Domain knowledge based priors for clustering

Distribuzioni a priori per l'analisi di raggruppamento basate sulla conoscenza di settore

Sally Paganin

Abstract The construction of informative priors based on domain knowledge is a delicate problem, complicated by the fact that the human mind finds it difficult to quantify qualitative knowledge. We focus on the situation in which a prior guess of the data partition is provided, and illustrate how to include such information in a Bayesian mixture model framework. The methodology builds on class of perturbed EPPFs (Exchangeable Partition Probability Function) which centers the prior probability on the most compatible set of partitions, according to the provided guess.

Abstract *La costruzione di distribuzioni a priori informative basate sulla conoscenza di settore è una problematica delicata, complicata dal fatto che la mente umana trova difficoltà nel quantificare le conoscenze di tipo qualitativo. Questo lavoro prende in esame il contesto in cui si ha disposizione una plausibile proposta di partizione dei dati, e illustra come includere tale informazione in modelli di mistura di tipo bayesiano. La metodologia considerata si basa su una classe di EPPF (Exchangeable Partition Probability Function) penalizzate che centrano la distribuzione di probabilità a priori intorno all'insieme di partizioni maggiormente compatibili con la partizione data.*

Key words: Bayesian clustering, centered process, domain knowledge, partition models.

1 Introduction

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. In particular, observations $y_1, \ldots, y_N$ can be divided

Sally Paganin
Department of Statistical Sciences, University of Padova, via Cesare Battisti 241, 35121 Padova, e-mail: paganin@stat.unipd.it

into $K \leq N$ groups, according to a partition $c = \{B_1, \dots, B_K\}$ with B_k comprising all the indices of data points in cluster k, for $k = 1, \dots, K$. The main underlying assumption of a mixture model is that observations are independent conditional on the partition c and on the vector of unknown parameters $\theta = (\theta_1, \dots, \theta_K)$ indexing the distribution of observations within each cluster. Hence the joint probability density of observations $y_1, \dots, y_N$ can be expressed as

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

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

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

The data partition c is conceived as a random object and elicitation of its prior distribution is a critical issue in Bayesian modeling since the space of all possible partitions grows exponentially fast given its combinatorial nature. Current Bayesian methods often relies on Species Sampling Models (SSM) [7], which avoid dealing with the clustering space directly by inducing a latent partitioning of the data. The induced probability distribution is known in literature as Exchangeable Partition Probability Function (EPPF).

Despite providing tractable tools to deal with mixture models, Bayesian nonparametric 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. In particular we focus on the situation in which a base partition c_0 is provided as a prior guess, and we wish to include this information in the prior distribution. To address this problem [6] propose a general strategy to modify a baseline EPPF to shrink the prior probability on partitions towards c_0. In particular, the prior distribution on all the possible clusterings is defined as proportional to a baseline EPPF multiplied by a penalization term of the type

$$p(c|c_0, \psi) \propto p_0(c) e^{-\psi d(c,c_0)}, \tag{1}$$

with $\psi > 0$ a tuning parameter, $d(c, c_0)$ a suitable distance measuring how far c is from c_0 and $p_0(c)$ indicates a baseline EPPF, that may depend on some parameters. Notice that as $\psi \to 0$ then $p(c|c_0, \psi)$ corresponds to the baseline EPPF $p_0(c)$, while as $\psi \to \infty$ then $p(c = c_0) \to 1$.

The general formulation given in (1) leads to different results on the basis of different choices of EPPF, tuning parameter and distance between partitions. While we refer to [6] for considerations about the choice of EPPFs and tuning parameter, this work focus on characterizing the distance term by providing the definition of a class of a suitable metric between partitions, along with a characterization of

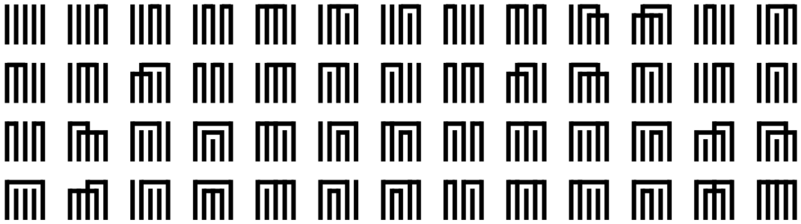

Fig. 1 Genji-mon symbols for all the possible grouping of 5 elements.

neighborhoods induced by the distance. In the following, Section 2 introduces the mathematical definition of set partitions along with concepts derived from lattice theory while Section 3 gives the general definition of distance between partitions. Finally Section 4 provides a characterization of the distance neighborhoods induced by the distance.

2 Set partitions

Let c be a generic clustering of indices $\{1,\ldots,N\} = [N]$. It can be either represented as a vector of indices $\{c_1,\ldots,c_N\}$ with $c_i \in \{1,\ldots,K\}$ for $i = 1,\ldots,N$ and $c_i = c_j$ when i and j belong to the same cluster, or as a collection of disjoint subsets (blocks) $\{B_1,B_2,\ldots,B_K\}$ 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 $c = \{B_1,\ldots,B_K\}$ 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 [8, 1] for an introduction to lattice theory, reporting here some of the base concepts.

According to [4], 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 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*

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

while the *Bell number* $\mathscr{B}_N = \sum_{K=1}^{N}\mathscr{S}_{N,K}$ describes the number of all possible set partitions of N elements. Refer to [4] for more information on history and algorithms related to set partitions and other combinatorial objects.

2.1 Poset representation of 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 $c = \{B_1,\dots,B_K\}, c' = \{B'_1,\dots,B'_{K'}\}$ belonging to Π_N, $c \leq c'$ 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 $c \in \Pi_N$, $c \leq c$,
2. Antisymmetry: if $c \leq c'$ and $c' \leq c$, then $c = c'$,
3. Transitivity: if $c \leq c'$ and $c' \leq c''$, then $c \leq c''$.

Let $<$ be the relation on Π_N such that $c < c'$ if and only if $c \leq c'$ and $c \neq c'$. For any $c, c' \in \Pi_N$, it is said that c is *covered* (or refined) by c' if $c \leq c'$ and there is no c'' such that $c < c'' < c'$ and indicate with $c \prec c'$ 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 c to c' when $c \prec c'$; in other words, there is a connection from a partition c to another one when the second can be obtained from the first by splitting or merging one of the blocks in c. See Figure 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 1, while the partition having every observation in its own cluster at the bottom and indicated with 0.

The space Π_N is also 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. $c \wedge c' = g.l.b.(c,c')$ and $c \vee c' = l.u.b.(c,c')$ and equality holds under a permutation of the cluster labels. An element $c \in \Pi_N$ is an upper bound for a subset $S \subseteq \Pi_N$ if $s \leq c$ for all $s \in S$, and it is the least

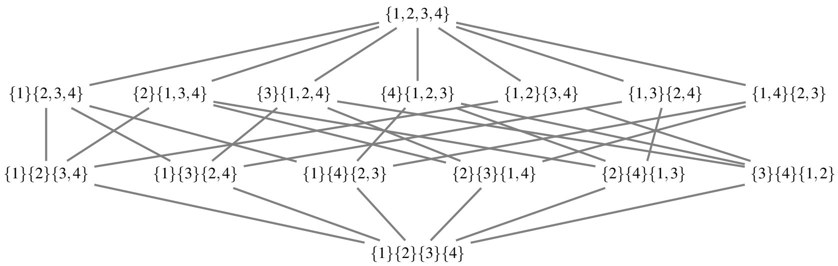

Fig. 2 Hasse diagram for the lattice of set partitions of 4 elements. A line is drawn when two partitions have a covering relation. For example $\{1\}\{2,3,4\}$ is connected with 3 partitions obtained by splitting the block $\{2,3,4\}$ in every possible way, and partition **1** obtained by merging the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

upper bound for a subset $S \subseteq \Pi_N$ if c is an upper bound for S and $c \leq c'$ for all upper bounds c' of 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 $c = \{1\}\{2,3,4\}, c' = \{3\}\{1,2,4\}$; their greatest lower bound (g.l.b.) is $c \wedge c' = \{1\}\{3\}\{2,4\}$ while the least upper bound (l.u.b.) is $c \vee c' = \{1,2,3,4\}$. Looking at the Hasse diagram in Fig 2 the g.l.b. and l.u.b. are in general the two partitions which reach both c and c' through the shortest path, respectively from below and from above.

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

We report here a useful result from lattice theory referring to [3] and [2]. Given a lattice function ν weights w_ν on edges between $\{c,c'\}$ are defined as

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

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 the function ν 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*

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

while if ν is submodular

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

4 Distance neighborhoods on the partition lattice

Due to the discrete nature of the space of partition, the distance $d_\nu(c,c_0)$ takes a finite number of discrete values $\Delta = \{\delta_0,\dots,\delta_L\}$, with L depending on c_0 and on the distance $d(\cdot,\cdot)$. We can define distance neighborhoods as

$$s_l(c_0) = \{c \in \Pi_N : d_\nu(c,c_0) = \delta_l\}, \quad l = 0,1,\dots,L, \tag{2}$$

hence sets of partitions having the same fixed distance from c_0. For $\delta_0 = 0$, $s_0(c_0)$ denotes the set of partitions equal to the base one, meaning that they differ from c_0 only by a permutation of the cluster labels. Then $s_1(c_0)$ denotes the set of partitions with minimum distance δ_1 from c_0, $s_2(c_0)$ the set of partitions with the second minimum distance δ_2 from c_0 and so on. Hence the exponential term in (1) penalizes equally partitions in the same set $s_l(c_0)$ for a given δ_l.

A trivial example can be obtained by considering the rank function, i.e. $r(\cdot) : \Pi_N \to \mathbb{Z}^+$ such that $r(c) = N - |c|$, which is a strictly order-preserving lattice function. For example, considering partitions in the Hasse diagram in Figure 2, the rank of the bottom partition 0 is equal to 0 and increases by 1 for each level of the graph up to 3 for top partition 1. Then the minimum-rank-weighted distance can be computed as $d_r(c,c') = 2r(c \vee c') - r(c) - r(c')$, 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.

Figure 3 provides a representation of the distance neighborhoods as defined in 2 induced by the rank function when the base partition corresponds to $c_0 =$

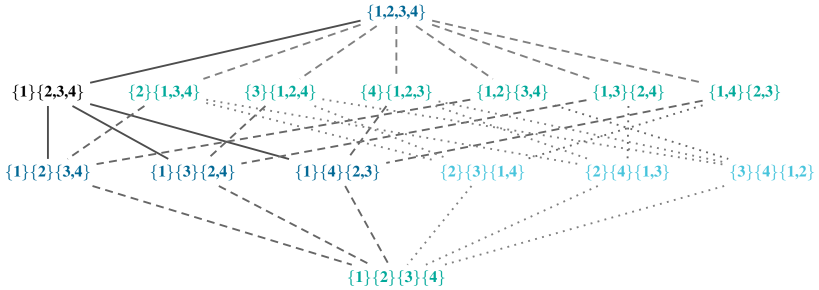

Fig. 3 Representation of distance neighborhoods on the poset lattice $\mathit{\Pi}_4$ for $c_0 = \{\mathbf{1}\}\{\mathbf{2,3,4}\}$ when the chosen distance is induced from the rank function. Partitions are colored from the closest to the most distant according to dark-light gradient.

$\{1\}\{2,3,4\}$. It can be notice that the closest partitions, besides c_0 itself, in $s_1(c_0)$ are the ones obtained with a single operation of split or merge on c_0, while the second closest ones by applying another operation of split or merge on the partitions in $s_1(c_0)$, and so on. This kind of behavior can be observed for all functions which induce unit weight on the edges of the partition lattice.

Another important measure of distance between two partitions, is the Variation of Information (VI), introduced axiomatically in information theory [5], which also belongs to the class of distances derived from a lattice function. In particular, consider the Shannon entropy $H(\cdot) : \mathit{\Pi}_N \to \mathbb{R}^+$ defined as $H(c) = -\sum_{i=1}^{K} |B_i|/N \log_2(|B_i|/N)$ which is a submodular and strictly order-reversing function, hence inducing distance

$$d_H(c,c') = VI(c,c') = 2H(c \wedge c') - H(c) - H(c'). \tag{3}$$

The VI ranges on a finite subset in $[0, \log_2 N]$, and in this case the weights assigned to the edges differs from the unit weight, leading to finer characterization of the distance neighborhoods as it can be seen from Figure 4. In general the closest partitions are the ones which differs from c_0 by merging two singleton clusters or splitting a cluster of size two into singletons. If neither is possible, the closest partitions differs from c_0 by a split operation on the smallest cluster of size k into a singleton and a cluster of size $k-1$ or, as in the example, by a merge operations on these last two clusters.

References

1. Davey, B. A. and Priestley, H. A.: Introduction to Lattices and Order. Cambridge university press. (2002)
2. Deza, M. M. and Deza, E.: Encyclopedia of Distances. Springer Berlin Heidelberg. (2009)
3. Grätzer, G.: General Lattice Theory. Springer Science & Business Media. (2002)
4. Knuth, D.: The Art of Computer Programming: Generating All Trees. History of Combinatorial Generation. Addison-Wesley. (2006)

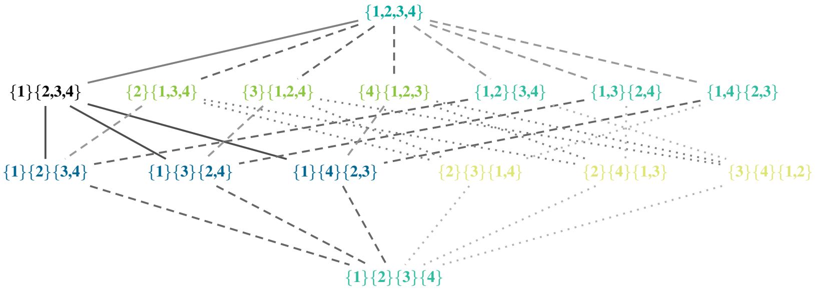

Fig. 4 Representation of distance neighborhoods on the poset lattice Π_4 for $c_0 = \{\mathbf{1}\}\{\mathbf{2,3,4}\}$ when the chosen distance is induced from the Shannon entropy function. Partitions are colored from the closest to the most distant according to dark-light gradient.

5. Meilă, M.: Comparing clusterings - an information based distance. *Journal of Multivariate Analysis* **98**(5), 873 – 895. (2007)
6. Paganin, S., Herring, A. H, Olshan, A. F., Dunson, D. B. and The National Birth Defect Study: Centered Partition Process: Informative Priors for Clustering. *arXiv preprint* arXiv:1901.10225. (2018)
7. Pitman J.: Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102**, 145–158. (1995)
8. Stanley, R. P.: Enumerative Combinatorics. Vol. 1. Cambridge University Press. (1997)