TMA4250 Spatial Statistics
Assignement 3: Markov Random Fields

IMF/NTNU

March 2017

Introduction

This assignment contains problems related to Markov Random Fields (MRF). The R-package should be used in solving the problems and relevant functions can be found in the R library `spatial`.

Problem 1: Markov RF

This problem is based on observations of seismic data over a domain $\mathcal{D}_s \subset \mathbb{R}^2$. The objective is to identify the underlying lithology ({sand, shale}) distribution over $\mathcal{D}_s$.

The data are collected on a regular (75×75) grid $\mathcal{L}_{\mathcal{D}_s}$, and the seismic data are denoted $Z(\boldsymbol{s}) : \{Z(\boldsymbol{s_i}); \boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}\}$. The data are available in the R library MASS in the file *seismic.dat*.

Moreover, observations of the lithology distribution ({sand, shale}) in a geologically comparable domain $\mathcal{D}_c \subset \mathbb{R}^2$ are available. The lithology distribution is collected on a regular (66×66) grid $\mathcal{L}_{\mathcal{D}_c}$, with the same spacing as $\mathcal{L}_{\mathcal{D}_s}$, over $\mathcal{D}_c$. The observations with code 0 for sand and 1 for shale are available in the R library MASS in the file *complit.dat*.

Assume that the underlying lithology surface over $\mathcal{D}_s$ can be represented by $\{Y(\boldsymbol{s}); \boldsymbol{s} \in \mathcal{D}_s \subset \mathbb{R}^2\}$ discretized into a lattice $\{Y(\boldsymbol{s_i}) : \boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}\}$ with $Y(\boldsymbol{s_i}) \in \{0, 1\}$ representing sand and shale respectively.

The seismic data stored in *seismic.dat* have uncertainty defined by:

$$Z(\boldsymbol{s_i})|Y(\boldsymbol{s_i}) = \begin{cases} 0.02 + \epsilon(\boldsymbol{s_i}) \text{ if } y(\boldsymbol{s_i}) = 0 \\ 0.08 + \epsilon(\boldsymbol{s_i}) \text{ if } y(\boldsymbol{s_i}) = 1 \end{cases} \quad ; \boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}$$

where $\epsilon(\boldsymbol{s_i})$ iid Gauss$(0, 0.06^2)$.

a) Plot the two datasets *seismic.dat* and *complit.dat*.

b) Consider a uniform prior model on $Y(\boldsymbol{s})$, i.e.

$$[Y(\boldsymbol{s_i})] = Unif\{0,1\} \quad \text{for } \boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}.$$

Specify the data model and the process model for the lithology ({sand,shale}) at an arbitrary position $\boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}$.

Find the posterior distribution $[Y(\boldsymbol{s_i})|Z(\boldsymbol{s_i})]$ for $Y(\boldsymbol{s_i}) \in \{0,1\}$. Simulate 10 realisations of this distribution on $\mathcal{L}_{\mathcal{D}_s}$ and visualize the realisations.

Derive expressions for the posterior mean $\mathrm{E}\{Y(\boldsymbol{s_i})|Z(\boldsymbol{s_i})\}$ and the posterior variance $\mathrm{Var}\{Y(\boldsymbol{s_i})|Z(\boldsymbol{s_i})\}$ and visualize the results. Compare these to the 10 realisations you made earlier.

Finally, use Maximum a posteriori estimation (MAP) to estimate the lithology distribution $Y(\boldsymbol{s_i})$ over $\mathcal{L}_{\mathcal{D}_s}$. Define the MAP-estimator for $Y(\boldsymbol{s_i})$, visualize the results and comment.

c) The Hammersley-Clifford Theorem and Brook's lemma allow us to express a MRF either through one multivariate pdf with respect to cliques (Gibbs formulation) or as a set of univariate conditional local pdf's with respect to neighborhoods (Markov formulation). The Gibbs formulation and the Markov formulation are equivalent.

Now, consider a new prior model for $Y(\boldsymbol{s})$ with neighborhood system denoted by $\mathcal{N}(\cdot) : \{\mathcal{N}(\boldsymbol{s_k}); \boldsymbol{s_k} \in \mathcal{L}_{\mathcal{D}_s}\}$ consisting of the four closest neighbors:

$$[\boldsymbol{Y};\beta] = \mathrm{const}_G \times \exp\{\beta \sum_{\boldsymbol{s_k} \in \mathcal{L}_{\mathcal{D}_s}} \sum_{\boldsymbol{s_j} \in \mathcal{N}(\boldsymbol{s_k})} I(y(\boldsymbol{s_k}) = y(\boldsymbol{s_j}))\}, \tag{1}$$

where $\boldsymbol{Y} = (Y(\boldsymbol{s_1}), ..., Y(\boldsymbol{s_n}))'$ and n is the number of grid nodes in $\mathcal{L}_{\mathcal{D}_s}$. The function $I(A)$ is an indicator function taking the value 1 if A is true and 0 otherwise. The above prior defines the Gibbs formulation of an Ising model.

Specify the Markov formulation corresponding to the Gibbs formulation in (1), i.e derive an expression for the discrete distribution $[Y(\boldsymbol{s_i})|\boldsymbol{Y_{-i}};\beta]$ for each $\boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}$ where $\boldsymbol{Y_{-i}} = (Y(\boldsymbol{s_1}), ..., Y(\boldsymbol{s_{i-1}}), Y(\boldsymbol{s_{i+1}}), ..., Y(\boldsymbol{s_n}))'$.

Derive expressions for the posterior distributions $[\boldsymbol{Y}|\boldsymbol{Z};\beta]$ and $[Y(\boldsymbol{s_i})|\boldsymbol{Y_{-i}}, \boldsymbol{Z};\beta]$ for all $\boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}$. Here, $\boldsymbol{Z} = (Z(\boldsymbol{s_1}), ..., Z(\boldsymbol{s_n}))'$, a vector containing the seismic observations for all $\boldsymbol{s_i} \in \mathcal{L}_{\mathcal{D}_s}$.

We are interested in realizations from $[\boldsymbol{Y}|\boldsymbol{Z};\beta]$ and estimates of $\mathrm{E}\{\boldsymbol{Y}|\boldsymbol{Z};\beta\}$, $\mathrm{Var}\{\boldsymbol{Y}|\boldsymbol{Z};\beta\}$ and $\mathrm{MAP}\{\boldsymbol{Y}|\boldsymbol{Z};\beta\}$. To obtain this, we need to estimate $\hat{\beta}$. Use the observations from the geologically comparable domain $\mathcal{D}_c$ to estimate β by a maximum pseudo-likelihood procedure. Denote the estimate $\hat{\beta}$.

Set the parameter $\beta = \hat{\beta}$ and use a Gibbs sampler MCMC-algorithm to simulate from the posterior $[\boldsymbol{Y}|\boldsymbol{Z};\hat{\beta}]$. You can for example use a single-site updating MCMC-algorithm. Use torus boundary conditions to avoid border problems. Document that the algorithm has converged. Include visualizations of 10 independent realizations in your report.

Estimate $\mathrm{E}\{\boldsymbol{Y}|\boldsymbol{Z};\beta\}$, $\mathrm{Var}\{\boldsymbol{Y}|\boldsymbol{Z};\beta\}$ and $\mathrm{MAP}\{\boldsymbol{Y}|\boldsymbol{Z};\beta\}$ based on several realizations from $[\boldsymbol{Y}|\boldsymbol{Z};\hat{\beta}]$, and visualize the results.

Comment on the results.

d) Compare the results in **b)** and **c)** and comment.