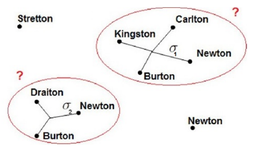
Bayesian complementary clustering, MCMC for Data Association and Anglo-Saxon Placenames

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 σ? Which portion of settlements is clustered? Which placenames tend to cluster together?

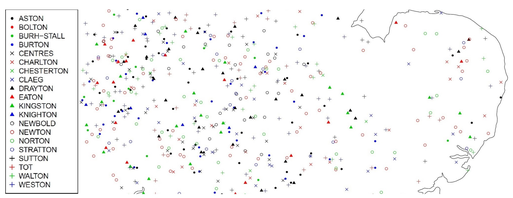
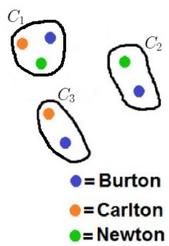


Figure: Anglo-Saxon settlements locations and names. The complete dataset is made of 1317 settlements.

3) MODEL REQUIREMENTS

Marked Point Process: marks (colors) represents placenames.
Complementary clustering: only different colors in the same cluster.
Inferences on cluster partition:

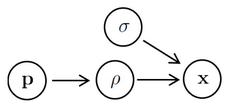


$$\pi(\rho|\mathbf{x}) \quad \propto \quad \pi(\rho)\pi(\mathbf{x}|\rho)$$

$\mathbf{x} = (x_1, ..., x_n) \leftrightarrow$ Observed points
$\rho = \{C_1, ..., C_{N(\rho)}\} \leftrightarrow$ Partition of $\mathbf{x}$

4) A RANDOM PARTITION MODEL

$k \leftrightarrow$ number of colors
$p_i \leftrightarrow$ probability of a point being in a cluster of size i
$\sigma \leftrightarrow$ intra-cluster dispersion



Prior distributions:

$$\mathbf{p} = (p_1, ..., p_k) \sim \text{Dir}(1, ..., 1) \qquad \sigma^2 \sim \text{Inv-Gamma}(2, \beta)$$

$$\pi(\rho|\mathbf{p}) \quad \propto \quad \frac{1}{\eta(\rho)} \frac{n(\mathbf{x})!}{Y_1(\rho)! \cdots Y_k(\rho)!} p_1^{Y_1(\rho)} \cdots p_k^{Y_k(\rho)},$$

where $Y_i(\rho)$ is the number of points in a cluster of size i and $\eta(\rho) = \#\{\tilde{\rho} \mid Y_1(\tilde{\rho}) = Y_1(\rho), ..., Y_k(\rho) = Y_k(\tilde{\rho})\}$.

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 σ^2.

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

Sampling from $\pi(\rho|rest)$ is a **Data Association Problem** (e.g. Multi-target tracking). It can be expressed with hypergraphs:

$\rho = \{\{1,2,4,7\},\{3\},\{5,6\}\} \leftrightarrow$ matching $X(\rho)$

$$\pi(X(\rho)) \;:=\; \pi(\rho|\mathbf{x},\sigma,\mathbf{p}) \;\propto\; \prod_{l=1}^{k} \left(\frac{(Y_l/l)!}{(Y_l)!}\right) \prod_{j=1}^{N(\rho)} \left(c_{|C_j|} \exp\left(-\frac{\delta^2_{C_j}}{2\sigma^2}\right)\right) \qquad (1)$$

where $\delta^2_{C_j} = \Sigma_{i \in C_j}(x_i - x_{C_j})^2$, and $c_{|C_j|} = \binom{k}{|C_j|}^{-1} \frac{p_{|C_j|}}{|C_j|\, W(2\pi\sigma^2)^{|C_j|-1}}$.

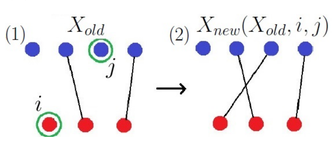
Jerrum and Sinclair (1995) proved a polynomial time bound for approximate sampling from the 2-color version of $\pi(\rho|rest)$. Unfortunately such bound is unfeasible in practice. For the k-color case finding the maximum likelihood is NP-hard.

Sampling with 2 colors

Sample space: matchings of a weighted bipartite graph.
Distribution: $\pi(X)$ 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.



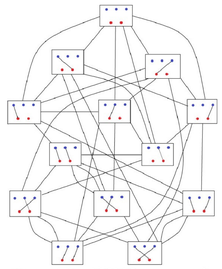


Figure: The MCMC as a graph.

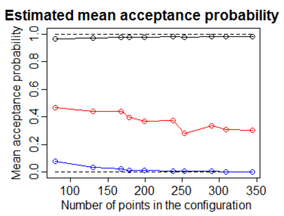
What is the optimal choice for $q(i,j)$?

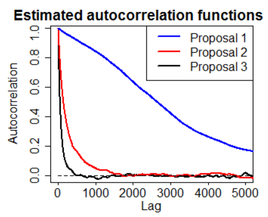
(A) Intuition: we want to have $\frac{Q(X_{old},X_{new})}{Q(X_{new},X_{old})} \approx \frac{\pi(X_{new})}{\pi(X_{old})}$.
This is obtained by $q(i,j) \propto \sqrt{\pi(X_{new})}$

$$\frac{Q(X_{old},X_{new})}{Q(X_{new},X_{old})} = \frac{q_{X_{old}}(i,j)}{q_{X_{new}}(i,j)} = \frac{\sqrt{\pi(X_{new})}/N(X_{old})}{\sqrt{\pi(X_{old})}/N(X_{new})} = \frac{\sqrt{\pi(X_{new})}}{\sqrt{\pi(X_{old})}} \frac{N(X_{new})}{N(X_{old})} \approx \frac{\sqrt{\pi(X_{new})}}{\sqrt{\pi(X_{old})}} \frac{\sqrt{\pi(X_{new})}}{\sqrt{\pi(X_{old})}} = \frac{\pi(X_{new})}{\pi(X_{old})},$$

where $N(X_{old}) = \Sigma_{i,j} \sqrt{\pi(X_{new}(X_{old}, i, j)}$.

(B) Simulations: we compare three choices of $q(i,j)$ on a bipartite graph with 100 red and 100 blue points.
1) $q(i,j) \propto 1$ 2) $q(i,j) \propto \pi(X_{new})$ 3) $q(i,j) \propto \sqrt{\pi(X_{new})}$

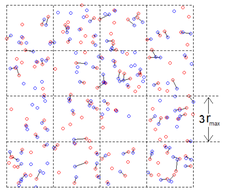




$q(i,j) \propto$	Integrated Autocorrelation Time	Time for G&R < 1.002	Time for D<0.05
$\sqrt{\pi(X_{new})}$	176.61	$20 \cdot 10^4$	$28 \cdot 10^4$
$\pi(X_{new})$	543.32	$49 \cdot 10^4$	$48 \cdot 10^4$

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$<r_{max}$).



Two Convergence Diagnostic tools

(1) Multi-dimensional Gelman and Rubin diagnostic.
(2) $D := \sup_{\{i,j\}} |\hat{p}^{(1)}_{ij} - \hat{p}^{(2)}_{ij}|$, where $p_{ij} = \mathbb{P}_\pi[\{i,j\} \in X]$ and $\hat{p}^{(1)}_{ij}$, $\hat{p}^{(2)}_{ij}$ are estimated by two independent MCMC runs.

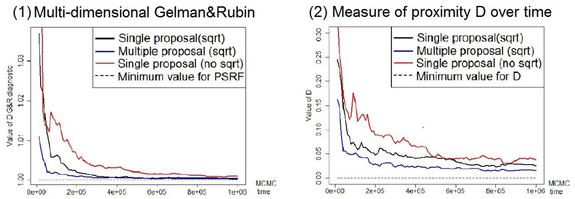
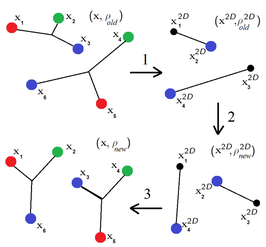


Figure : Left: G&R diagnostic for a 10-dimensional summary of 5 independent runs of the MCMC. Right: average of the value of D obtained by comparing 5 independent runs.

Sampling with k colors

Sample space: matchings of a weighted k-partite hypergraph.
Distribution: π given in (1).
Algorithm: Mixture of k Metropolis-Hastings (analogous to Gibbs Sampling on the colors).

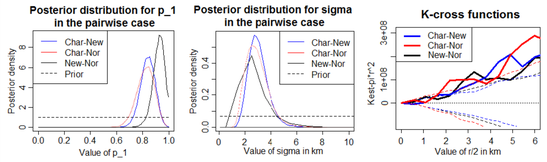
Proposal Distribution
Given $\mathbf{x}$, ρ_{old} and a color i:
1) Evaluate $\mathbf{x}^{2D}$ (keep the points of color i and replace the others with their cluster centroids).
2) Propose ρ^{2D}_{new} with the 2-color algorithm;
3) Evaluate the proposed state ρ_{new} from $\mathbf{x}^{2D}$ and ρ^{2D}_{new}.



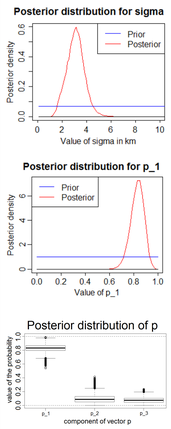
Remark: the projection of the target distribution π on the 2-colors space (x^{2D}, ρ^{2D}) is the 2-color version of π itself!

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.



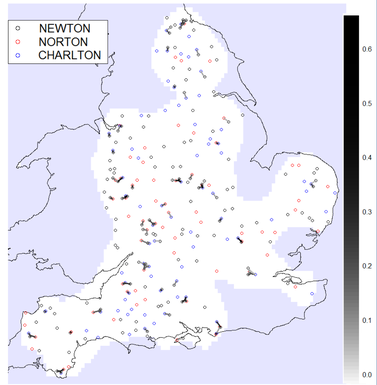


Figure: the intensity of gray represents the estimated posterior probability of the cluster being present.

- Mild evidences in favor of interaction ($p_1 \neq 1$). When more then two colors at the same time are considered evidences may increase (need to be explored more).
- With no prior information we estimate σ to lie in the range 3 km - 4 km (agrees with historian's hypothesis).

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 ($p_1 = 1$).
- Provide a **proof for the optimality of** $q(i,j) \propto \sqrt{\pi(X_{new})}$ in some simplified discrete setting.
- Consider influence of **edge-effects** and **spatial inhomogeneity**.

Acknowledgements

Prof. Wilfrid Kendall for the supervision, Prof. John Blair for the collaboration, CRiSM and EPRSC for funding.

EPSRC
Warwick Statistics