Bayesian Stochastic Partition Models For Markovian Dependence Structures

Väinö Jääskinen

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in the Linus Torvalds auditorium (B123), Exactum, on February 6th, 2015, at 12 o'clock noon.

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki

HELSINKI 2015

ISBN 978-951-51-0486-1 (paperback)
ISBN 978-951-51-0487-8 (PDF)
Unigrafia
http://ethesis.helsinki.fi/
HELSINKI 2015

Supervisor	Professor Jukka Corander Department of Mathematics and Statistics University of Helsinki Finland
Pre-examiners	Assistant Professor Harri Lähdesmäki Department of Information and Computer Science Aalto University Finland
	Professor Jaakko Nevalainen School of Health Sciences University of Tampere Finland
Custos	Professor Jukka Corander Department of Mathematics and Statistics University of Helsinki Finland
Opponent	Reader Korbinian Strimmer Department of Epidemiology and Biostatistics Imperial College London United Kingdom

Abstract

In various fields of knowledge we can observe that the availability of potentially useful data is increasing fast. A prime example is the DNA sequence data. This increase is both an opportunity and a challenge as new methods are needed to benefit from the big data sets. This has sparked a fruitful line of research in statistics and computer science that can be called machine learning. In this thesis, we develop machine learning methods based on the Bayesian approach to statistics. We address a fairly general problem called clustering, i.e. dividing a set of objects to non-overlapping group based on their similarity, and apply it to models with Markovian dependence structures. We consider sequence data in a finite alphabet and present a model class called the Sparse Markov chain (SMC). It is a special case of a Markov chain (MC) model and offers a parsimonious description of the data generating mechanism. A Variable length Markov chain (VLMC) is a popular sparse model presented earlier in the literature and it has a representation as an SMC model. We develop Bayesian clustering methodology for learning the SMC and other Markovian models.

Another problem that we study in this thesis is causal inference. We present a model and an algorithm for learning causal mechanisms from data. The model can be considered as a stochastic extension of the sufficient-component cause model that is popular in epidemiology. In our model there are several causal mechanisms each with its own parameters. A mixture distribution gives a probability that an outcome variable is associated with a mechanism.

Applications that are considered in this thesis come mainly from computational biology. We cluster states of Markovian models estimated from DNA sequences. This gives an efficient description of the sequence data when comparing to methods reported in the literature. We also cluster DNA sequences with Markov chains, which results in a method that can be used for example in the estimation of bacterial community composition in a sample from which DNA is extracted. The causal model and the related learning algorithm are able to estimate mechanisms from fairly challenging data. We have developed the learning algorithms with big data sets in mind. Still, there is a need to develop them further to handle ever larger data sets.

Acknowledgements

I am grateful for the funding provided by the Finnish Doctoral Programme in Computational Sciences (FICS). I also want to thank the Finnish Centre of Excellence in Computational Inference Research (COIN) for financial support.

I wish to thank my supervisor Jukka Corander for guiding me through the academic jungle. In all kinds of situations, he has been helpful and his enthusiasm continues to inspire me. He has taught me that doing science is both a serious business and a lot of fun.

I want to thank Elja Arjas for helping me to get started. I am thankful to my mentors Arto Klami and Jouko Lampinen for discussions and for giving me some perspective.

I want to express my gratitude to all my collaborators and co-authors: Jie, Ville, Lu, Helga, Jan Hillert and Timo Koski whom I want to also thank for hosting my visit at KTH in 2013. Working with these people has taught me about myself and scientific research, and overall it has been an enjoyable ride.

I want to thank present and former members of our research group: Jukka S., Jukka K., Paul, Elina, Minna, Alberto, Mikhail and others. Our discussions have meant a lot to me. I have not always been the keenest one to attend the statistics discussion club, but I appreciate your efforts to making our group more social and active. I think you have done a good job.

I am grateful for my mathematical friends Jarmo Jääskeläinen, Tanja Toivanen and Tuomas Orponen. You have taught me something about

mathematics, but more about life. I also want to thank Jan Cristina and other colleagues for making Kumpula a friendly place.

I am thankful to Jukka Luoma for the camaraderie that has developed as we have moved along our academic paths.

Finally, I want to express my gratitude to my friends and family. Your presence really makes a difference.

Helsinki, September 2014

Väinö Jääskinen

List of original articles

This thesis consists of four articles and an introductory part. We refer to the articles by Roman numerals I-IV.

I. Väinö Jääskinen, Jie Xiong, Jukka Corander, and Timo Koski. Sparse Markov chains for sequence data. *Scandinavian Journal of Statistics*, 41(3):639-655, 2013.

II. Väinö Jääskinen, Ville Parkkinen, Lu Cheng, and Jukka Corander. Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. *Statistical Applications in Genetics and Molecular Biology*, 13(1):105-121, 2013.

III. Jie Xiong, Väinö Jääskinen, and Jukka Corander. Recursive learning for sparse Markov models. Submitted, 2014.

IV. Helga Westerlind, Väinö Jääskinen, Jukka Corander, Jan Hillert, and Timo Koski. Estimation of Mixtures of Multicausal Interaction Networks. Submitted, 2014.

Author's contribution to Articles I-IV

I. VJ contributed to developing the method and had the main responsibility for the implementation and empirical testing, while JX also contributed to these. VJ also participated in writing the article, while JC had the main role.

II. VJ and VP contributed equally to the implementation, empirical results and writing the article. The method was mainly developed by JC and VP.

III. The method was jointly developed by the authors. VJ participated in the implementation, empirical testing and writing the article, while JX had the main responsibility for these.

IV. VJ participated in developing the method and writing the paper, while TK had the main role. VJ contributed to the implementation and testing the model, while HW had the main responsibility for these and empirical testing. This paper was part of the PhD thesis of Helga Westerlind at Karolinska Institut.

Contents

1

Introduction

One characteristic of the contemporary world is the abundance of potentially relevant data in various fields. This holds for scientific research, commercial applications and technology in general. There is a pressing need for new methods in data analysis. In statistics, prevailing attitude to big data in all of its forms has been somewhat ambivalent. This is because a given statistical method usually works for data sets that are large enough but not too large. In this thesis, the theme of large data sets and how to handle them is explored. Statistics is not the only discipline that is concerned with large data sets, as computer science is also intimately involved in these matters. Indeed, the relatively new field of machine learning combines perspectives from both statistics and computer science.

In this thesis, we start with the core ideas of Bayesian statistics and apply them to machine learning in applications including computational biology, especially analysis of genome sequence variation. This is natural as the Bayesian approach to statistics has become important for machine learning (Bishop, 2006). Bayesian statistics has a long and interesting history starting from Thomas Bayes in the 18th century (Bernardo and Smith, 1994).

A central motivation for this work comes from sequential data as it enables the use of Markovian models in their rich variety. These models have a history of over hundred years. The first definition and also application of Markov chains was by Andrei Markov to model probabilities of vowels and consonants in Alexander Pushkin's verse novel *Eugene Onegin* (Hayes,

2013). As an example of sequence data we have in Article II 91 240 DNA sequences each having after preprocessing an approximate length of 500 base pairs. Another issue we consider is the study of causal inference, especially in the context of epidemiology. For example, we can think of a situation where a disease has two causing mechanisms and the effective mechanism is chosen randomly for each individual. Then the probability of catching the disease depends on the chosen mechanism and the covariates, for example age and sex. One unifying theme in this thesis is the use of classification, clustering and probabilistic reasoning to make sense of large data sets.

Here, we present some facts about DNA as it is central for applications described in this thesis. In DNA (deoxyribonucleic acid), there are four types of nucleobases: adenine (A), guanine (G), thymine (T) and cytosine (C) (Kimball, 1994). These are nitrogen-containing ring-like structures. Bases take part in forming polymers of nucleotides. DNA itself is such a polymer of nucleotides and it has two strands giving it the double helix structure. The genetic code of the DNA includes information that is used in the synthesis of proteins and thus it controls development of living organisms together with environmental factors.

The following chapters provide background to the four articles and highlight important issues. In Chapter 2, the Bayesian approach to statistics is presented. Both key equations and modeling principles are covered. In Chapter 3, Markovian models are described. There are several models that each share a form of Markov property. Chapter 4 includes graphical representations of the Markovian models. These illustrate the similarity and differences between Markovian models. In Chapter 5, the framework of Bayesian clustering is introduced. This chapter includes a variety of concepts that are needed for modeling as well as description of algorithms. Chapter 6 describes causal inference from both epidemiological and probabilistic point of view, showing examples of theories related to causality. Finally, Chapter 7 presents conclusions based on the articles and other material as well as points directions for future research.

Here, we describe some of the notation used in this thesis. Overall, notation should be clear from the context. Notation in each chapter is chosen to bring out the subject matter with clarity. This leads to a situation where

there are several symbols for concepts that are similar and on the other hand multiple uses for a single symbol. This clarification is provided to help the reader.

Firstly, there are several symbols for data. In Chapter 2, X can be considered as a random variable. Its dimension is not defined explicitly. In Chapter 3, X_n is a random variable and n is an index variable. A realization of X_n is denoted by j_n or x_n. In Chapter 5, $\mathbf{x}$ denotes a set of observations with n data objects. The dimension of the data objects varies depending on the context. x represents the observed data in the context of expectation-maximization (EM) algorithm.

Generally, θ represents collectively quantitative parameters. Also its dimension varies depending on the context. In the context of Markov chains, n means the total number of observed transitions. In the context of clustering, it can also mean the number of data objects to be clustered. S denotes a partition variable. In Chapter 2, both H both and A are taken to mean a hypothesis.

In the description of the classification EM algorithm, M is the number of mechanisms and N the number of observations. j denotes an index of a mechanism and l an index of an observation (subject). $y^{(l)}$ denotes the health status of an observed subject l (ill or not) and $x^{(l)}$ denotes the covariate vector of the same subject. Assignment of each observation to a mechanism is denoted by $u^{(l)}$. An alternative representation is given by indicator variables of the form ${u_j}^{(l)}$.

2

Bayesian Statistics

2.1 The Bayesian Approach

Here, we present some central aspects of Bayesian statistics as a background for later chapters. Generally, in statistical inference we are concerned with hypothesis about quantitative parameters θ and we want to assess these hypothesis based on empirical data X. This means using data to draw conclusions about unobserved quantities (Gelman et al., 2004).

A fundamental result for Bayesian statistics is the Bayes' rule.

Theorem 2.1. *Bayes' rule. Assume we have defined probability distributions $P(A)$, $P(X|A)$, and $P(X) \neq 0$. Then*

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}. \tag{2.1}$$

$P(A)$ is the *prior* distribution for hypothesis A. Typically, A is a hypothesis about some quantitative parameters θ. The prior measures the degree of belief about A being true before we have observed X. $P(X|A)$ is the *likelihood* of observing X given A is true. *Evidence* $P(X)$ is the marginal probability of X. The rule of total probability yields

$$P(X) = P(X|A)P(A) + P(X|\neg A)P(\neg A).$$

$P(A|X)$ is the *posterior* probability of A given we have observed X. We can say that the Bayes' rule updates the degree of the prior belief $P(A)$ yielding the updated belief as measured by the posterior probability $P(A|X)$.

The Bayes' rule follows directly from the definition of conditional probability. A key characteristic of Bayesian inference in contrast with the frequentist paradigm is the use of probability for measuring uncertainty (Bernardo and Smith, 1994). In Bayesian inference, probabilities are interpreted as degrees of belief. In the frequentist interpretation, probabilities are considered as limits of relative frequencies. Especially, defining the prior distribution $P(A)$ with limits of relative frequencies can be problematic. Typically, it can be constructed with hypothetical repetitions under identical conditions (Gelman et al., 2004). But this can seem artificial, if for example we are modeling unique or almost unique events. Furthermore, in a strict frequentist interpretation an event that has not occurred would have probability zero, which can seem unintuitive in many situations (Hájek, 1996). In comparison, when $P(A)$ is interpreted as a prior belief about A, we get a coherent system of inference. The Bayesian approach to statistics can be derived from decision theoretic considerations (Bernardo and Smith, 1994). That way coherence of the inference system can be demonstrated. Alternatively, we can adopt the pragmatist view and state that Bayesian statistics has proven to be useful in the analysis of applied problems in many fields. This then justifies the use of (2.1) and the Bayesian approach in general.

2.2 Modeling

In practice, Bayesian inference includes but is not limited to specifying the likelihood and the prior. This is evident in Articles I, II and III which contain for example algorithms for learning the model in question. Defining these type of algorithms can be a central part of the statistical modeling effort.

When modeling real-world phenomena, all the relevant aspects of the process we are modeling cannot often be included in the model. There exists a separation between a "theoretical world" and a "real world" (Kass, 2011). In the "theoretical world" we have mathematics and statistical models while

in the "real world" exist the actual phenomena to be studied and data. Statistical modeling can be seen as building a bridge between these two worlds. As part of the modeling, simplifying assumptions have to be made. Also, the usefulness of the model is something that has to be considered. A statistical model can be an adequate description of a phenomenon but not very useful if it is computationally too burdensome to be used for any purpose. This is evident for example when high-order Markov chains are used for analysing sequence data. Increasing the model order easily leads to models that are not useful because of the huge volume of computations involved.

One feature of the Bayesian approach is predictive inference. We can calculate marginal probability distributions for hypothetical and observed data. Also, predictive power is a useful measure of the suitability of a model (Bernardo and Smith, 1994). If the model predicts accurately new observations, then it seems to be an efficient description of the underlying phenomenon. A principle called Occam's razor states that a simpler hypothesis should be preferred instead of unnecessarily complicated ones. In Bayesian statistics, this principle is at work in model comparison (MacKay, 1992). The idea is to calculate the marginal likelihood of the data for each competing hypothesis with (2.2). The probability distribution defined in (2.2) is also called the prior predictive distribution. These probabilities can be used for calculating Bayes factors (Kass and Raftery, 1995).

Definition 2.2. Marginal likelihood of the data and Bayes factor. For hypotheses, i.e. models H_1 and H_2 and data X, the marginal likelihood of data under the model $k, k = 1, 2$ is

$$P(X|H_k) = \int P(X|\theta_k, H_k)\pi(\theta_k|H_k)d\theta_k \tag{2.2}$$

where θ_k denotes collectively parameters of the model k, $P(X|\theta_k, H_k)$ is the likelihood and $\pi(\theta_k|H_k)$ is the prior. Then the Bayes factor is

$$B_{12} = \frac{P(X|H_1)}{P(X|H_2)}. \tag{2.3}$$

Here, the Bayes factor measures plausibility of H_1 in comparison to H_2. If the model has unnecessary parameters, this is penalized in the evidence (2.2) (MacKay, 1992).

A form of prediction is to calculate the posterior predictive distribution for future observations $\tilde{X}$ given data X and a model. Following previous notation, the posterior predictive distribution of $\tilde{X}$ under model k and given X is

$$P(\tilde{X}|X, H_k) = \int P(\tilde{X}|\theta_k, H_k)\pi(\theta_k|X, H_k)d\theta_k \tag{2.4}$$

where $P(\tilde{X}|\theta_k, H_k)$ is the likelihood of $\tilde{X}$ and $\pi(\theta_k|X, H_k)$ is the posterior distribution of θ_k.

In machine learning, data is often divided to subsets for learning and testing the model (Bishop, 2006). This is done to avoid over-fitting. If the model has a large number of parameters, it might describe well the training data but still fail to generalize to new data. The marginal likelihood of the data under a given model is a useful tool of Bayesian statistics that can be applied to many problems in machine learning. This is done in Articles I, II and III.

3

Markovian Models

Many fields of science and technology depend on the analysis of sequence data. Two prime examples are DNA sequences in biology and text in the context of processing natural language. Given empirical sequence data, a mathematical framework is needed for quantitative modeling (Koski, 2001, see also Article I). First, we consider a finite alphabet $S = \{s_1, s_2, \ldots, s_J\}$ with J symbols. An example of a finite alphabet is the DNA alphabet $\mathcal{X} = \{A, C, G, T\}$. We can label the symbols with integers and generally consider the following alphabet: $\mathcal{X} = \{1, \ldots, J\}$. Let $X_0, X_1, \ldots, X_n$ be a sequence of random variables that take values in $\mathcal{X}$.

3.1 Markov Chain

A fundamental model for sequence data is the Markov chain.

Definition 3.1. Time homogeneous Markov chain (MC). Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of random variables. If for all $n \geqslant 1$ and $j_0, j_1, \ldots, j_n \in \mathcal{X}$,

$$P(X_n = j_n | X_{n-1} = j_{n-1}, \ldots, X_0 = j_0) = P(X_n = j_n | X_{n-1} = j_{n-1}), \quad (3.1)$$

then $\{X_n\}_{n=0}^{\infty}$ is called a Markov chain.

Elements in $\mathcal{X}$ are called states of the Markov chain. A Markov chain has the *Markov property* defined by the equation (3.1). In essence, Markov property states that given the previous state $X_{n-1} = j_{n-1}$, the rest of the history

is irrelevant for predicting the current state X_n. Probabilities of transitions between states can be represented in a transition matrix with elements $p_{i|j} = P(X_n = j|X_{n-1} = i), i, j \in \mathcal{X}$. These probabilities are assumed to be independent of n making the Markov chain *time homogeneous*.

Another important model is a Markov chain of mth order.

Definition 3.2. Time homogeneous Markov chain (MC) of mth order. Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of random variables. If for all $n \geqslant m$ and $j_0, j_1, \ldots, j_n \in \mathcal{X}$,

$$P(X_n = j_n|X_{n-1} = j_{n-1}, \ldots, X_0 = j_0) = \\ P(X_n = j_n|X_{n-1} = j_{n-1}, \ldots, X_{n-m} = j_{n-m}), \tag{3.2}$$

for a positive integer m, then $\{X_n\}_{n=0}^{\infty}$ is called a Markov chain of mth order.

A Markov chain is called a first-order Markov chain in this context. A Markov chain $\{X_n\}_{n=0}^{\infty}$ of mth order can be transformed to a first-order Markov chain $\{Z_n\}_{n=0}^{\infty}$ with transition probabilities $p_{i|j}, i, j \in \mathcal{X}^m$ and an extended state space with $|\mathcal{X}|^m = J^*$ states. Transition matrix of the first-order Markov chain then has $|\mathcal{X}|^m$ rows and each row has exactly J transition probabilities that can have positive values.

Markov chains are useful for modeling different type of phenomena. Often, the Markov assumption describes reasonably well the dynamics of a real-world system. Heuristically, it is plausible to assume that what is close proximity affects the current state more than what is further apart.

In a Markov chain model, the order of the model controls how much of the history of the process is used at each step. In principle, it would be tempting to examine all of the history at once. However, realities of modeling limit the possibilities. The number of free parameters for a Markov chain of mth order is $|\mathcal{X}|^m(|\mathcal{X}| - 1)$. A Markovian model that has this number of free parameters is called a full Markov chain. This number grows exponentially with the order of the Markov chain. Thus, the number of observed transitions per state decreases when the order grows. This leads to difficulties in the estimation of transition probabilities. Also, this means

that finding an optimal order m for the Markov chain model is far from trivial.

Due to challenges associated with using mth order Markov chains, several alternatives have been presented. They are typically based on the idea of sparsity, aiming to find parsimonious descriptions of the data generating process. Starting from the mth order Markov chain model, we can try to find models that utilize the same information but in a more effective manner. This leads to the question of data compression. There is a rich literature on the Variable order Markov models (VOM) starting from Rissanen (1983). Here we focus on Variable length Markov chains, Sparse Markov chains and Mixture transition distribution models.

3.2 Variable Length Markov Chain

A realized value for a random variable X_t is denoted by x_t.

Definition 3.3. Variable length Markov chain (VLMC). Let $\{X_n\}_{n=0}^{\infty}$ be a time homogeneous Markov chain of mth order. Denote by $c_{\text{pre}} : \mathcal{X}^m \to \cup_{j=0}^{m} \mathcal{X}^j$ a function which maps $c_{\text{pre}} : x_{n-1}, \ldots, x_{n-m} \mapsto x_{n-1}, \ldots, x_{n-l}$ where

$$
\begin{aligned}
&l = l(x_{n-1}, \ldots, x_{n-m}) = \\
&\min\{k \in \mathbb{Z}_{\geq 0}; P(X_n = j_n | X_{n-1} = x_{n-1}, \ldots, X_{n-m} = x_{n-m}) = \\
&P(X_n = j_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}) \text{ for all } j_n \in \mathcal{X}\},
\end{aligned}
$$

such that $l = 0$ corresponds to independence. Then l is a variable length memory and $c_{\text{pre}}(\cdot)$ is the preliminary context function. Final context function $c(\cdot)$ is obtained by lumping together some of the values of $c_{\text{pre}}(\cdot)$ that share the second to last symbol. A Markov chain of mth order with a variable length memory l is called a Variable length Markov chain of order p where p is the smallest integer such that $l(x_{n-1}, \ldots, x_{n-m}) \leqslant p \leqslant m$ for all $x_{n-1}, \ldots, x_{n-m} \in \mathcal{X}^m$.

This definition of Variable length Markov chain is slightly different from those in Bühlmann and Wyner (1999) and Mächler and Bühlmann (2004)

as they start from a stationary process in finite alphabet instead of a Markov chain of mth order as is done here.

3.3 Mixture Transition Distribution

Another parsimonious model for high-order Markov chains is the Mixture transition distribution model (Raftery, 1985; Le et al., 1996; Berchtold and Raftery, 2002). It has been extended from modeling of high-order Markov chains in a finite state space to for example general state spaces. Here, the focus is on applying MTD models for high-order Markov chains in a finite state space.

Definition 3.4. Mixture transition distribution (MTD). Let $\{X_n\}_{n=0}^{\infty}$ be a time homogeneous Markov chain of mth order. In the corresponding MTD model we have

$$\begin{aligned} P(X_n = j_n | X_{n-1} = j_{n-1}, \ldots, X_{n-m} = j_{n-m}) = \\ \sum_{g=1}^{m} \lambda_g P(X_n = j_n | X_{n-g} = j_{n-g}) = \\ \sum_{g=1}^{m} \lambda_g p_{j_{n-g} | j_n} \end{aligned} \tag{3.3}$$

with constraints

$$\begin{aligned} \sum_{g=1}^{m} \lambda_g = 1, \\ \lambda_g \geqslant 0. \end{aligned}$$

In the MTD model, contributions of the lags $(1 \ldots m)$ are combined additively. This offers a parsimonious model of the Markov chain. The model has $|\mathcal{X}|(|\mathcal{X}|-1)+(m-1)$ free parameters which is clearly less than $|\mathcal{X}|^m(|\mathcal{X}|-1)$ of the full mth order Markov chain.

VLMC and MTD are complementary models in a sense that they have different strengths (Berchtold and Raftery, 2002). MTD model of mth order always deals with m lags while in the VLMC model part of the history is irrelevant depending on the context. Berchtold and Raftery report a comparison of these two models.

3.4 Sparse Markov Chain

Finally, we present the Sparse Markov chain (SMC). Such models are defined in Article I, and they offer another alternative to mth order Markov chains.

Definition 3.5. Sparse Markov chain (SMC). Let $\{X_n\}_{n=0}^{\infty}$ be a time homogeneous Markov chain of finite order m transformed to a first-order MC $\{Z_n\}_{n=0}^{\infty}$. Let $S = (s_1, \ldots, s_k)$ be a partition of $\mathcal{X}^m$ such that the transition probability vectors satisfy the equality $p_{i|\cdot} = p_{j|\cdot}$ for all pairs of states $\{i, j\} \in s_c, c = 1, \ldots, k$, and $\mathcal{P}$ the corresponding set of k transition probability distributions in $\mathcal{X}^m$. If $k < |\mathcal{X}^m|$, the pair $(S, \mathcal{P})$ is called an SMC (of order m).

The following theorem characterizes connection between SMC and VLMC models.

Theorem 3.6. *Let $(S, \mathcal{P})$ be an SMC. Then, there is an equivalent representation based on the set of contexts $\mathcal{B}$ of a VLMC model if and only if there exists a unique context B_c with $b^{(r)}$, which is a suffix to all states i assigned to the same class s_c for all $c = 1, \ldots, k$.*

The proof is given in Article I. Theorem (3.6) formally states that for a VLMC model that is not a full Markov chain, there is an equivalent representation as an SMC model. The reverse is not generally true. There are SMC models that do not have representation as a VLMC model as is shown in the proof. For example, in Article I an SMC model is described for which X_{n-1} is irrelevant for predicting X_n while X_{n-2} is relevant. This kind of probability model cannot be described with the context tree of a VLMC model.

For estimating Markovian models, there are plenty of methods. For any Markov chain of order m, maximum likelihood based methods can be used. For VLMC models, context algorithm can be used (Bühlmann and Wyner, 1999; Rissanen, 1983). For a MTD model, numerical maximization of the log-likelihood or the expectation-maximization (EM) algorithm can be used (Berchtold and Raftery, 2002). In Chapter 5, a Bayesian method for learning Markovian models is presented.

4

Graphical Representations

Earlier in Chapter 3, the following Markovian models were introduced: MC, VLMC and SMC. Here, we give a graphical representation of these models (see Article I). We find a DAG (directed acyclical graph) for the sample paths of each of the models. In all of them, we have a tree structure that represents probability distributions for the random variable X_n conditional on values of $X_{n-1}, \ldots, X_{n-m}$. Generally, a DAG consists of nodes and directed edges between them so that there are no cycles (Koski and Noble, 2009). Here, X_n corresponds to the root node in the graph. Other nodes in the tree consist of possible values (or a set of them) of $X_{n-1}, \ldots, X_{n-m}$. Here, m is the order of the Markovian model. Directions of edges are from X_{n-m} to X_n.

Firstly, we consider a full MC of order m in DNA alphabet. All the possible histories are represented in a tree. Example is given in Figure 4.1. The number of leaves is $|\mathcal{X}|^m$ while the number of free parameters is $|\mathcal{X}|^m(|\mathcal{X}| - 1)$. Here, leaves are those nodes for which no edges are directed at.

For a VLMC model of order m, part of the full tree has been pruned. This corresponds to the situation where for given a set of sample paths the same probability distribution always holds for the random variable X_n. In comparison with the SMC model, this set of sample paths has to be hierarchical, i.e. it has to have one common path to the root node X_n. Nodes can be lumped together in two ways. Firstly, history beyond some node can be irrelevant. For example, in Figure 4.2 if $X_{n-1} = T$, then all

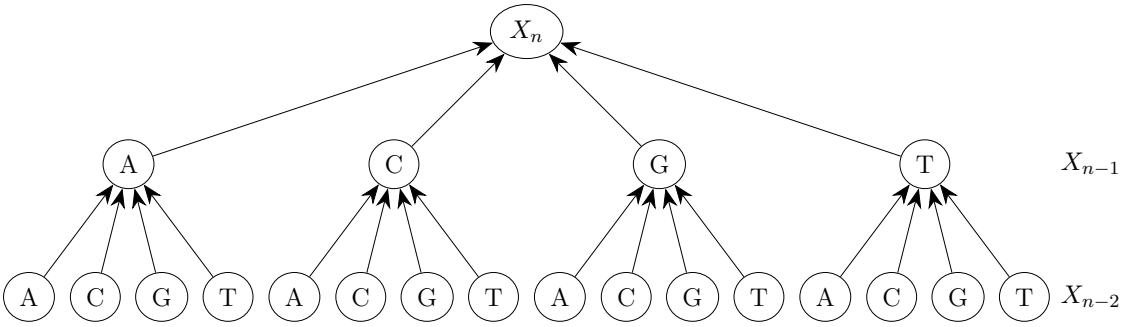

Figure 4.1: A full Markov chain of order 2 in DNA alphabet

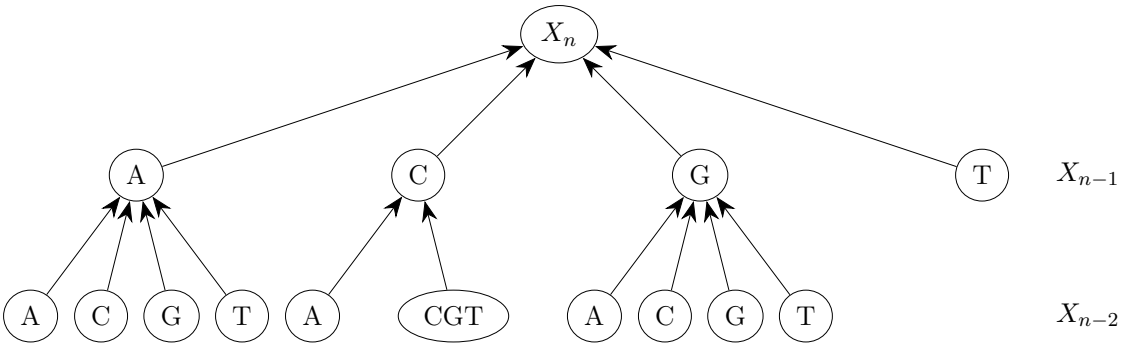

Figure 4.2: A variable length Markov chain of order 2 in DNA alphabet

values of X_{n-2} lead to the same probability distribution for X_n. Secondly, nodes sharing a second to last symbol can result in the same probability distribution for X_n. In Figure 4.2, histories CC, CG and CT are lumped together.

Also, for a VLMC model there exists an SMC representation which is a partition of the sample paths of the full tree of order m. For the VLMC model in Figure 4.2, there would be 11 clusters corresponding with the leaves of the pruned tree.

It should be also noted that the VLMC model's dependence structure

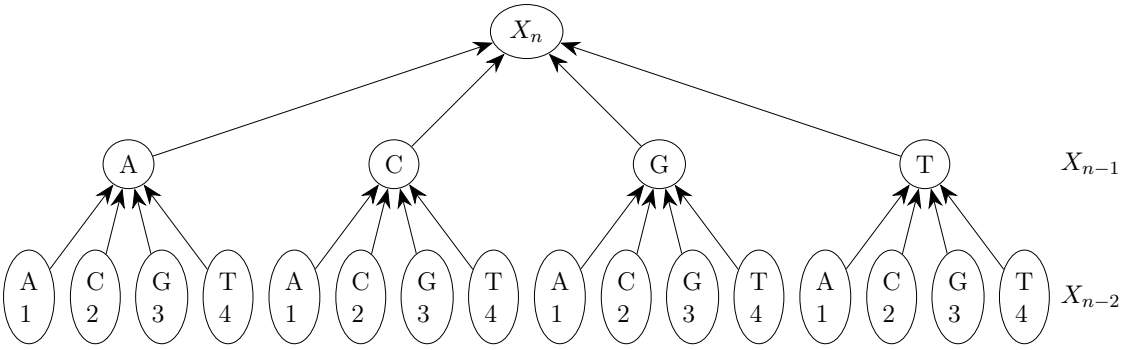

Figure 4.3: A sparse Markov chain of order 2 in DNA alphabet with 4 clusters

in the space of sample paths corresponds to the theory of context-specific DAGs, namely labeled DAGs, as presented by Pensar et al. (2014).

For an SMC model, it is possible that there does not exist a corresponding VLMC model. This is illustrated in Figures 4.3 and 4.4. Numbers inside the nodes denote to which cluster the sample paths belong to.

In Figure 4.3, we have the previously mentioned example, where X_{n-1} is irrelevant for predicting X_n while X_{n-2} is relevant. There are four clusters in the partition of the sample paths of the full three. In Figure 4.4, we have three clusters and no clear hierarchical structure as generally nodes in one cluster do not share a path to the root node X_n. In learning a Markovian model, we can estimate the partition of a SMC model. This enables the learning of MC and VLMC as well as SMC models.

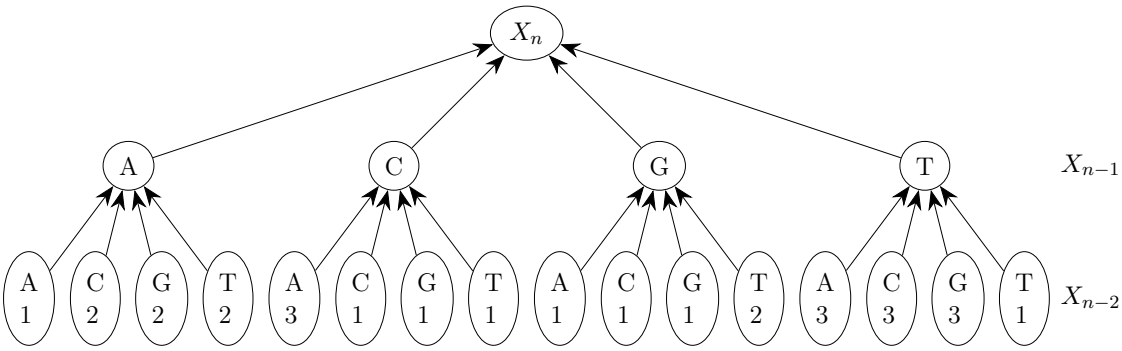

Figure 4.4: A sparse Markov chain of order 2 in DNA alphabet with 3 clusters

5

Bayesian Clustering

Here, we describe a Bayesian approach to clustering. In Articles I, II and III, Bayesian clustering is used for learning sequence models from data. However, there are certain differences between the considered problems. For example, in Articles I and III states of Markovian models are clustered while in Article II the objects to be clustered are sequences modeled with Markov chains of fixed order. We aim to present the general ideas of Bayesian clustering as well as to elaborate on some important details.

In Article II, we define a partition S of a set S_0 as "a collection of disjoint, non-empty subsets of S_0, whose union is S_0". The elements of the partition are called clusters. Clustering means finding a partition following some criterion. Typically, this is similarity. Then there should be a high probability that similar objects belong to the same cluster. Clustering is an example of a task in unsupervised machine learning (Bishop, 2006). The procedure is *unsupervised* because in principle no information besides the set of data objects is used.

5.1 Prior and Likelihood

In Bayesian clustering, the partition S is the variable of main interest. A central idea is that the partition S with high posterior probability $p(S|\boldsymbol{x})$ should be close to an optimal partition. This can be justified with results on predictive inference and classification theory (Corander et al., 2007, 2013;

Hartigan, 1990; Barry and Hartigan, 1992). Here, $\boldsymbol{x}$ denotes a set of observations with n objects to be clustered. To calculate the posterior distribution $p(S|\boldsymbol{x})$, a prior $p(S)$ and a marginal likelihood $p(\boldsymbol{x}|S)$ are needed. Then the Bayes' rule (2.1) yields

$$p(S|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|S)p(S)}{\sum_{S \in \mathcal{S}} p(\boldsymbol{x}|S)p(S)}, \tag{5.1}$$

where $\mathcal{S}$ is the set of all possible partitions of $\boldsymbol{x}$. In the expression for the marginal likelihood $p(\boldsymbol{x}|S)$, parameters of the model have been integrated over their prior distributions. A reasonable approach to identifying a good partition is to solve the maximum *a posteriori* (MAP) estimate of the partition parameter S. The MAP estimate is defined as

$$\begin{aligned} \hat{S} &= \operatorname*{argmax}_{S \in \mathcal{S}} p(S|\boldsymbol{x}) \\ &= \operatorname*{argmax}_{S \in \mathcal{S}} p(\boldsymbol{x}|S)p(S). \end{aligned} \tag{5.2}$$

Often, the MAP estimate can be solved only approximately. A simple solution to finding the MAP estimate would be to evaluate $p(\boldsymbol{x}|S)p(S)$ for all possible values of S. However, this is computationally infeasible with almost any realistic data set. The number of possible partitions for n objects is the Bell number $B(n)$ and it increases rapidly as a function of n (Bell, 1934; Rota, 1964). There are various methods that a stochastic algorithm could employ to maximize $p(\boldsymbol{x}|S)p(S)$. Markov chain Monte Carlo (MCMC) and similar stochastic simulation methods give consistent MAP estimates but they can be too slow in some cases when the number of data objects to be clustered is large. In Articles I and II, a stochastic greedy algorithm is used. Search operators like joining two clusters together and splitting one cluster into two are used in a data-driven manner. In Article III, a deterministic recursive learning algorithm is used. In later sections, algorithms for learning the MAP partition are presented.

The prior $p(S)$ can take many forms. The simplest form is the uniform prior $p(S) = 1/B(n)$. Then each partition is equally probable and it is

enough to maximize $p(\boldsymbol{x}|S)$ when calculating the MAP estimate. Another possibility is the Dirichlet process prior (DPP)

$$p(S) \propto \prod_{S_c} \eta_0 \Gamma(|S_c|), \tag{5.3}$$

where η_0 is a hyperparameter and $|S_c|$ is the number of items in cluster S_c. Here, $\Gamma(\cdot)$ denotes the Gamma function. With the Dirichlet process prior, sizes of individual clusters affect the prior probability of the partition so that partitions with larger clusters are favoured. This is desirable if we have *a priori* information that the number of clusters should be small relative to n, the number of items to be clustered.

The following description of the Dirichlet process prior is adapted from Article II. To derive the Dirichlet process prior (5.3) we can consider the Dirichlet process mixture (DPM) model under certain assumptions (Neal, 2000; Jain and Neal, 2004; Teh et al., 2006; Dahl, 2009). By representing the Dirichlet process (Ferguson, 1973) with the Pólya urn scheme (Blackwell and MacQueen, 1973), the prior can derived (see, e.g. Dahl, 2009). The stick-breaking construction of the Dirichlet process (Sethuraman, 1994) provides an alternative approach. Then we have the following form of the DPM model: (see, e.g. Teh et al., 2006)

$$\begin{aligned} \pi \,|\, \eta_0 &\sim \mathrm{GEM}(\eta_0) \\ z_i \,|\, \pi &\sim \pi \\ \varphi_k \,|\, G_0 &\sim G_0 \\ y_i \,|\, z_i, (\varphi_k)_{k=1}^{\infty} &\sim F(\varphi_{z_i}), \end{aligned} \tag{5.4}$$

where given the cluster membership indicators z_i and cluster parameters $\theta_i = \varphi_{z_i}$ we have conditional independence of observations y_i. The prior distribution for cluster parameters is G_0. We draw indicator variables z_i independently from the stick-breaking distribution π which is sometimes denoted by $\mathrm{GEM}(\eta_0)$. GEM stands for Griffiths, Engen, and McCloskey

(see, e.g. Pitman, 2006). The stick-breaking construction of the Dirichlet process $\mathrm{DP}(\eta_0, G_0)$ induces the random probability measure on the positive integers $\pi = (\pi_k)_{k=1}^{\infty}$. We have $G = \sum_{k=1}^{\infty} \pi_k \delta_{\varphi_k}$ in the construction. The distribution of G follows $\mathrm{DP}(\eta_0, G_0)$ (Sethuraman, 1994). By assuming n observations from the DPM model above, we get a finite number of clusters that have more than zero observations. We have a multinomial distribution with event probabilities defined by π and the cluster sizes follow this distribution. In the stick-breaking construction, there is a beta distributed variable β_k associated with each π_k (see, e.g. Teh et al., 2006). We can use the Bayes' theorem to get a posterior distribution for the cluster sizes and the unknown $(\beta_k)_{k=1}^{\infty}$. By integrating out each β_k over its beta prior distribution we get the Dirichlet process prior as given in (5.3).

Typically, in Bayesian clustering it is assumed that data objects in different clusters are conditionally independent given the partition. Then, we have a product partition model (Hartigan, 1990; Barry and Hartigan, 1992). In these models, the likelihood is expressed as a product

$$p(\boldsymbol{x}|S) = \prod_{S_c \in S} f(\boldsymbol{x}_{S_c}), \tag{5.5}$$

where $f(\boldsymbol{x}_{S_c})$ is the marginal likelihood contribution from cluster S_c, which can take a variety of forms. Here, the likelihood contribution from a cluster S_c is defined as

$$f(\boldsymbol{x}_{S_c}) = \int_{\Theta} p(\theta) p(\boldsymbol{x}_{S_c}|\theta) d\theta, \tag{5.6}$$

where θ denotes collectively the quantitative parameters of the model.

Next, we define the marginal likelihood $p(\boldsymbol{x}|S)$ for Markovian models excluding MTD. The marginal likelihood for SMC models is a basis for the learning algorithms. Because a VLMC model has a representation as an SMC model, the same definition for the marginal likelihood can be used when learning VLMC and SMC models. Also, the marginal likelihood for a full MC can be calculated with the same formulations.

For an MC of given order m, we have data on transitions from $|\mathcal{X}^m|$ states to $J = |\mathcal{X}|$ symbols. For an SMC model with a partition S and k clusters, there are $\{p_c|\cdot : c = 1, \ldots, k\}$ probability vectors. For a full MC, $k = |\mathcal{X}^m|$. We denote by $\theta \in \Theta$ the quantitative parameters of the model. When assuming the initial state fixed, the likelihood is a product of multinomial distributions

$$p(\boldsymbol{x}|\theta, S) \propto \prod_{i=1}^{|\mathcal{X}^m|} \prod_{j=1}^{J} p_{i|j}^{n_{i|j}} = \prod_{c=1}^{k} \prod_{j=1}^{J} p_{c|j}^{\sum_{i \in s_c} n_{i|j}}, \tag{5.7}$$

where the number of transitions from the state i to j in $\boldsymbol{x}$ is denoted by $n_{i|j}$. For transition probabilities $p_{c|j}, c = 1, \ldots, k, j = 1, \ldots, J$, we choose the canonical conjugate multivariate Dirichlet prior (Koski, 2001)

$$p(\theta|\alpha, q) = \prod_{c=1}^{k} \left[\frac{\Gamma(\alpha)}{\prod_{j=1}^{J} \Gamma(\alpha q_j)} \prod_{j=1}^{J} p_{c|j}^{\alpha q_j - 1} \right], \tag{5.8}$$

where the hyperparameters satisfy the following conditions: $\alpha > 0, q_j > 0$, $\sum_{j=1}^{J} q_j = 1$. Using the properties of Dirichlet distribution, the marginal likelihood $p(\boldsymbol{x}|S)$ can be calculated as

$$\begin{aligned} p(\boldsymbol{x}|S) &\propto \int_{\theta \in \Theta} p(\boldsymbol{x}|\theta, S) p(\theta|\alpha, q) d\theta \\ &\propto \int_{\theta \in \Theta} \left[\prod_{c=1}^{k} \frac{\Gamma(\alpha)}{\prod_{j=1}^{J} \Gamma(\alpha q_j)} \prod_{j=1}^{J} p_{c|j}^{\alpha q_j - 1} \prod_{j=1}^{J} p_{c|j}^{\sum_{i \in s_c} n_{i|j}} \right] d\theta \\ &\propto \prod_{c=1}^{k} \frac{\Gamma(\alpha)}{\prod_{j=1}^{J} \Gamma(\alpha q_j)} \frac{\prod_{j=1}^{J} \Gamma(\sum_{i \in s_c} n_{i|j} + \alpha q_j)}{\Gamma((\sum_{j=1}^{J} \sum_{i \in s_c} n_{i|j}) + \alpha)}. \end{aligned} \tag{5.9}$$

Also, predictive probability of future observations can be calculated analytically, as is demonstrated in Article I.

One important issue is choosing m, the order of the Markov chain used in estimation. In Article I, a uniform distribution is assigned over the values $m = 0, \ldots, M$, where M is an upper bound that has to be chosen and can be revised if necessary. Then the joint posterior distribution of m and S is

$$p(S, m \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid S)p(S)p(m) \tag{5.10}$$

and $p(m) = 1/(M+1)$. The approximate MAP estimate can be calculated for both S and m:

$$\left(\hat{S}, \hat{m}\right) = \underset{m \in \{0,\ldots,M\}}{\operatorname{argmax}} \left\{ \underset{S_m \in \mathcal{S}_m}{\operatorname{argmax}}\, p(\boldsymbol{x} \mid S_m)p(S_m) \right\}. \tag{5.11}$$

5.2 Expectation-Maximization Algorithm

Here, we discuss the expectation-maximization (EM) algorithm. Its first general formulation is due to Dempster et al. (1977). In Article II the EM-algorithm is used for learning a partition when the data objects are Markov chains of fixed order. In the EM-algorithm, we have unknown and latent variables. For example, a latent variable can be the partition S while the unknown variables can be the transition probability matrices collectively denoted by θ. Assume that x represents the observed data, z is a latent variable and θ is an unknown parameter. Then consider a posterior $p(\theta|x) = \sum_z p(\theta, z|x)$. Here, the description of the algorithm is adapted from Article II. The EM-algorithm for finding a MAP estimate of the posterior is then defined as follows. Starting with an initial value $\theta^{(0)}$ for θ and setting $k \leftarrow 0$, the following steps are applied until convergence:

- Expectation step

  Calculate

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E_{\theta^{(k)},x}\left[\ln p(\theta, Z|x)\right] \\ &= \sum_z \ln p(\theta, z|x) \cdot p(z|\theta^{(k)}, x) \end{aligned}$$

- Maximization step

  Set

$$\theta^{(k+1)} \leftarrow \arg\max_{\theta} Q(\theta|\theta^{(k)})$$

  and $k \leftarrow k+1$.

Convergence is achieved when the difference between $\theta^{(k+1)}$ and $\theta^{(k)}$ is below a threshold that has been set beforehand. Also, it can be useful to define an upper bound for the number of iterations that the algorithm is allowed to run.

In Article II, the EM-algorithm is used for estimating transition probability matrices of sequences modeled with Markov chains and for assigning sequences to the most appropriate clusters. The number of clusters is not changed by the EM-algorithm.

For iterations k and $k+1$ it holds that

$$p(\theta^{(k+1)}|x) \geqslant p(\theta^{(k)}|x).$$

Thus, the definition of the EM-algorithm leads to monotonic probabilities for the unknown parameter θ given the data x. This property of the algorithm makes the convergence possible. Here, we have marginalized out the latent variable z.

5.3 Classification EM Algorithm

In Article IV, we have M generating clusters (mechanisms) and N observations. The aim is then to estimate the prior probability that an observation is generated by cluster j, denoted by α_j for $j = 1, \ldots, M$. Also, we estimate parameter vectors for mechanisms $j = 1, \ldots, M$, denoted by $\underline{\psi}_j$. Finally, assignment of each observation to a cluster, denoted by $u^{(l)}$ for $l = 1, \ldots, N$ and $j = 1, \ldots, M$, has to be estimated. For an alternative representation, indicator variables of the form ${u_j}^{(l)}$ can be used. In article IV, a version of classification EM algorithm (CEM) (Celeux and Govaert, 1992; Redner and Walker, 1984) is used. The following description of the classification EM algorithm is adapted from Article IV:

- Initial step

  Choose initial values for ${\alpha_j}^{(0)}$ and ${\underline{\psi}_j}^{(0)}, j = 1, \ldots, M$. Move to E-step.

- First step

  ${\alpha_j}^{(0)}$ and ${\underline{\psi}_j}^{(0)}, j = 1, \ldots, M$ are our current parameters and ${u_j}^{(l)}, l = 1, \ldots, N$ are our current assignments.

- E-step

  Compute for $l = 1, \ldots, N$ and $j = 1, \ldots, M$ using the Bayes' rule

$$t_j(y^{(l)}) = \frac{\alpha_j^{(0)} p(y^{(l)} | \underline{\psi}_j^{(0)}, j, x^{(l)})}{\sum_{j=1}^{M} \alpha_j^{(0)} p(y^{(l)} | \underline{\psi}_j^{(0)}, j, x^{(l)})}.$$

  Here, $t_j(y^{(l)})$ is the current posterior probability that $(y^{(l)}, x^{(l)})$ belongs to the cluster j.

- C-step

  For $l = 1, \ldots, N$ we assign

$$\hat{u}_{j*}^{(l)} = \begin{cases} 1 & j* = \underset{1 \leqslant j \leqslant M}{\operatorname{argmax}}\, t_j(y^{(l)}) \\ 0 & \text{otherwise} \end{cases}$$

  to get a new assignment of $(y^{(l)}, x^{(l)})$. Thus, each observation is assigned to a cluster so that the posterior probability is maximized.

- M-step

$$\alpha_j^{(1)} = \frac{\sum_{l=1}^{N} u_{j*}^{(l)}}{N}$$

is maximum likelihood estimate of α_j from $L_2(\underline{\alpha}) = \sum_{j=1}^{M} n_j \log(\alpha_j)$. We need to find $\psi_j^{(1)}$ for $j = 1, \ldots, M$ by maximization of $L_1(\underline{\psi}) = \sum_{j=1}^{M} L_j(\underline{\psi}_j)$. Details of this maximum likelihood estimation are given in Article IV.

- Return to

  Now, we have the new estimates

  ${\alpha_j}^{(1)}, \underline{\psi}_j^{(1)}, u_{j*}^{(l)} \quad j = 1, \ldots, M; l = 1, \ldots, N.$

  We assign ${\alpha_j}^{(1)} \rightarrow {\alpha_j}^{(0)}, \underline{\psi}_j^{(1)} \rightarrow \underline{\psi}_j^{(0)}, u_{j*}^{(l)} \rightarrow u_j^{(l)}$.

- Stop

  We have $\ln(L) = L_1(\underline{\psi}) + L_2(\underline{\alpha})$. The algorithm is stopped when

$$|\ln L(\underline{\psi}^{(1)}, \underline{\alpha}^{(1)}) - \ln L(\underline{\psi}^{(0)}, \underline{\alpha}^{(0)})| \leqslant \varepsilon$$

  where ε is a small positive number that has to be specified in implementation. Also, if a preassigned maximum number of iterations has been reached, the algorithm stops.

5.4 Stochastic Search

Here, we present a stochastic search algorithm adapted from Article I. This greedy algorithm is based on algorithms presented in Corander and Marttinen (2006) and Marttinen et al. (2006). The idea is to find a good clustering of the $|\mathcal{X}|^m$ states of the Markov chain and thus estimate the Sparse Markov chain model. For a given value of m we have the following algorithm:

(i) Initialize $S_t, t = 0$ with $|\mathcal{X}|^m$ singleton clusters and store for all pairs of states $i, l \in \mathcal{X}^m$ the distances between posterior mean estimates of their transition probability vectors

$$d_{i,l} = \sum_{j=1}^{J} \left(\frac{n_{i|j} + \alpha q_j}{\sum_{j=1}^{J} n_{i|j} + \alpha q_j} - \frac{n_{l|j} + \alpha q_j}{\sum_{j=1}^{J} n_{l|j} + \alpha q_j} \right)^2. \tag{5.12}$$

(ii) Given the current value of $p(\boldsymbol{x}|S_t)$, apply the following operators sequentially until no change in S_t results in a higher marginal likelihood.

(iii) In a random order, move each state $i \in \mathcal{X}^m$ to the class c in S_t, which results in the S_{t+1} associated with a maximal increase in $p(\boldsymbol{x}|S_{t+1})$. If $p(\boldsymbol{x}|S_{t+1}) \leqslant p(\boldsymbol{x}|S_t)$ for all $c = 1, \ldots, k$, $S_{t+1} = S_t$.

(iv) For each pair of classes $c, c' = 1, \ldots, k$, calculate $p(\boldsymbol{x}|S^*)$ for the S^* obtained by merging classes c, c' in S_t. If any S^* satisfies $p(\boldsymbol{x}|S^*) - p(\boldsymbol{x}|S_t) > 0$, set S_{t+1} equal to the S^* for which $p(\boldsymbol{x}|S^*) - p(\boldsymbol{x}|S_t)$ is maximal, otherwise set $S_{t+1} = S_t$.

(v) For each class $c = 1, \ldots, k$, use the complete linkage algorithm (e.g Mardia et al., 1979) with distances (5.12) to split the class into two non-empty subsets of states and calculate $p(\boldsymbol{x}|S^*)$ for the resulting partition S^*. If $p(\boldsymbol{x}|S^*) - p(\boldsymbol{x}|S_t) > 0$, set S_{t+1} equal to S^*, otherwise set $S_{t+1} = S_t$.

This greedy algorithm converges to a local mode when the operators do not increase marginal likelihood any further. Several restarts from different initial conditions can be used to find clusterings that are closer to a global optimum. Also, this algorithm could be generalized to give a consistent posterior estimator using Markov chain Monte Carlo approach with a non-reversible Markov chain (Marttinen et al., 2006; Corander et al., 2006, 2008).

5.5 Recursive Search

In Article III, a heuristic deterministic algorithm is defined for searching the optimal SMC model $(S, \mathcal{P})$ for a given sequence $\{X_t\}_{t=1}^n$. Here, we present an

adapted description of the algorithm. The general idea is to apply Delaunay triangulation on $\mathcal{X}^m$ so that each transition state $i \in \mathcal{X}^m$ becomes a node in the triangulation graph. Then we recursively merge nodes to maximize the posterior probability. Bayes factor is used as the local criterion when choosing which nodes should be merged.

- Initial step

  Obtain the transition counts of $\{X_t\}_{t=1}^n$ for MC of order m. Estimate the transition probability distribution θ of the model by posterior mean estimation from the transition counts. Form Delaunay triangulation G of $\mathcal{X}^m$ by using values of free parameters in θ as coordinates. Calculate the log Bayes factor $\log BF_{uv}$ for each edge u, v in G. Find the edge (u^*, v^*) that gives maximal log Bayes factor value w. Set $\mathcal{U} = u^*$,$\mathcal{V} = v^*$ and $\mathcal{W} = w$.

- While $\mathcal{W} > 0$ do

  Merge $\mathcal{V}$ to $\mathcal{U}$ by the following steps:

  a) add the sufficient statistics counts of $\mathcal{V}$ to $\mathcal{U}$

  b) for each node r in G which has a connection with $\mathcal{V}$, if edge $(\mathcal{U}, r)$ does not exist, redirect the edge $(\mathcal{V}, r)$ to $(\mathcal{U}, r)$

  c) delete $\mathcal{V}$ from G. Update the Bayes factors for all the edges (include the edges added by merging) connected to $\mathcal{U}$. Find the edge $(u^{*'}, v^{*'})$ with a maximal log Bayes factor value w'. Set $\mathcal{U} = u^{*'}$, $\mathcal{V} = v^{*'}$ and $\mathcal{W} = w'$.

6

Causal Inference

Causal inference is a challenging issue for both scientific inquiry and philosophy of science (Rothman et al., 2008). In this chapter, we present two models of causality. The first one is Rothman's pie model, also called the sufficient-component cause model (Rothman, 1976; Rothman et al., 2008; Rothman, 2012). The second one is Pearl's Do-Calculus which gives a probabilistic account of causality (Pearl, 1995, 2000). In Article IV, a model that is a stochastic extension of the sufficient-component cause model is presented. In the article, we have a fixed number of causal mechanisms. The effect of the covariates on the probability distribution of the outcome variable depends on the mechanism-specific parameters. These are estimated from the data so that mechanisms can be identified. We can consider mechanisms analogous to sufficient causes in the sufficient-component cause model. When simulating data from the model, we first draw randomly a sufficient cause and then based on that we draw the value of the outcome variable. In this model, the individual component causes are modeled with parameters of the mechanisms.

There is a connection between the model in Article IV and Pearl's Do-Calculus. Also, Do-Calculus is described here as an example of a probabilistic framework for causality. Because the model in Article IV is essentially a Bayesian network, we can apply do-conditioning. The covariates do not have parents in the graph so forcing a subset of covariates to have certain value leads to a equivalent conditional probability distribution as see-conditioning, i.e. observing those values.

Traditionally, statistical inference has been concerned with associations and correlations instead of causality. In scientific inquiry, causal relations are important and often results of statistical methods are used for causal inference, with mixed success (Freedman, 1999). However, in statistics there are several important and relatively successful approaches to causal inference. Firstly, there is contrafactual causality (see, e.g. Rubin, 1974). For example, in a typical medical study each subject either gets the treatment or does not. We are then interested in the causal effect, i.e. the difference in outcome between a situation where the subject gets the treatment and the alternative situation where the subject does not get the treatment. Here, contrafactuality comes from the fact that the subject cannot both get the treatment and not get it. Randomization makes it possible to asses the causal effect when there are several trials. Rubin also discusses how an observational, non-randomized study can provide information on the causal effect. The contrafactual or *potential outcomes* approach has been extended for example with the use of structural equations and instrumental variables (Angrist et al., 1996). Another approach is the use of graphical models for causal inference (see, e.g. Pearl, 2000). Often, relations between variables are described with directed acyclic graphs (DAGs) and parents of a node are considered to be its direct causes. There has been symbiotic development between the contrafactual and graphical models approaches (Greenland and Brumback, 2002; Pearl, 2009). Finally, there is the predictive causality approach which takes explicitly into account the time between the cause and the effect (see, e.g. Arjas and Eerola, 1993; Arjas and Parner, 2004).

6.1 Sufficient-Component Cause Model

In epidemiology, causation is often modeled with the sufficient-component cause model. Here, we state the basic principles of the model (see e.g. Rothman et al., 2008). We define *cause* as a condition or event that precedes the disease and had the cause not been present the disease would not have occurred. By *sufficient cause* we mean a set of component causes that is minimal and complete and which is sufficient for the disease to appear.

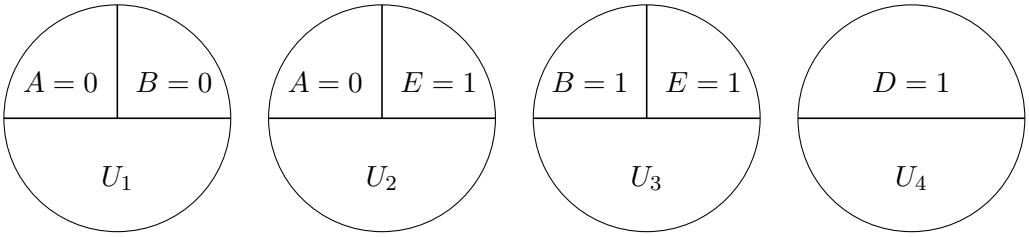

Figure 6.1: Four sufficient causes. Adapted from Rothman et al. (2008).

Minimality means that if one of the component causes is not present than the disease does not appear. Completeness means that if all the component causes are present than the disease occurs. There are typically several different sufficient causes for a disease. If a component cause is present in all sufficient causes, we can consider it to be *necessary*.

These principles are illustrated in Figure 6.1. There we have component causes A, E, B and D. These are assumed to be binary so that the condition of the component cause either is present (value 1) or not (value 0). Typically, in epidemiology there are always component causes that have not been identified. This is taken into consideration by having an unknown cause U_i as a part of each sufficient cause. We notice from the figure that the value $B = 0$ is part of the first sufficient cause while the value $B = 1$ is part of the third sufficient cause. This could be the situation for example if presence of some chemical substance would in one case be causing a disease and in another case preventing it. This depends on the other component causes of the sufficient cause. They are called collectively *causal complement*. Thus, the effect of chemical substance being present or not depends on its causal complement.

One insight that the sufficient-component cause framework gives is that the proportion of disease due to specific causes can add to over 100%. This is because there are multiple sufficient causes that have at least partially different component causes.

The sufficient-component cause model is deterministic but the methodology of epidemiology includes a variety of mathematical tools (Rothman

et al., 2008; Rothman, 2012). For example, the risk of contracting a disease for an individual can be modeled with probabilistic methods. This modeling can benefit from the sufficient-component cause model when different causes are assessed systematically. Also, a stochastic generalization of the sufficient-component cause model can be useful for epidemiological study.

It seems that for causal inference, a logical, qualitative foundation is needed. The sufficient-component cause model provides this. Then causal inference can proceed with either a model that includes causality or with more heuristic methods, perhaps comparing information from many sources or experiments. Article IV is an example of work that develops a probabilistic model of causality.

6.2 Do-Calculus

A notable framework for probabilistic causal inference is Pearl's Do-Calculus (Pearl, 1995, 2000). We present basic definitions and some properties of Do-Calculus following Koski and Noble (2009) as well as lecture notes by Koski and Noble and lecture slides from Koski's presentation from the Bayesian network course at KTH, year 2013.

Firstly, we define a Bayesian network as the following structure.

Definition 6.1. Bayesian Network. Let $\mathcal{G}$ be a directed acyclic graph (DAG) with nodes $V = \{1, \ldots, n\}$, $(X_v)_{v \in V}$ a set of finite discrete random variables and $\{P(X_v = x_{i_v} | X_{\pi(v)} = x_{\pi(v)})\}_{v \in V}$ a set of conditional probability distributions with

$$P(X_1 = x_{i_1}, \ldots, X_n = x_{i_n}) = \prod_{v=1}^{n} P(X_v = x_{i_v} | X_{\pi(v)} = x_{\pi(v)}),$$

where $\pi(v)$ is the set of parent nodes of v and $P(X_v | X_{\pi(v)}) = P(X_v)$ if $\pi(v) = \varnothing$.

We consider the parents $X_{\pi(v)}$ as the direct causes of X_v and $P(X_v | X_{\pi(v)})$ measures our belief about the strength of the causality. Often, in statistical

inference we observe a random variable having some value. In controlled experiments, a variable is forced to have some value (maybe after randomization). A central part of Do-Calculus is the intervention formula. It describes the joint probability distribution of the Bayesian network after a subset of variables has been forced some values. Observing a value is called see-conditioning and forcing a value is called do-conditioning.

When variables X_A, $A \subseteq V$ are forced to have values ${x_A}^*$ the resulting joint distribution is defined as

$$\begin{aligned} P(X_V = x_V | X_A \leftarrow {x_A}^*) &\doteq \frac{P(X_V = x_V)}{\prod_{v \in A} P(X_v = x_{i_v} | X_{\pi(v)} = x_{\pi(v)})} | x_A = {x_A}^* \\ &= \prod_{v \in V \setminus A} P(X_v = x_{i_v} | X_{\pi(v)} = x_{\pi(v)}) | x_A = {x_A}^*. \end{aligned}$$

In the notation, the arrow $\leftarrow$ and the double bar $||$ indicate conditioning by doing. In other words, $X \leftarrow x$ means that we intervene to force the value of the random variable X to be x. In the intervention formula, conditioning by $x_A = {x_A}^*$ is equivalent to a substitution. Applying the intervention formula means that edges from parents of nodes in A to nodes in A are removed. This 'local surgery' yields a mutilated graph and the joint distribution factorizes along it. Note that we can also define an alternative formulation $P(X_v = x_v | X_A \leftarrow {x_A}^*) \doteq P(X_{V \setminus A} = x_{V \setminus A} || X_A = {x_A}^*)$ as variables in X_A have known values after the intervention.

A property of Do-Calculus is that if node v has no parents then

$$P(X_V = x_V | X_v \leftarrow {x_v}^*) = P(X_V = x_V | X_v = {x_v}^*).$$

In this case, forcing a value results in same conditional probabilities as observing it. In other words, see and do probabilities are the same. Another property is the exogeneity:

$$P(X_v | X_{\pi(v)} \leftarrow x) = P(X_v | X_{\pi(v)} = x).$$

It means that forcing values to parents of a node v results in same conditional probability for v as observing those values in the parent variables $X_{\pi(v)}$. Finally, we have the invariance property. For all $v = 1, \ldots, n$ and $S \subseteq V$ such that $S \cap \{v, \pi(v)\} = \varnothing$, we have

$$P(X_v | X_{\pi(v)} \leftarrow x, X_S \leftarrow s) = P(X_v | X_{\pi(v)} \leftarrow x).$$

This means that once we control parents of v, no further interventions will affect the probability of X_v.

Here, we give an example of how an intervention affects a Bayesian network. The example is adapted from Pearl (2000) and Koski's lecture slides from the Bayesian network course at KTH, year 2013. We are modeling slipperiness of a pavement and we assume that the slipperiness is a binary random variable denoted by X_5. So the pavement either is slippery or not. Slipperiness is affected by several environmental attributes which we model as random variables. These include the wetness of the pavement (X_4), sprinkler being possibly on (X_3), the possible rain (X_2), and the season (X_1) which is assumed to have four possible values. All the other random variables in this example are binary. Relations between these variables are described by the DAG of the Bayesian network. For example, given information about the wetness of the pavement, the slipperiness is independent of the random variables associated with the season, the rain and the sprinkler. The joint distribution of the random variables factorizes along the DAG:

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5) \\ &= P(X_1 = x_1) P(X_2 = x_2 | X_1 = x_1) P(X_3 = x_3 | X_1 = x_1) \cdot \\ &\quad P(X_4 = x_4 | X_2 = x_2, X_3 = x_3) P(X_5 = x_5 | X_4 = x_4). \end{aligned}$$

The situation is illustrated in Figure 6.2.

Now we consider an intervention where the sprinkler is set on, i.e. we set X_3 equal to 1. We apply the intervention formula to get the do-conditioned probability:

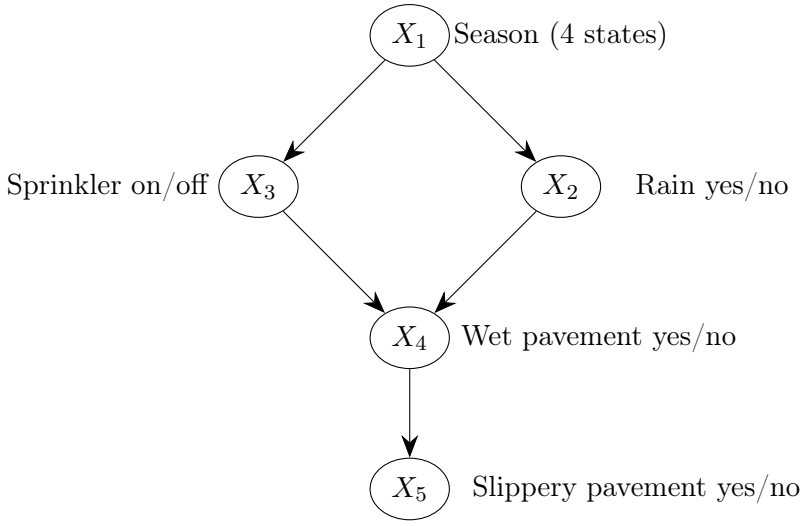

Figure 6.2: Wet pavement DAG before intervention. Adapted from Pearl (2000) and Koski's lecture slides.

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5 | X_3 \leftarrow 1) \\
=&\frac{P(X_1 = x_1, X_2 = x_2, X_3 = 1, X_4 = x_4, X_5 = x_5)}{P(X_3 = 1 | X_1 = x_1)} \\
=&P(X_1 = x_1)P(X_2 = x_2 | X_1 = x_1)P(X_4 = x_4 | X_2 = x_2, X_3 = 1)P(X_5 = x_5 | X_4 = x_4) \quad .
\end{aligned}
$$

The DAG in Figure 6.3 illustrates the situation. The edge from X_1 to X_3 has been removed. This is because after the intervention, i.e. setting the sprinkler on, the season has no effect on the sprinkler.

There are connections between Rothman's sufficient-component cause model and Pearl's Do-Calculus. As a part of probabilistic assessment of causal relations in the sufficient-component cause model, Do-Calculus could be used. We could describe a sufficient cause, i.e. a pie in the model, with a Bayesian network. Component causes would be parental nodes for the

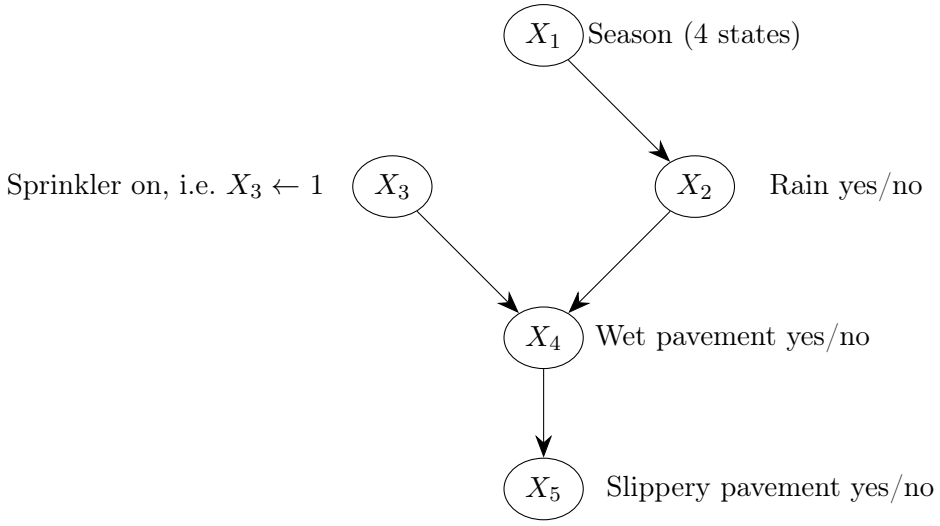

Figure 6.3: Wet pavement DAG after the sprinkler has been set on. Adapted from Pearl (2000) and Koski's lecture slides.

outcome variable, i.e. state of an individual. Then do-conditioning would correspond to having an experiment where values of putative component causes are controlled. More complicated modeling would be needed to consider several sufficient causes. Also, component causes themselves could have causes behind them as represented by parental nodes. Generally, the idea of intervention in Do-Calculus is potentially useful for inference that relies on controlling values of some variables, i.e. conducting controlled experiments.

7

Discussion

In this thesis, we have presented methods for sequence analysis and causal inference. In Article I, model class of Sparse Markov chains is defined and its properties are investigated. In Article II, a stochastic partition model of DNA sequences is considered so that the sequences are modeled as Markov chains of fixed order. Article III develops inference for SMC models further by introducing a recursive deterministic algorithm that uses Delaunay triangulation and Bayes factors. Finally, in Article IV causal inference is addressed and an algorithm for estimating interactions of multiple causes for a disease is given.

One underlying theme in this thesis is the use of the Bayesian approach for problems in machine learning. Especially, clustering is a task that we have addressed several times. A Bayesian approach to clustering is used in Articles I, II and III. Developing further the framework of Bayesian clustering could lead to improvements in the particular solutions we have presented in the articles. One possible direction is to use different prior distributions for partitions as now we have concentrated on the uniform prior and the Dirichlet process prior. For example, an uniform prior on the number of clusters could be used (Kohonen and Corander, 2014; Knorr-Held and Raßer, 2000; Quintana and Iglesias, 2003).

A second theme is the adaptation to the reality of large data sets. Our choice of learning algorithms reflects this. In Article I, a greedy stochastic algorithm is used instead of an MCMC approach that could provide consistent estimates (Marttinen et al., 2006; Corander et al., 2006, 2008). Even

the stochastic greedy algorithm proves to be relatively slow with very large data sets. In Article III, a heuristic deterministic recursive algorithm for the SMC models is presented. In future research, further approximations could be developed to deal with even larger data sets. In the articles, the use of point-estimation to obtain for example a MAP estimate has been a favoured method because estimating the posterior distribution would be computationally challenging. But for some future applications, estimating the posterior distribution with MCMC techniques could be valuable.

The data we have considered has been mostly sequence data. The use of Markovian models for sequence data is already a prominent tool for the scientist studying a variety of phenomena. As the data sets are growing, the need for sparse models will likely increase. The SMC model presented in Article I is a promising foundation for algorithms that process sequence data in a sparse manner. Article III provides a faster algorithm for learning an SMC model. Together these could be used in a variety of applications. For example, they could be combined with the clustering of Markov chains as presented in Article II.

Article IV presents a model for learning causal mechanisms from data. Although the Bayes' rule is used as a part of the learning algorithm, we can say that overall the solution is not Bayesian. This shows a pragmatic attitude to probabilistic modeling. Generally, we use Bayesian methods because they work in practice but we are at the same time interested in non-Bayesian methodology as well. Also, using greedy and heuristic algorithms when facing big data sets can be considered as pragmatic.

There are some attractive areas of application for SMC models that however seem to be computationally too burdensome. For example, processing of natural language with any Markov model is challenging because of the memory required for software implementation. For the analysis of DNA data, there is an abundant literature of methods. Our approaches belong to the category of alignment free methods. For more on aligning DNA sequences, see Cheng et al. (2012).

For all the four articles, there has been a need to implement the algo-

rithms by developing software. For Article I, BAPS[1] software was used as a part of the implementation. Based on Article II, a software package that is freely available was developed [2]. In future, software from Articles III and IV could also be made available for the public.

[1] http://www.helsinki.fi/bsg/software/BAPS/
[2] http://www.helsinki.fi/bsg/software/BACDNAS/

Bibliography

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Elja Arjas and Mervi Eerola. On predictive causality in longitudinal studies. *Journal of Statistical Planning and Inference*, 34(3):361–386, 1993.

Elja Arjas and Jan Parner. Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, 31(2):171–187, 2004.

Daniel Barry and John A Hartigan. Product partition models for change point problems. *The Annals of Statistics*, 20(1):260–279, 1992.

Eric T Bell. Exponential polynomials. *Annals of Mathematics*, 35(2):258–277, 1934.

André Berchtold and Adrian Raftery. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17(3):328–356, 2002.

Jose M Bernardo and Adrian FM Smith. *Bayesian theory*. John Wiley & Sons, 1994.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

Peter Bühlmann and Abraham J Wyner. Variable length Markov chains. *The Annals of Statistics*, 27(2):480–513, 1999.

Gilles Celeux and Gérard Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

Lu Cheng, Alan W Walker, and Jukka Corander. Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Research*, 40(12):5240–5249, 2012.

Jukka Corander and Pekka Marttinen. Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, 15(10): 2833–2843, 2006.

Jukka Corander, Mats Gyllenberg, and Timo Koski. Bayesian model learning based on a parallel MCMC strategy. *Statistics and Computing*, 16(4): 355–362, 2006.

Jukka Corander, Mats Gyllenberg, and Timo Koski. Random partition models and exchangeability for Bayesian identification of population structure. *Bulletin of Mathematical Biology*, 69(3):797–815, 2007.

Jukka Corander, Magnus Ekdahl, and Timo Koski. Parallell interacting MCMC for learning of topologies of graphical models. *Data Mining and Knowledge Discovery*, 17(3):431–456, 2008.

Jukka Corander, Yaqiong Cui, Timo Koski, and Jukka Sirén. Have I seen you before? Principles of Bayesian predictive classification revisited. *Statistics and Computing*, 23(1):59–73, 2013.

David B Dahl. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2):243–264, 2009.

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

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

David Freedman. From association to causation: some remarks on the history of statistics. *Statistical Science*, 14(3):243–258, 1999.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis: second edition.* CRC Press, 2004.

Sander Greenland and Babette Brumback. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31(5): 1030–1037, 2002.

Alan Hájek. "Mises redux"-redux: Fifteen arguments against finite frequentism. *Erkenntnis*, 45(2-3):209–227, 1996.

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

Brian Hayes. First links in the Markov chain. *American Scientist*, 101(2): 92–97, 2013.

Sonia Jain and Radford M Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

Robert E Kass. Statistical inference: the big picture. *Statistical Science*, 26 (1):1–9, 2011.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

John W Kimball. *Biology: 6th edition.* Wm. C. Brown Publishers, 1994.

Leonhard Knorr-Held and Günter Raßer. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.

Jukka Kohonen and Jukka Corander. Computing exact clustering posteriors with subset convolution. *Communications in Statistics: Theory and Methods*, 2014. doi: 10.1080/03610926.2014.894070.

Timo Koski. *Hidden Markov models for bioinformatics*. Springer, 2001.

Timo Koski and John Noble. *Bayesian networks: an introduction*. John Wiley & Sons, 2009.

Nhu D Le, R Douglas Martin, and Adrian E Raftery. Modeling flat stretches, bursts outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91(436):1504–1515, 1996.

Martin Mächler and Peter Bühlmann. Variable length Markov chains: methodology, computing, and software. *Journal of Computational and Graphical Statistics*, 13(2):435–455, 2004.

David JC MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate analysis*. Academic press, 1979.

Pekka Marttinen, Jukka Corander, Petri Törönen, and Liisa Holm. Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, 22(20):2466–2474, 2006.

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

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688, 1995.

Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

Johan Pensar, Henrik Nyman, Timo Koski, and Jukka Corander. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. *Data Mining and Knowledge Discovery*, pages 1–31, 2014. doi: 10.1007/s10618-014-0355-0. URL `http://dx.doi.org/10.1007/s10618-014-0355-0`.

Jim Pitman. *Combinatorial stochastic processes*. Springer, 2006.

Fernando A. Quintana and Pilar L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574, 2003.

Adrian E Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3):528–539, 1985.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

Jorma Rissanen. A universal data compression system. *IEEE Transactions on Information Theory*, 29(5):656–664, 1983.

Gian-Carlo Rota. The number of partitions of a set. *American Mathematical Monthly*, 71(5):498–504, 1964.

Kenneth J Rothman. Causes. *American Journal of Epidemiology*, 104(6): 587–592, 1976.

Kenneth J Rothman. *Epidemiology: an introduction*. Oxford University Press, 2012.

Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology: third edition*. Lippincott Williams & Wilkins, 2008.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688–701, 1974.

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

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.