Approximate Bayesian Inference for latent Gaussian models by using integrated nested Laplace approximations

Havard Rue, Sara Martino, and Niholas Chopin

The paper concerns the development of a fast approximate Bayesian inference in a subclass of structured additive regression models, named *latent Gaussian models.* The common approach to inference for these models is the Markov chain Monte Carlo (MCMC) sampling. Despite all the developments, MCMC sampling remains painfully slow from the end user's point of view.

The main contribution of the paper is the approach to inference based on integrated nested Laplace approximations for the posterior marginals of the Gaussian field as an alternative approach to overcome the drawbacks of the MCMC methods.

In the structured additive regression models, the response variable y_i is assumed to belong to an exponential family, where the mean μ_i is linked to a structured additive predictor η_i through a link function $g(\cdot)$

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \tag{1}$$

The structured additive predictor η_i accounts the effects of various covariates in an additive way, where $\{f^{(j)}(\cdot)\}$s are unknown, often nonlinear, functions of the covariates $\mathbf{u}$, the $\{\beta_k\}$ represent the linear effect of the covariates $\mathbf{z}$ and the ε_is are unstructured terms. These models are very flexible and have a lot of applications due to the different forms that $\{f^{(j)}(\cdot)\}$s can take. The latent Gaussian models are those models which assign a Gaussian prior to α, $\{f^{(j)}(\cdot)\}$, $\{\beta_k\}$ and $\{\varepsilon_i\}$, which are called latent Gaussian variables. The hyperparameters are not necessarily Gaussian. The latent Gaussian models have a wide range of applications like linear regression, dynamic, spatial and spatiotemporal models. In many practical applications the final model is a sum of various components, such as spatial component, random effects, linear and smooth effects for some covariates. Sometimes linear constraints on the latent Gaussian variables are imposed to separate the effects of different components of the additive predictor $\{\eta_i\}$.

The paper provides fast deterministic approximations to all the posterior marginals for x_i, the components of the n-dimensional latent Gaussian vector $\mathbf{x}$, and the posterior marginals for θ_j, the components of the m-dimensional vector θ of hyperparameters, based on so-called *Laplace approximation.* The paper also discusses how to use these marginals to provide adequate approximations to the posterior marginals for subvectors $\mathbf{x}_S$ for any subset S, how to compute the marginal likelihood and the deviance information criterion (DIC) for model comparison, and how to compute Bayesian predictive measures.

The posterior of $(\mathbf{x}, \theta)$ given $\mathbf{y}$ is

$$\begin{aligned} \pi(\mathbf{x}, \theta | \mathbf{y}) &\propto \pi(\theta)\pi(\mathbf{x}|\theta) \prod_{i \in \mathcal{I}} \pi(y_i | x_i, \theta) \\ &\propto \pi(\theta) |\mathbf{Q}(\theta)|^{n/2} \exp\left[-\frac{1}{2} \mathbf{x}^T \mathbf{Q}(\theta) \mathbf{x} + \sum_{i \in \mathcal{I}} \log\{\pi(y_i | x_i, \theta)\} \right]. \end{aligned}$$

$\pi(\mathbf{x}|\theta)$ is the Gaussian prior on the latent variables, assumed having zero-mean and non-singular precision matrix $\mathbf{Q}(\theta)$. The distribution for the n_d response variables $\mathbf{y} = \{y_i : i \in \mathcal{I}\}$ is denoted by $\pi(\mathbf{y}|\mathbf{x}, \theta_2)$ and y_i's are assumed conditionally independent given $\mathbf{x}$ and θ.

Often, the latent Gaussian vector $\mathbf{x}$ has a very large dimension n, which increases the computational complexity of the inference. The latent Gaussian models considered in the paper are assumed to be

Gaussian Markov random fields (GMRF), which have sparse precision matrices, and to have a small number of hyperparameters, $\dim(\theta) = m \leq 6$. These two properties are usually required for fast inference.

When the GMRF is subject to additional linear constraints $\mathbf{Ax} = \mathbf{e}$ for a $k \times n$ matrix $\mathbf{A}$ of rank k, a sample $\mathbf{x}^c$ from the constrained GMRF can be obtained from a sample $\mathbf{x}$ from the unconstrained GMRF by using "conditioning by kriging" $\mathbf{x}^c = \mathbf{x} - \mathbf{Q}^{-1}\mathbf{A}^T(\mathbf{AQ}^{-1}\mathbf{A}^T)^{-1}(\mathbf{Ax} - \mathbf{e})$.

The aim of the paper is to approximate the posterior $\pi(x_i|\mathbf{y})$, $\pi(\theta|\mathbf{y})$, and $\pi(\theta_j|\mathbf{y})$.

The common approach to inference for latent Gaussian models is MCMC sampling. However, the MCMC methods tend to exhibit poor performance when applied to such models, firstly, because the components of the latent field $\mathbf{x}$ are strongly dependent on each other and, secondly, $\mathbf{x}$ and θ are also strongly dependent, when n is large. The paper discuss alternative approaches to try to overcome these problems, including the Gaussian approximations to the full conditional of $\mathbf{x}$ and joint updates of both θ and $\mathbf{x}$. The central part of the paper is the novelty of the approximate inference using the *integrated nested Laplace approximations* (INLA) for approximating the posterior marginals of the latent Gaussian field, $\pi(x_i|\mathbf{y})$, $i = 1, \dots, n$.

The posterior marginals of interest are

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\theta, \mathbf{y})\pi(\theta|\mathbf{y})d\theta,$$

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta_{-j},$$

and the key feature of the new approach is to use this form to construct nested approximations

$$\tilde{\pi}(x_i|\mathbf{y}) = \int \tilde{\pi}(x_i|\theta, \mathbf{y})\tilde{\pi}(\theta|\mathbf{y})d\theta, \tag{2}$$

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \int \tilde{\pi}(\theta|\mathbf{y})d\theta_{-j}. \tag{3}$$

The fast inference is based on the following Laplace approximation

$$\tilde{\pi}(\theta|\mathbf{y}) \propto \left.\frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})}\right|_{x=x^*(\theta)} \tag{4}$$

where $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ is the Gaussian approximation to the full conditional of $\mathbf{x}$, and $\mathbf{x}^*(\theta)$ is the mode of the full conditional $\mathbf{x}$, for given θ.

$$\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \sum_{i \in \mathcal{I}} g_i(x_i)\right\}$$

with $g_i(x_i) = \log\{\pi(y_i|x_i, \theta)\}$. The Gaussian approximation is obtained by matching the modal configuration and the curvature at the mode. The mode is computed iteratively by using a Newton-Raphson method (the Fisher scoring algorithm).

For the posterior marginals of the latent field, it is proposed to start from $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ and approximate the density of $x_i|\theta, \mathbf{y}$ with the Gaussian marginal derived from $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$, i.e.

$$\tilde{\pi}(x_i|\theta, \mathbf{y}) = \mathcal{N}\{x_i; \mu_i(\theta), \sigma^2_{ii}(\theta)\}. \tag{5}$$

where $\mu(\theta)$ is the mean vector of the Gaussian approximation, whereas $\boldsymbol{\sigma^2}(\theta)$ is the covariance matrix.

To obtain the marginals of interest for the latent field, the above approximation can be integrated numerically with respect to θ (see (2))

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}(x_i|\theta_k, \mathbf{y}) \times \tilde{\pi}(\theta_k|\mathbf{y}) \times \Delta_k. \tag{6}$$

The INLA is computed in three steps.

1. Approximate the posterior marginal of θ by using the Laplace approximation (4).

2. Compute the Laplace approximation, or simplified Laplace approximation, of $\pi(x_i|\theta, \mathbf{y})$ for selected values of θ to improve the Gaussian approximation (5).

$$\tilde{\pi}_{LA}(x_i|\theta, \mathbf{y}) \propto \left. \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})} \right|_{x_{-i}=x^*_{-i}(x_i,\theta)}. \tag{7}$$

   The optimization step in computing $\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})$ is avoided by approximating the modal configuration

$$\mathbf{x}^*_{-i}(x_i, \theta) \approx E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i).$$

   The right-hand side is evaluated under the conditional density that is derived from the Gaussian approximation $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$.

   The simplified Laplace approximation $\tilde{\pi}_{SLA}(x_i|\theta, \mathbf{y})$ is derived form the series expansion of $\tilde{\pi}_{LA}(x_i|\theta, \mathbf{y})$ around $x_i = \mu_i(\theta)$ and keeping the terms up to the third order. This allows to correct the Gaussian approximation $\tilde{\pi}_G(x_i|\theta, \mathbf{y})$ for location and skewness. The benefit is purely computational. The simplified Laplace approximation appears to be highly accurate for many observational models.

3. Combine steps 1 and 2 by using the numerical integration (6).

Obviously, the only one way to assess with certainty the approximation error of the INLA method is to run an MCMC sampler for infinite time. Two strategies to asses the approximation error are proposed.

1. Verify overall approximation $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$, for each $\theta_{\mathbf{k}}$ that is used in the numerical integration. This is done by computing $p_D(\theta)$, the *effective number of parameters*, defined by Spiegelhalter *et al.* (2002). Since $\mathbf{x}$ given $\mathbf{y}$ and θ is roughly Gaussian, $p_D(\theta)$ is conveniently approximated by $p_D(\theta) \approx n - tr\{\mathbf{Q}(\theta)\mathbf{Q}^*(\theta)^{-\mathbf{1}}\}$, the trace of the prior precision matrix times the posterior covariance of the Gaussian approximation. The smaller p_D, the better.
2. Comparing elements of a sequence of increasingly accurate approximations, i.e. the Gaussian approximations (5), followed by the simplified Laplace approximation, then by the Laplace approximation (7). Specifically, the integrated marginals (6) are computed on the bases of both the Gaussian approximation and the simplified Laplace Laplace approximation, and compute their symmetric Kullback-Leibler divergence. If the divergence is small, then both approximations are considered acceptable. Otherwise, compute the divergence based on the Laplace approximations (7) and the simplified Laplace approximations. If this divergence is small, simplified Laplace and Laplace approximations appear to be acceptable; otherwise the Laplace approximation is claimed to be the best estimate, but it could be problematic. The last option has not happened in the examples ran by the authors.

The paper provides examples of applications of the INLA approach, with comparisons to results obtained from intensive MCMC runs. Comparisons are expressed in terms of computational time. The new approach provides precise estimates in seconds and minutes, even for models involving thousands of variables, in situations where any MCMC computation typically takes hours or even days.

The authors acknowledge that their work goes against a general trend of favoring exact Monte Carlo methods over non-random approximations. However, the authors point is that, in the specific case of latent Gaussian models, the orders of magnitude involved in the computational cost of both approaches are such that this idealistic point of view is simply untenable for these models. As A. Gelman - professor of statistics at Columbia University - noted, in addition to being a competitor to Gibbs and Metropolis algorithms, the INLA ultimately can be used to make these stochastic algorithms more efficient.

The advantages of the INLA approach are not only computational. It also allows for greater automation and parallel implementation. A challenge remains with problems with many hyperparameters, which are often themselves modeled hierarchically. The computational cost in the INLA approach is exponential in the number of hyperparameters m. In most applications m is small (≤ 6), but applications where m goes up to 10 do exist.

In conclusion, the authors view is that the prospects of the work are more important than this work itself. Near instant inference will make latent Gaussian models more applicable, useful and appealing for the end user.