Department of Mathematics and Statistics

Bayesian cluster analysis with applications to pathogen population genomics

Alberto Pessia

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in Auditorium XII, Main building, on Friday, October 27, 2017 at 12:00 p.m.

University of Helsinki
Finland

Supervisor

Professor Jukka Corander, University of Helsinki, Helsinki, Finland

Pre-examiners

Professor Lasse Holmström, University of Oulu, Oulu, Finland
Docent Jing Tang, FIMM, University of Helsinki, Helsinki, Finland

Opponent

Professor Guido Consonni, Università Cattolica del Sacro Cuore, Milan, Italy

Custos

Professor Jukka Corander, University of Helsinki, Helsinki, Finland

Contact information

Department of Mathematics and Statistics
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: mathstat-info@helsinki.fi
URL: http://math.helsinki.fi/
Telephone: +358 2 941 51502, +358 2 941 51506

ISBN 978-951-51-3675-6 (paperback)
ISBN 978-951-51-3676-3 (PDF)
Helsinki 2017
Unigrafia

To the memory of my father Elio

Acknowledgments

This work would not exist without the help I received from a large number of people. It is extremely valuable to have people to rely on in difficult times, and I am most fortunate to have had plenty of them. The list is long and unfortunately I cannot thank each and every one that would deserve it.

A big thank you goes obviously to my supervisor Jukka Corander, who believed in me already when I was an exchange student and afterwards throughout my journey. He always supported me in everything I did with a positive and optimistic attitude. I owe him a lot.

I am profoundly grateful to my love Elli for all the support she gives me. She helped me finalizing this dissertation during the most difficult and toughest period of my life and I will always be deeply indebted to her. I would also like to thank my family for always being present and supportive whenever I needed. I feel extremely lucky to have such a loving mother, two wonderful sisters, and brilliant nephews. Thank you also Seija, Veikko, Eija, and Hannu for treating me with love and kindness.

My sincere gratitude goes to all the friends that accompanied me during this period of my life. I am particularly grateful to my dearest friends Andrea, Ilenia, Ilaria, Sabrina, Daniela, Maria Alessandra, Leo, Linda, Giuseppe, Roberto, and Monica. My warmest thanks go to Janica, Alex, Eikku, Lena, Kata, Antti, and Kanerva. To my mökki friends Maija, Miika, Ville, Fanni, Pekkis, Laura, Mira, Jusu, Laura, Lauri, Anssi, Markus, Tuomo, and Veronika. To my Kempo friends Kaisa, Tommi, Sami, Pauliina, Jiri, Matias, Tuulia, Jaakko, Joonas, Kirsi, Timo, and Tero. To my salsa friends Laura, Alexandra, Max, Solomon, Kati, Sini, Valtteri, Nuria, Jukkis, Bettina, Cristina, Hanna, Mikko, and Liz. To my ex-colleagues Jie, Lu, Elina, Minna, Paul, Jukka K., Jukka S., Mikhail, Michael, and Väinö.

Finally, I would like to express my immense gratitude to my former professors Enzo, who made me truly fall in love with Statistics, and Luca, who introduced me to computational statistics and inspired me to continue my studies.

Helsinki, October 2017
Alberto Pessia

Abstract

Identifying similarity patterns in heterogeneous observations is a very common problem in many branches of science. When the similarities and dissimilarities are encoded by a group structure, the task of dividing the observed sample into an unknown number of homogeneous groups is known as cluster analysis. Among the many types of statistical data analyses, it is one of the most widely applied.

In evolutionary biology, for example, the population structure plays an important role. Groups naturally arise as the result of evolutionary processes and depending on the resolution of the study, clusters might represent similar molecules, organisms, or even species. With the huge amount of genetic data now freely available in on-line databases, cluster analysis is a valuable technique to better understand the evolution of organisms.

In this dissertation we focus our attention on Bayesian approaches to model-based clustering. We review the mathematical formalization of the two most common methods, finite mixture models and product partition models, together with algorithms needed to draw inferences. We then introduce a novel Bayesian model which has been specifically designed to partition categorical data matrices. Finally, we show how cluster analysis is a very effective method for understanding the evolution of pathogens, and how this information is relevant to public health.

List of articles

This doctoral dissertation consists of four original articles and a summarising part. The articles, which in the text are referred to with their corresponding Roman numeral, are:

I. Chewapreecha, C., Harris, S. R., Croucher, N. J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D. M., Mather, A. E., Page, A. J., Salter, S. J., Harris, D., Nosten, F., Goldblatt, D., Corander, J., Parkhill, J., Turner, P., and Bentley, S. D. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics* 46(3): 305–309.

II. Pessia, A., Grad, Y., Cobey, S., Puranen, J. S., and Corander, J. (2015). K-Pax2: Bayesian identification of cluster-defining amino acid positions in large sequence datasets. *Microbial Genomics*, 1(1).

III. Pessia, A., and Corander, J. (2017). Kpax3: Bayesian bi-clustering of large sequence datasets. *Submitted manuscript.*

IV. Méric, G., McNally, A., Pessia, A., Mageiros, L., Mourkas, E., Vehkala, M., Corander, J., and Sheppard, S. K. (2017). Convergent amino acid signatures in paraphyletic *Campylobacter jejuni* sub-populations suggest human niche tropism. *Submitted manuscript.*

Author contribution

In article I and IV AP had the main responsibility in performing the statistical analyses and interpreting the results, jointly with JC.

In article II and III AP had the main responsibility in designing the statistical model, implementing the method, performing the analysis, and writing the manuscript.

Contents

Chapter 1

Introduction

The task of allocating statistical units into a discrete number of homogeneous groups, or clusters, is a common problem in statistics. The group structure of the population is often interesting *per se*, such as in biological taxonomy, and cluster analysis techniques (Rokach, 2010) are routinely applied in many different branches of science. Even when clusters are not the primary target of the study, their existence has an indirect effect on the observed values. This fact should obviously be taken into account if accurate estimates of parameters, for a model thought to approximate the data generating process, are to be sought.

It is not surprising that applications following this approach date back as far as the end of the 19th century, with the seminal work of Karl Pearson (1894). In his study, Pearson employed a mixture of two Normal distributions for the estimation of biological parameters of crabs from the Bay of Naples (Figure 1.1). Despite the fact that clustering was not the aim of the statistical analysis, it was nevertheless implicitly modelled by assuming the presence of two sub-populations having different size.

This classic example highlights very clearly a key concept on which this work is based, that is of filtering out the effect of the group structure in order to get a better view of the underlying stochastic process. Taken to the extreme, this is the same reason why explanatory variables are included in a statistical model for a clinical trial. By removing as much variability as possible not imputable to the treatment (noise), we expect to recover its true effect (signal). Indeed, we might even think of the discrete clinical parameters as defining a group structure of the sample according to combinations of, for example, gender, treatment, age, etc. (Figure 1.2). The main concept of this example is that the information we possess about the clustering might vary, from complete control to none.

When groups are defined beforehand and the task is to classify new observa-

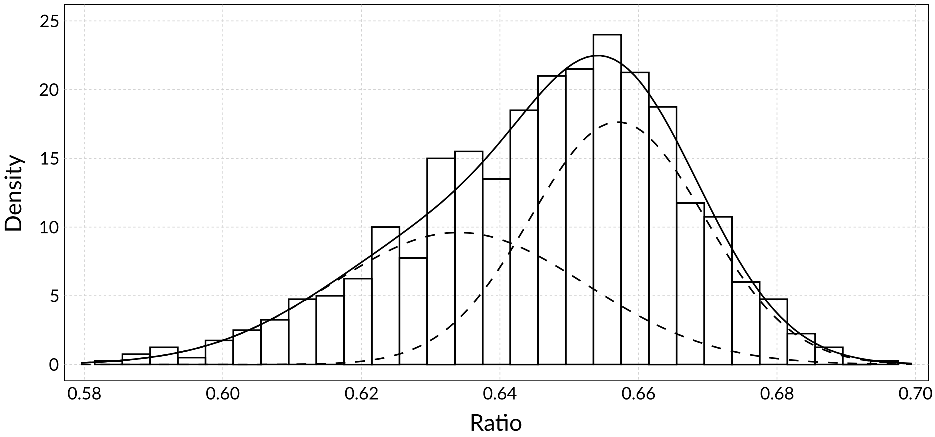

Figure 1.1: Histogram of forehead to body length ratio for 1000 crabs sampled from the Bay of Naples. The solid line represents the maximum likelihood density of a mixture of two normal distributions (dashed lines).

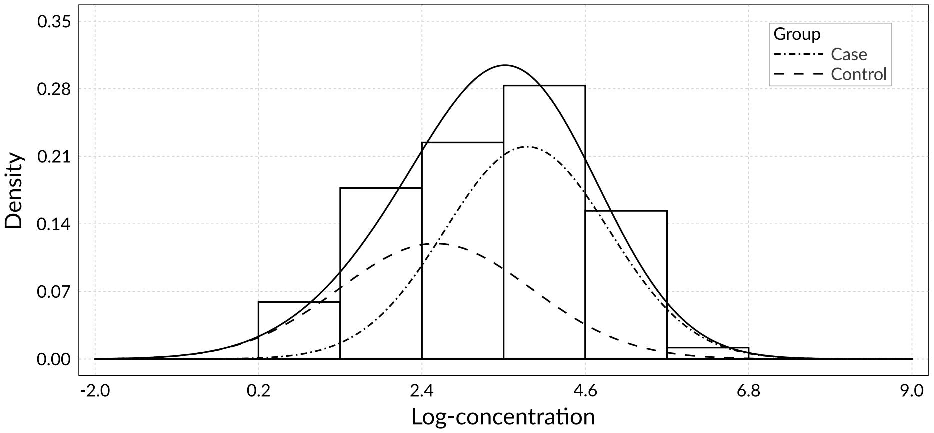

Figure 1.2: Histogram of Succinate concentration levels (micromoles per litre on the log scale) from healthy and cancer patients. Data from Eisner et al. (2011). Solid line represents the maximum likelihood density of a mixture of two normal distributions (dashed lines) when group association is known.

tions into such clusters on the basis of a training set, we refer to this approach as *supervised learning* (Bishop, 2006). It is the opposite situation of *unsupervised learning* (Hastie et al., 2009) which is considered and described in this thesis. In a setting of (partial) ignorance about the population structure, our aim will be that of identifying interesting patterns within and between the recovered groups.

1.1 Notation

Observations produced by a statistical experiment are denoted by n multivariate random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of dimension m, collected into a n-by-m data matrix $\mathbf{X}$. The observed sample $\mathbf{X}$ is just a single point in $\mathcal{X}$, the set of all possible results that the experiment could have produced. If $\boldsymbol{\psi}$ represents all the unknown characteristics of the phenomenon under study, then Ψ denotes the set of all the possible *hypotheses* regarding the phenomenon.

A probabilistic model creates a link between the hypothesis $\boldsymbol{\psi}$ and the observed sample $\mathbf{X}$ through a mathematical formula: a probability measure. Let $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, P_{\boldsymbol{\psi}})$ be a probability space, where $\mathcal{F}_{\mathcal{X}}$ is a σ-algebra of subsets of $\mathcal{X}$ and $P_{\boldsymbol{\psi}}$ is a probability measure indexed by $\boldsymbol{\psi}$. For each hypothesis $\boldsymbol{\psi} \in \Psi$, a different probabilistic explanation is given to the observed sample $\mathbf{X}$. The basic assumption of a probabilistic model is that only one of the many hypotheses $\boldsymbol{\psi} \in \Psi$ is the "truth", and that some information about it is contained in $\mathbf{X}$.

In a basic clustering problem the hypothesis $\boldsymbol{\psi} \in \Psi$ is decomposed into two main *parameters*: the partition $\mathbf{R}$ of the rows of $\mathbf{X}$ and the parameter $\boldsymbol{\theta}$ associated with the probability distributions that generated the data.

Let $\mathcal{N} = \{1, \ldots, n\}$ be the set of indices of the statistical units. A partition $\mathbf{R}$ of $\mathcal{N}$ is defined as a collection of non-empty pairwise disjoint subsets of $\mathcal{N}$, such that their union is $\mathcal{N}$. Formally, $\mathbf{R}$ is a partition of $\mathcal{N}$ if

- $\emptyset \notin \mathbf{R}$
- $\bigcup_{\mathbf{R}_g \in \mathbf{R}} \mathbf{R}_g = \mathcal{N}$
- $\mathbf{R}_g \cap \mathbf{R}_h = \emptyset, \forall\, \mathbf{R}_g, \mathbf{R}_h \in \mathbf{R}, g \neq h$

Refer to Table 1.1 for an example.

The total number of blocks k in the partition $\mathbf{R}$ has an important role in the inferential problem: it defines the dimensions of the parameter space. For this reason, we will explicitly denote with $\mathbf{R}_k$ a partition of $\mathcal{N}$ into k parts and with $\boldsymbol{\theta}_k$ the corresponding distribution parameters. The possible values for the number of blocks k is the set of integers $\{1, \ldots, n\}$, meaning that k cannot be greater than the sample size. This dependency also implies that the cardinality of the parameter space is a direct function of n. Another problem arising in the

Table 1.1: Partitions of $n = 4$.

k	Configuration class	Partition
1	$\star\star\star\star$	$(1,2,3,4)$
2	$\star\mid\star\star\star$	$(1)(2,3,4)$
		$(2)(1,3,4)$
		$(3)(1,2,4)$
		$(4)(1,2,3)$
	$\star\star\mid\star\star$	$(1,2)(3,4)$
		$(1,3)(2,4)$
		$(1,4)(2,3)$
3	$\star\mid\star\mid\star\star$	$(1)(2)(3,4)$
		$(1)(3)(2,4)$
		$(1)(4)(2,3)$
		$(2)(3)(1,4)$
		$(2)(4)(1,3)$
		$(3)(4)(1,2)$
4	$\star\mid\star\mid\star\mid\star$	$(1)(2)(3)(4)$

estimation of $\mathbf{R}$ is that its single realization is never directly observed and it should be treated as a *latent* variable: its effect is only seen through the values of sample $\mathbf{X}$.

Example 1.1. Suppose we want to infer the evolutionary process of a virus based on the four observed viral strains

$$\mathbf{X} = \begin{pmatrix} \mathrm{T} & \mathrm{T} & \mathrm{T} \\ \mathrm{T} & \mathrm{T} & \mathrm{C} \\ \mathrm{A} & \mathrm{T} & \mathrm{T} \\ \mathrm{A} & \mathrm{T} & \mathrm{C} \end{pmatrix}$$

A simple evolutionary model would consider the four strains originating from the same lineage, defining a single vector $\boldsymbol{\theta} = (\theta_A, \theta_C, \theta_G, \theta_T)'$ for the nucleotide probabilities. The observation that the first two strains encode the amino acid Phenylalanine while the last two translate into Isoleucine suggests the existence of two different sub-populations. Based on this new information it might be worth defining $k = 2$ groups and associate a vector $\boldsymbol{\theta}_1$ to $(\mathbf{x}_1, \mathbf{x}_2)$ and a different vector $\boldsymbol{\theta}_2$ to $(\mathbf{x}_3, \mathbf{x}_4)$. Note that the total number of probabilities to estimate is now doubled.

There are many different ways to represent the partition $\mathbf{R}$ in a mathematical form. Here, two approaches that are popular in the statistical clustering literature, and that will be used in later chapters, are illustrated. The first straightforward approach is to represent the partition as a matrix of indicator variables. If $\mathbf{Z}_k$ is a n-by-k binary matrix, define $z_{ig} = 1$ if and only if unit i ($i = 1, \dots, n$) belongs to cluster g ($g = 1, \dots, k$). By definition, $\sum_{g=1}^{k} z_{ig} = 1$ for all i. Block size is represented by n_g and easily computed as $n_g = \sum_{i=1}^{n} z_{ig}$. The main problem with this representation, which will become evident in the context of mixture models, is that $\mathbf{R}_k$ is invariant to permutations of the columns of $\mathbf{Z}_k$. Expressed in a different way, cluster labels are arbitrary in the sense that rearranging the columns of $\mathbf{Z}_k$ in any discretionary order will result in the same partition $\mathbf{R}_k$.

The second approach is encoding the partition as an adjacency matrix $\mathbf{A}_k$, whose generic element a_{it} is equal to 1 if units i and t belong to the same group, or 0 if they do not. It is easy to show that $\mathbf{A}_k = \mathbf{Z}_k \mathbf{Z}_k'$ and that its values do not depend on the cluster labels.

1.2 Bayesian inference

Following a subjective view of probability, every uncertain event can be associated with a probability that quantifies its uncertainty. According to this approach, since the "true" hypothesis $\boldsymbol{\psi} \in \Psi$ is unknown, it is always possible to elicit a probability measure $P_{\boldsymbol{\phi}}$ on the joint space $\mathcal{X} \times \Psi$ of observations and hypotheses. If $p(\cdot|\boldsymbol{\phi})$ is the density function (or mass function in case of discrete random variables) associated with $P_{\boldsymbol{\phi}}$, then

$$p(\boldsymbol{\psi}, \mathbf{X}|\boldsymbol{\phi}) = p(\boldsymbol{\psi}|\boldsymbol{\phi})p(\mathbf{X}|\boldsymbol{\psi}, \boldsymbol{\phi})$$

where $p(\boldsymbol{\psi}|\boldsymbol{\phi})$ is the *prior* distribution and $p(\mathbf{X}|\boldsymbol{\psi}, \boldsymbol{\phi})$ is the *sampling* distribution (or *likelihood* function if seen as a function of $\boldsymbol{\psi}$). Hyperparameter $\boldsymbol{\phi}$ encodes all the prior information we possess about the phenomenon and is often assumed to be a known constant. If there is uncertainty also in the hyperparameter $\boldsymbol{\phi}$, such as in multilevel/hierarchical models, it is usually straightforward to expand the model and accommodate another prior probability distribution on $\boldsymbol{\phi}$.

Conditional on the observed sample $\mathbf{X}$, every conclusion about $\boldsymbol{\psi}$ is given in terms of probabilities. Through the Bayes' theorem, from which the approach got its name, prior information about $\boldsymbol{\psi}$ is updated according to the new evidence contained in $\mathbf{X}$:

$$p(\boldsymbol{\psi}|\mathbf{X}, \boldsymbol{\phi}) = \frac{p(\boldsymbol{\psi}, \mathbf{X}|\boldsymbol{\phi})}{p(\mathbf{X}|\boldsymbol{\phi})} = \frac{p(\boldsymbol{\psi}|\boldsymbol{\phi})p(\mathbf{X}|\boldsymbol{\psi}, \boldsymbol{\phi})}{\int_{\Psi} p(\boldsymbol{\psi}|\boldsymbol{\phi})p(\mathbf{X}|\boldsymbol{\psi}, \boldsymbol{\phi})\mathrm{d}\boldsymbol{\psi}} \propto p(\boldsymbol{\psi}|\boldsymbol{\phi})p(\mathbf{X}|\boldsymbol{\psi}, \boldsymbol{\phi})$$

Answers to inferential problems are usually defined as functions of the *posterior* distribution. For example, it is common to look for *credibility sets* defined as any set S that satisfies

$$\int_S p(\boldsymbol{\psi}|\mathbf{X}, \boldsymbol{\phi})\mathrm{d}\boldsymbol{\psi} = \alpha$$

for some fixed probability α. Point estimates are usually defined as measures of central tendency, such as the posterior mean, median, or mode. In a statistical decision theory framework, they are defined as the values minimizing the posterior expected loss

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\xi}\in\Psi} \mathbb{E}_{\boldsymbol{\psi}}[L(\boldsymbol{\psi}, \boldsymbol{\xi})|\mathbf{X}] = \arg\min_{\boldsymbol{\xi}\in\Psi} \int_\Psi L(\boldsymbol{\psi}, \boldsymbol{\xi})p(\boldsymbol{\psi}|\mathbf{X}, \boldsymbol{\phi})\mathrm{d}\boldsymbol{\psi}$$

for some loss function $L(\boldsymbol{\psi}, \boldsymbol{\xi})$.

When expected values with respect to the posterior distribution are not available in closed form, approximate solutions to quantities of interest can be obtained through simulation techniques (Robert and Casella, 2004).

Example 1.2. Let $\mathbf{t} = (t_1, \ldots, t_n)'$ be a collection of i.i.d. random variables representing the time to first occurrence of some event of interest. Each t_i is here modelled with a Gamma distribution with known shape parameter α but unknown rate parameter β, meaning that

$$p(\mathbf{t}|\alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} t_i^{\alpha-1} e^{-\beta t_i} = \left(\frac{\tilde{t}^{\alpha-1}}{\Gamma(\alpha)}\right)^n \beta^{n\alpha} e^{-\beta n \bar{t}}$$

where $\bar{t}$ is the arithmetic mean and $\tilde{t}$ is the geometric mean. A maximum likelihood approach would estimate β by maximizing the previous expression, leading to $\hat{\beta} = \alpha/\bar{t}$. From a Bayesian point of view, instead, our uncertainty about β is encoded by a probability distribution. If the unknown parameter is given a prior Gamma distribution with parameters ω and ρ, then

$$p(\beta|\alpha, \omega, \rho, \mathbf{t}) \propto \beta^{\omega+n\alpha-1} e^{-(\rho+n\bar{t})\beta}$$

and its posterior distribution is again a Gamma distribution with shape parameter $\omega' = \omega + n\alpha$ and rate parameter $\rho' = \rho + n\bar{t}$. From the properties of the Gamma distribution, its posterior expected value is simply ω'/ρ' while the posterior variance is ω'/ρ'^2. Credible intervals for β might be obtained from the quantile function of the Gamma distribution. Finally, note that the maximum likelihood estimate is recovered when both ω and ρ approach zero on the limit.

1.3 Probability distributions

Important probability distributions, used extensively in later chapters, are now presented.

1.3.1 Exponential family

Let $(\mathcal{X}, \mathcal{Z}_{\mathcal{X}}, \mu)$ be a measure space, where $\mathcal{Z}_{\mathcal{X}}$ is a σ-algebra of subsets of $\mathcal{X}$ and μ is a σ-finite measure on $\mathcal{X}$. Let S be a subset of $\mathcal{X}$, representing the support of a random variable $\mathbf{X}$. A family $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega\}$ of probability measures on $\mathcal{X}$ is called *exponential* if the probability of any measurable set $E \subseteq \mathcal{X}$ is equal to the Lebesgue integral $\int_E p(\mathbf{x}|\boldsymbol{\theta})\mu(\mathrm{d}\mathbf{x})$, where

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta}) \exp\{\boldsymbol{\lambda}(\boldsymbol{\theta})'\mathbf{u}(\mathbf{x})\} = h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left\{\sum_{l=1}^{d} \lambda_l(\boldsymbol{\theta})u_l(\mathbf{x})\right\} \quad (\mathbf{x} \in S)$$

for an arbitrary choice of functions $h, g, \lambda_1, \dots, \lambda_d, u_1, \dots, u_d$. $F_{\boldsymbol{\theta}}$ is an absolutely continuous probability distribution if μ is the standard Lebesgue measure (length, area, etc.) while it is a discrete probability distribution if μ is a counting measure.

The importance of the exponential family in statistics is connected to the theory of sufficiency (see, for example, Andersen (1970)). From a Bayesian point of view, its conjugacy property (Bernardo and Smith, 1994) is also of particular interest.

The likelihood for a sample of (conditionally) independent and identically distributed random variables is

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n|\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})^n \exp\left\{\boldsymbol{\lambda}(\boldsymbol{\theta})' \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right\}$$

where $\sum_i \mathbf{u}(\mathbf{x}_i)$ is the vector of sufficient statistics for $\boldsymbol{\theta}$. If the prior distribution has the exponential family form

$$p(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})^\beta \exp\{\boldsymbol{\lambda}(\boldsymbol{\theta})'\boldsymbol{\alpha}\}$$

then it is easy to show that the posterior distribution is again in the exponential family:

$$p(\boldsymbol{\theta}|\mathbf{x}_1, \dots, \mathbf{x}_n) \propto g(\boldsymbol{\theta})^{\beta+n} \exp\left\{\boldsymbol{\lambda}(\boldsymbol{\theta})' \left(\boldsymbol{\alpha} + \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right)\right\}$$

Having a tractable posterior distribution is especially convenient when the parameter $\boldsymbol{\theta}$ is not the primary target of the inferential process, meaning that it can be integrated out to obtain the predictive distribution

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n) = \int_\Theta p(\boldsymbol{\theta})p(\mathbf{x}_1, \dots, \mathbf{x}_n|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

Dirichlet-Categorical distribution

Categorical random variables are commonly encountered in statistical modelling of genetic data. Let $\mathcal{A} = \{a_1, \dots, a_d\}$ be the support of independent and identically distributed discrete random variables Y_i $(i = 1, \dots, n)$. In case of DNA data, we might define $\mathcal{A} = \{A, C, G, T\}$. If $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$ is the vector of corresponding probabilities, then $\Pr\{Y_i = a_l | \boldsymbol{\theta}\} = \theta_l$.

An alternative formalization is obtained by introducing an indicator vector $\mathbf{X}_i = (X_{i1}, \dots, X_{id})'$, where $X_{il} = 1$ if and only if Y_i is equal to a_l. Obviously, $\sum_l X_{il} = 1$ for all $i = 1, \dots, n$. The likelihood function is simply

$$\prod_{i=1}^{n} \Pr\{\mathbf{X}_i = \mathbf{x}_i | \boldsymbol{\theta}\} = \prod_{i=1}^{n} \prod_{l=1}^{d} \theta_l^{x_{il}} = \prod_{l=1}^{d} \theta_l^{t_l}$$

where $t_l = \sum_i x_{il}$ $(l = 1, \dots, d)$ are the sufficient statistics, i.e. the total number of observations equal to l.

The exponential family form is recovered by choosing $h(\mathbf{x}) = 1$ for all $\mathbf{x}$, $g(\boldsymbol{\theta}) = 1$ for all $\boldsymbol{\theta} \in \Omega$, $\lambda_l(\boldsymbol{\theta}) = \log \theta_l$, and $u_l(\mathbf{x}_i) = x_{il}$:

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n | \boldsymbol{\theta}) = \exp\left\{ \sum_{l=1}^{d} t_l \log \theta_l \right\}$$

The conjugate Dirichlet distribution

$$p(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{l=1}^{d} \alpha_l)}{\prod_{l=1}^{d} \Gamma(\alpha_l)} \exp\left\{ \sum_{l=1}^{d} (\alpha_l - 1) \log \theta_l \right\}$$

is also a member of the exponential family. Multiplying the previous two equations and integrating out $\boldsymbol{\theta}$ leads to the predictive distribution

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n | \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{l=1}^{d} \alpha_l)}{\prod_{l=1}^{d} \Gamma(\alpha_l)} \frac{\prod_{l=1}^{d} \Gamma(\alpha_l + t_l)}{\Gamma(n + \sum_{l=1}^{d} \alpha_l)} = \frac{B(\boldsymbol{\alpha} + \mathbf{t})}{B(\boldsymbol{\alpha})} \tag{1.1}$$

where $B(\cdot)$ is the multivariate Beta function. From (1.1), which is known as the Dirichlet-Categorical distribution, we can easily obtain conditional predictive distributions for units yet to be observed:

$$p(\mathbf{x}_{n+1} | \mathbf{x}_n, \dots, \mathbf{x}_1, \boldsymbol{\alpha}) = \frac{B(\boldsymbol{\alpha} + \mathbf{t} + \mathbf{x}_{n+1})}{B(\boldsymbol{\alpha} + \mathbf{t})} = \frac{\prod_{l=1}^{d} (\alpha_l + t_l)^{x_{(n+1)l}}}{n + \sum_{l=1}^{d} \alpha_l} \tag{1.2}$$

Beta-Bernoulli distribution

When the set $\mathcal{A}$ has cardinality $d = 2$, the Dirichlet distribution reduces to the Beta distribution and the Categorical distribution reduces to the Bernoulli distribution. In this case, equations (1.1) and (1.2) become

$$p(\mathbf{x}_1,\dots,\mathbf{x}_n|\alpha_1,\alpha_2) = \frac{B(\alpha_1+t_1,\alpha_2+n-t_1)}{B(\alpha_1,\alpha_2)} \tag{1.3}$$

$$p(\mathbf{x}_{n+1}|\mathbf{x}_n,\dots,\mathbf{x}_1,\alpha_1,\alpha_2) = \frac{(\alpha_1+t_1)^{x_{(n+1)1}}(\alpha_2+n-t_1)^{1-x_{(n+1)1}}}{n+\alpha_1+\alpha_2} \tag{1.4}$$

where $B(\cdot,\cdot)$ is the standard Beta function.

1.3.2 Uniform distributions on partitions of n

The most straightforward (prior) probability distribution over the set of partitions of $\mathcal{N}=\{1,\dots,n\}$ is to assume that each one of them has the same probability of being selected. Considering that the total number of partitions of n is equal to the Bell number

$$B(n) = \frac{1}{e}\sum_{h=0}^{\infty}\frac{h^n}{h!}$$

then the uniform distribution is simply equal to

$$p(\mathbf{R}) = \frac{1}{B(n)} \propto 1 \tag{1.5}$$

Despite its uniform nature, distribution (1.5) induces very different probability masses on the number of blocks k, favouring partitions with k around the value of $n/\log(n)$. This can be explained by observing that the total number of partitions of n into k parts is equal to the Stirling number of the second kind

$$S(n,k) = \frac{1}{k!}\sum_{h=0}^{k}(-1)^{k-h}\binom{k}{h}h^n$$

and that the Bell number can be rewritten as the sum

$$B(n) = \sum_{g=1}^{n} S(n,g)$$

The probability of observing a partition with k groups is then equal to

$$p(k) = \frac{S(n,k)}{\sum_{g=1}^{n} S(n,g)}$$

attaining its maximum value at $k \approx n/\log(n)$, that is the maximum of $S(n,\cdot)$.

As seen in Section 1.1, a clustering of the set $\mathcal{N}=\{1,\dots,n\}$ can be represented by the total number of clusters k and the actual partition of n units into k

parts. Therefore, it is natural to model the probability distribution on $\mathbf{R}$ with a hierarchical approach:

$$p(\mathbf{R}|a,b) = p(k|a,b)p(\mathbf{R}_k|k) = \frac{\mathbb{I}(a \leq k \leq b)}{b-a+1}\frac{1}{S(n,k)} \tag{1.6}$$

where $\mathbb{I}(A)$ is the indicator function, equal to 1 if A is true and 0 otherwise, and $1 \leq a \leq b \leq n$ are, respectively, the minimum and maximum number of groups allowed. When $a = 1$ and $b = n$ the total number of clusters k is uniformly distributed on the set $\{1,\dots,n\}$. Conditional on k, $\mathbf{R}_k$ is also uniformly distributed on the set of partitions of exactly k groups.

Kohonen and Corander (2016) illustrated the stable behaviour of this kind of prior distribution and compared it with other kind of distributions on the partitions of n. Cheng et al. (2013) employed this approach with a population genomics model, as it offers both an easy implementation and a reasonable level of penalty for an increase in the number of clusters.

Another interesting hierarchical approach has been recently proposed by Casella et al. (2014) as a new option for an objective prior. Their Hierarchical Uniform Prior (HUP) is defined as

$$p(\mathbf{R}) = p(k)\binom{n}{n_1 \cdots n_k}^{-1}\frac{R(n_1,\dots,n_k)}{b(n,k)} \tag{1.7}$$

where $R(n_1,\dots,n_k) = \prod_{h=1}^{n}[\sum_{g=1}^{k}\mathbb{I}(n_g = h)]!$ and $b(n,k)$ is the number of configuration classes for partitions of n into k parts. The closed form of $b(n,k)$ is not yet known, and the high computational cost for its evaluation decreases HUP applicability to real situations with big n.

1.3.3 Ewens-Pitman distribution

The Ewens-Pitman distribution is an important family of probability distributions indexed by two real parameters (α,η) (Crane, 2016) and its origin lies in evolutionary molecular genetics with the seminal paper of Ewens (1972). It can be found in many fields of statistics and mathematics, including Bayesian nonparametric models (Antoniak, 1974) and combinatorial stochastic processes (Pitman, 2006). Its close relationship to coalescent theory (Kingman, 1978; Hoppe, 1987) and its many desirable mathematical properties make it a good candidate as a prior distribution on partitions of biological samples.

Its general form is

$$p(\mathbf{R}) = \frac{(\eta+\alpha)_{k-1,\alpha}}{(\eta+1)_{n-1,1}}\prod_{g=1}^{k}(1-\alpha)_{n_g-1,1} \tag{1.8}$$

Table 1.2: Probability distributions on partitions of $n = 4$. "U" refers to equation (1.5), "HU" to equation (1.6) with $a = 1$ and $b = n = 4$, "HUP" to equation (1.7) with a Uniform distribution on k, "EP" to equation (1.8) with $\alpha = 0.5$ and $\eta = -0.25$.

k	Configuration class	Partition	U	HU	HUP	EP
1	$\star\star\star\star$	$(1,2,3,4)$	1/15	1/4	1/4	0.5195
2	$\star\mid\star\star\star$	$(1)(2,3,4)$	1/15	1/28	1/32	0.0519
		$(2)(1,3,4)$	$\vert$	$\vert$	$\vert$	$\vert$
		$(3)(1,2,4)$	$\vert$	$\vert$	$\vert$	$\vert$
		$(4)(1,2,3)$	$\vert$	$\vert$	$\vert$	$\vert$
	$\star\star\mid\star\star$	$(1,2)(3,4)$	1/15	1/28	1/24	0.0173
		$(1,3)(2,4)$	$\vert$	$\vert$	$\vert$	$\vert$
		$(1,4)(2,3)$	$\vert$	$\vert$	$\vert$	$\vert$
3	$\star\mid\star\mid\star\star$	$(1)(2)(3,4)$	1/15	1/24	1/24	0.0260
		$(1)(3)(2,4)$	$\vert$	$\vert$	$\vert$	$\vert$
		$(1)(4)(2,3)$	$\vert$	$\vert$	$\vert$	$\vert$
		$(2)(3)(1,4)$	$\vert$	$\vert$	$\vert$	$\vert$
		$(2)(4)(1,3)$	$\vert$	$\vert$	$\vert$	$\vert$
		$(3)(4)(1,2)$	$\vert$	$\vert$	$\vert$	$\vert$
4	$\star\mid\star\mid\star\mid\star$	$(1)(2)(3)(4)$	1/15	1/4	1/4	0.0649

where

$$(x)_{y,z} = \prod_{t=1}^{y} (x + (t-1)z)$$

is the Pochhammer k-symbol. In order to result in a proper probability mass function, the hyperparameters must satisfy either $0 \leq \alpha < 1$ and $\eta > -\alpha$, or $\alpha < 0$ and $\eta = -L\alpha$ for some $L \in \{1, 2, \ldots\}$.

The expected number of groups, when $\alpha \geq 0$, equals

$$\mathbb{E}[k] = \begin{cases} \dfrac{\Gamma(\eta+n+\alpha)\Gamma(\eta+1)}{\alpha\Gamma(\eta+n)\Gamma(\eta+\alpha)} - \dfrac{\eta}{\alpha} & \text{if } \alpha > 0 \text{ and } \eta > -\alpha \\ \displaystyle\sum_{g=1}^{n} \frac{\eta}{\eta+g-1} & \text{if } \alpha = 0 \text{ and } \eta > 0 \end{cases}$$

See Table 1.2 and Table 1.3 for an example comparison between distributions on partitions of $n = 4$.

Table 1.3: Probability on the number of groups k of $n = 4$. “U” refers to equation (1.5), “HU” to equation (1.6) with $a = 1$ and $b = n = 4$, “HUP” to equation (1.7) with a Uniform distribution on k, “EP” to equation (1.8) with $\alpha = 0.5$ and $\eta = -0.25$.

k	U	HU	HUP	EP
1	1/15	1/4	1/4	0.5195
2	7/15	1/4	1/4	0.2597
3	6/15	1/4	1/4	0.1559
4	1/15	1/4	1/4	0.0649

1.4 Loss functions

In the point estimation of parameter $\mathbf{R}$, two loss functions are generally considered in Bayesian cluster analysis: the unit loss and the Binder loss (Binder, 1978). Both approaches have their strengths and weaknesses as soon will be shown. The choice of which one is most suitable for the problem at hand is often simply based on how much computational power is available.

1.4.1 Unit loss

Let

$$L(\mathbf{R}, \hat{\mathbf{R}}) = \begin{cases} 1 & \text{if } \hat{\mathbf{R}} \neq \mathbf{R} \\ 0 & \text{otherwise} \end{cases}$$

be the unit loss function. According to this loss function, every wrong estimate of $\mathbf{R}$ is given the same loss of 1 regardless of how distant the estimation is from the true value.

The posterior expected loss is

$$\mathbb{E}[L(\mathbf{R}, \hat{\mathbf{R}})|\mathbf{X}] = \sum_{\mathbf{R}} L(\mathbf{R}, \hat{\mathbf{R}}) p(\mathbf{R}|\mathbf{X}) = 1 - p(\hat{\mathbf{R}}|\mathbf{X})$$

and it is obviously minimized when $\hat{\mathbf{R}}$ is equal to the mode of the posterior distribution. For this reason, this estimator is also known as the Maximum A Posteriori (MAP) estimator of $\mathbf{R}$.

The simplicity of this estimator lies on the fact that it does not require the knowledge of the whole posterior distribution in order to be determined. Instead, optimization techniques are often necessary and employed to find the solution.

1.4.2 Binder loss

Representing the partition $\mathbf{R}$ with the adjacency matrix $\mathbf{A}$, the simplest form of the Binder loss (Binder, 1978) is equal to

$$L(\mathbf{R}, \hat{\mathbf{R}}) = \begin{cases} l_1 & \text{if } a_{it} = 1 \text{ and } \hat{a}_{it} = 0 \\ l_2 & \text{if } a_{it} = 0 \text{ and } \hat{a}_{it} = 1 \\ 0 & \text{otherwise} \end{cases}$$

for arbitrary values $l_1, l_2 > 0$. This kind of loss function gives different weights to the pairwise classification errors. If two units belong to the same cluster but are wrongly put apart, a loss l_1 occurs. On the other hand, if they should be in different clusters but are instead put together, a loss l_2 occurs. The ratio l_1/l_2 is a measure of how important internal cohesion is compared to external isolation. When $l_1 = l_2$, which is generally the case, the posterior expected loss is simply equal to (Fritsch and Ickstadt, 2009)

$$\mathbb{E}[L(\mathbf{R}, \hat{\mathbf{R}})|\mathbf{X}] = \sum_{i<t} |\hat{a}_{it} - \pi_{it}|$$

where π_{it} is the posterior probability of i and t being in the same group.

Minimization of the posterior expected Binder loss is known to be a NP-complete problem (Binder, 1981). Not only samples from the posterior distribution are usually required for approximating the posterior similarity matrix, but optimization techniques are also needed in order to minimize the posterior expected loss.

Chapter 2

Bayesian cluster analysis

2.1 Finite mixture models

A common approach to model-based cluster analysis is to assume data points to be generated by a finite mixture of k probability distributions (McLachlan and Peel, 2000). To simplify the notation the index k will be omitted, but it is important to remember that since k defines the dimensions of the parameter space, each parameter depends on it.

If $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ is a vector of probabilities ($\pi_g \geq 0 \; \forall g; \sum_g \pi_g = 1$) and $\boldsymbol{\theta}_g$ is the parameter corresponding to component g, the probability of observing the sample is

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n | k, \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i | k, \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{g=1}^{k} \pi_g p_g(\mathbf{x}_i | \boldsymbol{\theta}_g) \tag{2.1}$$

leading to the posterior distribution

$$p(k, \boldsymbol{\pi}, \boldsymbol{\theta} | \mathbf{X}) \propto p(k) p(\boldsymbol{\pi}, \boldsymbol{\theta} | k) \prod_{i=1}^{n} \sum_{g=1}^{k} \pi_g p_g(\mathbf{x}_i | \boldsymbol{\theta}_g) \tag{2.2}$$

Note that the partition $\mathbf{R}$ does not appear explicitly in equation (2.2), meaning that we cannot use this particular model for evaluating its posterior distribution. Analytical formulas are greatly simplified if equation (2.1) is rewritten in terms of the latent cluster association matrix $\mathbf{Z}$ (see Section 1.1). If each unit is independently put *a priori* into its own cluster according to $\boldsymbol{\pi}$, then

$$p(\mathbf{Z} | k, \boldsymbol{\pi}) = \prod_{g=1}^{k} \prod_{i=1}^{n} \pi_g^{z_{ig}}$$

and the posterior distribution becomes

$$p(k, \boldsymbol{\pi}, \mathbf{Z}, \boldsymbol{\theta}|\mathbf{X}) \propto p(k)p(\boldsymbol{\pi}, \boldsymbol{\theta}|k) \prod_{g=1}^{k} \prod_{i=1}^{n} (\pi_g p_g(\mathbf{x}_i|\boldsymbol{\theta}_g))^{z_{ig}} \tag{2.3}$$

MAP point estimation of $(k, \boldsymbol{\pi}, \boldsymbol{\theta})$, but not $\mathbf{Z}$, can be obtained with the EM algorithm (see Section 3.1). Since the parameter of interest is actually $\mathbf{Z}$, a common approach is to set $\hat{z}_{ig} = 1$ if

$$\frac{\hat{\pi}_g p_g(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_g)}{\sum_{h=1}^{\hat{k}} \hat{\pi}_h p_h(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_h)}$$

is maximized. Note that this approach does not guarantee that the resulting $\hat{\mathbf{Z}}$ would have exactly $\hat{k}$ clusters nor that it would be the MAP estimator, making it a sub-optimal choice.

For a proper Bayesian approach to the estimation of $(k, \mathbf{Z})$, it is thus necessary to compute the normalizing constant of (2.3). Unfortunately, even in the case of a conjugate prior for $(\boldsymbol{\pi}, \boldsymbol{\theta})$, the summation over all the possible values of k and $\mathbf{Z}$ make the computation not feasible for any real dataset. To overcome this problem, a huge amount of literature has been devoted to simulations from the posterior distribution. For an extensive survey, see Frühwirth-Schnatter (2006).

The major problem with (2.3) is the invariance to relabeling of the mixture components when the parameters are considered exchangeable, producing a posterior with at least $k!$ modes (one for each possible reordering of the columns of $\mathbf{Z}$). Obviously, no MCMC algorithm will be able to properly explore each and every mode in a reasonable amount of time. A general solution is to reject the assumption that parameters are exchangeable and introduce clear constraints within the prior distribution, such as proposed by Robert and Mengersen (1999).

2.2 Product partition models

Taking into account their identifiability problems and considering that in finite mixture models the partition of the sample is not the main parameter, but rather a nuisance parameter, product partition models (Hartigan, 1990; Barry and Hartigan, 1992) are the preferred way to perform Bayesian cluster analysis in this work.

An immediate benefit obtained by following this approach is the explicit modelling of the partition $\mathbf{R}$, to which a prior distribution representing our belief is given and with which a posterior distribution for drawing inferences is directly associated. Conditional on the knowledge of how the sample is partitioned, each cluster is assumed to be stochastically independent from each other, meaning that the sampling distribution simply factorizes into a product of k components

$$p(\mathbf{X}|\mathbf{R}) = \prod_{g=1}^{k} p(\mathbf{X}_g) = \prod_{g=1}^{k} \int_{\Theta_g} p(\boldsymbol{\theta}_g) \prod_{i \in g} p(\mathbf{x}_i|\boldsymbol{\theta}_g) \mathrm{d}\boldsymbol{\theta}_g \tag{2.4}$$

where $\mathbf{X}_g$ is the data matrix associated with cluster g. Opposite to finite mixture models, parameter $\boldsymbol{\theta}$ is not of interest and is usually integrated out. The posterior distribution is then equal to

$$p(\mathbf{R}|\mathbf{X}) \propto p(\mathbf{R}) \prod_{g=1}^{k} p(\mathbf{X}_g) \tag{2.5}$$

A classic product partition model would require the definition of the prior distribution similar to

$$p(\mathbf{R}) = c_0 \prod_{g=1}^{k} f(\mathbf{R}_g) \tag{2.6}$$

where c_0 is the normalizing constant and $f(\cdot)$ is the so called *cohesion* of the cluster. By construction, the posterior distribution is again a product of k cohesion functions

$$p(\mathbf{R}|\mathbf{X}) \propto \prod_{g=1}^{k} f(\mathbf{R}_g) p(\mathbf{X}_g)$$

Although the Ewens-Pitman distribution (1.8) can be rewritten to resemble (2.6), in general the same does not apply to other kind of distributions. To extend their applicability, we will refer to product partition models as any model whose sampling distribution decomposes into a product of independent components, regardless of the shape of the prior.

2.2.1 Clustering categorical data

When data is categorical, each observation in the n-by-m data matrix $\mathbf{X}$ is represented as a sequence of characters. Let $\mathcal{A}_j = \{a_1, \dots, a_{d_j}\}$ be the set of possible letters that can be observed at column j. If θ_{gju} is the probability of observing the letter $u \in \mathcal{A}_j$ at column j in cluster g, then

$$\Pr\{X_{ij} = u|\boldsymbol{\theta}, \mathbf{R}\} = \prod_{g=1}^{k} \theta_{gju}^{\mathbb{I}(i \in g)}$$

for all $i = 1, \dots, n$. Let $\boldsymbol{\theta}_{gj} = (\theta_{gj1}, \dots, \theta_{gjd_j})'$ be the vector of probabilities associated with cluster g at column j. Assuming they are *a priori* independent, we obtain

$$p(\mathbf{R}, \boldsymbol{\theta}) = p(\mathbf{R})p(\boldsymbol{\theta}|\mathbf{R}) = p(\mathbf{R}) \prod_{g=1}^{k} \prod_{j=1}^{m} p(\boldsymbol{\theta}_{gj})$$

As shown in Section 1.3.1, if $p(\boldsymbol{\theta}_{gj})$ is the density of a Dirichlet distribution with parameter $\boldsymbol{\alpha}_j$, and the columns are stochastically independent also in the likelihood, then

$$p(\mathbf{X}|\mathbf{R}) = \prod_{g=1}^{k} p(\mathbf{X}_g) = \prod_{g=1}^{k} \prod_{j=1}^{m} \frac{B(\boldsymbol{\alpha}_j + \mathbf{t}_{gj})}{B(\boldsymbol{\alpha}_j)}$$

where $\mathbf{t}_{gj}$ is the sufficient statistic counting the occurrences of each letter in cluster g at column j. The posterior distribution is then simply

$$p(\mathbf{R}|\mathbf{X}) \propto p(\mathbf{R}) \prod_{j=1}^{m} \frac{1}{B(\boldsymbol{\alpha}_j)^k} \prod_{g=1}^{k} B(\boldsymbol{\alpha}_j + \mathbf{t}_{gj}) \tag{2.7}$$

which is easy to interpret: $p(\mathbf{R})$ controls our prior information about the number of clusters and their relative size, while the Dirichlet prior regularizes the composition of letters within each cluster. This model, with the hierarchical uniform prior (1.6), is employed in article I to analyse bacterial DNA strains (see Section 4.1).

> **Example 2.1.** Consider the binary case $\mathcal{A}_j = \{\mathrm{A}, \mathrm{G}\}$, where the Dirichlet distribution becomes a $\mathrm{Beta}(\alpha_{j1}, \alpha_{j2})$. If $\alpha_{j1} > \alpha_{j2}$ we favour configurations where each cluster has more letters A than G. If we believe that each cluster contains either one of the two letters, we would then define a symmetric distribution with modes on its extremities by choosing $\boldsymbol{\alpha}_j = (\alpha, \alpha)'$ with $\alpha < 1$. On the other hand, when $\alpha_1 = \alpha_2 = \alpha > 1$ the prior puts more weight on configurations where the letters A and G have approximately the same proportion within each cluster.

2.2.2 Bi-clustering categorical data

In the case of multivariate data, it is often necessary to identify which statistical variables are the best predictors of cluster association. Here, the task is accomplished by classifying the observed features according to their amount of discrimination power.

The simultaneous classification of both the rows and columns of a data matrix is known as bi-clustering (Mirkin, 1996), whose basic idea is to rearrange the data matrix across its two dimensions in such a way that non-random patterns present

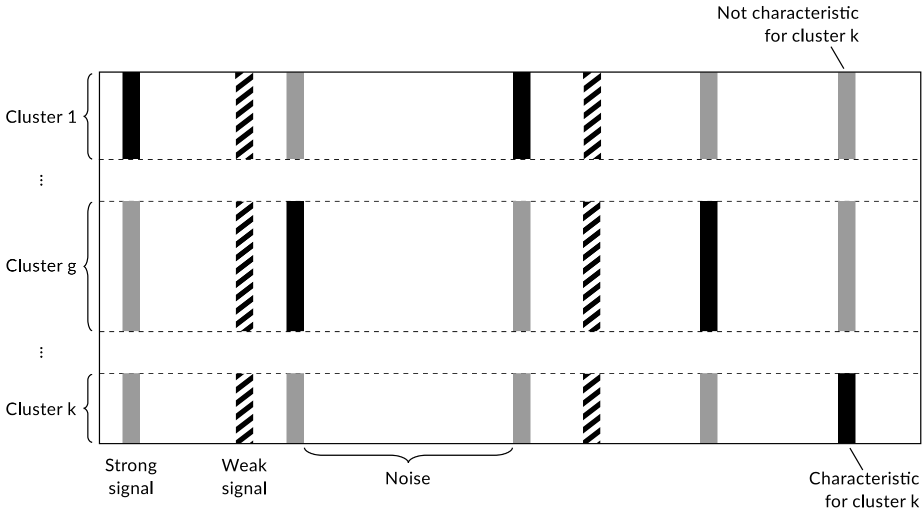

Figure 2.1: Point estimate of model parameters. The data matrix is partitioned into a grid of row and column clusters. Row blocks represent groups of homogeneous sample vectors while column blocks express similar levels of feature discrimination power.

in the data are correctly highlighted. Depending on the problem at hand, data can be restructured into different shapes (Van Mechelen et al., 2004; Madeira and Oliveira, 2004). The major contribution of this dissertation is the model initially introduced in article II and subsequently improved in article III, according to which the categorical data matrix is decomposed into a grid of non-overlapping sub-matrices (Figure 2.1).

The basic assumption of the proposed bi-clustering model is that, conditioned on the knowledge of partition $\mathbf{R}$, there is a subset of features that best discriminate the clusters. In its simplest case, the columns might be classified as either noise or signal. In the general case, we assume that each feature can be associated with only one of v possible classes or *statuses*. The classification is further refined by allowing each feature, that was given the status u $(u = 1, \ldots, v)$, to possess a particular *property* l $(l = 1, \ldots, s_u)$ within each cluster g $(g = 1, \ldots, k)$. For example, we might give a feature which was classified as signal (status) the property of either "low" or "high" probability of success in cluster g.

Formally, define the binary variable c^u_{jgl} to be equal to 1 if and only if variable j $(j = 1, \ldots, m)$ possesses status u $(u = 1, \ldots, v)$ with the property l $(l = 1, \ldots, s_u)$ in cluster g $(g = 1, \ldots, k)$. The whole collection of binary variables c^u_{jgl} can be represented by the array $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_j, \ldots, \mathbf{C}_m)$ where

$\mathbf{C}_j = (\mathbf{C}_j^1, \ldots, \mathbf{C}_j^u, \ldots, \mathbf{C}_j^v)$ and

$$\mathbf{C}_j^u = \begin{pmatrix} c_{j11}^u & \cdots & c_{j1l}^u & \cdots & c_{j1s_u}^u \\ \vdots & & \vdots & & \vdots \\ c_{jg1}^u & \cdots & c_{jgl}^u & \cdots & c_{jgs_u}^u \\ \vdots & & \vdots & & \vdots \\ c_{jk1}^u & \cdots & c_{jkl}^u & \cdots & c_{jks_u}^u \end{pmatrix}$$

The total number of properties, statuses, and their interpretation is assumed to be known and not a part of the inferential problem.

When all the parameters of the model are fixed, the samples are assumed to be independent and identically distributed

$$p(\mathbf{X}|\mathbf{R}, \mathbf{C}, \boldsymbol{\theta}) = \prod_{g=1}^{k} \prod_{i \in g} p(\mathbf{x}_i|\mathbf{R}, \mathbf{C}, \boldsymbol{\theta}_g)$$

Within this setting, $\boldsymbol{\theta}$ is considered a nuisance parameter and therefore integrated out of the model:

$$\begin{aligned} p(\mathbf{X}|\mathbf{R}, \mathbf{C}) &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}, \mathbf{X}|\mathbf{R}, \mathbf{C}) \mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathbf{R}, \mathbf{C}) p(\mathbf{X}|\mathbf{R}, \mathbf{C}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \\ &= \int_{\boldsymbol{\Theta}_1} \cdots \int_{\boldsymbol{\Theta}_k} \prod_{g=1}^{k} p(\boldsymbol{\theta}_g|\mathbf{R}, \mathbf{C}) \prod_{i \in g} p(\mathbf{x}_i|\mathbf{R}, \mathbf{C}, \boldsymbol{\theta}_g) \mathrm{d}\boldsymbol{\theta}_1 \ldots \mathrm{d}\boldsymbol{\theta}_k = \\ &= \prod_{g=1}^{k} \int_{\boldsymbol{\Theta}_g} p(\boldsymbol{\theta}_g|\mathbf{R}, \mathbf{C}) \prod_{i \in g} p(\mathbf{x}_i|\mathbf{R}, \mathbf{C}, \boldsymbol{\theta}_g) \mathrm{d}\boldsymbol{\theta}_g = \prod_{g=1}^{k} p(\mathbf{X}_g|\mathbf{R}, \mathbf{C}) \end{aligned}$$

Finally, conditional on the knowledge of which class each feature belongs to, the statistical variables are themselves stochastically independent:

$$p(\mathbf{X}|\mathbf{R}, \mathbf{C}) = \prod_{g=1}^{k} \prod_{j=1}^{m} p(\mathbf{x}_{gj}|\mathbf{R}, \mathbf{C}) \tag{2.8}$$

where $\mathbf{x}_{gj}$ is the vector observed at column j in cluster g.

We write the joint prior distribution explicitly as $p(\mathbf{R}, \mathbf{C}) = p(\mathbf{R})p(\mathbf{C}|\mathbf{R})$. The choice of $p(\mathbf{R})$ is left unspecified and free to be chosen among the many available possibilities as discussed in Section 1.3. Instead, we will focus here on $p(\mathbf{C}|\mathbf{R})$.

Elements of array $\mathbf{C}$ are considered *a priori* stochastically independent, so that the prior distribution factorizes similar to the likelihood function. From an analytical point of view, this choice allows tractable models. From a practical point of view, having m independent stochastic variables in the posterior distribution opens the way to parallelization of the computations, speeding up the algorithms employed.

As shown in Section 3.2 of article III, if each column status is independently sampled according to the probability distribution $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{ju}, \dots, \gamma_{jv})'$ and then, conditional on the realization of this status, each cluster property is independently sampled according to the (multinomial) probability distribution $\boldsymbol{\omega}^u_{jg} = \left(\omega^u_{jg1}, \dots, \omega^u_{jgl}, \dots, \omega^u_{jgs_u}\right)'$, the prior distribution becomes

$$p(\mathbf{C}|\mathbf{R}) = \prod_{j=1}^{m} \prod_{g=1}^{k} \prod_{u=1}^{v} \prod_{l=1}^{s_u} \left(\gamma_{ju}^{\frac{1}{k}} \omega^u_{jgl}\right)^{c^u_{jgl}} \tag{2.9}$$

The joint posterior distribution of $(\mathbf{R}, \mathbf{C})$, up to a normalizing constant, is obtained by multiplying the likelihood (2.8), the chosen distribution $p(\mathbf{R})$ and the prior (2.9):

$$p(\mathbf{R}, \mathbf{C}|\mathbf{X}) \propto p(\mathbf{R}) p(\mathbf{C}, \mathbf{X}|\mathbf{R}) = p(\mathbf{R}) \prod_{j=1}^{m} \prod_{g=1}^{k} \prod_{u=1}^{v} \prod_{l=1}^{s_u} \left(\gamma_{ju}^{\frac{1}{k}} \omega^u_{jgl} p^u_{jgl}\right)^{c^u_{jgl}} \tag{2.10}$$

where $p^u_{jgl} = p(\mathbf{x}_{gj}|c^u_{jgl} = 1)$.

By integrating out $\mathbf{C}$ from $p(\mathbf{C}, \mathbf{X}|\mathbf{R})$, we obtain

$$p(\mathbf{X}|\mathbf{R}) = \prod_{j=1}^{m} \sum_{u=1}^{v} \gamma_{ju} \prod_{g=1}^{k} \sum_{l=1}^{s_u} \omega^u_{jgl} p^u_{jgl} \tag{2.11}$$

from which follows that the probability of observing the data associated with a generic cluster g is

$$p(\mathbf{X}_g|\mathbf{R}) = \prod_{j=1}^{m} \sum_{u=1}^{v} \sum_{l=1}^{s_u} \gamma_{ju} \omega^u_{jgl} p^u_{jgl} \tag{2.12}$$

and that the conditional distribution $p(\mathbf{x}_i|\mathbf{x}_{i-1}, \dots, \mathbf{x}_1, k)$ is equal to the product of weighted averages

$$\prod_{j=1}^{m} \sum_{u=1}^{v} \sum_{l=1}^{s_u} \frac{\gamma_{ju} \omega^u_{jgl} p(x_{(1:(i-1))j}|k, c^u_{jgl} = 1)}{\sum_{a=1}^{v} \sum_{t=1}^{s_a} \gamma_{ja} \omega^a_{jgt} p(x_{(1:(i-1))j}|k, c^a_{jgt} = 1)} p(x_{ij}|x_{(1:(i-1))j}, k, c^u_{jgl} = 1)$$

Combining all the previous results, it is easy to show (see supplementary material S2 of article III) that

$$p(\mathbf{C}|\mathbf{R},\mathbf{X}) = \prod_{j=1}^{m}\prod_{u=1}^{v}\left(\frac{\gamma_{ju}\prod\limits_{h=1}^{k}\sum\limits_{t=1}^{s_u}\omega^u_{jht}p^u_{jht}}{\sum\limits_{a=1}^{s_u}\gamma_{ja}\prod\limits_{h=1}^{k}\sum\limits_{t=1}^{s_u}\omega^a_{jht}p^a_{jht}}\right)^{\frac{c^u_{j..}}{k}}\prod_{g=1}^{k}\prod_{l=1}^{s_u}\left[\frac{\omega^u_{jgl}p^u_{jgl}}{\sum\limits_{t=1}^{s_u}\omega^u_{jgt}p^u_{jgt}}\right]^{c^u_{jgl}} \tag{2.13}$$

from which we can directly sample values of **C** conditioned on a particular realization of **R**.

Chapter 3

Inference algorithms

3.1 Expectation-Maximization

The Expectation-Maximization (EM) algorithm is a common procedure for finding (local) optima of probabilistic models in the presence of latent variables (Gupta and Chen, 2011). It was originally applied to maximum likelihood estimation (Dempster et al., 1977), but it can be easily extended to MAP point estimation (McLachlan and Krishnan, 2008).

Expectation-Maximization is an iterative algorithm in which two steps are alternated until convergence is reached. In the first step, known as the E-step, the expected value of the latent variable, conditioned on the current estimate of the parameter of interest, is computed. In the second step, known as the M-step, the objective function is maximized with respect to the parameter of interest.

The solution to the optimization problem is the mode $\hat{\boldsymbol{\psi}}$ of the posterior distribution $p(\boldsymbol{\psi}|\mathbf{X})$. Since the logarithm is a monotonic function, the same solution can be obtained by maximizing the logarithm of the joint distribution $p(\boldsymbol{\psi}, \mathbf{X})$. The normalization constant $1/p(\mathbf{X})$, not depending on $\boldsymbol{\psi}$, can be safely ignored.

If $\boldsymbol{\psi} \in \Psi$ is the parameter we seek to optimize and $\mathbf{Z} \in \mathcal{Z}$ is the unobservable parameter, then

$$p(\boldsymbol{\psi}, \mathbf{X}) = \int_{\mathcal{Z}} p(\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X}) \frac{p(\boldsymbol{\psi}, \mathbf{Z}, \mathbf{X})}{p(\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X})} \mathrm{d}\mathbf{Z} = \mathbb{E}_{\mathbf{Z}|\boldsymbol{\xi},\mathbf{X}} \left[\frac{p(\boldsymbol{\psi}, \mathbf{Z}, \mathbf{X})}{p(\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X})} \right]$$

for all $\boldsymbol{\xi} \in \Psi$. Define

$$f(\boldsymbol{\psi}|\boldsymbol{\xi}) = \mathbb{E}_{\mathbf{Z}|\boldsymbol{\xi},\mathbf{X}} \left[\log \frac{p(\boldsymbol{\psi}, \mathbf{Z}, \mathbf{X})}{p(\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X})} \right] \tag{3.1}$$

By Jensen's inequality

$$\log p(\boldsymbol{\psi}, \mathbf{X}) \geq f(\boldsymbol{\psi}|\boldsymbol{\xi})$$

for all $\boldsymbol{\psi}, \boldsymbol{\xi} \in \Psi$. It is clear that any increment in the *surrogate* function (3.1) will also increase, or leave unchanged, the objective function $\log p(\boldsymbol{\psi}, \mathbf{X})$. It is also easy to show that the log-joint distribution is recovered when $\boldsymbol{\xi}$ is exactly equal to $\boldsymbol{\psi}$:

$$f(\boldsymbol{\psi}|\boldsymbol{\psi}) = \int_{\mathcal{Z}} p(\mathbf{Z}|\boldsymbol{\psi}, \mathbf{X}) \log \frac{p(\boldsymbol{\psi}, \mathbf{X}) p(\mathbf{Z}|\boldsymbol{\psi}, \mathbf{X})}{p(\mathbf{Z}|\boldsymbol{\psi}, \mathbf{X})} \mathrm{d}\mathbf{Z} = \log p(\boldsymbol{\psi}, \mathbf{X})$$

The EM algorithm, as shown in A.1, is guaranteed to converge to a local optimum or a saddle point (Wu, 1983). To increase the chances of reaching a global optimum, the recommended approach is to restart the algorithm from many different starting values.

Algorithm A.1 Expectation-Maximization

```
Input: Initial estimate $\boldsymbol{\psi}^{(0)} \in \Psi$
Output: Stationary point $\hat{\boldsymbol{\psi}}$ of the posterior distribution $p(\boldsymbol{\psi}|\mathbf{X})$
Initialization:
    define $f(\boldsymbol{\psi}|\boldsymbol{\xi})$ as in equation (3.1)
    $t \leftarrow 0$

repeat
    $t \leftarrow t + 1$
    $g(\boldsymbol{\psi}) \leftarrow f(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)})$                    (E-step)
    $\boldsymbol{\psi}^{(t)} \leftarrow \arg\max_{\boldsymbol{\psi} \in \Psi} g(\boldsymbol{\psi})$                    (M-step)
until convergence
```

In the case of a finite mixture model, $\boldsymbol{\psi} = (\boldsymbol{\pi}, \boldsymbol{\theta})$ is the parameter of interest and $\mathbf{Z}$ is the latent matrix of cluster associations. For simplicity we consider here a fixed value for k. In order to find the global maximum, the algorithm must obviously be repeated for each $k = 1, \ldots, n$, weighting each solution with the corresponding probability $p(k)$.

From the joint probability distribution (2.3) we easily obtain

$$p(\mathbf{Z}|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}, \mathbf{X}) = \prod_{i=1}^{n} \prod_{g=1}^{k} \left(\frac{\pi_g^{(t-1)} p_g(\mathbf{x}_i|\boldsymbol{\theta}_g^{(t-1)})}{\sum_{h=1}^{k} \pi_h^{(t-1)} p_h(\mathbf{x}_i|\boldsymbol{\theta}_h^{(t-1)})} \right)^{z_{ig}} = \prod_{i=1}^{n} \prod_{g=1}^{k} q_{ig}^{z_{ig}}$$

For any fixed value of $\mathbf{Z}$, the logarithm of (2.3) can be written as the sum

$$\begin{aligned} \log p(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{Z}, \mathbf{X}) &= \log p(\boldsymbol{\theta}, \boldsymbol{\pi}) + \sum_{i=1}^{n} \sum_{g=1}^{k} z_{ig} \log p(\mathbf{x}_i|\boldsymbol{\theta}_g) + \sum_{i=1}^{n} \sum_{g=1}^{k} z_{ig} \log \pi_g = \\ &= \log p(\boldsymbol{\theta}, \boldsymbol{\pi}) + h(\boldsymbol{\theta}|\mathbf{Z}) + u(\boldsymbol{\pi}|\mathbf{Z}) \end{aligned}$$

and the function (3.1) computed in the E-step is

$$f(\boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{\theta}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}) = \log p(\boldsymbol{\theta}, \boldsymbol{\pi}) + \sum_{i=1}^{n} \sum_{g=1}^{k} q_{ig} \log p(\mathbf{x}_i | \boldsymbol{\theta}_g) + \\ + \sum_{i=1}^{n} \sum_{g=1}^{k} q_{ig} \log \pi_g - \sum_{i=1}^{n} \sum_{g=1}^{k} q_{ig} \log q_{ig}$$

which is then maximized, with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, in the M-step of the algorithm.

3.2 Stochastic optimization

Computation of (3.1) in the E-step of the EM algorithm is sometimes not feasible. In the context of product partition models, the expected value requires to sum over all the possible partitions of n, which we already know is equal to the Bell number of n. Stochastic approaches are valid alternatives for finding solutions to optimization problems (Spall, 2003). In their simplest form, they differ from deterministic algorithms by proposing new solutions through the use of a random number generator. In general, they do not provide any guarantee that the given solution is a stationary point of the objective function. Nevertheless, if properly designed, they often produce outputs in a neighbourhood of the global optimum. Here, we will look in detail the algorithms used in articles I and II.

Let $\mathbf{R}$ be a partition of the rows of matrix $\mathbf{X}$ and $\boldsymbol{\psi}$ another parameter of interest, such as the partition $\mathbf{C}$ of the columns of matrix $\mathbf{X}$ as defined in Section 2.2.2. We seek the MAP estimate

$$(\hat{\mathbf{R}}, \hat{\boldsymbol{\psi}}) = \underset{(\mathbf{R}, \boldsymbol{\psi})}{\arg\max}\, p(\mathbf{R}, \boldsymbol{\psi} | \mathbf{X}) \tag{3.2}$$

In a standard coordinate ascend algorithm (Wright, 2015), similarly to what has been done with the EM algorithm, we would maximize the posterior distribution by alternating maximization of one variable conditioned on the last value of the other variable. For any given partition $\mathbf{R}^{(t)}$ at iteration t, it is usually easy to find $\boldsymbol{\psi}^{(t)}$ that maximizes (3.2). Taking model (2.10) as an example, the implementation of an exhaustive search for this task is straightforward as long as the total number of combinations is not too high. The opposite problem of searching in the partition space is, instead, a difficult problem.

The greedy algorithm derived from Corander and Marttinen (2006) and employed in Cheng et al. (2013) was used in the statistical analysis of article I. The idea behind this algorithm is to explore the space of partitions by performing local moves from the current known partition. These moves can be summarized by *split*, *merge*, or *transfer* operators and their basic forms are shown in algorithm A.2. The simplest version of the greedy stochastic search is outlined in algorithm A.3. It is important to note that it employs a data-driven approach in the form of a pre-computed distance matrix, whose use in guiding proposal moves was shown to considerably boost the algorithm performance. Without changing its core structure, a more complex algorithm would modify and combine the three operators in order to perform a wide variety of different kind of explorations in the solution neighbourhood.

An alternative version of the transfer operator, for example, is implemented in article II. First, a reference cluster is chosen at random. Subsequently, in a sequential manner, units from other clusters are collected until the objective function is not increased any more. By construction, the log-posterior distribution

Algorithm A.2 Move operators for Greedy Stochastic Search

```
// Note: $l$ is the current value of the posterior distribution
function MERGE($l, \mathbf{R}^{(t-1)}$)
    $u \leftarrow l,\ \mathbf{R}^{(t)} \leftarrow \mathbf{R}^{(t-1)}$
    for all pair of clusters $(g, h)$ in $\mathbf{R}^{(t-1)}$ do
        $\mathbf{S} \leftarrow$ partition obtained by merging $(g, h)$
        $v \leftarrow p(\mathbf{S}|\boldsymbol{\psi}^{(t-1)}, \mathbf{X})$
        if $v > u$ then
            $u \leftarrow v,\ \mathbf{R}^{(t)} \leftarrow \mathbf{S}$
        end if
    end for
    return $(u, \mathbf{R}^{(t)})$
end function

// Note: $\mathbf{D}$ is a matrix of pairwise distances
function SPLIT($l, \mathbf{R}^{(t-1)}, \mathbf{D}$)
    $u \leftarrow l,\ \mathbf{R}^{(t)} \leftarrow \mathbf{R}^{(t-1)}$
    for all clusters $g$ in $\mathbf{R}^{(t-1)}$ do
        $H \leftarrow$ dendrogram of cluster $g$ built from $\mathbf{D}$
        $\mathbf{S} \leftarrow$ partition obtained by cutting $H$ at some height $h$
        $v \leftarrow p(\mathbf{S}|\boldsymbol{\psi}^{(t-1)}, \mathbf{X})$
        if $v > u$ then
            $u \leftarrow v,\ \mathbf{R}^{(t)} \leftarrow \mathbf{S}$
        end if
    end for
    return $(u, \mathbf{R}^{(t)})$
end function
```

Algorithm A.2 Move operators for Greedy Stochastic Search (continued)

```
function TRANSFER($l, \mathbf{R}^{(t-1)}$)
    $u \leftarrow l,\ \mathbf{R}^{(t)} \leftarrow \mathbf{R}^{(t-1)}$
    for all units $i$ do
        for all clusters $g$ in $\mathbf{R}^{(t-1)}$ do
            $\mathbf{S} \leftarrow$ partition obtained by moving $i$ to $g$
            $v \leftarrow p(\mathbf{S}|\boldsymbol{\psi}^{(t-1)}, \mathbf{X})$
            if $v > u$ then
                $u \leftarrow v,\ \mathbf{R}^{(t)} \leftarrow \mathbf{S}$
            end if
        end for
    end for
    return $(u, \mathbf{R}^{(t)})$
end function
```

is guaranteed to not decrease at each iteration. The advantage of this kind of approach comes from the natural use of the univariate conditional probabilities, which are updated at each step.

Algorithm A.3 Greedy Stochastic Search

```
Input: Initial estimate $\mathbf{R}^{(0)}$
Output: Approximate stationary point $\hat{\mathbf{R}}$ of $p(\mathbf{R}|\boldsymbol{\psi}^{(t-1)}, \mathbf{X})$
Initialization:
    compute matrix $\mathbf{D}$ of pairwise distances between data points
    $t \leftarrow 0$
    $l \leftarrow p(\mathbf{R}^{(0)}|\boldsymbol{\psi}^{(t-1)}, \mathbf{X})$

repeat
    $t \leftarrow t + 1$
    $w \leftarrow$ random order of operators described in A.2
    for all operators in $w$ do
        compute new values for $l$ and $\mathbf{R}^{(t)}$
        if $l$ is increased then
            break for
        end if
    end for
until all operators do not produce a better solution
```

3.3 Markov Chain Monte Carlo

When other summaries other than the MAP are desired, simulation techniques are often employed for complex models (Robert and Casella, 2004). Similar to the greedy stochastic search algorithm A.3, also in this context global and local moves are alternated. In this case the aim is to produce samples from the posterior distribution. Split and merge operators are implemented in the MCMC setting with the Metropolis-Hastings proposal introduced by Jain and Neal (2007). See, in particular, the variant proposed by Dahl (2005). Transfer moves are done by the Gibbs sampler (Neal, 2000), or by the biased random walk of Booth et al. (2008) when the former is not available. In any case, we assume that $p(\boldsymbol{\theta}|\mathbf{R},\mathbf{X})$ is available in closed form and it is easy to simulate from.

3.3.1 Split-Merge procedure

The Metropolis-Hastings acceptance probability has the general form

$$\alpha(\boldsymbol{\psi},\boldsymbol{\psi}^\star) = \min\left\{1, \frac{p(\boldsymbol{\psi}^\star)}{p(\boldsymbol{\psi})}\frac{p(\mathbf{X}|\boldsymbol{\psi}^\star)}{p(\mathbf{X}|\boldsymbol{\psi})}\frac{q(\boldsymbol{\psi}|\boldsymbol{\psi}^\star)}{q(\boldsymbol{\psi}^\star|\boldsymbol{\psi})}\right\} \tag{3.3}$$

where $\boldsymbol{\psi}$ is the current value of the parameter and $\boldsymbol{\psi}^\star$ is the new proposed value, which has been sampled from the proposal distribution $q(\boldsymbol{\psi}^\star|\boldsymbol{\psi})$.

Each step of the split-merge algorithm starts by selecting two different units, i and j, completely at random. If the chosen units belong to the same cluster, say g, a split operator is performed. On the other hand, if i and j happen to be in two different groups, say g_i and g_j, a merge operator is implemented.

In case of a merge, all the units that are currently in clusters g_i and g_j are reallocated, with probability 1, into the new cluster $\mathbf{R}^\star_g = \mathbf{R}_{g_i} \cup \mathbf{R}_{g_j}$. In case of a split, all the units $\mathcal{U} = \{u : u \in \mathbf{R}_g, u \neq i, j\}$ are sequentially reallocated at random to either g_i or g_j, the clusters *founded* by i and j. After iteration s of the reallocation process, $|\mathbf{R}^{\star(s)}_{g_i}|$ and $|\mathbf{R}^{\star(s)}_{g_j}|$ units have been associated with i and j respectively. By convention we set $\mathbf{R}^{\star(0)}_{g_i} = \{i\}$ and $\mathbf{R}^{\star(0)}_{g_j} = \{j\}$. Obviously, $|\mathbf{R}^{\star(s)}_{g_i}| + |\mathbf{R}^{\star(s)}_{g_j}| = s + 2$. The next unit is allocated to g_i with probability

$$\Pr\{(s+1) \in \mathbf{R}^{\star(s+1)}_{g_i}|\mathbf{R}^{\star(s)}_{g_i}, \mathbf{R}^{\star(s)}_{g_j}, \mathbf{X}\} \propto |\mathbf{R}^{\star(s)}_{g_i}|p(\mathbf{x}_{s+1}|\{\mathbf{x}_u : u \in \mathbf{R}^{\star(s)}_{g_i}\})$$

and to g_j with probability $1-\Pr\{(s+1) \in \mathbf{R}^{\star(s+1)}_{g_i}|\mathbf{R}^{\star(s)}_{g_i}, \mathbf{R}^{\star(s)}_{g_j}, \mathbf{X}\}$. The complete joint probability distribution can be rewritten in compact form as

$$q(\mathbf{R}^\star_{g_i}, \mathbf{R}^\star_{g_j}|\mathbf{X}) = \prod_{s=1}^{|\mathbf{R}_g|-2} \Pr\{s \in \mathbf{R}^{\star(s)}_h|\mathbf{R}^{\star(s-1)}_{g_i}, \mathbf{R}^{\star(s-1)}_{g_j}, \mathbf{X}\} \tag{3.4}$$

where h is either g_i or g_j, depending on the final allocation.

For product partition models (see Section 2.2), the Metropolis-Hastings acceptance probability becomes

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}^\star) = \min \left\{ 1, \frac{f(\mathbf{R}^\star_g)}{f(\mathbf{R}_{g_i})f(\mathbf{R}_{g_j})} \frac{p(\mathbf{X}^\star_g)}{p(\mathbf{X}_{g_i})p(\mathbf{X}_{g_j})} q(\mathbf{R}_{g_i}, \mathbf{R}_{g_j} | \mathbf{X}) \right\} \tag{3.5}$$

in case of a merge operator, or

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}^\star) = \min \left\{ 1, \frac{f(\mathbf{R}^\star_{g_i})f(\mathbf{R}^\star_{g_j})}{f(\mathbf{R}_g)} \frac{p(\mathbf{X}^\star_{g_i})p(\mathbf{X}^\star_{g_j})}{p(\mathbf{X}_g)} \frac{1}{q(\mathbf{R}^\star_{g_i}, \mathbf{R}^\star_{g_j} | \mathbf{X})} \right\} \tag{3.6}$$

in case of a split operator. The new value of $\boldsymbol{\theta}$ is finally sampled from the posterior distribution, after the Metropolis-Hastings step is completed.

For the proposed model (2.10), Split-Merge Metropolis-Hastings algorithm is implemented by computing probability (3.4) with equation (2.12), while $\mathbf{C}$ is directly sampled from the posterior distribution (2.13).

3.3.2 Gibbs sampler and Biased Random Walk

A local move in the partition space is accomplished by transferring a single observation from one cluster to another. By concatenating multiple local moves, the resulting Markov chain is able to explore in greater detail the surface of the posterior distribution. Since only one unit is transferred at any step, the process is usually very slow in escaping high-density regions of the posterior distribution and states of the Markov chain are often highly autocorrelated. In order to increase the effective sample size, very long chains are therefore required and the recommended strategy for a good mixing of the Markov chain is to combine both local and global moves.

Suppose $t-1$ complete iterations of the Markov chain have been performed and that, during iteration t, $i-1$ units have been already reallocated. We seek to sample the new cluster for unit i from the full conditional distribution

$$\Pr\{Z^{(t)}_{gi} = 1 | \mathbf{Z}^{(t)}_{1:(i-1)}, \mathbf{Z}^{(t-1)}_{(i+1):n}, \mathbf{X}\} \propto f(\mathbf{R}^\star_g \cup \{i\}) p(\mathbf{x}_i | \{\mathbf{x}_u : u \in \mathbf{R}^\star_g\}) \tag{3.7}$$

where $\mathbf{Z}$ is an indicator matrix of cluster associations, $f(\cdot)$ is the cohesion function of the product partition model, and $\mathbf{R}^\star_g$ is the current group g (not considering i). If g is an empty cluster, the probability is simply proportional to $f(\{i\})p(\mathbf{x}_i)$. For example, Neal (2000) reviews different ways of obtaining full conditional distributions associated to Dirichlet Process mixture models.

When the full conditional distributions are not be available in closed form, it is still possible to implement local moves with a Metropolis-Hastings correction

(Booth et al., 2008). The basic idea is to select a unit at random, with probability $1/n$, among the n available and to reallocate it into a group different from its own. In the general case, the new cluster is sampled with probability $1/k$ from the set of existing clusters plus the empty cluster (in which case, if selected, it would form a new group). Two special cases might also arise. If $k = 1$, the unit is put with probability 1 into its own cluster. If the unit is instead a singleton, it is allocated with probability $1/(k-1)$ into one of the other existing clusters.

The transition matrix constructed following this procedure is symmetric and the Metropolis-Hastings acceptance probability is simply

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}^{\star}) = \min\left\{1, \frac{p(\mathbf{R}^{\star})p(\mathbf{X}|\mathbf{R}^{\star})}{p(\mathbf{R})p(\mathbf{X}|\mathbf{R})}\right\} \tag{3.8}$$

which might greatly simplify, depending of the functional forms of the distributions involved.

Chapter 4

Applications

4.1 *S. pneumoniae* population structure

Streptococcus pneumoniae (pneumococcus) is a pathogenic bacterium, normally colonizing the nasopharynx of healthy individuals in an asymptomatic way. However, in the presence of a weak immune system, its carriage poses a serious risk for the development of diseases such as, but not restricted to, pneumonia (Bogaert et al., 2004). Carriage rates are usually higher in developing countries (Adetifa et al., 2012) and it is a leading cause of mortality of children worldwide (Rudan et al., 2008).

In article I, whole-genome sequence data of 3,085 pneumococci isolates were analysed using the statistical model explained in Section 2.2.1. The aim of the study was to improve our understanding of the mechanism of recombination at the pneumococcus population level and quantifying its impact in the development of antibiotic resistance.

The population structure found by means of unsupervised learning has been used as basis of the recombination analysis. 33 primary clusters were found by the algorithm, dividing the sample according to their serotypes. In a hierarchical fashion, each one of the 33 clusters was independently re-analysed obtaining a total of 183 secondary clusters in the whole dataset. The sub-clusters were mostly clonal complexes, that is groups of strains with very little variation, with non-typeable (NT) bacteria being the most numerous.

Analysis of recombination events focused on the seven largest primary clusters, comprising enough samples to obtain reliable results. Mutation rates within each cluster were very similar and compatible with the null hypothesis of equal rates. Recombination rates, instead, seemed to vary across groups with the NT cluster being the most recombinogenic. High recombination regions were found to be located at positions occupied by genes encoding antigens or associated with antibiotic resistance.

The estimated population structure, combined with phylogenetic analyses of these important genes, led to the hypothesis that specific lineages of pneumococci take the role of gene hubs. In particular, the NT group possessed high rates of both acquisition and donation of recombinant DNA, making them putative reservoir of genetic material for the whole population.

4.2 Haemagglutinin of influenza A/H3N2

Haemagglutinin (HA) is a protein located on the surface of the influenza virion with the role of binding the virus to the target cell and promoting its entrance by fusing the viral envelope together with the cell membrane. These two tasks are accomplished by two separate parts of the protein: HA1, which contains the binding sites, and HA2, which is involved in membrane fusion (Skehel and Wiley, 2000).

Initial observations of A/H3N2 date back to 1968. Since then, in order to escape the immune system of the target host, the HA protein has undergone rapid evolution with regular antigenic changes. Its short coalescent times can be clearly observed from its ladder-like phylogeny (Bedford et al., 2014; Fitch et al., 1991; Smith et al., 2004).

The enormous amount of viral strains available in public databases (Bao et al., 2008; Benson et al., 2005; Bogner et al., 2006; Squires et al., 2012) and the detailed information about its structure and evolution found in the literature (Bedford et al., 2014; Bizebard et al., 1995; Fleury et al., 1999; Knossow et al., 2002; Koel et al., 2013; Smith et al., 2004; Suzuki, 2006; Wolf et al., 2006) make the HA protein of influenza A/H3N2 a good candidate for testing the proposed model (2.10).

In article II, 4,898 unique HA sequences made of 567 amino acids were collected and analysed. MAP estimates suggested that viral proteins were split into a total of 57 different groups, discriminated by a joint set of 117 sites. This subset of protein amino acid positions were consistent with previous research (Smith et al., 2004), demonstrating how automatic procedures of unsupervised learning might highlight interesting spots of genetic evolution without huge effort from the practitioner.

Post-hoc analyses of the 57 original groups discovered the existence of 23 “core” clusters of strains, representing the backbone clades of the A/H3N2 HA phylogeny. Reassuringly, previous knowledge regarding HA A/H3N2 was reflected through these core clusters, in particular the dominant role of the B-cell epitopes in contrast to T-cell epitopes (Suzuki, 2006) and their distinct antigenic variation (Bedford et al., 2014).

4.3 VP4 protein of Rotavirus A

Rotavirus belongs to the family *Reoviridae* and is a double-stranded RNA virus causing gastroenteritis, of which species A is the most common variant across humans. Mostly affecting children and infants, it is still a major cause of death in developing countries (Bernstein, 2009).

Having a spike shape, VP4 is located on the surface of the virion. Similar to influenza A/H3N2 HA protein, its role is to bind the virus to receptors of the target cell and promote its entry. For this reason, it is constantly under selective pressure from the immune system.

For the statistical analysis conducted in article III, a total of 841 unique VP4 sequences, 783 amino acids long, were retrieved from NCBI's Virus Variation Resource database (Brister et al., 2013). Together with MAP estimation, 10^6 MCMC samples were also collected with default prior hyperparameters.

Posterior inference of the total number of groups placed the value of k between 11 and 17 (95% credible interval) with a most likely value of 13. Reassuringly, estimate of $\mathbf{R}$ shown to reflect already known serotypes of rotavirus A, splitting the dataset into a total of 11 groups. Posterior distribution of parameter $\mathbf{C}$ highlighted regions of most discriminating sites at the extremities of the protein, suggesting possible sites under heavy selection pressure.

Following a hierarchical approach similar to that applied in article I, the most common serotype (P[8]) was independently analysed. In this case, the dataset was perfectly split into 14 clusters with a posterior probability close to 1. In the global analysis, the clear population structure of P[8] was masked by general noise introduced by the other strains.

4.4 *C. jejuni* population structure

Campylobacter jejuni is a species of pathogenic bacteria that is generally found in the gastrointestinal tract of wild and domesticated animals (Sheppard et al., 2011). Humans are usually infected by ingestion of contaminated food (Friedman et al., 2004) and *C. jejuni* is the most common cause of bacterial food-borne diarrhoeal disease (Blaser, 1997). Gastroenteritis caused by *C. jejuni* is rarely life-threatening and antibiotics are usually not needed, but symptoms are severely debilitating and may last for a long period of time (Allos, 2001).

In article IV, 601 whole-genome protein sequences of *C. jejuni* isolates (Yahara et al., 2017) were analysed using the statistical model explained in Section 2.2.2. Sequences were 153,911 amino acids long with 17,405 polymorphic sites, resulting in a final binary data matrix of 601 rows and 37,666 columns. Available metadata included host information (poultry, ruminant, humans), sequence type (ST), and clonal complex (CC).

Bi-clustering results led to the discovery of 36 main bacterial groups. Integrating this information with each isolate source, we were able to further classify each cluster into one of 4 possible macro-groups: human (12 clusters), human and chicken (10 clusters), human and cattle (4 clusters), and human, chicken, and cattle (10 clusters). The most surprising discovery was the existence of groups of bacteria associated only with human patients and all belonging to the same clonal complex 21. This led to the hypothesis that a genetic bottleneck might be the main cause of the increase in relative frequency of particular kinds of bacterial strains in clinical isolates. Important amino acids in key genes are likely to be responsible of the higher fitness in human hosts and might be worth investigating. Phylogenetic analysis also showed an incongruence between DNA and protein group structures, where distinct lineages at the DNA level were instead convergent into the same protein group. These results show clearly how important it is to integrate both DNA and amino acid statistical analyses in any population genomics study.

Chapter 5

Conclusions

The steady increase in the amount of available genomics data poses great challenges to statistical inference but also brings many opportunities in understanding the evolution and transmission of pathogens, for a hope to develop better prevention and treatment of diseases. In this thesis, we showed how cluster analysis can be a valuable statistical tool for pathogen population genomics. Bayesian model-based unsupervised learning, in particular, gives the statistician complete freedom in building a model to target specific types of heterogeneity patterns that are known *a priori* to be present in the data.

In article I, for example, we demonstrated how the knowledge of the population structure of pneumococcus allows the biologists to test important biological hypotheses. In particular, we were able to observe how a specific lineage (NT) of pneumococci has a central role in the exchange of genetic material associated with antibiotics resistance.

A novel statistical model for identification of cluster-defining features in large categorical data was introduced in article II. Put to test with nearly 5000 HA protein sequences of influenza A/H3N2, we demonstrated its efficacy in highlighting amino acids under selective pressure from the host immune system. Further, the bi-clustering structure recovered the core evolution of the virus which was hidden by the noise introduced by the other sequences.

In article III the model was generalized and formulas were derived for implementing MCMC simulations. Posterior inference other than MAP estimation was done for a real dataset of 841 rotavirus A VP4 proteins. Results were compatible to current knowledge of rotavirus A, showing the reliability of inference results.

A real dataset of *Campylobacter jejuni* was finally analysed in article IV. Inference results allowed the discovery of particular lineages, and the amino acids involved, that are associated with increased virulence. Thanks to the method introduced in article II and III, novel biological hypotheses were made and tested, showing the usefulness of the model in pathogen population genomics.

Still a lot of potential remains for further research in this area. Current MCMC techniques are not able to cope with the ever increasing size of datasets. At the moment, for big datasets, we still recommend the use of MAP estimates as efficient alternatives to simulation techniques. Nevertheless, knowledge of the posterior distribution would be valuable for a better understanding of our uncertainty. For this reason, novel alternatives to the standard Gibbs sampler employed in Bayesian cluster analysis, or split-merge moves in Metropolis-Hastings algorithm, should be developed.

MAP estimation can be improved too. Stochastic search done by split-merge-transfer moves, despite its efficacy, is still computationally costly and might not scale well with new generation high-throughput data. At the moment, days or even weeks of computation is the norm for big genetic dataset. Implementation of parallel computation and recent advances in stochastic optimization, such as evolutionary algorithms, might be the winning approaches to this problem. The challenge that this branch has to overcome to become routinely applied is to provide good solutions in a reasonable amount of time, possibly minutes or even seconds.

References

I. M. O. Adetifa, M. Antonio, C. A. N. Okoromah, C. Ebruke, V. Inem, D. Nsekpong, A. Bojang, and R. A. Adegbola. Pre-vaccination nasopharyngeal pneumococcal carriage in a nigerian population: Epidemiology and population biology. *PLoS ONE*, 7(1):e30548, 2012. doi: 10.1371/journal.pone.0030548.

B. M. Allos. *Campylobacter Jejuni* infections: Update on emerging issues and trends. *Clinical Infectious Diseases*, 32(8):1201–1206, 2001. doi: 10.1086/319760.

E. B. Andersen. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331):1248–1255, 1970. doi: 10.1080/01621459.1970.10481160.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974. doi: 10.1214/aos/1176342871.

Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. The Influenza Virus Resource at the National Center for Biotechnology Information. *Journal of Virology*, 82(2):596–601, 2008. doi: 10.1128/JVI.02005-07.

D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, 20(1):260–279, 1992. doi: 10.1214/aos/1176348521.

T. Bedford, M. A. Suchard, P. Lemey, G. Dudas, V. Gregory, A. J. Hay, J. W. McCauley, C. A. Russell, D. J. Smith, and A. Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914, 2014. doi: 10.7554/eLife.01914.

D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 33(suppl 1):D34–D38, 2005. doi: 10.1093/nar/gki063.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Chichester, UK, 1994. ISBN 978-0-471-49464-5.

D. I. Bernstein. Rotavirus overview. *The Pediatric Infectious Disease Journal*, 28(3):S50–S53, 2009. doi: 10.1097/INF.0b013e3181967bee.

D. A. Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 1978. doi: 10.1093/biomet/65.1.31.

D. A. Binder. Approximations to Bayesian clustering rules. *Biometrika*, 68(1): 275–285, 1981. doi: 10.1093/biomet/68.1.275.

C. M. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer New York, 2006. ISBN 978-0-387-31073-2.

T. Bizebard, B. Gigant, P. Rigolet, B. Rasmussen, O. Diat, P. Bösecke, S. A. Wharton, J. J. Skehel, and M. Knossow. Structure of influenza virus haemagglutinin complexed with a neutralizing antibody. *Nature*, 376(6535):92–94, 1995. doi: 10.1038/376092a0.

M. J. Blaser. Epidemiologic and clinical features of *Campylobacter jejuni* infections. *Journal of Infectious Diseases*, 176(Supplement 2):S103–S105, 1997. doi: 10.1086/513780.

D. Bogaert, R. de Groot, and P. W. M. Hermans. *Streptococcus Pneumoniae* colonisation: The key to pneumococcal disease. *The Lancet Infectious Diseases*, 4(3):144–154, 2004. doi: 10.1016/S1473-3099(04)00938-7.

P. Bogner, I. Capua, D. J. Lipman, N. J. Cox, and others. A global initiative on sharing avian flu data. *Nature*, 442(7106):981, 2006. doi: 10.1038/442981a.

J. G. Booth, G. Casella, and J. P. Hobert. Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):119–139, 2008. doi: 10.1111/j.1467-9868.2007.00629.x.

J. R. Brister, Y. Bao, S. A. Zhdanov, Y. Ostapchuck, V. Chetvernin, B. Kiryutin, L. Zaslavsky, M. Kimelman, and T. A. Tatusova. Virus Variation Resource - recent updates and future directions. *Nucleic Acids Research*, 42(D1):D660–D665, 2013. doi: 10.1093/nar/gkt1268.

G. Casella, E. Moreno, and F. J. Girón. Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9(3):613–658, 2014. doi: 10.1214/14-BA869.

L. Cheng, T. R. Connor, J. Sirén, D. M. Aanensen, and J. Corander. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution*, 30(5):1224–1228, 2013. doi: 10.1093/molbev/mst028.

J. Corander and P. Marttinen. Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, 15(10):2833–2843, 2006. doi: 10.1111/j.1365-294X.2006.02994.x.

H. Crane. The ubiquitous Ewens sampling formula. *Statistical Science*, 31(1): 1–19, 2016. doi: 10.1214/15-STS529.

D. B. Dahl. Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models. Technical report, Department of Statistics, Texas A&M University, 2005.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38, 1977.

R. Eisner, C. Stretch, T. Eastman, J. Xia, D. Hau, S. Damaraju, R. Greiner, D. S. Wishart, and V. E. Baracos. Learning to predict cancer-associated skeletal muscle wasting from ^{1}H-NMR profiles of urinary metabolites. *Metabolomics*, 7 (1):25–34, 2011. doi: 10.1007/s11306-010-0232-9.

W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112, 1972. doi: 10.1016/0040-5809(72)90035-4.

W. M. Fitch, J. M. E. Leiter, X. Q. Li, and P. Palese. Positive darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 88(10):4270–4274, 1991. doi: 10.1073/pnas.88.10.4270.

D. Fleury, B. Barrère, T. Bizebard, R. S. Daniels, J. J. Skehel, and M. Knossow. A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site. *Nature Structural & Molecular Biology*, 6(6):530–534, 1999. doi: 10.1038/9299.

C. R. Friedman, R. M. Hoekstra, M. Samuel, R. Marcus, J. Bender, B. Shiferaw, S. Reddy, S. D. Ahuja, D. L. Helfrick, F. Hardnett, M. Carter, B. Anderson, and R. V. Tauxe. Risk factors for sporadic *Campylobacter* infection in the United States: A case-control study in FoodNet sites. *Clinical Infectious Diseases*, 38 (Supplement_3):S285–S296, 2004. doi: http://dx.doi.org/10.1086/381598.

A. Fritsch and K. Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391, 2009. doi: 10.1214/09-BA414.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models.* Springer Series in Statistics. Springer New York, New York, NY, USA, 2006. ISBN 978-0-387-35768-3.

M. R. Gupta and Y. Chen. Theory and use of the EM algorithm. *Foundations and Trends in Signal Processing*, 4(3):223–296, 2011. doi: 10.1561/2000000034.

J. A. Hartigan. Partition models. *Communications in Statistics - Theory and Methods*, 19(8):2745–2756, 1990. doi: 10.1080/03610929008830345.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, Berlin, Germany, 2 edition, 2009. ISBN 978-0-387-84858-7.

F. M. Hoppe. The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology*, 25(2):123–159, 1987. doi: 10.1007/BF00276386.

S. Jain and R. M. Neal. Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007. doi: 10.1214/07-BA219.

J. F. C. Kingman. Random partitions in population genetics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 361(1704):1–20, 1978. doi: 10.1098/rspa.1978.0089.

M. Knossow, M. Gaudier, A. Douglas, B. Barrère, T. Bizebard, C. Barbey, B. Gigant, and J. J. Skehel. Mechanism of neutralization of influenza virus infectivity by antibodies. *Virology*, 302(2):294–298, 2002. doi: 10.1006/viro.2002.1625.

B. F. Koel, D. F. Burke, T. M. Bestebroer, S. van der Vliet, G. C. M. Zondag, G. Vervaet, E. Skepner, N. S. Lewis, M. I. J. Spronken, C. A. Russell, M. Y. Eropkin, A. C. Hurt, I. G. Barr, J. C. de Jong, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, R. A. M. Fouchier, and D. J. Smith. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979, 2013. doi: 10.1126/science.1244730.

J. Kohonen and J. Corander. Computing exact clustering posteriors with subset convolution. *Communications in Statistics - Theory and Methods*, 45(10):3048–3058, 2016. doi: 10.1080/03610926.2014.894070.

S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004. doi: 10.1109/TCBB.2004.2.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. Wiley, 2 edition, 2008. ISBN 978-0-470-19160-6.

G. McLachlan and D. Peel. *Finite Mixture Models.* Wiley Series in Probability and Statistics. Wiley, 2000. ISBN 978-0-471-00626-8.

B. Mirkin. *Mathematical Classification and Clustering*, volume 11 of *Nonconvex optimization and its applications.* Springer US, 1996. ISBN 978-0-7923-4159-8.

R. M. Neal. Markov Chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. doi: 10.1080/10618600.2000.10474879.

K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110, 1894. doi: 10.1098/rsta.1894.0003.

J. Pitman. *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics.* Springer, 2006. ISBN 978-3-540-34266-3.

C. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer Texts in Statistics. Springer New York, 2 edition, 2004. ISBN 978-0-387-21239-5.

C. P. Robert and K. L. Mengersen. Reparameterisation issues in mixture modelling and their bearing on MCMC algorithms. *Computational Statistics & Data Analysis*, 29(3):325–343, 1999. doi: 10.1016/S0167-9473(98)00058-9.

L. Rokach. A survey of clustering algorithms. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 269–298. Springer US, Boston, MA, USA, 2 edition, 2010. ISBN 978-0-387-09823-4.

I. Rudan, C. Boschi-Pinto, Z. Biloglav, K. Mulholland, and H. Campbell. Epidemiology and etiology of childhood pneumonia. *Bulletin of the World Health Organization*, 86(5):408–416, 2008. doi: 10.1590/S0042-96862008000500019.

S. K. Sheppard, F. M. Colles, N. D. McCarthy, N. J. C. Strachan, I. D. Ogden, K. J. Forbes, J. F. Dallas, and M. C. J. Maiden. Niche segregation and genetic structure of *Campylobacter jejuni* populations from wild and agricultural host species. *Molecular Ecology*, 20(16):3484–3490, 2011. doi: 10.1111/j.1365-294X.2011.05179.x.

J. J. Skehel and D. C. Wiley. Receptor binding and membrane fusion in virus entry: The influenza hemagglutinin. *Annual Review of Biochemistry*, 69(1):531–569, 2000. doi: 10.1146/annurev.biochem.69.1.531.

D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004. doi: 10.1126/science.1097211.

J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley Series in Discrete Mathematics and Optimization. Wiley, Hoboken, New Jersey, USA, 2003. ISBN 978-0-471-33052-3.

R. B. Squires, J. Noronha, V. Hunt, A. García-Sastre, C. Macken, N. Baumgarth, D. Suarez, B. E. Pickett, Y. Zhang, C. N. Larsen, A. Ramsey, L. Zhou, S. Zaremba, S. Kumar, J. Deitrich, E. Klem, and R. H. Scheuermann. Influenza Research Database: An integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses*, 6(6):404–416, 2012. doi: 10.1111/j.1750-2659.2011.00331.x.

Y. Suzuki. Natural selection on the influenza virus genome. *Molecular Biology and Evolution*, 23(10):1902–1911, 2006. doi: 10.1093/molbev/msl050.

I. Van Mechelen, H.-H. Bock, and P. De Boeck. Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research*, 13(5):363–394, 2004. doi: 10.1191/0962280204sm373ra.

Y. I. Wolf, C. Viboud, E. C. Holmes, E. V. Koonin, and D. J. Lipman. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct*, 1(1):34, 2006. doi: 10.1186/1745-6150-1-34.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34, 2015. doi: 10.1007/s10107-015-0892-3.

C.-F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

K. Yahara, G. Méric, A. J. Taylor, S. P. W. de Vries, S. Murray, B. Pascoe, L. Mageiros, A. Torralbo, A. Vidal, A. Ridley, S. Komukai, H. Wimalarathna, A. J. Cody, F. M. Colles, N. McCarthy, D. Harris, J. E. Bray, K. A. Jolley, M. C. J. Maiden, S. D. Bentley, J. Parkhill, C. D. Bayliss, A. Grant, D. Maskell, X. Didelot, D. J. Kelly, and S. K. Sheppard. Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork. *Environmental Microbiology*, 19(1):361–380, 2017. doi: 10.1111/1462-2920.13628.