Part 2: Model Description

Likelihood

Consider a geographic area divided into n regions. A country with n districts, for example. For each of these regions we have two data points: y_i, $i \in 1, ..., n$, is the observed number of individuals suffering from a given disease in region i, and the expected number of cases in each region e_i.

The authors' model partitions a geographic region into k clusters, and assumes that the relative risk h_j, $j \in 1, ..., k$, is constant within each cluster. Let H_k denote the vector of relative risks. Clusters are denoted C_j, and the number of clusters k is unknown.

The authors assume each y_i has a Poisson distribution with $\lambda = e_i h_j$. Therefore, given k clusters, a corresponding cluster configuration, and a relative risk h_j for each cluster, the likelihood of observed data y_i and e_i is:

$$L(y|H_k, k, G_k) = \prod_{j=1}^{k} \prod_{i \in C_j} \frac{(e_i h_j)^{y_i}}{y_i!} e^{-e_i h_j}$$

Clustering

Consider again a geographic area divided into n regions. The authors define the distance between regions i and j as the minimum number of borders one must cross in order to travel from the interior of region i to the interior of region j.

To form a clustering configuration, select without replacement k regions from the set of n regions. Preserve the order in which this selection occurs. These selected regions are cluster centers, and they completely determine a cluster configuration. Let G_k represent this vector of cluster centers. A cluster configuration is formed by assigning each region that is not a cluster center to the cluster3 center that is of minimum distance away.

The distance between a non-center and two separate cluster centers could be equal. Such regions are assigned to the cluster center that appears first in G_k.

Priors

The authors assume a truncated geometric prior for k: $p(k) \propto (1-c)^k$ for some fixed c. In their analysis of the German dataset $c = .02$.

The vector G_k is assumed to be uniform across all possible lists of cluster centers. Given k there are $\binom{n}{k}$ possible selections for cluster centers, and for each of these $k!$ orderings. Thus $Pr(G_k|k) = \frac{1}{\binom{n}{k} * k!} = \frac{(n-k)!}{n!}$.

Heights follow a hierarchical structure. $H_k = \{h_1, ..., h_k\}$ are assumed to follow a lognormal distribution with $p(\mu)$ proportional to the whole real line and $p(\sigma^2)$ an inverse gamma distribution. In their analysis of the German data set, $p(\sigma^2) = \text{Inv-Gam}\ (1, .01)$.

MCMC

Each iteration of their MCMC executed one of six steps: birth, death, shift, switch, height, and hyperparameter. The probability each step is selected is $(.4, .4, .05, .05, .05, .05)$ respectively.

Birth: Given k cluster centers and a corresponding vector G_k there are $n - k$ regions that are not cluster centers. Choose one at random, and insert it at random into G_k. This is the proposed vector of cluster centers. Draw a proposed height h_j^* for this new region from a gamma distribtuion which the authors use to approximate the full conditional $p(h_j|\cdot)$

$$\alpha = \Big(\sum_{i \in C_j^*} y_i \Big) + \frac{(\mu + \frac{1}{2}\sigma^2)^2}{(e^{\sigma^2} * (e^{\sigma^2} - 1) * e^{2\mu})^2}$$

$$\beta = \Big(\sum_{i \in C_j^*} e_i \Big) + \frac{(\mu + \frac{1}{2}\sigma^2)}{(e^{\sigma^2} * (e^{\sigma^2} - 1) * e^{2\mu})^2}$$

Where C_j^* represents the new, birthed cluster. Only regions whose heights have changed will influence the likelihood. So, after some simplification, the birth step is accepted with probability:

$$r = \min\Big\{1, \frac{L(y_i^*|H_k^*, \mu, \sigma^2, k, G_k^*)}{L(y_i^*|H_k, \mu, \sigma^2, k, G_k)} * \frac{p(h^*|\mu, \sigma^2)}{q(h^*|\mu, \sigma^2)}\Big\}$$

Here $p(h^*)$ is the density of a lognormal distribution evaulauted at h^*, while $q(h^*$ is the density of the approximating gamma distribution at h^{**}. And y_i^* represents those clusters whose heights changed due to the birth of a new cluster.

Death: Select a cluster at random from your list of k cluster centers G_k. Eliminiate this cluster center and its corresponding height. You are left with a proposal configuration of $k-1$ clusters and $k-1$ heights.

This step is the reverse of a birth step, its acceptance probability is the inverse of the birth step acceptance probability.

Shift: In the list of cluster centers, let $n(G_k)$ be the number of cluster centers that are adjacent to regions which are not in G_k. These cluster centers are eligible for the shift step.

Select one of these eligible cluster centers at random and call it g_j. Let $m(g_j)$ be the number of clusters adjacent to g_j that are not cluster centers. Select one of these regions at random and call it g^*. Replace g_j with g^* to form a new cluster configuration G_k^*.

The shift step is accepted with probability

$$r = \min\Big\{1, \frac{L(y|H_k^*, \mu, \sigma^2, G_k^*)}{L(y|H_k^*, \mu, \sigma^2, G_k^*)} * \frac{n(G_k)m(g_j)}{n(G_k^*)m(g^*)}\Big\}$$

Switch: From the list of k cluster centers, select two at random. Switch their order in the list of cluster centers G_k to create a new vector G_k^*. Because order matters in this list, this will create a new cluster configuration C_j^*.

The authors do not make clear if the heights corresponding to these switched cluster centers are also swapped. However, in my runs of their MCMC leaving the vector of cluster centers in its original order yields a low acceptance rate, about .05, while switching the heights to follow the cluster centers yields an acceptance rate in their range, about .41. Therefore in my implementation I switched the heights corresponding to the switched cluster centers.

Note that the switch step is symmetric with respect to its jumping probability. The probability of jumping from G_k to G_k^* is identical to jumping from G_k^* to G_k. Therefore the probability of accepting the switch step depends only on its likelihood ratio.

Hyperparameters: The hyperparameters μ and σ^2 are updated by a Gibbs step:

$$\mu|\cdot \sim N\Big(\frac{1}{k}\sum_{j=1}^{k}\log(h_j), \frac{1}{k}\sigma^2\Big)$$

$$\sigma^2|\cdot \sim \text{IG}\Big(a + \frac{k}{2}, b + \frac{1}{2}\sum_{j=1}^{k}\big(\log(h_j) - \mu\big)^2\Big)$$