Bayesian complementary clustering, MCMC for Data Association and Anglo-Saxon Placenames

Giacomo Zanella
Department of Statistics, University of Warwick, UK.

A statistical problem from Anglo-Saxon history

1) HISTORIANS’ HYPOTHESIS
Anglo-Saxon settlements were organized into administrative clusters involving settlements with different names (i.e. functions) in each cluster.

2) RELEVANT QUESTIONS
Do the geographical locations of placenames support the “administrative clusters hypothesis”? What is the typical intra-cluster dispersion σ²? Which portion of settlements is clustered? Which placenames tend to cluster together?

3) MODEL REQUIREMENTS
Marked Point Process: marks (colors) represents placenames. Complementary clustering: only different colors in the same cluster.

Inferences on cluster partition:
\[ \pi(p|\sigma) \propto \pi(p)\pi(\sigma|p) \]
where \( p \) = \{p₁, ..., pₖ\} ↔ Observes points \( p \) = \{C₁, ..., Cₖ\} ↔ Partition of \( X \)

4) A RANDOM PARTITION MODEL
\[ \pi = \{p|\sigma\} \sim Dir(1, ..., 1) \]
\[ \sigma \sim Inv-Gamma(2, J) \]
\[ \pi(p|\sigma) \propto 1 \] for \( \pi(p) Y₁(\sigma) Y₂(\sigma) \cdot \cdot \cdot Yₖ(\sigma) \cdot p₁ \cdot p₂ \cdot \cdot \cdot pₖ \)

where \( Y₁(\sigma) \) is the number of points in a cluster of size \( i \) and \( Y₂(\sigma) = \# \{ p | Y₁(\sigma) = Y₁(p) \} \).

Likelihood: each cluster independent of the others with 1) a uniformly distributed center, 2) points distributed around the center according to a Gaussian distribution with variance \( \sigma^2 \).

Problem: the posterior of the partition \( \pi(p|rest) \) is intractable

Sampling from \( \pi(p|rest) \) is a Data Association Problem (e.g. Multi-target tracking). It can be expressed with hypergraphs:
\[ \pi(p|\sigma) \propto \pi(p)\pi(\sigma|p) \propto \prod_i \pi(Y_i|\sigma) \]
where \( Y_i|\sigma \) is clustered. Which placenames tend to cluster together?

\[ \pi(p|\sigma) \propto \pi(p)\pi(\sigma|p) \propto \prod_i \pi(Y_i|\sigma) \]

\[ \pi(X|p) = \frac{1}{Z(p)} \sum_{\sigma} \pi(p|\sigma) \prod_i \pi(Y_i|\sigma) \]
where \( Z(p) \) is the partition function.

\[ \pi(X|p) = \frac{1}{Z(p)} \sum_{\sigma} \pi(p|\sigma) \prod_i \pi(Y_i|\sigma) \]

Jerrum and Sinclair (1995) proved a polynomial time bound for approximate sampling from the 2-color version of \( \pi(p|rest) \).

Unfortunately such bound is unfeasible in practice. For the \( k \)-color case finding the maximum likelihood is NP-hard.

Sampling with 2 colors

Sample space: matchings of a weighted bipartite graph.
Distrbution: \( \pi(X) \) given in (1).

Algorithm: Metropolis-Hastings.

Proposal Distribution
1) Pick a red and a blue point, \( i \) and \( j \) according to a distribution \( q(i,j) \).
2) Propose the corresponding state.

What is the optimal choice for \( q(i,j) \)?

(A) Intuition: we want to have \( Q(i,j) = \pi(X|p) \)
This is obtained by \( q(i,j) \propto \pi(X|p) \)
\[ Q(X_{old}, X_{new}) = \frac{\pi(X_{new})}{\pi(X_{old})} \]
\[ = \frac{\pi(X_{new}|p)\pi(p)}{\pi(X_{old}|p)\pi(p)} \]
\[ = \frac{\pi(X_{new}|p)}{\pi(X_{old}|p)} \]

(B) Simulation: we compute three choices of \( q(i,j) \) on a bipartite graph with 100 red and 100 blue points.

1) \( q(i,j) \propto 1 \) 2) \( q(i,j) \propto \pi(X|p) \) 3) \( q(i,j) \propto \pi(X_{new}|p) \)

Multiple proposal scheme

Idea: propose a red-blue couple for each non-adjacent square of a grid.

Pros: simultaneous and independent proposal of many links (parallel computation).

Cons: truncation requirement (link length = \( c_{maks} \)).

Two Convergence Diagnostic tools

1) Multi-dimensional Gelman and Rubin diagnostic.
2) \( D = \text{sup} \{\| \hat{p}_i - \hat{p}_j \| | i \neq j \} \) where \( \hat{p}_i = \hat{p}_i \{i,j \in X \} \) and \( \hat{p}_i \) and \( \hat{p}_j \) are estimated by two independent MCMC runs.

Sampling with \( k \) colors

Sample space: matchings of a weighted \( k \)-partite hypergraph.

Distribution: \( \pi \) given in (1).

Algorithm: Mixture of \( k \) Metropolis-Hastings (analogous to Gibbs Sampling on the colors).

Multiple proposal scheme

1) Propose \( \hat{p}_i \) from \( x^{2D} \) and \( \hat{p}_j \).
2) Propose \( \hat{p}_i \) from \( x^{2D} \) and \( \hat{p}_j \).
3) Evaluate the proposed state \( \hat{p}_e \).

Remark: the projection of the target distribution \( \pi \) on the 2-colors space (\( \pi^{2D}(x, \hat{p}_e) \)) is the 2-color version of \( \pi \) itself.

Results on the Dataset

We consider three placenames. Firstly we analyze them pairwisely with our model and K-cross functions. Secondly we analyze the same placenames altogether with our model.

Future Steps

• Analyze the full dataset with our model.
• Use Bayesian model selection techniques to test for clustering \( p_1 \neq 1 \) against no interaction \( p_2 = 1 \).
• Provide a proof for the optimality of \( q(i,j) \propto \pi(X_{new}|p) \) in some simplified discrete setting.
• Consider influence of edge-effects and temporal inhomogeneity.

Acknowledgements

Prof. Wilfrid Kendall for the supervision, Prof. John Blair for the collaboration, CRISM and EPSRC for funding.