Centered Partition Process: Informative Priors for Clustering

Sally Paganin* **Amy H. Herring†** **David B. Dunson†**
paganin@stat.unipd.it
University of Padova* Italy. Duke University†, Durham, USA.
Università degli Studi di Padova

[1] Introduction

- Clustering is one of the building blocks in Bayesian nonparametric modeling.
- Discrete nonparametric priors typically induce a **latent partitioning** c of the data $\Rightarrow$ described by mean of an **Exchangeable Partition Probability Function** (EPPF)

How to incorporate concrete prior knowledge into the clustering process?

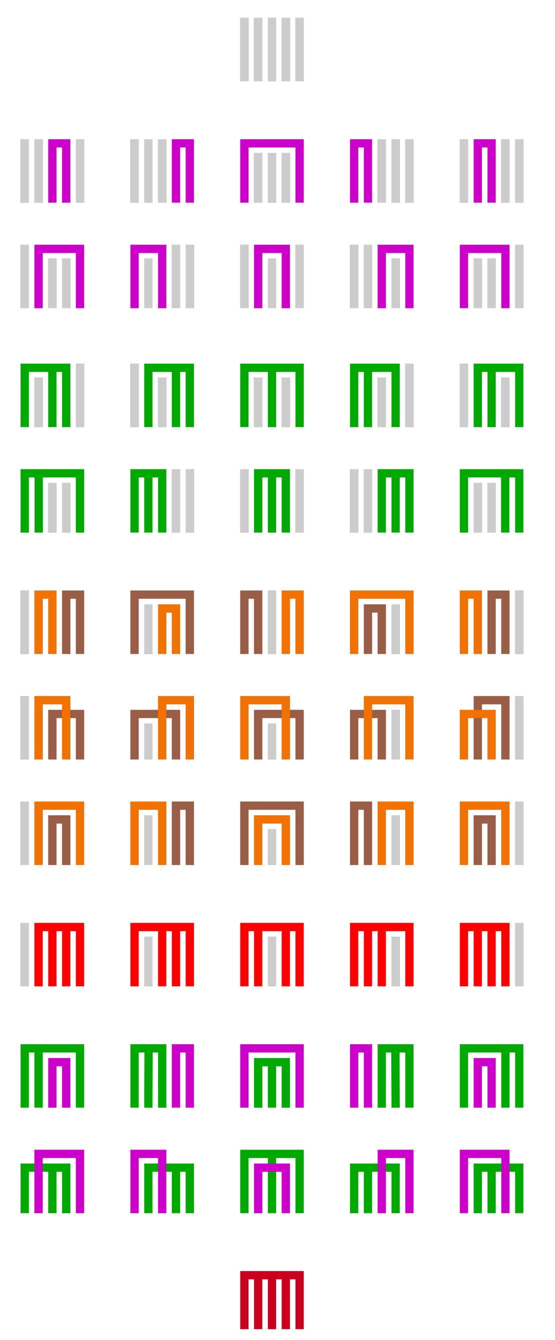
Motivation arises from an epidemiological problem in which experts provide grouping information on the basis of biological knowledge

- n different diseases, indexed by $j \in \{1, \dots, n\}$, varying from common to rare ones
- access to an informed prior guess $\boldsymbol{c}_0$ based on biological knowledge
- interest in inserting this information in the model to perform data analysis

Diseases in the same group will have similar coefficients in logistic regression analysis relating exposure factors to the risk of developing the disease

[2] Set Partitions

A **set partition** $\boldsymbol{c}$ of an integer $[n]$ is a collection of non-empty disjoint subsets $\{B_1, B_2, \dots, B_K\}$ such that $\cup_i^K B_i = [n]$.



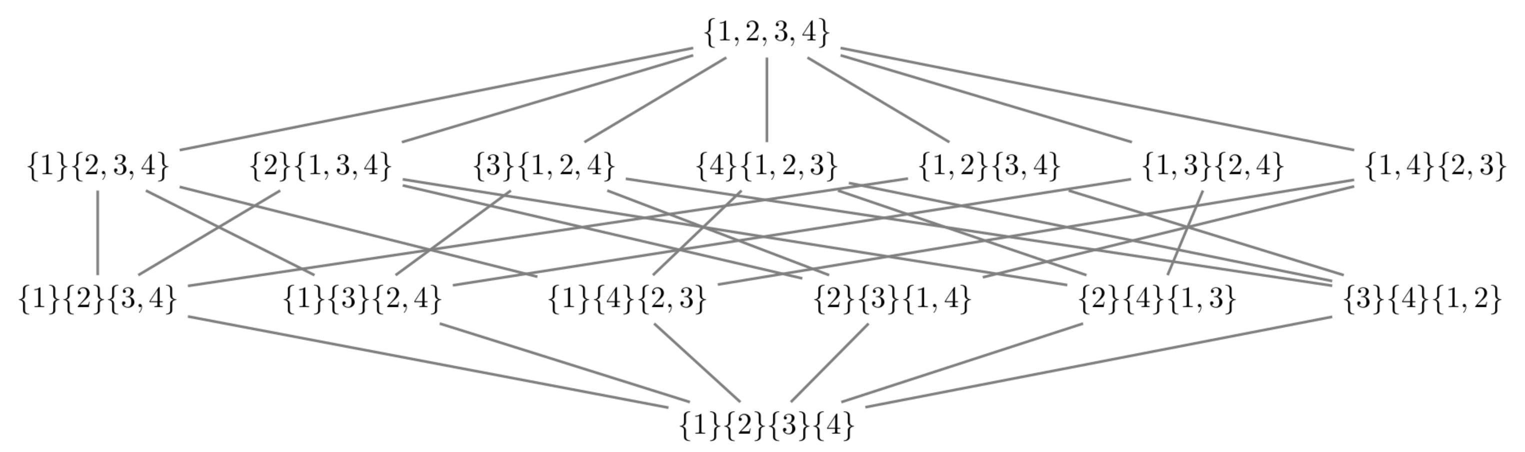
First know application in 1500 AD, in the context of popular games in Japanese upper-class society (tea ceremonies, Genji-ko game)[1]

Which of the 5 incense sticks are the same?

There are 52 possible answers!

Set partitions space Π_n

- Number of partitions of $[n]$ into k blocks $\Rightarrow$ Stirling numbers $S(n,k)$
- Total number of set partitions $\Rightarrow$ Bell number $\mathcal{B}_n = \sum_{k=1}^n S(n,k)$
- Blocks sizes $\{|B_1|, \dots, |B_K|\}$ $\Rightarrow$ individuate an **integer partition**, a set of positive integers $\{\lambda_1, \dots, \lambda_K\}$ such that $\sum_{i=1}^K \lambda_i = n$
- Such space $(\Pi_n, \leq)$ endowed with a relation of set containment is a ***partially ordered set*** (poset) which allows to represent the space by means of the ***Hasse diagram***
- The space Π_n is also a lattice, with upper bound $\mathbf{1} = \{n\}$ and lower bound $\mathbf{0} = \{1\}\{2\} \dots \{n\}$



Hasse diagram for the lattice of set partitions of 4 elements. A line is drawn when a partition is covered by the other. For example $\{1\}\{2,3,4\}$ is connected with 3 partitions obtained by splitting the block $\{2,3,4\}$ in any possible way.

[3] Centered Partition process

The Centered Partition process defines a probability distribution over the space of set partitions as

$$p(\boldsymbol{c}, \boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c}) e^{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)} \quad (1)$$

- $p_0(\boldsymbol{c})$ indicates a **baseline distribution** (EPPF) on the set partitions space
- $d(\boldsymbol{c}, \boldsymbol{c}_0)$ **distance** measuring how much a generic partition $\boldsymbol{c}$ is far form the base one $\boldsymbol{c}_0$ $\Rightarrow$ ideally a suitable metric on the set partitions lattice
- ψ **penalization parameter** controlling for the centering $\psi = 0; p(\boldsymbol{c}, \boldsymbol{c}_0, \psi) \to p_0(\boldsymbol{c})$; $\psi \to \infty; p(\boldsymbol{c}, \boldsymbol{c}_0, \psi) = \delta_{\boldsymbol{c}_0}$

Consider sets of partitions with a fixed distance from $\boldsymbol{c}_0$

$$s_l(\boldsymbol{c}_0) = \{\boldsymbol{c} \in \Pi_n : d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l\}, \quad l = 0, 1, \dots, L \quad (2)$$

- L the maximum possible distance from $\boldsymbol{c}_0$
- $\delta_0 = 0$, hence $s_0(\boldsymbol{c}_0)$ is set of partitions differing from $\boldsymbol{c}_0$ by a permutation of the cluster labels.

Analytic form for (1)

$$p(\boldsymbol{c}, \boldsymbol{c}_0, \psi) = p_0(\boldsymbol{c}) \frac{e^{-\psi s_l(\boldsymbol{c}_0)}}{\sum_{m=1}^L n_m e^{-\psi s_m(\boldsymbol{c}_0)}}, \quad \text{for } \boldsymbol{c} \in s_l(\boldsymbol{c}_0)$$

$n_m = |s_m(\boldsymbol{c}_0)|$ denotes the cardinality of the set $s_m(\boldsymbol{c}_0)$ typically not possible to be calculated analytically
$\Rightarrow$ but can nonetheless be used in Bayesian models relying on Monte Carlo methods.

Baseline EPPF

Come from different process depending on the assumed exchangeable behavior

- Uniform $p_0 = 1/\mathcal{B}_n$
- Dirichlet Process $p_0 \propto \alpha^{|\boldsymbol{c}|} \prod_{j=1}^{|\boldsymbol{c}|} (|B_j| - 1)!$
- generic Gibbs-type priors

Choosing the distance

We employed the **Variation of information** [3]

- Entropy-based metric $VI(\boldsymbol{c}, \boldsymbol{c}') = -H(\boldsymbol{c}) - H(\boldsymbol{c}') + 2H(\boldsymbol{c}, \boldsymbol{c}')$
- Alignment properties
- Easy to compute (block dependent)

Tuning parameter ψ

Depends on n and *where* $\boldsymbol{c}_0$ is located in the space

- Exact values computed up to $n = 8$
- For $n > 8$ we consider prior calibration using a Monte Carlo estimate

[4] Logistic regression borrowing

Model specification

- $j = 1, \dots, n$ diseases, $i = 1, \dots, n_j$ observations related to the disease
- $y_i^{(j)} = 1$ if observation i has the disease j while $y_i^{(j)} = 0$ is a control
- $\mathbf{X}^{(j)}$ data matrix associated to diasease j, with each row $\mathbf{x}_i^{(j)} = (x_{i1}^{(j)}, \dots, x_{ip}^{(j)})$ being the observed values for ith observation of p dichotomous variables.

$$\begin{aligned} y_i^{(j)} &\sim Ber(\pi_i^{(j)}) \\ \pi_i^{(j)} &= \alpha^{(j)} + \mathbf{x}_i^{(j)T} \boldsymbol{\beta}^{c_j} \\ (\boldsymbol{\beta}^{(j)} | c_j = h) &= \boldsymbol{\beta}^{(h)}, \qquad j = 1, \dots, n, \\ \boldsymbol{\beta}^{(h)} &\sim N_p(\mathbf{b}, \mathbf{Q}) \qquad h = 1, \dots, H, \\ \boldsymbol{c} = (c_1, \dots, c_n) &\sim \mathrm{CP}(\boldsymbol{c}_0, \psi, p_0) \end{aligned}$$

Posterior computation

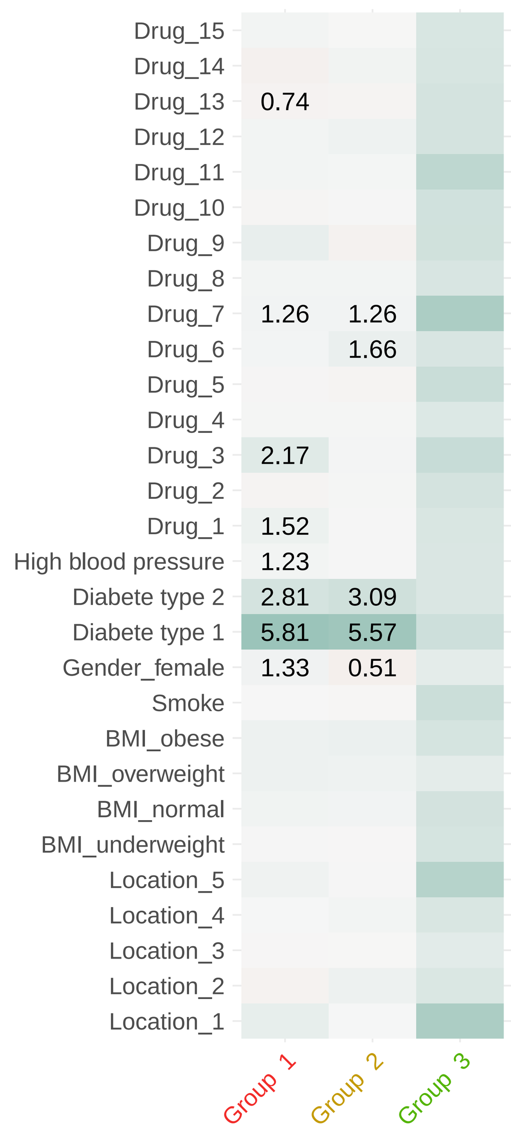
Posterior distributions are obtained via **MCMC** algorithm, with key steps

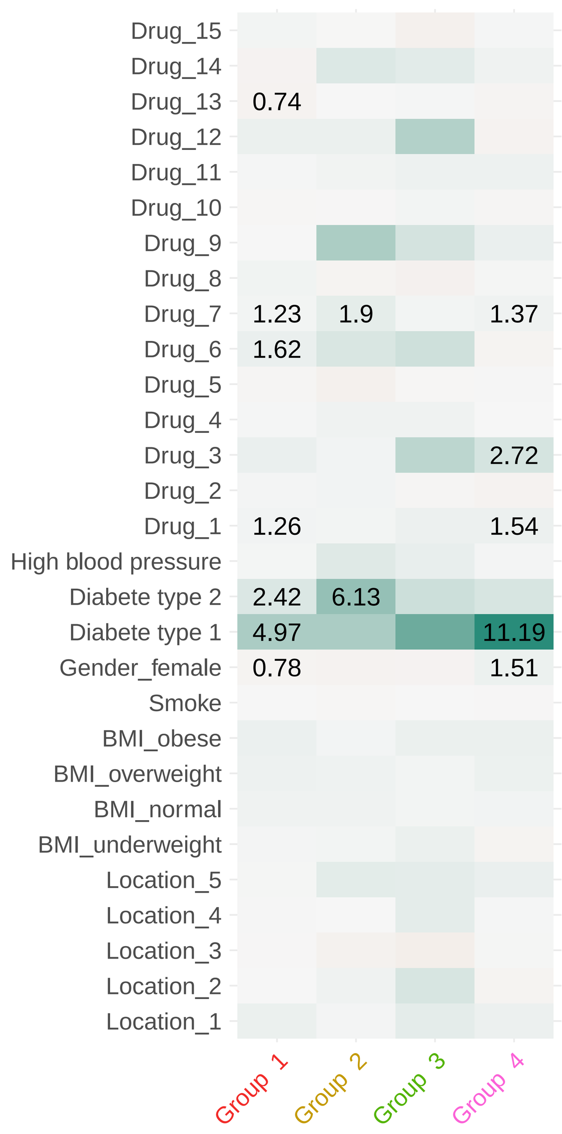
- A **Polya-gamma data augmentation**[4] for Bayesian logistic regression, introducing latent variables $\omega_i^{(j)} \sim PG(1, \alpha^{(j)} + \mathbf{x}_i^{(j)T} \boldsymbol{\beta}^{c_j})$
- Class allocation step involving the CP process penalization, easily adapt widely used sampling algorithm (eg. marginal sampling, split-merge moves [2])

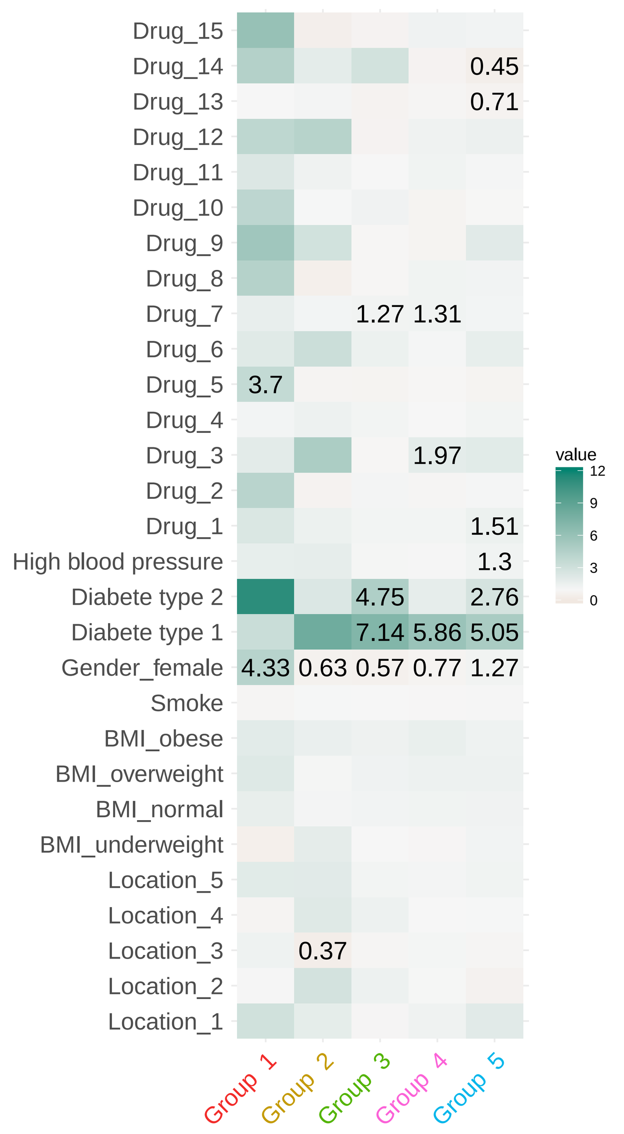
[5] Grouping diseases

- Data comprises $n = 26$ different diseases, with classification providing 6 groups with sizes $\{8, 7, 4, 4, 2, 1\}$
- Around 80 exposures, comprising demographics, drugs, habits
- Considered different values for $\psi \in \{300, 700, 1110\}$ (note that $\mathcal{B}_{26} = O(10^{19})$)

		1	2	3	4	5	6	
ψ_1	Group 1	0	1	0	0	4	7	12
	Group 3	1	0	8	4	0	0	13
	Group 5	1	0	0	0	0	0	1
ψ_2	Group 2	0	0	0	4	0	7	11
	Group 3	0	0	8	0	0	0	8
	Group 9	2	0	0	0	0	0	2
	Group 10	0	1	0	0	4	0	5
ψ_3	Group 1	0	1	0	0	0	0	1
	Group 3	2	0	0	0	0	0	2
	Group 6	0	0	8	0	0	0	8
	Group 9	0	0	0	4	4	0	8
	Group 10	0	0	0	0	0	7	7







References

[1] Knuth, D. E. (2006) Generating all trees - history of combinatorial generation. *The art of computer programming. Vol. 4, Fasc. 4*.

[2] Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

[3] Meilă, M. (2007). Comparing clusterings - an information based distance *Journal of multivariate analysis* **98**, 873–895.

[4] Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.