Prior-Driven Cluster Allocation in Bayesian Mixture Models

Sally Paganin
sally.paganin@berkeley.edu

JSM 2020

August 03, 2020

Berkeley
UNIVERSITY OF CALIFORNIA

Amy Herring
Duke University

David Dunson
Duke University

Andrew Olshan
UNC at Chapel Hill

Introduction

Clustering is one of the canonical data analysis goal in statistics

- **Distance based methods**: distance metric between data points
- **Model-based clustering**: rely on discrete mixture models

Bayesian perspective : allow to incorporate prior information

Introduction

Clustering is one of the canonical data analysis goal in statistics

- **Distance based methods**: distance metric between data points
- **Model-based clustering**: rely on discrete mixture models

Bayesian perspective : allow to incorporate prior information

What if, we have prior information on the clustering itself?

Introduction

Clustering is one of the canonical data analysis goal in statistics

- **Distance based methods**: distance metric between data points
- **Model-based clustering**: rely on discrete mixture models

Bayesian perspective : allow to incorporate prior information

What if, we have prior information on the clustering itself?

Motivating application - **Birth defects data**

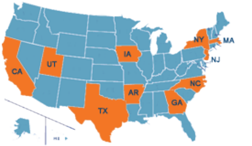
- Relate exposure factors to the development risk of a defect
- **Prior information** available (biology/expert's judgments)

➔ We aim to provide methods to facilitate data-adaptive clustering, both using **information in the data** and **external knowledge**.

National Birth Defect Prevention Study

- **Population-based case-control study**
  - ➔300 controls/100 cases per year since 1997
  - ➔monthly n. of controls $\propto$ n. of births previous year
- **Cases** (37 major birth defect)
  - ➔Birth defects surveillance system
  - $+$clinical genetist review
  - ➔Cases with known etiology were excluded
- **Controls**
  - ➔Non-malformed live birth
  - ➔Birth certificates or hospital delivery records
- **Data collection**
  - ➔CATI (English/Spanish) within 24 months

national • birth • defects • prevention • study



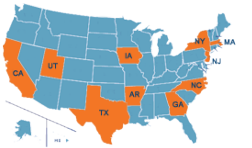
http://www.nbdps.org/

National Birth Defect Prevention Study

- **Population-based case-control study**
  - →300 controls/100 cases per year since 1997
  - →monthly n. of controls $\propto$ n. of births previous year
- **Cases** (37 major birth defect)
  - →Birth defects surveillance system
  - +clinical genetist review
  - →Cases with known etiology were excluded
- **Controls**
  - →Non-malformed live birth
  - →Birth certificates or hospital delivery records
- **Data collection**
  - →CATI (English/Spanish) within 24 months

nbdps

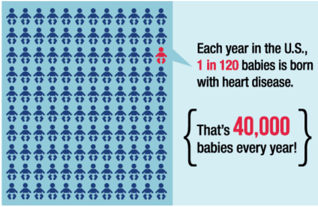
national • birth • defects • prevention • study



http://www.nbdps.org/

We focus on the **Congenital Heart Defects** (**CDH**) which are problems in the structure of the heart that are present at birth.

Congenital Heart Defects



Clinical importance

priority in public health

→most frequent class of defects

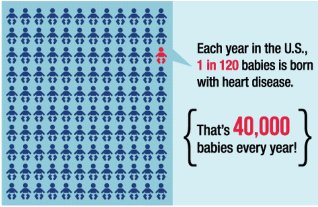
→high impact on pediatric mortality

Statistical relevance: challenge in birth defects modeling

→Most defects are too rare for individual study

→Difficult to determine how best to group birth defects

Congenital Heart Defects



Clinical importance

priority in public health

→most frequent class of defects

→high impact on pediatric mortality

Statistical relevance: challenge in birth defects modeling

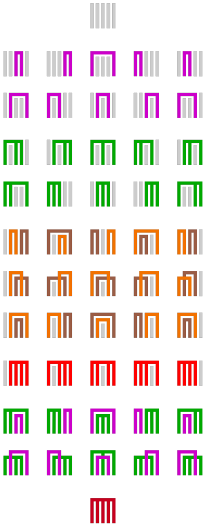
→Most defects are too rare for individual study

→Difficult to determine how best to group birth defects

Experts have provided a **mechanistic classification** of the defects

- → relies on biological knowledge and embryologic development
- → **translates in a prior guess c_0 for the clustering**

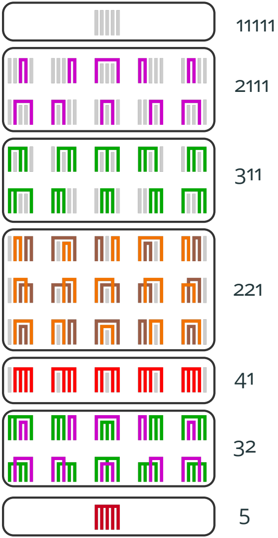
Set partitions



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

- Number of partitions of $[n]$ into k blocks
  →Stirling numbers $S(n, k)$
- Total number of set partitions
  →Bell number $\mathcal{B}_n = \sum_{k=1}^{n} S(n, k)$

Set partitions



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

- Number of partitions of $[n]$ into k blocks
  →Stirling numbers $S(n, k)$
- Total number of set partitions
  →Bell number $\mathcal{B}_n = \sum_{k=1}^{n} S(n, k)$
- **Configuration** $\boldsymbol{\lambda} = \{|B_1|, \ldots, |B_K|\}$
  →sequence of block cardinalities
  →individuate an **integer partition**, a set of positive integers $\{\lambda_1, \ldots, \lambda_K\}$ such that $\sum_{i=1}^{K} \lambda_i = n$

Modeling birth defects

- $i = 1, \ldots, N$ heart defects, $j = 1, \ldots, n_i$ observations
- $y_{ij} = 1$ if observation j has the b.d. i while $y_{ij} = 0$ is a control
- $\mathbf{x}_{ij}^T = (x_{ij1}, \ldots, x_{ijp})$ observed values for p dichotomous variables

Grouped logistic regression

$$
\begin{aligned}
y_{ij} &\sim Ber(\pi_{ij}) \qquad \text{logit}(\pi_{ij}) = \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i}, \quad j = 1, \ldots, n_i, \\
\alpha_i &\sim \mathcal{N}(a_0, \tau_0^{-1}) \qquad \boldsymbol{\beta}_{c_i} | \boldsymbol{c} \sim \mathcal{N}_p(\mathbf{b}, \mathbf{Q}) \quad i = 1, \ldots, N,
\end{aligned}
$$

Bayesian framework: assign a prior probability $p(\boldsymbol{c})$

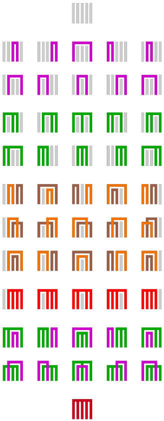
➔*Exchangeable Partition Probability Function* (EPPF)

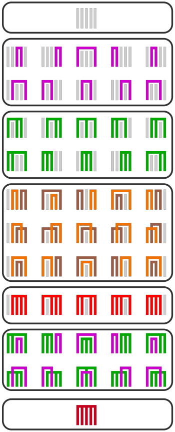
Uniform distribution

$p(\boldsymbol{c}) \propto 1/\mathcal{B}_N$

Dirichlet Process: $p(\boldsymbol{c}) \propto \prod_{i=1}^{K}(|B_i| - 1)!$

Pitman-Yor Process: $p(\boldsymbol{c}) \propto \prod_{i=1}^{K}(1-\sigma)_{|B_i|}$





How to account for c_0?

Base idea: penalize a baseline EPPF in order to center the prior distribution on the given partition $\boldsymbol{c}_0$

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c}) \exp\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\} \tag{1}$$

- $p_0(\boldsymbol{c})$ indicates a **baseline distribution** (EPPF) on Π_N
- $d(\boldsymbol{c}, \boldsymbol{c}_0)$ a suitable **distance** between partitions
  ➔ideally a metric on the set partitions lattice
- ψ **penalization parameter** controlling for the centering
  ➔ $\psi = 0 \quad p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \to p_0(\boldsymbol{c})$
  ➔ $\psi \to \infty \quad p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) = \delta_{\boldsymbol{c}_0}$

How to account for c_0?

Base idea: penalize a baseline EPPF in order to center the prior distribution on the given partition $\boldsymbol{c}_0$

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c}) \exp\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\} \tag{1}$$

- $p_0(\boldsymbol{c})$ indicates a **baseline distribution** (EPPF) on Π_N
- $d(\boldsymbol{c}, \boldsymbol{c}_0)$ a suitable **distance** between partitions
  ➔ideally a metric on the set partitions lattice
- ψ **penalization parameter** controlling for the centering
  ➔ $\psi = 0 \quad p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \to p_0(\boldsymbol{c})$
  ➔ $\psi \to \infty \quad p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) = \delta_{\boldsymbol{c}_0}$

Choice of the distance ➔ Variation of information [Meila (2007)]

- $\text{VI}(\boldsymbol{c}, \boldsymbol{c}') = -H(\boldsymbol{c}) - H(\boldsymbol{c}') + 2H(\boldsymbol{c} \wedge \boldsymbol{c}')$
- $H(\cdot)$ information entropy
- metric on set partition lattice

Centered Partition Processes

Define sets of partitions with distance δ_l from $\boldsymbol{c}_0$ and configuration $\boldsymbol{\lambda}_m$

$$s_{lm}(\boldsymbol{c}_0) = \{\boldsymbol{c} \in \Pi_N : d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l, \boldsymbol{\Lambda}(\boldsymbol{c}) = \boldsymbol{\lambda}_m\}$$

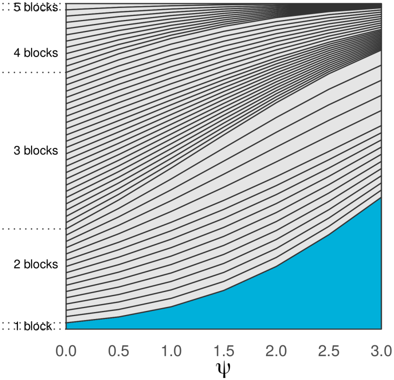
for $l = 0, \ldots, L$ and $m = 1, \ldots, M$.

Centered Partition Processes - analytic form

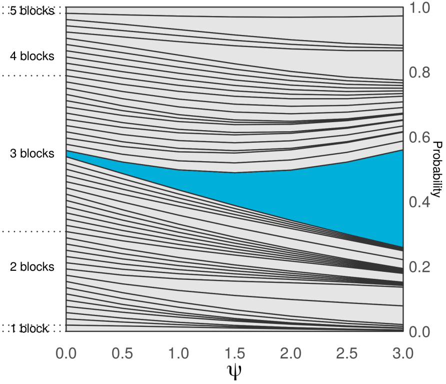
$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) = \frac{g(\boldsymbol{\lambda}_m)e^{-\psi\delta_l}}{\sum_{u=0}^{L}\sum_{v=1}^{M} |s_{uv}(\boldsymbol{c}_0)| g(\boldsymbol{\lambda}_v) e^{-\psi\delta_u}}, \quad \text{for } \boldsymbol{c} \in s_{lm}(\boldsymbol{c}_0)$$

- $g(\cdot)$ function of the configuration $\boldsymbol{\Lambda}(\boldsymbol{c})$
  → e.g. Uniform $g(\boldsymbol{\Lambda}(\boldsymbol{c})) = 1$, DP $g(\boldsymbol{\Lambda}(\boldsymbol{c})) = \alpha^K \prod_{j=1}^{K} \Gamma(\lambda_j)$
- $|\cdot|$ cardinality of the set $s_{lm}(\boldsymbol{c}_0)$, not analytically tractable
  → but can nonetheless be used in Bayesian models relying on Monte Carlo methods

CP Process - Uniform EPPF

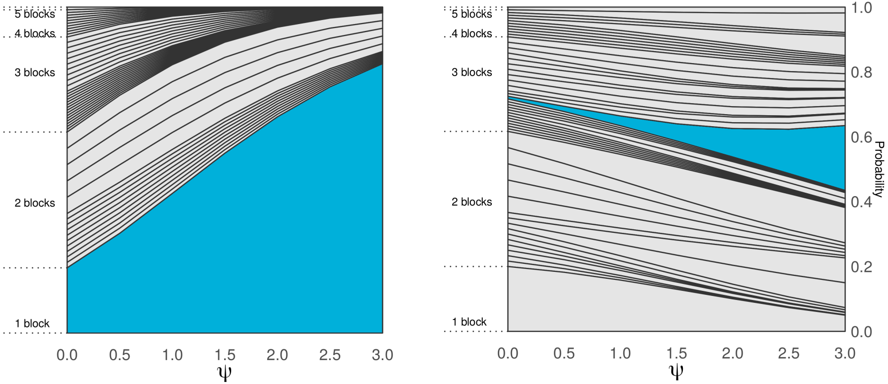


$$c_0 = \{1, 2, 3, 4, 5\}$$



$$c_0 = \{1, 2\}\{3, 4\}\{5\}$$

CP Process - DP EPPF ($\alpha = 1$)



$$\boldsymbol{c}_0 = \{1, 2, 3, 4, 5\}$$

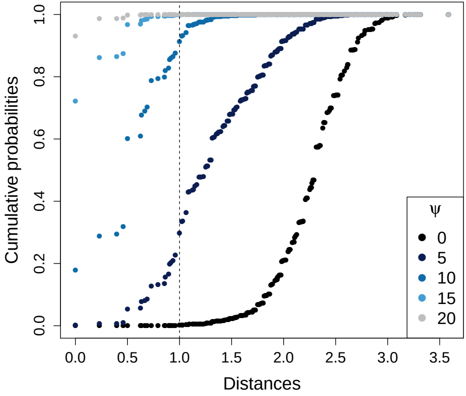
$$\boldsymbol{c}_0 = \{1, 2\}\{3, 4\}\{5\}$$

Prior calibration

We consider to estimate the distribution of **distance** $\delta \in \{\delta_l\}_{l=0}^{L}$

$$p(\delta = \delta_l) = \frac{\sum_{m=1}^{M} n_{lm} g(\boldsymbol{\lambda}_m) e^{-\psi\delta_l}}{\sum_{u=0}^{L} \sum_{v=1}^{M} n_{uv} g(\boldsymbol{\lambda}_v) e^{-\psi\delta_u}}$$

- **Monte Carlo procedure**
  ➔ uniform sampler on the set partition space Π_N [Stam (1983)]
- **Deterministic local search**
  ➔ for small values of the distance $\delta \in \{\delta_0, \ldots, \delta_{L^*}\}$
  ➔ greedy search algorithm



Modeling birth defects

$N = 26$ birth defects, $4,047$ cases, $8,125$ controls, 90 potential risk factors

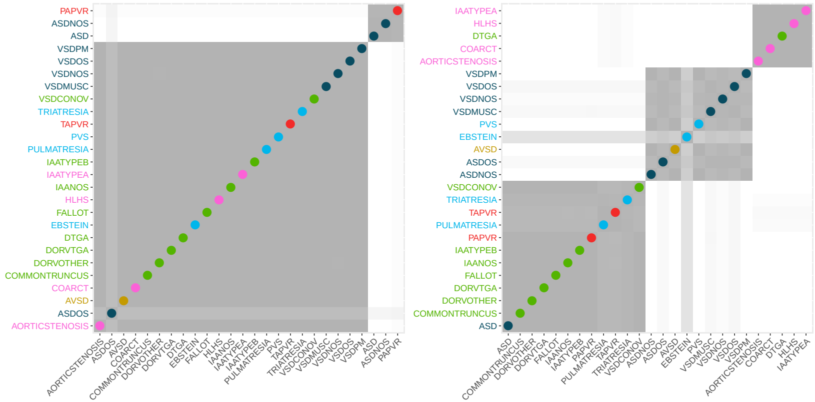
$$y_{ij} \sim Ber(\pi_{ij}) \qquad \text{logit}(\pi_{ij}) = \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i}, \quad j = 1, \ldots, n_i,$$
$$\alpha_i \sim \mathcal{N}(a_0, \tau_0^{-1}) \qquad \boldsymbol{\beta}_{c_i} | \boldsymbol{c} \sim \mathcal{N}_p(\mathbf{b}, \mathbf{Q}) \quad i = 1, \ldots, N,$$
$$p(\boldsymbol{c}) \sim CP(\boldsymbol{c}_0, \psi, p_0(\boldsymbol{c})) \qquad p_0(\boldsymbol{c}) \propto \alpha^K \prod_{k=1}^{K} (\lambda_k - 1)!$$

from the prior calibration: $\psi = 40$ (90% partitions with $d = 0.8$ ($d_{max} = 4.70$)

Posterior estimation (MCMC)

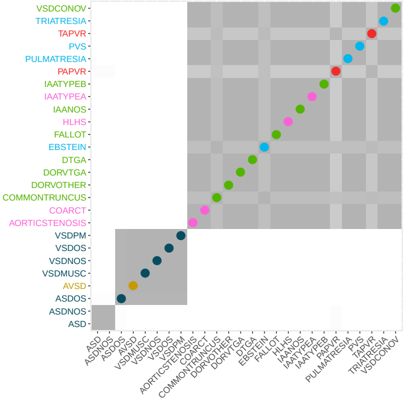
- A **Polya-gamma data augmentation** for Bayesian logistic regression, introducing latent variables $\omega_i^{(j)} \sim PG(1, \alpha^{(j)} + \mathbf{x}_i^{(j)T} \boldsymbol{\beta}^{c_j})$
- Class allocation step involving prior penalization easily adapt marginal sampling for DP process

Clustering results

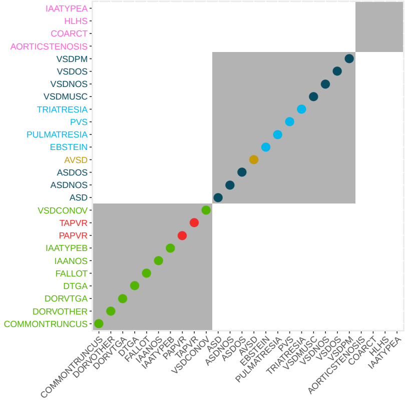


(a) ψ = 0, VI$(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 2.43$ **(b)** ψ = 40, VI$(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 1.78$

Clustering results

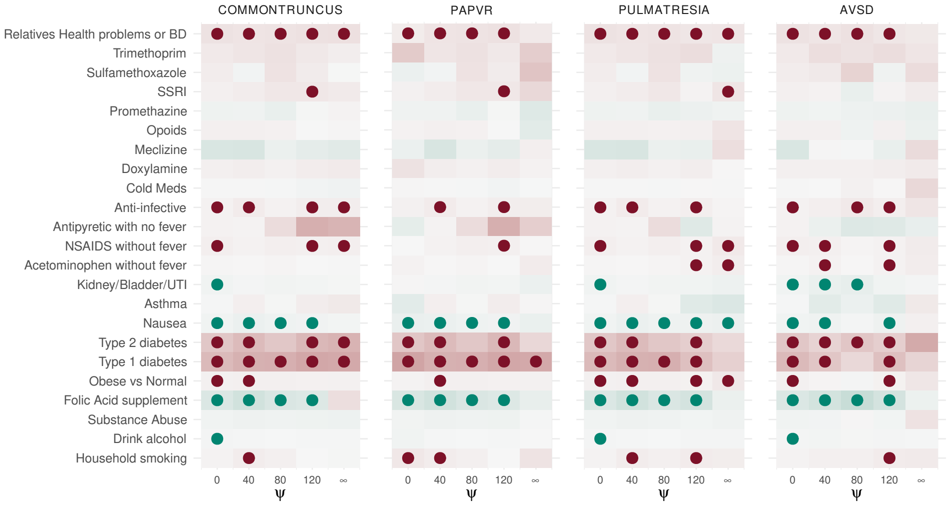


(c) ψ = 80, VI$(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 1.65$



(d) ψ = 120, VI$(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 0.86$

Exposure effects



Future work

Data analysis

- Variable selection in order to account for shared effects.
- Inclusion of information favoring relation between specific outcomes and exposure factors.

Methodology

- Building prediction rules for new observations/clusters.
- Formalize inclusion of partial information, number/sizes of clusters.

Software

- Provide sampling methods via NIMBLE

Thanks!

Centered Partition Processes: Informative Priors for Clustering.
Paganin S., Herring A. H., Olshan A. F. & Dunson B. D. (2020)
Bayesian Analysis (Advanced publication)

sally.paganin@berkeley.edu

@sampling_sally

salleuska

`https://salleuska.github.io/`

References i

Hartigan, J. A. (1990).
Partition models
Commun. Statist. A **19**, 2745–2756.

Meila M. (2007).
Comparing clusterings - an information based distance.
J. of Mult. Analysis **98**, 873–895.

Müller, P., Quintana, F. & Rosner, G. L. (2011).
A Product Partition Model With Regression on Covariates.
J. Comput. Graph. Statist. **20**, 260–278.

Neal, R. M. (2000).
Markov chain sampling methods for Dirichlet process mixture models
J. Comput. Graph. Statist. **9**, 249–265.

Park, J.-H. & Dunson, D. B. (2010).
Bayesian Generalize Product Partition Models.
Stat. Sin. **20**, 1203–1226.

References ii

RODRIGUEZ, A. & DAVID B. D. (2011).
Nonparametric Bayesian models through probit stick-breaking processes
Bayesian analysis (Online) 6.1.

STAM, A.J. (1983).
Generation of a random partition of a finite set by an urn model
J. of Comb. Theory, Series A **35**, 231–240.