THE EFFICIENT SELECTION OF AN INITIAL MODE FOR GAUSSIAN APPROXIMATION

Ji Won Yoon and Simon P. Wilson

Statistics Department, Trinity College Dublin, Dublin, Ireland

ABSTRACT

For approximating distributions of unknown form, a Gaussian approximations (GA) is a popular technique. The approximation is often quick and almost always computed iteratively. However, if used in a highly non-linear or non-Gaussian system, then convergence of the approximation may be slow. In this work we propose an efficient approach to determine a good initial mode for a GA method. With such a good initial mode we can reduce the number of iterations of GA for convergence and this enable us to speed up GA. The approach is illustrated through an example of Bayesian inference for disease mapping using A Gaussian (Markov) Random Field (GRF/GMRF) [1].

Index Terms— Gaussian approximation, Non-linear system, mode selection, Gaussian Markov random field

1. GAUSSIAN MODEL

Gaussian Markov random fields (GMRFs) are defined as discrete Gaussian fields with a Markov property that the density of one component conditioned on all the other components depends only on its kth order neighbours [2]. Several application domains are modelled and resolved by GMRF, such as spatio-temporal models [1] and dynamic linear models [3]. [4] introduced the Integrated Nested Laplace Approximation (INLA) for approximating the posterior distribution of both the GMRF and its parameters θ in a latent model of the form:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x},\theta) &= \prod_i p(y_i \,|\, x_i,\theta) \\ \mathbf{x} &\sim GMRF(\theta). \end{aligned} \tag{1}$$

A Gaussian approximation is often used to approximate the marginal posterior of θ by

$$\tilde{\pi}(\theta|\mathbf{y}) \propto \left.\frac{\pi(\mathbf{x},\theta,\mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\theta,\mathbf{y})}\right|_{\mathbf{x}=\mathbf{x}^*(\theta)} \tag{2}$$

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 for $\mathbf{x}$ for a given θ.

This work is part of the STATICA project, funded by the Principal Investigator programme of Science Foundation Ireland, grant number 08/ IN.1/I1879.

Challenges that arise with Gaussian approximations are to construct a more accurate precision matrix while maintaining computational speed and to suggest transformations of the matrix for efficient computation [5]. After exploring the matrix, a Gaussian approximation starts from an initial mode which is generated either randomly or manually. In addition, better approximation is obtained by using Gaussian approximation and its integrand [6]. However, computation time also depends on which mode is selected at the initial step. In this paper, we propose a new approach to speeding up the calculation of the Gaussian approximation by selecting a good initial mode especially for the highly nonlinear problem such as Bayesian mapping of Disease.

1.1. Model for Bayesian Mapping of Disease

Suppose that we have observations $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_i, \cdots, \mathbf{y}_n)$ where i denotes the index of spatial location. The observations are generated from the Poisson distribution given the expected number of cases, λ_i and the relative risk, $\mathbf{x}_i$ in the area i. This model is well studied in epidemiology [7]. The measurement space in this model is defined by

$$p(\mathbf{y}_i|\mathbf{x}_i,\theta) = \mathcal{P}\{\mathbf{y}_i; \lambda_i \exp(\mathbf{x}_i,\theta)\} \tag{3}$$

where λ_i is assumed known and $\mathbf{x}$ follows Gaussian Markov Random process. Here $\mathcal{P}(\cdot)$ represents Poisson distribution.

1.2. Gaussian Markov Random Field

There are several possible designs for GMRFs including intrinsic GMRFs, where the $\mathbf{Q}$ matrix is not of full rank [5]. Set $\mathbf{t} = \mathbf{x} - \mu$ in a one dimensional lattice and assume that it is Gaussian (Markov) Random Field. We define Δ_i as differenced values of $\mathbf{t}$ at site i on the lattice. For example, we can use $\Delta_i = \mathbf{t}_{i+1} + \mathbf{t}_{i-1} - 2\mathbf{t}_i$ for the second order random walk of intrinsic GMRFs [5] and $\Delta_i = \mathbf{t}_i$ for a simple iid Gaussian Random process. The prior distribution for $\mathbf{t}$ is then

$$p(\mathbf{t}|\theta) \propto \exp\left\{-\frac{\kappa}{2}\sum_{i=1}^{n-1}\Delta_i^2\right\} = \exp\left\{-\frac{1}{2}\mathbf{t}^T\mathbf{Q}\mathbf{t}\right\}. \tag{4}$$

Substituting $\mathbf{t}$ into $\mathbf{x} - \mu$, we obtain

$$p(\mathbf{x}|\theta) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{Q}(\mathbf{x}-\mu)\right\} \tag{5}$$

where $\theta = \{\kappa\}$ denotes a set of hidden parameters.

1.3. Gaussian Approximation

When we have the following target distribution

$$\pi(\mathbf{x}) = \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{Q}(\mathbf{x}-\mu) + \sum_{i\in\mathbf{I}} g_i(\mathbf{x}_i)\right\}, \quad (6)$$

the Gaussian approximation of interest $\pi_G(\mathbf{x})$ is obtained by matching the modal configuration and the curvature at the mode where $g_i(\mathbf{x}_i) = \log\pi(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{y}_i\log\lambda + \mathbf{y}_i\mathbf{x}_i - \lambda e^{\mathbf{x}_i} - log(\mathbf{y}_i!)$. Eq. (6) can be transformed to canonical form and we have

$$\begin{aligned}\pi(\mathbf{x}) &\propto \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{Q}(\mathbf{x}-\mu)\right. \\ &\quad \left. -\sum_i\{\lambda_i\exp(\mathbf{x}_i) - \mathbf{y}_i\mathbf{x}_i\}\right] = \exp\{f(\mathbf{x})\} \\ &= \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{c}\mathbf{x} + \mathbf{b}\mathbf{x} + \text{const.}\right\}. \quad (7)\end{aligned}$$

In order to obtain the parameters $\mathbf{c}$ and $\mathbf{b}$ of the canonical form, we use the first and the second derivatives:

$$\begin{aligned}f^{'}(\mathbf{x}) &= -\mathbf{x}^T\mathbf{Q} - \lambda\exp(\mathbf{x}^T) + \mu^T\mathbf{Q} + \mathbf{y}^T \\ f^{''}(\mathbf{x}) &= -\mathbf{Q} - \text{diag}(\lambda\exp(\mathbf{x})) \quad (8)\end{aligned}$$

where $\mathbf{Q}$ is the precision matrix of the prior distribution, i.e. Σ^{-1}. With an empirically chosen mode $\mathbf{m}$, we can use Taylor expansion for $f(\mathbf{x})$ as follows:

$$\begin{aligned}f(\mathbf{x}) &= (\mathbf{x}-\mathbf{m})^T\frac{f^{''}(\mathbf{m})}{2}(\mathbf{x}-\mathbf{m}) + f^{'}(\mathbf{m})(\mathbf{x}-\mathbf{m}) \\ &\quad +\text{constant} \\ &= -\frac{1}{2}\mathbf{x}^T\mathbf{c}\mathbf{x} + \mathbf{b}\mathbf{x} + \text{constant} \quad (9)\end{aligned}$$

Now, we obtain the $\mathbf{c}$ and $\mathbf{b}$ by

$$\begin{aligned}\mathbf{c} &= \mathbf{Q} + \text{diag}(\lambda\exp(\mathbf{m})) \quad (10) \\ \mathbf{b} &= \mathbf{m}^T\text{diag}(\lambda\exp(\mathbf{m})) - \lambda\exp(\mathbf{m}^T) + \mu^T\mathbf{Q} + \mathbf{y}^T\end{aligned}$$

Using $-\frac{1}{2}\mathbf{x}^T\mathbf{c}\mathbf{x} + \mathbf{b}\mathbf{x} + \text{constant} = -\frac{1}{2}(\mathbf{x}-\mathbf{m}^*)^T\mathbf{Q}^*(\mathbf{x}-\mathbf{m}^*)$, we can obtain

$$\begin{aligned}\mathbf{Q}^* &= \mathbf{c} = \mathbf{Q} + \text{diag}(\lambda\exp(\mathbf{m})) \quad (11) \\ \mathbf{m}^* &= \mathbf{Q}^{*-1}\mathbf{b}^T \quad (12)\end{aligned}$$

In order to obtain the optimal mode of $\mathbf{Q}^*$ and $\mathbf{m}^*$, we run Eq. (11) and (12) recursively until convergence.

2. FINDING A GOOD INITIAL MODE

Suppose that the optimal mode is $\mathbf{m}^*$ which is obtained by Eq. (11) and (12) with initial mode $\mathbf{m}_0$. It is very important to find an initial mode $\mathbf{m}_0$ which is close to desired mode $\mathbf{m}^*$ in high dimensional space since the number of recursive iterations of (11) and (12) is dramatically reduced as $\mathbf{m}_0$ becomes closer to $\mathbf{m}^*$. We estimate a close mode to $\mathbf{m}^*$ through use of Cholesky decomposition and appropriate Taylor expansions.

A derivative of $\log\pi(\mathbf{x})$ in terms of $\mathbf{x}$ is given by

$$\frac{\delta l(\cdot)}{\delta\mathbf{x}} \propto \lambda e^{\mathbf{x}} - \mathbf{y} + \mathbf{Q}(\mathbf{x}-\mu). \quad (13)$$

Let $h(\mathbf{x}_t)$ be this equation and then the optimal mode of interest is

$$h(\mathbf{x}) = \lambda e^{\mathbf{x}} - \mathbf{y} + \mathbf{Q}(\mathbf{x}-\mu) = 0. \quad (14)$$

Unfortunately, there is no solution with a closed form for $h(\mathbf{x}) = 0$ so that we approximate it to the simple form and find a neighbour of the mode. In order to do this, we use Taylor series for $e^{\mathbf{x}}$. Since our $\mathbf{x}$ is a $N\times 1$ vector, it is not easy to make Taylor series form so that we transform the vector to a $N\times N$ matrix as follows: $\Lambda : \mathbf{a} \rightarrow A$ where $\mathbf{a} = \begin{bmatrix}a_1 & a_2 & \cdots & a_N\end{bmatrix}^T$ and $\mathbf{A} = \text{diag}(\mathbf{a}) = \Lambda(\mathbf{a})$. For clearance, the bold capital notation stands for the transformed matrix from a vector. Therefore, $\Lambda(e^{\mathbf{x}}) = \mathbf{I} + \mathbf{X} + \frac{\mathbf{X}^T\mathbf{X}}{2} + \cdots$. Now, we have

$$\begin{aligned}\Lambda[h(\mathbf{x})] &= \lambda\left\{I + \mathbf{X} + \frac{\mathbf{X}^T\mathbf{X}}{2} + \cdots\right\} \\ &\quad -\Lambda(\mathbf{y}) + \Lambda\left(\mathbf{Q}(\mathbf{x}-\mu)\right) = 0. \quad (15)\end{aligned}$$

The gradient function $h(\mathbf{x})$ becomes

$$\begin{aligned}\Lambda[h(\mathbf{x})] &= \lambda\left\{I + \mathbf{X} + \frac{\mathbf{X}^T\mathbf{X}}{2} + \cdots\right\} - \mathbf{Y} \\ &\quad +\Lambda\left(\mathbf{Q}\mathbf{x}\right) - \Lambda\left(\mathbf{Q}\mu\right) \\ &\approx \lambda\left\{I + \mathbf{X} + \frac{\mathbf{X}^T\mathbf{X}}{2}\right\} - \mathbf{Y} \\ &\quad +\Lambda\left(\mathbf{Q}\mathbf{x}\right) - \Lambda\left(\mathbf{Q}\mu\right). \quad (16)\end{aligned}$$

Here, we assume that $\Lambda(\mathbf{Q}\mathbf{x}) \approx \mathbf{Q}\Lambda(\mathbf{x}) = \mathbf{Q}\mathbf{X}$. This assumption gives huge benefit to make a simple approximate form:

$$\Lambda[h(\mathbf{x})] \approx \lambda\left\{I + \mathbf{X} + \frac{\mathbf{X}^T\mathbf{X}}{2}\right\} - \mathbf{Y} + \mathbf{Q}\mathbf{X} - \Lambda\left(\mathbf{Q}\mu\right). \quad (17)$$

This equation can be written with regard to $\mathbf{X}$ and we have

$$\Lambda[h(\mathbf{x})] \approx \frac{\lambda}{2}(\mathbf{X}-\alpha)^T(\mathbf{X}-\alpha) + \beta \quad (18)$$

where

$$\begin{aligned}\alpha &= -(\mathbf{I}_{N\times N} + \lambda^{-1}\mathbf{Q}) \\ \beta &= -\frac{\lambda}{2}\alpha^T\alpha + \lambda\mathbf{I}_{N\times N} - \mathbf{Y} - \Lambda(\mathbf{Q}\mu). \quad (19)\end{aligned}$$

Our goal is to find $\mathbf{X}$ to satisfy $\Lambda[h(\mathbf{x})] = 0$. Using Eq. (19) we can achieve it by

$$\Lambda[h(\mathbf{x})] \approx \frac{\lambda}{2}(\mathbf{X}-\alpha)^T(\mathbf{X}-\alpha) + \beta = 0. \quad (20)$$

Thus, we have $(\mathbf{X}-\alpha)^T(\mathbf{X}-\alpha) = -\frac{2}{\lambda}\beta = \mathbf{U}^T\mathbf{U}$ where $\mathbf{U}$ is obtained by cholesky decomposition of $-\frac{2}{\lambda}\beta$. From this equation, we derive $\bar{\mathbf{X}} = \alpha + \mathbf{U}$. Eventually, we can obtain a good initial mode $\mathbf{m}_0$ via the inverse function:

$$\mathbf{m}_0 \quad = \quad \Lambda^{-1}(\bar{\mathbf{X}}) = \Lambda^{-1}(\mathbf{U}+\alpha). \tag{21}$$

Algorithm 1 An efficient initial mode selection in Gaussian approximation: Bayesian mapping of disease

Require: $\mathbf{Q} > 0$
  Set up ϵ which is a threshold for convergence.
  $\alpha = -(\mathbf{I}_{n\times n} + \lambda^{-1}\mathbf{Q})$.
  $\beta = -\frac{\lambda}{2}\alpha^T\alpha + \lambda\mathbf{I} - \Lambda(\mathbf{y}) - \Lambda(\mathbf{Q}\mu)$.
  $\mathbf{U} = \text{chol}\left(-\frac{2}{\lambda}\beta\right)$.
  $\mathbf{m}_0 = \Lambda^{-1}\left(U^T + \alpha\right)$.
  $\mathbf{m} = \mathbf{m}_0$.
  while $true$ **do**
    $\mathbf{Q}^* = \mathbf{Q} + \Lambda(\lambda\exp(\mathbf{m}))$.
    $\mathbf{b} = \mathbf{m}^T\text{diag}(\lambda\exp(\mathbf{m})) - \lambda\exp(\mathbf{m}^T) + \mu^T\mathbf{Q} + \mathbf{y}^T$.
    $\mathbf{R} = \text{chol}(\mathbf{Q}^*)$ where $\mathbf{R}^T\mathbf{R} = \mathbf{Q}$
    $\mathbf{a} = \left(\mathbf{R}^T\right)^{-1}\mathbf{b}^T$.
    $\mathbf{m}^* = \mathbf{R}^{-1}\mathbf{a}$
    if $|\mathbf{m}^* - \mathbf{m}| < \epsilon$ **then**
      break
    else
      $\mathbf{m} = \mathbf{m}^*$ and $\mathbf{Q} = \mathbf{Q}^*$.
    end if
  end while

3. SIMULATION

We have tested our approach with varying dimensions. We set up a set of system parameters used as in Table 1. In order to easily monitor the performance of our approach compared to the conventional approach, a simple Gaussian Random Process is tested with $\mathbf{Q} = \kappa\mathbf{I}$ where $\kappa = 1$.

Table 1. Parameter setting

μ:	the mean of Gaussian process prior	$\mathbf{1}_{d_x}$
$\mathbf{Q}$:	the precision of Gaussian process prior	$\mathbf{I}_{d_x\times d_x}$
λ:	the expected number of cases	1
ϵ:	the threshold for convergence checking	10^{-10}

As we can see in Fig. 1, the modified GA (MGA) also converged to the desired results and it is almost overlapped to the approximation by the conventional GA (CGA).

Fig. 2 shows the comparison of the approximations in each iterations. In this figure, the initial mode estimated by MGA is close to the actual desired mode of the target distribution and it approximates the distribution only after a few

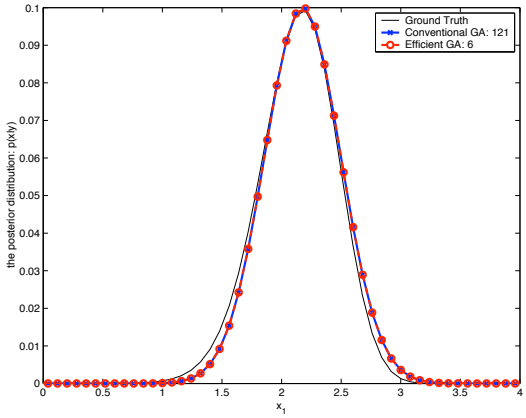


Fig. 1. 1D fitting: target Distribution $\pi(\mathbf{x})$ (solid black line), the distribution by conventional Gaussian approximation (solid red line with square marker) and the distribution by modified Gaussian approximation (dashed line with cross marker)

iterations. Whereas, CGA takes more iterations to obtain reasonable approximation.

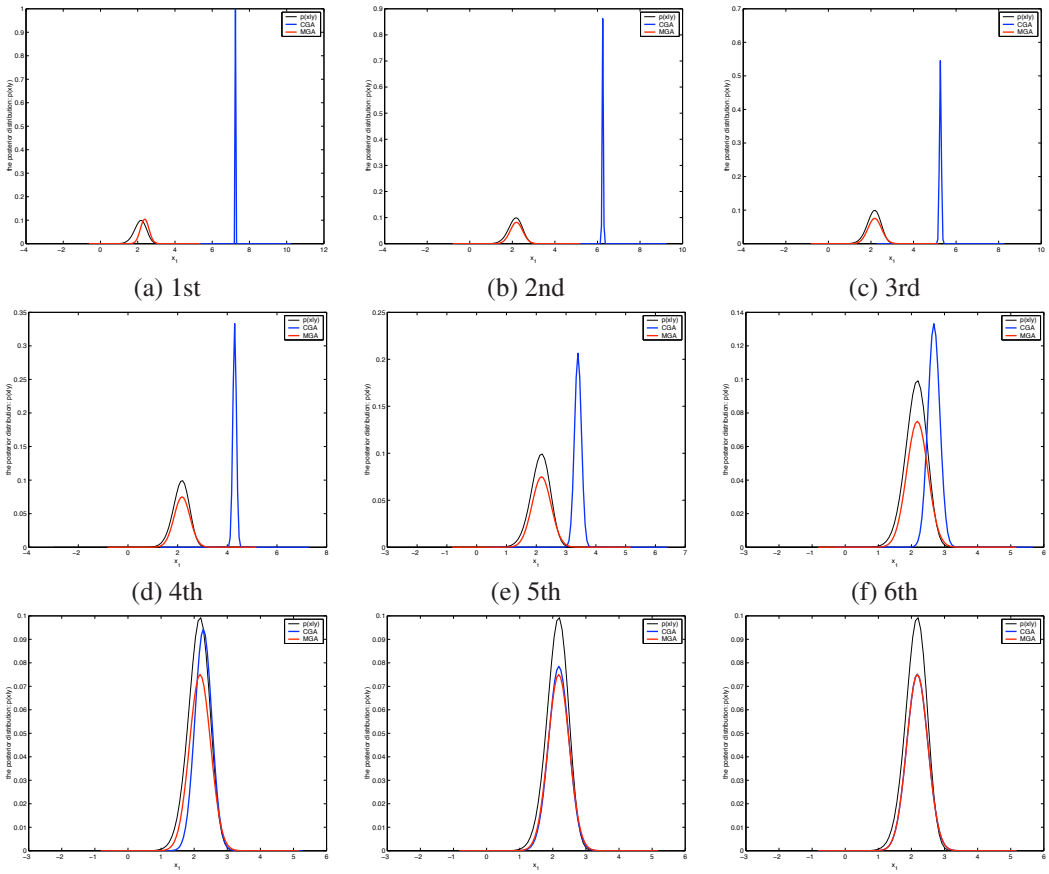
We also compared the number of iterations for both approach with 100 randomly generated samples. For CGA, the initial modes are selected randomly. Table 2 represents means and standard deviations of the numbers of iterations for convergence respectively. As we can see in this table, Modified GA runs relatively small number of iterations for convergence compared to the conventional GA.

Table 2. Comparison of the number of iterations

Dimension	CGA	MGA
1	46.12 ± 60.27	3.830 ± 1.718
2	73.41 ± 63.27	4.810 ± 1.419
3	98.40 ± 67.58	5.180 ± 1.123
5	123.4 ± 65.21	5.740 ± 1.177
10	159.2 ± 64.16	6.430 ± 1.320
50	220.8 ± 41.47	8.100 ± 1.521
100	252.9 ± 44.80	8.970 ± 1.642
200	286.4 ± 43.56	9.400 ± 1.564
500	305.7 ± 35.26	10.91 ± 1.918

4. CONCLUSION

We proposed a fast Gaussian approximation (GA) by embedding an efficient selection of the initial mode. This algorithm does not change any procedure and any results in the conventional GA. However, it saves the number of iterations of GA a lot and it results in speeding up the GA. The power of this algorithm is more effectively shown in higher dimensional problem such as Bayesian mapping of disease. We showed



(a) 1st (b) 2nd (c) 3rd

(d) 4th (e) 5th (f) 6th

(g) 7th (h) 8th (i) 9th

Fig. 2. Comparing the means and covariances of two different approach with 9 iterations: target distribution (solid black line), the approximated distribution by conventional Gaussian approximation (solid blue line) and the approximated distribution by modified Gaussian approximation (solid red line)

that for $n = 500$ our proposed approach has around 10 iterations while the conventional approach has roughly 300 iterations to reach the convergence.

5. REFERENCES

[1] J. Besag, J. York, and A. Mollie, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics*, vol. 43, no. 1, pp. 1–20, March 1991.

[2] K. V. Mardia, "Multi-dimensional multivariate gaussian markov random fields with application to image processing," *J. Multivar. Anal.*, vol. 24, no. 2, pp. 265–284, 1988.

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

[4] 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*, vol. 71, no. 2, pp. 319–392, 2009.

[5] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, vol. 104 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London, 2005.

[6] H. Rue, I. Steinsland, and S. Erland, "Approximating hidden gaussian markov random fields," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 4, pp. 877–892, November 2004.

[7] A. Mollie, *Bayesian Mapping of disease*, Chapman & Hall/CRC, December 1995, in Markov Chain Monte Carlo in Practice.