UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXI

Prior-driven Cluster Allocation in Bayesian Mixture Models

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Bruno Scarpa

Co-supervisore: Prof. Amy H. Herring

Dottoranda: Sally Paganin

Abstract

There is a very rich literature proposing Bayesian approaches for clustering starting with a prior probability distribution on partitions. Most approaches assume exchangeability, leading to simple representations of such prior in terms of an Exchangeable Partition Probability Function (EPPF). Gibbs-type priors encompass a broad class of such cases, including Dirichlet and Pitman-Yor processes. Even though there have been some proposals to relax the exchangeability assumption, allowing covariate-dependence and partial exchangeability, limited consideration has been given on how to include concrete prior knowledge on the partition. Our motivation is drawn from an epidemiological application, in which we wish to cluster birth defects into groups and we have a prior knowledge of an initial clustering provided by experts. The underlying assumption is that birth defects in the same group may have similar coefficients in logistic regression analysis relating different exposures to risk of developing the defect. As a general approach for including such prior knowledge, we propose a Centered Partition (CP) process that modifies a base EPPF to favor partitions in a convenient distance neighborhood of the initial clustering. This thesis focus on providing characterization of such new class, along with properties and general algorithms for posterior computation. We illustrate the methodology through simulation examples and an application to the motivating epidemiology study of birth defects.

Sommario

Esiste una letteratura molto vasta che propone approcci di tipo bayesiano per l'analisi di raggruppamento (clustering), a partire da una distribuzione di probabilià a priori definita sullo spazio delle possibili partizioni. La maggior parte degli approcci si basa sull'assunzione di scambiabilità, tale da permettere una semplice rappresentazione della distribuzione a priori in termini di una funzione nota come *Exchangeable Partition Probability Function* (EPPF). Le distribuzioni a priori di tipo Gibbs rappresentano un'ampia classe di priori aventi tale caratterizzazione e include i processi di Dirichlet e Pitman-Yor. Sono state fatte in letteratura alcune proposte volte a rilassare l'assunzione di scambiabilità, inserendo delle dipendenze dalle covariate o proponendo distribuzioni basate sul concetto di scambiabilità parziale. Tuttavia è stata data limitata considerazione alla definizione di metodologie in grado di includere una concreta informazione a priori sulla partizione dei dati. La nostra motivazione deriva da un'applicazione in ambito epidemiologico, nella quale è di interesse ottenere un raggruppamento di dati riguardati malformazioni congenite in nuovi nati, e per la quale è disponibile un raggruppamento iniziale fornito dagli esperti del settore. L'assunzione di base è data dal fatto che malformazioni nello stesso gruppo sono probabilmente associate a simili coefficienti di regressione, quando derivanti da un'analisi di tipo logistico che mette in relazione le diverse esposizioni con il rischio di sviluppare un difetto. In questo lavoro, proponiamo un metodo generale per includere una tale informazione a priori, definito come processo *Centered Partition*, volto a modificare una distribuzione di base di tipo EPPF in modo da favorire le partizioni in un desiderato intervallo di distanza rispetto alla partizione iniziale. Questa tesi si concentra sul fornire una caratterizzazione di questa nuova classe di distribuzioni a priori, descrivendone le proprietà e gli algoritmi di calcolo. La metodologia è illustrata per mezzo di simulazione e applicazione ai dati sulle malformazioni nei neonati.

Acknowledgements

I have few thoughts about people I met during this PhD that I guess belong to the acknowledgments.

First of all Bruno, who had the role of my PhD supervisor in this story. In spite of that, he has always claimed he was not able to answer my statistical questions, even before knowing them; as a result he always had a cool answer. I constantly found him supportive and comprehensive, with good advice to offer.

I am really grateful to Amy, who is not only a great professor but one of the nicest person I ever met. I admire a lot her positive attitude towards research, she was always supporting with me and made me feel welcomed when I first came to UNC. Not to mention that she makes a super pecan pie!

David is also to mention, who I found to be truthfully enthusiastic about research, and always able to find some time for discussion. I learned a lot from working together.

A special thanks to my PhD buddies, especially to Massimiliano and Charlotte, always up for a beer, and to all the companions I found in the department, in particular Giulio, Emanuele, Alessandro. A thought also to the other PhD students and post-docs I met at UNC, Duke, and around conferences, who made me realizing that we are all in the same boat.

My acknowledgments also to the Department of Statistical Sciences, Professors Monica Chiogna and Nicola Sartori who directed the PhD program, and Patrizia Piacentini that helped us to not get lost.

And a final big thanks to my family and my historical friends, Alessia, Camilla, Diana rigorously in alphabetical order - I know you will check on this - for having always being there for me.

Contents

List of Figures

List of Tables

Introduction

Overview

Clustering is one of the canonical data analysis goals in statistics. There are two main strategies that have been used for clustering, namely, distance and model-based clustering. Distance-based methods leverage upon a distance metric between data points and do not in general require a generative probability model of the data, while model-based methods rely on discrete mixture models, which model the data in different clusters as arising from kernels having different parameter values. The majority of the model-based literature reckon on maximum likelihood estimation, commonly relying on the EM algorithm. Bayesian approaches instead aim to approximate a full posterior distribution on the clusters, with advantages in terms of uncertainty quantification, while also having the ability to incorporate prior information.

Most of the Bayesian methods assume *exchangeability*, which means that the prior probability of a partition $\boldsymbol{c}$ of indices $\{1, \ldots, N\}$ into clusters depends only on the number of clusters and the cluster sizes; the indices on the clusters play no role. Under the exchangeability assumption, one can define what is referred to in the literature as an Exchangeable Partition Probability Function (EPPF) (Pitman, 1995). This EPPF provides a prior distribution on the random partition $\boldsymbol{c}$. One direction to obtain a specific form for the EPPF is to start with a nonparametric Bayesian discrete mixture model with a prior for the mixing measure P, and then marginalize over this prior to obtain an induced prior on partitions. Standard choices for P, such as the Dirichlet (Ferguson, 1973) and Pitman-Yor process (Pitman and Yor, 1997), lead to relatively simple analytic forms for the EPPF. There has been some recent literature studying extensions to broad classes of Gibbs-type processes (Gnedin and Pitman, 2006; De Blasi *et al.*, 2015), but mostly related on improving flexibility and the ability to predict the number of new clusters in a future sample of data (Cesari *et al.*, 2014; Arbel and Favaro, 2017).

There is also a rich literature on relaxing exchangeability in various ways. Most of the emphasis has been on the case in which a vector of features $\mathbf{x}_i$ is available for index i, which stimulated the construction of priors to generate feature-dependent random partitions. Building on the stick-breaking representation of the DP (Sethuraman, 1994) MacEachern (1999, 2000) proposed a class of dependent DP (DDP) priors introducing dependence in the base measure, while maintaining fixed the mixing weights. Many extensions of the DPP priors were employed in ANOVA modeling (De Iorio *et al.*, 2004), spatial data analysis (Gelfand *et al.*, 2005), time series (Caron *et al.*, 2006) and functional data analysis (Petrone *et al.*, 2009; Scarpa and Dunson, 2009) applications, with theoretical properties highlighted in Barrientos *et al.* (2012). However such priors turned out lacking of flexibility for feature-dependent clustering, requiring the introduction of too many mixture components in practice as noted in MacEachern (2000). This has motivated more general formulations which allow also the mixing weights to change with the features, with some example including the order-based dependent Dirichlet process (Griffin and Steel, 2006), kernel- (Dunson and Park, 2008), and probit- (Rodriguez and Dunson, 2011) stick breaking processes.

Alternative approaches build on random partition models (RPM), working directly with the probability distribution $p(\boldsymbol{c})$ on the partition $\boldsymbol{c}$ of indices $\{1, \ldots, N\}$ into clusters. Particular attention has been given to the class of product partitions models (PPM) (Barry and Hartigan, 1992; Hartigan, 1990) in which $p(\boldsymbol{c})$ can be factorized into a product of cluster-dependent functions, known as *cohesion functions*. A common strategy modifies such function in order to let features influence a priori the probability for random partitions such as in Park and Dunson (2010), Müller *et al.* (2011) and Dahl *et al.* (2017).

Main contributions of the thesis

Although a plethora of methods has been presented in literature to allow covariate-dependent clustering, there has been limited consideration of the problem of how to effectively include concrete prior knowledge on the partition. In many application settings such as epidemiology, genomics but also business intelligence and sociology, there is often some prior information about the clustering, pertaining the number of clusters, the sizes, or even an actual partition of the data. This thesis focuses on providing a broad new class of methods for improving clustering performance in practice.

Available methods mostly rely on additional covariates, but our setting is totally different. In particular, we do not have features x_i on indices i but have access to an

informed prior guess $\boldsymbol{c}_0$ for the partition $\boldsymbol{c}$; apart from this information it is plausible to rely on exchangeable priors. To address this problem, we propose a novel strategy to modify a baseline EPPF to include centering on $\boldsymbol{c}_0$. In particular, our proposed Centered Partition (CP) process defines the partition prior as proportional to an EPPF multiplied by an exponential factor which depends on a distance function $d(\boldsymbol{c}, \boldsymbol{c}_0)$ measuring how far $\boldsymbol{c}$ is from $\boldsymbol{c}_0$. The proposed framework should be broadly useful in including extra information into EPPFs, which tend to face issues in lacking incorporation of real prior information from applications.

Our motivation arises from a particular application involving birth defects epidemiology. Specifically we have data coming from The National Birth Defects Prevention Study (NBDPS), which is the largest multi-state population-based, case-control study of birth defects ever conducted in the United States (Yoon *et al.*, 2001). Participants in the study included mothers with expected dates of delivery from 1997-2011, with cases identified using birth defects surveillance systems in recruitment areas within ten US states. Because birth defects are highly heterogeneous, a relatively large number of defects of unknown etiology are included in the NBDPS, with the aim to provide new insights on the leading causes of such defects.

We are particularly interested in the class of Congenital Heart Defects (CHD), being the most common type of birth defect and the leading cause of infant death due to birth defects. Because some of these defects are relatively rare, in many cases we may lack precision for investigating associations between potential risk factors and birth defect outcomes. For this reason, researchers typically lump heterogeneous defects in order to increase power (e.g., grouping all heart defects together), even knowing the underlying mechanisms may differ substantially. In fact, how best to group defects is subject to uncertainty, despite a variety of proposed groupings available in the literature.

In this context, there are N different birth defects, which we can index using $i \in \{1, \ldots, N\}$, and we are interested in clustering these birth defects into mechanistic groups. The underlying assumption is that birth defects in the same group may have similar coefficients in logistic regression analysis relating different exposures to risk of developing the defect. Investigators have provided us with an initial partition $\boldsymbol{c}_0$ of the defects $\{1, \ldots, N\}$ into groups. It is appealing to combine this prior knowledge with information in the data from a grouped logistic regression to produce a posterior distribution on clusters, which characterize uncertainty. The motivating question of this thesis is how to do this, with the resulting method ideally having broader impact to other types of *centering* of priors for clustering; for example, we may want to center the prior based on information on the number of clusters or cluster sizes.

To achieve such goal, we built on Bayesian priors for model-based clustering, while exploiting notions belonging to combinatorics. A first part of the thesis is dedicated to the review of such notions that, despite being known in combinatorial and information theory, have been received little attention from the statistics community. Chapter 3 will provide the general formulation of the Centered Partition process, illustrating its properties and behavior under different settings. A general strategy for prior tuning and posterior sampling is presented. Finally in Chapter 4 results on the motivating data from the NBDPS are presented and discussed, along with conclusions on future directions of research.

Chapter 1

Motivation

1.1 Modeling with prior information

The construction of informative priors based on domain knowledge or expert opinions is a delicate problem, complicated by the fact that the human mind finds it difficult to quantify qualitative knowledge. Most of the literature has been focused on combining opinions from experts, with the final objective being a consensus opinion (O'Hagan *et al.*, 2006; Albert *et al.*, 2012). Elicitation of the prior based on such opinions has been found to be broadly useful in many applied contexts, with examples in decision theory (Spetzler and Stael von Holstein, 1975), political sciences (Gill and Walker, 2005) and rare event modeling (Choi, 2016) among others.

Little attention has been given to the role of prior information in data clustering. Some proposals have been made in machine learning, mostly dealing with the case of information on the pairwise allocations. Inclusion of such information, is often handled with the introduction of hard/soft constraints on a base clustering method, such as the k-means algorithm (Wagstaff *et al.*, 2001; Klein *et al.*, 2002), spectral clustering (Kamvar *et al.*, 2003) or model based clustering using graphical models Shental *et al.* (2004); Law *et al.* (2004).

However there are many areas of application in which substantial information about data grouping can be derived from the domain knowledge, from customers segmentation in targeted marketing (Mazanec, 1997), typically searching for few clusters profiling buyers behavior, to investigation of phenotypic associations with partially known groups of regulator genes (Sarup *et al.*, 2016). Our motivating context deals with the presence of a prior guess on the data partitions, with data coming from an epidemiological study,

the National Birth Defects Prevention Study, aimed to individuate the potential risk factors associated with birth defects of unknown etiology.

The diversity and number of birth defects is a major challenge in this field of research, and authors have for many years noted the tradeoffs between lumping birth defects together into larger groups to facilitate analysis and splitting defects into finer, more homogeneous subcategories (Khoury *et al.*, 1992). While most individual defects are too rare for individual study, there is considerable difficulty in determining how best to group defects and even in what to call a defect. We aim to provide a new methodology to facilitate data-adaptive probabilistic clustering of defects, both based on information in the data and also based on incorporating established biological knowledge on embryologic development. Such knowledge will constitute a baseline backbone to aid shrinkage in a manner that allows defects to be analyzed individually when needed, while also grouping effects to facilitate analysis.

1.2 The National Birth Defects Prevention Study

The National Birth Defects Prevention Study (NBDPS) is the largest population-based study ever conducted on the causes of birth defects in the United States (Yoon *et al.*, 2001). It was designed to identify infants with major birth defects and evaluate genetic and environmental factors associated with their occurrence. Because of the study focus on finding the causes of birth defects, cases with known etiology (eg. chromosomal or genetic conditions) are excluded. Diagnostic case information was obtained from medical records and verified by a standardized clinician review specific to the study (Rasmussen *et al.*, 2003). The study has been enrolling families of infants with one or more of 37 selected major birth defects, together with control families, for over 14 years starting in 1998, recruiting over $40,000$ families. Data collection is based on a computer-assisted telephone interview with case and control mothers, focusing on a wide range of demographic, lifestyle, medical, nutrition, occupational and environmental exposure history information. (Reefhuis *et al.*, 2015b).

Cases were identified using birth defects surveillance systems, while controls were randomly selected from birth certificates or hospital records in recruitment areas within ten US states: Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, North Carolina, Texas, and Utah. The annual birth population covered in these states is roughly 10% of all births in the United States, with the demographic distribution of subjects recruited to NBDPS being comparable to that of the general U.S. population with respect to maternal age, race, ethnicity, and education level (Cogswell

et al., 2009). Each state site attempted to recruit 300 cases and 100 (unmatched) controls annually, with the number of controls chosen to be proportional to the number of births registered in the same month during the previous year.

The prevalence of congenital malformations is currently believed to be around 1 in 33 (Murphy *et al.*, 2017), though this quantity is difficult to estimate due to spontaneous abortion and miscarriage of malformed fetuses, induced abortion and subclinical abnormalities that are rarely diagnosed. However birth defects or congenital malformation are the leading causes of infant mortality (Murphy *et al.*, 2017), and research focused on providing new insights on potential risk factors is one major concern in public health. On this side, the NBPDS has provided an unprecedented source of information on birth defects of unknown etiology.

In this thesis, particular attention is given on the class of Congenital Heart Defects (CHD), that accounts for over the 50% of all birth defects. In the NBDPS, heart defects are provided at different levels of detail, according to a hierarchical grouping (Botto *et al.*, 2007). We will consider 26 different heart defects and a prior guess on the grouping comprising 6 clusters. For each defect we observe a different number of cases, and we wish to obtain a mechanistic grouping of the defects in order to better assess the important risk factors in comparison with the controls.

Chapter 2

Preliminaries

2.1 Clustering and Bayesian mixture models

Mixture models have become increasingly popular tools to model data characterized by the presence of subpopulations, in which each observation belongs to one of a certain number of groups, while providing a flexible class for density estimation. In particular, observations $y_1, \ldots, y_N$ can be divided into $K < N$ groups, according to a partition $\boldsymbol{c} = \{B_1, \ldots, B_K\}$ with B_k comprising all the indices of data points in cluster k, for $k = 1, \ldots, K$. The main underlying assumption of a mixture model is that observations are independent conditional on the partition $\boldsymbol{c}$ and on vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ indexing the distribution of observations within each cluster. Hence the joint probability density of observations $y_1, \ldots, y_N$ can be expressed as

$$p(\mathbf{y}|\boldsymbol{c}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{i \in B_k} p(y_i|\theta_k) = \prod_{k=1}^{K} p(\mathbf{y}_k|\theta_k) \tag{2.1}$$

with $\mathbf{y}_k = \{y_i\}_{i \in B_k}$ indicating all the observations in cluster k for $k = 1, \ldots, K$. In the full Bayesian formulation, a prior distribution is assigned to each possible partition $\boldsymbol{c}$, leading to a posterior of the form

$$p(\boldsymbol{c}|\mathbf{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{c}) \prod_{k=1}^{K} p(\mathbf{y}_k|\theta_k). \tag{2.2}$$

Hence the data partitioning $\boldsymbol{c}$ is conceived as a random object and elicitation of its prior distribution is a critical issue in Bayesian modeling. The main issue is that the space of all possible clustering grows exponentially fast given its combinatorial nature. Current Bayesian methodology typically relies on discrete nonparametric priors, which provides

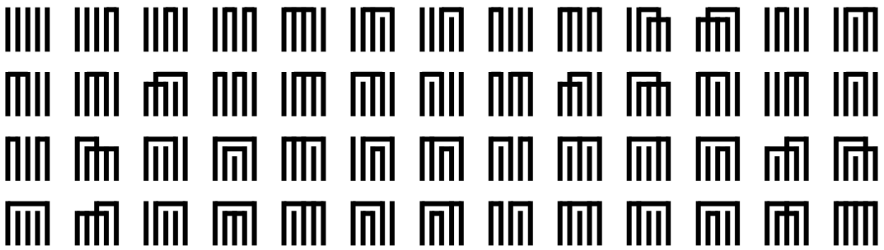


FIGURE 2.1: Genji-mon symbols for all the possible grouping of 5 elements.

tractable tools to deal with mixture models, by avoiding dealing with the clustering space directly, but inducing a latent partitioning of the data. However such priors may be too flexible especially when relevant prior information is available about the clustering, since they lack of a simple way to include this type of information.

With the aim of providing a novel class of priors which allow inclusion of external information on the data, we found a useful theoretical framework in lattice theory and combinatorics. This section introduces the definition of clustering from a combinatorial perspective, providing theoretical background on structural characterization of the corresponding space, while providing a review of the commonly employed priors on the clustering space.

2.1.1 Set partitions

Let $\boldsymbol{c}$ be a generic clustering of indices $\{1, \dots, N\} = [N]$. It can be either represented as a vector of indices $\{c_1, \dots, c_N\}$ with $c_i \in \{1, \dots, K\}$ for $i = 1, \dots, N$ and $c_i = c_j$ when i and j belong to the same cluster, or as a collection of disjoint subsets (blocks) $\{B_1, B_2, \dots, B_K\}$ where B_k contains all the indices of data points in the k-th cluster and K is the number of clusters in the sample of size N. From a mathematical perspective $\boldsymbol{c} = \{B_1, \dots, B_K\}$ is a combinatorial object known as *set partition* of $[N]$. In denoting a set partition, we either write $\{\{1, 2, 4\}, \{3, 5\}\}$ or $124|35$ using a vertical bar to indicate a break in blocks. By convention, elements are ordered from least to greatest and from left to right within a block; we then order the blocks by their least element from left to right. The collection of all possible set partitions of $[N]$, denoted with Π_N, is known as *partition lattice*. We refer to (Stanley, 1997; Davey and Priestley, 2002) for an introduction to lattice theory, reporting here some of the base concepts.

According to Knuth (2006), set partitions seem to have been systematically studied for the first time in Japan (1500 A.D.), due to a parlor game popular in the upper class society known as *genji-ko*; 5 unknown incense were burned and players were asked to

identify which of the scents were the same, and which were different. Ceremony masters soon developed symbols to represent all the possible 52 outcomes, so called *genji-mon* represented in Figure 2.1. Each symbol consists of five vertical bars, with some of them connected by horizontal bars, in correspondence of grouped elements. As an aid to memory, each of the patterns was made after a famous 11th-century novel, *Tales of Genji* by Lady Murasaki, whose original manuscript is now lost, but has made genji-mon an integral part of the Japanese culture. In fact, such symbols continued to be employed as family crests or in Japanese kimono patterns until the early 20th century, and can be found printed in many dresses sold today.

First results in combinatorics focused on enumerating the elements of the space, making their appearance during the 17th century, still in Japan. For example, the number of ways to assign N elements to a fixed number of K groups is described by the *Stirling number of the second kind*

$$\mathcal{S}_{N,K} = \frac{1}{K!}\sum_{j=0}^{K}(-1)^j\binom{K}{j}(K-j)^N,$$

while the *Bell number* $\mathcal{B}_N = \sum_{K=1}^{N}\mathcal{S}_{N,K}$ describes the number of all possible set partitions of N elements. Refer to Knuth (2006) for more information on history and algorithms related to set partitions and other combinatorial objects.

2.1.2 Poset representation and partition lattice

The interest progressively shift from counting elements of the space to characterizing the structure of space partitions using the notion of partial order. Consider Π_N endowed with the set containment relation $\leq$, meaning that for $\boldsymbol{c} = \{B_1, \dots, B_K\}, \boldsymbol{c}' = \{B'_1, \dots, B'_{K'}\}$ belonging to Π_N, $\boldsymbol{c} \leq \boldsymbol{c}'$ if for all $i = 1, \dots, K, B_i \subseteq B'_j$ for some $j \in \{1, \dots, K'\}$. Then the space $(\Pi_N, \leq)$ is a *partially ordered set* (poset), which satisfies the following properties:

1. Reflexivity: for every $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$, $\boldsymbol{c} \leq \boldsymbol{c}$,
2. Antisymmetry: if $\boldsymbol{c} \leq \boldsymbol{c}'$ and $\boldsymbol{c}' \leq \boldsymbol{c}$ then $\boldsymbol{c} = \boldsymbol{c}'$,
3. Transitivity: if $\boldsymbol{c} \leq \boldsymbol{c}'$ and $\boldsymbol{c}' \leq \boldsymbol{c}''$, then $\boldsymbol{c} \leq \boldsymbol{c}''$.

Moreover, for any $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$, it is said that $\boldsymbol{c}$ is *covered* (or refined) by $\boldsymbol{c}'$ if $\boldsymbol{c} \leq \boldsymbol{c}'$ and there is no $\boldsymbol{c}''$ such that $\boldsymbol{c} < \boldsymbol{c}'' < \boldsymbol{c}'$ and indicate with $\boldsymbol{c} \prec \boldsymbol{c}'$ such relation. This covering relation allows one to represent the space of partitions by means of the *Hasse diagram*, in which the elements of Π_N correspond to nodes in a graph and a line is drawn from $\boldsymbol{c}$

to $\boldsymbol{c}'$ when $\boldsymbol{c} \prec \boldsymbol{c}'$; in other words, there is a connection from a partition $\boldsymbol{c}$ to another one when the second can be obtained from the first by splitting or merging one of the blocks in $\boldsymbol{c}$. See Figure 2.2 for an example of Hasse diagram of Π_4. If two elements are not connected, as for example partitions $\{1,2\}\{3,4\}$ and $\{1,3\}\{2,4\}$, they are said to be *incomparable*. Conventionally the partition with just one cluster is represented at the top of the diagram and denoted as $\mathbf{1}$, while the partition having every observation in its own cluster at the bottom and indicated with $\mathbf{0}$.

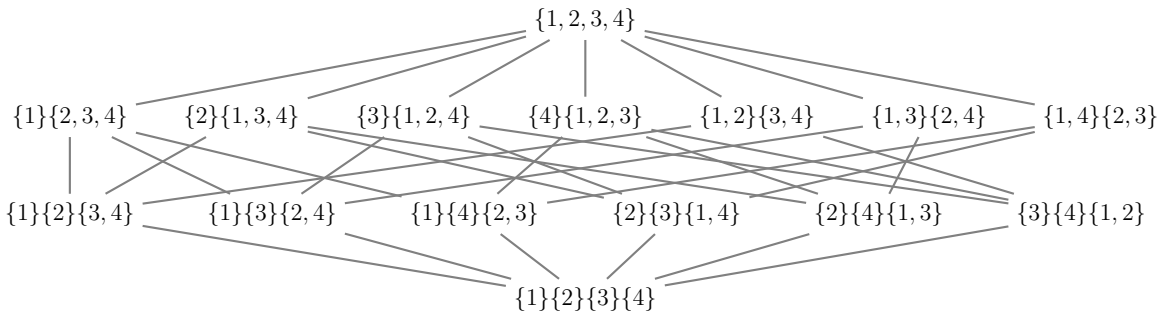


FIGURE 2.2: Hasse diagram for the lattice of set partitions of 4 elements. A line is drawn when two partitions have a covering relation. For example $\{1\}\{2,3,4\}$ is connected with 3 partitions obtained by splitting the block $\{2,3,4\}$ in every possible way, and partition $\mathbf{1}$ obtained by merging the two clusters.

This representation of the set partitions space Π_N as a partially ordered set provides a useful framework to characterize its elements. As already mentioned, two partitions connected in the Hasse diagram can be obtained from one another by means of a single operation of split or merge; a sequence of connections is *path*, linking the two extreme partitions $\mathbf{0}$ and $\mathbf{1}$. A path starting from $\mathbf{0}$ connects partitions with an increasing *rank*, which is related to the number of blocks through $r(\boldsymbol{c}) = N - |\boldsymbol{c}|$. Set partitions with the same rank, may differ in terms of their *configuration* $\boldsymbol{\Lambda(c)}$, the sequence of block cardinalities $\{|B_1|, \ldots, |B_K|\}$, which corresponds to another combinatorial object known as *integer partition* of N. In combinatorics, an integer partition is defined as multiset of positive integers $\{\lambda_1 \ldots \lambda_K\}$, listed in decreasing order by convention, such that $\sum_{i=1}^{K} \lambda_i = N$. The space I_N of all possible integer partitions, is also a partially ordered set, making the definition of configuration a poset mapping $\boldsymbol{\Lambda}(\cdot) : \boldsymbol{c} \in \Pi_N \rightarrow \boldsymbol{\lambda} \in I_N$.

Finally, the space Π_N is a *lattice*, for the fact that every pair of elements has a *greatest lower bound* (g.l.b.) and a *least upper bound* (l.u.b.) indicated with the "meet" $\wedge$ and the "join" $\vee$ operators, i.e. $\boldsymbol{c} \wedge \boldsymbol{c}' = g.l.b.(\boldsymbol{c}, \boldsymbol{c}')$ and $\boldsymbol{c} \vee \boldsymbol{c}' = l.u.b.(\boldsymbol{c}, \boldsymbol{c}')$ and equality holds under a permutation of the cluster labels. An element $\boldsymbol{c} \in \Pi_N$ is an upper bound for a subset $\boldsymbol{S} \subseteq \Pi_N$ if $\boldsymbol{s} \leq \boldsymbol{c}$ for all $\boldsymbol{s} \in \boldsymbol{S}$, and it is the least upper bound for a subset $\boldsymbol{S} \subseteq \Pi_N$

if $\boldsymbol{c}$ is an upper bound for $\boldsymbol{S}$ and $\boldsymbol{c} \leq \boldsymbol{c}'$ for all upper bounds $\boldsymbol{c}'$ of $\boldsymbol{S}$. The lower bound and the greatest lower bound are defined similarly, and the definition applies also to the elements of the space I_N. Consider as an example $\boldsymbol{c} = \{1\}\{2,3,4\}, \boldsymbol{c}' = \{3\}\{1,2,4\}$; their greatest lower bound (g.l.b.) is $\boldsymbol{c} \wedge \boldsymbol{c}' = \{1\}\{3\}\{2,4\}$ while the least upper bound (l.u.b.) is $\boldsymbol{c} \vee \boldsymbol{c}' = \{1,2,3,4\}$. Looking at the Hasse diagram in Fig 2.2 the g.l.b. and l.u.b. are in general the two partitions which reach both $\boldsymbol{c}$ and $\boldsymbol{c}'$ through the shortest path, respectively from below and from above.

2.1.3 Distances on the partition lattice

The representation of the space of set partitions Π_N from lattice theory, provides a useful framework to define metrics between partitions. In fact, the distance between any two partitions can be defined by means of the Hasse diagram as the length of any shortest path between them, which necessarily passes through the meet or join of two partitions.

More general distances arise when the graph is weighted, meaning that every edge is associated with a strictly positive weight; then the distance between any two elements is the weight of the lightest path between them, where the weight of a path is the sum over its edges of their weight. Weights over the edges of the Hasse diagram are usually defined starting from a function ν on the lattice Π_N having the following properties.

Definition 1. A lattice function $\nu : \Pi_N \to \mathbb{R}^+$, is said to be

- *strictly order-preserving* if $\nu(\boldsymbol{c}) > \nu(\boldsymbol{c}')$, for $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$ such that $\boldsymbol{c} > \boldsymbol{c}'$.
- *strictly order-reversing* if $\nu(\boldsymbol{c}) > \nu(\boldsymbol{c}')$, for $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$ such that $\boldsymbol{c} < \boldsymbol{c}'$.
- *supermodular* if $\nu(\boldsymbol{c} \vee \boldsymbol{c}') + \nu(\boldsymbol{c} \wedge \boldsymbol{c}') - \nu(\boldsymbol{c}) - \nu(\boldsymbol{c}') \geq 0$, for any $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$.
- *submodular* if $\nu(\boldsymbol{c} \vee \boldsymbol{c}') + \nu(\boldsymbol{c} \wedge \boldsymbol{c}') - \nu(\boldsymbol{c}) - \nu(\boldsymbol{c}') \leq 0$, for any $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$.

We report here a useful result from lattice theory referring to Grätzer (2002) and Deza and Deza (2009). Given a lattice function ν weights w_ν on edges between $\{\boldsymbol{c}, \boldsymbol{c}'\}$ are defined as

$$w_\nu(\{\boldsymbol{c}, \boldsymbol{c}'\}) = |\nu(\boldsymbol{c}) - \nu(\boldsymbol{c}')|,$$

with distance between two partitions being the minimum-ν-weighted path. Properties outlined in Definition 1 guarantee that such path visits either the meet or the join of any two incomparable partitions; which one of the two depends on whether is supermodular or submodular.

Proposition 1. *For any strictly order-preserving (order-reversing) function ν, if ν is supermodular, the minimum-ν-weight partition distance is*

$$d_\nu(\boldsymbol{c}, \boldsymbol{c}') = \nu(\boldsymbol{c}) + \nu(\boldsymbol{c}') - 2\nu(\boldsymbol{c} \wedge \boldsymbol{c}') \quad (d_\nu(\boldsymbol{c}, \boldsymbol{c}') = \nu(\boldsymbol{c}) + \nu(\boldsymbol{c}') - 2\nu(\boldsymbol{c} \vee \boldsymbol{c}')),$$

while if ν is submodular

$$d_\nu(\boldsymbol{c}, \boldsymbol{c}') = 2\nu(\boldsymbol{c} \vee \boldsymbol{c}') - \nu(\boldsymbol{c}) - \nu(\boldsymbol{c}') \quad (d_\nu(\boldsymbol{c}, \boldsymbol{c}') = 2\nu(\boldsymbol{c} \wedge \boldsymbol{c}') - \nu(\boldsymbol{c}) - \nu(\boldsymbol{c}')).$$

Moreover the defined distance is a metric on the space of partitions Π_N. Some widely used distances can be recast in this framework, such for example, the Hamming distance (see Rossi, 2015, for proof), however different functions can be employed to obtain a suitable distance.

A trivial example of lattice function can be draw from the definition of the rank, i.e. $r : \Pi_N \to \mathbb{Z}^+$ such that $r(\boldsymbol{c}) = N - |\boldsymbol{c}|$, which is strictly order-preserving. For example, considering partitions in the Hasse diagram in Figure 2.2, the rank of the bottom partition $\mathbf{0}$ is equal to 0 and increases by 1 for each level of the graph up to 3 for top partition $\mathbf{1}$. Then the minimum-rank-weighted distance can be computed as $d_r(\boldsymbol{c}, \boldsymbol{c}') = 2r(\boldsymbol{c} \vee \boldsymbol{c}') - r(\boldsymbol{c}) - r(\boldsymbol{c}')$, since the function is also submodular. Notice that the rank assigns to every edge between partitions a unit weight, and then d_r is indeed the shortest path distance.

2.2 Distributions on the set partition lattice

2.2.1 Uniform distribution

The first distribution one may use in absence of prior information, is the uniform prior, which gives the same probability to every partitions with $p(\boldsymbol{c}) = 1/\mathcal{B}_N$; however, even for small values of N the Bell number $\mathcal{B}_N$ is very large, making computation of the posterior intractable even for simple choices of the likelihood. This motivated the definition of alternative prior distributions based on different concepts of uniformity, with Jensen and Liu (2008) prior favoring uniform placement of new observation in one of the existing clusters, and Casella *et al.* (2014) proposing a hierarchical uniform prior, which gives equal probability to set partitions having the same configuration.

2.2.2 Discrete nonparametric priors

In Bayesian nonnparametric settings, a probability distribution on the space of partitions is typically obtained by means of discrete nonparametric prior, i.e. priors that have discrete realization almost surely. Due to the discreteness of the process, any random probability measure associated with the discrete prior induces an exchangeable random partition on the data indices, which is described in terms of Exchangeable Probability Partition Function. Popular choices of such priors comprises the Dirichlet (Ferguson, 1973) and Pitman-Yor (Pitman and Yor, 1997) processes, which are instances of the more general class of the Gibbs-type priors. We refer to Gnedin and Pitman (2006) and De Blasi *et al.* (2015) for characterization of the Gibbs-type priors, limiting here on aforementioned processes that lead to tractable EPPFs on the space of set partitions.

Starting from a Dirichlet process prior with concentration parameter α and base measure G_0, the corresponding EPPF is obtained by marginalizing out the process, (Lo, 1984; Kuo, 1986)

$$p(\boldsymbol{c}) = \frac{\alpha^K}{(\alpha)^N} \prod_{j=1}^{K} (\lambda_j - 1)!, \tag{2.3}$$

where $\lambda_j = |B_j|$ is the cardinality of the j-th clusters composing the partition, and $(x)_r = x(x+1)\cdots(x+r-1)$ denoting the rising factorial. Similar is the case of the Pitman-Yor process, which involves an extra discount parameter σ modifying the EPPF into

$$p(\boldsymbol{c}) = \frac{\prod_{j=1}^{K-1}(\alpha + j\sigma)}{(\alpha+1)_{(N-1)}} \prod_{j=1}^{K} (1-\sigma)_{(\lambda_j - 1)}. \tag{2.4}$$

Additionally, an explicit formulation of the EPPF is obtained in the finite-dimensional case by starting from symmetric Dirichlet distribution over κ components and parameters α/κ. In this case the EPPF is restricted to the space of partitions of κ elements and corresponds to

$$p(\boldsymbol{c}) = \frac{\kappa!}{(\kappa - K)!} \prod_{j=1}^{K} \frac{\Gamma(\alpha/\kappa + \lambda_j)}{\Gamma(\alpha/\kappa)}, \tag{2.5}$$

which for $\kappa \to \infty$ becomes 2.3.

2.2.3 Product partition models

There is a strong connection with the exchangeable random partitions induced by Gibbs-type priors and product partition models (Barry and Hartigan, 1992; Hartigan, 1990). A product partition model assumes that the prior probability for the partition $\boldsymbol{c}$ has the

Random probability measure	Parameters	$p(\boldsymbol{c}) \propto$
Dirichlet process	(α)	$\alpha^K \prod_{j=1}^{K} (\lambda_j - 1)!$
Pitman-Yor process	(α, σ)	$\prod_{j=1}^{K-1} (\alpha + j\sigma)(1-\sigma)_{(\lambda_j - 1)}$
Symmetric Dirichlet	(κ, γ)	$\frac{\kappa!}{(\kappa - K)!} \prod_{j=1}^{K} \frac{\Gamma(\gamma/\kappa + \lambda_j)}{\Gamma(\gamma/\kappa)}$

TABLE 2.1: Cohesion functions for Dirichlet, Pitman-Yor processes and Symmetric Dirichlet distribution. $\lambda_j = |B_j|$ is the cardinality of the clusters composing the partition, while $(x)_r = x(x+1)\cdots(x+r-1)$ denotes the rising factorial.

following form

$$p(\boldsymbol{c} = \{B_1, \ldots, B_K\}) \propto \prod_{j=1}^{K} \rho(B_j), \tag{2.6}$$

with $\rho(\cdot)$ known as cohesion function. The underlying assumption is that the prior distribution for the set partition $\boldsymbol{c}$ can be factorized in the product of functions that depends only on the blocks composing it. Such definition, in conjunction with formulation (2.1) for the data likelihood, guarantees the property that the posterior distribution for $\boldsymbol{c}$ is still in the class of product partition models.

Distributions in Table 2.1 are all characterized by a cohesion function that depends on the clusters through their cardinality. Such characterization results too strict in many applied context in which there are reasonable assumptions about the grouping, since the same prior probability is given to partitions having the same configuration, i.e. the same sequence of cluster sizes, independently on the observations assigned to the clusters.

Chapter 3

Centered Partition Processes

Although prior distributions presented in the previous chapter have been proved to provide a suitable framework for Bayesian mixture modeling, they still lack of flexibility in including external information. Our focus is on incorporating structured knowledge about data partitioning in the prior distribution. As a first approach, we consider as source of information a given potential clustering, but our proposal should be useful also to accommodate other types of prior information such as the number of clusters and cluster sizes.

3.1 General formulation

Assume that a base partition $\boldsymbol{c}_0$ is given and we wish to include this information in the prior distribution. To address this problem, we propose a general strategy to modify a baseline EPPF to shrink towards $\boldsymbol{c}_0$. In particular, our proposed CP process defines the prior on set partitions as proportional to a baseline EPPF multiplied by a penalization term of the type

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c}) e^{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)}, \tag{3.1}$$

with $\psi > 0$ a penalization parameter, $d(\boldsymbol{c}, \boldsymbol{c}_0)$ a suitable distance measuring how far $\boldsymbol{c}$ is from $\boldsymbol{c}_0$ and $p_0(\boldsymbol{c})$ indicates a baseline EPPF, that may depend on some parameters that are not of interest at the moment. For $\psi \to 0$, $p(\boldsymbol{c}|\boldsymbol{c}_0, \psi)$ corresponds to the baseline EPPF $p(\boldsymbol{c}_0)$, while for $\psi \to \infty$, $p(\boldsymbol{c} = \boldsymbol{c}_0) \to 1$.

Note that $d(\boldsymbol{c}, \boldsymbol{c}_0)$ takes a finite number of discrete values $\Delta = \{\delta_0, \ldots, \delta_L\}$, with L depending on $\boldsymbol{c}_0$ and on the distance $d(\cdot, \cdot)$. We can define sets of partitions having the

same fixed distance from $\boldsymbol{c}_0$ as

$$s_l(\boldsymbol{c}_0) = \{\boldsymbol{c} \in \Pi_N : d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l\}, \quad l = 0, 1, \ldots, L. \tag{3.2}$$

Hence, for $\delta_0 = 0$, $s_0(\boldsymbol{c}_0)$ denotes the set of partitions equal to the base one, meaning that they differ from $\boldsymbol{c}_0$ only by a permutation of the cluster labels. Then $s_1(\boldsymbol{c}_0)$ denotes the set of partitions with minimum distance δ_1 from $\boldsymbol{c}_0$, $s_2(\boldsymbol{c}_0)$ the set of partitions with the second minimum distance δ_2 from $\boldsymbol{c}_0$ and so on. The introduced exponential term penalizes equally partitions in the same set $s_l(\boldsymbol{c}_0)$ for a given δ_l, but the resulting probabilities may differ depending on the chosen baseline EPPF.

3.1.1 Choiche of the distance and related properties

The proposed CP process modifies a baseline EPPF to include a distance-based penalization term, which aims to shrink the prior distribution towards a prior partition guess. The choice of distance plays a key role in determining the behavior of the prior distribution. A variety of different distances and indices have been employed in clustering procedures and comparisons. We consider in this paper the Variation of Information (VI), obtained axiomatically in Meilă (2007) using information theory, and shown to nicely characterize neighborhoods of a given partition by Wade and Ghahramani (2018). The variation of information is based on the Shannon entropy $H(\cdot)$, and can be computed as

$$\begin{aligned}\mathrm{VI}(\boldsymbol{c}, \boldsymbol{c}') &= -H(\boldsymbol{c}) - H(\boldsymbol{c}_0) + 2H(\boldsymbol{c}, \boldsymbol{c}_0) \\ &= \sum_{j=1}^{K} \frac{\lambda_j}{N} \log\left(\frac{\lambda_j}{N}\right) + \sum_{l=1}^{K'} \frac{\lambda'_l}{N} \log\left(\frac{\lambda'_l}{N}\right) - 2\sum_{j=1}^{K}\sum_{l=1}^{K'} \frac{\lambda^{\wedge}_{jl}}{N} \log\left(\frac{\lambda^{\wedge}_{jl}}{N}\right),\end{aligned}$$

where log denotes log base 2, and $\lambda^{\wedge}_{jl}$ the size of blocks of the intersection $\boldsymbol{c} \wedge \boldsymbol{c}'$ and hence the number of indices in block j under partition $\boldsymbol{c}$ and block k under $\boldsymbol{c}'$. Notice that VI ranges from 0 to $\log_2(N)$.

Although normalized versions have been proposed (Vinh *et al.*, 2010), some desirable properties are lost under normalization. We refer to Meilă (2007) and Wade and Ghahramani (2018) for additional properties and empirical evaluations.

The Variation of Information belongs to class of metrics described in Section 2.1.3 in which the weighting edged function is the Shannon entropy $H(\cdot)$. In general, a distance between any two different partitions $\boldsymbol{c}, \boldsymbol{c}' \in \Pi_N$ can be defined by means of the Hasse diagram via their minimum weighted path, which corresponds to the shortest path

length when edges are equally weighted. Instead, when edges depend on the entropy function through $w(\{\boldsymbol{c}, \boldsymbol{c}'\}) = |H(\boldsymbol{c}) - H(\boldsymbol{c}')|$, the minimum weighted path between two partitions is the Variation of Information. In fact, notice that two connected partitions are in a covering relation then $\boldsymbol{c} \wedge \boldsymbol{c}'$ is either equal to $\boldsymbol{c}$ or $\boldsymbol{c}'$ and then $VI(\boldsymbol{c}, \boldsymbol{c}') = w(\boldsymbol{c}, \boldsymbol{c}')$. This means that splitting or merging smaller clusters has less impact on the VI then splitting or merging larger ones. In fact the minimum weight corresponds to $2/N$ which is attained when two singleton clusters are merged, or conversely, a cluster consisting of two points is split.

3.1.2 Effect of the prior penalization

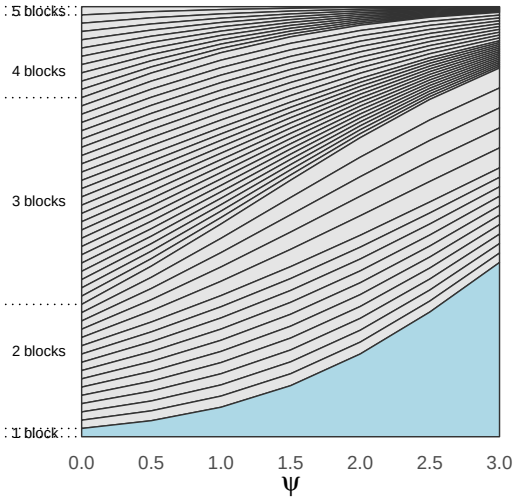
We first consider the important special case in which the baseline EPPF is $p_0(\boldsymbol{c}) = 1/\mathcal{B}_N$ and the CP process reduces to $p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto \exp\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\}$ with equation (3.1) simplifying to

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) = \frac{e^{-\psi \delta_l}}{\sum_{u=0}^{L} n_u e^{-\psi \delta_u}}, \quad \text{for } \boldsymbol{c} \in s_l(\boldsymbol{c}_0), \quad l = 0, 1, \ldots, L, \tag{3.3}$$

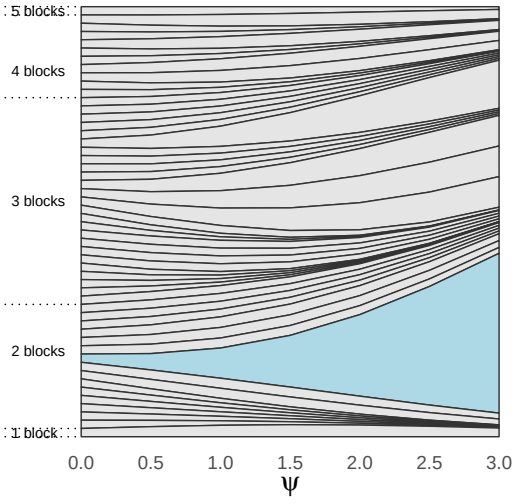
where $n_u = |s_u(\boldsymbol{c}_0)|$, the number of partitions in the set $s_u(\boldsymbol{c}_0)$ as defined in (3.2).

Considering $N = 5$, there are 52 possible set partitions; Figure 3.1 shows the prior cumulative probabilities for the CP process for different values of $\psi \in (0, 3)$ with $\psi = 0$ corresponding to the uniform prior. In order to compute the cumulative probabilities, we ordered partitions arbitrarily grouping them according to the number of blocks and block sizes. Notice that since the order of the partitions is arbitrary, the same applies to values of the cumulative probabilities, that are hence omitted; the interesting information is in how the probability associated to each partition, given from the area between curves, varies according to different values of ψ. Notice also that base partitions with the same configuration (e.g. for $\boldsymbol{c}_0 = \{1, 2\}\{3, 4, 5\}$ all the partitions with blocks sizes $\{3, 2\}$), will behave in the same way, with the same probabilities assigned to partitions different in composition.

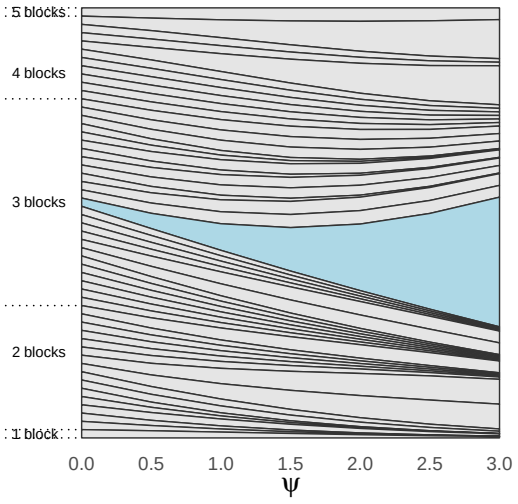
Non-zero values of ψ increase the prior probability of partitions $\boldsymbol{c}$ that are relatively close to the chosen $\boldsymbol{c}_0$. However, the effect is not uniform but depends on the structure of both $\boldsymbol{c}$ and $\boldsymbol{c}_0$. For example, consider the inflation that occurs in the blue region as ψ increases from 0 to 3. When $\boldsymbol{c}_0$ has 1 block (Figure 3.1a) versus 4 (Figure 3.1d) there is a bigger increase. This is because the space of set partitions Π_N is not "uniform" with respect configurations; an integer partition $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ can be rewritten as $\boldsymbol{\lambda} = (1^{f_1}, 2^{f_2}, \ldots, K^{f_K})$, with the notation indicating that there are f_j elements of $\boldsymbol{\lambda}$



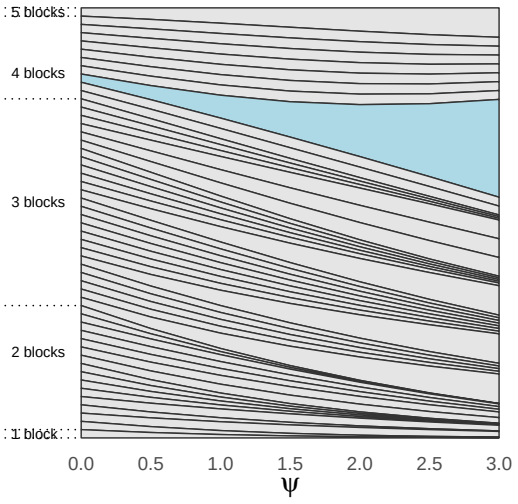
(A) $\boldsymbol{c}_0 = \{1,2,3,4,5\}$



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



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



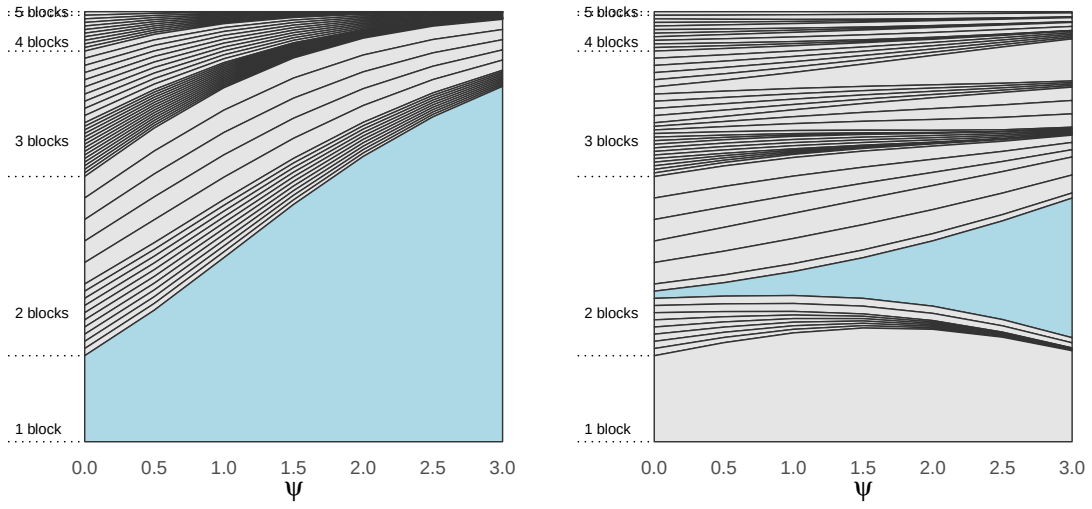
(D) $\boldsymbol{c}_0 = \{1\}\{2\}\{3\}\{4,5\}$

FIGURE 3.1: The cumulative probabilities of the 52 set partitions of $N = 5$ elements for the CP process with uniform base EPPF. In each graph the CP process is centered on a different partition c_0 highlighted in blue. For each partition, the cumulative probabilities across different values of the penalization parameter ψ are joined to form the curves. The probability of a given partition corresponds to the difference between the curves.

equal to j; then the number of set partitions having configuration $\boldsymbol{\lambda}$ is

$$\frac{N!}{\prod_{j=1}^{K} \lambda_j! \prod_{i=1}^{N} f_i!}.$$

For example, for $\{32\} = 1^0 2^1 3^1 4^0 5^0$, the number of corresponding set partitions is 10,



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

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

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

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

FIGURE 3.2: The cumulative probabilities of the 52 set partitions of $N = 5$ elements for the CP process with Dirichlet Process of $\alpha = 1$ base EPPF. In each graph the CP process is centered on a different partition c_0 highlighted in blue. For each partition, the cumulative probabilities across different values of the penalization parameter ψ are joined to form the curves. The probability of a given partition corresponds to the difference between the curves.

while there are 5 set partitions of type $\{41\}$. When $\boldsymbol{c}_0 = \mathbf{1}$ the closest partitions are ones having this last configuration, and as it can be seen Figure 3.1a, for increasing values of ψ their probability is inflated more and more. Instead, if the closest set partitions would correspond to ones with configuration $\{32\}$, there would need an higher value of ψ to similarly increase the probability of those partitions, simply because they are more

in numerosity.

While the uniform distribution gives the same probability to each partition in the space, the EPPF induced by Gibbs-type priors distinguishes between different configurations, but not among partitions with the same configuration. We focus on the Dirichlet process case, being the most popular process employed in applications. Under the DP the induced EPPF $p_0(\boldsymbol{c}) \propto \alpha^K \prod_{j=1}^K \Gamma(\lambda_j)$ is a function of the configuration $\boldsymbol{\Lambda}(\boldsymbol{c})$, which is one of $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$ since the possible configurations are finite and correspond to the number of integer partitions. Letting $g(\boldsymbol{\Lambda}(\boldsymbol{c})) = \alpha^K \prod_{j=1}^K \Gamma(\lambda_j)$, the formulation in (3.1) can be written as

$$p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) = \frac{g(\boldsymbol{\lambda}_m) e^{-\psi \delta_l}}{\sum_{u=0}^{L} \sum_{v=1}^{M} n_{uv} g(\boldsymbol{\lambda}_v) e^{-\psi \delta_u}}, \quad \text{for } \boldsymbol{c} \in s_{lm}(\boldsymbol{c}_0), \tag{3.4}$$

where $s_{lm}(\boldsymbol{c}_0) = \{\boldsymbol{c} \in \Pi_N : d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l, \boldsymbol{\Lambda}(\boldsymbol{c}) = \boldsymbol{\lambda}_m\}$, the set of partitions with distance δ_l from $\boldsymbol{c}_0$ and configuration $\boldsymbol{\lambda}_m$ for $l = 0, 1, \ldots, L$ and $m = 1, \ldots, M$, with n_{lm} indicating the cardinality of such set. The factorization (3.4) applies for the family of Gibbs-type priors in general, with different expressions of $g(\boldsymbol{\Lambda}(\boldsymbol{c}))$.

In Figure 3.2 we consider the prior distribution induced by the CP process when the baseline EPPF $p_0(\boldsymbol{c})$ comes from a Dirichlet process with concentration parameter $\alpha = 1$, considering the same base partitions and values for ψ as in Figure 3.1. For the same values of the parameter ψ, the behavior of the CP process changes significantly due to the effect of the base prior. In particular, in the top left panel the CP process is centered on $\boldsymbol{c}_0 = \{1, 2, 3, 4, 5\}$, the partition with only one cluster, which is *a priori* the most likely one for $\psi = 0$. In general, for small values of ψ the clustering process will most closely resemble that for a DP, and as ψ increases the DP prior probabilities are decreased for partitions relatively far from $\boldsymbol{c}_0$ and increased for $\boldsymbol{c}_0$ relatively close.

3.2 Posterior computation

In classical Bayesian mixture models, inference is conducted by generating samples from the posterior distribution, obtained by means of Markov Chain Monte Carlo (MCMC) algorithms. The proposed CP process modifies the prior distribution on the space of set partitions, favoring partitions in neighborhood of the available prior guess. Hence, in defining algorithms for posterior computation, we need to account for the distance-dependent penalization when sampling a new partition $\boldsymbol{c}^*$. In this section we illustrate how to adapt some MCMC scheme to sample from the CP processes when the base EPPF is uniform or a Gibbs-type.

3.2.1 Posterior computation under uniform distribution

Recall that the Metropolis-Hastings algorithm for sampling from the posterior distribution $p(\boldsymbol{c}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{c})p(\boldsymbol{c})$ uses a distribution $q(\boldsymbol{c}^*|\boldsymbol{c})$ to draw a proposal value $\boldsymbol{c}^*$ and update it according to the acceptance probability

$$\min\left\{1, \frac{q(\boldsymbol{c}|\boldsymbol{c}^*)}{q(\boldsymbol{c}^*|\boldsymbol{c})}\frac{p(\boldsymbol{c}^*)}{p(\boldsymbol{c})}\frac{p(\mathbf{y}|\boldsymbol{c}^*)}{p(\mathbf{y}|\boldsymbol{c})}\right\}. \tag{3.5}$$

In general the prior ratio under the CP process becomes

$$\frac{p(\boldsymbol{c}^*)}{p(\boldsymbol{c})} = \exp\{-\psi[d(\boldsymbol{c}^*, \boldsymbol{c}_0) - d(\boldsymbol{c}, \boldsymbol{c}_0)]\},$$

being higher for proposal partitions closer to $\boldsymbol{c}_0$ than the current one, with the difference between the distances being equal to $H(\boldsymbol{c}) - H(\boldsymbol{c}^*) + 2H(\boldsymbol{c}^* \wedge \boldsymbol{c}_0) - 2H(\boldsymbol{c} \wedge \boldsymbol{c}_0)$ using the Variation of Information.

When Metropolis-Hastings (MH) algorithms are defined on a combinatorial structure, the candidate Markov chain is often taken to be a random walk on a graph that defines a neighborhood structure for the combinatorial set in question (Booth *et al.*, 2008). The relationship associated with the edges in the graph determines the possible moves. The Hasse diagram provides a natural representation of the set partitions, that allows to define candidate partitions obtained with split and merge moves, described in Algorithm 1. We randomly decide between a merge move with probability $p_m \in (0, 1)$ and a split move with probability $(1 - p_m)$. A split move is automatically proposed whenever the currents state consists of a single cluster and likewise a merge move when the current state consists of n clusters. Assume that a proposal partition $\boldsymbol{c}^*$ is obtained from a partition $\boldsymbol{c}$ with K clusters, by merging two of them; then the transition probability of a merge operation is

$$q(\boldsymbol{c}^*|\boldsymbol{c}) = \frac{2p_m}{K(K-1)}, \tag{3.6}$$

while the one of a split corresponds to

$$q(\boldsymbol{c}|\boldsymbol{c}^*) = \frac{1 - p_m}{(2^{\lambda^*-1} - 1)\sum_{j=1}^{K^*}\mathcal{I}\{\lambda_j \geq 2\}}, \tag{3.7}$$

where λ^* is the size of the cluster in $\boldsymbol{c}^*$ to split to obtain $\boldsymbol{c}$. When the proposal partition is obtained with a split operation, the ratio between the two probabilities in (3.6)- (3.7) is inverted. Finally the likelihood ratio $p(\mathbf{y}|\boldsymbol{c}^*)/p(\mathbf{y}|\boldsymbol{c})$ depends only on observations

belonging to the clusters involved in the split and merge operations.

Algorithm 1 : MH proposal based on split and merge moves

0. Let the partition $\boldsymbol{c}$ with K clusters be the current state of the Markov Chain.

1. Sample $u \sim \text{Bern}(p_m)$.

if $u = 1$ **then**

Select uniformly at random 2 clusters in $\{1, \ldots, K\}$ and define the proposal $\boldsymbol{c}^*$ by merging the two clusters.

else if $u = 0$ **then**

Select uniformly at random 1 cluster in $\{1, \ldots, K\}$ and define the proposal $\boldsymbol{c}^*$ by splitting the chosen cluster into two clusters conditionally on neither being empty.

end if

Following Booth *et al.* (2008), we paired these moves with a second proposal based on an alternative graph definition, in which there is a connection between two partitions if one can be obtained from the other by moving exactly one of the N objects to a different cluster. The resulting proposal is a biased random walk described in Algorithm 2. It can be assessed that $q(\boldsymbol{c}|\boldsymbol{c}^*) = q(\boldsymbol{c}^*|\boldsymbol{c})$, which simplifies computation of the acceptance probability in 3.5. Moreover, in this case, the likelihood ratio involves only the observation moved from one clusters to the other.

Algorithm 2 : MH proposal based on biased random walk

0. Let the partition $\boldsymbol{c}$ with K clusters be the current state of the Markov Chain.

if $K = 1$ **then**

1. Choose one observation i in $\{1, \ldots, N\}$ uniformly at random and move it to its own cluster.

else if $K > 1$ **then**

Choose one observation i in $\{1, \ldots, N\}$ uniformly at random.

if i is a singleton **then**

Move i to one of the other $K - 1$ clusters with probability $1/(K - 1)$.

else if i is not a singleton **then**

Move i to one of the other $K - 1$ clusters or its own one with probability $1/K$.

end if

end if

We suggest to alternate between these two proposals at each iteration, by making a transition according to the split and merge algorithm or the biased random walk with probabilities p_b and $1 - p_b$ respectively. Indeed the MH algorithm results in mixture of small scale changes in the partition acting on one element at time, and large scale moves involving entire clusters.

3.2.2 Posterior computation under Gibbs-type priors

Certain MCMC algorithms for Bayesian nonparametric mixture models can be easily modified for posterior computation in CP process models. In particular, we adapt the so-called "marginal algorithms" developed for Dirichlet and Pitman-Yor processes. These methods are called marginal since the mixing measure is integrated out of the model and the predictive distribution is used within a MCMC sampler. In the following, we recall Algorithm 2 in Neal (2000) and illustrate how it can be adapted to sample from the CP process posterior. We refer to Neal (2000) and references therein for an overview and discussion of methods for both conjugate and nonconjugate cases, and to Fall and Barat (2014) for adaptation to Pitman-Yor processes.

Let $\boldsymbol{c}$ be represented as an N-dimensional vector of indices $\{c_1, \dots, c_N\}$ encoding cluster allocation and let θ_k be the set of parameters currently associated to cluster k. The prior predictive distribution for a single c_i conditionally on $\boldsymbol{c}^{-i} = \{c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_N\}$ is exploited to perform the Gibbs sampling step allocating observations to either a new cluster or one of the existing ones. Algorithm 2 in Neal (2000) updates each c_i sequentially for $i = 1, \dots, N$ via a reseating procedure, according to the conditional posterior distribution

$$p(c_i = k|\boldsymbol{c}^{-i}, \boldsymbol{\theta}, y_i) \propto \begin{cases} p(c_i = k|\boldsymbol{c}^{-i})p(y_i|\theta_k) & k = 1, \dots, K^- \\ p(c_i = k|\boldsymbol{c}^{-i}) \int p(y_i|\theta) dG_0(\theta) & k = K^- + 1, \end{cases} \tag{3.8}$$

with K^- the number of clusters after removing observation i. The conditional distribution $p(c_i = k|\boldsymbol{c}^{-i})$ is reported in Table 3.1 for different choices of the prior EPPF. Notice that, for the case of finite Dirichlet prior, the update consists only in the first line of equation (3.8), since the number of classes is fixed. For Dirichlet and Pitman-Yor processes, when observation i is associated to a new cluster, a new value for θ is sampled from its posterior distribution based on the base measure G_0 and the observation y_i. This approach is straightforward when we can compute the integral $\int p(y_i|\theta) dG_0(\theta)$, as will generally be the case when G_0 is a conjugate prior.

Considering the proposed CP process, the conditional distribution for c_i given $\boldsymbol{c}^{-i}$ can still be computed, but it depends both on the base prior and the penalization term accounting for the distance between the base partition $\boldsymbol{c}_0$ and the one obtained by assigning the observation i to either one of the existing classes $k \in \{1, \dots, K^-\}$ or a new one. Hence, the step in equation (3.8) can be easily adapted by substituting the

Random probability measure	Parameters	$p(c_i = k\|\boldsymbol{c}^{-i}) \propto$
Dirichlet process	(α)	$\begin{cases} \frac{\lambda_k^{-i}}{\alpha+N-1} & k = 1,\dots,K^- \\ \frac{\alpha}{\alpha+N-1} & k = K^-+1 \end{cases}$
Pitman-Yor process	(α,σ)	$\begin{cases} \frac{\lambda_k^{-i}-\sigma}{\alpha+N-1} & k = 1,\dots,K^- \\ \frac{\alpha+\sigma K^-}{\alpha+N-1} & k = K^-+1 \end{cases}$
Symmetric Dirichlet	(κ,γ)	$\frac{\lambda_k^{-i}+\gamma/\kappa}{\alpha+N-1} \quad k = 1,\dots,\kappa$

TABLE 3.1: Conditional prior distribution for c_i given $\boldsymbol{c}^{-i}$ under different choices of the EPPF. With K^- we denote the total number of clusters after removing the i-th observation, while λ_k^{-i} is the corresponding number of observations in cluster k.

conditional distribution for $p(c_i = k|\boldsymbol{c}^{-i})$ with

$$p(c_i = k|\boldsymbol{c}^{-i}, \boldsymbol{c}_0, \psi) \propto p_0(c_i = k|\boldsymbol{c}^{-i}) \exp\{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)\} \quad k = 1,\dots,K^-,K^-+1$$

with $\boldsymbol{c} = \{\boldsymbol{c}^{-i} \cup \{c_i = k\}\}$ the current state of the clustering and $p_0(c_i = k|\boldsymbol{c}^{-i})$ one of the conditional distributions in Table 3.1. Additional steps on the implementation using the Variation of Information as a distance are given in Algorithm 3), but the same procedure applies when using other distances based on blocks sizes, such as the Hamming distance or the Rand Index.

Extension to the non-conjugate context can be similarly handled exploiting algorithm 8 in Neal (2000) based on auxiliary parameters, which avoids the computation of the integral $\int p(y_i|\theta)dG_0(\theta)$. The only difference is in that, when c_i is updated, m temporary auxiliary variables are introduced that represent possible values for the parameters of components that are not associated with any other observations. Such variables are simply sampled from the base measure G_0, with the probabilities of a new cluster in Table 3.1 changing into $(\alpha/m)/(\alpha+N-1)$ for the Dirichlet Process and to $[(\alpha+\sigma K^-)/m]/(\alpha+N-1)$ for $k = K^-+1,\dots,K^-+1$.

Algorithm 3 : Computation strategy for the penalization term in marginal sampling

Let K^- and K_0^- denote respectively the number of clusters in $\boldsymbol{c}^{-i}$ and $\boldsymbol{c}_0^{-i}$, i.e. partitions $\boldsymbol{c}$ and $\boldsymbol{c}_0$ after removing the i observation.

for $i = 1,\dots,N$ **do**

1. Compute cardinalities $\{\lambda_1^{-i},\dots,\lambda_{K^-}^{-i}\}$ representing the number of observations in each cluster for $\boldsymbol{c}^{-i}$.

2. Compute λ_{lm}^{-i}, the number of observations in cluster l under $\boldsymbol{c}^{-i}$ and cluster m under $\boldsymbol{c}_0^{-i}$ for $l = 1,\dots,K^-$ and $m = 1,\dots,K_0^-$.

for $k = 1,\dots,K^-,K^-+1$ **do**

Let $c_{i,0}$ be the cluster of index i under partition $\boldsymbol{c}_0$.
Compute $d(\boldsymbol{c}, \boldsymbol{c}_0) \propto -H(\boldsymbol{c}) + 2H(\boldsymbol{c} \wedge \boldsymbol{c}_0)$ for $\boldsymbol{c} = \{\boldsymbol{c}^{-i} \cup k\}$ using

$$-H(\boldsymbol{c}) = \sum_{l \neq k}^{K} \left\{ \frac{\lambda_l^{-i}}{N} \log \frac{\lambda_l^{-i}}{N} \right\} + \left(\frac{\lambda_k^{-i} + 1}{N} \right) \log \left(\frac{\lambda_k^{-i} + 1}{N} \right)$$

$$H(\boldsymbol{c} \wedge \boldsymbol{c}_0) = -\left\{ \sum_{l=1}^{K} \sum_{m=1}^{K_0^-} \frac{\lambda_{lm}^{-i}}{N} \log \left(\frac{\lambda_{lm}^{-i}}{N} \right) - \frac{\lambda_{kc_{i,0}}^{-i}}{N} \log \left(\frac{\lambda_{kc_{i,0}}^{-i}}{N} \right) + \frac{\lambda_{kc_{i,0}}^{-i} + 1}{N} \log \left(\frac{\lambda_{kc_{i,0}}^{-i} + 1}{N} \right) \right\}$$

end for
end for

3.3 Prior calibration

As the number of observations N increases, the number of partitions explodes, and higher values of ψ are needed to place non-negligible prior probability in small to moderate neighborhoods around $\boldsymbol{c}_0$. The prior concentration around $\boldsymbol{c}_0$ depends on three main factors: i) N through $\mathcal{B}_N$, i.e. the cardinality of the space of set partitions, ii) the baseline EPPF $p_0(\boldsymbol{c}_0)$ and iii) where $\boldsymbol{c}_0$ is located in the space. We hence propose a general method to evaluate the prior behavior under different settings, while suggesting how to choose the parameter ψ.

One may evaluate the prior distribution for different values of ψ and check its behavior using graphs such as those in Section 3.1.2, however they become difficult to interpret as the space of partitions grows. We propose to evaluate the probability distribution of the distances $\delta = d(\boldsymbol{c}, \boldsymbol{c}_0)$ from the known partition $\boldsymbol{c}_0$. The probability assigned to different distances by the prior is

$$p(\delta = \delta_l) = \sum_{\boldsymbol{c} \in \Pi_N} p(\boldsymbol{c}) \mathcal{I}\left\{ d(\boldsymbol{c}, \boldsymbol{c}_0) = \delta_l \right\} = \sum_{\boldsymbol{c} \in s_l(\boldsymbol{c}_0)} p(\boldsymbol{c}) \quad l = 0, \ldots, L,$$

with $\mathcal{I}(\cdot)$ the indicator function and $s_l(\boldsymbol{c}_0)$ denoting the set of partitions distance δ_l from $\boldsymbol{c}_0$, as defined in (3.2). Consider the uniform distribution on set partitions, $p(\delta = \delta_l) = |s_l(\boldsymbol{c}_0)|/\mathcal{B}_N$, the proportion of partitions distance δ_l from $\boldsymbol{c}_0$. Under the general definition of the CP process, the resulting distribution becomes

$$p(\delta = \delta_l) = \sum_{\boldsymbol{c} \in s_l(\boldsymbol{c}_0)} \frac{p_0(\boldsymbol{c}) e^{-\psi \delta_l}}{\sum_{u=0}^{L} \sum_{\boldsymbol{c}^* \in s_u(\boldsymbol{c}_0)} p_0(\boldsymbol{c}^*) e^{-\psi \delta_u}} \quad l = 0, \ldots, L, \tag{3.9}$$

with the case of Gibbs-type EPPF corresponding to

$$p(\delta = \delta_l) = \frac{\sum_{m=1}^{M} n_{lm} g(\boldsymbol{\lambda}_m) e^{-\psi \delta_l}}{\sum_{u=0}^{L} \sum_{v=1}^{M} n_{uv} g(\boldsymbol{\lambda}_v) e^{-\psi \delta_u}}, \quad l = 0, \ldots, L. \tag{3.10}$$

Notice that the uniform EPPF case is recovered when $g(\boldsymbol{\lambda}_m) = 1$ for $m = 0, \ldots, M$, so that $\sum_{m=1}^{M} n_{lm} = n_l$. Hence the probability in (3.9) simplifies to

$$p(\delta = \delta_l) = \frac{n_l e^{-\psi \delta_l}}{\sum_{u=0}^{L} n_u e^{-\psi \delta_u}} \quad l = 0, \ldots, L. \tag{3.11}$$

In general, since distances are naturally ordered, the corresponding cumulative distribution function can be simply defined as $F(\delta) = \sum_{\delta_l \leq \delta} p(\delta_l)$ for $\delta \in \{\delta_0, \ldots, \delta_L\}$ and used to assess how much mass is placed in different size neighborhoods around $\boldsymbol{c}_0$ under different values of ψ. Hence we can choose ψ to place a specified probability q (e.g. $q = 0.9$) on partitions within a specified distance δ^* from $\boldsymbol{c}_0$. This would correspond to calibrating ψ so that $F(\delta^*) \approx q$, with $F(\delta^*) \geq q$. In other words, partitions generated from the prior would have at least probability q of being within distance δ^* from $\boldsymbol{c}_0$.

The main problem is in computing the probabilities in equations (3.10)-(3.11), which depend on all the set partitions in the space. In fact, one needs to count all the partitions having distance δ_l for $l = 0, \ldots, L$ when the base EPPF is uniform, while taking account also of configurations in the case of the Gibbs-type priors. Even if there are quite efficient algorithms to list all the possible set partitions of N (see Knuth, 2005; Nijenhuis and Wilf, 2014), it becomes computationally infeasible due to the extremely rapid growth of the space; for example from $N = 12$ to 13, the number of set partitions grows from $\mathcal{B}_{12} = 4,213,597$ to $\mathcal{B}_{13} = 27,644,437$. Another option is to approximate counts by means of a Monte Carlo procedure based on uniform sampling, but with the main drawback is that small values of counts are highly underestimated, or even missed, given the very low probability to sample the corresponding partitions. In particular, we will miss information about partitions close to $\boldsymbol{c}_0$, which are the ones we wish to favor.

We propose a general strategy to approximate prior probabilities assigned to different distances from $\boldsymbol{c}_0$ focused on obtaining estimates of distance values and related counts, which represent the sufficient quantities to compute (3.10)-(3.11) under different values of ψ. We consider a targeted Monte Carlo procedure which augments uniform sampling on the space of set partitions with a deterministic local search using the Hasse diagram to provide exact counts for small values of the distance.

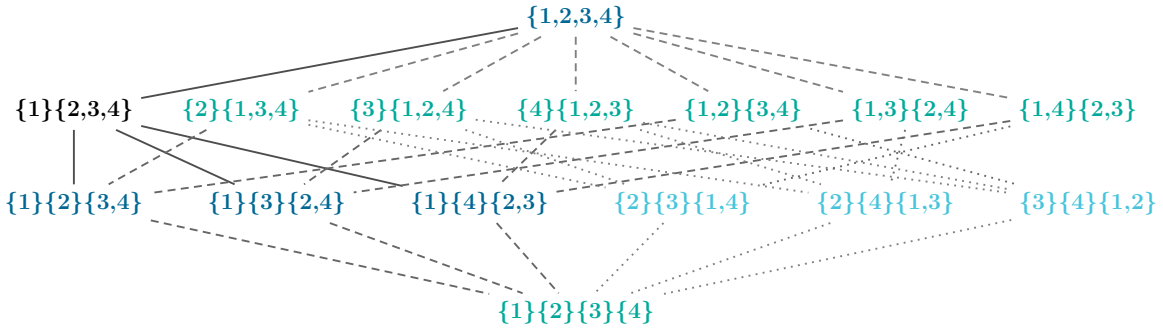


FIGURE 3.3: Illustration of results from the local search algorithm based on the Hasse diagram of Π_4 starting from $\boldsymbol{c}_0 = \mathbf{\{1\}\{2,3,4\}}$. Partitions are colored according the exploration order according to dark-light gradient. Notice that after 3 iterations the space is entirely explored.

3.3.1 Deterministic local search

Poset theory provides a nice representation of the space of set partitions by means of the Hasse diagram illustrated in Section 2.1.2, along with suitable definition of metrics. A known partition $\boldsymbol{c}_0$ can be characterized in terms of number of blocks K_0 and configuration $\boldsymbol{\Lambda}(\boldsymbol{c}_0)$. These elements allows one to locate $\boldsymbol{c}_0$ in the Hasse diagram and then explore connected partitions by means of split and merge operations on the clusters in $\boldsymbol{c}_0$.

As an illustrative example, consider the Hasse diagram of Π_4 in Figure 3.3 and $\boldsymbol{c}_0 = \{1\}\{2,3,4\}$, having 2 clusters and configuration $\boldsymbol{\Lambda}(\boldsymbol{c}_0) = \{31\}$. Let $\mathcal{N}_1(\boldsymbol{c}_0)$ denote the sets of partitions directly connected with $\boldsymbol{c}_0$, i.e. partitions covering $\boldsymbol{c}_0$ and those covered by $\boldsymbol{c}_0$. In general, a partition $\boldsymbol{c}_0$ with K_0 clusters is covered by $\binom{K_0}{2}$ partitions and covers $\sum_{j=1}^{K_0} 2^{\lambda_j - 1} - 1$. In the example, $\mathcal{N}_1(\boldsymbol{c}_0)$ contains $\{1,2,3,4\}$ obtained from $\boldsymbol{c}_0$ with a merge operation on the two clusters, and all the partitions obtained by splitting the cluster $\{2,3,4\}$ in any possible way. The base idea underlying the proposed local search, consists in exploiting the Hasse diagram representation to find all the partitions in increasing distance neighborhoods of $\boldsymbol{c}_0$. One can list partitions at T connections from $\boldsymbol{c}_0$ starting from $\mathcal{N}_1(\boldsymbol{c}_0)$ by recursively applying split and merge operations on the set of partitions explored at each step. Potentially, with enough operations one can reach all the set partitions, since the space is finite with lower and upper bounds.

In practice, the space is too huge to be explored entirely, and a truncation is needed. From the example in Figure 3.3, $\mathcal{N}_1(\boldsymbol{c}_0)$ contains 3 partitions with distance 0.69 from $\boldsymbol{c}_0$ and one with distance 1.19. Although $\mathcal{N}_2(\boldsymbol{c}_0)$ may contain partitions closer to $\boldsymbol{c}_0$ than this last, the definition of distance in Section 2.1.3 guarantees that there are no other partitions with distance from $\boldsymbol{c}_0$ less than 0.69. Since the VI is the minimum weighted path between two partitions, all the partitions reached at the second exploration step add

a nonzero weight to distance computation. This consideration extends to an arbitrary number of explorations T, with $\delta_{L^*} = \min\{d(\boldsymbol{c}^*, \boldsymbol{c}_0)\}_{\boldsymbol{c}^* \in \mathcal{N}_T(\boldsymbol{c}_0)}$ being the upper bound on the distance value. By discarding all partitions with distance greater that δ_{L^*}, one can compute exactly the counts in equations (3.10)-(3.11) related to distances $\delta_0, \dots, \delta_{L^*}$. Notice that $2/N$ is the minimum distance between two different partitions, and $2T/N$ is a general lower bound on the distances from $\boldsymbol{c}_0$ that can be reached in T iterations.

3.3.2 Monte Carlo approximation

We pair the local exploration with a Monte Carlo procedure to estimate the counts and distances greater that δ_{L^*}, in order to obtain a more refined representation of the prior distance probabilities. Sampling uniformly from the space of partitions is not in general a trivial problem, but a nice strategy has been proposed in Stam (1983), in which the probability of a partition with K clusters is used to sample partitions via an urn model. Derivation of the algorithm starts from the *Dobiński formula* (Dobiński, 1877) for the Bell numbers

$$\mathcal{B}_N = e^{-1} \sum_{k=1}^{\infty} \frac{k^N}{k!}, \tag{3.12}$$

which from a probabilistic perspective corresponds to the k-th moment of the Poisson distribution with expected value equal to 1. Then a probability distribution for the number of clusters $K \in \{1, \dots, N\}$ of a set partition can be defined as

$$P(K = k) = e^{-1} \frac{k^N}{\mathcal{B}_N k!}, \tag{3.13}$$

which is a well defined law thanks to (3.12). To simulate a uniform law over Π_N, Stam (1983)'s algorithm first generates the number of clusters K according to (3.13) and, conditionally on the sampled value, it allocates observations to the clusters according a discrete uniform distribution over $\{1, \dots, K\}$. We refer to Stam (1983) and Pitman (1997) for derivations and proof of the validity of the algorithm.

We adapt the uniform sampling to account for the values already computed by rejecting all the partitions with distance less that δ_{L^*}, restricting the space to $\left\{\Pi_N \setminus \{\mathcal{N}_t(\boldsymbol{c}_0)\}_{t=0}^T\right\}$. In practice, few samples are discarded since the probability to sample one such partition corresponds to $|\{\mathcal{N}_t(\boldsymbol{c}_0)\}_{t=0}^T|/\mathcal{B}_N$, which is negligible for small values of exploration steps T that are generally used in the local search. A sample of partitions $\boldsymbol{c}^{(1)}, \dots, \boldsymbol{c}^{(R)}$, can be used to provide an estimate of the counts. Let R^* denote the number of accepted partitions and $\mathcal{B}^* = \mathcal{B}_N - |\{\mathcal{N}_t(\boldsymbol{c}_0)\}_{t=0}^T|$ be the number of partitions in the restricted space. Conditionally on the observed values of distances

in the sample, $\hat{\delta}_{(L^*+1)}, \ldots, \hat{\delta}_L$, an estimate of the number of partitions with distance $\hat{\delta}_l$ to use in the uniform EPPF case is

$$\hat{n}_l = \mathcal{B}^* \frac{1}{R^*} \sum_{r=1}^{R^*} \mathcal{I}\left\{ d(\boldsymbol{c}^{(r)}, \boldsymbol{c}_0) = \hat{\delta}_l \right\}, \tag{3.14}$$

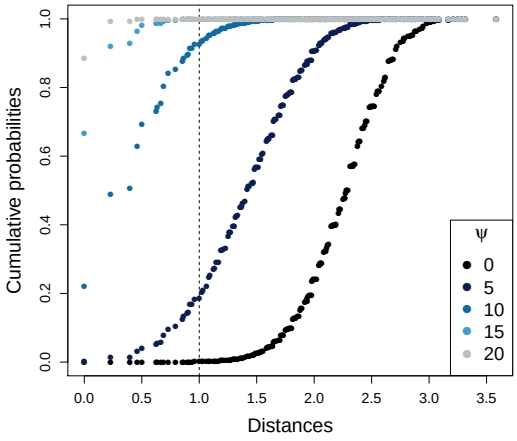
obtained by multiplying the proportions of partitions in the sample by the total known number of partitions. For the Gibbs-type EPPF case one needs also to account for the configurations $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M$ in a given orbital of the distance; hence, the estimates are

$$\hat{n}_{lm} = \mathcal{B}^* \frac{1}{R^*} \sum_{r=1}^{R^*} \mathcal{I}\left\{ d(\boldsymbol{c}^{(r)}, \boldsymbol{c}_0) = \hat{\delta}_l \right\} \mathcal{I}\left\{ \boldsymbol{\Lambda}(\boldsymbol{c}^{(r)}) = \boldsymbol{\lambda}_m \right\}. \tag{3.15}$$

Pairing these estimates with the counts obtained via the local search, one can evaluate the distributions in equations (3.10)-(3.11) for different values of ψ. The entire procedure is summarized in Algorithm 4. Although it requires a considerable number of steps, the procedure can be performed one single time providing information for different choices of ψ and EPPFs. Moreover the local search can be implemented in parallel to reduce computational costs.

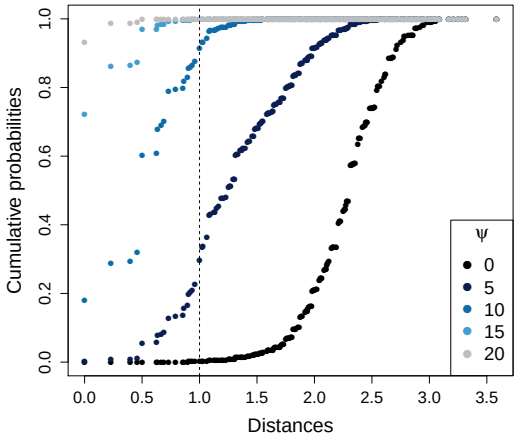
We consider an example for $N = 12$ and $\boldsymbol{c}_0$ with configuration $\{3, 3, 3, 3\}$. Figure 3.4 shows the resulting cumulative probability estimates of the CP process under uniform and DP($\alpha = 1$) base distributions, estimated with $T = 4$ iterations of the local search and $20,000$ samples. Dots represent values of the cumulative probabilities, with different colors in correspondence to different values of the parameter ψ. Using these estimates one can assess how much probability is placed in different distance neighborhoods of $\boldsymbol{c}_0$; tables in Figure 3.4 show the distance values in terms of VI defining neighborhoods around $\boldsymbol{c}_0$ with 90% prior probability. If one wishes to place such probability mass on partitions within distance 1 from $\boldsymbol{c}_0$, a value of ψ around 10 and 15 is needed, respectively, under uniform and DP base prior.

Figures 3.5a-3.5b provide some insights about the motivation underlying the proposed procedure. In particular Figure 3.5a shows the differences in percentage between the true distance probabilities and the corresponding Monte Carlo estimates for increasing number of samples, considering the same setting as above. Obviously, best performances are obtained for the biggest sample, however it is worth noticing that a sample of $100,000$ partitions corresponds to the 0.02% of the total, but even the 0.01% providing very good estimates.



ψ	0	5	10	15	20
$\min \delta : F(\delta) \geq 0.9$	2.68	1.97	0.90	0.5	0.22

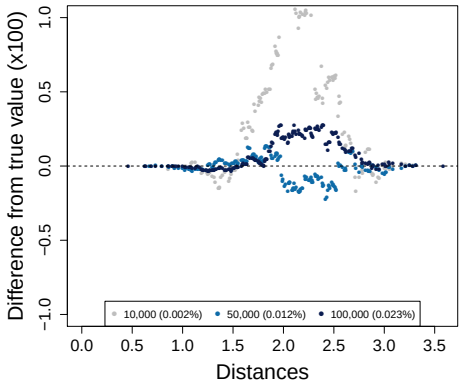
Uniform EPPF



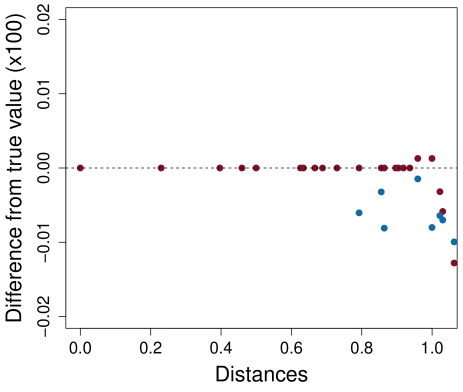
ψ	0	5	10	15	20
$\min \delta : F(\delta) \geq 0.9$	2.71	2.04	1.27	0.5	0

DP ($\alpha = 1$) EPPF

FIGURE 3.4: Estimate of the cumulative prior probabilities assigned to different distances from $\boldsymbol{c}_0$ for $N = 12$ and $\boldsymbol{c}_0$ with configuration $\{3, 3, 3, 3\}$, under the CP process with uniform prior on the left and Dirichlet process on the right. Black dots correspond to the base prior with no penalization, while dots from bottom-to-top correspond to increasing values of $\psi \in \{5, 10, 15, 20\}$. Tables report the minimum distance values in terms of VI such that $F(\delta) \geq 0.9$.



(A) Monte Carlo estimates



(B) Local search

FIGURE 3.5: Differences, in percentage, between the estimated cumulative distance probabilities and the true ones for $N = 12$ ($\mathcal{B}_{12} = 4,213,597$), $\boldsymbol{c}_0$ with configuration $\{3, 3, 3, 3\}$. Figure 3.5a shows differences between true values and Monte Carlo estimates, for different values of the sample sizes $(10,000,\ 50,000,\ 100,000)$. Figure 3.5b compares estimates of small distance values probabilities obtained via Monte Carlo (blue dots) and our procedure including a deterministic local search (red dots).

Nevertheless the main drawback of using solely Monte Carlo estimates, is shown in Figure 3.5b. Red dots represent differences between the true values and the ones computed by our proposed procedure, while blue dots corresponds to the Monte Carlo ones. As it can be noticed, by sampling partitions uniformly, probabilities related to small values of distances are highly underestimated, or even missed; moreover such probabilities are the ones we wish to inflate.

Algorithm 4 : Estimation of counts statistics related to distances neighborhoods of $\boldsymbol{c}_0$

Local search

0. Start from the base partition $\boldsymbol{c}_0$ with $|K_0|$ clusters and configuration $\boldsymbol{\lambda}_{m_0}$ and set $\delta_0 = 0$ and $\mathcal{N}_0(\boldsymbol{c}_0) = \boldsymbol{c}_0$.

for $t = 1, \dots, T$ **do**

Obtain $\mathcal{N}_t(\boldsymbol{c}_0)$ from partitions in $\mathcal{N}_{t-1}(\boldsymbol{c}_0)$ by exploring all directed connections, i.e. partitions obtained with one operation of split/merge on elements $\mathcal{N}_{t-1}(\boldsymbol{c}_0)$.

end for

2. Compute the distance from $\boldsymbol{c}_0$ and all partitions in $\mathcal{N}_T(\boldsymbol{c}_0)$ and take the minimum distance, δ_{L^*}; discard all partitions having distances greater than δ_{L^*}.

3. Obtain counts n_l and n_{lm} relative to distances $\delta_1, \dots, \delta_{L^*}$ for $m = 1, \dots, M$.

Monte Carlo approximation

for $r = 1, \dots, R$ **do**

4. Sample the number of clusters K from the discrete probability distribution

$$p(K = k) = e^{-1}k^N/(k!\mathcal{B}_N), \quad k \in \{1, \dots, N\}.$$

5. Conditional on K generate a partition $\boldsymbol{c}^{(r)} = \{c_1^{(r)}, \dots, c_N^{(r)}\}$ by sampling each $c_i^{(r)}$ from a discrete uniform distribution on $\{1, \dots, K\}$.

6. If $d(\boldsymbol{c}^{(r)}, \boldsymbol{c}_0) > \delta_{L^*}$ reject the partition.

end for

7. Let R^* be the number of accepted partitions, and estimate counts $\hat{n}_l$ and $\hat{n}_{lm}$ for $m = 1, \dots, M$ and according to (3.14)-(3.15) conditional on the observed distance values $\hat{\delta}_{(L^*+1)}, \dots, \hat{\delta}_L$.

8. Using R^* be the number of accepted partitions, and estimate counts $\hat{n}_l$ and $\hat{n}_{lm}$ relative to distances $\hat{\delta}_{L^*+1}, \dots, \hat{\delta}_L$ for $m = 1, \dots, M$.

Chapter 4

Application to birth defects epidemiology

4.1 Congenital heart defects

Birth defects or congenital malformations are the leading causes of infant mortality (Murphy *et al.*, 2017), with the class of Congenital Heart Defects (CHD) accounting for over the 50% of all birth defects. Because some of these defects are relatively rare, in many cases we lack precision for investigating associations between potential risk factors and individual birth defects. For this reason, researchers typically lump together heterogeneous defects in order to increase power (e.g., grouping all heart defects together), even knowing the underlying mechanisms may differ substantially. In fact, how best to group defects is still subject to uncertainty, despite a variety of proposed groupings available in the literature (Lin *et al.*, 1999).

While it may seem natural at first to group defects by anatomic features, this type of grouping may obscure important developmental relationships, and mechanistic classification of the defects has become increasingly popular (Clark, 2001). Such classification is based on the knowledge gained on the embryonic development, that has roots in a cluster of cells that are intended to contribute to each organ or structure. Small errors in the coordination of cells differentiation and migration can produce defects that are linked to specific developmental mechanism. The concept can be summarized in the following way: one molecular abnormality, one mechanism, one group of heart diseases potentially different from an anatomic perspective but homogeneous in their embryonic development.

Congenital Heart Defect	Abbreviation	Frequencies	Percentage of cases
Septal			
Atrial septal defect	ASD	765	0.15
Perimembranous ventricular septal defect	VSDPM	552	0.11
Atrial septal defect, type not specified	ASDNOS	225	0.04
Muscular ventricular septal defect	VSDMUSC	68	0.02
Ventricular septal defect, otherwise specified	VSDOS	12	0.00
Ventricular septal defect, type not specified	VSDNOS	8	0.00
Atrial septal defect, otherwise specified	ASDOS	4	0.00
Conotruncal			
Tetralogy of Fallot	FALLOT	639	0.12
D-transposition of the great arteries	DTGA	406	0.08
Truncus arteriosus	COMMONTRUNCUS	61	0.01
Double outlet right ventricle	DORVTGA	35	0.01
Ventricular septal defect reported as conoventricular	VSDCONOV	32	0.01
D-transposition of the great arteries, other type	DORVOTHER	22	0.00
Interrupted aortic arch type B	IAATYPEB	13	0.00
Interrupted aortic arch, not otherwise specified	IAANOS	5	0.00
Left ventricular outflow			
Hypoplastic left heart syndrome	HLHS	389	0.08
Coarctation of the aorta	COARCT	358	0.07
Aortic stenosis	AORTICSTENOSIS	224	0.04
Interrupted aortic arch type A	IAATYPEA	12	0.00
Right ventricular outflow			
Pulmonary valve stenosis	PVS	678	0.13
Pulmonary atresia	PULMATRESIA	100	0.02
Ebstein anomaly	EBSTEIN	66	0.01
Tricuspid atresia	TRIATRESIA	46	0.01
Anomalous pulmonary venous return			
Total anomalous pulmonary venous return	TAPVR	163	0.03
Partial anomalous pulmonary venous return	PAPVR	21	0.01
Atrioventricular septal defect			
Atrioventricular septal defect	AVSD	112	0.02

TABLE 4.1: Summary statistics of the distribution of congenital heart defects among cases. Defects are divided according the grouping provided from investigators.

In this particular application, we consider 26 individual heart defects, which have been previously grouped into 6 categories by investigators on the basis of the classification provided in Clark (2001) and reviewed in Botto *et al.* (2007) specifically for the NBPDS. The prior grouping is shown in Table 4.1, along with basic summary statistics of the distribution of defects in the analyzed data. Interest is in evaluating the impact of about 90 potential risk factors related to maternal health status, pregnancy experience, lifestyle and family history. We extracted a subset of data from NBDPS, excluding observations with missing covariates, obtaining a dataset with $8,125$ controls, while all heart defects together comprise $4,947$ cases.

4.1.1 A base modeling approach

Standard approaches assessing the impact of exposure factors on the risk to develop a birth defect often rely on logistic regression analysis. Let $i = 1, \dots, N$ index birth defects, while $j = 1, \dots, n_i$ indicates observations related to birth defect i, with $y_{ij} = 1$ if observation j has birth defect i and $y_{ij} = 0$ if observation j is a control, i.e. does not have any birth defect. Let $\mathbf{X}_i$ denote the data matrix associated to defect i, with each row $\mathbf{x}_{ij}^T = (x_{ij1}, \dots, x_{ijp})$ being the vector of the observed values of p categorical variables for the j-th observation. At first one may consider N separate logistic regressions of the type

$$\log\left(\frac{\Pr(y_{ij} = 1|\mathbf{x}_{ij})}{\Pr(y_{ij} = 0|\mathbf{x}_{ij})}\right) = \text{logit}(\boldsymbol{\pi}_{ij}) = \alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta}_i, \tag{4.1}$$

with α_i denoting the defect-specific intercept, and $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})$ the $p \times 1$ vector of regression coefficients. However, Table 4.1 highlights the heterogeneity of heart defects prevalences, with some of them being so few as to preclude separate analyses.

A Bayesian version of model in 4.1 is obtained by considering a Pólya-gamma data augmentation scheme for Bayesian logistic regression (Polson *et al.*, 2013). A latent variable $\omega_{ij} \sim PG(1, \alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta}_i)$ is introduced for each observation j in defect-specific dataset i for $i = 1, \dots, N$ and $j = 1, \dots, n_j$. Following Polson *et al.* (2013) the likelihood contribution for the j-th observation in dataset i conditionally on the ω_{ij} is

$$\propto \exp\left[-\frac{\omega_{ij}}{2}\{(y_{ij} - 0.5)/\omega_{ij} - \alpha_i - \mathbf{x}_{ij}^T\boldsymbol{\beta}_i\}\right]. \tag{4.2}$$

Equation 4.2 is the kernel of a Gaussian distribution for data $(y_{ij} - 0.5)/\omega_{ij}$ with mean $\alpha_i - \mathbf{x}_{ij}^T\boldsymbol{\beta}_i$ and variance $1/\omega_{ij}$. Letting $\boldsymbol{\beta}_i \sim \mathcal{N}_p(\mathbf{b}, \mathbf{Q})$ be the prior for the coefficient vector, the logistic regression can be recasted in terms of a Bayesian linear regression with Gaussian response $(y_{ij} - 0.5)/\omega_{ij}$.

A weakly informative prior may be placed on the coefficient vector (Gelman *et al.*, 2013), e.g. $\boldsymbol{\beta} \sim \mathcal{N}_p(0, \text{diag}(2))$, since it is reasonable to not expect extreme values of the odds ratio resulting from coefficient estimates. Although the introduction of a prior distribution may help in regularizing the estimates, still we need to group the defects somehow, in order to deal with low numerosity cases.

4.1.2 Sharing information across defects

A first step in introducing uncertainty about clustering of the defects may rely on a standard Bayesian nonparametric approach, placing a Dirichlet process prior on the distribution of regression coefficient vector $\boldsymbol{\beta}_i$ in order to borrow information across

multiple defects while letting the data inform on the number and composition of the clusters. A similar approach has been previously proposed in MacLehose and Dunson (2010), with the aim being to shrink the coefficient estimates towards multiple unknown means. In our setting, an informed guess on the group mean values is available through $\boldsymbol{c}_0$, available in Table 4.1.

We consider a simple approach building on the Bayesian version of the model in (4.1), and allowing the exposure coefficients $\boldsymbol{\beta}_i$ for $i = 1, \dots, N$ to be shared across regressions while accounting for the partition $\boldsymbol{c}_0$ by means of our proposed Centered Partition process. The model, written in a hierarchical specification

$$\begin{aligned} y_{ij} &\sim Ber(\pi_{ij}) & \text{logit}(\pi_{ij}) &= \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i}, \quad j = 1, \dots, n_i, \\ \alpha_i &\sim \mathcal{N}(a_0, \tau_0^{-1}) & \boldsymbol{\beta}_{c_i} | \boldsymbol{c} &\sim \mathcal{N}_p(\mathbf{b}, \mathbf{Q}) \quad i = 1, \dots, N, \\ p(\boldsymbol{c}) &\sim CP(\boldsymbol{c}_0, \psi, p_0(\boldsymbol{c})) & p_0(\boldsymbol{c}) &\propto \alpha^K \prod_{k=1}^{K} (\lambda_k - 1)! \end{aligned} \tag{4.3}$$

where $CP(\boldsymbol{c}_0, \psi, p_0(\boldsymbol{c}_0))$ indicates our proposed Centered Partition process, with base partition $\boldsymbol{c}_0$, tuning parameter ψ and baseline EPPF $p_0(\boldsymbol{c}_0)$. We specify the baseline EPPF so that when $\psi = 0$ the prior distribution reduces to a Dirichlet Process with concentration parameter α. Instead, for $\psi \to \infty$ the model would corresponds to K separate logistic regressions, one for each group composing $\boldsymbol{c}_0$. Estimation is carried by following Algorithm 5, with the step for clustering update following indications in Section 3.2.2 including Algorithm 8 variation in Neal (2000) to with $m = 1$ auxiliary variables.

Algorithm 5 : Gibbs sampling for posterior computation

Conditionally on the cluster allocation vector $\mathbf{c} = (c_1, \dots, c_n)$ and data $\{\mathbf{y}_i, \mathbf{X}_i\}$ for $i = 1, \dots, N$, update mixture related parameters and Pólya-gamma latent variables as follow

---

[1] Sample Pólya-gamma latent variables for each observation in each dataset
for $i = 1, \dots, N$ and $j = 1, \dots, n_i$ **do**

$$(\omega_{ij}|-) \sim PG(1, \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i})$$

end for

---

[2] Update defect-specific intercept, exploiting Pólya-gamma conjugancy
for $i = 1, \dots, N$ **do**

$$(\alpha_i|-) \sim \mathcal{N}(a^*, \tau^*)$$

with $\tau^* = \tau_0 + \sum_{j=1}^{n_i} \omega_{ij}$ and $a^* = [a_0\tau_0 + \sum_{j=1}^{n_i}(y_{ij} - 1/2 - \omega_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta}_{c_i})]/\tau^*$

end for

---

[3] Defining $\kappa_{ij} := y_{ij} - 1/2 - w_{ij}\alpha_i$, then the vector $(\kappa_{ij}/\omega_{ij}|c_i = h, \omega_{ij}) \sim \mathcal{N}(\mathbf{x}_i^T\boldsymbol{\beta}^{(k)}, 1/\omega_{ij})$, and each cluster-specific coefficient vector $\boldsymbol{\beta}_h$ can be updated by aggregating all observations and augmented data relative to birth defects that are in the same cluster.

for $k = 1, \ldots, K$ **do**

Let $\mathbf{X}^{(k)}$, $\mathbf{y}^{(k)}$, $\boldsymbol{\kappa}^{(k)}$ be the obtained quantities relative to cluster h, and $\boldsymbol{\Omega}^{(k)}$ a diagonal matrix with the corresponding Pólya-gamma augmented variables. Then update cluster-specific coefficients vector from

$$(\boldsymbol{\beta}^{(k)}|-) \sim \mathcal{N}_p(\mathbf{b}^{(k)}, \mathbf{Q}^{(k)})$$

with $\mathbf{Q}^{(k)} = (\mathbf{X}^{(h)T}\boldsymbol{\Omega}^{(k)}\mathbf{X}^{(k)} + \mathbf{Q}^{-1})^{-1}$ and $\mathbf{b}^{(k)} = \mathbf{Q}^{(k)}(\mathbf{X}^{(k)T}\boldsymbol{\kappa}^{(k)} + \mathbf{Q}^{-1}\mathbf{b})$.

end for

---

[4] Allocate each birth defect i to one of the clusters

for $i = 1, \ldots, N$ **do**

Sample the class indicator c_i conditionally on $\boldsymbol{c}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$ from the discrete distribution with probabilities

$$\Pr(c_i = k|\boldsymbol{c}_{-i}, -) \propto \Pr(c_i = k|\boldsymbol{c}_{-i})\Pr(\mathbf{y}_i|\mathbf{X}_i, \alpha_i, c_i = k, \boldsymbol{\beta}^{(k)})$$

with

$$\Pr(\mathbf{y}_i|\mathbf{X}_i, \alpha_i, c_j = k, \boldsymbol{\beta}^{(k)}) = \prod_{j=1}^{n_i} \left[\exp(\alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta}^{(k)})^{y_{ij}}\right] \left[1 + \exp(\alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta}^{(k)})\right]^{(-1)}$$

being the model likelihood evaluated for cluster h and $\Pr(c_i = h|\boldsymbol{c}^{-(i)})$ computed as described in Section 3.2 depending on the base EPPF.

end for

4.2 Simulation studies

We conduct a simulation study to evaluate the performance of our approach in accurately estimating the impact of the covariates across regressions. In simulating data, we choose a scenario mimicking the structure of our application, producing datasets with highly variable dimensions, favoring a low number of observations. We consider 4 underlying groups and generate 12 equally divided datasets, with $\{n_1, n_2, n_3\} = \{100, 600, 200\}$, $\{n_4, n_5, n_6\} = \{300, 100, 100\}$, $\{n_7, n_8, n_9\} = \{500, 100, 200\}$, $\{n_{10}, n_{11}, n_{12}\} = \{200, 200, 200\}$ and $p = 10$ dichotomous explanatory variables. Each data matrix $\mathbf{X}_i$ for $i = 1, \ldots, 12$ is generated by sampling each of the variables from a Bernoulli distribution with probability of success equal to 0.5, and fixing most of coefficients β_i, for $i = 1, \ldots, 10$, to 0, while defining a challenging scenario with small to moderate changes across different groups.

In particular we fix $\{\beta_1, \beta_2, \beta_3, \beta_4\} = \{\log(1.3), \log(1.2), \log(1.2), \log(1.3)\}$ for group 1, $\{\beta_4, \beta_5, \beta_6\} = \{\log(1.5), \log(0.7), \log(1.5)\}$ for group 2, $\{\beta_9, \beta_{10}\} = \{\log(1.5), \log(0.5)\}$

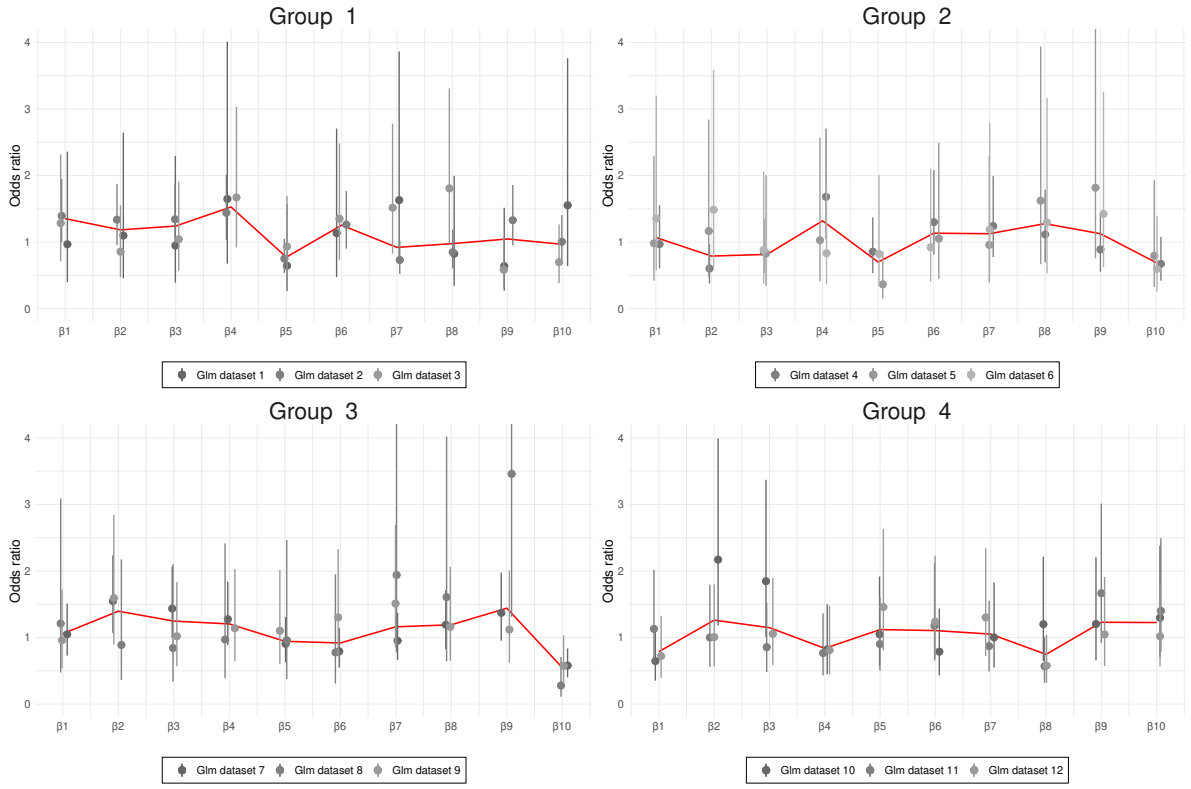


FIGURE 4.1: Estimates of the coefficients obtained using separate logistic regressions. For each group, the red line correspond to maximum likelihood estimates obtained from logistic regression using all the observation generated for the groups. Instead grey dots corresponds to maximum likelihood estimates obtained from separate regressions, with bars indicating confidence intervals at 95%. Wider confidence intervals corresponds to dataset with fewer observations.

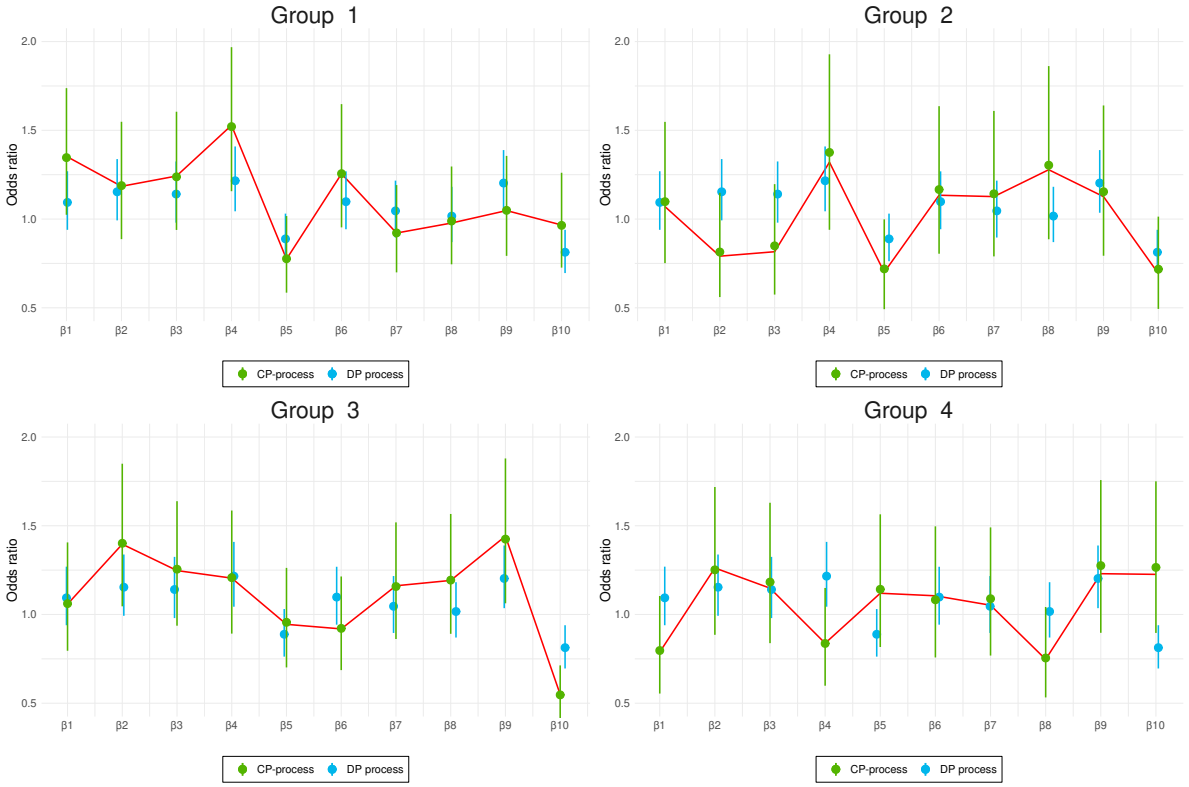


FIGURE 4.2: Estimates of the coefficients obtained using a grouped logistic regression with Dirichlet process prior and CP process prior centered on the true partition ($\psi = 20$). Red lines correspond to maximum likelihood estimates obtained from logistic regression using the true grouping. Bars indicates credibility intervals at 95%.

for group 3 and $\{\beta_1, \beta_2, \beta_9, \beta_{10}\} = \{\log(0.8), \log(1.2), \log(1.5), \log(1.2)\}$ for group 4. We consider the following models: i) separate logistic regressions, ii) a grouped logistic regression with a DP prior with $\alpha = 1$ and iii) a grouped logistic regression using a CP prior with a DP base EPPF with concentration parameter $\alpha = 1$. We fixed $\psi = 20$ according to the considerations made in Section 3.3, and evaluate the CP prior behavior both centering it on the true known grouping and on a wrong guess of the partition. Posterior estimates are obtained using the Gibbs sampler described in the Appendix. We run the algorithm for $5,000$ iterations and used a burn-in of $1,000$, with inspection of trace-plots suggesting convergence of the parameters, and estimate the partition as as maximum a posteriori. We consider a multivariate normal distribution with zero mean vector and covariance matrix $\mathbf{Q} = \text{diag}_p(2)$ as base measure for the Dirichlet process, while we assume the defect-specific intercepts $\alpha_i \sim N(0, 2)$ for $i = 1, \ldots, N$.

We first centered the CP prior on the true known grouping and resulting estimates are shown in Figures 4.1-4.2 for each of the true groups, with corresponding 95% confidence and credibility intervals for maximum likelihood and Bayesian estimates, respectively. Maximum likelihood estimates have quite large confidence intervals, especially

for datasets with a small number of observations. The grouped logistic regression using the Dirichlet Process prior, although borrowing information across the datasets, does not distinguish between the groups, while the CP process recovers the true grouping with good performances in estimating the coefficients.

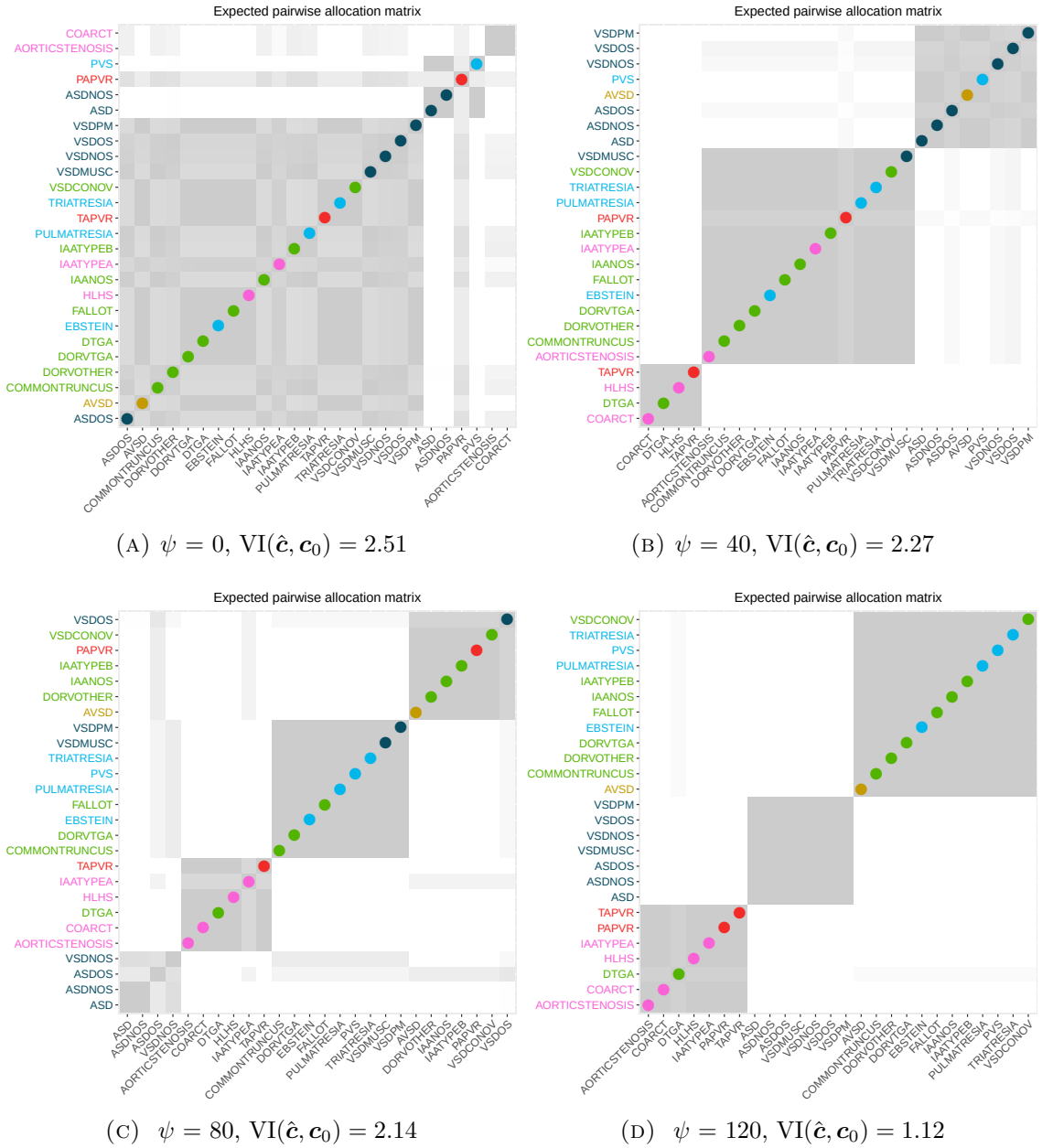
We also evaluated the CP prior performances when centered on a wrong guess $\boldsymbol{c}_0'$ of the base partition. In particular, we set $\boldsymbol{c}_0' = \{1,5,9\}\{2,6,10\}\{3,7,11\}\{4,8,12\}$. Despite having the same configuration of $\boldsymbol{c}_0$, it has distance from $\boldsymbol{c}_0$ in terms of VI of approximately 3.16, where the maximum possible distance is $\log_2(12) = 4.70$. In this case, we estimate 2 clusters, with the first one comprising a dataset for each of the groups $2, 3, 4$, and the second putting together all the remaining datasets. Although the estimated clustering is quite different from $\boldsymbol{c}_0$, having distance 2.58, the estimated partition is closer to the one induced by the DP (0.81) than $\boldsymbol{c}_0'$ (1.18). In this case, partitions distances may provide an indication about the goodness of our prior guess.

4.3 Application to NBDPS data

We estimated the model in (4.3) on the NBDPS data, considering the controls as shared with the aim of grouping cases into informed groups on the basis of the available $\boldsymbol{c}_0$. In order to choose a value for the penalization parameter, we consider the prior calibration illustrated in Section 3.3, finding a value of $\psi = 40$ assigning a 90% probability to partitions within a distance around 0.8, where the maximum possible distance is equal to 4.70. In terms of moves on the Hasse diagram we are assigning 90% prior probability to partitions at most at 11 split/merge operations, given that the minimum distance from $\boldsymbol{c}_0$ is $2/N \approx 0.07$. To assess sensitivity of the results, we performed the analysis under different values of $\psi \in \{0, 40, 80, 120, \infty\}$. In particular, for $\psi = 0$ the clustering behavior is governed by a Dirichlet process prior, while $\psi \to \infty$ corresponds to fixing the groups to $\boldsymbol{c}_0$.

In analyzing the data we run the Gibbs sampler for $10,000$ iterations and use a burn-in of $4,000$, under the same prior settings as in Section 4.2. Figure 4.3 summarizes the posterior estimates of the pairwise allocation matrices under different values of ψ with partition estimated as maximum at posteriori. Colored dots emphasize differences with the base partition $\boldsymbol{c}_0$. Under the DP process ($\psi = 0$) the estimated partition differs substantially from the given prior clustering.

Due to the immense space of the possible clusterings, this is likely reflective of limited information in the data, combined with the tendency of the DP to strongly favor certain types of partitions, typically characterized from few large clusters along many small ones (Miller and Harrison, 2018). When increasing the value of the tuning parameter ψ the estimated clustering is closer to $\boldsymbol{c}_0$. In particular, for $\psi = 120$ one of the clusters in $\boldsymbol{c}_0$ is recovered, while the others are merged in two different groups.



(A) $\psi = 0$, $\text{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 2.51$

(B) $\psi = 40$, $\text{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 2.27$

(C) $\psi = 80$, $\text{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 2.14$

(D) $\psi = 120$, $\text{VI}(\hat{\boldsymbol{c}}, \boldsymbol{c}_0) = 1.12$

FIGURE 4.3: Posterior allocation matrices obtained using the CP prior with DP($\alpha = 1$) for different values of $\psi \in \{0, 40, 80, 120\}$. On the y-axis labels are colored according base grouping information $\boldsymbol{c}_0$, with dots on the diagonal highlighting differences between $\boldsymbol{c}_0$ and the estimated partition $\hat{\boldsymbol{c}}$.

Details on the results for each of the estimated models are given in at the end of the chapter (Figures 4.5-4.9) and summarized here. Figure 4.4 shows a heatmap of the mean posterior log odds-ratios for increasing values of the penalization parameter ψ, with dots indicating if they are significant according to a 95% credibility interval. In general, the sign of the effects does not change for most of the exposure factors across the different clusterings. Figure 4.4 focuses on pharmaceutical use in the period from 1 month before the pregnancy and 3 months during, along with some exposures related to maternal behavior and health status.

We found consistent results for known risk factors for CHD in general, including for diabetes (Correa *et al.*, 2008) and obesity (Waller *et al.*, 2007). The finding that nausea is associated with positive outcomes is consistent with prior literature (Koren *et al.*, 2014). The association between use of SSRIs and pulmonary atresia was also noted in Reefhuis *et al.* (2015a). It is worth noticing that estimates obtained under the DP prior are less consistent with prior work. Also, there apparent artifacts such as the protective effect of alcohol consumption, which is mitigated from an informed borrowing across the defects. On the other side, estimates for the AVSD or PAPVR, which corresponds to 0.02% and 0.01% of cases respectively, show how a separate analysis of cases with low prevalence misses even widely assessed risk factors, as for example diabetes.

It is also worth noticing that, since we are analyzing data pertaining the same class of

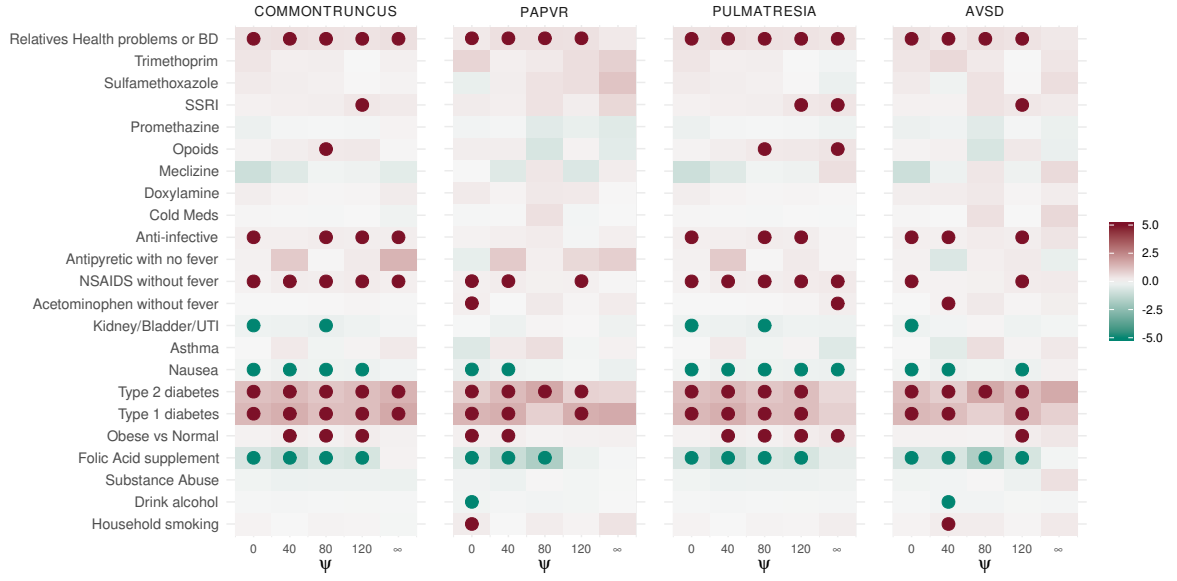


FIGURE 4.4: Comparison of significant odds ratio under $\psi \in \{0, 40, 80, 120, \infty\}$ for some exposure factors and 4 selected heart defects in 4 different groups under $\boldsymbol{c}_0$. Dots are in correspondence of significant mean posterior log odds ratio (log-OR) at 95% with red encoding risk factors (log-OR > 0) and green protective factors (log-OR < 0).

defects, many of the estimated effects are pretty similar across different data partitions. An immediate extension of the model in 4.3 would consists in modifying the prior distribution for the coefficient vector to allow a selection procedure for the covariates, by distinguishing between common effects and group-specific ones. This may lead to a better comprehension of which might be interesting factor to better monitor during pregnancy, and contribute to the development of tools for early diagnosis.

Additionally we stress that in this application we employed a model belonging to the class of infinite mixture models, given that we have specified a Dirichlet process baseline EPPF. A finite Dirichlet for the baseline EPPF could be a suitable alternative, given that in the NBDPS the number of defects to monitor is fixed a by the study requirements. However the chosen framework results to be more general, allowing for inclusion of new defects with motivation drawn from new versions of study currently running.

Additional results for NBDPS data application

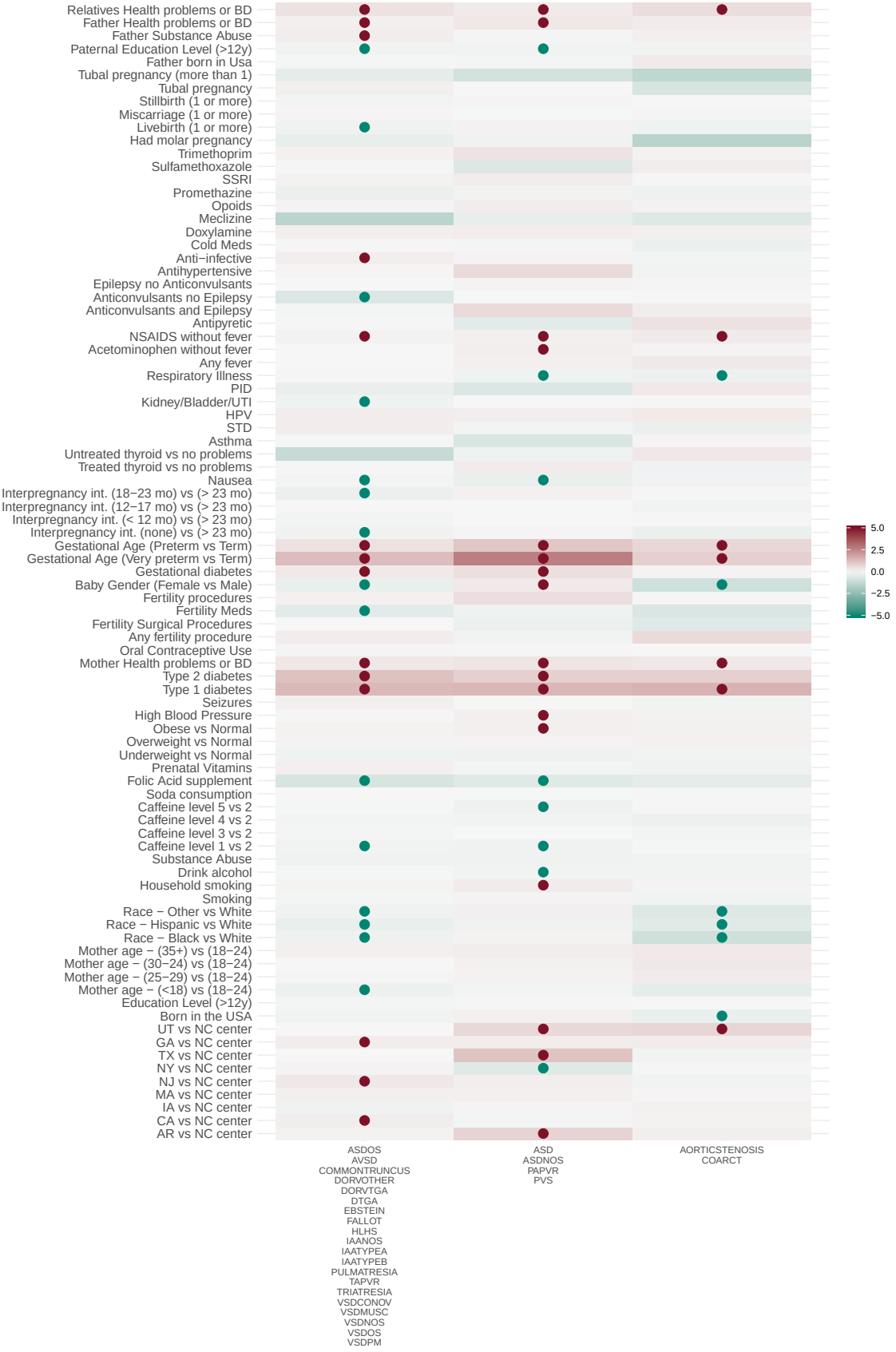


FIGURE 4.5: **CP process with $\psi = 0$.** Posterior mean estimates of log odds-ratios, with values shown if significant at 95% using credibility intervals. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor with green a protective effect.

FIGURE 4.6: **CP process with $\psi = 40$.** Posterior mean estimates of log odds-ratios, with values shown if significant at 95% using credibility intervals. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor with green a protective effect.

FIGURE 4.7: **CP process with $\psi = 80$.** Posterior mean estimates of log odds-ratios, with values shown if significant at 95% using credibility intervals. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor with green a protective effect.

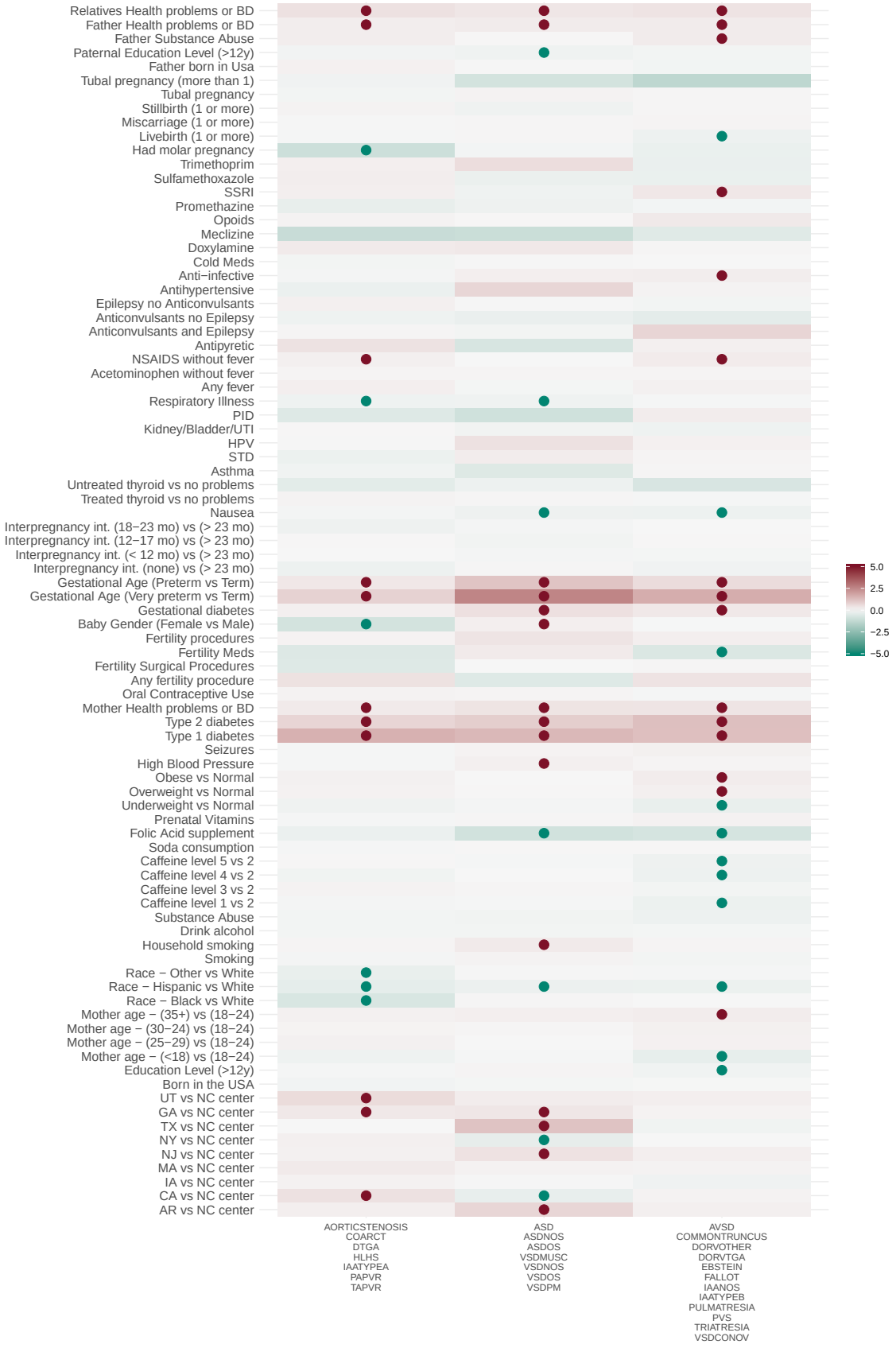


FIGURE 4.8: **CP process with** $\psi = 120$. Posterior mean estimates of log odds-ratios, with values shown if significant at 95% using credibility intervals. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor with green a protective effect.

FIGURE 4.9: **CP process with $\psi = \infty$.** Posterior mean estimates of log odds-ratios, with values shown if significant at 95% using credibility intervals. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor with green a protective effect.

Conclusions

Discussion

There is a very rich literature on priors for clustering, with almost all of the emphasis on exchangeable approaches and a smaller literature focused on including dependence on known features (e.g. temporal or spatial structure or covariates). The main contribution of this thesis is to propose what is seemingly a first attempt at including prior information on an informed guess at the clustering structure. We were particularly motivated by a concrete application on birth defects study in proposing our method, which is based on shrinking an initial clustering prior towards the prior guess.

There are many immediate interesting directions for future research. One thread pertains to developing better theoretical insight and analytical tractability into the new class of priors. For existing approaches, such as product partition models and Gibbs-type partitions, there is a substantial literature providing simple forms of prediction rules and other properties. It is an open question whether such properties can be modified to our new class. This may yield additional insight into the relative roles of the base prior, centering value and hyperparameters in controlling the behavior of the prior and its impact on the posterior.

Another important thread relates to applications of the proposed framework beyond the setting in which we have an exact guess at the complete clustering structure. In many cases, we may have an informed guess or initial clustering in a subset of the objects under study, with the remaining objects (including future ones) completely unknown. Conceptually the proposed approach can be used directly in such cases, and also when one has different types of prior information on the clustering structure than simply which objects are clustered together.

Bibliography

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K. and Rousseau, J. (2012) Combining expert opinions in prior elicitation. *Bayesian Analysis* **7**(3), 503–532.

Arbel, J. and Favaro, S. (2017) Approximating predictive probabilities of gibbs-type priors. *arXiv preprint arXiv:1707.08053* .

Barrientos, A. F., Jara, A., Quintana, F. A. *et al.* (2012) On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* **7**(2), 277–310.

Barry, D. and Hartigan, J. A. (1992) Product partition models for change point problems. *The Annals of Statistics* pp. 260–279.

Booth, J. G., Casella, G. and Hobert, J. P. (2008) Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society - Series B* **70**(1), 119–139.

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., Correa, A. and National Birth Defects Prevention Study (2007) Seeking causes: classifying and evaluating congenital hearth defects in etiologic studies. *Birth Defects Research Part A: Clinical and Molecular Teratology* **79**(10), 714–727.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2006) Bayesian inference for dynamic models with dirichlet process mixtures. In *International Conference on Information Fusion*. Florence, Italy.

Casella, G., Moreno, E., Girón, F. J. *et al.* (2014) Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis* **9**(3), 613–658.

Cesari, O., Favaro, S. and Nipoti, B. (2014) Posterior analysis of rare variants in Gibbs-type species sampling models. *Journal of Multivariate Analysis* **131**, 79 – 98.

Choi, H.-s. (2016) Expert information and nonparametric bayesian inference of rare events. *Bayesian Analysis* **11**(2), 421–445.

Clark, E. B. (2001) Etiology of congenital cardiovascular malformations: epidemiology and genetics. In *Moss and Adams' Heart Disease in Infants, Children and Adolescents*, pp. 64–79. Allen H., Cark E., Gutgesell H., Driscoll D., editors.

Cogswell, M. E., Bitsko, R. H., Anderka, M., Caton, A. R., Feldkamp, M. L., Sherlock, S. M. H., Meyer, R. E., Ramadhani, T., Robbins, J. M., Shaw, G. M., Mathews, T. J., Royle, M., Reefhuis, J. and the National Birth Defects Prevention Study (2009) Control selection and participation in an ongoing, population-based, case-control study of birth defects: the National Birth Defects Prevention Study. *American Journal of Epidemiology* **170**(8), 975–985.

Correa, A., Gilboa, S. M., Besser, L. M., Botto, L. D., Moore, C. A., Hobbs, C. A., Cleves, M. A., Riehle-Colarusso, T. J., Waller, D. K., Reece, E. A. *et al.* (2008) Diabetes mellitus and birth defects. *American journal of Obstetrics and Gynecology* **199**(3), 237.e1–237.e9.

Dahl, D. B., Day, R. and Tsai, J. W. (2017) Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association* **112**(518), 721–732.

Davey, B. A. and Priestley, H. A. (2002) *Introduction to Lattices and Order*. Cambridge university press.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2015) Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE transactions on pattern analysis and machine intelligence* **37**(2), 212–229.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An anova model for dependent random measures. *Journal of the American Statistical Association* **99**(465), 205–215.

Deza, M. M. and Deza, E. (2009) Encyclopedia of Distances. In *Encyclopedia of Distances*, pp. 1–583. Springer.

Dobiński, G. (1877) Summirung der reihe $\sum$ nm/n! für m= 1, 2, 3, 4, 5, ... *Archiv der Mathematik und Physik* **61**, 333–336.

Dunson, D. B. and Park, J.-H. (2008) Kernel stick-breaking processes. *Biometrika* **95**(2), 307–323.

Fall, M. D. and Barat, É. (2014) Gibbs sampling methods for Pitman-Yor mixture models. working paper or preprint.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**(2), 209–230.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**(471), 1021–1035.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gill, J. and Walker, L. D. (2005) Elicited priors for bayesian model specifications in political science research. *The Journal of Politics* **67**(3), 841–872.

Gnedin, A. and Pitman, J. (2006) Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences* **138**(3), 5674–5685.

Grätzer, G. (2002) *General Lattice Theory*. Springer Science & Business Media.

Griffin, J. E. and Steel, M. F. (2006) Order-based dependent dirichlet processes. *Journal of the American Statistical Association* **101**(473), 179–194.

Hartigan, J. (1990) Partition models. *Communications in Statistics - Theory and Methods* **19**(8), 2745–2756.

Jensen, S. T. and Liu, J. S. (2008) Bayesian clustering of transcription factor binding motifs. *Journal of the American Statistical Association* **103**(481), 188–200.

Kamvar, S. D., Klein, D. and Manning, C. D. (2003) Spectral learning. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pp. 561–566. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Khoury, M. J., Moore, C. A., James, L. M. and Cordero, J. F. (1992) The interaction between dysmorphology and epidemiology: methodologic issues of lumping and splitting. *Teratology* **45**(2), 133–138.

Klein, D., Kamvar, S. D. and Manning, C. D. (2002) From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 307–314. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-55860-873-7.

Knuth, D. (2006) *The Art of Computer Programming: Generating All Trees. History of Combinatorial Generation.* Addison-Wesley.

Knuth, D. E. (2005) *The Art of Computer Programming. Generating all combinations and partitions.* Addison-Wesley.

Koren, G., Madjunkova, S. and Maltepe, C. (2014) The protective effects of nausea and vomiting of pregnancy against adverse fetal outcome. a systematic review. *Reproductive Toxicology* **47**, 77 – 80.

Kuo, L. (1986) Computations of mixtures of dirichlet processes. *SIAM Journal on Scientific and Statistical Computing* **7**(1), 60–71.

Law, M. H., Topchy, A. and Jain, A. K. (2004) Clustering with soft and group constraints. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 662–670.

Lin, A. E., Herring, A. H., Amstutz, K. S., Westgate, M.-N., Lacro, R. V., Al-Jufan, M., Ryan, L. and Holmes, L. B. (1999) Cardiovascular malformations: changes in prevalence and birth status, 1972–1990. *American Journal of Medical Genetics* **84**(2), 102–110.

Lo, A. Y. (1984) On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* pp. 351–357.

MacEachern, S. N. (1999) Dependent nonparametric processes. In *Proceedings of the Bayesian Section.*, pp. 50–55. Alexandria, VA: American Statistical Association.

MacEachern, S. N. (2000) Dependent nonparametric processes. Technical report, Department of Statistics, The Ohio State University.

MacLehose, R. F. and Dunson, D. B. (2010) Bayesian semiparametric multiple shrinkage. *Biometrics* **66**(2), 455–462.

Mazanec, J. A. (1997) Segmenting city tourists into vacation styles. *International City Tourism: Analysis and Strategy* pp. 114–128.

Meilă, M. (2007) Comparing clusterings - an information based distance. *Journal of Multivariate Analysis* **98**(5), 873 – 895.

Miller, J. W. and Harrison, M. T. (2018) Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**(521), 340–356.

Müller, P., Quintana, F. and Rosner, G. L. (2011) A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics* **20**(1), 260–278.

Murphy, S. L., Xu, J., Kochanek, K. D., Curtin, S. C. and Arias, E. (2017) Deaths: Final data for 2015. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* **66**(6), 1.

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

Nijenhuis, A. and Wilf, H. S. (2014) *Combinatorial Algorithms: for Computers and Calculators*. Elsevier.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons.

Park, J.-h. and Dunson, D. B. (2010) Bayesian generalize product partition models. *Statistica Sinica* **20**, 1203–1226.

Petrone, S., Guindani, M. and Gelfand, A. E. (2009) Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(4), 755–782.

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**(2), 145–158.

Pitman, J. (1997) Some probabilistic aspects of set partitions. *The American Mathematical Monthly* **104**(3), 201–209.

Pitman, J. and Yor, M. (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**(2), 855–900.

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**(504), 1339–1349.

Rasmussen, S. A., Olney, R. S., Holmes, L. B., Lin, A. E., Keppler-Noreuil, K. M. and Moore, C. A. (2003) Guidelines for case classification for the National Birth Defects

Prevention Study. *Birth Defects Research Part A: Clinical and Molecular Teratology* **67**(3), 193–201.

Reefhuis, J., Devine, O., Friedman, J. M., Louik, C. and Honein, M. A. (2015a) Specific ssris and birth defects: bayesian analysis to interpret new data in the context of previous reports. *British Medical Journal* **351**.

Reefhuis, J., Gilboa, S. M., Anderka, M., Browne, M. L., Feldkamp, M. L., Hobbs, C. A., Jenkins, M. M., Langlois, P. H., Newsome, K. B., Olshan, A. F. *et al.* (2015b) The National Birth Defects Prevention Study: a review of the methods. *Birth Defects Research Part A: Clinical and Molecular Teratology* **103**(8), 656–669.

Rodriguez, A. and Dunson, D. B. (2011) Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**(1).

Rossi, G. (2015) Weighted paths between partitions. *arXiv preprint* .

Sarup, P., Jensen, J., Ostersen, T., Henryon, M. and Sørensen, P. (2016) Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred danish duroc pigs. *BMC Genetics* **17**(1), 11.

Scarpa, B. and Dunson, D. B. (2009) Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics* **65**(3), 772–780.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statistica Sinica* **4**(2), 639–650.

Shental, N., Bar-Hillel, A., Hertz, T. and Weinshall, D. (2004) Computing gaussian mixture models with em using equivalence constraints. In *Advances in Neural Information Processing Systems*, pp. 465–472.

Spetzler, C. S. and Stael von Holstein, C.-A. S. (1975) Exceptional paper - probability encoding in decision analysis. *Management Science* **22**(3), 340–358.

Stam, A. (1983) Generation of a random partition of a finite set by an urn model. *Journal of Combinatorial Theory, Series A* **35**(2), 231–240.

Stanley, R. P. (1997) *Enumerative Combinatorics. Vol. 1.* Cambridge University Press.

Vinh, N. X., Epps, J. and Bailey, J. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* **11**(Oct), 2837–2854.

Wade, S. and Ghahramani, Z. (2018) Bayesian cluster analysis: point estimation and credible balls (with discussion). *Bayesian Analysis* **13**(2), 559–626.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S. *et al.* (2001) Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pp. 577–584.

Waller, D. K., Shaw, G. M., Rasmussen, S. A., Hobbs, C. A., Canfield, M. A., Siega-Riz, A.-M., Gallaway, M. S. and Correa, A. (2007) Prepregnancy obesity as a risk factor for structural birth defects. *Archives of Pediatrics & Adolescent Medicine* **161**(8), 745–750.

Yoon, P. W., Rasmussen, S. A., Lynberg, M. C., Moore, C. A., Anderka, M., Carmichael, S. L., Costa, P., Druschel, C., Hobbs, C. A., Romitti, P. A., Langlois, P. H. and Edmonds, L. D. (2001) The National Birth Defects Prevention Study. *Public Health Reports* **116**, 32–40.