

Review: Variational Bayesian methods for spatial data analysis

Chicas Reading Group

Erick A. Chacon-Montalvan, Claudio Fronterre

11th August 2017

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

Introduction

Variational Bayes

Variational Bayes on Geostatistical Models

Models Definition

Comparisons

Convergence

Remarks

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 $w(\mathbf{s})$ with an approximation $\tilde{w}(\mathbf{s})$ 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 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 \mathbf{y} and a set of parameters θ that govern the model, we are interested in obtaining the posterior distribution

$$p(\theta | \mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})}$$

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

$$q^*(\theta) = \arg \min_{q(\theta) \in \mathcal{Q}} \text{KL}(q(\theta) \parallel p(\theta | \mathbf{y}))$$

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 $\log p(\mathbf{y})$. To see why, recall that KL divergence is

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

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

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

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

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 $p(\theta)$ and $q(\theta)$,

$$\begin{aligned}\text{ELBO}(q) &= \mathbb{E}[\log p(\theta)] + \mathbb{E}[\log p(\mathbf{y} \mid \theta)] - \mathbb{E}[\log q(\theta)] \\ &= \mathbb{E}[\log p(\mathbf{y} \mid \theta)] - \text{KL}(q(\theta) \parallel p(\theta)).\end{aligned}$$

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

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

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

Consider the j th 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, $p(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathbf{y})$.
- Fix the other variational factors $q_i(\theta_i)$, $i \neq j$. The optimal $q_j(\theta_j)$ is then proportional to the exponentiated expected log of the full conditional,

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

The algorithm

Algorithm 1: Coordinate ascent variational inference

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

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

Initialize Variational factors $q_j(\boldsymbol{\theta}_j)$

while the ELBO has not converged **do**

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

 Set $q_j(\boldsymbol{\theta}_j) \propto \exp \{ \mathbb{E}_{-j} [(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \mathbf{y})] \}$

end

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

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 $\mathbf{x}(s)$ at location $s \in D$, such as

$$Y(s) = \mathbf{x}^\top(s)\boldsymbol{\beta} + w(s) + \epsilon(s), \quad (1)$$

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

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

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

Two Ways of Specifying the Bayesian Geostatistical Model

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

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

Two Bayesian models: marginal and hidden GP

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

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

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 $\mathbf{R}(\phi)^{-1}$: $E^{(0)}(1/\tau^2) = (1/\tau^2)^{(0)}$, $E^{(0)}(\phi) = \phi^{(0)}$, $E^{(0)}(\mathbf{R}(\phi)^{-1}) = \mathbf{R}(\phi^{(0)})^{-1}$.

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

Step 1: Update the distribution of $\beta \sim MVN(\mu_\beta^{(t)}, \mathbf{V}_\beta^{(t)})$, where

$$\mathbf{V}_\beta^{(t)} = [E^{(t-1)}(1/\tau^2)]^{-1} (\mathbf{X}'\mathbf{X})^{-1} \text{ and } \mu_\beta^{(t)} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{Y} - \mu_{\mathbf{w}}^{(t-1)}).$$

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

$$b_\tau + \frac{1}{2} \left[\text{Tr}(\mathbf{V}_{\mathbf{w}}^{(t-1)}) + p E^{(t-1)}(1/\tau^2) + (\mathbf{Y} - \mu_{\mathbf{w}}^{(t-1)})' (\mathbf{I}_n - \mathbf{H}) (\mathbf{Y} - \mu_{\mathbf{w}}^{(t-1)}) \right], \text{ where } \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'.$$

$$\text{calculate } m_{\tau^2}^{(t)} = E^{(t)}(1/\tau^2).$$

Step 3: Update the distribution of $\sigma^2 \sim IG$ with parameters $a_\sigma + \frac{n}{2}$ and

$$b_\sigma + \frac{1}{2} \left\{ \text{Tr} [E^{(t-1)}(\mathbf{R}(\phi)^{-1}) \mathbf{V}_{\mathbf{w}}^{(t-1)}] + \mu_{\mathbf{w}}^{(t-1)'} E^{(t-1)}(\mathbf{R}(\phi)^{-1}) \mu_{\mathbf{w}}^{(t-1)} \right\};$$

$$\text{calculate } m_{\sigma^2}^{(t)} = E^{(t)}(1/\sigma^2).$$

Step 4: Update the distribution of $\mathbf{w} \sim MVN(\mu_{\mathbf{w}}^{(t)}, \mathbf{V}_{\mathbf{w}}^{(t)})$, where

$$\mathbf{V}_{\mathbf{w}}^{(t)} = [m_{\sigma^2}^{(t)} E^{(t-1)}(\mathbf{R}(\phi)^{-1}) + m_{\tau^2}^{(t)} \mathbf{I}_n]^{-1} \text{ and}$$

$$\mu_{\mathbf{w}}^{(t)} = m_{\tau^2}^{(t)} [m_{\sigma^2}^{(t)} E^{(t-1)}(\mathbf{R}(\phi)^{-1}) + m_{\tau^2}^{(t)} \mathbf{I}_n]^{-1} (\mathbf{Y} - \mathbf{X} \mu_\beta^{(t)}).$$

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

$$|\mathbf{R}(\phi)|^{-\frac{1}{2}} \exp \left\{ -\frac{m_{\sigma^2}^{(t)} \left[\text{Tr}(\mathbf{R}(\phi)^{-1} \mathbf{V}_{\mathbf{w}}^{(t)}) + \mu_{\mathbf{w}}^{(t)'} \mathbf{R}(\phi)^{-1} \mu_{\mathbf{w}}^{(t)} \right]}{2} \right\}$$

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

end for

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

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 $\mathbf{C}(\phi, r)^{-1}$: $E^{(0)}(1/\tau^2) = (1/\tau^2)^{(0)}$, $E^{(0)}(r)$ and $E^{(0)}(\mathbf{C}(\phi, r)^{-1}) = \mathbf{C}(\phi^{(0)}, r^{(0)})^{-1}$.

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

Step 1: Update the distribution of $\beta \sim \text{MVN}(\mu_\beta^{(t)}, \mathbf{V}_\beta^{(t)})$

$$\mathbf{V}_\beta^{(t)} = [\mathbf{E}^{(t-1)}(1/\tau^2)]^{-1} [\mathbf{X}'\mathbf{E}^{(t-1)}(\mathbf{C}^{-1})\mathbf{X}]^{-1} \text{ and } \mu_\beta^{(t)} = [\mathbf{X}'\mathbf{E}^{(t-1)}(\mathbf{C}^{-1})\mathbf{X}]^{-1} \mathbf{X}'\mathbf{E}^{(t-1)}(\mathbf{C}^{-1})\mathbf{Y}.$$

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

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

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

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

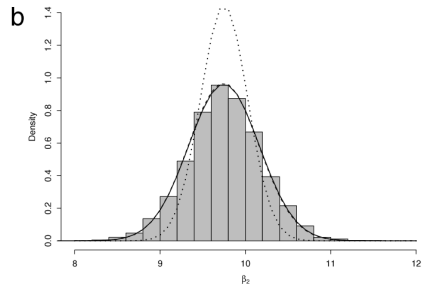
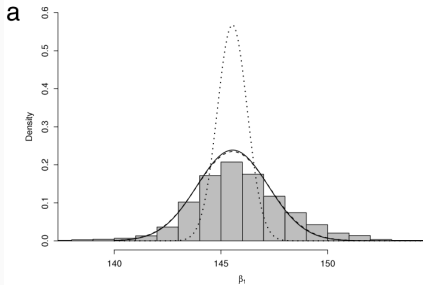
$$|\mathbf{C}|^{-\frac{1}{2}} \times \exp \left\{ E^{(t)}(1/\tau^2) \left[-\frac{\text{Tr}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{V}_\beta^{(t)}) + (\mathbf{X}\mu_\beta^{(t)} - \mathbf{Y})' \mathbf{C}^{-1} (\mathbf{X}\mu_\beta^{(t)} - \mathbf{Y})}{2} \right] \right\}$$

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

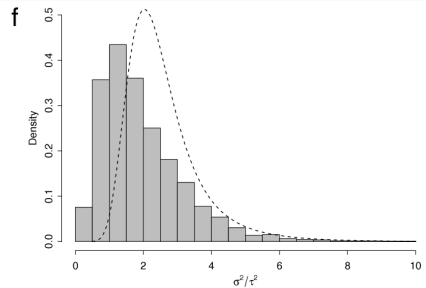
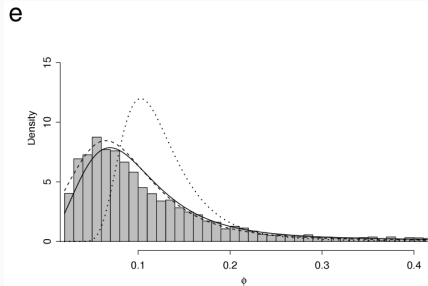
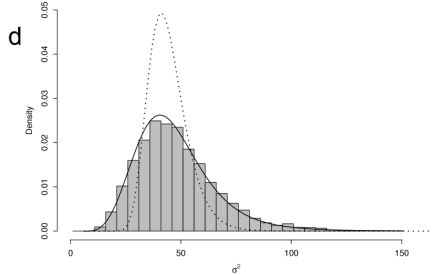
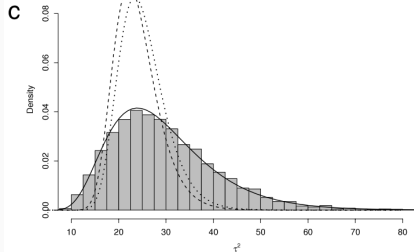
end for

Comparisons of the VB models and MCMC

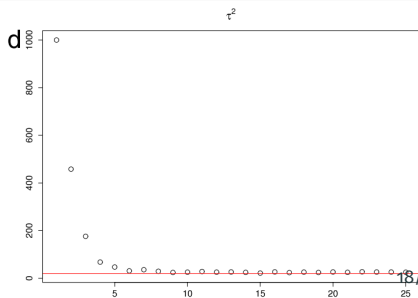
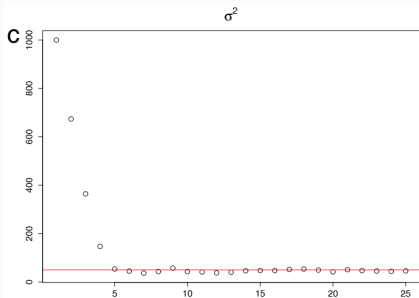
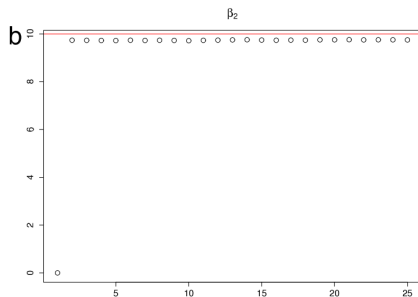
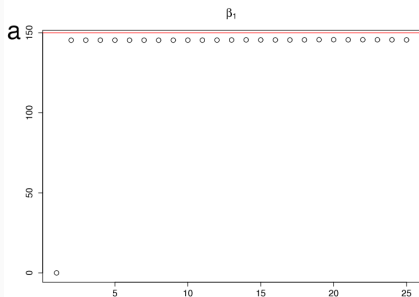
- MCMC samples: histogram
- VB treating \mathbf{w} as latent: dotted
- VB marginal model with $\pi(\theta | \mathbf{Y}) \simeq q(\beta)q(\sigma^2, \tau^2, \phi)$: solid
- VB marginal model with $\pi(\theta | \mathbf{Y}) \simeq q(\beta)q(r = \sigma^2/\tau^2, \phi)q(\tau^2)$: dashed



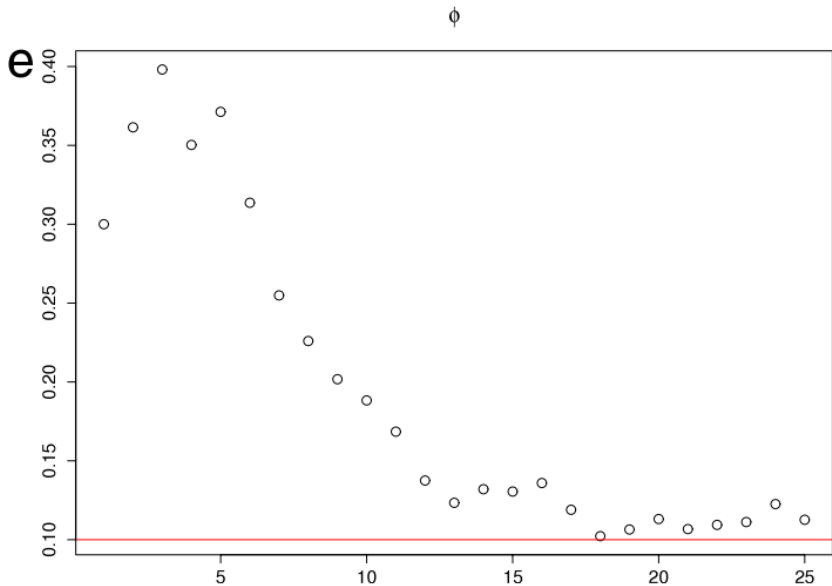
Comparisons of the VB models and MCMC



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



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



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.