Inference for latent variable models with many hyperparameters

Yoon, Ji Won
Statistics Department
Lloyd Building, College Green
Dublin 2, Ireland
E-mail: yoonj@tcd.ie

Wilson, Simon P.
Statistics Department
Lloyd Building, College Green
Dublin 2, Ireland
E-mail: swilson@tcd.ie

Abstract

The integrated nested Laplace approximation (INLA) [10] is now a well-known functional approximation algorithm for implementing Bayesian inference in latent Gaussian models but has some limitations: it is unable to handle a high dimensional model parameter θ, and makes a poor approximation when the posterior is multi-modal and the likelihood is highly non-Gaussian. A combination of INLA and Monte Carlo methods is proposed to address these limitations. We test the performance of algorithms on a factor analysis model which can be applied to multi-spectral extra-terrestrial microwave maps.

1 Introduction

Latent Gaussian models, a subclass of structured additive regression models, have a wide range of application domains in statistics, signal processing and machine learning. Examples are regression with an additive mixed linear model [5, 8] and random walk model [9], dynamic linear models [13], spatial and spatio-temporal models [1, 2, 3]. The integrated nested Laplace approximation (INLA) [10] is a fast and accurate functional approximation algorithm for implementing Bayesian inference with such models when compared to conventional Monte Carlo simulation.

The latent Gaussian model has the following structure for observations $\boldsymbol{y}$ in terms of latent variables $\boldsymbol{x}$ and hyperparameters θ: $p(\boldsymbol{y} \mid \boldsymbol{x}, \theta) = \prod_j p(y_j \mid x_j, \theta)$ (so observations are conditionally independent), $p(\boldsymbol{x} \mid \theta)$ is a Gaussian Markov random field (GMRF) with precision matrix $\boldsymbol{Q}$, or at least a Gaussian with sparse precision matrix, and there is a some prior $p(\theta)$ on the model parameters. A sparse precision matrix as the prior of the latent Gaussian model speeds up computation, of which the most popular is through the Gaussian Markov random field (GMRF) model [12, 6, 7, 11].

The integrated nested Laplace approximation (INLA) [10] approximates the marginal posterior $p(\mathbf{x}|\mathbf{y})$ by

$$(1) \quad p(\mathbf{x}|\mathbf{y}) = \int p(\mathbf{x}|\mathbf{y},\theta)p(\theta|\mathbf{y})d\theta \approx \int \tilde{p}(\mathbf{x}|\mathbf{y},\theta)\tilde{p}(\theta|\mathbf{y})d\theta \approx \sum_{\theta_j} \tilde{p}(\mathbf{x}|\mathbf{y},\theta_j)\tilde{p}(\theta_i|\mathbf{y})\Delta_{\theta_j}$$

where $\tilde{p}(\theta|\mathbf{y}) = \left.\frac{p(\mathbf{x},\mathbf{y},\theta)}{p_G(\mathbf{x}|\mathbf{y},\theta)}\right|_{\mathbf{x}=\mathbf{x}^*(\theta)}$. Here, $p_G(\mathbf{x}|\mathbf{y},\theta)$ denotes a Gaussian approximation and $\mathbf{x}^*(\theta)$ is its mode. The mode $\mathbf{x}^*(\theta)$ is typically calculated by a numerical optimization of the log posterior.

A set of discrete values θ_j is defined over which $\tilde{p}(\theta|\mathbf{y})$ is computed by first finding a mode μ_θ^* of $\log \tilde{p}(\theta|\mathbf{y})$ by a quasi-Newton optimization and computing the Hessian $\mathbf{H}_\theta^*$ at that mode. A grid search is then conducted from the mode in all directions until $\log \tilde{p}(\mu_\theta^*|\mathbf{y}) - \log \tilde{p}(\theta|\mathbf{y}) > \delta_\pi$ where δ_π is a given threshold. This gives a region over which a grid of θ_j values may be defined.

The contribution of this paper is to propose several variants of Laplace Approximation (LA) to overcome some of limitations of INLA: inability to deal with high dimensionality of the model parameters, posterior multi-modality and highly skewed likelihoods.

2 Proposed algorithms

Problems with INLA occur if the dimension of θ is even modestly large e.g. much above 5 because the grid over θ space becomes too large. There are also problems if $p(\theta|\mathbf{y})$ or $p(\mathbf{x}|\mathbf{y},\theta)$ are multi-modal or severely skewed (i.e. far from Gaussian). In order to address the above problems, we propose two classes of approach. The first addresses a high dimension θ through a Monte Carlo approach to generate samples of θ but not $\mathbf{x}$. The other approach is still a functional approximation, where higher order moments of the function are used. In the multi-modal case, the Gaussian approximation results in large error in the approximated distribution, i.e. given a reasonable threshold δ, $|p(\mathbf{x}^*|\mathbf{y},\theta) - p_G(\mathbf{x}^*|\mathbf{y},\theta)| > \delta$. Several variant LA algorithms are proposed in Algorithm 1.

2.1 IS-LA

The first approach is a hybrid approach between importance sampling and LA, named IS-LA. It uses:

$$p(\mathbf{x}|\mathbf{y}) \approx \int p(\mathbf{x}|\mathbf{y},\theta)\tilde{p}(\theta|\mathbf{y})d\theta \approx \int p(\mathbf{x}|\mathbf{y},\theta)\frac{\tilde{p}(\theta|\mathbf{y})}{q(\theta|\mathbf{y})}q(\theta|\mathbf{y})d\theta \approx \frac{1}{\sum_i w_i}\sum_i p(\mathbf{x}|\mathbf{y},\theta_i)w_i,$$

where $\theta_i \sim q(\theta|\mathbf{y})$ and its weights is defined as $w_i = \frac{\tilde{p}(\theta_i|\mathbf{y})}{q(\theta_i|\mathbf{y})} = \frac{\left.\frac{p(\mathbf{y},\mathbf{x},\theta_i)}{p(\mathbf{x}|\mathbf{y},\theta_i)}\right|_{\mathbf{x}=\mathbf{x}^*(\theta_i)}}{q(\theta_i|\mathbf{y})}$. The proposal function $q(\theta_i|\mathbf{y})$ can be defined in different ways.

- **The prior distribution (IS-LA$^{(1)}$):** The proposal distribution is the prior distribution: $q(\theta|\mathbf{y}) = p(\theta)$. In general importance sampling has a serious problem in obtaining incorrect weights in the tail area when the tail of the proposal distribution is too small. This problem is avoided since the terms are analytically cancelled off as $w_i = \left.\frac{p(\mathbf{y}|\mathbf{x},\theta)p(\mathbf{x}|\theta)}{p_G(\mathbf{x}|\mathbf{y},\theta)}\right|_{\mathbf{x}=\mathbf{x}^*(\theta)}$. In addition, using the prior as the proposal distribution does not require Newton-style optimization.
- **An approximation to the posterior distribution (IS-LA$^{(2)}$):** Although the prior distribution has many benefits, it may not be close to the posterior. It may be better to use a proposal that depends on the data; ideally $q(\theta|\mathbf{y}) = p(\theta|\mathbf{y})$. However, it is rather difficult to obtain a reasonable $p(\theta|\mathbf{y})$ if $p(\mathbf{y} \mid \mathbf{x},\theta)$ is either non-linear or non-Gaussian. A Laplace approximation can be used: $q(\theta|\mathbf{y}) = p_G(\theta|\mathbf{y})$ via the quasi-Newton approach. Although this approach works efficiently, it can suffer from incorrect weights when values are sampled in the tail.

2.2 MH-LA

Another hybrid approach combines LA and the Metropolis-Hastings (MH) algorithm. Samples are generated from $p(\theta|\mathbf{y})$ using the Laplace approximation as the MH proposal distribution. The accept probability is $\mathcal{A} = \min\left(1, \frac{\tilde{p}(\theta'|\mathbf{y})q(\theta|\theta')}{\tilde{p}(\theta|\mathbf{y})q(\theta'|\theta)}\right)$ where θ and θ' denote current and new samples respectively, and $\tilde{p}(\theta'|\mathbf{y}) = \left.\frac{p(\mathbf{y},\mathbf{x},\theta')}{p(\mathbf{x}|\mathbf{y},\theta')}\right|_{\mathbf{x}=\mathbf{x}^*(\theta')}$.
There are several possibilities for the proposal function, as below.

- **A random walk:** One of the simplest proposal functions is a random walk $q(\theta'|\theta) = \mathcal{N}(\theta';\theta,\Sigma_\theta)$ where θ and θ' denote the previous and current samples respectively. In addition, Σ_θ can be chosen either manually or systematically.

- **An approximation to the posterior distribution:** The proposal function used in IS-LA can be used in MH-LA. For instance, the optimal proposal function is designed as $q(\theta^{'}|\theta) = p_G(\theta^{'}|\mathbf{y})$, as by the quasi-Newton approach.

2.3 Multiple initial points

An alternative approach to the multiple mode problem is to repeat the search for modes of $p(\mathbf{x}\,|\,\mathbf{y},\theta)$ from different initial points. Each search yields a mode and Hessian. Of course there is no guarantee that all modes will be discovered, even if the number of them is known.

Algorithm 1 Variant Laplace Approximations

1: Choose an approach, *type*$\in$ {INLA, IS-LA, MH-LA}.
Obtain a mode and its negative hessian matrix by a quasi newton approach for $\tilde{p}(\theta|\mathbf{y})$.
2: $(\mu^*_\theta, \mathbf{H}^*_\theta) = \arg\max_\theta \log\left(\frac{p(\mathbf{y},\mathbf{x},\theta)}{p_G(\mathbf{x}|\mathbf{y},\theta)}\Big|_{\mathbf{x}=\mathbf{x}^*(\theta)}\right)$
where $\mathbf{x}^*(\theta)$ be a mode of $p_G(\mathbf{x}|\mathbf{y},\theta)$.
Obtain θ_is given the following strategies:
3: **if** type is INLA **then**
4: After finding $(\mu^*_\theta, \mathbf{H}^*_\theta)$, do a grid search from the mode in all directions until the log $\tilde{p}(\mu^*_\theta|\mathbf{y}) - \log\tilde{p}(\theta|\mathbf{y}) > \delta_\pi$ where δ_π is a given threshold.
5: **else if** IS-LA **then**
6: Draw samples from the optimal proposal function, $q(\theta|\mathbf{y}) = p(\theta|\mu^*_\theta, \mathbf{H}^{*-1}_\theta)$.
7: Calculate the weights by $w_i = \frac{\tilde{p}(\theta^{(i)}|\mathbf{y})}{q(\theta^{(i)}|\mathbf{y})} = \frac{\frac{p(\mathbf{y},\mathbf{x},\theta^{(i)})}{p_G(\mathbf{x}|\mathbf{y},\theta^{(i)})}\Big|_{\mathbf{x}=\mathbf{x}^*(\theta^{(i)})}}{p(\theta^{(i)}|\mu^*_\theta,\mathbf{H}^{*-1}_\theta)}$.
8: **else if** MH-LA **then**
9: Draw samples from the proposal function such as optimal proposal function $q(\theta^{'}|\theta) = p\left(\theta|\mu^*_\theta, \mathbf{H}^{*-1}_\theta\right)$.
10: Calculate the acceptance ratio by $\mathcal{A} = \min\left\{1, \frac{\tilde{p}(\theta^{'}|\mathbf{y})q(\theta;\theta^{'})}{\tilde{p}(\theta|\mathbf{y})q(\theta^{'};\theta)}\right\}$
11: **end if**
12: Estimate $p(\mathbf{x}|\mathbf{y}) = \sum_{\theta_i} p(\mathbf{x}|\mathbf{y},\theta_i)\tilde{p}(\theta_i|\mathbf{y})\Delta\theta_i$.

3 Application to multi-spectral image source separation

In the multi-spectral source separation problem, the data consists of n_f images of J pixels, that usually correspond to images at different frequencies $v_1, \cdots, v_{n_f}$. The data at pixel j are denoted $\mathbf{y}_j \in \Re^{n_f}$, $j = 1, 2, \cdots, J$, while $\mathbf{Y}_k = (y_{1k}, \cdots, y_{Jk})^T$ denotes the image at frequency v_k. The observed images are believed to be built up of linear combinations of n_s sources, represented by the vectors $\mathbf{x}_j \in \Re^{n_s}$. We assume that the $\mathbf{y}_j$ follow a standard statistical independent components analysis model, so that they can be represented as a linear combination of the $\mathbf{x}_j$ such that $\mathbf{y}_j = \mathbf{A}\mathbf{x}_j + \mathbf{e}_j$ where $\mathbf{A}$ is an $n_f \times n_s$ mixing matrix and $\mathbf{e}_j$ is a vector of n_f independent Gaussian error terms with precisions $\tau = \{\tau_1, \cdots, \tau_{n_f}\}$. We also define $\mathbf{X}_i = \{x_{ji}|j = 1, \cdots, J\}$ to be the image of the ith source. Stacking the $\mathbf{Y}_k$ and $\mathbf{X}_i$ as $\mathbf{y} = (\mathbf{Y}_1^T, \cdots, \mathbf{Y}_{n_f}^T)^T$ and $\mathbf{x} = (\mathbf{X}_1^T, \cdots, \mathbf{X}_{n_s}^T)^T$, and stacking the error terms by frequency $\mathbf{E}$, we have $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{E}$ where $\mathbf{B} = \mathbf{A} \otimes \mathbf{I}_{J\times J}$ is the Kronecker product of $\mathbf{A}$ with the $J \times J$ identity matrix. The vector of error terms $\mathbf{E}$ is zero-mean Gaussian with precision matrix $\mathbf{C} = \text{diag}(\tau_1\mathbf{1}_J, \tau_2\mathbf{1}_J, \ldots, \tau_{n_f}\mathbf{1}_J)$, where $\mathbf{1}_J$ is a vector of ones of length J.

A so-called intrinsic GMRF prior is used for each source $\mathbf{X}_i$, defined as follows. Let Δ_{ji} be differenced values of $\mathbf{X}_i$ at pixel j. A second order random walk can be defined by letting $\Delta_{ji} = \sum_{j'\in \text{ne}(j)} \mathbf{X}_{j'i} - |\text{ne}(j)|\mathbf{X}_{ji}$ be independent zero-mean Gaussians with precision ϕ_i, for $i = 1, \ldots, n_s$ where $|\cdot|$ represents the

cardinality of a particular set and ne(j) is the set of neighbouring pixels of j. It can be shown that the resulting distribution of $\mathbf{X}_i$ is of Gaussian form: $p(\mathbf{X}_i|\theta) \propto \exp\left\{-\frac{\phi_i}{2}{\mathbf{X}_i}^T\mathbf{Q}\mathbf{X}_i\right\}$ where $\mathbf{Q}_f = \mathbf{D}^T\mathbf{D}$ and $\mathbf{D}$ is a $J \times J$ Toeplitz matrix. The stacked set of latent variables $\mathbf{x} = \left[{\mathbf{X}_1}^T, {\mathbf{X}_2}^T, \ldots, {\mathbf{X}_{n_s}}^T\right]^T$ is zero-mean Gaussian with precision $\mathbf{Q} = \text{diag}(\phi_1\mathbf{Q}_f, \cdots, \phi_{n_s}\mathbf{Q}_f)$. However the precision matrix is not of full rank; this is the intrinsic GMRF and is widely used as a prior distribution in Bayesian latent models; see [9].

This is a model of the form where INLA can be applied, with the latent variables $\mathbf{x}$ having precision matrix $\mathbf{Q}$. Original INLA is the best choice in this simple linear model. It is also noted that, in this model, the likelihood is $p(\mathbf{y}|\mathbf{x}, \theta) = \mathcal{N}(\mathbf{Y}; \mathbf{B}\mathbf{x}, \mathbf{C}^{-1})$.

3.1 Mixing Matrix Structure

In this application, $\mathbf{A}$ is parameterized and denoted $\mathbf{A}(\theta)$. One of the most significant applications with this model is the separation of Cosmic Microwave Background (CMB) signals from full-sky map via Plank satilite. Each column of $\mathbf{A}(\theta)$ is the contribution to the observation of a source at different frequencies, which is written as a function of the frequencies and θ. These parameterizations are approximations that come from the current state of knowledge about how the sources are generated. Here, we merely state the parameterization that we are going to use, and refer to [4] for a more detailed exposition on the background to them. It is assumed that the CMB is the first source and therefore, it corresponds to the first column of $\mathbf{A}(\theta)$. It is modelled as a black body at a temperature, and its contribution is a known constant at each frequency. The parametrization of the mixing matrix is given as

$$\mathbf{A}_{i1}(\theta) = \frac{g(v_i)}{g(v_1)}, \mathbf{A}_{i2}(\theta) = \left(\frac{v_i}{v_1}\right)^{\kappa_s} \text{and} \mathbf{A}_{i3}(\theta) = \frac{\exp(\eta v_1/k_B T_1) - 1}{\exp(\eta v_i/k_B T_1) - 1}\left(\frac{v_i}{v_1}\right)^{1+\kappa_d}$$

where $g(v_i) = \left(\frac{\eta v_i}{k_B T_0}\right)^2 \frac{\exp(\eta v_i/k_B T_0)}{(\exp(\eta v_i/k_B T_0)-1)^2}$, $T_0 = 2.725$ is the average CMB temperature in Kelvin, $T_1 = 18.1$, η is the Plank constant and k_B is Boltzmann's constant. The ratio $g(v_i)/g(v_4)$ is designed to ensure that $\mathbf{A}_{41}(\theta) = 1$ as we constraint the fourth row of $\mathbf{A}(\theta)$ to be ones. There are two unknown model parameters for $\mathbf{A}$, for synchrotron $\kappa_s \in \{\kappa_s : -3.0 \leq \kappa_s \leq -2.3\}$ and the spectral indeces for dust $\kappa_d \in \{\kappa_d : 1 \leq \kappa_d \leq 2\}$.

There are two types of hidden variables: sources $\mathbf{x}$ and model parameters θ, which consists of κ_s, κ_d and a precision parameter ϕ_i for the GMRF prior of each source. The prior for θ is designed as $p(\tau_{1:n_s}, \kappa_s, \kappa_d) = [\prod_{i=1}^{n_s} p(\tau_i)]\, p(\kappa_s)p(\kappa_d)$ where $p(\tau_i) = \mathcal{G}(\tau_i; \alpha_i, \beta_i)$, $p(\kappa_s = \mathcal{U}(\kappa_s; \alpha_{\kappa_s}, \beta_{\kappa_s})$ and $p(\kappa_d) = \mathcal{U}(\kappa_d; \alpha_{\kappa_d}, \beta_{\kappa_d})$. Let $\mathcal{G}$ and $\mathcal{U}$ denote the Gamma distribution and the uniform distribution and α and β are hyper-parameters which are fixed in this paper: we assign this hyper parameter to generate flat prior.

3.2 Data description and comparison of the performance of IS-LA and MH-LA

The first source is a second-order isotropic GMRF with mean 0, standard deviation 0.00389 and interaction parameter 0.2. The second source is a mixture of two second-order GMRFs with means 0.003 and 0.006, standard deviation 0.0003 and 0.001 and interaction parameters (horizontal, SW-NE, vertical and NW-SE) of $[-0.1, 0.05, 0.7, 0.05]$ and $[-0.05, 0.05, 0.05, 0.05]$ respectively. The pixel component label (e.g. which GMRF the pixel takes its value from) is generated by an Ising model with temperature $1/0.8$. Pixels of the third source are *i.i.d.* mixture of two Gaussians with means 0.0003 and 0.0006, standard deviations 0.0001 and 0.0001 and weights 0.6 and 0.4. The observations were also generated at 6 channels at the frequencies to be observed by Planck ($30, 44, 70, 100, 143$ and 217GHz), using the mixing matrix model of Section 3.1 with special indices $\kappa_s = -2.8$ and $\kappa_d = 1.4$ as shown in Fig. 1. The measurement error standard deviations used were $0.00126, 0.00120, 0.00113, 0.00028, 0.00018$ and 0.00018mK at each channel, respectively, which are those that are expected to be attained by the Planck detectors.

The extracted sources obtained by the different algorithms are plotted in Fig. 2. The ground truth was well recovered in most approaches. Tables 1 and 2 demonstrate the performance based on a measure of

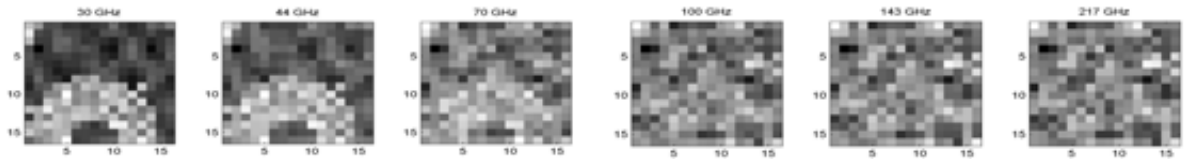

Figure 1: Observed signals

goodness of fit Peak Signal-to-Inference Ratio (PSIR) and running time respectively. It can be seen that the proposed variant LAs have similar performance to MCMC in accuracy while they are considerably faster.

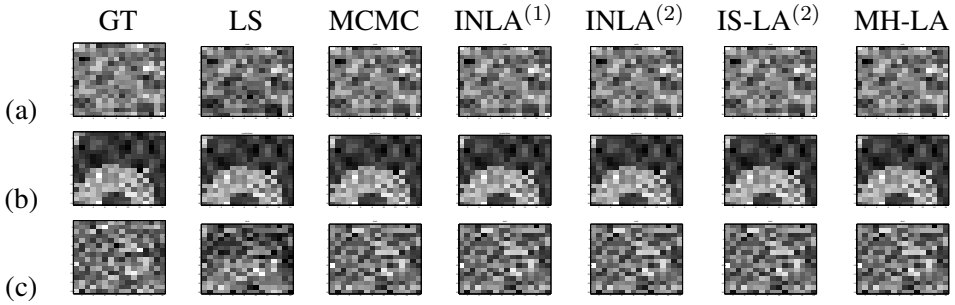

Figure 2: Comparison of restored three signals: GT (Ground Truth) and LS (Least Square)

	CMB	Synchrotron	Dust
LS	28.13± 6.26	45.10±6.66	18.28±6.19
MCMC	41.47±18.0	45.17±7.51	27.80±11.3
INLA$^{(1)}$	40.87±16.3	44.90± 6.25	27.90± 11.3
INLA$^{(2)}$	40.93±17.2	44.90±6.23	27.89±11.2
IS-LA$^{(1)}$	41.19±16.4	44.94±6.29	27.78± 11.7
IS-LA$^{(2)}$	41.11±15.1	44.93±6.37	27.90± 11.3
MH-LA	41.13± 16.1	44.92± 6.27	27.89±11.2

Table 1: PSIR comparison: INLA$^{(1)}$ and INLA$^{(2)}$ used GF and GMRF priors respectively. Also, LS is acronym of (General) Least Square.

LS	MCMC	INLA$^{(1)}$	INLA$^{(1)}$	IS-LA$^{(2)}$	MH-LA
0.00785	1729.2	2.3283	3.1743	23.824	26.836
±0.0315	±140.42	±0.7104	±0.6661	±3.4309	±5.25

Table 2: Run time comparison for the synthetic example

4 Conclusion

The well-known integrated nested Laplace approximation (INLA) has several limitations. In this paper, several variants of Laplace Approximation (LA) are proposed to tackle such serious limitation of the conventional INLA. In order to solve the high dimensionality over the model parameter space, Monte Carlo (MC) simulation are hybridized with LA. This approach still practically fast and efficient since MC draws samples only from model parameter space.

References

[1] J. Besag, J. York, and A. Mollie. Bayesian image restoration with two applications in spatial statistics. *Annuals of the Institute of Statistical Mathematics*, 43:1–59, 1991.

[2] N. Best, S. Richardson, and A. Thomson. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59, 2005.

[3] A. Brix and P. J. Diggle. Spatiotemporal prediction for log-gaussian cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):823–841, 2001.

[4] H. K. Eriksen, C. Dickinson, C. R. Lawrence, and C. Baccigalupi. Cosmic microwave background component separation by parameter estimation. *The Astrophysical Journal*, 2007.

[5] L. Fahrmeir and S. Lang. Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 134, 1999.

[6] K. M. Hanson and G. W. Wecksung. Bayesian approach to limited-angle reconstruction in computed tomography. *Journal of the Optical Society of America*, 73:1501–1509, November 1983.

[7] B. R. Hunt. Bayesian methods in nonlinear digital image restoration. *IEEE Transactions on Computers*, 26(3):219–229, March 1977.

[8] C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models, Second Edition*. Wiley Series in Probabilisty and statistics, 2008.

[9] H. Rue and L. Held. *Gaussian Markov Random Field: Theory and Applications*. Chapman & Hall/CRC, 2005.

[10] H. Rue, S. Martino, and N.s Chopin. Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal Of The Royal Statistical Society Series B*, 71(2):319–392, 2009.

[11] C. Therrien. An approximate method of evaluating the joint likelihood for first-order gmrfs. *IEEE Transactions on Image Processing*, 2(4):520–523, October 1993.

[12] I. S. Weir. Fully Bayesian Reconstruction from Single-Photon Emission Computed Tomography Data. *Journal of the Ameerican Statistical Association*, 92(437):49–60, March 1997.

[13] M. West and J. Harrison. *Bayesian forecasting and dynamic models (2nd ed.)*. Springer-Verlag New York, Inc., New York, NY, USA, 1997.