Monday, 2020-11-23

# Home Assignment 2: GMRF:s with Non-Gaussian Data

A common case is that we have non-Gaussian data. Possible examples include: count data, e.g. that we have observed the number of events in each of several regions; 0/1 data, e.g. if an event has occured in each of a number of areas; positive data, e.g. rainfall or concentrations that we know has to be $\geq 0$.

## Model

For 0/1 data one possible model is that the observations are Bernoulli (Binomial with n=1) distributed where the transformed probabilties are spatially dependent:

$$y_i|z_i \in \mathsf{Be}(p_i) \qquad p_i = \frac{e^{z_i}}{1+e^{z_i}} \qquad z_i = \log(p_i) - \log(1-p_i)$$

where $p_i$ is the probability at each location and the latent field, $z_i$, is modeled as

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\boldsymbol{\beta} = \widetilde{\mathbf{A}}\widetilde{\mathbf{x}} \qquad \widetilde{\mathbf{x}} = \begin{bmatrix}\mathbf{x}\\ \boldsymbol{\beta}\end{bmatrix} \in \mathsf{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{Q} & 0\\ 0 & 10^{-6}\cdot\mathbf{I}\end{bmatrix}^{-1}\right).$$

A suitable model for spatial smoothing is that $\mathbf{Q}$ is a CAR or SAR model with

$$\mathbf{Q}_{\text{CAR}} = \tau\left(\varkappa^2\mathbf{C} + \mathbf{G}\right) \qquad \mathbf{Q}_{\text{SAR}} = \tau\left(\varkappa^4\mathbf{C} + 2\varkappa^2\mathbf{G} + \mathbf{G}_2\right)$$

where $\mathbf{G}_2 = \mathbf{G}^2$ can be precomputed and the parameters (to avoid bounded optimisation) are taken as $\theta_1 = \log(\tau)$ and $\theta_2 = \log(\varkappa^2)$

We will define the latent GMRF for a regular grid in longitude and latitude. However, we are only interested in reconstructions for land pixels and those reconstructions can be obtained using the gridded covariates:

$$\mathsf{E}(\mathbf{z}_{\text{grid}}|\mathbf{y}) = \mathsf{E}\left(\begin{bmatrix}\mathbf{A}_{\text{grid}} & \mathbf{B}_{\text{grid}}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \boldsymbol{\beta}\end{bmatrix}\middle|\mathbf{y}\right)$$

For plotting the reconstruction can be projected back to the grid using the transpose of $\mathbf{A}_{\text{grid}}$, see examples in the `proj2_data.m` file.

## Data

Our data is the presence or absence (1/0) of pollen related to agriculture during the 500 years between 1700 and 2200 before present (BP[1], or before 1950), i.e. from 250 BCE[2] to 250 CE. Data has only been obtained for gridcell with lakes or bogs suitable for extracting pollen.

In addition to the observations we have a set of covariates in the `B` and `B_grid` matrices:

[1] `https://en.wikipedia.org/wiki/Before_Present`

[2] `https://en.wikipedia.org/wiki/Common_Era`

**intercept** A column of ones (it is adviasable to **always** include an intercept in the models).

**longitude** Longitude of grid centeroid.

**latitidue** Latitude of grid centeroid.

**elevation** Average elevation for each grid cell, in km.

**d_coast** Distance to coast for each grid cell (in degrees).

**KK10_2000BP** Estimates of human land use for the time-period based on historic population estimates.

**logit_KK10** Logit transformation of the above covariate (consider how the covariates enter the modelling!).

## Theory

Given a model with Bernoulli observations

$$y_i|z_i \in \mathsf{Be}\left(\tfrac{e^{z_i}}{1+e^{z_i}}\right)$$

a few theory-assignments need to be solved **before** we can implement the model (There maybe hints in the skeleton functions `GMRF_negloglike_skeleton.m` and `GMRF_taylor_skeleton.m`):

1. Compute the observation log-density $\log p(y_i|z_i)$ and it's first and second derivatives wrt. $z_i$.

2. Write down the log-posterior of $\widetilde{\mathbf{x}}$

$$\mathrm{p}\left(\widetilde{\mathbf{x}}|\mathbf{y},\boldsymbol{\theta}\right) = \frac{\mathrm{p}\left(\widetilde{\mathbf{x}},\mathbf{y},\boldsymbol{\theta}\right)}{\mathrm{p}\left(\mathbf{y},\boldsymbol{\theta}\right)} \propto \mathrm{p}\left(\mathbf{y}|\widetilde{\mathbf{x}},\boldsymbol{\theta}\right)\mathrm{p}\left(\widetilde{\mathbf{x}}|\boldsymbol{\theta}\right)$$

3. Compute the first and second derivatives of the log-posterior wrt. $\widetilde{\mathbf{x}}$ (used in the internal optimization to find the mode).

4. Write down the approximate log-likelihood (as a function of $\boldsymbol{\theta}$)

$$\mathrm{p}\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \frac{\mathrm{p}\left(\mathbf{y}\middle|\widehat{\mathbf{x}}^{(0)},\boldsymbol{\theta}\right)\mathrm{p}\left(\widehat{\mathbf{x}}^{(0)}\middle|\boldsymbol{\theta}\right)}{\mathrm{p}_G\left(\widehat{\mathbf{x}}^{(0)}\middle|\mathbf{y},\boldsymbol{\theta}\right)}$$

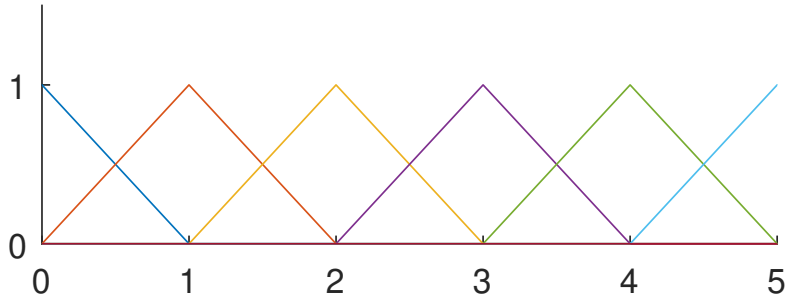
5. To illustrate computations of the **C** and **G** matrices we consider the 1D-case with regularly spaced locations (0,1,2,3,4,5). The 6 basis functions, $\psi_i(t)$, are illustrated below, compute the elements of the 6-by-6 **C** and **G** matrices:

$$C_{ii} \approx \langle \psi_i(t),\, 1\rangle = \int_{-\infty}^{\infty} \psi_i(t)\cdot 1\,\mathrm{d}t \qquad G_{ij} = \left\langle \psi_i(t),\, -\psi_j''(t)\right\rangle$$

Partial-integration and Neumann boundary conditions gives:

$$G_{ij} = \left\langle \psi_i(t),\ -\psi_j''(t) \right\rangle = \left\langle \psi_i'(t),\ \psi_j'(t) \right\rangle$$

and the integrals can be computed "graphically" by drawing suitable figures.




## Estimation

Use the skeleton functions `GMRF_negloglike_skeleton.m` and `GMRF_taylor_skeleton.m` to estimate parameters in a SAR and/or CAR model.

A suitable mean model can be found by studying which $\beta$-values are significant. Given the posterior Gaussian approximation

$$\widetilde{\mathbf{x}}|\mathbf{y},\boldsymbol{\theta} \mathrel{\underset{\sim}{\in}} \mathsf{N}\left(\boldsymbol{\mu}_{x|y}, \mathbf{Q}_{x|y}^{-1}\right)$$

the uncertainty in the $\beta$ can be computer as

$$\mathsf{V}(\beta|\mathbf{y},\boldsymbol{\theta}) \approx \begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix} \mathbf{Q}_{x|y}^{-1} \begin{bmatrix}\mathbf{0} \\ \mathbf{I}\end{bmatrix}$$

or, in MATLAB,

```
e = [zeros(size(Q_xy,1)-size(B,2), size(B,2); eye(size(B,2))];
V_beta0 = e'*(Q_xy\e);
```

It is adviasable to **always** include an intercept in the models!

## Reconstruction

A reconstruction of the latent field (i.e. log-risk) can be recovered as the mode of the last iteration in the optimisation, since

$$\mathsf{E}\left(\widetilde{\mathbf{x}}\middle|\mathbf{y}_{\text{obs}}, \widehat{\boldsymbol{\theta}}\right) = \widetilde{\mathbf{x}}_{\text{mode}}$$

and for different components of the model we have

**Mean component:** $\mathsf{E}(\mathbf{z}_{\text{grid}}|\mathbf{y}) = \mathsf{E}\left(\begin{bmatrix}\mathbf{0} & \mathbf{B}_{\text{grid}}\end{bmatrix} \widetilde{\mathbf{x}}|\mathbf{y}\right)$

**Spatial smooth:** $\mathsf{E}(\mathbf{z}_{\text{grid}}|\mathbf{y}) = \mathsf{E}\left(\begin{bmatrix}\mathbf{A}_{\text{grid}} & \mathbf{0}\end{bmatrix}\widetilde{\mathbf{x}}|\mathbf{y}\right)$

**All spatial structure:** $\mathsf{E}(\mathbf{z}_{\text{grid}}|\mathbf{y}) = \mathsf{E}\left(\begin{bmatrix}\mathbf{A}_{\text{grid}} & \mathbf{B}_{\text{grid}}\end{bmatrix}\widetilde{\mathbf{x}}|\mathbf{y}\right)$

The posterior variance (and confidence intervals) can be obtained through simulation of the posterior to avoid inverting $\mathbf{Q}_{\widetilde{\mathbf{x}}|\mathbf{y}}$. Use the simulated data to find confidence confidence intervals for the components listed above (or there transformation to probabilities).

Consider the following:

- Which models seems most reasonable (covariates, CAR/SAR)?
- How much of the logit-probaiblity is explained by each of the components (give posterior mean and standard errors)?
- What can be said about the spatial structure?

## The report

Write a clear report presenting your approach to the assignment, discussing the methods, model selection and results. Include figures with explanatory texts. The report should be ***no more than 3 pages of text and equations*** excluding figures and answers to the theory questions; place these in an appendix.

Report submissions should be as PDF. Also submit your Matlab .m-files, with a file named `proj2.m` that can be used to run your analysis. (Note that the report should be understandable *without* the .m-files. Also remember to submit *all* of the .m-files you have created.). Report and files should be submitted in **Canvas**.

The report is due on Tuesday 2020-12-08, 12:00. Late reports will only be graded pass/fail. Use the computer labs, office hours and discussion forum in **Canvas** if you need help. Work in groups of two (does not need to be the same groups as in project 1). Discussion between groups is permitted (and encouraged), as long as your report reflects your own work.

## Summary

**IMPORTANT:** `fminsearch` does **MINIMIZATION**, add a suitable number of minus signs to the log-likelihood!

Some reasonable steps to follow are:

1. Start by computing all the components in the theory section
2. Complete `gmrf_taylor_skeleton.m`.
3. Complete `gmrf_negloglike_skeleton.m`.
4. Estimate parameters for the different models and decide on covariates.
5. Reconstruct the different components of the latent field, e.g. $\mathsf{E}(\mathbf{z}|\mathbf{y})$, and compute the reconstruction uncertainty, $\mathsf{V}(\mathbf{z}|\mathbf{y})$.
6. Transform the reconstructions (and uncertainties) to probabilities.