Homework 4: STA465/ STA2016

Homework 4 is due on Wednesday, April 13th. The homework assignment is worth 25 points in total.

Question 1: Sloths in Costa Rica

We will fit a spatial point process model to the sloth occurrence data set. The data is available in the Hwk5Files folder in Quercus: `Hwk5Data.RData`. The data can also be obtained by following the steps in the Sloth Case Study: https://www.paulamoraga.com/tutorial-point-patterns/#1_abstract

The INLA code to fit the model with default priors is:

```
library(INLA)

formula <- Y ~ 1 + cov +
  f(id, model="rw2d", nrow = nrow, ncol = ncol) +
  f(id2, model="iid")

res <- inla(formula, family = "poisson", data = grid@data,
            E = cellarea, control.predictor = list(compute = TRUE))
```

Question 1.1

What class is the `grid` data set? What is the CRS?

Question 1.2

Fit two models to the sloth occurence data.

- Model 1 – set weakly informative priors for the 'iid' and 'rw2d' components
- Model 2 – set noninformative priors for the 'iid' and 'rw2d' components

Include all R + INLA code. Present the results of the estimates along with 95% credible intervals in a table. Comment on any differences across the estimates of the models.

Question 1.3

For each model, create maps of the random effects (both `iid` and `rw2d`), the predicted counts per cell along with the lower and upper limits of the 95% credible interval of predicted counts.

Question 2:

Question 2.1

For the data set in Homework 2 (lung cancer in Pennsylvania), fit the following five models in INLA using the default priors:

- Complete pooling and smoking covariate (no random effects)
- Hierarchical random effect (iid) - (intercept only)
- Hierarchical random effect (iid) + smoking covariate
- Spatial + iid random effect
- Spatial + iid random effect + smoking covariate

Include all INLA code. For each model, compute the CPO, PIT values and create maps of predicted prevalence and standard deviation of predicted prevalence. Comment on any major differences in predicted prevalence across models.

Question 2.2

Organize the results of the estimates, 95% credible intervals, $\sum$log(CPO) for each model in a table. Plot a histogram of the PIT values. Which model has the best predictive performance as measured by $\sum$log(CPO)? What does the histogram of PIT values tell you?

Question 3:

Leave-one-out cross-validation is useful for checking the influence of individual observations and can work well to measure predictive performance if the conditional independence conditions hold (not practical in spatial stats!). An alternative to LOO-CV is to use k-fold CV. For data with dependence (spatial/temporal), it is important to create folds that are independent of one another.

Using the data set in Homework 4 (Malaria prevalence in The Gambia), perform an approximate 4-fold CV that takes spatial dependence into account.

- Include all R and INLA code.
- Make a map that shows how the data are partitioned into the 4 folds.
- Compute $\frac{1}{N}\sum(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the out-of-sample predicted value of y_i given N_i (use a point estimate for $\hat{y}_i$).