Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA

Leonhard Held, Birgit Schrödle and Håvard Rue

Abstract Model criticism and comparison of Bayesian hierarchical models is often based on posterior or leave-one-out cross-validatory predictive checks. Cross-validatory checks are usually preferred because posterior predictive checks are difficult to assess and tend to be too conservative. However, techniques for statistical inference in such models often try to avoid full (manual) leave-one-out cross-validation, since it is very time-consuming. In this paper we will compare two approaches for estimating Bayesian hierarchical models: Markov chain Monte Carlo (MCMC) and integrated nested Laplace approximations (INLA). We review how both approaches allow for the computation of leave-one-out cross-validatory checks without re-running the model for each observation in turn. We then empirically compare the two approaches in an extensive case study analysing the spatial distribution of bovine viral diarrhoe (BVD) among cows in Switzerland.

Key words: Bayesian hierarchical models; INLA; Leave-one-out cross-validation; MCMC; Posterior predictive model checks

1 Introduction

Bayesian hierarchical models are widely used in applied statistics. Inference is typically based on Markov chain Monte Carlo (MCMC), a computer-intensive simulation-based approach. However, integrated nested Laplace approximations

Leonhard Held and Birgit Schrödle
Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland,
e-mail: leonhard.held@ifspm.uzh.ch, birgit.schroedle@ifspm.uzh.ch

Håvard Rue
Department of Mathematical Science, Norwegian University of Science and Technology, Trondheim, Norway, e-mail: havard.rue@math.ntnu.no

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_6,

(INLA) are a promising alternative to inference via MCMC in latent Gaussian models (Rue et al. 2009). The methodology is particularly attractive if the latent Gaussian model is a Gaussian Markov random field (GMRF) (Rue & Held 2005). In contrast to empirical Bayes approaches (Fahrmeir et al. 2004), the INLA approach incorporates posterior uncertainty with respect to hyperparameters. Examples where INLA is applicable include generalized linear mixed models (Breslow & Clayton 1993), disease mapping (Besag et al. 1991) including ecological regression (Clayton & Bernardinelli 1992, Natário & Knorr-Held 2003), spatial and spatio-temporal GMRF models (Gössl et al. 2001), dynamic (generalized) linear models (Fahrmeir 1992) and structured additive regression (Fahrmeir & Lang 2001).

A particularly interesting feature of INLA is that it provides leave-one-out cross-validatory model checks without re-running the model for each observation in turn. In this paper we review the computation of the conditional predictive ordinate (CPO) and the probability integral transform (PIT) in INLA and compare it with computation of the corresponding quantities using MCMC. We also consider posterior predictive model checks based on the whole data as an alternative to cross-validation. Section 2 reviews INLA and gives a detailed description how cross-validatory model checks are computed with INLA. Section 3 describes how these quantities are computed with MCMC. An extensive case study using an example from spatial epidemiology is described in Section 4 to compare the two approaches. We close with some discussion in Section 5.

2 The INLA Approach

The following section reviews INLA as an approach for approximate Bayesian inference in latent Gaussian models and shows how posterior and cross-validatory predictive checks can be computed using INLA.

2.1 Parameter Estimation with INLA

Consider a three-stage Bayesian hierarchical model based on an observation model $\pi(y|x) = \prod_i \pi(y_i|x_i)$, a parameter model $\pi(x|\theta)$, and a hyperprior $\pi(\theta)$. Here $y = (y_1, \ldots, y_n)$ denotes the observed data, x are unknown parameters which typically follow a GMRF, and θ are unknown hyperparameters. Note that reparametrization and parameter augmentation can be used to achieve $\pi(y_i|x) = \pi(y_i|x_i)$. The dimension of x will often be larger than n and we assume in the following that only the first n components of x are directly linked to the observations y.

Consider now the marginal posterior density

$$\pi(x_i|y) = \int_\theta \pi(x_i|\theta, y)\, \pi(\theta|y)\, d\theta$$

of the i-th component x_i of x. INLA approximates this by

$$\widetilde{\pi}(x_i|y) = \sum_k \widetilde{\pi}(x_i|\theta_k, y)\ \widetilde{\pi}(\theta_k|y)\ \Delta_k$$

using an approximation $\widetilde{\pi}(x_i|\theta, y)$ of $\pi(x_i|\theta, y)$ and an additional approximation $\widetilde{\pi}(\theta|y)$ of the marginal posterior density $\pi(\theta|y)$ of the hyperparameters θ. The weights Δ_k are chosen appropriately.

We first describe how $\pi(\theta|y)$ is approximated. Clearly,

$$\pi(x, \theta, y) = \pi(x|\theta, y) \times \pi(\theta|y) \times \pi(y), \tag{1}$$

so it follows that

$$\pi(\theta|y) \propto \frac{\pi(x, \theta, y)}{\pi(x|\theta, y)} \text{ for all } x.$$

INLA approximates $\pi(\theta|y)$ using a Laplace approximation (Tierney & Kadane 1986):

$$\widetilde{\pi}(\theta|y) \propto \frac{\pi(x, \theta, y)}{\widetilde{\pi}_G(x|\theta, y)}|_{x=x^*(\theta)}.$$

The numerator can be easily evaluated based on (1). The denominator $\widetilde{\pi}_G(x|\theta, y)$ is the Gaussian approximation (Rue et al. 2009, Section 2.2) of $\pi(x|\theta, y)$ and $x^*(\theta)$ is the mode of the full conditional $\pi(x|\theta, y)$, obtained through a suitable iterative algorithm. The approximate posterior density $\widetilde{\pi}(\theta|y)$ is "numerically explored" to obtain suitable support points θ_k and the respective weights Δ_k.

For approximating the first component $\pi(x_i|\theta, y)$, a Gaussian approximation (Rue & Martino 2007), easily extractable from $\widetilde{\pi}_G(x|\theta, y)$,

$$\tilde{\pi}_G(x_i|\theta, y) = \mathrm{N}(x_i; \mu_i(\theta), \sigma_i^2(\theta))$$

can be used. The approximation can be improved using a Laplace approximation

$$\tilde{\pi}_{LA}(x_i|\theta, y) \propto \mathrm{N}(x_i; \mu_i(\theta), \sigma_i^2(\theta)) \times \exp(\text{cubic spline}(x_i)),$$

or a simplified Laplace approximation based on the skew-normal distribution (Azzalini & Capitano 1999), for details see Rue et al. (2009).

As suggested in Fahrmeir & Kneib (2009), it is instructive to compare the INLA approach with a REML/Empirical Bayes estimation in mixed models. In the empirical Bayes approach no hyperprior $\pi(\theta)$ is necessary, so the (RE)ML marginal likelihood corresponds to the marginal posterior $\pi(\theta|y)$. The (RE)ML marginal likelihood is maximized and only the (RE)ML estimate of θ is used, so no uncertainty with respect to θ is taken into account. The empirical Bayes estimate of x_i corresponds to the Gaussian approximation of $\pi(x_i|\theta, y)$ with θ fixed at the (RE)ML estimate. Hierarchical likelihood (Lee et al. 2006) is a variation of this.

2.2 Posterior Predictive Model Checks with INLA

In order to check the fit of a Bayesian model posterior predictive checks were proposed by Gelman et al. (1996). The underlying concept of such checks is the posterior predictive distribution of a replicate observation Y_i which has density

$$\pi(y_i|y) = \int \pi(y_i|x_i, y) \cdot \pi(x_i|y)\, dx_i. \tag{2}$$

In Stern & Cressie (2000) it is suggested to use the posterior predictive p-value

$$\text{Prob}(Y_i \leq y_i^{obs}|y)$$

as a measure of model fit, here y_i^{obs} denotes the actually observed count. If data are discrete, the posterior predictive mid-p-value (Berry & Armitage 1995, Marshall & Spiegelhalter 2003)

$$\text{Prob}(Y_i < y_i^{obs}|y) + \frac{1}{2}\text{Prob}(Y_i = y_i^{obs}|y)$$

can be used instead. An alternative quantity that may be of interest is the posterior predictive ordinate $\pi(y_i^{obs}|y)$. Small values of $\pi(y_i^{obs}|y)$ will indicate an outlying observation.

Extreme posterior predictive (mid-)p-values can be used to identify observations that diverge from the assumed model. However, one drawback concerning the interpretation of posterior predictive p-values is that they do not have a uniform distribution even if the data come from the assumed model. See Hjort et al. (2006), Marshall & Spiegelhalter (2007) and references therein for further details.

We will now explain how posterior p-values can be computed with INLA (Rue et al. 2009). INLA returns an estimate of the posterior marginal of x_i in a discretised way: For $j = 1, \ldots, J$ support points $x_i^{(j)}$ an estimate $\widetilde{\pi}(x_i^{(j)}|y)$ of the posterior density $\pi(x_i^{(j)}|y)$ is given. The support points are chosen such that they cover all areas with non-negligible posterior density. The value of the posterior predictive density (2) can then be approximated using the trapezoidal rule:

$$\hat{\pi}(y_i|y) \approx \sum_{j=2}^{J} \pi(y_i|\frac{1}{2}(x_i^{(j-1)} + x_i^{(j)})) \cdot \frac{1}{2}(x_i^{(j)} - x_i^{(j-1)})(\widetilde{\pi}(x_i^{(j)}|y) + \widetilde{\pi}(x_i^{(j-1)}|y)). \tag{3}$$

Of course, alternative techniques such as Simpson's rule can also be used. For discrete data, the posterior predictive (mid-)p-value can easily be derived as the sum of such probabilities. For $y_i = y_i^{obs}$ we obtain an estimate of the posterior predictive ordinate.

2.3 Leave-one-out Cross-validation with INLA

INLA routinely computes the DIC (Spiegelhalter et al. 2002), a commonly used Bayesian model choice criterion. However, DIC may underpenalize complex models with many random effects (Plummer 2008, Riebler & Held 2009). Alternatively, the conditional predictive ordinate (CPO) (Pettit 1990, Geisser 1993) and the cross-validated probability integral transform (PIT) (Dawid 1984) are available in INLA:

$$\begin{aligned} \text{CPO}_i &= \pi(y_i^{obs}|y_{-i}), \\ \text{PIT}_i &= \text{Prob}(Y_i \leq y_i^{obs}|y_{-i}). \end{aligned}$$

Here y_{-i} denotes the observations y with the i-th component omitted. This facilitates the computation of the cross-validated log-score (Gneiting & Raftery 2007) for model choice. Similarly, PIT histograms (Czado et al. 2009) can be computed to assess calibration of out-of-sample predictions.

We will now describe how these quantities are computed in INLA without re-running the model. Throughout we assume that $y_{-i} = y_{-i}^{obs}$. However, we keep the explicit notation y_i^{obs} for the i-th observation to avoid confusion with other possible realisations of the corresponding random variable Y_i. As before, the vector y will always contain the observed data including y_i^{obs}.

First note that

$$\text{CPO}_i = \int \pi(y_i^{obs}|y_{-i},\theta)\pi(\theta|y_{-i})d\theta, \tag{4}$$

$$\text{PIT}_i = \int \text{Prob}(Y_i \leq y_i^{obs}|y_{-i},\theta)\pi(\theta|y_{-i})d\theta. \tag{5}$$

The first term in the integral in (4) now equals

$$\pi(y_i^{obs}|y_{-i},\theta) = 1 \Big/ \int \frac{\pi(x_i|y,\theta)}{\pi(y_i^{obs}|x_i,\theta)} dx_i. \tag{6}$$

To see this, first note that

$$\pi(x_i|y_{-i},\theta) = \frac{\pi(x_i|y,\theta)\pi(y_i^{obs}|y_{-i},\theta)}{\pi(y_i^{obs}|x_i,\theta)}. \tag{7}$$

Integration with respect to x_i gives (6).

In practice, (6) is computed using numerical integration. The denominator of the ratio in the integral in (6) is the likelihood contribution of the i-th observation and known. However, only an approximation $\widetilde{\pi}(x_i|y,\theta)$ of the numerator $\pi(x_i|y,\theta)$ is known using INLA, as described in Section 2.1. It depends on the accuracy of this approximation how accurate the numerical integration is. In particular, it may happen that the ratio $\widetilde{\pi}(x_i|y,\theta)/\pi(y_i^{obs}|x_i,\theta)$ is multimodal or tends to infinity for extreme values of x_i. It may also be difficult to locate the region of interest, i.e. the region with non-negligible contributions of $\pi(x_i|y,\theta)/\pi(y_i^{obs}|x_i,\theta)$. Such features are an artefact

and a consequence of an imprecise approximation of the numerator $\pi(x_i|y,\theta)$ in the tails. Fortunately, INLA flags such problematic cases, for details see Section 4.

The first term in the integral in (5) can be written as

$$\text{Prob}(Y_i \leq y_i^{obs}|y_{-i},\theta) = \int \text{Prob}(Y_i \leq y_i^{obs}|x_i,\theta)\pi(x_i|y_{-i},\theta)dx_i.$$

The first term in this integral can be computed easily from the likelihood. The second term is available from (7) using $\pi(y_i^{obs}|y_{-i},\theta)$ as computed in (6). As before, $\pi(x_i|y,\theta)$ is available approximately through INLA.

Finally, we need to compute

$$\pi(\theta|y_{-i}) = \frac{\pi(\theta|y)\pi(y_i^{obs}|y_{-i})}{\pi(y_i^{obs}|y_{-i},\theta)}. \tag{8}$$

The denominator $\pi(y_i^{obs}|y_{-i},\theta)$ is known from (6). An approximation to $\pi(\theta|y)$ is available from Section 2.1. Therefore, the normalizing constant

$$\pi(y_i^{obs}|y_{-i}) = 1\Big/\int \frac{\pi(\theta|y)}{\pi(y_i^{obs}|y_{-i},\theta)}d\theta \tag{9}$$

of (8) can be approximately calculated as

$$\widetilde{\pi}(y_i^{obs}|y_{-i}) = 1\Big/\sum_k \frac{\widetilde{\pi}(\theta_k|y)}{\widetilde{\pi}(y_i^{obs}|y_{-i},\theta_k)}\Delta_k. \tag{10}$$

Here the θ_k's are support points of the approximate marginal posterior density $\widetilde{\pi}(\theta|y)$, which has been obtained in the first step of the INLA fitting procedure as described in Section 2.1. So the estimate $\widetilde{\pi}(y_i^{obs}|y_{-i})$ is the *weighted harmonic mean* of the $\widetilde{\pi}(y_i^{obs}|y_{-i},\theta_k)$'s, $k = 1,\ldots,K$, with weights $w_k = \widetilde{\pi}(\theta_k|y)\Delta_k$.

All terms appearing in (4) and (5) are now computed. Final approximation of PIT_i using (5) is based on support points θ_k as in (10) by replacement of the integral with a finite sum. Concerning CPO_i, note that (4) has been approximated already in (10), so the additional integration is not necessary.

3 Predictive Model Checks with MCMC

MCMC delivers samples $x^{(1)},\ldots,x^{(S)}$ from the posterior distribution $\pi(x|y)$. Similarly, samples $\theta^{(1)},\ldots,\theta^{(S)}$ from the posterior distribution $\pi(\theta|y)$ of the hyperparameters can be obtained on a routine basis. These samples are typically dependent, but suitable "thinning" can be be applied to obtain approximately independent samples.

3.1 Posterior Predictive Model Checks with MCMC

Within MCMC the posterior predictive p-values can be derived by drawing a replicate observation $Y_i^{(s)}$ for each of the $s = 1, ..., S$ samples $x_i^{(s)}$ of the MCMC run and counting, how many replicated observations are less than or equal to the actually observed count y_i^{obs}. For discrete data, the posterior predictive mid-p-value and the posterior predictive ordinate can be computed analogously.

If the likelihood $\pi(y_i|x_i)$ is available in closed form, an alternative approach is to average the likelihood across all samples $x_i^{(s)}$ from $\pi(x_i|y)$:

$$\widehat{\pi}(y_i|y) = \frac{1}{S}\sum_{s=1}^{S} \pi(y_i|x_i^{(s)}).$$

This technique is known as Rao-Blackwellization (Gelfand & Smith 1990, Robert & Casella 2004, Casella & Robert 1996) and is typically more accurate than the approach based on replicates $Y_i^{(s)}$ from the predictive density. However, the Monte-Carlo error of the sample-based version is easier to assess so we have used this estimate in Section 4.

3.2 Leave-one-out Cross-validation with MCMC

Omitting the dependence on θ in (6) we obtain

$$\pi(y_i^{obs}|y_{-i}) = 1 \Big/ \int \frac{\pi(x_i|y)}{\pi(y_i^{obs}|x_i)} dx_i. \tag{11}$$

The immediate Monte-Carlo estimate of (11) is simply the harmonic mean of the likelihood values $\pi(y_i^{obs}|x_i)$,

$$\hat{\pi}(y_i^{obs}|y_{-i}) = 1 \Big/ \frac{1}{S}\sum_{s=1}^{S} \frac{1}{\pi(y_i^{obs}|x_i^{(s)})}, \tag{12}$$

evaluated at samples $x_i^{(1)}, \ldots, x_i^{(S)}$ from $\pi(x_i|y)$. This estimate goes back at least to Gelfand (1996) and is very easy to use in MCMC applications. However, the harmonic mean can be numerically unstable and may not even follow a central-limit theorem (Newton & Raftery 1994). This manifests itself by the occasional occurrence of a value $x_i^{(s)}$ with small likelihood $\pi(y_i^{obs}|x_i^{(s)})$ and hence large effect on the estimate (12). Indeed, Raftery (1996) has noted that the reciprocal of (12) may not even have finite variance.

However, for the computation of (mid-)p-values the value of $\pi(y_i|y_{-i})$ needs to be known for all $y_i \leq y_i^{obs}$. An importance sampling approach (Robert & Casella 2004) can be adopted to compute $\pi(y_i|y_{-i})$ for any y_i, not necessarily equal to y_i^{obs}. First

rewrite $\pi(y_i|y_{-i})$ as

$$\begin{aligned}\pi(y_i|y_{-i}) &= \int \pi(y_i|x_i)\pi(x_i|y_{-i})dx_i \\ &= \int \pi(y_i|x_i)\frac{\pi(x_i|y_{-i})}{\pi(x_i|y)}\pi(x_i|y)dx_i.\end{aligned}$$

The importance sampling estimate of $\pi(y_i|y_{-i})$ based on samples $x_i^{(1)},\ldots,x_i^{(S)}$ from $\pi(x_i|y)$ is hence

$$\hat{\pi}(y_i|y_{-i}) = \frac{\sum_{s=1}^{S}\pi(y_i|x_i^{(s)})w_i^{(s)}}{\sum_{s=1}^{S}w_i^{(s)}} \tag{13}$$

with importance weights

$$w_i^{(s)} = \frac{\pi(x_i^{(s)}|y_{-i})}{\pi(x_i^{(s)}|y)} \propto \frac{1}{\pi(y_i^{obs}|x_i^{(s)})},$$

compare Robert & Casella (2004, Equation (3.10)). For count data, the computation of cross-validatory (mid-)p-values reduces then to summing up the estimates $\hat{\pi}(y_i|y_{-i})$ for $y_i = 0,\ldots,y_i^{obs}$ (Marshall & Spiegelhalter 2003). Note that the importance sampling estimate (13) reduces to the harmonic mean (12), if $y_i = y_i^{obs}$.

The variance of importance sampling estimators is difficult to assess; in fact the estimate may not even have finite variance. In particular, if the weights $w_i^{(s)}$ vary widely, they will give too much importance to only a few values of $\pi(y_i|x_i^{(s)})$ and the estimator (13) will be quite unstable, even for large S. However, we have investigated the weights $w_i^{(s)}$ in Section 4 and have found no weight particularly large relative to the others.

3.3 Approximate Cross-validation with MCMC

We now describe an alternative approach, based on an idea originally presented by Marshall & Spiegelhalter (2003) for approximate cross-validation in disease mapping models via MCMC. The method is based on the assumption that

$$\pi(\theta|y_{-i}) \approx \pi(\theta|y).$$

This assumption is plausible for moderate to large dimension of y, since θ is a *global* hyperparameter. Its posterior distribution based on all observations y should not change much if a single observation y_i is omitted.

The Marshall & Spiegelhalter (2003) *mixed predictive approach* is to generate additional samples

$$\tilde{x}_i^{(s)} \sim \pi(x_i|\theta^{(s)}, y_{-i})$$

$s = 1, \ldots, S$, where $\theta^{(s)}$ is a sample from $\pi(\theta|y)$. The samples $\tilde{x}_i^{(s)}$ do not directly depend on y_i, only indirectly because $\theta^{(s)} \sim \pi(\theta|y)$ does depend on y_i. The $\tilde{x}_i^{(s)}$'s are therefore approximately cross-validated and can be used in various ways to compute the predictive model checks discussed earlier.

A straightforward approach to compute PIT values is to draw additional samples $\tilde{y}_i^{(s)}$ from the pseudo-cross-validated predictive distribution and to compute the proportion of samples which are not larger than the observed value y_i^{obs}. Similarly, CPO values can be estimated based on the proportion of samples equal to y_i^{obs}. Alternatively a Rao-Blackwell approach as described in Section 3.1 can be used. In our application the PIT and CPO values resulting from the sampling strategy and the Rao-Blackwellization were almost identical. Mixed predictive PIT and CPO values shown in the following section are computed using Rao-Blackwellization.

4 Application

In our application we consider a typical example from spatial epidemiology. The data considered are cases of bovine viral diarrhoe (BVD) among cows in Switzerland collected during the year 2008. On behalf of an eradication program each cow in Switzerland was tested and the herd was marked as infected, if one or more diseased cows within this herd were detected. As Switzerland is divided in 184 administrative regions, the number of cases is available aggregated on regional level. Additionally, the Principality of Liechtenstein was included in the analysis. A number of 7164 cases was detected in total. For one region the number of cases is missing.

Under the rare disease assumption the usual starting point is to assume that the number of disease cases y_i in region $i = 1, \ldots, 185$ is Poisson distributed with parameter λ_i, which can be interpreted as the relative risk of the disease in the respective region. Additionally, the number of herds m_i is included in the model as an offset to adjust for the different number of herds living in each region. Using a standard formulation with Poisson observation model and a logarithmic link the relative risk parameter λ_i is modelled using the specification

$$\eta_i = \log(\lambda_i) = \log(m_i) + \psi_i + v_i. \tag{14}$$

The spatially unstructured component v_i is assumed to be i.i.d. normally distributed with zero mean and unknown precision τ_v whereas ψ_i is assumed to be structured in space. To account for the assumption that geographically close areas have similar incidence rates the spatially structured component ψ_i is modelled as an intrinsic Gaussian Markov random field with unknown precision τ_ψ (Rue & Held 2005). This model was proposed by Besag et al. (1991), an extension to include covariates has been considered in Clayton & Bernardinelli (1992). The hyperpriors are chosen as

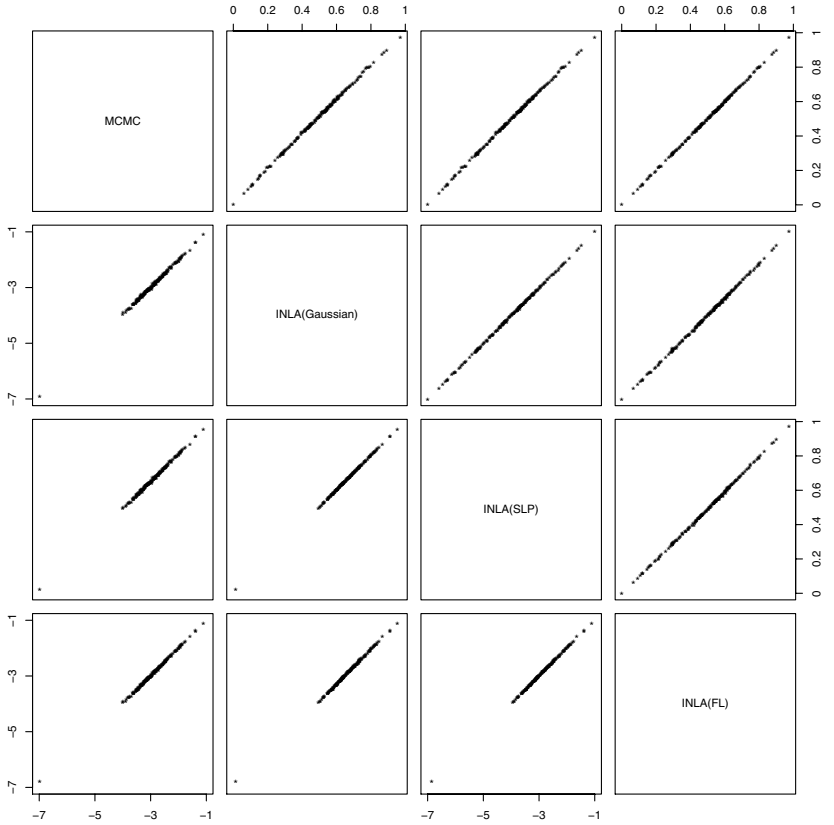

Fig. 1 Scatterplots of posterior predictive mid-p-values (above diagonal) and log posterior predictive ordinates (below diagonal) computed by MCMC and INLA using the Gaussian (Gaussian), simplified Laplace (SLP) and full Laplace (FL) approximation

$\tau_\psi \sim \text{Ga}(1, 0.018)$ and $\tau_v \sim \text{Ga}(1, 0.01)$, compare Bernardinelli et al. (1995) and Schrödle & Held (2009) for some motivation.

For the following analyses an MCMC run of length 930 000 was performed. Using every 30th iteration and a burn-in of 30 000 iterations, 30 000 MCMC samples have been stored. We also tested all three approximation methods available within INLA, as they are known to be differently accurate (Rue & Martino 2007, Rue et al. 2009). All calculations were done using the `inla` program version number 1.526.

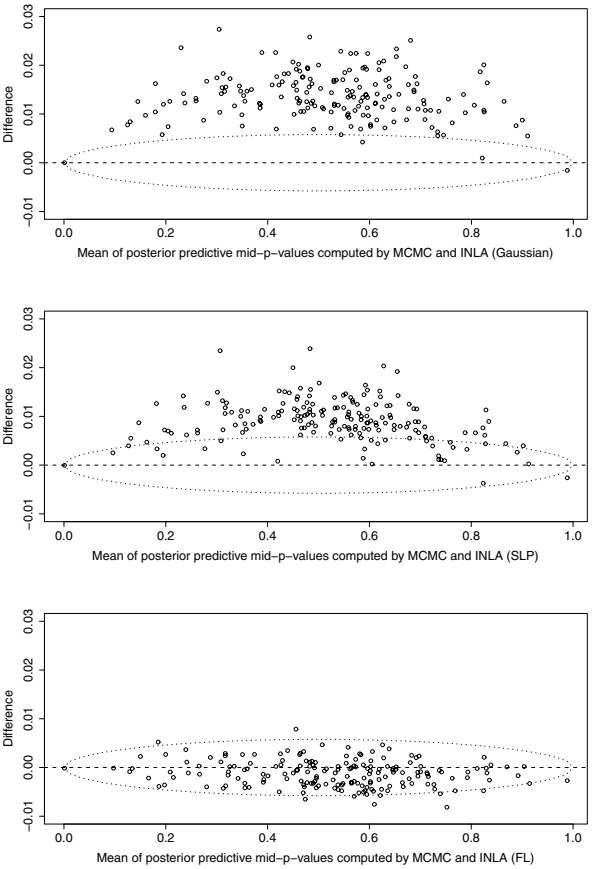

Fig. 2 Bland-Altman plot to investigate the agreement between posterior predictive mid-p-values computed by MCMC vs. INLA using the Gaussian, simplified Laplace and full Laplace approximation. The dotted lines indicate pointwise 95%-confidence intervals based on the Monte-Carlo error attached to the MCMC estimates

4.1 A Comparison of Posterior Predictive Model Checks

In the following the difference between the posterior predictive ordinates and posterior predictive mid-p-values computed by MCMC and INLA using three different approximation methods for the latent Gaussian field will be assessed.

Pairwise scatterplots are shown in Figure 1. The distribution of the posterior predictive ordinates is quite skewed and therefore shown on the log-scale. As can be seen from the plot, the estimates obtained with the four different methods look virtually identical.

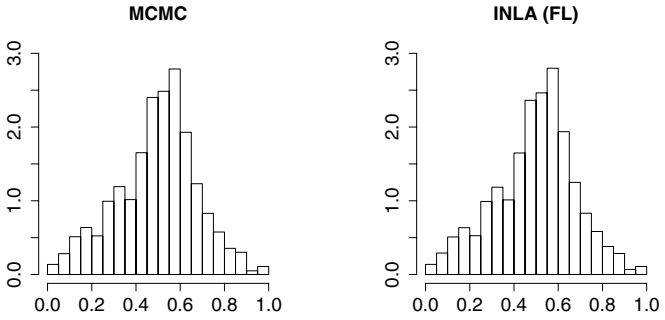

Fig. 3 Adjusted histograms of posterior predictive *p*-values computed by MCMC and INLA using the full Laplace approximation

The extent of agreement between any two methods can be visually examined in more detail using a plot suggested in Bland & Altman (1986), see also Kirkwood & Sterne (2003). The difference between two estimates is plotted on the vertical axis against the mean of each pair on the horizontal axis, see Figure 2. Also shown are 95%-confidence intervals indicating the Monte Carlo error attached to the MCMC estimates. The Monte Carlo standard error has been computed based on the assumption that the MCMC samples are independent. This assumption has been checked by visually inspecting the corresponding empirical autocorrelation functions.

Using this plot systematic bias can be detected and it can be examined if the differences between pairs of estimates depend on the actual value of the estimate. Posterior predictive mid-*p*-values obtained using the Gaussian and simplified Laplace approximation are slightly biased and typically smaller than the corresponding MCMC estimates. The bias is largest for mid-*p*-values around 0.5. For the full Laplace approximation the differences are close to zero and do not show any specific pattern. In fact, nearly all differences are now within the Monte Carlo confidence limits, i.e. the differences can be explained solely by the Monte Carlo error attached to the MCMC estimates. The MCMC estimates based on Rao-Blackwell were even closer to the INLA estimates.

Histograms of posterior predictive mid-*p*-values can be computed in analogy to the PIT histogram (Czado et al. 2009), which was recently proposed for count data. The results are shown in Figure 3 based on MCMC and INLA using the full Laplace approximation. There is virtually no difference to see.

The histograms can be compared with histograms of the cross-validated PIT values in Figure 6. As mentioned in Stern & Cressie (2000) and Marshall & Spiegelhalter (2007) posterior predictive *p*-values are not uniformly distributed and tend to be too conservative as the data are used twice. Indeed, the histograms in Figure 3 are far from uniformity with too many observations having mid-*p*-values around 0.5.

Table 1 Number of unreliable CPO/PIT values for the Gaussian, simplified Laplace and full Laplace approximation

Gaussian	56 unreliable CPO/PIT values
Simplified Laplace	18 unreliable CPO/PIT values
Full Laplace	13 unreliable CPO/PIT values

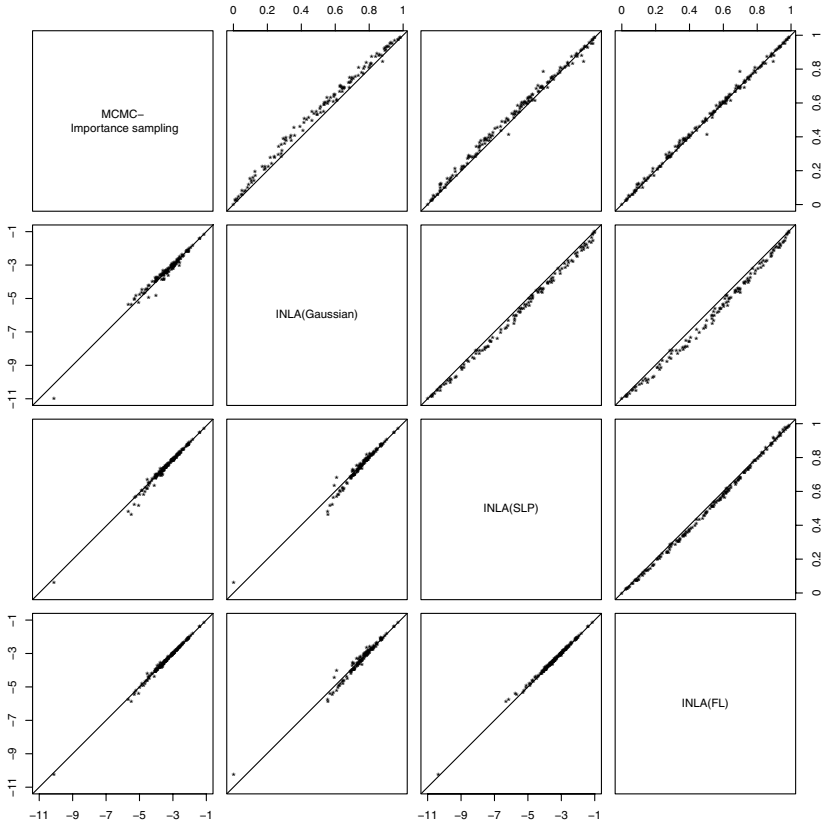

Fig. 4 Scatterplots of leave-one-out cross-validatory predictive mid-p-values (above diagonal) and log conditional predictive ordinates (below diagonal) computed by MCMC vs. INLA using the Gaussian (Gaussian), simplified Laplace (SLP) and full Laplace (FL) approximation

4.2 A Comparison of Leave-one-out Cross-validated Predictive Checks

Leave-one-out cross-validated predictive checks overcome the difficulties of posterior predictive checks mentioned in Section 4.1 and can be used to assess the predictive quality of a model (Marshall & Spiegelhalter 2003, Czado et al. 2009).

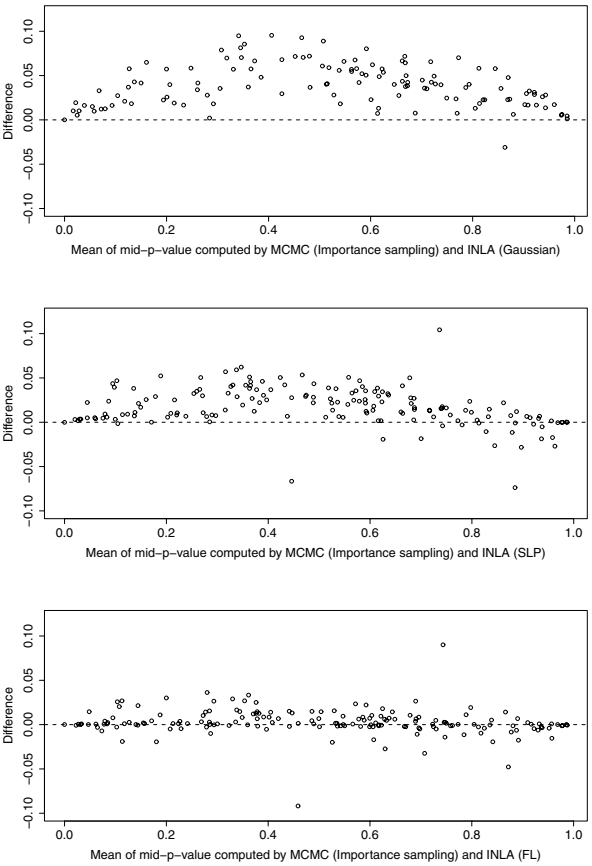

Fig. 5 Bland-Altman plot to investigate the agreement between leave-one-out cross-validated mid-*p*-values computed by MCMC (importance sampling) vs. INLA using the Gaussian, simplified Laplace and full Laplace approximation

Histograms of the PIT values have been proposed to assess the calibration of a model (Czado et al. 2009), the logarithmic score (Gneiting & Raftery 2007), the sum of the log CPO values, can be used for model choice.

INLA returns the CPO and PIT values, as described in Section 2.3. Since the approximation methods for the latent Gaussian field are known to be differently accurate (Rue & Martino 2007, Rue et al. 2009), an empirical comparison is conducted. However, numerical problems may occur when CPO and PIT values are computed in INLA. Some of the CPO and PIT values might not be reliable due to numerical problems in evaluating the integral in (6). INLA automatically stores a file `failure.dat` which contains failure flags for each observation. We considered

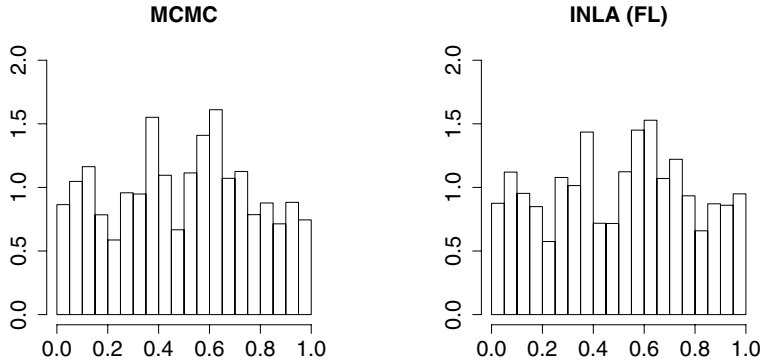

Fig. 6 Adjusted histogram of PIT values computed by MCMC and INLA using the full Laplace approximation

CPO/PIT values with flag equal to 1 as unreliable. Further details on this issue can be found in Martino & Rue (2009).

In Table 1 it is listed for how many observations the computation failed. Most failures occur based on the Gaussian approximation, the full Laplace approximation performs best.

In order to assess the performance of INLA the output will be compared with results from a MCMC analysis based on the estimates (12) and (13). Mid-p- and log CPO values calculated with INLA and MCMC are shown in Figure 4. Each sub-figure is based on all those observations where CPO and PIT values could be computed without failure with the corresponding INLA approximation technique(s) considered.

Figure 4 reveals that the full Laplace approximation is closest to MCMC concerning bias and the differences between the full Laplace and the MCMC output do not show any specific pattern. More details can be seen on the corresponding Bland-Altman plots of the leave-one-out cross-validated mid-p-values, see Figure 5. First of all, a comparison with the corresponding plot showing the posterior predictive mid-p-values (Figure 2) reveals that the differences between MCMC and INLA have increased. However, a similar pattern as in Figure 2 can be seen, with mainly positive differences for the Gaussian and simplified Laplace approximation. In contrast, the mid-p-values computed with the full Laplace approach are closest to the MCMC estimates and do not exhibit a systematic bias. The corresponding PIT histograms are shown in Figure 6 and are quite similar. Note that the PIT histograms are much closer to a uniform distribution than the corresponding posterior predictive histograms shown in Figure 3.

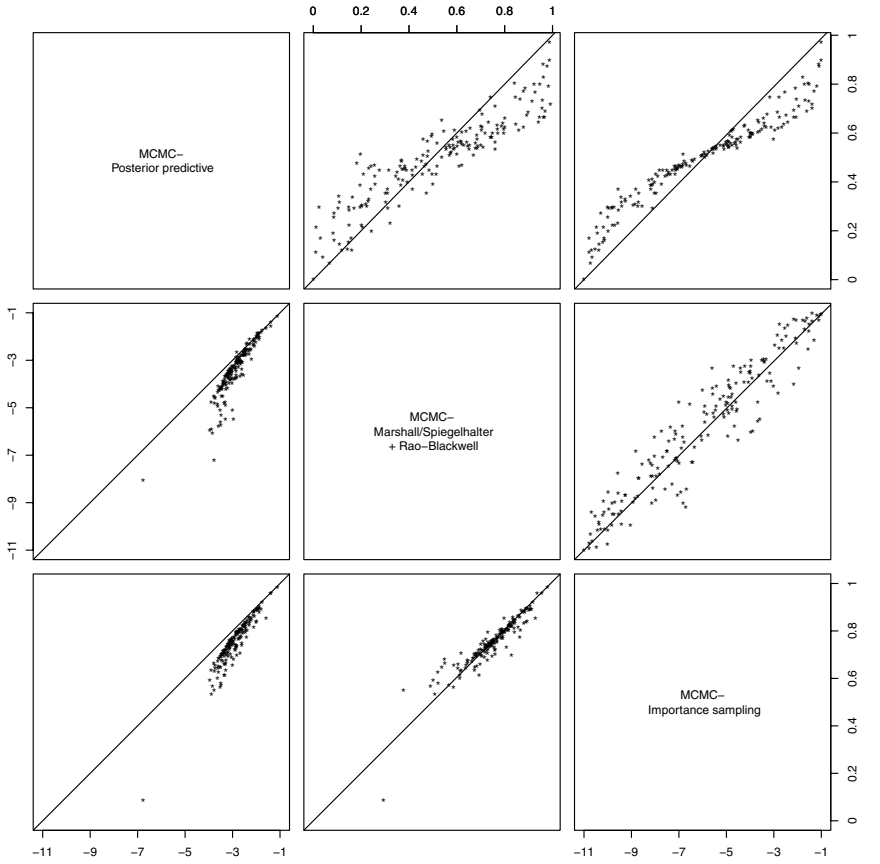

Fig. 7 Scatterplots of mid-p- (above diagonal) and log CPO-values (below diagonal) computed by MCMC using three different approaches: The posterior predictive approach, the mixed predictive approach proposed by Marshall and Spiegelhalter in combination with a Rao-Blackwellization, and importance sampling

4.3 A Comparison of Approximate Cross-validation with Posterior and Leave-one-out Predictive Checks using MCMC

CPO and mid-p-values resulting from a MCMC analysis have also been computed using the mixed predictive approach by Marshall & Spiegelhalter (2003) as described in Section 3.3. The approach is based on posterior samples of the precisions $\tau_v^{(s)}$ and $\tau_\psi^{(s)}$ based on the full data.

Approximately cross-validated samples of η_i and ψ_i are generated in a two-stage procedure based on a reparametrization of model (14) described in Knorr-Held & Rue (2002): First, $\tilde{\psi}_i^{(s)}$ is drawn from the conditional density

$$\tilde{\psi}_i^{(s)} | \psi_{-i}^{(s)}, \tau_\psi^{(s)} \sim N(\frac{1}{n_i} \sum_{j:j\sim i} \psi_j^{(s)}, \frac{1}{n_i \cdot \tau_\psi^{(s)}}).$$

Here n_i denotes the number of neighbours of region i. In a second step, a sample $\tilde{\eta}_i^{(s)}$ of the linear predictor is drawn using

$$\tilde{\eta}_i^{(s)} | \tilde{\psi}_i^{(s)}, \tau_v^{(s)} \sim N(\tilde{\psi}_i^{(s)}, \frac{1}{\tau_v^{(s)}}).$$

This gives pseudo-cross-validated samples $\tilde{\eta}_i^{(s)}$ of the linear predictor, as proposed in Marshall & Spiegelhalter (2003).

Figure 7 compares the mixed predictive approach with the posterior predictive and the cross-validatory approach based on importance sampling. Compared with the importance sampling and the mixed predictive estimates, the posterior predictive estimates are systematically biased. As suspected, the mid-p-values are shrunk towards 0.5. Interestingly, the mixed predictive approach is closer to the ("exact") cross-validatory approach based on importance sampling. There is no systematic bias, although there is some variation in the estimates. This is in contrast to Marshall & Spiegelhalter (2003), who report that the mixed predictive approach performs better than the importance sampling approach in a similar disease mapping model using the well-known Scotland lip cancer data.

5 Discussion

The case study revealed that the cross-validatory checks provided by INLA are close to "exact" importance sampling estimates based on MCMC. The agreement is best if the full Laplace approximation is used. However, the relatively large number of failures is a drawback. Fortunately, these failures are flagged by INLA and it is straightforward to "manually" remove such an observation and to compute the desired leave-one-out quantities directly. The predictive distribution for the observation removed can be calculated in exactly the same way as the posterior predictive distribution, see Section 2.2. For illustration, Figure 8 compares manually computed mid-p-values with the mid-p-values calculated based on the techniques described in Section 2.3 using the full Laplace approximation. The amount of agreement is remarkable.

We finally illustrate how the cross-validated log-score can be used for model comparison. To do so, we have considered two alternative models with either the unstructured or the structured component removed. The logarithmic score in the full model is -3.459, while in the reduced model with no unstructured component the score is even slightly larger (-3.454). However, the score of the model with only an unstructured component is considerably smaller (-3.779). This indicates that the structured component in the model is important, whereas the unstructured

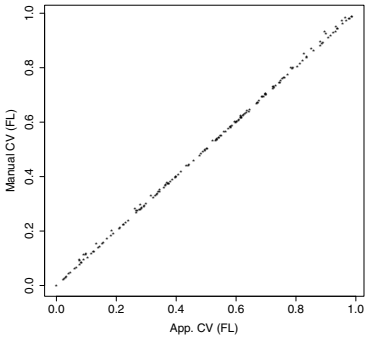

Fig. 8 Scatterplot of manually computed mid-*p*-values using INLA vs. approximate mid-*p*-values obtained from the standard INLA output; the comparison was conducted for the full Laplace approximation

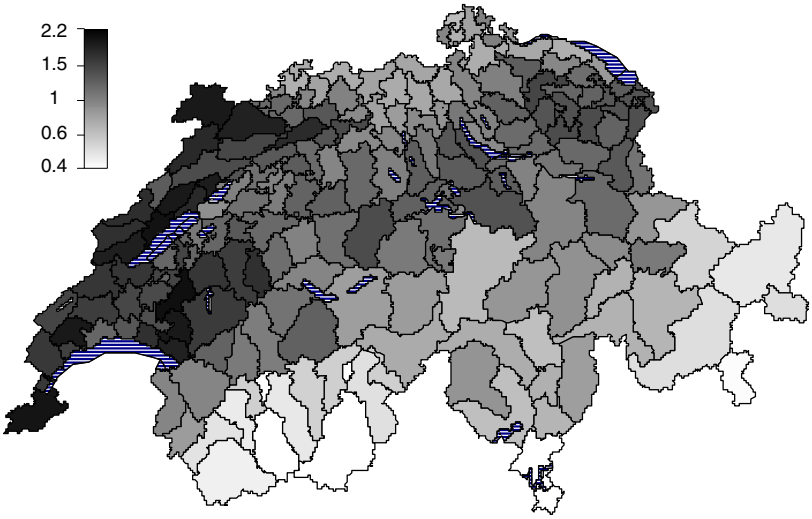

Fig. 9 Fitted relative incidence of BVD in Switzerland, 2008

component can be omitted. The estimated relative incidence obtained from the best model without unstructured component is finally shown in Figure 9.

References

Azzalini, A. & Capitano, A. (1999). Statistical applications of the multivariate skew normal distribution., *Journal of the Royal Statistical Society: Series B* **61**: 579–602.

Bernardinelli, L., Clayton, D. & Montomoli, C. (1995). Bayesian estimates of disease maps: How important are priors?, *Statistics in Medicine* **14**: 2411–2431.

Berry, G. & Armitage, P. (1995). Mid-p confidence intervals: a brief review, *Statistician* **44**: 417–423.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**(1): 1–59.

Bland, J. & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **i**: 307–310.

Breslow, N. & Clayton, D. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**: 9–25.

Casella, G. & Robert, C. (1996). Rao-Blackwellisation of sampling schemes, *Biometrika* **83**: 81–94.

Clayton, D. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, *in* J. Cuzick et al. (eds), *Geographical and Environmental Epidemiology. Methods for Small Area Studies*, Oxford University Press, pp. 205–220.

Czado, C., Gneiting, T. & Held, L. (2009). Predictive model assessment for count data, *Biometrics* . In press.

Dawid, A. P. (1984). Statistical theory: The prequential approach, *Journal of the Royal Statistical Society. Series A (General)* **147**: 278–292.

Fahrmeir, L. (1992). Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models, *Journal of the American Statistical Association* **87**: 501–509.

Fahrmeir, L. & Kneib, T. (2009). Discussion of Rue et al. (2009), *Journal of the Royal Statistical Society: Series B* **71**: 367.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective, *Statistica Sinica* **14**(3): 731–761.

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Journal of the Royal Statistical Society. Series C. Applied Statistics* **50**(2): 201–220.

Geisser, S. (1993). *Predictive Inference: An Introduction*, Chapman & Hall, London.

Gelfand, A. E. (1996). Model determination using sampling-based methods, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in Practice*, Chapman & Hall, pp. 145–161.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**(410): 398–409.

Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica* **6**: 733–807.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**: 359–378.

Gössl, C., Auer, D. P. & Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging, *Biometrics* **57**(2): 554–562.

Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2006). Post-processing posterior predictive p-values, *Journal of the American Statistical Association* **101**(475): 1157–1174.

Kirkwood, B. & Sterne, J. (2003). *Medical Statistics*, 2nd edn, Blackwell Publishing, Oxford.

Knorr-Held, L. & Rue, H. (2002). On block updating in Markov random field models for disease mapping, *Scandinavian Journal of Statistics* **29**(4): 597–614.

Lee, Y., Nelder, J. A. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects - Unified Analysis via H-likelihood*, Chapman & Hall/CRC.

Marshall, E. C. & Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease-mapping methods, *Statistics in Medicine* **22**(4): 1649–1660.

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

Martino, S. & Rue, H. (2009). Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the INLA program. `http://www.math.ntnu.no/~hrue/GMRFLib`.

Natário, I. & Knorr-Held, L. (2003). Non-parametric ecological regression and spatial variation, *Biometrical Journal* **45**(6): 670–688.

Newton, M. A. & Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion), *Journal of the Royal Statistical Society: Series B* **56**: 3–48.

Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution, *Journal of the Royal Statistical Society: Series B* **52**: 175–184.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison, *Biostatistics* **9**(3): 523–539.

Raftery, A. E. (1996). Hypothesis testing and model selection, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in Practice*, Chapman & Hall, pp. 163–187.

Riebler, A. & Held, L. (2009). The analysis of heterogeneous time trends in multivariate age-period-cohort models, *Technical report*, University of Zurich, Biostatistics Unit. Conditionally accepted for *Biostatistics*.

Robert, C. & Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC Press, London.

Rue, H. & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models, *Journal of Statistical Planning and Inference* **137**: 3177–3192.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion), *Journal of the Royal Statistical Society: Series B* **71**: 319–392.

Schrödle, B. & Held, L. (2009). Evaluation of case reporting data from Switzerland: Spatio-temporal disease mapping using INLA, *Technical report*, University of Zurich, Biostatistics Unit.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society: Series B* **64**: 583–639.

Stern, H. & Cressie, N. (2000). Posterior predictive model checks for disease mapping models, *Statistics in Medicine* **19**: 2377–2397.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *J. Amer. Statist. Assoc.* **81**(393): 82–86.