Review: Variational Bayesian methods for spatial data analysis

Chicas Reading Group

Erick A. Chacon-Montalvan, Claudio Fronterrè

11th August 2017

Centre for Health Informatics, Computing, and Statistics (CHICAS)

Contents

Introduction

Introduction

- Fitting spatial models often involves expensive computations like matrix inversions, whose computational complexity increases in cubic order with the number of spatial locations.

- This situation is aggravated in Bayesian settings where such computations are required once at every iteration of the Markov chain Monte Carlo (MCMC) algorithms.

Alternatives to MCMC

1. Approximating the spatial process using kernel convolutions, moving averages, low-rank splines or basis functions. Essentially, these methods replace the process $w(\boldsymbol{s})$ with an approximation $\tilde{w}(\boldsymbol{s})$ that represents the realizations in a lower-dimensional subspace.
2. A second approach seeks to approximate the likelihood either by working in the spectral domain of the spatial process and avoiding the matrix computations or by forming a product of appropriate conditional distributions to approximate the likelihood (composite likelihoods).
3. Replacing the process (random field) model by a Markov random field or approximating the random field model by a Markov random field (SPDE and INLA).

Variational Bayes

Variational Inference

- Variational inference is a method extensively used in machine learning that approximates probability densities through optimization.
- It has been used in many applications and tends to be faster than classical methods, such as MCMC.
- The idea behind is to first posit a family of densities and then to find the member of that family which is close to the target, where closeness is measured by Kullback-Leibler divergence.

The general problem

Given a set of data $\boldsymbol{y}$ and a set of parameters $\boldsymbol{\theta}$ that govern the model, we are interested in obtaining the posterior distribution

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})}$$

Rather than use sampling throug MCMC, optimization is used. First, we posit a family of approximate densities $\mathcal{Q}$. This is a set of densities over $\boldsymbol{\theta}$. Then, we try to find the member of that family that minimizes the Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in \mathcal{Q}}{\arg\min} \mathrm{KL}\left(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \boldsymbol{y})\right)$$

Finally, we approximate the posterior with the optimized member of the family $q^*(\cdot)$.

The evidence lower bound

The objective function is not computable because it requires computing the evidence $\log p(\boldsymbol{y})$. To see why, recall that KL divergence is

$$\begin{aligned}\mathrm{KL}\left(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\right) &= \mathbb{E}\left[\log q(\boldsymbol{\theta})\right] - \mathbb{E}\left[\log p(\boldsymbol{\theta} \mid \boldsymbol{y})\right] \\ &= \mathbb{E}\left[\log q(\boldsymbol{\theta})\right] - \mathbb{E}\left[\log p(\boldsymbol{\theta}, \boldsymbol{y})\right] + \log p(\boldsymbol{y}).\end{aligned}$$

Because we cannot compute the KL, we optimize an alternative objective that is equivalent to the KL up to an added constant,

$$\mathrm{ELBO}(q) = \mathbb{E}\left[\log p(\boldsymbol{\theta}, \boldsymbol{y})\right] - \mathbb{E}\left[\log q(\boldsymbol{\theta})\right].$$

This function is called the evidence lower bound (ELBO). The ELBO is the negative KL divergence plus $\log p(\boldsymbol{y})$, which is a constant with respect to $q(\boldsymbol{\theta})$. Maximizing the ELBO is equivalent to minimizing the KL divergence.

Interpretation

Examining the ELBO gives intuitions about the optimal variational density. We rewrite the ELBO as a sum of the expected log likelihood of the data and the KL divergence between the prior $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$,

$$\begin{aligned} \mathrm{ELBO}(q) &= \mathbb{E}[\log p(\boldsymbol{\theta})] + \mathbb{E}[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})] - \mathbb{E}[\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})] - \mathrm{KL}(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})). \end{aligned}$$

Hence, the variational objective mirrors the usual balance between likelihood and prior.

The mean-field variational family

The complexity of the family determines the complexity of the optimization. We want a family to be flexible enough to capture a density close to the true posterior, but simple enough for efficient optimization.

The **mean-field variational family** considers the parameters as mutually independent and each governed by a distinct factor in the variational density. A generic member of the mean-field variational family is

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{m} q_j(\boldsymbol{\theta}).$$

Each parameter is governed by its own density. In optimization, these density are chosen to maximize the ELBO.

Coordinate ascent mean-field variational inference

Consider the jth parameter θ_j

- The full conditional of θ_j is its conditional density given all of the other parameters in the model and the observations, $p\left(\theta_j \mid \boldsymbol{\theta}_{-\mathbf{j}}, \boldsymbol{y}\right)$.
- Fix the other variational factors $q_i\left(\theta_i\right)$, $i \neq j$. The optimal $q_j\left(\theta_j\right)$ is then proportional to the exponentiated expected log of the full conditional,

$$
\begin{aligned}
q_j^*\left(\theta_j\right) &\propto \exp\left\{\mathbb{E}_{-j}\left[\log p\left(\theta_j \mid \boldsymbol{\theta}_{-\mathbf{j}}, \boldsymbol{y}\right)\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{-j}\left[\log p\left(\theta_j, \boldsymbol{\theta}_{-\mathbf{j}}, \boldsymbol{y}\right)\right]\right\}.
\end{aligned}
$$

The algorithm

Algorithm 1: Coordinate ascent variational inference

Input: A model $p(\boldsymbol{y}, \boldsymbol{\theta})$, a data set $\boldsymbol{y}$

Output: A variational dentisity $q(\boldsymbol{\theta}) = \prod_{j=1}^{m} q_j(\boldsymbol{\theta})$

Initialize *Variational factors* $q_j(\theta_j)$

while *the* ELBO *has not converged* **do**

  for $j \in \{1, \ldots, m\}$ **do**

    Set $q_j(\theta_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\left(\theta_j \mid \boldsymbol{\theta_{-j}}, \boldsymbol{y}\right)\right]\right\}$

  end

  Compute $\mathrm{ELBO}(q) = \mathbb{E}[\log p(\boldsymbol{\theta}, \boldsymbol{y})] - \mathbb{E}[\log q(\boldsymbol{\theta})]$

end

return $q(\boldsymbol{\theta})$

Variational Bayes on Geostatistical Models

Univariate Geostatistical Model

On a geostatistical setting we usually assume an outcome $Y(s)$ with covariates $\boldsymbol{x}(s)$ at location $s \in D$, such as

$$Y(s) = \boldsymbol{x}^\intercal(s)\boldsymbol{\beta} + w(s) + \epsilon(s), \tag{1}$$

where $\boldsymbol{\beta}$ is the covariates effect vector, $w(s) \sim \mathrm{SGP}(0, \sigma^2, \rho(\phi))$ and $\epsilon(s) \sim \mathcal{N}(0, \tau^2)$.

For Bayesian inference, we can define prior distributions for the parameters

$$\begin{aligned}
\boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\
1/\sigma^2 &\sim Gamma(a_\sigma, b_\sigma) \\
1/\tau^2 &\sim Gamma(a_\tau, b_\tau) \\
\phi &\sim \pi(\phi)
\end{aligned} \tag{2}$$

Two Ways of Specifying the Bayesian Geostatistical Model

Considering a set of locations $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$ and Eq. 1,

$$\begin{aligned} \boldsymbol{w} &\sim \mathcal{N}(0, \sigma^2 \boldsymbol{R}_\phi) \\ \boldsymbol{Y} \mid \boldsymbol{w} &\sim \mathcal{N}(\boldsymbol{X}\beta + \boldsymbol{w}, \tau^2 \boldsymbol{I}_n) \\ \boldsymbol{Y} &\sim \mathcal{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{R}_\phi + \tau^2 \boldsymbol{I}_n) \end{aligned} \tag{3}$$

Two Bayesian models: marginal and hidden GP

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi \mid \boldsymbol{Y}) \propto\ & \pi(\phi) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times \\ & \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\beta, \sigma^2 \boldsymbol{R}_\phi + \tau^2 \boldsymbol{I}_n) \\ \pi(\boldsymbol{\beta}, \boldsymbol{w}, \sigma^2, \tau^2, \phi \mid \boldsymbol{Y}) \propto\ & \pi(\phi) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times \\ & \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times \mathcal{N}(\boldsymbol{w} \mid 0, \sigma^2 \boldsymbol{R}_\phi) \times \\ & \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\beta, \tau^2 \boldsymbol{I}_n) \end{aligned}$$

Model with w as latent: $\pi(\theta \mid Y) \simeq q(\beta)q(\mathbf{w})q(\sigma^2)q(\tau^2)q(\phi)$

Specify hyper-parameters of the prior distributions for σ^2, τ^2 and ϕ.
Give initial values to the expectation of $1/\tau^2$, ϕ, $\mathbf{w}$ and $\mathbf{R}(\phi)^{-1}$: $\mathrm{E}^{(0)}(1/\tau^2) = (1/\tau^2)^{(0)}$, $\mathrm{E}^{(0)}(\phi) = \phi^{(0)}$, $\mathrm{E}^{(0)}(\mathbf{R}(\phi)^{-1}) = \mathbf{R}\left(\boldsymbol{\phi}^{(0)}\right)^{-1}$.

for $t = 1$ to T **do**

Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN\left(\boldsymbol{\mu}_\beta^{(t)}, \mathbf{V}_\beta^{(t)}\right)$, where
$\mathbf{V}_\beta^{(t)} = \left[\mathrm{E}^{(t-1)}\left(1/\tau^2\right)\right]^{-1}(\mathbf{X}'\mathbf{X})^{-1}$ and $\boldsymbol{\mu}_\beta^{(t)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{w}^{(t-1)})$.

Step 2: Update the distribution of $\tau^2 \sim IG$ with parameters $a_\tau + \frac{n}{2}$ and
$b_\tau + \frac{1}{2}\left[\mathrm{Tr}\left(\mathbf{V}_\mathbf{w}^{(t-1)}\right) + p\mathrm{E}^{(t-1)}\left(1/\tau^2\right) + \left(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{w}^{(t-1)}\right)'(\mathbf{I}_n - \mathbf{H})\left(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{w}^{(t-1)}\right)\right]$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
calculate $m_{\tau^2}^{(t)} = \mathrm{E}^{(t)}(1/\tau^2)$.

Step 3: Update the distribution of $\sigma^2 \sim IG$ with parameters $a_\sigma + \frac{n}{2}$ and
$b_\sigma + \frac{1}{2}\left\{\mathrm{Tr}\left[\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right)\mathbf{V}_\mathbf{w}^{(t-1)}\right] + \boldsymbol{\mu}_\mathbf{w}^{(t-1)\prime}\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right)\boldsymbol{\mu}_\mathbf{w}^{(t-1)}\right\}$;
calculate $m_{\sigma^2}^{(t)} = \mathrm{E}^{(t)}(1/\sigma^2)$.

Step 4: Update the distribution of $\mathbf{w} \sim MVN\left(\boldsymbol{\mu}_\mathbf{w}^{(t)}, \mathbf{V}_\mathbf{w}^{(t)}\right)$, where
$\mathbf{V}_\mathbf{w}^{(t)} = \left[m_{\sigma^2}^{(t)}\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right) + m_{\tau^2}^{(t)}\mathbf{I}_n\right]^{-1}$ and
$\boldsymbol{\mu}_\mathbf{w}^{(t)} = m_{\tau^2}^{(t)}\left[m_{\sigma^2}^{(t)}\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right) + m_{\tau^2}^{(t)}\mathbf{I}_n\right]^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_\beta^{(t)}\right)$.

Step 5: Update the distribution of ϕ which is proportional to

$$|\mathbf{R}(\phi)|^{-\frac{1}{2}}\exp\left\{-\frac{m_{\sigma^2}^{(t)}\left[\mathrm{Tr}\left(\mathbf{R}(\phi)^{-1}\mathbf{V}_\mathbf{w}^{(t)}\right) + \boldsymbol{\mu}_\mathbf{w}^{(t)\prime}\mathbf{R}(\phi)^{-1}\boldsymbol{\mu}_\mathbf{w}^{(t)}\right]}{2}\right\}$$

and calculate $\mathrm{E}^{(t)}(\phi)$ and $\mathrm{E}^{(t)}\left(\mathbf{R}(\phi)^{-1}\right)$.

end for

Marginal model: $\pi(\theta \mid Y) \simeq q(\beta)q(r = \sigma^2/\tau^2, \phi)q(\tau^2)$

Specify hyper-parameters of the prior distribution for τ^2, r and ϕ.
Give initial values to the expectation of $1/\tau^2$, ϕ, r and $\mathbf{C}(\phi, r)^{-1}$: $\mathrm{E}^{(0)}(1/\tau^2) = (1/\tau^2)^{(0)}$, $\mathrm{E}^{(0)}(r)$
and $\mathrm{E}^{(0)}(\mathbf{C}(\phi, r)^{-1}) = \mathbf{C}\left(\boldsymbol{\phi}^{(0)}, r^{(0)}\right)^{-1}$.
for $t = 1$ to T **do**

Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN\left(\boldsymbol{\mu}_\beta^{(t)}, \mathbf{V}_\beta^{(t)}\right)$

$\mathbf{V}_\beta^{(t)} = \left[\mathrm{E}^{(t-1)}\left(1/\tau^2\right)\right]^{-1}\left[\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\right]^{-1}$ and $\boldsymbol{\mu}_\beta^{(t)} = \left[\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{Y}$.

Step 2: Update the distribution of $\tau^2 \sim IG$ with parameters $a_\tau + \frac{n}{2}$ and

$b_\tau + \frac{1}{2}\left[\mathrm{Tr}\left(\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\mathbf{V}_\beta^{(t)}\right) + \left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)\right]$;

calculate $\mathrm{E}^{(t)}(1/\tau^2)$.

Step 3: Update the joint distribution of ϕ and r, which is proportional to

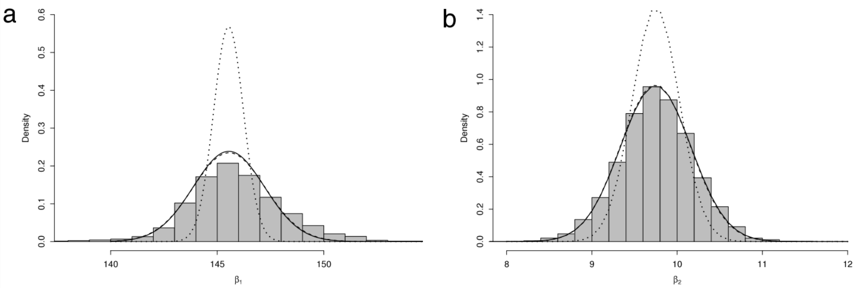
$$|\mathbf{C}|^{-\frac{1}{2}} \times \exp\left\{\mathrm{E}^{(t)}\left(1/\tau^2\right)\left[-\frac{\mathrm{Tr}\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{V}_\beta^{(t)}\right) + \left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)'\mathbf{C}^{-1}\left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)}{2}\right]\right\}$$

and calculate $\mathrm{E}^{(t)}(r)$, $\mathrm{E}^{(t)}(\phi)$ and $\mathrm{E}^{(t)}\left(\mathbf{C}^{-1}\right)$.

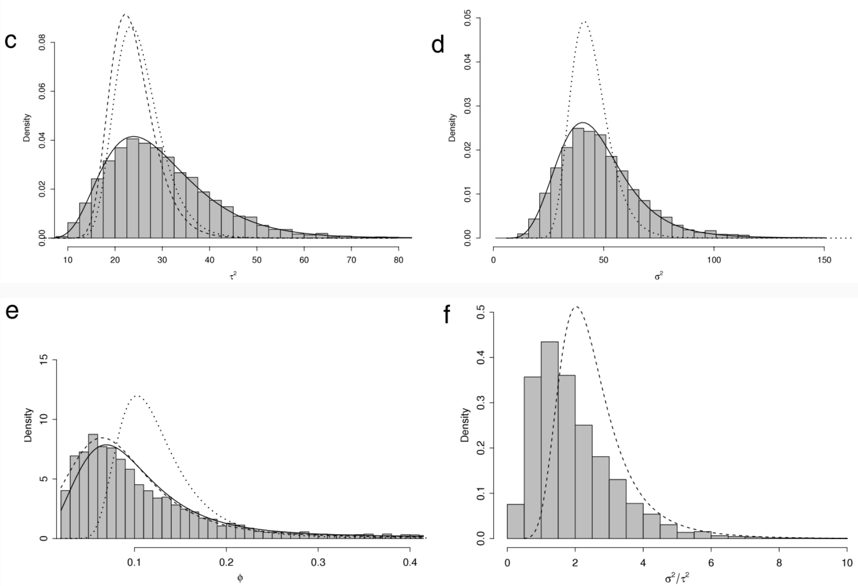
end for

Comparisons of the VB models and MCMC

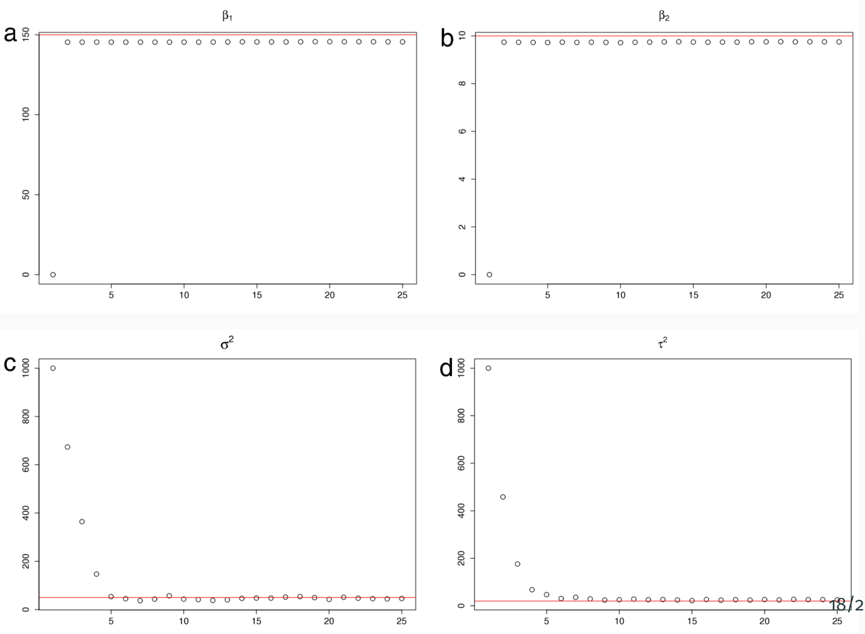
- MCMC samples: histogram
- VB treating $\boldsymbol{w}$ as latent: dotted
- VB marginal model with $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) \simeq q(\boldsymbol{\beta})q(\sigma^2, \tau^2, \phi)$: solid
- VB marginal model with $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) \simeq q(\boldsymbol{\beta})q(r = \sigma^2/\tau^2, \phi)q(\tau^2)$: dashed



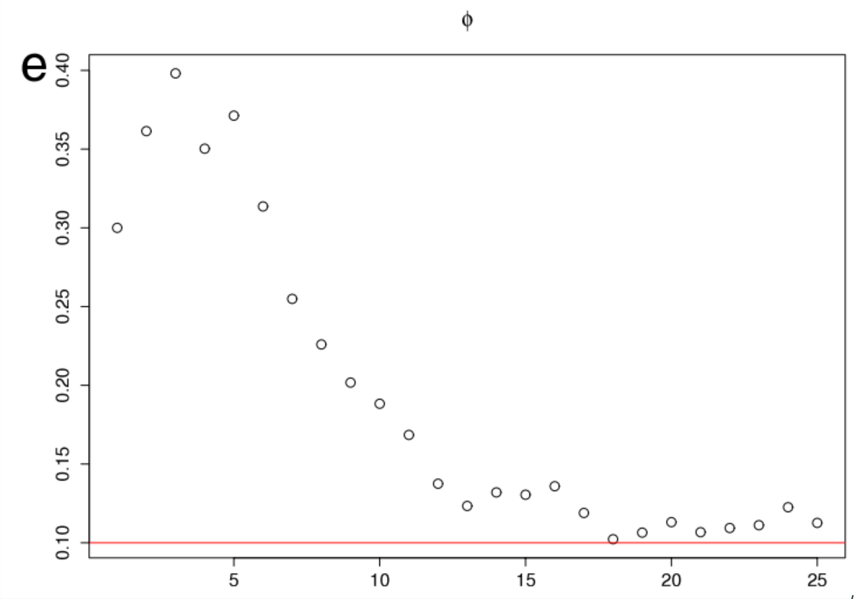
Comparisons of the VB models and MCMC



Convergence of marginal model with $\pi(\theta \mid \boldsymbol{Y}) \simeq q(\beta)q(\sigma^2, \tau^2, \phi)$



Convergence of marginal model with $\pi(\theta \mid \boldsymbol{Y}) \simeq q(\beta)q(\sigma^2, \tau^2, \phi)$



Remarks

Some remarks about Variational Bayes for Geostatistics

- VB methods can provide precise posterior estimates for the parameters in a relative shorter compared to MCMC.
- The **unmarginalized model** offers the advantage of closed form expressions for β, τ^2 and σ^2, but the approximated posteriors tend to underestimate the variance.
- Although it was required to use importance sampling, **marginal models** closely approximated posterior obtained with MCMC.
- Success using variational Bayes strongly depend on the selected **variational family** to approximate the posterior distribution.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877.

Ren, Q., Banerjee, S., Finley, A. O., and Hodges, J. S. (2011). Variational bayesian methods for spatial data analysis. *Comput. Stat. Data Anal.*, 55(12):3197–3217.