, with fairly large variance, which is well-known by many researchers (Spiegelhalter et al., 2002) [19]. In summary, the integration applied to latent variables associated with each validation unit substantially improves the estimates of CVIC given by nWAIC and nIS.

The good approximates of CVIC by iIS may not be surprising because our derivation in Section 2.3.2 has shown their equivalence in these models. It is surprising to note that the heuristic iWAIC also gives estimates very close to CVIC for model (3.14) and (3.15), which contain correlated random effects. Furthermore, note that iWAIC has smaller standard deviations and biases than iIS. Therefore, the equivalence of iWAIC to iIS (or CVIC) deserves more empirical and theoretical investigations in the future.

Chapter 4

Empirical Results on Detecting Divergent Regions

4.1 Detecting Divergent Regions Using p-value Based on CV Posterior Predictive Distribution of $\boldsymbol{y}_i$

In this section, we are interested in detecting the outliers in Scotland lip cancer data. This data set is the same as the data set shown in Table 3.5 from section 3.2. We wish to remind the reader that the data set has the following fields: (1) the number of observed cases of lip cancer, y_i; (2) the number of expected cases, E_i; (3) the standardized morbidity ratio (SMR_i) for the ith districts, $SMR_i \equiv y_i/E_i$; and (4) the percent of the population employed in agriculture, fishing and forestry, x_i. By the conclusion presented in section 3.2, the lip cancer data set is well-fit to the full model (spatial + linear):

$$y_i|E_i, \lambda_i \sim \text{Poisson}(\lambda_i E_i) \tag{4.1}$$

$$\boldsymbol{s} \sim N_n(\alpha + \boldsymbol{X}\beta, \Phi\tau^2). \tag{4.2}$$

Note that $\boldsymbol{s} = \log \boldsymbol{\lambda}$. The full model is a member of the Bayesian latent variable models depicted by Figure 2.1. The observable variable is y_i, the latent variable b_i is s_i, the covariate variable vector is $\boldsymbol{X}$, and the model parameter vector $\boldsymbol{\theta}$ is $(\alpha, \beta, \tau, \phi)$. We consider comparing different methods for computing posterior p-values in order to identify outliers in Scotland lip cancer data. We used OpenBUGS through R package R2OpenBUGS to run MCMC for fitting the full model to Scottish lip cancer data. For each simulation, we ran two parallel chains, each with 15000 iterations with 5000 iterations for burning in and 10000 for sampling.

The p-value(given parameters and latent variable) defined by (2.9) for this example is:

$$\begin{aligned} a_0(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) &= \text{p-value}(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) && (4.3)\\ &= Pr(y_i > y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) + 0.5Pr(y_i = y_i^{\text{obs}}|\boldsymbol{\theta}, s_i) && (4.4)\\ &= 1 - \text{ppoisson}(y_i^{\text{obs}}; \lambda_i E_i) + 0.5\text{dpoisson}(y_i^{\text{obs}}; \lambda_i E_i), && (4.5) \end{aligned}$$

where y_i^{obs} is the actual observation for y_i, and ppoisson and dpoisson denote CDF and PMF of Poisson distribution. Very small or very large p-values indicate that the actual observed y_i^{obs} falls on the lower tail of Poisson($\lambda_i E_i$) (i.e. is unusual to Poisson). Those extreme p-values identify outliers in Scotland lip cancer data. We also evaluate two probabilities to have a more thorough picture:

$$\begin{aligned} a_1(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) &= Pr(y_i > y_i^{\text{obs}}|\boldsymbol{\theta}, s_i), && (4.6)\\ a_2(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) &= Pr(y_i = y_i^{\text{obs}}|\boldsymbol{\theta}, s_i). && (4.7) \end{aligned}$$

Therefore, it is clear that $a_o(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) = a_1(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i) + 0.5a_2(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$.

In this example, CV posterior p-value (Marshall and Spiegelhalter, 2003) [13] for observation y_i^{obs} is the mean of p-value$(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$ with respect to the CV posterior distribution $P(\boldsymbol{\theta}, s_i|\boldsymbol{y}_{-i}^{\text{obs}})$:

$$\text{p-value}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = Pr(y_i > y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) + 0.5Pr(y_i = y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}). \qquad (4.8)$$

To detect outliers, we are interested in computing the following three probabilities (considering the p-value as a probability), which can be written as mean of $a_o(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$, $a_1(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$ and $a_2(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$, with respect to CV posterior distribution of $(\boldsymbol{\theta}, s_i)$ defined by (2.6):

$$\begin{aligned} Pr(y_i > y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) &= E_{\text{post(-i)}}\left[a_1(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)\right] && (4.9)\\ Pr(y_i = y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) &= E_{\text{post(-i)}}\left[a_2(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)\right] && (4.10)\\ \text{p-value}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) &= E_{\text{post(-i)}}\left[a_0(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)\right]. && (4.11) \end{aligned}$$

Note that:

$$\text{p-value}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}) = E_{\text{post(-i)}}\big[a_1(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)\big] + 0.5E_{\text{post(-i)}}\big[a_2(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)\big]. \tag{4.12}$$

We carried out 56 actual cross-validatory MCMC simulations, and used the MCMC samples of $(\boldsymbol{\theta}, \boldsymbol{s}_i)$ to estimate the two probabilities in (4.9) and (4.10), and the p-value in (4.11). The results of estimating the two probabilities and p-value are shown in column "CV" of Table B.1.

In Table 4.1, we show some selected results of the CV p-values and two probabilities. We can see that we get very small or very large CV posterior p-value for some observation y_i^{obs}. This indicates that the y_i^{obs} is unusual to the predictive distribution of y_i given $\boldsymbol{y}_{-i}^{\text{obs}}$. In this example, when CV posterior p-value for y_i^{obs} is very small or very large, the ith district is probably an outlier to other districts (e.g. CV posterior p-values of district 2 and district 55 are 0.03 and 0.99).

Table 4.1: The results of quantities of three probabilities: $Pr(y_i > y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$, $Pr(y_i = y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$ and p-value$(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$, which are means of $a_0(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$, $a_1(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$ and $a_2(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)$ with respect to posterior distribution of $(\boldsymbol{\theta}, s_i)$ for different methods under selected districts.

ID	$Pr(y_i > \boldsymbol{y}_i^{\text{obs}}\|\boldsymbol{y}_{-i}^{\text{obs}})$					$Pr(y_i = y_i^{\text{obs}}\|\boldsymbol{y}_{-i}^{\text{obs}})$					p-value$(y^{\text{obs}}\|\boldsymbol{y}_{-i}^{\text{obs}})$				
	CV	PCH	GHO	nIS	iIS	CV	PCH	GHO	nIS	iIS	CV	PCH	GHO	nIS	iIS
1	0.29	0.37	0.30	0.28	0.29	0.03	0.09	0.03	0.05	0.03	0.31	0.42	0.32	0.30	0.31
2	0.03	0.30	0.05	0.03	0.03	0.00	0.04	0.01	0.00	0.00	**0.03**	**0.32**	**0.05**	**0.03**	**0.03**
3	0.08	0.29	0.09	0.10	0.08	0.02	0.08	0.02	0.02	0.02	0.09	0.33	0.10	0.12	0.09
11	0.12	0.30	0.12	0.10	0.11	0.02	0.07	0.02	0.02	0.02	0.13	0.34	0.13	0.11	0.12
15	0.06	0.24	0.06	0.07	0.05	0.01	0.06	0.01	0.02	0.01	0.06	0.27	0.07	0.07	0.06
17	0.55	0.36	0.54	0.46	0.55	0.11	0.20	0.12	0.14	0.11	0.60	0.47	0.60	0.53	0.61
26	0.04	0.20	0.05	0.05	0.04	0.01	0.06	0.02	0.02	0.01	0.05	0.22	0.06	0.05	0.05
38	0.06	0.18	0.07	0.05	0.06	0.03	0.08	0.04	0.03	0.03	0.07	0.22	0.09	0.07	0.07
42	0.99	0.82	0.97	0.98	0.99	0.00	0.06	0.01	0.01	0.00	0.99	0.85	0.98	0.99	0.99
45	0.95	0.78	0.89	0.95	0.96	0.01	0.05	0.02	0.01	0.01	0.96	0.80	0.91	0.96	0.96
49	0.99	0.85	0.94	0.98	0.99	0.00	0.03	0.01	0.00	0.00	0.99	0.87	0.95	0.99	0.99

Table 4.1 – *Continued from previous page*

ID	$Pr(y_i > y_i^{\text{obs}} \vert \boldsymbol{y}_{-i}^{\text{obs}})$					$Pr(y_i = y_i^{\text{obs}} \vert \boldsymbol{y}_{-i}^{\text{obs}})$					p-value$(y_i^{\text{obs}} \vert \boldsymbol{y}_{-i}^{\text{obs}})$				
	CV	PCH	GHO	nIS	iIS	CV	PCH	GHO	nIS	iIS	CV	PCH	GHO	nIS	iIS
50	0.94	0.78	0.91	0.93	0.94	0.02	0.08	0.03	0.03	0.02	0.96	0.82	0.93	0.95	0.96
55	0.98	0.85	0.97	0.97	0.98	0.02	0.15	0.03	0.03	0.02	**0.99**	**0.92**	**0.99**	**0.99**	**0.99**
56	0.68	0.45	0.67	0.64	0.68	0.32	0.55	0.33	0.36	0.32	0.84	0.73	0.83	0.82	0.84

Abbreviations: CV: cross validation, PCH: posterior checking, GHO: Ghosting, nIS: non-integrated Importance Sampling, iIS: integrated Importance Sampling

There are three methods proposed in the literature for estimating the p-values (and the two probabilities) with only MCMC simulation based on the full data set. One method is to apply the posterior checking concept of Gelman et al. (1996) [5] without considering bias-correction. That is, to average each $a_0(y_i^{\text{obs}}|\boldsymbol{\theta}, s_1)$ with respect to the posterior distribution of $(\boldsymbol{\theta}, s_i)$ given the full data set $\boldsymbol{y}_{1:56}^{\text{obs}}$ (full data posterior distribution). We will call this method *posterior checking*:

$$\widehat{\text{p-value}}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{Post.check}} = \hat{E}_{\text{post(-i)}}^{\text{Post.check}}\left[a_0(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)\right] \tag{4.13}$$

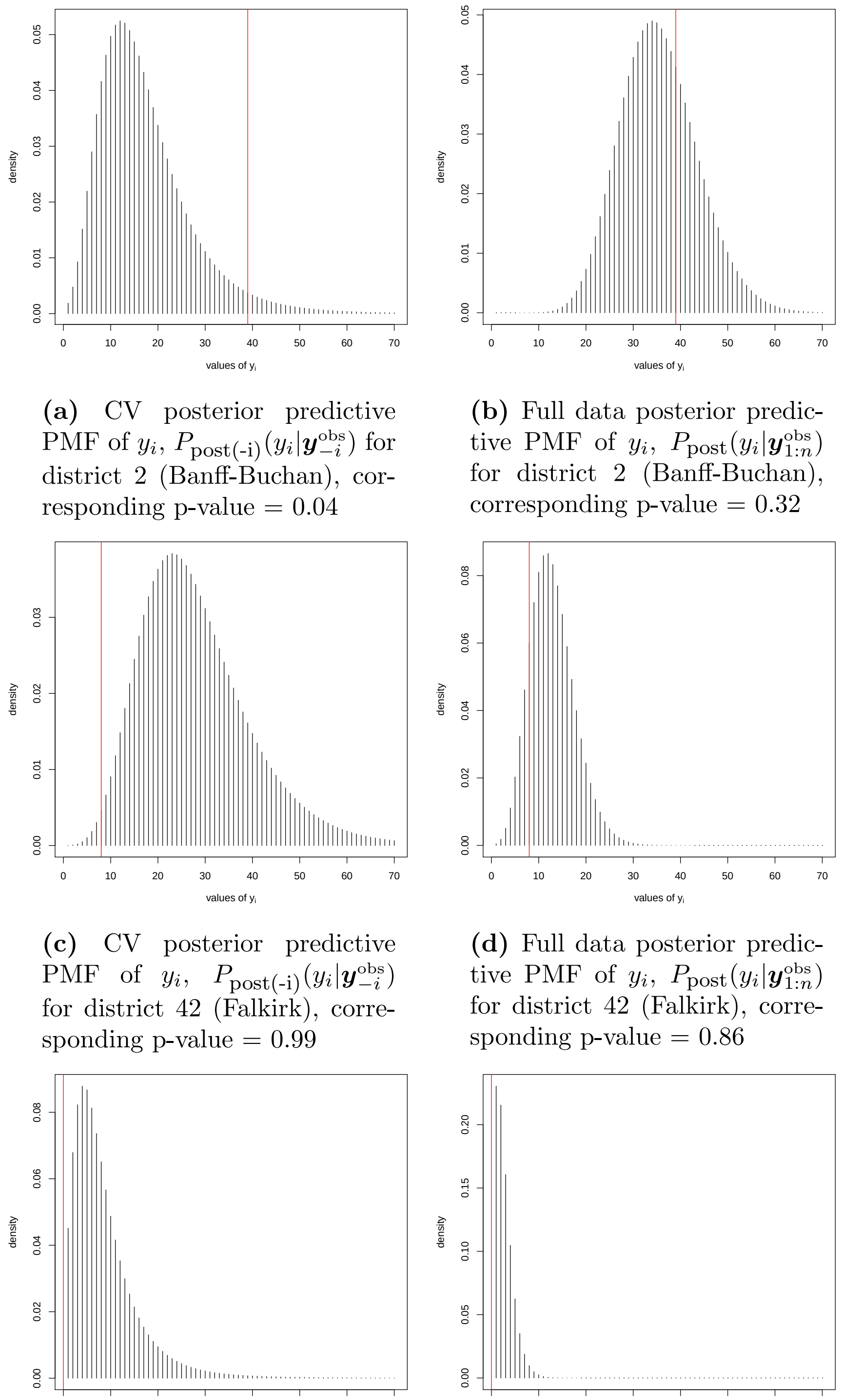
$$= \hat{E}_{\text{post}}\left[a_1(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)\right] + 0.5\hat{E}_{\text{post}}\left[a_2(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)\right]. \tag{4.14}$$

We drew a single MCMC sample of $(\boldsymbol{\theta}, s_i)$ from the full data posterior distribution $P_{\text{post}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}|y_{1:n}^{\text{obs}})$ to estimate the probabilities and p-value. The results of the approximation using posterior check are shown in Table B.1 and Table 4.1 in the "PCH" column.

We also portray the two p-values values given by CV posterior distribution and full data posterior distribution as tails of Poisson$(\lambda_i E_i)$. We first ran 56 actual cross-validatory MCMC simulations with each of the 56 observations removed (set y_i^{obs} to NA in OpenBUGS) and then computed actual CV posterior predictive probability mass function (PMF) for y_i using the equation:

$$P_{\text{post(-i)}}(y_i|\boldsymbol{y}_{-i}^{\text{obs}}) = \int P(y_i|\boldsymbol{\theta}, s_i)P(\boldsymbol{\theta}, s_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})d\boldsymbol{\theta}ds_{1:n}, \tag{4.15}$$

for $y_i = 0, 1, 2, 3, \ldots$. For each value of y_i (up to a limit), the above integration was found by using MCMC samples of $\boldsymbol{\theta}, s_i$ from the CV posterior $P(\boldsymbol{\theta}, s_{1:n}|\boldsymbol{y}_{-i})$. Secondly, we ran a MCMC simulation based on the full data set and computed full data posterior predictive



(a) CV posterior predictive PMF of y_i, $P_{\text{post(-i)}}(y_i|\boldsymbol{y}_{-i}^{\text{obs}})$ for district 2 (Banff-Buchan), corresponding p-value $= 0.04$

(b) Full data posterior predictive PMF of y_i, $P_{\text{post}}(y_i|\boldsymbol{y}_{1:n}^{\text{obs}})$ for district 2 (Banff-Buchan), corresponding p-value $= 0.32$

(c) CV posterior predictive PMF of y_i, $P_{\text{post(-i)}}(y_i|\boldsymbol{y}_{-i}^{\text{obs}})$ for district 42 (Falkirk), corresponding p-value $= 0.99$

(d) Full data posterior predictive PMF of y_i, $P_{\text{post}}(y_i|\boldsymbol{y}_{1:n}^{\text{obs}})$ for district 42 (Falkirk), corresponding p-value $= 0.86$

(e) CV posterior predictive PMF of y_i, $P_{\text{post(-i)}}(y_i|\boldsymbol{y}_{-i}^{\text{obs}})$ for district 55 (Annandale), corresponding p-value $= 0.99$

(f) Full data posterior predictive PMF of y_i, $P_{\text{post}}(y_i|\boldsymbol{y}_{1:n}^{\text{obs}})$ for district 55 (Annandale), corresponding p-value 0.92

Figure 4.1: Posterior predictive distribution of y_i, given full data compared to CV posterior predictive distribution of $\boldsymbol{y}_i$ and data set with y_i^{obs} omitted. The red vertical line represents the value of y_i^{obs}.

probability mass function for $y_i = 0, 1, 2, 3, \ldots$:

$$\hat{P}^{\text{Post.check}}_{\text{post(-i)}}(y_i|\boldsymbol{y}^{\text{obs}}_{-i}) = P^{\text{Post.check}}_{\text{post}}(y_i|\boldsymbol{y}^{\text{obs}}_{1:n}) \tag{4.16}$$

$$= \int P(y_i|\boldsymbol{\theta}, s_i)P(\boldsymbol{\theta}, s_{1:n}|\boldsymbol{y}^{\text{obs}}_{1:n})d\boldsymbol{\theta}ds_{1:n}. \tag{4.17}$$

Similarly, for each value of y_i, the integration is found by using MCMC samples of $\boldsymbol{\theta}, s_i$ from the full data posterior $P(\boldsymbol{\theta}, s_{1:n}|\boldsymbol{y}^{\text{obs}}_{1:n})$. We compared the above two PMFs in Figure 4.1 with red vertical lines indicating the actual observed values of y^{obs}_i for the three selected districts, 2, 42 and 55. Figure 4.1 draws the shape of tails of Poisson($\lambda_i E_i$) cut by observation y^{obs}_i, in which the p-values (using either CV or full data posterior) are read as the sum of PMFs on the right of the actual observation y^{obs}_i, with only halp of the PMF at y^{obs}_i. In Figure 4.1, we can see that the full data posterior fits the actual observations better than the CV-posterior. This is called optimistic bias. The consequence is that the posterior p-values will tend to be more concentrated to 0.5 than the CV posterior p-values, as shown by Figure 4.2a.

In order to reduce the bias of including y^{obs}_i in model fitting, Marshall and Spiegelhalter (2003) [13] propose *Ghosting method*: for each MCMC sample, one averages p-value($y^{\text{obs}}_i, \boldsymbol{\theta}, s_i$) with respect to the conditional distribution of s_i given $\boldsymbol{\theta}$ (but without y^{obs}_i) to obtain Ghosting p-value. Ghosting method discards s_i associated with the y^{obs}_i, and re-generates it from the distribution without reference to the actual observation of y^{obs}_i using Monte Carlo method to compute the p-value.

The third method is the *non-integrated importance sampling* method (nIS) which averages p-value($y^{\text{obs}}_i, \boldsymbol{\theta}, s_i$) after being weighted with the inverse of probability density (mass) of y^{obs}_i. For computing integrated p-values and predictive densities as needed by nIS and Ghosting method, we generated 100 of s_i from

$$s_i|s_{-i}, \boldsymbol{\theta} \sim N(\alpha + x_i\beta + \phi \sum_{j \in N_i}(c_{ij}(s_j - \alpha - x_j\beta)), \tau^2 m_{ii}) \tag{4.18}$$

for each district and each MCMC sample.

The fourth method is the *integrated importance sampling* method (iIS). For each MCMC sample, we must first average p-value($y^{\text{obs}}_i, \boldsymbol{\theta}, s_i$) with respect to $P(s_i|\boldsymbol{\theta}, \boldsymbol{s}_{-i})$ in order to find integrated evaluation p-value (equation(2.16)) and integrate predictive density (equation(2.19)). Then we must find the weighted average of integrated p-values with reversed integrated predictive density as weights over all MCMC samples using formula (2.20). We see that the difference between Ghosting method and iIS is that iIS uses importance weighting to correct the bias in full data posterior of $\boldsymbol{\theta}$, but Ghosting method does not. Therefore, Ghosting method can be viewed as a partial implementation of iIS method presented here. We write the four methods estimating CV posterior p-values as expectations of evaluation functions, in which expectations are defined by (2.11) - (2.20):

$$\begin{aligned}
\widehat{\text{p-value}}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{nIS}} &= \frac{\hat{E}_{\text{post}}\left[a_0(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]} && (4.19)\\
&= \frac{\hat{E}_{\text{post}}\left[a_1(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]} && (4.20)\\
&+ \frac{0.5\hat{E}_{\text{post}}\left[a_2(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]} && (4.21)\\
\widehat{\text{p-value}}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{iIS}} &= \frac{\hat{E}_{\text{post}}\left[A_0(y_i^{\text{obs}}|\boldsymbol{\theta}, s_{-i})W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]} && (4.22)\\
&= \frac{\hat{E}_{\text{post}}\left[A_1(y_i^{\text{obs}}|\boldsymbol{\theta}, s_{-i})W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]} && (4.23)\\
&+ \frac{0.5\hat{E}_{\text{post}}\left[A_2(y_i^{\text{obs}}|\boldsymbol{\theta}, s_{-i})W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]} && (4.24)\\
\widehat{\text{p-value}}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{Ghost}} &= \hat{E}_{\text{post}}\left[A_0(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right] && (4.25)\\
&= \hat{E}_{\text{post}}\left[A_1(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right] + 0.5\hat{E}_{\text{post}}\left[A_2(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right], && (4.26)
\end{aligned}$$

where

$$
\begin{aligned}
A_0(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i}) &= \int a_0(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})ds_i && (4.27)\\
A_1(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i}) &= \int a_1(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})ds_i && (4.28)\\
A_2(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i}) &= \int a_2(y_i^{\text{obs}}, \boldsymbol{\theta}, s_i)P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})ds_i && (4.29)\\
W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}) &= \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)} && (4.30)\\
W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i}) &= \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i})} && (4.31)\\
&= \frac{1}{\int P(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})ds_i}, && (4.32)
\end{aligned}
$$

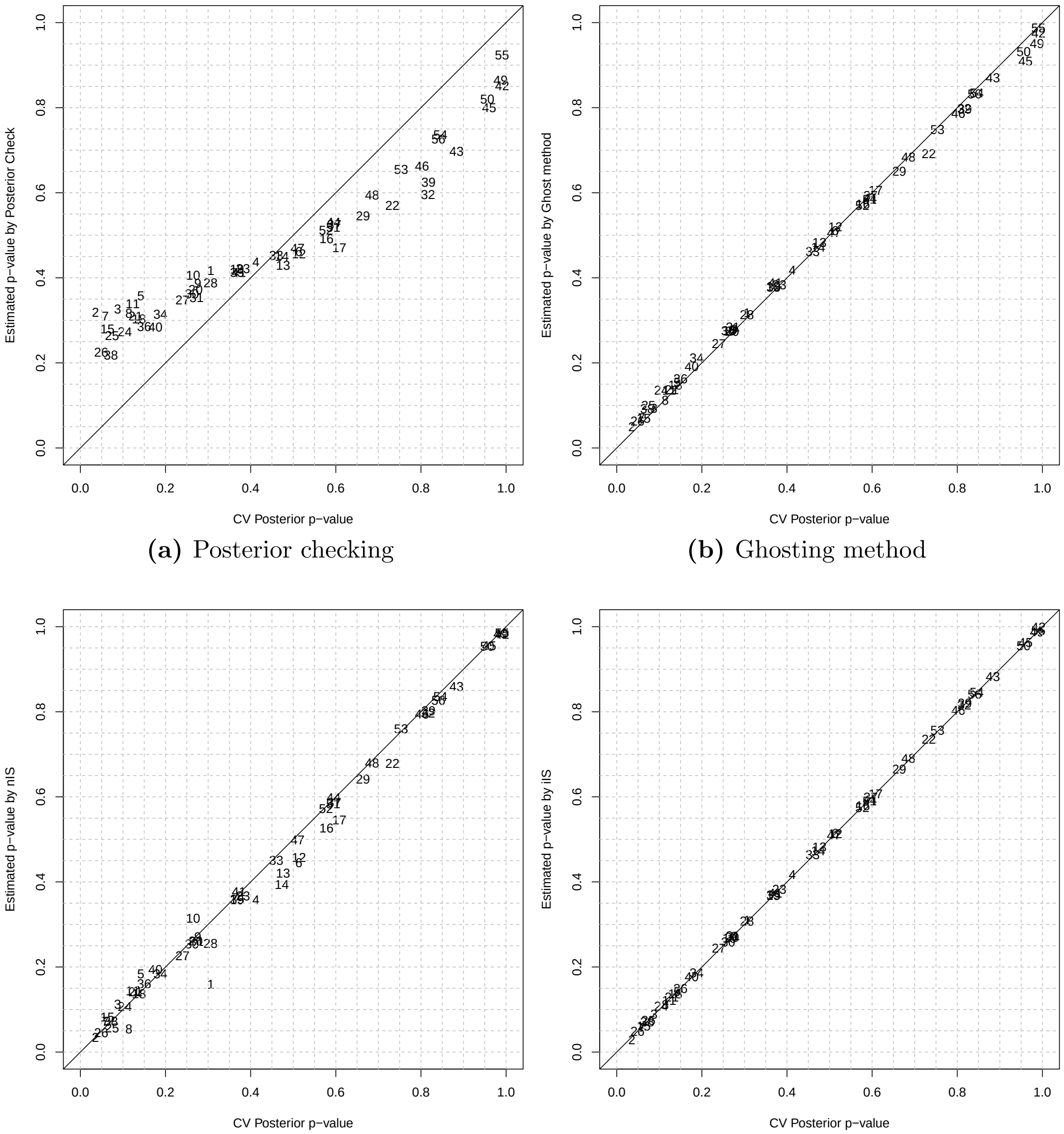
noting that $a_0()$, $a_1()$ and $a_2()$ are defined by equation (4.3),(4.6) and (4.7).

Using the above four methods, we calculated 56 posterior p-values, given a MCMC simulation based on the full data set shown by Table B.1 and p-values of selected districts shown in Table 4.1. Figure 4.2 shows scatter-plots of four sets of estimated CV posterior p-values, given by four different methods against the actual CV posterior p-values from one MCMC simulation. In Figure 4.2, we can see that the actual p-values given by posterior checking are more concentrated around 0.5 than the actual CV posterior p-value, and do not appear to be uniformly distributed (Gelman et al., 2013) [6]. Ghosting method reduces the bias; hence, the estimated p-values are closer to the actual CV p-values, and more spread out over $(0, 1)$. However, for this example, the bias is still visible from Figure 4.2b. Both nIS and iIS give estimate p-values very close to the actual values found by CV. However, nIS is less stable than iIS, and sometimes gives very bad estimate (e.g for the district Skye-Lochalsh with ID = 1 shown in Figure 4.2c).

Table 4.2: Comparisons of Relative Errors of Estimated CV p-values

iIS	nIS	Ghost	Post. check
1.501(0.210)	12.481(1.586)	19.212(0.359)	160.580(1.101)

To measure more precisely the accuracy of estimated p-values to the actual CV p-values,



(a) Posterior checking

(b) Ghosting method

(c) Non-integrated IS (nIS)

(d) Integrated IS (iIS)

Figure 4.2: Scatterplots of estimated posterior p-values in full data from an MCMC simulation against actual CV posterior p-values. The number points show indices of districts.

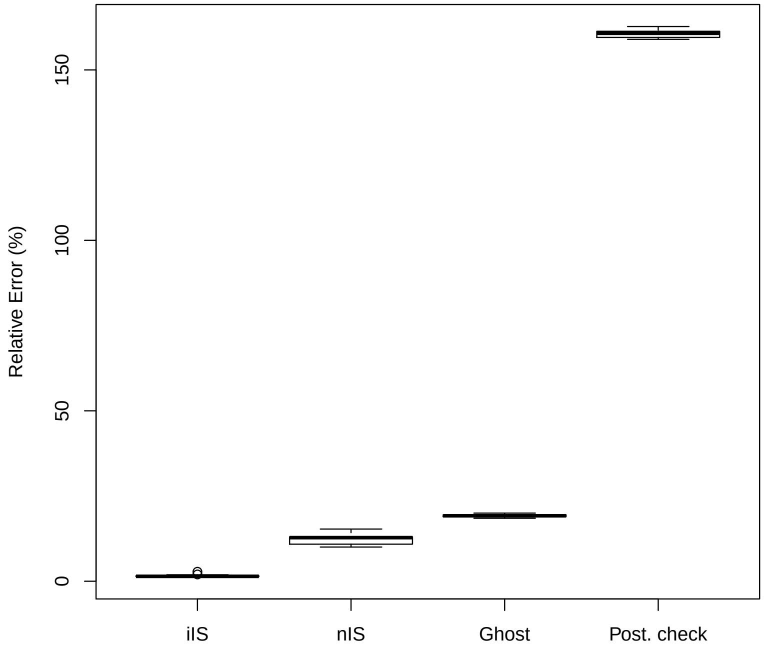


Figure 4.3: Box-plots of relative errors of P-value in different methods in 100 replicates of MCMC simulations given the full data

we use absolute relative error, defined as:

$$RE = (1/n) \sum_{i=1}^{n} \frac{|\hat{p}_i - p_i|}{\min(p_i, 1 - p_i)} \times 100, \tag{4.33}$$

where $\hat{\boldsymbol{p}}_{1:n}$ are estimates of $\boldsymbol{p}_{1:n}$. This measure greatly emphasize the error between $\hat{p}_i$ and p_i when p_i is very small or very large, for which we demand more absolute error than when p_i is close to 0.5. A similar measure (only using p_i in denominator) is used by Marshall and Spiegelhalter (2007) [14]. Here, we modify the denominator because large p-values are also important. Table 4.2 and Figure 4.3 show the averages of REs over 100 independent simulations for each method. Clearly, we see that iIS is the best method among the four, as it is a significant improvement from Ghosting and posterior checking methods.

4.2 Detecting Divergent Regions using p-value based on CV posterior predictive distribution of $\boldsymbol{\lambda}_i$

In this section, we will use another definition of p-value which computes probability of λ_i with CV posterior predictive distribution greater to actual observation SMR_i. The data set and model we used here is the same as in Section 3.2 and in Section 4.1. This CV posterior p-value for λ_i is defined as follows:

$$\text{p-value}(SMR_i|\boldsymbol{y}_{-i}^{\text{obs}}) = Pr(\lambda_i > SMR_i|\boldsymbol{y}_{-i}^{\text{obs}}). \tag{4.34}$$

In order to compute the above p-value with different approximating methods, we further define evaluation function of p-value of λ_i:

$$b(SMR_i, \boldsymbol{\theta}, s_i) = I(\lambda_i > SMR_i) \text{ noting } \lambda_i = \exp(s_i). \tag{4.35}$$

In this way, the above p-value can be expressed as:

$$\text{p-value}(SMR_i|\boldsymbol{y}_{-i}^{\text{obs}}) = E_{\text{post(-i)}}\left[b(SMR_i, \boldsymbol{\theta}, s_i)\right]. \tag{4.36}$$

We carried out 56 actual cross-validatory MCMC simulations, and used the MCMC samples of $(\boldsymbol{\theta}, \boldsymbol{s}_i)$ to estimate the p-value (4.36). The results of estimating the p-value are shown in column "CV" of Table 4.3.

Table 4.3: p-value based on $P(\lambda_i|\boldsymbol{y}_{-i}^{\text{obs}})$ for all districts

ID	District name	CV	Posterior checking	Ghosting	nIS	iIS
1	Skye-Lochalsh	0.31	0.40	0.31	0.26	0.30
2	Banff-Buchan	0.03	0.26	0.04	0.05	0.02
3	Caithness	0.07	0.27	0.08	0.09	0.07
4	Berwickshire	0.42	0.43	0.42	0.39	0.41
5	Ross-Cromarty	0.13	0.31	0.14	0.14	0.12

Table 4.3 – *Continued from previous page*

ID	District name	CV	Posterior checking	Ghosting	nIS	iS
6	Orkney	0.52	0.47	0.51	0.42	0.52
7	Moray	0.05	0.24	0.06	0.07	0.05
8	Shetland	0.09	0.26	0.10	0.08	0.09
9	Lochaber	0.26	0.36	0.27	0.21	0.26
10	Gorden	0.26	0.37	0.26	0.23	0.25
11	Western Isles	0.11	0.29	0.12	0.12	0.10
12	Sutherland	0.52	0.47	0.52	0.47	0.52
13	Nairn	0.49	0.44	0.48	0.43	0.48
14	Wigtown	0.47	0.45	0.48	0.47	0.47
15	North East Fife	0.05	0.20	0.05	0.04	0.04
16	Kincardine	0.60	0.52	0.59	0.61	0.59
17	Badenoch	0.62	0.50	0.62	0.56	0.62
18	Ettrick	0.11	0.25	0.11	0.11	0.11
19	Inverness	0.36	0.40	0.36	0.34	0.35
20	Roxburgh	0.24	0.33	0.24	0.21	0.24
21	Angus	0.10	0.24	0.10	0.11	0.09
22	Aberdeen	0.77	0.61	0.72	0.72	0.76
23	Argyll-Bute	0.37	0.41	0.37	0.38	0.37
24	Clydesdale	0.07	0.20	0.10	0.07	0.07
25	Kirkcaldy	0.05	0.17	0.06	0.04	0.04
26	Dunfermline	0.03	0.13	0.03	0.03	0.02
27	Nithsdale	0.20	0.29	0.20	0.17	0.20
28	East Lothian	0.27	0.34	0.27	0.28	0.27
29	Perth-Kinross	0.70	0.59	0.69	0.67	0.70
30	West Lothian	0.22	0.31	0.24	0.22	0.22
31	Cumnock-Doon	0.24	0.31	0.24	0.23	0.23
32	Stewartry	0.86	0.69	0.85	0.84	0.86

Table 4.3 – *Continued from previous page*

ID	District name	CV	Posterior checking	Ghosting	nIS	iIS
33	Midlothian	0.46	0.44	0.46	0.46	0.46
34	Stirling	0.14	0.24	0.17	0.14	0.14
35	Kyle-Carrick	0.34	0.39	0.35	0.33	0.34
36	Inverclyde	0.10	0.20	0.11	0.11	0.10
37	Cunninghame	0.62	0.55	0.62	0.67	0.63
38	Monklands	0.03	0.12	0.05	0.03	0.03
39	Dumbarton	0.88	0.73	0.85	0.86	0.87
40	Clydebank	0.12	0.20	0.13	0.12	0.12
41	Renfrew	0.33	0.38	0.35	0.35	0.35
42	Falkirk	1.00	0.96	0.99	1.00	1.00
43	Clackmannan	0.95	0.86	0.95	0.94	0.95
44	Motherwell	0.63	0.58	0.65	0.63	0.65
45	Edinburgh	0.98	0.93	0.97	0.99	0.99
46	Kilmarnock	0.89	0.81	0.88	0.89	0.90
47	East Kilbride	0.53	0.49	0.53	0.53	0.53
48	Hamilton	0.79	0.72	0.79	0.79	0.80
49	Glasgow	0.99	0.96	0.99	1.00	1.00
50	Dundee	0.99	0.95	0.98	0.99	0.99
51	Cumbernauld	0.67	0.61	0.67	0.67	0.68
52	Bearsden	0.64	0.58	0.64	0.65	0.65
53	Eastwood	0.92	0.87	0.92	0.92	0.92
54	Strathkelvin	0.98	0.95	0.98	0.98	0.98
55	Annandale	1.00	1.00	1.00	1.00	1.00
56	Tweeddale	1.00	1.00	1.00	1.00	1.00

In Table 4.3, we can see that CV p-values of district 2, 7, 15, 25, 26 and 38 are smaller than 0.05 and CV p-values of district 42, 49, 50, 55 and 56 are greater than 0.99. Whereas in

Section 3.2, Table 4.1 shows CV p-values of district 2 and 26 are smaller than 0.05 and CV p-values of districts 42, 49 and 55 are greater than 0.99.

We used three other methods to estimate the CV posterior p-value: posterior checking, non-integrated importance sampling(nIS) and integrated importance sampling(iIS). The expressions for these methods are as follows:

$$\begin{aligned}
\widehat{\text{p-value}}(SMR_i|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{Post.check}} &= \hat{E}_{\text{post}}\left[b(SMR_i, \boldsymbol{\theta}, s_i)\right] && (4.37)\\
\widehat{\text{p-value}}(SMR_i|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{Ghost}} &= \hat{E}_{\text{post}}\left[B(SMR_i, \boldsymbol{\theta}, \boldsymbol{s}_{-i})\right] && (4.38)\\
\widehat{\text{p-value}}(SMR_i|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{nIS}} &= \frac{\hat{E}_{\text{post}}\left[b(SMR_i|\boldsymbol{\theta}, s_i)W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n})\right]} && (4.39)\\
\widehat{\text{p-value}}(SMR_i|\boldsymbol{y}_{-i}^{\text{obs}})^{\text{iIS}} &= \frac{\hat{E}_{\text{post}}\left[B(SMR_i, \boldsymbol{\theta}, \boldsymbol{s}_{-i})W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]}{\hat{E}_{\text{post}}\left[W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i})\right]}, && (4.40)
\end{aligned}$$

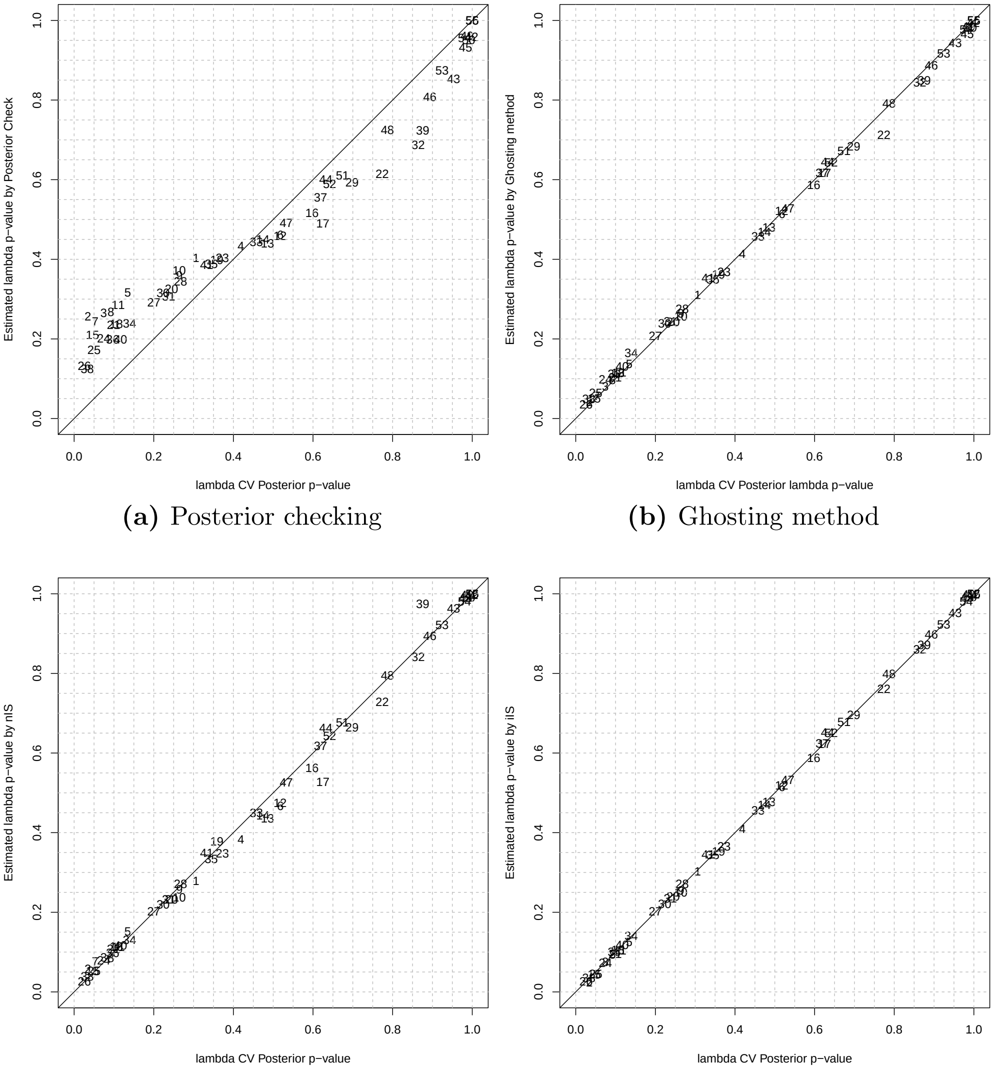
where

$$\begin{aligned}
B(SMR_i, \boldsymbol{\theta}, \boldsymbol{s}_{-i}) &= \int b(SMR_i, \boldsymbol{\theta}, s_i)P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})ds_i && (4.41)\\
W_i^{\text{nIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{1:n}) &= \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)} && (4.42)\\
W_i^{\text{iIS}}(\boldsymbol{\theta}, \boldsymbol{s}_{-i}) &= \frac{1}{P(y_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{s}_{-i})} && (4.43)\\
&= \frac{1}{\int P(y_i^{\text{obs}}|\boldsymbol{\theta}, s_i)P(s_i|\boldsymbol{s}_{-i}, \boldsymbol{\theta})ds_i}, && (4.44)
\end{aligned}$$

as $b(SMR_i, \boldsymbol{\theta}, s_i)$ is defined in equation (4.35).

Given an MCMC simulation based on the full data set, we calculated 56 posterior p-values with the four methods and repeated this calculation for 100 independent MCMC simulations. The results of quantities of p-values with the four approximating CV methods are shown in Table 4.3.

We present a scatterplot of posterior p-values on $\boldsymbol{\lambda}$ using four methods with full data against actual CV posterior p-values in Figure 4.4. We can see that in Figure 4.4a, p-values given by posterior checking are more concentrated around 0.5 than the actual CV posterior p-value, and do not appear to be uniformly distributed (Gelman et al., 2013) [7]. It can be observed that extreme p-values (identifying outliers) in both lower and upper tails in this graph are quite different in CV and posterior checking. For example, in Table 4.3, we see that the CV posterior p-value is 0.03 compared to 0.26 in posterior checking for district 2. We call this gap between CV posterior p-values and posterior checking p-values *optimistic bias*.



(a) Posterior checking

(b) Ghosting method

(c) Non-integrated IS (nIS)

(d) Integrated IS (iIS)

Figure 4.4: Scatterers of posterior p-values on $\boldsymbol{\lambda}$ in four methods given full data against actual CV posterior p-values on $\boldsymbol{\lambda}$.

To assess the bias more visually, we portray p-values of λ_i with respect to CV posterior distribution and respect to full data posterior distribution. Tail of both densities of $\lambda_i s$ are taking SMR_i as critical value indicating $I(\lambda_i > SMR_i)$. We plot a histogram that approximates the density of λ_i using MCMC simulations of s_i from posterior distribution, marking SMR_i in this histogram shown in Figure 4.5. We observe that these shapes of full data posterior densities of λ_i are much closer than the shapes of CV posterior densities. This is because when computing each p-value, the observed value y_i^{obs} itself is included in model fitting, resulting in optimistic bias.

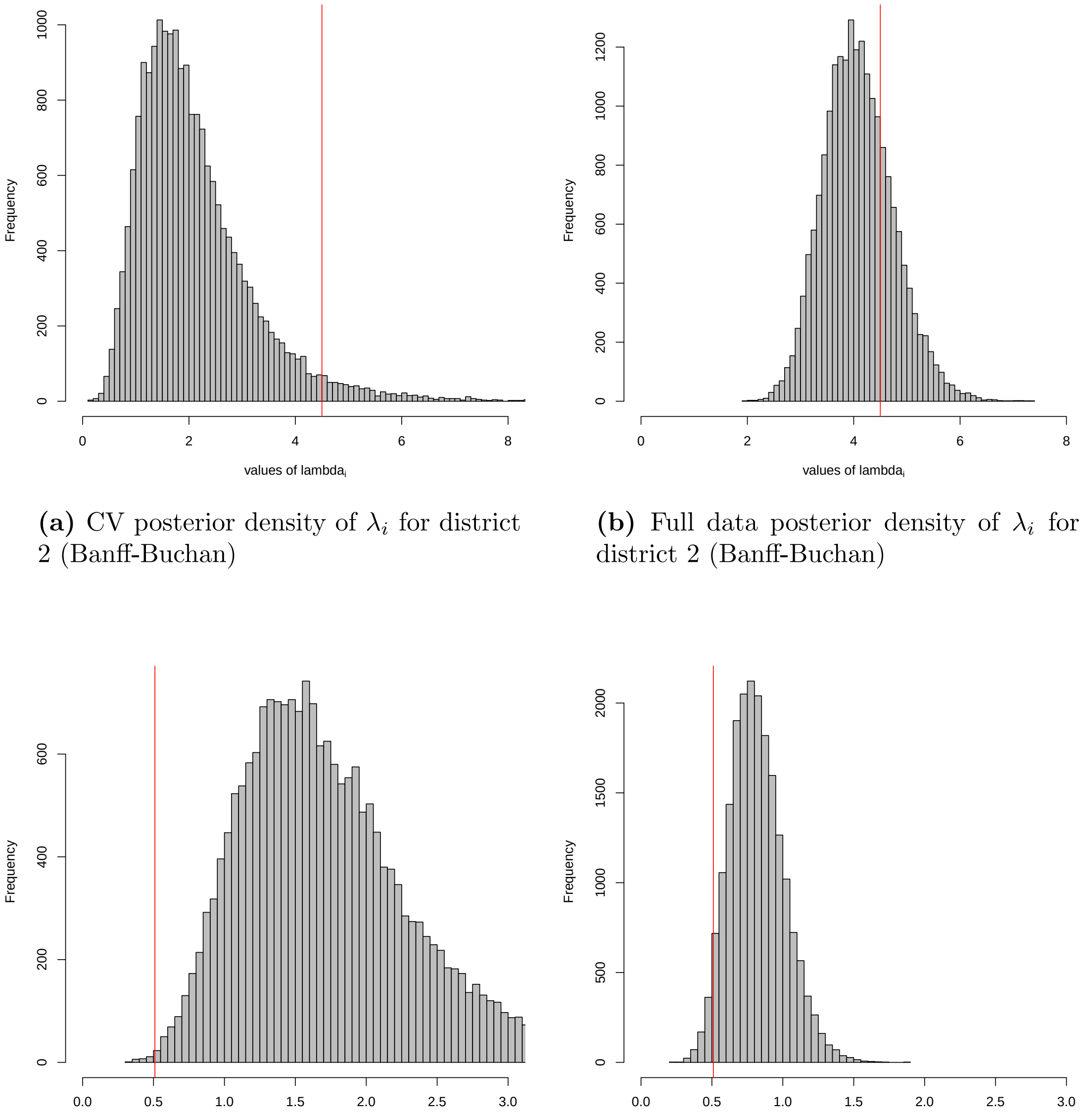
In Figures 4.4b, 4.4c and 4.4d, we see that Ghosting, nIS and iIS methods approximate CV posterior p-values much better than posterior checking and greatly reduce the optimistic bias. In Figure 4.4b, p-values with Ghosting method are more spread out in $(0, 1)$ so that the bias is still visible. Compared with p-values with iIS, p-values with nIS are unstable in graph 4.4c, which provides the best approximates of CV posterior p-values of the four methods.

We also apply absolute relative error formula (4.33) to measure the accuracy of estimated p-values to the actual CV p-values of $\boldsymbol{\lambda}$. Because the estimated and actual p-values for district 55 and 56 are equal to 1, denominator of relative error (4.33) is invalid. In this case, we discard the two districts' p-values then use the remaining p-values to compute the relative error. We display the results in Table 4.4 and Figure 4.6.

Table 4.4: Comparisons of relative errors of estimated CV p-values of $\boldsymbol{\lambda}$

iIS	nIS	Ghost	Post. check
5.486(0.195)	13.780(1.507)	18.726(0.995)	166.241(3.724)

Our purpose in researching CV approximation methods is to lower computing resource cost by CV method, especially in computing time. We recorded execution time of computing p-values and running MCMC for fitting the full model to Scottish lip cancer data. For each simulation, we ran two parallel chains, each with 15000 iterations, 5000 iterations for burning in, and 10000 for sampling. We considered time consumed in two parts: 1) running MCMC simulations of $(\boldsymbol{\theta}, s_i)$, and 2) computing means of evaluation function $(b(SMR_i, \boldsymbol{\theta}, s_i))$. In Table 4.5, we observe that the time spent on MCMC simulations in CV method is about 56 times that of other methods because cross validation computes MCMC simulation for each of the 56 observation unit. In contrast, the other methods use only one MCMC simulation, given full data. We also observe that for Ghosting and iIS methods, time is mainly spent on computing integrations of evaluation functions: re-generating s_i or integrating away the $(s_{1:n}, \boldsymbol{\theta})$ in Ghosting and iIS. Considering the total time, we see that iIS saves 85% computing



(a) CV posterior density of λ_i for district 2 (Banff-Buchan)

(b) Full data posterior density of λ_i for district 2 (Banff-Buchan)

(c) CV posterior of λ_i for district 42 (Falkirk)

(d) Full data posterior of λ_i for district 42 (Falkirk)

Figure 4.5: Histograms of λ_i simulated from CV posterior distribution and full data posterior distribution.

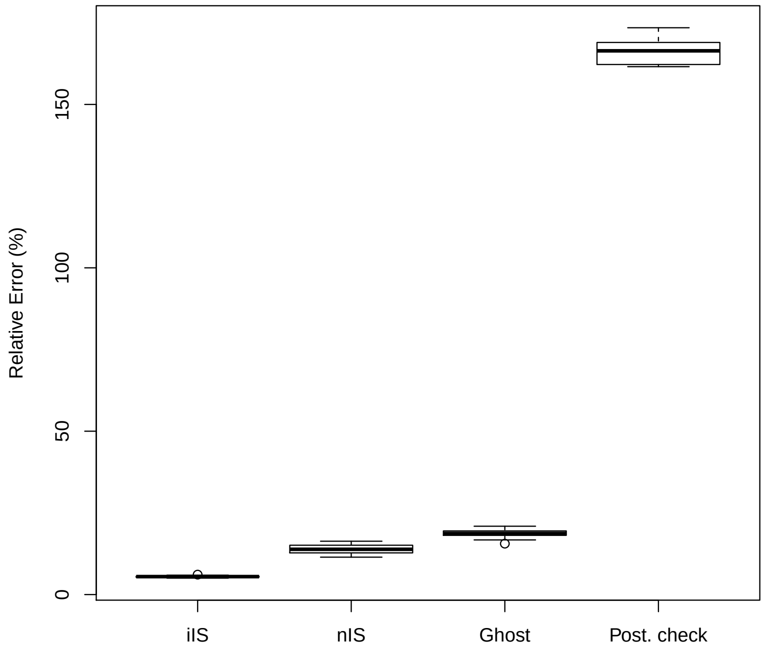


Figure 4.6: Box-plots of relative errors of p-valus of $\boldsymbol{\lambda}$ in different methods

time of p-value with CV, and it preforms best in approximating p-values to CV.

Table 4.5: Comparisons of user execution time(in seconds). "p-values" represents time spned on work process of calculating p-values

	CV	iIS	nIS	Ghost	Post. Check
MCMC	1037.35	19.95	19.98	20.14	19.91
p-values	0.12	133.61	10.94	113.91	0.18
Total	1037.47	153.56	30.92	134.04	20.08

Chapter 5
Conclusions and Future Work

The new proposed iIS and iWAIC significantly reduce the bias of nIS and nWAIC in evaluating Bayesian spatial models with unit-specific latent variables. We provide formulas for iWAIC and iIS that are applicable to general evaluation function.

In Chapter 3, we saw that iIS and iWAIC produce very close results for comparing competing models as what the actual CVIC provides. The results answer the hypothesis questions about whether spatial or linear effects are present in models. On the other hand, iIS and iWAIC save a great deal of computing resources compared to CVIC. In addition, iWAIC works well in the spatial random effect models, the result in which is surprising and encouraging. iIS and iWAIC provide new options with better performance in correcting optimistic bias than DIC, ordinary IS and WAIC.

In Chapter 4, we used iIS to compute CV posterior p-values, which aims to detect the outlier from Scotland lip cancer data. iIS has the lowest relative errors in approximating the CV posterior predictive p-values of $\boldsymbol{y}_i$ and $\boldsymbol{\lambda}_i$.

Although our empirical results show that iIS and iWAIC provide better approximates of CVIC than DIC, we notice that the implementations of iIS and iWAIC is much more complicated, and requires users to have background knowledge in statistics and scientific computing. To automate applications is a direction for future research one can pursue. One may consider investigating the validity of iWAIC theoretically due to the lack of applicability of iWAIC to models. In the future, we will empirically test iWAIC in many other models using correlated latent variables, such as stochastic volatility models, multivariate spatial models, etc.

References

[1] Peter Congdon. A model framework for mortality and health data classified by age, area, and time. *Biometrics*, 62(1):269278, 2006. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2005.00419.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2005.00419.x/abstract.

[2] Peter Congdon. *Bayesian statistical modelling*. Wiley, 2007.

[3] A E Gelfand, D K Dey, and H Chang. Model determination using predictive distributions with implementation via Sampling-Based methods (with discussion). Technical report, DTIC Document, 1992.

[4] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185, 1998.

[5] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

[6] Andrew Gelman, Christian P Robert, and Judith Rousseau. Inherent difficulties of non-bayesian likelihood-based inference, as revealed by an examination of a recent book by aitkin. *Statistics & Risk Modeling with Applications in Finance and Insurance*, 30(2): 105–120, 2013.

[7] Andrew Gelman et al. Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electronic Journal of Statistics*, 7:2595–2602, 2013.

[8] John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[9] Leonhard Held, Birgit Schrdle, and Hvard Rue. Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In Thomas Kneib and Gerhard Tutz, editors, *Statistical Modelling and Regression Structures*, pages 91–110. Physica-Verlag HD, jan 2010. ISBN 978-3-7908-2412-4, 978-3-7908-2413-1. URL http://link.springer.com/chapter/10.1007/978-3-7908-2413-1_6.

[10] Phaisarn Jeefoo, Nitin Kumar Tripathi, and Marc Souris. Spatio-temporal diffusion pattern and hotspot detection of dengue in chachoengsao province, thailand. *International journal of environmental research and public health*, 8(1):51–74, 2010.

[11] Longhai Li, Shi Qiu, Bei Zhang, and Cindy X. Feng. Approximating cross-validatory predictive evaluation in bayesian latent variables models with integrated IS and WAIC. *arXiv:1404.2918 [stat]*, apr 2014. URL http://arxiv.org/abs/1404.2918.

[12] J S Liu. *Monte Carlo Strategies in Scientific Computing.* Springer-Verlag, 2001.

[13] E C Marshall and D J Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Stat. Med.*, 22(10):1649–1660, 2003.

[14] E C Marshall and D J Spiegelhalter. Identifying outliers in bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, 2(2):409–444, 2007.

[15] Kari S McLeod. Our sense of snow: the myth of john snow in medical geography. *Social Science & Medicine*, 50(7):923–935, 2000.

[16] Simon Reading Moore, NW Johnson, Angela Mary Pierce, and David Francis Wilson. The epidemiology of lip cancer: a review of global incidence and aetiology. *Oral diseases*, 5(3):185–195, 1999.

[17] Radford M Neal. Probabilistic inference using markov chain monte carlo methods. 1993.

[18] Martyn Plummer. Penalized loss functions for bayesian model comparison. *Biostatistics*, 9(3):523–539, 2008.

[19] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. Bayesian measures of model complexity and fit. *JRSSB*, 64(4):583–639, oct 2002.

[20] Hal S. Stern and Noel Cressie. Posterior predictive model checks for disease mapping models. *Statistics in medicine*, 19(17-18):23772397, 2000. URL http://onlinelibrary.wiley.com/doi10.1002/1097-0258(20000915/30)19:17/18%3C2377::AID-SIM576%3E3.0.CO;2-1/abstract.

[21] Jarno Vanhatalo, Jaakko Riihimki, Jouni Hartikainen, Pasi Jylnki, Ville Tolvanen, and Aki Vehtari. Bayesian modeling with gaussian processes using the GPstuff toolbox. *arXiv:1206.5754 [cs, stat]*, jun 2012. URL http://arxiv.org/abs/1206.5754.

[22] Jarno Vanhatalo, Jaakko Riihimki, Jouni Hartikainen, Pasi Jylnki, Ville Tolvanen, and Aki Vehtari. GPstuff: bayesian modeling with gaussian processes. *The Journal of Machine Learning Research*, 14(1):11751179, 2013. URL http://dl.acm.org/citation.cfm?id=2502617.

[23] Aki Vehtari and Jouko Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput.*, 14(10):2439–2468, oct 2002.

[24] Aki Vehtari et al. *Bayesian model assessment and selection using expected utilities.* Helsinki University of Technology, 2001.

[25] Sumio Watanabe. *Algebraic geometry and statistical learning theory.* Cambridge University Press, 2009.

[26] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[27] Sumio Watanabe. Equations of states in singular statistical estimation. *Neural Networks*, 23(1):20–34, 2010.

[28] Sumio Watanabe. Equations of states in statistical learning for an unrealizable and regular case. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(3):617–626, 2010.

Appendix A

Working Procedures of iIS and iWAIC

A.1 Working procedure of iIS

1. Generate MCMC samples $\{(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{1:n}^{(s)}); \text{s}= 1,\ldots, \text{S}\}$ from $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$

2. For each $s = 1, \ldots, S$

   (a) for each $i = 1, \ldots, n$, generate $\{\boldsymbol{b}_i^{(s,r)}; r = 1, \ldots, R\}$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}^{(s)}, \boldsymbol{\theta}^{(s)})$, and estimate $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})$ by

$$\hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = (1/R)\sum_{r=1}^{R} P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}, \boldsymbol{b}_i^{(s,r)}). \tag{A.1}$$

   Then, we can find iIS weight:

$$W_i^{\text{iIS}}(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = \frac{1}{\hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})} \tag{A.2}$$

   (b) For each $i = 1, \ldots, n$, generate $\{\tilde{\boldsymbol{b}}_i^{(s,k)}; k = 1, \ldots, K\}$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}^{(s)}, \boldsymbol{\theta}^{(s)})$, and estimate integrated evaluation function A by

$$A(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = (1/K)\sum_{k=1}^{K} a\left(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}^{(s)}, \tilde{\boldsymbol{b}}_i^{(s,k)}\right) \tag{A.3}$$

   (c) Estimate expected evaluation function a with respect to $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{-i}^{\text{obs}})$ by

$$\hat{E}_{\text{post(-i)}}^{\text{iIS}}(a(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}, \boldsymbol{b}_i)) = \frac{(1/S)\sum_{s=1}^{S}\left[A(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})W_i^{\text{iIS}}(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})\right]}{(1/S)\sum_{s=1}^{S} W_i^{\text{iIS}}(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})}. \tag{A.4}$$

Note that, if we are only interested in computing CVIC, don't need to do step 2(b), and take the numerator in (A.4) to be 1 as warranted by theory.

A.2 Working procedure of iWAIC

1. Generate MCMC sampels $\{(\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{1:n}^{(s)});$ s= 2,. . . ,S$\}$ from $P(\boldsymbol{\theta}, \boldsymbol{b}_{1:n}|\boldsymbol{y}_{1:n}^{\text{obs}})$

2. For each $s = 1, \dots, S$

   (a) For each $i = 1, \dots, n$, generate $\{\boldsymbol{b}_i^{(s,r)}; r = 1, \dots, R\}$ from $P(\boldsymbol{b}_i|\boldsymbol{b}_{-i}^{(s)}, \boldsymbol{\theta}^{(s)})$, and estimate integrated predictive density $P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{b}_{-i})$ by

   $$\hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}) = (1/R) \sum_{r=1}^{R} P(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)}, \boldsymbol{b}_i^{(s,r)}). \tag{A.5}$$

   (b) Estimate log CV posterior predictive density:

   $$\log(\hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})) = \log((1/S) \sum_{s=1}^{S} \hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})) - V_{s=1}^{S} \log(\hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}^{(s)}, \boldsymbol{b}_{-i}^{(s)})), \tag{A.6}$$

   where $V_{s=1}^{S} a^{(s)}$ stands for sample variance of $(a^{(1)}, \dots, a^{(S)})$.

3. Find iWAIC:

   $$\text{iWAIC} = -2 \sum_{i=1}^{n} \log(\hat{P}(\hat{P}(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}))) \tag{A.7}$$

Appendix B

Probabilities of p-values Table and Posterior inference Table

Table B.1: The results of quantities of three probabilities: $Pr(y_i > y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$, $Pr(y_i = y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$ and p-value($y_i^{\text{obs}}|\boldsymbol{y}_i^{\text{obs}}$), which are means of $a_0()$, a_1 and $a_2()$ with respect to posterior distribution of $(\boldsymbol{\theta}, s_i)$ for different methods under selected districts.

ID	$Pr(y_i > \boldsymbol{y}_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$					$Pr(y_i = y_i^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}})$					p-value($y^{\text{obs}}|\boldsymbol{y}_{-i}^{\text{obs}}$)				
	CV	POH	GHO	nIS	iIS	CV	PCH	GHO	nIS	iIS	CV	PCH	GHO	nIS	iIS
1	0.29	0.37	0.30	0.28	0.29	0.03	0.09	0.03	0.05	0.03	0.31	0.42	0.32	0.30	0.31
2	0.03	0.30	0.05	0.03	0.03	0.00	0.04	0.01	0.00	0.00	0.03	0.32	0.05	0.03	0.03
3	0.08	0.29	0.09	0.10	0.08	0.02	0.08	0.02	0.02	0.02	0.09	0.33	0.10	0.12	0.09
4	0.39	0.39	0.40	0.37	0.39	0.04	0.10	0.05	0.06	0.04	0.42	0.44	0.42	0.40	0.42
5	0.14	0.32	0.14	0.11	0.13	0.02	0.07	0.02	0.02	0.02	0.15	0.36	0.15	0.12	0.14
6	0.49	0.41	0.49	0.39	0.49	0.05	0.10	0.05	0.06	0.05	0.51	0.46	0.51	0.42	0.52
7	0.05	0.29	0.07	0.08	0.05	0.01	0.05	0.01	0.01	0.01	0.06	0.31	0.07	0.09	0.06
8	0.10	0.27	0.10	0.10	0.10	0.03	0.10	0.03	0.03	0.03	0.11	0.32	0.11	0.11	0.11
9	0.25	0.33	0.26	0.25	0.25	0.05	0.11	0.05	0.06	0.05	0.27	0.39	0.28	0.28	0.27
10	0.25	0.37	0.26	0.23	0.25	0.03	0.06	0.03	0.03	0.03	0.26	0.40	0.27	0.25	0.27
11	0.12	0.30	0.12	0.10	0.11	0.02	0.07	0.02	0.02	0.02	0.13	0.34	0.13	0.11	0.12
12	0.48	0.40	0.48	0.44	0.48	0.07	0.13	0.07	0.08	0.07	0.51	0.46	0.52	0.48	0.52
13	0.45	0.35	0.44	0.37	0.44	0.09	0.16	0.09	0.10	0.09	0.49	0.43	0.48	0.42	0.48
14	0.44	0.40	0.44	0.40	0.45	0.06	0.10	0.06	0.06	0.06	0.47	0.45	0.47	0.43	0.47
15	0.06	0.24	0.06	0.07	0.05	0.01	0.06	0.01	0.02	0.01	0.06	0.27	0.07	0.07	0.06
16	0.55	0.44	0.55	0.52	0.55	0.06	0.10	0.06	0.06	0.06	0.58	0.49	0.58	0.55	0.58
17	0.55	0.36	0.54	0.46	0.55	0.11	0.20	0.12	0.14	0.11	0.60	0.47	0.60	0.53	0.61
18	0.12	0.25	0.12	0.12	0.11	0.04	0.10	0.04	0.04	0.04	0.14	0.30	0.15	0.14	0.14
19	0.34	0.37	0.35	0.36	0.34	0.06	0.10	0.06	0.06	0.06	0.37	0.42	0.38	0.39	0.37
20	0.24	0.32	0.24	0.19	0.24	0.06	0.11	0.06	0.05	0.06	0.27	0.37	0.27	0.21	0.27
21	0.11	0.27	0.12	0.13	0.11	0.03	0.07	0.03	0.03	0.03	0.13	0.30	0.14	0.14	0.13
22	0.72	0.55	0.68	0.73	0.73	0.03	0.05	0.03	0.03	0.03	0.74	0.57	0.69	0.74	0.74
23	0.35	0.38	0.36	0.35	0.35	0.06	0.09	0.06	0.06	0.06	0.38	0.42	0.39	0.38	0.38
24	0.09	0.23	0.11	0.09	0.09	0.04	0.10	0.04	0.04	0.04	0.11	0.28	0.14	0.11	0.11
25	0.07	0.23	0.09	0.04	0.06	0.02	0.06	0.02	0.01	0.02	0.08	0.26	0.10	0.05	0.07
26	0.04	0.20	0.05	0.05	0.04	0.01	0.06	0.02	0.02	0.01	0.05	0.22	0.06	0.05	0.05
27	0.21	0.29	0.21	0.21	0.21	0.07	0.11	0.07	0.07	0.07	0.24	0.35	0.25	0.24	0.24

Table B.1 – *Continued from previous page*

ID	$Pr(y_i > y_i^{\text{obs}} \vert \boldsymbol{y}_{-i}^{\text{obs}})$					$Pr(y_i = y_i^{\text{obs}} \vert \boldsymbol{y}_{-i}^{\text{obs}})$					p-value$(y_i^{\text{obs}} \vert \boldsymbol{y}_{-i}^{\text{obs}})$				
	CV	PCH	CHO	nIS	iIS	CV	PCH	GHO	nIS	iIS	CV	PCH	GHO	nIS	iIS
28	0.28	0.34	0.28	0.27	0.28	0.06	0.09	0.06	0.06	0.06	0.31	0.39	0.31	0.30	0.31
29	0.64	0.51	0.63	0.58	0.64	0.05	0.08	0.05	0.05	0.05	0.66	0.55	0.65	0.60	0.67
30	0.23	0.32	0.25	0.23	0.23	0.05	0.09	0.06	0.05	0.05	0.25	0.36	0.28	0.26	0.26
31	0.23	0.29	0.24	0.21	0.23	0.09	0.13	0.09	0.08	0.09	0.27	0.36	0.28	0.25	0.27
32	0.78	0.51	0.76	0.77	0.78	0.07	0.17	0.08	0.08	0.07	0.82	0.60	0.80	0.81	0.82
33	0.42	0.40	0.42	0.41	0.42	0.08	0.12	0.09	0.09	0.09	0.46	0.45	0.46	0.46	0.46
34	0.16	0.27	0.18	0.16	0.16	0.06	0.10	0.06	0.06	0.06	0.19	0.31	0.21	0.19	0.19
35	0.34	0.37	0.34	0.33	0.33	0.07	0.09	0.07	0.06	0.07	0.37	0.42	0.37	0.36	0.37
36	0.13	0.24	0.14	0.13	0.13	0.05	0.09	0.05	0.05	0.05	0.15	0.29	0.16	0.16	0.15
37	0.56	0.48	0.56	0.54	0.57	0.07	0.09	0.07	0.07	0.07	0.60	0.52	0.59	0.58	0.60
38	0.06	0.18	0.07	0.05	0.06	0.03	0.08	0.04	0.03	0.03	0.07	0.22	0.09	0.07	0.07
39	0.79	0.57	0.76	0.75	0.79	0.06	0.12	0.06	0.07	0.06	0.82	0.63	0.80	0.78	0.82
40	0.13	0.22	0.14	0.13	0.13	0.09	0.13	0.09	0.08	0.09	0.18	0.28	0.19	0.17	0.18
41	0.34	0.36	0.34	0.33	0.33	0.08	0.10	0.08	0.08	0.08	0.38	0.41	0.38	0.37	0.37
42	0.99	0.82	0.97	0.98	0.99	0.00	0.06	0.01	0.01	0.00	0.99	0.85	0.98	0.99	0.99
43	0.84	0.61	0.83	0.85	0.85	0.08	0.18	0.08	0.07	0.08	0.88	0.70	0.87	0.89	0.88
44	0.54	0.47	0.53	0.53	0.54	0.11	0.13	0.11	0.11	0.11	0.60	0.53	0.59	0.59	0.59
45	0.95	0.78	0.89	0.95	0.96	0.01	0.05	0.02	0.01	0.01	0.96	0.80	0.91	0.96	0.96
46	0.75	0.58	0.74	0.75	0.75	0.10	0.16	0.11	0.10	0.10	0.80	0.66	0.79	0.80	0.80
47	0.41	0.36	0.40	0.41	0.41	0.20	0.23	0.20	0.20	0.20	0.51	0.47	0.50	0.51	0.51
48	0.62	0.51	0.61	0.61	0.62	0.14	0.18	0.14	0.14	0.14	0.69	0.60	0.68	0.69	0.69
49	0.99	0.85	0.94	0.98	0.99	0.00	0.03	0.01	0.00	0.00	0.99	0.87	0.95	0.99	0.99
50	0.94	0.78	0.91	0.93	0.94	0.02	0.08	0.03	0.03	0.02	0.96	0.82	0.93	0.95	0.96
51	0.46	0.37	0.45	0.45	0.46	0.26	0.31	0.27	0.27	0.26	0.59	0.52	0.59	0.58	0.59
52	0.44	0.36	0.43	0.44	0.44	0.27	0.31	0.27	0.27	0.27	0.57	0.51	0.57	0.57	0.57
53	0.65	0.52	0.64	0.67	0.65	0.21	0.28	0.21	0.20	0.21	0.76	0.66	0.75	0.77	0.76
54	0.77	0.62	0.76	0.76	0.77	0.15	0.24	0.16	0.16	0.15	0.84	0.74	0.84	0.84	0.85
55	0.98	0.85	0.97	0.97	0.98	0.02	0.15	0.03	0.03	0.02	0.99	0.92	0.99	0.99	0.99
56	0.68	0.45	0.67	0.64	0.68	0.32	0.55	0.33	0.36	0.32	0.84	0.73	0.83	0.82	0.84

Abbreviations: CV: cross validation, PCH: posterior checking, GHO: Ghosting, nIS: non-integrated Importance Sampling, iIS: integrated Importance Sampling

Table B.2: Posterior inference for $\boldsymbol{\lambda}$ for the spatial + linear model of Scotland lip cancer data.

Parameter	Mean	2.5%	Median	97.5%	Parameter	Mean	2.5%	Median	97.5%
λ_1	6.26	2.96	6.04	10.85	λ_{29}	1.19	0.76	1.17	1.72
λ_2	4.10	2.92	4.06	5.50	λ_{30}	0.98	0.55	0.95	1.55
λ_3	3.12	1.57	3.02	5.24	λ_{31}	0.92	0.38	0.87	1.76
λ_4	3.48	1.71	3.36	5.97	λ_{32}	1.40	0.49	1.30	2.84
λ_5	3.19	1.84	3.10	5.00	λ_{33}	0.99	0.49	0.96	1.69
λ_6	3.37	1.57	3.24	5.90	λ_{34}	0.79	0.39	0.76	1.36
λ_7	2.85	1.87	2.81	4.04	λ_{35}	0.85	0.50	0.83	1.32
λ_8	2.52	1.03	2.40	4.74	λ_{36}	0.73	0.39	0.70	1.20

Table B.2 – *Continued from previous page*

Parameter	Mean	2.5%	Median	97.5%	Parameter	Mean	2.5%	Median	97.5%
λ_9	2.78	1.10	2.62	5.28	λ_{37}	0.92	0.55	0.90	1.39
λ_{10}	2.87	1.81	2.82	4.18	λ_{38}	0.62	0.31	0.59	1.09
λ_{11}	2.61	1.42	2.55	4.15	λ_{39}	1.04	0.53	1.01	1.77
λ_{12}	2.86	1.07	2.70	5.55	λ_{40}	0.57	0.22	0.53	1.20
λ_{13}	2.82	0.73	2.55	6.42	λ_{41}	0.51	0.30	0.49	0.77
λ_{14}	2.41	1.16	2.32	4.21	λ_{42}	0.82	0.48	0.80	1.23
λ_{15}	1.83	1.09	1.79	2.80	λ_{43}	0.80	0.29	0.75	1.61
λ_{16}	2.07	1.06	2.01	3.44	λ_{44}	0.45	0.25	0.44	0.73
λ_{17}	2.12	0.42	1.85	5.32	λ_{45}	0.48	0.34	0.48	0.65
λ_{18}	1.38	0.60	1.31	2.52	λ_{46}	0.53	0.24	0.50	0.96
λ_{19}	1.56	0.79	1.51	2.61	λ_{47}	0.39	0.13	0.36	0.83
λ_{20}	1.42	0.65	1.36	2.54	λ_{48}	0.41	0.19	0.39	0.75
λ_{21}	1.33	0.80	1.30	2.00	λ_{49}	0.41	0.31	0.41	0.52
λ_{22}	1.45	1.04	1.44	1.93	λ_{50}	0.49	0.28	0.48	0.75
λ_{23}	1.21	0.69	1.18	1.90	λ_{51}	0.39	0.10	0.34	0.97
λ_{24}	0.97	0.43	0.92	1.79	λ_{52}	0.37	0.09	0.32	0.91
λ_{25}	1.03	0.66	1.01	1.51	λ_{53}	0.33	0.11	0.30	0.71
λ_{26}	0.94	0.56	0.92	1.46	λ_{54}	0.33	0.12	0.31	0.67
λ_{27}	1.01	0.48	0.97	1.80	λ_{55}	0.58	0.16	0.53	1.27
λ_{28}	1.03	0.57	1.00	1.67	λ_{56}	0.42	0.05	0.33	1.33
α	-0.57	-0.89	-0.57	-0.23	β	6.31	3.57	6.30	9.17
ϕ	0.14	0.02	0.15	0.17	τ	1.40	0.99	1.39	1.90

Appendix C

R Code for CV

C.1 Utility Function

```
###################################
## Three functions to calculate mean and harmonic mean
## of exponential of log functions
###################################
## log_sum_exp --- A function to calculate the value of log of
##                 sum over exponential of log function.
## log_mean_exp --- A function to calculate the value of log of
##                  mean of exponential of log function.
## log_hmean_exp --- A function to calculate the value of log of
##                   harmonic mean of exponential of log function.

log_sum_exp <- function (lx)
{
    mlx <- max (lx)
    log (sum (exp (lx - mlx))) + mlx
}
log_mean_exp <- function (lx)
{
    log_sum_exp (lx) - log(length (lx))
}
log_hmean_exp <- function (lx)
{
    - log_mean_exp (-lx)
}
```

C.2 MCMC Simulation for CV Posterior

```
## load R package "R2OpenBUGS"
library("R2OpenBUGS")

##################################
## A function to specify initial values for MCMC simulation.
## Initials are for starting points of  prior parameters
```

```
##################################
## alpha --- A latent variable for linear effect representing intercept.
## prec ---  (inverse variance) A scalar parameter representing
##             the overall precision parameter.
## beta ---  a latent variable  for linear effect with covariate X.



inits <- function()
    {
     list(alpha = 0, prec = 0.5, gamma =0,

       S=rnorm(56,0,20)
    )

     }

#########################################
## Load data for MCMC simulation
########################################
## N --- number of districts
## sumNumNeigh --- summation of number of neighbourhood for each district
## O --- number of observation cases for each district
## E --- expected number of cases for each district
## X --- covariate, percent of population employed in
##       agriculture, fishing and forestry
## adj --- adjacent ID of each district
## num --- number of adjacent ID of each district


raw_data <- list(N = 56,

        sumNumNeigh = 264,

  O  = c(    9,   39,   11,    9,   15,    8,   26,    7,    6,   20,
                   13,    5,    3,    8,   17,    9,    2,    7,    9,    7,
                   16,   31,   11,    7,   19,   15,    7,   10,   16,   11,
                    5,    3,    7,    8,   11,    9,   11,    8,    6,    4,
                   10,    8,    2,    6,   19,    3,    2,    3,   28,    6,
                    1,    1,    1,    1,    0,    0),
        E = c( 1.38, 8.66, 3.04, 2.53, 4.26, 2.40, 8.11, 2.30, 1.98, 6.63,
               4.40, 1.79, 1.08, 3.31, 7.84, 4.55, 1.07, 4.18, 5.53, 4.44,
              10.46,22.67, 8.77, 5.62,15.47,12.49, 6.04, 8.96,14.37,10.20,
               4.75, 2.88, 7.03, 8.53,12.32,10.10,12.68, 9.35, 7.20, 5.27,
              18.76,15.78, 4.32,14.63,50.72, 8.20, 5.59, 9.34,88.66,19.62,
               3.44, 3.62, 5.74, 7.03, 4.16, 1.76),

       X= c(16,16,10,24,10,24,10, 7, 7,16,
               7,16,10,24, 7,16,10, 7, 7,10,
               7,16,10, 7, 1, 1, 7, 7,10,10,
               7,24,10, 7, 7, 0,10, 1,16, 0,
               1,16,16, 0, 1, 7, 1, 1, 0, 1,
               1, 0, 1, 1,16,10),

        adj = c( 5, 9,11,19,
                 7,10,
                 6,12,
                 18,20,28,
                  1,11,12,13,19,
                  3, 8,
                  2,10,13,16,17,
                  6,
                  1,11,17,19,23,29,
                  2, 7,16,22,
                  1, 5, 9,12,
                  3, 5,11,
                  5, 7,17,19,
                 31,32,35,
                 25,29,50,
                  7,10,17,21,22,29,
                  7, 9,13,16,19,29,
                  4,20,28,33,55,56,
                  1, 5, 9,13,17,
                  4,18,55,
                 16,29,50,
                 10,16,
                  9,29,34,36,37,39,
                 27,30,31,44,47,48,55,56,
                 15,26,29,
                 25,29,42,43,
                 24,31,32,55,
                  4,18,33,45,
                  9,15,16,17,21,23,25,26,34,43,50,
                 24,38,42,44,45,56,
                 14,24,27,32,35,46,47,
                 14,27,31,35,
                 18,28,45,56,
                 23,29,39,40,42,43,51,52,54,
                 14,31,32,37,46,
                 23,37,39,41,
                 23,35,36,41,46,
                 30,42,44,49,51,54,
                 23,34,36,40,41,
                 34,39,41,49,52,
                 36,37,39,40,46,49,53,
                 26,30,34,38,43,51,
                 26,29,34,42,
                 24,30,38,48,49,
                 28,30,33,56,
                 31,35,37,41,47,53,
                 24,31,46,48,49,53,
                 24,44,47,49,
                 38,40,41,44,47,48,52,53,54,
                 15,21,29,
                 34,38,42,54,
                 34,40,49,54,
                 41,46,47,49,
                 34,38,49,51,52,
                 18,20,24,27,56,
                 18,24,30,33,45,55),

         num = c(4, 2, 2, 3, 5, 2, 5, 1,  6,
                 4, 4, 3, 4, 3, 3, 6, 6, 6 ,5,
                 3, 3, 2, 6, 8, 3, 4, 4, 4,11,
                 6, 7, 4, 4, 9, 5, 4, 5, 6, 5,
                 5, 7, 6, 4, 5, 4, 6, 6, 4, 9,
                 3, 4, 4, 4, 5, 5, 6)
     )

#####################################
## Identify latent variables and  parameters to be MCMC sampled.
#######################################
## S --- latent variable, the logarithm of relative risk


parameters<-c("alpha","S","sigma","prec")

##################################
## Set NA to validity unit of observation
##################################
## ifold --- ifold in CV process is to identify validity unit
##           that is, e.g. ifold = 2 means that we omit the observation of
##           y for district 2 (set NA).

if (!exists ("ifold")) ifold <- 2
data <- raw_data
O_t<-data$O
E_t<-data$E
Coef<-rep(0,length(O_t))
O.ts <- data$O[ifold]
data$O[ifold] <- NA
Coef[ifold]<- sum(O_t[-ifold])/sum(E_t[-ifold])
data$E <- data$E*Coef[ifold]
```

```
#########################################
## Define the full model
## One may also define alternative model in this way.
## The model description is saved in model file. 'lipcancer_prop_full.txt'
##########################################

'model {


    for(i in 1 : N) {
        m[i] <- 1/E[i]
        }

    cumsum[1] <- 0

    for(i in 2:(N+1)) {
       cumsum[i] <- sum(num[1:(i-1)])
       }

    for(k in 1 : sumNumNeigh) {

       for(i in 1:N) {
           pick[k,i] <- step(k - cumsum[i] - epsilon) * step(cumsum[i+1] - k)
                      }
           C[k] <- sqrt(E[adj[k]] / inprod(E[], pick[k,]))
       }

    epsilon <- 0.0001


   for (i in 1 : N) {
       O[i] ~ dpois(mu[i])
       log(mu[i]) <- log(E[i]) + S[i]
       RR[i] <- exp(S[i])
       theta[i] <- alpha + beta*X[i]/100
       }
   # Proper CAR prior distribution for spatial random effects:

   S[1:N] ~ car.proper(theta[], C[], adj[], num[], m[], prec, gamma)
   # Other priors:
   alpha ~ dnorm(0, 0.0001)
   beta ~ dnorm(0, 0.001)
   prec ~ dgamma(0.5, 0.0005)
   v <- 1/prec
   sigma <- sqrt(1 / prec)
   # prior on precision
   # variance
   # standard deviation
   gamma.min <- min.bound(C[], adj[], num[], m[])
   gamma.max <- max.bound(C[], adj[], num[], m[])
   gamma ~ dunif(gamma.min, gamma.max)

}'


##########################################
## A function from R2OpenBUGS package to do MCMC sample simulation with
## arguments data, initial values, sampled parameters, iterations number,
## model file, number of chains, number of thin iteration,
## number of burning iteration,
## bugs.seed and return of DIC.
###########################################


fit<-bugs(data,inits,parameters, n.iter = 15000,
model.file='/home/shq471/lipcancer/models/lipcancer_prop_full.txt',
n.chains=2, n.thin=1, n.burnin = 5000, DIC=FALSE)

##################################
## save MCMC sample
##################################

mcmc<-fit$sims.matrix
alpha_hat<-mcmc[,"alpha"]
s_hat<-mcmc[,sprintf("S[%d]",ifold)]
mu_hat <- exp(s_hat)*data$E[ifold]
```

C.3 CVIC and CV p-value

```
#######################################
## calculate CVIC for district 'ifold' and save it
######################################

log_prob_CV <-dpois(O.ts, mu_hat,log=TRUE)
CVIS_sub <- -2*(log_mean_exp(log_prob_CV))
cat(CVIS_sub,file =sprintf("/lipcancer_CV_full_E%d.txt",ifold))

###########################################
## calculate CV p-value for district 'ifold' and save it
#########################################

log_pv <- log_sum_exp(c( log_mean_exp(ppois(O.ts, mu_hat,log=TRUE,
lower.tail=FALSE)),log_mean_exp(dpois(O.ts,mu_hat,log=TRUE))-log(2) ))
p_value <- exp(log_pv)
```

```
cat(p_value,file =sprintf("/nPvcv%d.txt",ifold))
```

Appendix D

R Code for iIS and iWAIC

D.1 MCMC Simulation for Full Data Posterior

```
## load R package "R2OpenBUGS"

library("R2OpenBUGS")


##########################################
## A function to specify initial values for MCMC simulation.
## Initials are prior parameters
#########################################
## alpha --- A latent variable for linear effect representing intercept.
## prec ---  (inverse variance) A scalar parameter representing
##           the overall precision parameter.
## beta ---  a latent variable  for linear effect with covariate X.

inits <- function()
    {
     list(alpha = rnorm(1,0,2), prec = runif(1,0.1,2), gamma =0,
           S = rep(0,56), beta=rnorm(1,5,5))
     }



##############################################
## Load data for MCMC simulation
###########################################
## N --- number of districts
## sumNumNeigh --- summation of number of neighbourhood for each district
## O --- number of observation cases for each district
## E --- expected number of cases for each district
## X --- covariate, percent of population employed in agriculture, fishing and forestry
## adj --- adjacent ID of each district
## num --- number of adjacent ID of each district
```

```
data <- list(N = 56,

        sumNumNeigh = 264,

  O  = c(   9,   39,   11,    9,   15,    8,   26,    7,    6,   20,
             13,    5,    3,    8,   17,    9,    2,    7,    9,    7,
             16,   31,   11,    7,   19,   15,    7,   10,   16,   11,
              5,    3,    7,    8,   11,    9,   11,    8,    6,    4,
             10,    8,    2,    6,   19,    3,    2,    3,   28,    6,
              1,    1,    1,    1,    0,    0),

        E = c( 1.38, 8.66, 3.04, 2.53, 4.26, 2.40, 8.11, 2.30, 1.98, 6.63,
               4.40, 1.79, 1.08, 3.31, 7.84, 4.55, 1.07, 4.18, 5.53, 4.44,
              10.46,22.67, 8.77, 5.62,15.47,12.49, 6.04, 8.96,14.37,10.20,
               4.75, 2.88, 7.03, 8.53,12.32,10.10,12.68, 9.35, 7.20, 5.27,
              18.76,15.78, 4.32,14.63,50.72, 8.20, 5.59, 9.34,88.66,19.62,
               3.44, 3.62, 5.74, 7.03, 4.16, 1.76),

           X= c(16,16,10,24,10,24,10, 7, 7,16,
               7,16,10,24, 7,16,10, 7, 7,10,
               7,16,10, 7, 1, 1, 7, 7,10,10,
               7,24,10, 7, 7, 0,10, 1,16, 0,
               1,16,16, 0, 1, 7, 1, 1, 0, 1,
               1, 0, 1, 1,16,10),

        adj = c( 5, 9,11,19,
                 7,10,
                 6,12,
                 18,20,28,
                  1,11,12,13,19,
                  3, 8,
                  2,10,13,16,17,
                  6,
                  1,11,17,19,23,29,
                  2, 7,16,22,
                  1, 5, 9,12,
                  3, 5,11,
                  5, 7,17,19,
                 31,32,35,
                 25,29,50,
                  7,10,17,21,22,29,
                  7, 9,13,16,19,29,
                  4,20,28,33,55,56,
                  1, 5, 9,13,17,
                  4,18,55,
                 16,29,50,
                 10,16,
                  9,29,34,36,37,39,
                 27,30,31,44,47,48,55,56,
                 15,26,29,
                 25,29,42,43,
                 24,31,32,55,
                  4,18,33,45,
                  9,15,16,17,21,23,25,26,34,43,50,
                 24,38,42,44,45,56,
                 14,24,27,32,35,46,47,
                 14,27,31,35,
                 18,28,45,56,
                 23,29,39,40,42,43,51,52,54,
                 14,31,32,37,46,
                 23,37,39,41,
                 23,35,36,41,46,
                 30,42,44,49,51,54,
                 23,34,36,40,41,
                 34,39,41,49,52,
                 36,37,39,40,46,49,53,
                 26,30,34,38,43,51,
                 26,29,34,42,
                 24,30,38,48,49,
                 28,30,33,56,
                 31,35,37,41,47,53,
                 24,31,46,48,49,53,
                 24,44,47,49,
                 38,40,41,44,47,48,52,53,54,
                 15,21,29,
                 34,38,42,54,
                 34,40,49,54,
                 41,46,47,49,
                 34,38,49,51,52,
                 18,20,24,27,56,
                 18,24,30,33,45,55),

         num = c(4, 2, 2, 3, 5, 2, 5, 1,  6,
                 4, 4, 3, 4, 3, 3, 6, 6, 6 ,5,
                 3, 3, 2, 6, 8, 3, 4, 4, 4,11,
                 6, 7, 4, 4, 9, 5, 4, 5, 6, 5,
                 5, 7, 6, 4, 5, 4, 6, 6, 4, 9,
                 3, 4, 4, 4, 5, 5, 6)
     )


########################################
## Identify latent variables and  parameters to be MCMC sampled.
#######################################
## S --- latent variable, the logarithm of relative risk
```

```
parameters<-c("alpha","S","sigma","prec","gamma" ,"beta","theta")



##############################################
## nIter --- number of iterations for each chain of MCMC simulation
## nBur --- number of iterations for burning
## Sims --- total number of iterations sampled in two chains
nIter <- 15000
nBur  <- 5000
Sims  <- 2*(nIter - nBur)



#######################################
## A function from R2OpenBUGS package to do MCMC sample simulation with
## arguments data, initial values, sampled parameters, iterations number,
## model file, number of chains, number of thin iteration, number of
## burning iteration, bugs.seed and return of DIC.
#######################################

fit<-bugs(data,inits,parameters, n.iter = nIter, model.file='lipcancer_prop_full.txt',
n.chains=2, n.thin=1, n.burnin = nBur, DIC=TRUE , bugs.seed = sample(1:14,1))

####################################
## Save MCMC sample posteriors
####################################

mcmc<-fit$sims.matrix
alpha_hat<-mcmc[,"alpha"]
beta_hat<-mcmc[,"beta"]
gamma_hat<-mcmc[,"gamma"]
prec_hat<-mcmc[,"prec"]
sigma_hat<-mcmc[,"sigma"]
S_hat<-mcmc[,sprintf("S[%d]",c(1:56))]
theta_hat<-mcmc[,sprintf("theta[%d]",c(1:56))]


#################################
## Save data to another name
################################
adj<-data$adj
num<-data$num
E<-data$E
X<-data$X
logE<- log(data$E)
O.data<-data$O
```

```
########################################
## C --- a vector the same length as adj[] giving normalised weights
##       associated with each pair of districts

C <- table(adj,rep(1:56,num))
for(i in 1:56){
   for(j in 1:56){
    C[i,j]<-sqrt(C[i,j]*E[j]/E[i])
    }
}
########################################
## R --- a scalar that indicates volume of integral of evaluation for iIS and iWAIC

R <- 200
```

D.2 Approximating CVIC with DIC, nWAIC, nIS and iIS

```
###########################################
## Define MCMC sample size and integrated sample size
#########################################
log_prob_hat <- rep(0,56)
log_prob_rep <- rep(0,56)
log_P_matrix_hat<-matrix(,nrow=Sims,ncol=56,byrow=TRUE)
log_P_matrix_rep<-matrix(,nrow=Sims,ncol=56,byrow=TRUE)


##########################################
## Following code are details of calculation of iWAIC, nWAIC, nIS and iIS
## information criterion. We operate on full data posteriors through MCMC
## samples. Therefore, we execute MCMC sample with loop for each of the
## simulations
##########################################

for(k in 1:Sims){

    for( i in 1:56 ) {

     O.ts<- O.data[i]


## the mean of the Gaussian distribution that S belongs to

     mu_S <- alpha_hat[k] + beta_hat[k]*X[i]/100 +
             gamma_hat[k]*sum(C[i,]*(S_hat[k,]-alpha_hat[k]-beta_hat[k]*X/100))
## the sigma of the Gaussian distribution


    sigma_S <- 1/sqrt(prec_hat[k]*E[i])


## work process of integrated predictive density of observation y_i :
## First, to regenerate sample of latent variable S_i from
## conditional distribution of S_-i.
## Second, plug in the regenerated mean into predictive Poisson density of y_i.

     S_rep <- rnorm(R, mu_S,sigma_S)
     log_mu_rep <- S_rep + logE[i]
     log_prob_raw <- dpois(O.ts,exp(log_mu_rep),log=T)
     log_prob_rep[i] <- log_mean_exp(log_prob_raw)

## work process of non-integrated predictive density of observation y_i:

     log_mu_hat <-  S_hat[k,i] + logE[i]
     log_prob_hat[i]<-dpois(O.ts,exp(log_mu_hat),log=T)
                    }

    log_P_matrix_hat[k,] <- log_prob_hat
    log_P_matrix_rep[k,] <- log_prob_rep

    }

###########################################
## calculate iWAIC, nWAIC, iIS and nIS using integrated predictive density or
## non-integrated predictive density.
#######################################

logp_iwaic<-rep(0,56)
for(i in 1:56){
    logp_iwaic[i] <- log_mean_exp(log_P_matrix_rep[,i])

    - var(log_P_matrix_rep[,i])
    }
logp_iwaic_t <- -2*sum(logp_iwaic)

logp_nwaic<-rep(0,56)
for(i in 1:56) {
    logp_nwaic[i] <- log_mean_exp(log_P_matrix_hat[,i])

    - var(log_P_matrix_hat[,i])
    }
logp_nwaic_t <- -2*sum(logp_nwaic)

logp_iis<-rep(0,56)

for(i in 1:56){
    logp_iis[i] <- log_hmean_exp(log_P_matrix_rep[,i])
   }
logp_iis_t <- -2*sum(logp_iis)

logp_nis <- rep(0,56)

for(i in 1:56){
logp_nis[i] <- log_hmean_exp(log_P_matrix_hat[,i])
   }
logp_nis_t<- -2*sum(logp_nis)


#####################################
## return the value of DIC through function ’bugs’
######################################

DIC<-fit$DIC


###########################################
## Save results of information criterion into files
#############################################
## ifold --- a scaler, indicator of repeated number. We repeat the process
##           calculation for 100 times to test the reliability.

cat(logp_iwaic_t,file = sprintf("/full_IWAI%d.txt",ifold))

cat(logp_nwaic_t,file = sprintf("/full_NWAIC%d.txt",ifold))

cat(logp_iis_t,file = sprintf("/full_IIS%d.txt",ifold))

cat(logp_nis_t,file = sprintf("/full_NIS%d.txt",ifold))

cat(DIC, file = sprintf("//full_DIC%d.txt",ifold))
```

D.3 Approximating CV p-value with Posterior checking, Ghosting, nIS and iIS

```
###################################
## Define the size of p-values
```

```
###################################
p_nis <- p_post <- rep(0,56)
p_ghost <- p_iis <- rep(0,56)

#############################################
## Define two vectors for integrated weight and integrated p-value
#############################################

log_weight_rep <- rep(0,Sims)
log_p_value_rep <- rep(0,Sims)


##########################################
## Following code are details of calculation of posterior checking,
## Ghosting, nIS and iIS methods for p-values.
###########################################

for( i in 1:56 )
{

     O.ts<- O.data[i]

     for(k in 1:Sims) {


## the mean of the Gaussian distribution that S belongs to

mu_S <- alpha_hat[k] + beta_hat[k]*X[i]/100 +
gamma_hat[k]*sum(C[i,]*(S_hat[k,]-alpha_hat[k]- beta_hat[k]*X/100))


## the standard error of the Gaussian distribution
         sigma_S <- 1/sqrt(prec_hat[k]*E[i])

## the integrated weighting
         s_rep_w <-rnorm(100,mu_S,sigma_S)

## the integrated evaluation function

         S_rep <- rnorm(10, mu_S,sigma_S)

         log_mu_rep <- S_rep + logE[i]

         log_mu_rep_w<- s_rep_w + logE[i]

## the log of p-value with integrated evaluation function

         log_average_raw <- log_mean_exp(dpois(O.ts,exp(log_mu_rep_w),log=T))
         log_weight_rep[k] <- - log_average_raw

         log_p_value_rep[k] <- log_sum_exp(c(log_mean_exp(

         ppois(O.ts,exp(log_mu_rep),
         lower.tail=FALSE,log.p=TRUE)
         ),

        log_mean_exp(dpois(O.ts,exp(log_mu_rep),log=TRUE))

        - log(2)

                                        ))
                        }

## the log of p-value with non-integrated evaluation function

     log_mu_hat <- S_hat[,i] + logE[i]
     log_prob_hat <- dpois(O.ts,exp(log_mu_hat),log=T)
     log_weight_hat<- -log_prob_hat
     log_p_value_hat<- log(ppois(O.ts,exp(log_mu_hat),lower.tail=FALSE) +
                                     0.5*dpois(O.ts,exp(log_mu_hat)))

## the p-value of each methods

     p_ghost[i] <- exp(log_mean_exp(log_p_value_rep))

     p_iis[i] <- exp(log_mean_exp(log_weight_rep + log_p_value_rep)-
     log_mean_exp(log_weight_rep))

     p_nis[i] <-exp(log_mean_exp(log_weight_hat + log_p_value_hat) -
     log_mean_exp(log_weight_hat))

     p_post[i] <- exp(log_mean_exp(log_p_value_hat))

}

############################################
## ifold --- the indicator of number of replicates
##############################################

cat(p_ghost,file = sprintf("/pvghs%d.txt", ifold))
cat(p_iis, file = sprintf("/pviis%d.txt",ifold))
cat(p_nis,file = sprintf("/pvnis%d.txt",ifold))
cat(p_post,file = sprintf("/pvpch%d.txt",ifold))
```