Master's Thesis

Variational Inference for Dirichlet Process Mixtures and Application to Gaussian Estimation

for obtaining the academic degree

Diplom-Ingenieur

in the masters's degree program

Telecommunications

submitted by

Thomas Lipovec
matriculation number: 01529232

Institute of Telecommunications
Faculty of Electrical Engineering and Information Technology
TU Wien

Supervision:
Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Franz Hlawatsch
Dipl.-Ing. Thomas John Bucco

Statement on Academic Integrity

Hiermit erkläre ich, dass die vorliegende Arbeit gemäß dem Code of Conduct – Regeln zur Sicherung guter wissenschaftlicher Praxis (in der aktuellen Fassung des jeweiligen Mitteilungsblattes der TU Wien), insbesondere ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel, angefertigt wurde. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In– noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Wien, 21. November 2023

Thomas Lipovec

Abstract

Bayesian nonparametric models have emerged as a flexible tool for learning patterns and structures in complex data. A well-known Bayesian nonparametric model is the Dirichlet process mixture (DPM) model. The DPM model extends the Bayesian mixture model with a finite number of mixture components to a Bayesian mixture model with a countably infinite number of mixture components while permitting the use of efficient Bayesian inference methods. Thereby, it becomes possible to cluster data and estimate unknown parameters without specifying the number of clusters and unknown parameters a priori. Practical Bayesian inference methods for DPM models include Markov chain Monte Carlo (MCMC) and variational inference (VI) methods. In this thesis, we focus on the VI methodology and provide a detailed derivation of the coordinate ascent variational inference (CAVI) algorithm for DPM models. Subsequently, we apply the CAVI algorithm to a Gaussian estimation problem that involves noisy observations of object features modeled by a DPM with Gaussian mixture components. We present simulation results that compare the improvement of the estimation accuracy of the object features due to clustering achieved by our CAVI algorithm and by a previously proposed MCMC algorithm. Our simulation results demonstrate to which extent we have to sacrifice estimation accuracy when opting for the less accurate but more efficient CAVI algorithm as opposed to the MCMC algorithm.

Contents

List of Figures

List of Tables

List of Abbreviations

BMM	Bayesian mixture model
CAVI	coordinate ascent variational inference
CG	clustering gain
CRP	Chinese restaurant process
DP	Dirichlet process
DPM	Dirichlet process mixture
EF	exponential family
ELBO	evidence lower bound
KLD	Kullback-Leibler divergence
MAP	maximum a-posteriori
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
MF	mean field
MMSE	minimum mean square error
VI	variational inference
i.i.d.	independent and identically distributed
pdf	probability density function
pmf	probability mass function

1 Introduction

In this chapter, we explain the motivation behind this thesis and provide an overview of related work. Additionally, we outline the contributions and structure of the thesis and define the mathematical notation to be used.

1.1 Motivation and Contribution

Modern statistical analysis methods rely on complex statistical models for which intractable probability distributions are approximated. A prominent example is Bayesian inference [1], which has become a crucial component of probabilistic machine learning [2] and statistical signal processing [3]. The central object of interest in Bayesian inference is the posterior distribution, because it allows us to estimate unknown random parameters of the statistical model from observed data. Unfortunately, it is not possible to calculate the exact posterior distribution for many statistical models (e.g., Bayesian mixture models, state-space models, Bayesian neural networks), and thus we must resort to using approximate inference techniques. This is especially true in the case of high-dimensional statistical models which are suited to large-scale applications.

In this thesis, we consider the Dirichlet process mixture (DPM) model [4], which is a Bayesian nonparametric model, i.e., a statistical model of infinite dimension. More specifically, the DPM model is a Bayesian mixture model (BMM) that consists of a countably infinite number of mixture components, where each component itself is parameterized by a finite or infinite set of parameters. The DPM is obtained by using the Dirichlet process (DP) as the prior distribution of the parameters of the mixture model. The main advantage of the DPM model is that the task of selecting the number of mixture components, which arises in the conventional finite case, is removed. Consequently, the number of clusters in the data (components of the mixture on which the data actually depends) can be jointly inferred along with the cluster assignments and cluster parameters and does not need to be constrained to a prespecified value.

For DPM models, obtaining the posterior distribution involves integrating over an infinite number of parameters, which makes the direct calculation of the posterior distribution intractable. Thus, we use approximate inference methods. The two most widely used approaches to approximate inference are Markov chain Monte Carlo (MCMC) [5] and variational inference (VI) [6]. In the MCMC approach, the posterior distribution is approximately represented by a set of samples, which is generated using a mechanism involving a Markov chain. The desired estimates can then be approximated by empirical estimates constructed from the generated samples. For an increasing number of samples, this approximation converges to the true posterior distribution. On the other hand, the convergence can be slow and difficult to diagnose. The VI methodology is a deterministic, optimization-based approach. Its principle is to approximate

the posterior distribution by a tractable distribution by minimizing a divergence measure between the posterior distribution and the tractable distribution. Compared to MCMC methods, VI methods tend to be faster but also tend to obtain less accurate results since they may suffer from oversimplified approximations of the posterior distribution.

This thesis centers around the VI approach. We present a detailed derivation, based on [7], of a VI method called coordinate ascent variational inference (CAVI) for DPM models with conjugate priors and component distributions from the exponential family. The CAVI algorithm approximates the posterior distribution of a DPM model in an iterative manner, seeking the best-fitting distribution within a predefined family of distributions. Additionally, we specialize the DPM of exponential family distributions to the DPM of Gaussians and develop an estimation method based on the CAVI algorithm for the Gaussian estimation problem proposed in [8]. This Gaussian estimation problem considers object features that are modeled according to a DPM of Gaussians, and estimated from noisy observations of the object features. As was shown in [8], by implicitly clustering the observations and estimating the parameters of the clusters, the accuracy of the estimates of the object features is improved compared to a conventional estimation that does not use clustering. This improvement is referred to as the clustering gain.

The Gaussian estimation problem was solved in [8] using an MCMC method known as the Gibbs sampler. We experimentally compare the estimation accuracy and clustering gain obtained using the CAVI algorithm and the Gibbs sampler. For our simulations, we use synthetic data generated by a DPM of Gaussians with values for the hyperparameters that match the values chosen in [8]. Moreover, we investigate the impact of different initialization procedures on the convergence of the CAVI algorithm and compare the case where the DPM hyperparameters conform to those underlying the data with the case where they do not.

1.2 Related Work

As mentioned above, we consider a specific type of BMM called the DPM model. A treatment of BMMs including DPM models is provided in [9]. A more detailed view on the DP and DPM models as well as a general discussion of inference for Bayesian nonparametric models are given in [10]. Further work considers representations of the DPM model and related statistical processes, namely the stick-breaking process [11], the Blackwell-MacQueen Urn Scheme [12], and the Chinese restaurant process [13]. These representations are important for the derivation of the CAVI algorithm for DPM models and for generating data from a DPM model.

A review of VI methods is given in [6], and recent advances in VI are discussed in [14]. A VI method for DPM models based on CAVI is introduced in [7]. Later works [15]–[21] discuss various improvements of CAVI for DPM models in terms of estimation accuracy, efficiency, scalability for large data sets, and the adaptation of VI methods to streaming data.

For a review of MCMC methods we refer to [5] and [22]. Various methods for DPM models based on Gibbs sampling are presented in [23]–[28]. The works [8], [22], [29] use the DPM model and MCMC methods to improve the estimation of model parameters and object features by an implicit clustering of the data. As previously mentioned, we will compare our simulation results to those of [8], which where obtained using a Gibbs sampler.

1.3 Thesis Outline

Following this introductory chapter, Chapter 2 presents an introduction to BMMs and the exponential family. In particular, the chapter discusses the estimation of the parameters of a BMM from observed data.

In Chapter 3, we focus on the DP and DPM models. We explain different representations of the DP and define a DPM model with component distributions from the exponential family. This model forms the basis for the statistical models considered in later chapters.

In Chapter 4, we describe the VI method and the CAVI algorithm. We derive the CAVI algorithm for DPM models and discuss practical considerations like initialization and convergence.

In Chapter 5, we focus on a Gaussian estimation problem and specialize the CAVI algorithm of Chapter 4 to a DPM with Gaussian components. We present and discuss simulation results obtained with the CAVI algorithm and compare them to results obtained with Gibbs sampling in [8]. Furthermore, we compare the performance gain due to clustering achieved by the two methods.

Finally, Chapter 6 concludes this thesis by summarizing the main results and suggesting future research directions.

1.4 Notation

This section summarizes the mathematical notation used throughout the remaining chapters of this thesis. Scalar-valued quantities are denoted by lowercase letters, e.g., x, vectors by lowercase boldface letters, e.g., $\boldsymbol{x}$, and matrices by uppercase boldface letters, e.g., $\boldsymbol{X}$. Our notation does not distinguish between a random variable and a realization drawn from its distribution. To indicate the distribution of a continuous random variable x, we write $x \sim f(x)$; e.g., $x \sim \mathcal{N}(x; \mu_x, \sigma_x^2)$ means that the random variable x has a Gaussian probability distribution with mean μ_x and variance σ_x^2. A summary of our mathematical notation is given in Table 1.1.

$\mathbb{N}$	natural numbers
$\mathbb{R}$	real numbers
Δ_K	$(K-1)$-dimensional probability simplex
x	scalar
$\boldsymbol{x}$	vector
$\boldsymbol{X}$	matrix
$\boldsymbol{x}_{1:N}$	N column vectors $\boldsymbol{x}_n$, $n = 1, \ldots, N$ stacked into a column vector, i.e., $\boldsymbol{x}_{1:N} = \left(\boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}}\right)^{\mathrm{T}}$
$\boldsymbol{x}_{\sim m}$	vector $\boldsymbol{x}$ without the m-th entry, i.e., $\boldsymbol{x}_{\sim m} = \left(x_1 \ \cdots \ x_{m-1} \ x_{m+1} \ \cdots \ x_M\right)^{\mathrm{T}}$
$\mathbf{1}_M$	all-ones vector of dimension $M \times 1$
$\mathbf{I}_M$	$M \times M$ identity matrix
$\mathbb{1}(\cdot)$	indicator function
$\ln(\cdot)$	natural logarithm
$\Psi(\cdot)$	digamma function
$\delta(x)$	Dirac delta function
$\det(\cdot)$	determinant
$\mathrm{tr}(\cdot)$	trace
$\cdot^{-1}$	inverse
$\dim(\cdot)$	vector/matrix dimension
$\cdot^{\mathrm{T}}$	vector/matrix transpose
$x \perp\!\!\!\perp y$	independence of two random variables x and y
$x \perp\!\!\!\perp y \mid z$	conditional independence of two random variables x and y given z
$\mathrm{E}^{(f(\boldsymbol{x}))}\{\cdot\}$	expectation with respect to the pdf $f(\boldsymbol{x})$
μ_x, $\boldsymbol{\mu}_{\boldsymbol{x}}$	mean, mean vector
σ_x^2, σ_x	variance, standard deviation
$\boldsymbol{\Sigma}_{\boldsymbol{x}}$	covariance matrix
$f(x)$, $f(\boldsymbol{x})$	probability density function (pdf)
$p(x)$, $p(\boldsymbol{x})$	probability mass function (pmf)
$f(x\|\theta)$, $f(\boldsymbol{x}\|\boldsymbol{\theta})$	conditional probability density function
$p(x\|\theta)$, $p(\boldsymbol{x}\|\boldsymbol{\theta})$	conditional probability mass function
$q(\boldsymbol{x})$	variational approximation of a probability distribution
$G(\boldsymbol{\theta})$	discrete mixing distribution
$\mathcal{B}(\,\cdot\,; \alpha, \beta)$	beta distribution with shape parameters α and β
$\mathcal{D}(\,\cdot\,; \boldsymbol{\alpha})$	Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = \left(\alpha_1 \ \cdots \ \alpha_K\right)^{\mathrm{T}}$
$\mathcal{DP}(\,\cdot\,; \alpha, G_0)$	Dirichlet process with concentration parameter α and base distribution G_0
$\mathrm{GEM}(\,\cdot\,; \alpha)$	GEM distribution with concentration parameter α
$\mathcal{N}(\,\cdot\,; \mu, \sigma^2)$	univariate Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{C}(\,\cdot\,; \boldsymbol{\pi})$	categorical distribution with event probabilities $\boldsymbol{\pi} = \left(\pi_1 \ \cdots \ \pi_K\right)^{\mathrm{T}}$

Table 1.1: Summary of notation.

2 Bayesian Mixture Models and the Exponential Family

In this chapter, we provide an introduction to Bayesian mixture models (BMM), which involve considering the parameters of a mixture model as random variables. Furthermore, we review the exponential family (EF) of probability distributions and Bayesian estimation. We will first define BMMs and two equivalent ways to represent them. Following that, we will proceed to define the EF and introduce the concept of conjugate priors. Lastly, we will outline two popular Bayesian estimators, i.e., the minimum mean square error (MMSE) and maximum a-posterior (MAP) estimators, and address the problem of estimating the unknown parameters of a BMM given observed data. Our objective is to help readers become acquainted with the relevant theory, terminology, and notation while establishing a solid foundation for the remainder of the thesis.

2.1 Bayesian Mixture Models

Mixture models [9], [30], [31] provide a flexible and parametric framework for statistical modeling of a wide variety of random phenomena, especially when sampling from a population that consists of a number of subpopulations. Examples are image processing where mixture models can be used to segment an image into different regions or objects [32], modeling the mixture of topics within a collection of text documents [33], and modeling the geographical distribution of disease occurrence [34]. Mixture models assume that the observed data is generated by a superposition of underlying component distributions (also called mixands), with their respective contributions governed by proportions called the mixing proportions. In a Bayesian approach we model the parameters of the component distributions and the mixing proportions as random variables with prior distributions that represent our initial beliefs about these values. We therefore refer to such models as BMMs. In what follows, we will first define a BMM and then present two alternative representations that are important for the remaining chapters of this thesis.

2.1.1 Definition

Consider a random vector $\boldsymbol{x} = (x_1 \ \cdots \ x_M)^{\mathrm{T}} \in \mathbb{R}^M$ of dimension $M \in \mathbb{N}$ and $K \in \mathbb{N}$ component distributions $f(\boldsymbol{x}|\boldsymbol{\theta}_k^*)$, $k = 1, \ldots, K$, with random parameter vectors $\boldsymbol{\theta}_k^* = (\theta_1^* \ \cdots \ \theta_p^*)^{\mathrm{T}} \in \mathbb{R}^p$. The component distributions (mixands) are conditional probability density functions (pdfs) from the same parametric family, which is represented by the set of functions $\{f(\boldsymbol{x}|\boldsymbol{\theta}_k^*) \,|\, \boldsymbol{\theta}_k^* \in \mathbb{R}^p\}$, and each component distribution is indexed by its respective component parameter $\boldsymbol{\theta}_k^*$. We

say that $\boldsymbol{x}$ is distributed according to a mixture distribution with K components when the pdf $f(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K})$ is a convex combination of the K component distributions $f(\boldsymbol{x}|\boldsymbol{\theta}^*_k)$, i.e.,

$$f(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}|\boldsymbol{\theta}^*_k). \tag{2.1}$$

Here, $\boldsymbol{\theta}^*_{1:K} = \left(\boldsymbol{\theta}^{*\mathrm{T}}_1 \ \cdots \ \boldsymbol{\theta}^{*\mathrm{T}}_K\right)^{\mathrm{T}}$ is a vector containing the parameters $\boldsymbol{\theta}^*_k$, $k = 1, \ldots, K$ and $\boldsymbol{\pi} = (\pi_1 \ \cdots \ \pi_K)^{\mathrm{T}}$ is a vector containing random mixing proportions $\pi_k \in [0, 1]$, also called the mixing weights or the mixing probabilities.

The fact that (2.1) is a convex combination means the proportions π_k have to satisfy the constraints $\pi_k \geq 0$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. Additionally, this guarantees that (2.1) is a valid probability distribution, i.e., integrating $f(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K})$ over the domain $\mathbb{R}^M$ results in 1. Thus, the random vector $\boldsymbol{\pi}$ exists in the $(K-1)$-dimensional probability simplex defined by

$$\Delta_K = \left\{ \boldsymbol{\pi} \in [0, 1]^K \,\middle|\, \pi_k \geq 0 \text{ and } \sum_{k=1}^{K} \pi_k = 1 \right\}, \tag{2.2}$$

which is a $(K-1)$-dimensional object that lies in the K-dimensional space $[0, 1]^K$. Each vector $\boldsymbol{\pi} \in \Delta_K$ can be thought of as a probability mass function (pmf) with probabilities π_k, for $k = 1, \ldots, K$, i.e., π_k is the probability that a vector $\boldsymbol{x}$ arises from the k-th mixture component $f(\boldsymbol{x}|\boldsymbol{\theta}^*_k)$. We assume the unknown vector $\boldsymbol{\pi}$ to be distributed according to a prior distribution $f(\boldsymbol{\pi})$ and, therefore, $f(\boldsymbol{\pi})$ is a distribution over a distribution, specifically a distribution over the pmf with probabilities π_k for $k = 1, \ldots, K$. Typical choices for $f(\boldsymbol{\pi})$ are the beta distribution, the Dirichlet distribution, and, in the case of $K = \infty$, the GEM [35] distribution. The choices for the prior distribution $f(\boldsymbol{\pi})$ will be discussed in more detail in Chapter 3. Furthermore, we assume that the parameter vectors $\boldsymbol{\theta}^*_k$ are independent and identically distributed (i.i.d.) according to a prior distribution $f(\boldsymbol{\theta}^*)$ and that $\boldsymbol{\theta}^*_k$ and $\boldsymbol{\pi}$ are statistically independent, i.e.,

$$\boldsymbol{\theta}^*_k \perp\!\!\!\perp \boldsymbol{\pi}, \quad \text{for all } k = 1, \ldots, K. \tag{2.3}$$

We summarize the BMM described above as

$$\boldsymbol{\pi} \sim f(\boldsymbol{\pi}), \tag{2.4a}$$

$$\boldsymbol{\theta}^*_k \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}^*_k), \tag{2.4b}$$

$$\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K} \sim f(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}|\boldsymbol{\theta}^*_k), \tag{2.4c}$$

for all $k = 1, \ldots, K$. Note that we consider the vectors $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*_{1:K}$ to be hidden (unknown) parameters of the BMM which determine the statistical behavior of the random vector $\boldsymbol{x}$.

2.1.2 Representation Using Random Indicator Variables

We now consider multiple conditionally i.i.d. observations $\boldsymbol{x}_n$, $n = 1, \ldots, N$, of the random vector $\boldsymbol{x}$, where we will refer to $\boldsymbol{x}_{1:N} = \left(\boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}}\right)^{\mathrm{T}}$ as the observed data. An equivalent way of generating an observation $\boldsymbol{x}_n$ that is distributed according to the K-component mixture $f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K})$ given by (2.1) is as follows. Let z_n be a categorical random variable taking on the values $k = 1, \ldots, K$ with probabilities $\pi_1, \ldots, \pi_K$. We will refer to z_n as indicator variable and use it to indicate which component of the mixture model is responsible for generating the n-th observation $\boldsymbol{x}_n$. Similar to the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ being conditionally i.i.d., we assume the variables z_n to be i.i.d. conditioned on the mixing proportions $\boldsymbol{\pi}$. First, $z_n \in \{1, \ldots, K\}$ is realized from the categorical distribution

$$p(z_n|\boldsymbol{\pi}) = \mathcal{C}(z_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\mathbb{1}(z_n=k)} = \pi_{z_n}, \tag{2.5}$$

where $\mathbb{1}(\cdot)$ represents the indicator function that takes a condition as its input and returns 1 if the condition is satisfied and 0 if the condition is not satisfied. Then, conditioned on z_n and the component parameters $\boldsymbol{\theta}^*_{1:K}$, the observation $\boldsymbol{x}_n$ is realized from the corresponding component distribution, i.e.,

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{1:K}, z_n) = \prod_{k=1}^{K} f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_k)^{\mathbb{1}(z_n=k)} = f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}). \tag{2.6}$$

Incorporating the hidden indicator variables z_n into the BMM in (2.4) yields the following hierarchical representation:

$$\boldsymbol{\pi} \sim f(\boldsymbol{\pi}), \tag{2.7a}$$
$$\boldsymbol{\theta}^*_k \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}^*_k), \tag{2.7b}$$
$$z_n|\boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n|\boldsymbol{\pi}), \tag{2.7c}$$
$$\boldsymbol{x}_n|\boldsymbol{\theta}^*_{1:K}, z_n \sim f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}), \tag{2.7d}$$

for $k = 1, \ldots, K$ and $n = 1, \ldots, N$, with conditional or unconditional independence relations among the parameters, indicator variables, and observations given by

$$\boldsymbol{\theta}^*_k \perp\!\!\!\perp \boldsymbol{\pi}, \quad \text{for all } k = 1, \ldots, K, \tag{2.8a}$$
$$\boldsymbol{\theta}^*_k \perp\!\!\!\perp z_n \mid \boldsymbol{\pi}, \quad \text{for all } k = 1, \ldots, K \text{ and } n = 1, \ldots, N, \tag{2.8b}$$
$$\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{\pi}, \boldsymbol{z}_{n'}, \boldsymbol{x}_{n'} \mid \boldsymbol{\theta}^*_{1:K}, z_n, \quad \text{where } n, n' = 1, \ldots, N \text{ with } n \neq n'. \tag{2.8c}$$

This representation of the BMM again gives a mixture distribution in the form of (2.1), because by marginalizing the indicator variables z_n from the conditional joint distribution $f(\boldsymbol{x}_n, z_n|\boldsymbol{\theta}^*_{1:K}, \boldsymbol{\pi})$ we obtain

$$f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) = \sum_{z_n=1}^{K} f(\boldsymbol{x}_n, z_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K})$$

$$= \sum_{z_n=1}^{K} f(\boldsymbol{x}_n|z_n, \boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) p(z_n|\boldsymbol{\pi})$$
$$= \sum_{z_n=1}^{K} \pi_{z_n} f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}). \tag{2.9}$$

Note that the set of hidden parameters also includes the indicator variables $\boldsymbol{z}_{1:N} = (z_1 \ \cdots \ z_N)^{\mathrm{T}}$ in this representation. The mixing proportions $\boldsymbol{\pi}$ and the component parameters $\boldsymbol{\theta}^*_{1:K}$ are "global" parameters that determine the shape of the mixture distribution. The count of the global parameters increases with the number of components K in the mixture model. The indicator variables $\boldsymbol{z}_{1:N}$ are "local" parameters in that each z_n associates the corresponding data point $\boldsymbol{x}_n$ with a certain component k. Thus, the dimension of $\boldsymbol{z}_{1:N}$ increases with the number N of observed data points $\boldsymbol{x}_{1:N}$.

Based on the indicator variables $\boldsymbol{z}_{1:N}$, the observed data $\boldsymbol{x}_{1:N}$ can be organized into $L \leq K$ groups which we will refer to as clusters. Clusters are formed by indicator variables z_n that share the same value k. All data points $\boldsymbol{x}_n$ that are associated with indicator variables z_n of the same cluster are considered to be part of that cluster. Assuming that we know $\boldsymbol{z}_{1:N}$ (or have an estimate thereof), we can count the number of observations $\boldsymbol{x}_n$ generated by component k with

$$N_k = \sum_{n=1}^{N} \mathbb{1}(z_n = k), \tag{2.10}$$

and determine the number of clusters L, i.e., the number of mixture components used to generate the observed data, by

$$L = \sum_{k=1}^{K} \mathbb{1}(N_k > 0). \tag{2.11}$$

Note that the number of observations N is reobtained from

$$\sum_{k=1}^{K} N_k = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}(z_n = k) = \sum_{n=1}^{N} 1 = N. \tag{2.12}$$

Furthermore note that, L may be smaller than K, which means that not every one of the K components was used to generate a data point $\boldsymbol{x}_n$. This is trivially true if the number of observed data points N is smaller than the number of components K. For $N > K$, this means at least two $\boldsymbol{x}_n$ belong to the same cluster. Still it can happen that $L < K$, i.e., when there are components in the mixture model with a small mixing proportion π_k, that is, components that are not likely to generate a sample $\boldsymbol{x}_n$.

2.1.3 Representation Using a Mixing Distribution

Recall that the mixture proportions $\boldsymbol{\pi}$ represent a probability distribution because $\boldsymbol{\pi}$ is an element of the $(K-1)$-dimensional probability simplex defined by (2.2). This distribution can be defined over the component parameters $\boldsymbol{\theta}^*_{1:K}$ using a sum of weighted Dirac delta functions,

i.e.,

$$G(\boldsymbol{\theta}) \triangleq \sum_{k=1}^{K} \pi_k \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*), \tag{2.13}$$

where we associate the probability π_k with the event $\boldsymbol{\theta} = \boldsymbol{\theta}_k^*$. Note that $G(\boldsymbol{\theta})$ is a random pdf because it includes the random vectors $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_{1:K}^*$ of the BMM (2.4). In the context of mixture models, $G(\boldsymbol{\theta})$ is referred to as the mixing distribution (cf. [9]). We can represent the BMM for N conditionally independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ using the mixing distribution $G(\boldsymbol{\theta})$ and local parameters $\boldsymbol{\theta}_n \in \{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*\}$, $n = 1, \ldots, N$, that associate each data point $\boldsymbol{x}_n$ with a certain component k, in the following hierarchical manner:

$$\boldsymbol{\pi} \sim f(\boldsymbol{\pi}), \tag{2.14a}$$

$$\boldsymbol{\theta}_k^* \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}_k^*), \tag{2.14b}$$

$$G(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\theta}_{1:K}^*) = \sum_{k=1}^{K} \pi_k \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*), \tag{2.14c}$$

$$\boldsymbol{\theta}_n|G \overset{\text{i.i.d.}}{\sim} G(\boldsymbol{\theta}_n|\boldsymbol{\pi}, \boldsymbol{\theta}_{1:K}^*), \tag{2.14d}$$

$$\boldsymbol{x}_n|\boldsymbol{\theta}_n \sim f(\boldsymbol{x}_n|\boldsymbol{\theta}_n), \tag{2.14e}$$

for $k = 1, \ldots, K$ and $n = 1, \ldots, N$. A realization of $G(\boldsymbol{\theta})$ is determined by sampling from the prior distributions $f(\boldsymbol{\pi})$ and $f(\boldsymbol{\theta}^*)$. Given the mixing distribution $G(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\theta}_{1:K}^*)$, we consider N samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ that are conditionally i.i.d. according to this mixing distribution (see (2.14d)). This means that

$$f(\boldsymbol{\theta}_n|G) = G(\boldsymbol{\theta}_n|\boldsymbol{\pi}, \boldsymbol{\theta}_{1:K}^*). \tag{2.15}$$

Because of the discrete nature of $G(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\theta}_{1:K}^*)$ in (2.14c), each sample $\boldsymbol{\theta}_n$ is equal to a component parameter $\boldsymbol{\theta}_k^*$ with probability π_k, i.e., $\Pr(\boldsymbol{\theta}_n = \boldsymbol{\theta}_k^* \mid \boldsymbol{\pi}, \boldsymbol{\theta}_{1:K}^*) = \pi_k$. By using the conditional pdf $f(\boldsymbol{x}_n|\boldsymbol{\theta}_n)$ in (2.14e) we associate the n-th sample $\boldsymbol{\theta}_n$ with the n-th observation $\boldsymbol{x}_n$ and determine from which component distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*)$ each data point $\boldsymbol{x}_n$ is realized. We assume that the n-th observation $\boldsymbol{x}_n$ statistically depends on $\boldsymbol{\theta}_n$ but not on the random pdf G and not on any other local parameter $\boldsymbol{\theta}_{n'}$ with $n' \neq n$, i.e.,

$$\boldsymbol{x}_n \perp\!\!\!\perp G, \boldsymbol{\theta}_{n'} \mid \boldsymbol{\theta}_n \quad \text{where} \quad n, n' = 1, \ldots, N \quad \text{with} \quad n' \neq n.$$

Hence, we have

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}_{1:N}, G) = f(\boldsymbol{x}_n|\boldsymbol{\theta}_n, G) = f(\boldsymbol{x}_n|\boldsymbol{\theta}_n). \tag{2.16}$$

The mixture distribution (2.1) can be formally rewritten as $f(\boldsymbol{x}_n|G)$. Indeed,

$$\begin{aligned} f(\boldsymbol{x}_n|G) &= \int_{\mathbb{R}^p} f(\boldsymbol{x}_n|\boldsymbol{\theta}_n, G) f(\boldsymbol{\theta}_n|G) \, \mathrm{d}\boldsymbol{\theta}_n \\ &= \int_{\mathbb{R}^p} f(\boldsymbol{x}_n|\boldsymbol{\theta}_n) G(\boldsymbol{\theta}_n|\boldsymbol{\pi}, \boldsymbol{\theta}_k^*) \, \mathrm{d}\boldsymbol{\theta}_n \end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^p} f(\boldsymbol{x}_n|\boldsymbol{\theta}_n)\left(\sum_{k=1}^{K}\pi_k\delta(\boldsymbol{\theta}_n-\boldsymbol{\theta}_k^*)\right)\mathrm{d}\boldsymbol{\theta}_n \\
&= \sum_{k=1}^{K}\pi_k\int_{\mathbb{R}^p} f(\boldsymbol{x}_n|\boldsymbol{\theta}_n)\delta(\boldsymbol{\theta}_n-\boldsymbol{\theta}_k^*)\,\mathrm{d}\boldsymbol{\theta}_n \\
&= \sum_{k=1}^{K}\pi_k f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*), \qquad (2.17)
\end{aligned}
$$

where we have used (2.16) and (2.15) in the first step. Thus, $f(\boldsymbol{x}_n|G)$ is equal to (2.1), i.e., $f(\boldsymbol{x}_n|G) = f(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}_{1:K}^*)$. This means that the set of hidden parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_{1:K}^*$ is equivalent to the mixing distribution $G(\boldsymbol{\theta})$.

The representation (2.14) of a BMM also allows us to determine the clustering structure of the data $\boldsymbol{x}_{1:N}$. Similar to the indicator variables $\boldsymbol{z}_n$ taking on $L \leq K$ distinct values $1,\dots,L$ for $n = 1,\dots,N$, the local parameters $\boldsymbol{\theta}_n$ take on $L \leq K$ distinct values $\boldsymbol{\theta}_l'$, $l = 1,\dots,L$. Each $\boldsymbol{\theta}_l'$ is equal to one of the parameters $\boldsymbol{\theta}_1^*,\dots,\boldsymbol{\theta}_K^*$. The number of times $\boldsymbol{\theta}_k^*$ appears in $\boldsymbol{\theta}_{1:N}$, which is equal to the number of observations $\boldsymbol{x}_n$ generated by component k (cf. (2.10)), is given by

$$
N_k = \sum_{n=1}^{N}\mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\theta}_k^*). \qquad (2.18)
$$

This again determines the number of clusters L by evaluating (2.11).

Finally, we note that the number of components K in a mixture model does not have to be finite. Countably infinite mixtures with mixing distribution $G(\boldsymbol{\theta}) = \sum_{k=1}^{\infty}\pi_k\delta(\boldsymbol{\theta}-\boldsymbol{\theta}_k^*)$ are well-defined and often used in practice when it is difficult to choose K in advance. Using a mixture model with an infinite number of components and observing a finite data set $\boldsymbol{x}_{1:N}$ means there will always be "inactive components," i.e., components k for which $\boldsymbol{\theta}_n \neq \boldsymbol{\theta}_k^*$ for all $n = 1,\dots,N$ or, equivalently, $z_n \neq k$ for all $n = 1,\dots,N$. A prominent example of an infinite mixture is the Dirichlet process mixture, which will be discussed in Chapter 3. A different approach is to consider the number of components K to be random according to a prior distribution $f(K)$. Such models are called mixture of finite mixtures [36], [37] and are outside the scope of this thesis.

2.2 Exponential Family

Throughout this thesis our focus will be on mixtures consisting of component distributions belonging to the exponential family (EF), i.e., a set of parameterized probability distributions with a certain mathematical structure. An important example is constituted by mixtures of multivariate Gaussian distributions. Therefore, we will now provide a concise introduction to the EF. Many of the commonly used distributions, e.g. Gaussian, Poisson, binomial, exponential, and gamma, belong to the EF. When the parameters of a distribution belonging to the EF are considered to be random variables, such distributions are conditional probability distributions or, when viewed as a function of the parameters, they can be interpreted as likelihood functions involving observed data. In the Bayesian context, likelihood functions can be combined with conjugate priors to obtain posterior distributions that have the same functional form as the

prior distribution. The concept of conjugate priors is important because it allows for an easy determination of the posterior distribution and, thereby, an efficient inference of the parameters. In this regard, we will see that likelihood functions belonging to the EF always have conjugate priors. A more detailed introduction to the EF and its properties can be found in [2], [38], [39].

2.2.1 Definition

A conditional pdf $f(\boldsymbol{x}|\boldsymbol{\theta}^*)$ with M-dimensional vector $\boldsymbol{x} \in \mathbb{R}^M$ and p-dimensional parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is element of an EF if it is of the following form [2]:

$$f(\boldsymbol{x}|\boldsymbol{\theta}^*) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^{*\mathrm{T}}(\boldsymbol{\theta}^*)\boldsymbol{t}(\boldsymbol{x}) - a(\boldsymbol{\eta}^*(\boldsymbol{\theta}^*))). \tag{2.19}$$

Here, $h(\boldsymbol{x}) \geq 0$ is called the base measure, $\boldsymbol{\eta}^*(\boldsymbol{\theta}^*) = \begin{pmatrix} \eta_1^*(\boldsymbol{\theta}^*) & \cdots & \eta_p^*(\boldsymbol{\theta}^*) \end{pmatrix}^{\mathrm{T}}$ is a p-dimensional vector consisting of p parameter functions, $\boldsymbol{t}(\boldsymbol{x}) = \begin{pmatrix} t_1(\boldsymbol{x}) & \cdots & t_p(\boldsymbol{x}) \end{pmatrix}^{\mathrm{T}}$ is a p-dimensional vector called the sufficient statistic, and $a(\boldsymbol{\eta}^*(\boldsymbol{\theta}^*))$ is a real function called the log-partition function. If we consider the parameter in (2.19) to be $\boldsymbol{\eta}^* = \boldsymbol{\eta}^*(\boldsymbol{\theta}^*)$ rather than $\boldsymbol{\theta}^*$, the EF is said to be in canonical form and $\boldsymbol{\eta}^* = \begin{pmatrix} \eta_1^* & \cdots & \eta_p^* \end{pmatrix}^{\mathrm{T}}$ is referred to as canonical or natural parameter. Here, (2.19) reduces to

$$f(\boldsymbol{x}|\boldsymbol{\eta}^*) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}) - a(\boldsymbol{\eta}^*)). \tag{2.20}$$

The log-partition function $a(\boldsymbol{\eta}^*)$ ensures that the pdf is normalized, which means it is automatically determined by the functions $h(\boldsymbol{x})$ and $\boldsymbol{t}(\boldsymbol{x})$. Integrating (2.20) yields

$$\int_{\mathbb{R}^M} f(\boldsymbol{x}|\boldsymbol{\eta}^*)\,\mathrm{d}\boldsymbol{x} = \exp(-a(\boldsymbol{\eta}^*)) \int_{\mathbb{R}^M} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} = 1,$$

and thus

$$\exp(a(\boldsymbol{\eta}^*)) = \int_{\mathbb{R}^M} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x},$$

or equivalently

$$a(\boldsymbol{\eta}^*) = \ln\left(\int_{\mathbb{R}^M} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\right). \tag{2.21}$$

For the practically important case of observing a sequence of data $\boldsymbol{x}_{1:N}$, where $\boldsymbol{x}_n$ is conditionally i.i.d. given $\boldsymbol{\eta}^*$ and individually distributed according to (2.20), the likelihood function is obtained as

$$\begin{aligned} f(\boldsymbol{x}_{1:N}|\boldsymbol{\eta}^*) &= \prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*) \\ &= \prod_{n=1}^{N} h(\boldsymbol{x}_n) \exp(\boldsymbol{\eta}^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}) - a(\boldsymbol{\eta}^*)) \\ &= \left(\prod_{n=1}^{N} h(\boldsymbol{x}_n)\right) \exp\left(\boldsymbol{\eta}^{*\mathrm{T}} \sum_{n=1}^{N} \boldsymbol{t}(\boldsymbol{x}_n) - Na(\boldsymbol{\eta}^*)\right) \end{aligned}$$

$$= h'(\boldsymbol{x}_{1:N}) \exp(\boldsymbol{\eta}^{*\mathrm{T}} \boldsymbol{t}'(\boldsymbol{x}_{1:N}) - a'(\boldsymbol{\eta}^*)), \tag{2.22}$$

which is again an EF pdf with the same parameter vector $\boldsymbol{\eta}^*$ and

$$h'(\boldsymbol{x}_{1:N}) = \prod_{n=1}^{N} h(\boldsymbol{x}_n), \tag{2.23}$$

$$\boldsymbol{t}'(\boldsymbol{x}_{1:N}) = \sum_{n=1}^{N} \boldsymbol{t}(\boldsymbol{x}_n), \tag{2.24}$$

$$a'(\boldsymbol{\eta}^*) = N a(\boldsymbol{\eta}^*). \tag{2.25}$$

Examples of distributions used in this thesis that belong to the EF include the Gaussian, beta, Dirichlet, and categorical distributions. Although a product of distributions belonging to the EF also belongs to the EF (see (2.22) for example), a mixture of distributions belonging to the EF, e.g. a mixture of Gaussian distributions, does not necessarily belong to the EF [38].

2.2.2 Conjugate Prior

In what follows we assume the EF distribution to be in canonical form (see (2.20)). An important property of the EF is that all of its members have a conjugate prior, i.e., for every EF likelihood function $f(\boldsymbol{x}|\boldsymbol{\eta}^*)$ there exists a prior $f(\boldsymbol{\eta}^*)$ that is conjugate to the likelihood $f(\boldsymbol{x}|\boldsymbol{\eta}^*)$. Conjugacy of the prior $f(\boldsymbol{\eta}^*)$ and likelihood $f(\boldsymbol{x}|\boldsymbol{\eta}^*)$ means that the posterior $f(\boldsymbol{\eta}^*|\boldsymbol{x})$ has the same functional form as the prior $f(\boldsymbol{\eta}^*)$. This simplifies Bayesian inference of the parameters $\boldsymbol{\eta}^*$ because it leads to closed-form solutions of the posterior distributions where otherwise an approximation may be necessary.

Let us consider N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ that are conditionally i.i.d. and individually distributed according to (2.20), so that the likelihood function of $\boldsymbol{x}_{1:N}$ is of the form (2.22). Then, the corresponding conjugate prior can be written as

$$f(\boldsymbol{\eta}^*) = b(\boldsymbol{\lambda}) \exp(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}^* - \lambda_2 a(\boldsymbol{\eta}^*)), \tag{2.26}$$

where $b(\boldsymbol{\lambda}) \in \mathbb{R}^+$ is a normalization coefficient given by

$$b(\boldsymbol{\lambda}) = \frac{1}{\int_{\mathbb{R}^p} \exp(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}^* - \lambda_2 a(\boldsymbol{\eta}^*)) \, \mathrm{d}\boldsymbol{\eta}^*}. \tag{2.27}$$

The hyperparameters $\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1^{\mathrm{T}} & \lambda_2 \end{pmatrix}^{\mathrm{T}}$, with $\boldsymbol{\lambda}_1 \in \mathbb{R}^p$ and $\lambda_2 \in \mathbb{R}$, represent our prior belief about $\boldsymbol{\eta}^*$, and $a(\boldsymbol{\eta}^*)$ is the same function as appears in (2.20) and (2.21).

Through Bayes' theorem, the posterior $f(\boldsymbol{\eta}^*|\boldsymbol{x}_{1:N})$ is obtained by updating the prior (2.26) with information summarized by the likelihood function (2.22), i.e.,

$$f(\boldsymbol{\eta}^*|\boldsymbol{x}_{1:N}) = \frac{f(\boldsymbol{x}_{1:N}|\boldsymbol{\eta}^*) f(\boldsymbol{\eta}^*)}{f(\boldsymbol{x}_{1:N})} \tag{2.28}$$

$$\propto f(\boldsymbol{x}_{1:N}|\boldsymbol{\eta}^*) f(\boldsymbol{\eta}^*) \tag{2.29}$$

$$\propto \exp(\boldsymbol{\eta}^{*\mathrm{T}} \boldsymbol{t}'(\boldsymbol{x}_{1:N}) - a'(\boldsymbol{\eta}^*)) \exp(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}^* - \lambda_2 a(\boldsymbol{\eta}^*)) \tag{2.30}$$

$$\propto \exp\left(\left(\boldsymbol{\lambda}_1 + \sum_{n=1}^{N} \boldsymbol{t}(\boldsymbol{x}_n)\right)^{\mathrm{T}} \boldsymbol{\eta}^* - (\lambda_2 + N) a(\boldsymbol{\eta}^*)\right), \tag{2.31}$$

where (2.24) and (2.25) were used and the symbol $\propto$ denotes equality up to a normalization factor that does not depend on $\boldsymbol{\eta}^*$. Equivalently, we have

$$f(\boldsymbol{\eta}^* | \boldsymbol{x}_{1:N}) = b(\boldsymbol{\tau}) \exp(\boldsymbol{\tau}_1^{\mathrm{T}}(\boldsymbol{x}_{1:N}) \boldsymbol{\eta}^* - \tau_2 a(\boldsymbol{\eta}^*)) \tag{2.32}$$

with updated hyperparameters $\boldsymbol{\tau} = \left(\boldsymbol{\tau}_1^{\mathrm{T}}(\boldsymbol{x}_{1:N}) \;\; \tau_2\right)^{\mathrm{T}}$ given by

$$\boldsymbol{\tau}_1(\boldsymbol{x}_{1:N}) = \boldsymbol{\lambda}_1 + \sum_{n=1}^{N} \boldsymbol{t}(\boldsymbol{x}_n), \tag{2.33a}$$

$$\tau_2 = \lambda_2 + N, \tag{2.33b}$$

and normalization constant $b(\boldsymbol{\tau}) = 1/\int_{\mathbb{R}^p} \exp(\boldsymbol{\tau}_1^{\mathrm{T}}(\boldsymbol{x}_{1:N}) \boldsymbol{\eta}^* - \tau_2 a(\boldsymbol{\eta}^*)) \, \mathrm{d}\boldsymbol{\eta}^*$. Comparing (2.32) with (2.26), we recognize that the posterior pdf $f(\boldsymbol{\eta}^* | \boldsymbol{x}_{1:N})$ is of the same functional form as the prior pdf $f(\boldsymbol{\eta}^*)$, thus confirming conjugacy; however the hyperparameters $\boldsymbol{\lambda}$ are updated according to (2.33). This update relation can be interpreted as an augmentation of our prior beliefs, represented by $\boldsymbol{\lambda}$, by the information provided by the data $\boldsymbol{x}_{1:N}$, represented by the sum $\sum_{n=1}^{N} \boldsymbol{t}(\boldsymbol{x}_n)$ and the sample size N.

2.3 Bayesian Estimation of Mixture Models

Throughout the thesis, we will summarize all hidden parameters of a statistical model in a vector denoted by $\boldsymbol{w} \in \mathbb{R}^P$. For the BMM (2.4), we have $\boldsymbol{w} = \left(\boldsymbol{\pi}^{\mathrm{T}} \;\; \boldsymbol{\theta}_{1:K}^{*\mathrm{T}}\right)^{\mathrm{T}}$ with

$$P = \dim(\boldsymbol{w}) = \dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\theta}_{1:K}^*) = K + Kp, \tag{2.34}$$

and for the BMM (2.7) including the indicator variables $\boldsymbol{z}_{1:N}$, we have $\boldsymbol{w} = \left(\boldsymbol{\pi}^{\mathrm{T}} \;\; \boldsymbol{z}_{1:N}^{\mathrm{T}} \;\; \boldsymbol{\theta}_{1:K}^{*\mathrm{T}}\right)^{\mathrm{T}}$ with

$$P = \dim(\boldsymbol{w}) = \dim(\boldsymbol{\pi}) + \dim(\boldsymbol{z}_{1:N}) + \dim(\boldsymbol{\theta}_{1:K}^*) = K + N + Kp. \tag{2.35}$$

Here, the operator $\dim(\cdot)$ takes a vector as argument and returns its dimension, and P denotes the number of scalar parameters contained in $\boldsymbol{w}$.

Bayesian estimation is a statistical approach to obtaining an estimate of the unknown parameter vector $\boldsymbol{w}$ based on both prior knowledge and observed data [40]. An estimator $\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}(\boldsymbol{x}_{1:N})$ is said to be a Bayes estimator if it minimizes the posterior expected value of a loss function (or cost function), which means it involves the posterior pdf $f(\boldsymbol{w} | \boldsymbol{x}_{1:N})$ and a function that takes into account the estimation error $\boldsymbol{e} = \hat{\boldsymbol{w}} - \boldsymbol{w}$. Compared to the prior pdf $f(\boldsymbol{w})$, the posterior pdf $f(\boldsymbol{w} | \boldsymbol{x}_{1:N})$ describes the distribution of $\boldsymbol{w}$ after the data $\boldsymbol{x}_{1:N}$ has been observed.

For a BMM defined by (2.4), and for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ conditionally i.i.d. and individually dis-

tributed according to (2.1), the prior is given by

$$f(\boldsymbol{w}) = f(\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) = f(\boldsymbol{\pi}) \prod_{k=1}^{K} f(\boldsymbol{\theta}^*_k), \tag{2.36}$$

where (2.3) and (2.4b) were used. Furthermore, by using Bayes' rule, the posterior is obtained as

$$f(\boldsymbol{w}|\boldsymbol{x}_{1:N}) = f(\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}|\boldsymbol{x}_{1:N}) = \frac{f(\boldsymbol{x}_{1:N}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) f(\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K})}{f(\boldsymbol{x}_{1:N})}. \tag{2.37}$$

Here, the constant $f(\boldsymbol{x}_{1:N})$ is known as the evidence, and

$$f(\boldsymbol{x}_{1:N}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) = \prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:K}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{\pi}_k f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_k) \tag{2.38}$$

represents a likelihood function called the mixture likelihood. Inserting (2.38) and (2.36) into (2.37) yields

$$f(\boldsymbol{\theta}^*_{1:K}, \boldsymbol{\pi}|\boldsymbol{x}_{1:N}) \propto \left(\prod_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{\pi}_k f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_k) \right) f(\boldsymbol{\pi}) \prod_{k=1}^{K} f(\boldsymbol{\theta}^*_k). \tag{2.39}$$

For a given posterior distribution (and loss function), various Bayesian estimators can be obtained, the two most popular being the minimum mean square error (MMSE) estimator and the maximum a-posteriori (MAP) estimator. The MMSE estimator is equal to the posterior expectation or mean of $\boldsymbol{w}$, i.e., the expectation of $\boldsymbol{w}$ with respect to the posterior:

$$\hat{\boldsymbol{w}}_{\mathrm{MMSE}}(\boldsymbol{x}_{1:N}) = \mathrm{E}^{(f(\boldsymbol{w}|\boldsymbol{x}_{1:N}))}\{\boldsymbol{w}\} = \int_{\mathbb{R}^P} \boldsymbol{w} f(\boldsymbol{w}|\boldsymbol{x}_{1:N}) \, \mathrm{d}\boldsymbol{w}. \tag{2.40}$$

The MAP estimator is the posterior mode, i.e., the position of the global maximum of the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x}_{1:N})$:

$$\hat{\boldsymbol{w}}_{\mathrm{MAP}}(\boldsymbol{x}) = \underset{\boldsymbol{w} \in \mathbb{R}^P}{\arg\max} \, f(\boldsymbol{w}|\boldsymbol{x}_{1:N}). \tag{2.41}$$

Since any positive constant factor will not change the position of the maximum, the MAP estimator is also obtained by

$$\hat{\boldsymbol{w}}_{\mathrm{MAP}}(\boldsymbol{x}) = \underset{\boldsymbol{w} \in \mathbb{R}^P}{\arg\max} \{ f(\boldsymbol{x}_{1:N}|\boldsymbol{w}) f(\boldsymbol{w}) \}, \tag{2.42}$$

where we used (2.37) (with $\boldsymbol{w}$ in place of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*_{1:K}$) and the fact that $f(\boldsymbol{x}_{1:N})$ is a positive constant with respect to the parameter $\boldsymbol{w}$.

Even though explicit derivations of the Bayesian estimators considered above may be formally available when using conjugate priors, the representation of the mixture distribution given by (2.1) does not allow for an efficient use of Bayesian estimators because the corresponding posterior involves the expansion of the mixture likelihood (2.38) into K^N terms [40], [41]. The computational difficulty is seen to increase at an exponential rate with the sample size N. Thus,

a Bayesian estimator is often computationally prohibitive for more than a few observations $\boldsymbol{x}_n$. To circumvent the problem of an intractable posterior, we resort to approximation techniques. The two most prominent are Monte Carlo (MC) sampling, particularly, Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling [5], and variational inference (VI) methods such as coordinate ascent variational inference [14]. While MCMC methods provide a numerical and stochastic approximation of the exact posterior through a set of samples, VI methods provide a locally-optimal, analytical solution to a deterministic approximation of the posterior. MCMC methods are able to asymptotically approximate the posterior with arbitrary accuracy, but they tend to be more computationally expensive than VI methods and do not scale easily to high-dimensional models. In contrast, VI methods may suffer from oversimplified posterior approximations, but they are usually faster than MC sampling methods and thus make Bayesian inference computationally efficient and scalable to large-scale applications. VI methods aim at calculating an approximated posterior pdf/pmf $q(\boldsymbol{w}) \approx f(\boldsymbol{w}|\boldsymbol{x}_{1:N})$, which is commonly referred to as the variational distribution. This variational distribution can then be used to approximate the MMSE or MAP estimator by replacing $f(\boldsymbol{w}|\boldsymbol{x}_{1:N})$ with $q(\boldsymbol{w})$ in (2.40) or (2.41), respectively. In this thesis, we focus on VI methods.

3 Dirichlet Process Mixture Models

This chapter gives an introduction to the Dirichlet process mixture (DPM). The DPM is a Bayesian nonparametric mixture model, i.e., a mixture model similar to what we discussed in Section 2.1. "Nonparametric" means that the mixture model has an infinite number of parameters, and therefore an infinite number of components. This is a result of using the Dirichlet process (DP), which is also a Bayesian nonparametric model, as a prior distribution for the mixing distribution in a mixture model. In what follows, we first present the DP and then demonstrate how finite mixture models can be generalized to infinite mixture models, resulting in the DPM.

3.1 Stick-breaking Representation of the Dirichlet Distribution

As the DP can be viewed as a generalization of the Dirichlet distribution and the Dirichlet distribution can be viewed as a generalization of the beta distribution, we will first discuss an alternative representation of the Dirichlet and beta distributions. Both distributions are examples of distributions over distributions, which are suitable to model the random behavior of a finite set of proportions. Because of this interpretation and the conjugacy to the binomial and categorical distribution, they are often used as a prior distribution for the mixing proportions $\boldsymbol{\pi}$ in mixture models [42]. We will now describe both distributions in the context of modeling the random proportions $\boldsymbol{\pi} \in \Delta_K$.

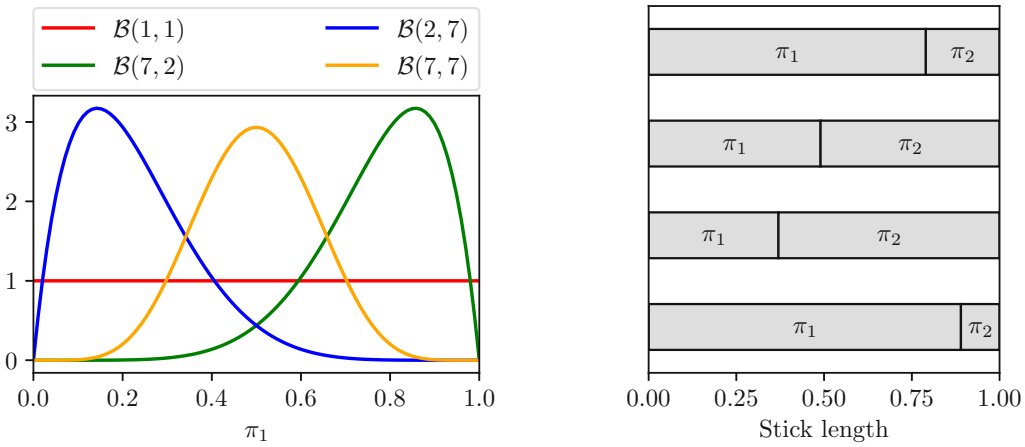
3.1.1 Beta Distribution

The beta distribution is given by

$$f(\pi) = \mathcal{B}(\pi; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1 - 1}(1 - \pi)^{\alpha_2 - 1}, \quad \pi \in [0, 1], \tag{3.1}$$

where $\Gamma(x)$ is the gamma function and $\alpha_1 > 0$ and $\alpha_2 > 0$ are scalar parameters that control the shape of the distribution. Figure 3.1a shows examples of (3.1) for different shape parameters. We note that the support of the beta distribution is $[0, 1]$ and for $\alpha_1 = \alpha_2 = 1$, the uniform distribution on $[0, 1]$ is obtained. Small values of α_1 and large values of α_2 result in higher probabilities for low values of π and vice versa. If $\alpha_1 = \alpha_2$, then for larger values of $\alpha_1 = \alpha_2$ more probability mass is concentrated around $\pi = 0.5$. The mean of the beta distribution is given by

$$\mathrm{E}^{(f(\pi))}\{\pi\} = \frac{\alpha_1}{\alpha_1 + \alpha_2} > 0. \tag{3.2}$$



(a) (b)

Figure 3.1: (a) Beta distribution for four different parameter vectors $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2)^{\mathrm{T}}$. (b) Visualization of four realizations of the beta distribution as breaking a unit-length stick into $K = 2$ pieces using the distribution $\mathcal{B}(1,1)$ in (a) for π_1 and setting $\pi_2 = 1 - \pi_1$.

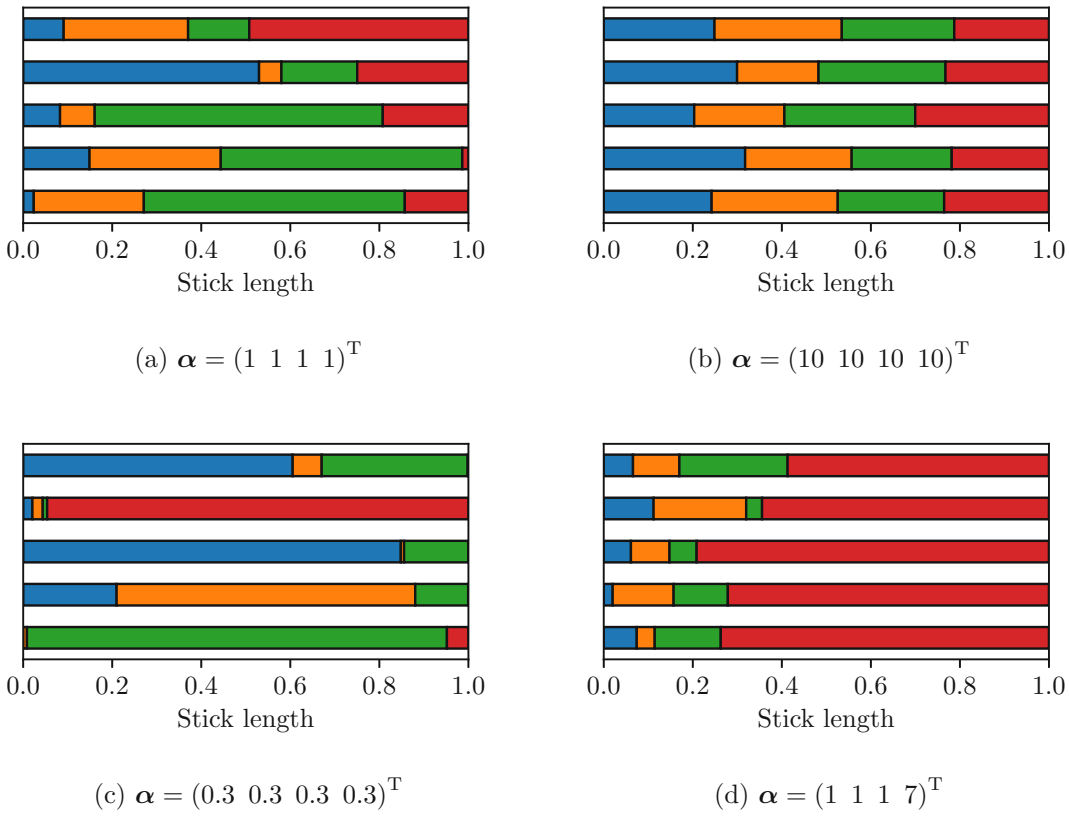
We now consider a mixture model with $K = 2$ components and mixing probabilities $\boldsymbol{\pi} = (\pi_1 \ \pi_2)^{\mathrm{T}} = (\pi_1 \ 1 - \pi_1)^{\mathrm{T}}$, and we use the beta distribution given by (3.1) as a prior for π_1. Figure 3.1b shows four different realizations from $\mathcal{B}(\pi_1; 1, 1)$. Each realization results in two mixing proportions π_1 and π_2 with $\pi_1 + \pi_2 = 1$. This can be interpreted as breaking a unit-length stick into $K = 2$ pieces where the first piece has an average length according to (3.2). Drawing conditionally i.i.d. data from a mixture distribution given $\boldsymbol{\pi}$, where $\boldsymbol{\pi}$ corresponds to a broken stick, means that each new data point is realized either from component 1 with probability π_1 or from component 2 with probability π_2, i.e., we will observe two clusters in the data set. Note that for $\alpha_1 = \alpha_2$, the means of π_1 and of $\pi_2 = 1 - \pi_1$ (see (3.2)) are equal.

3.1.2 Dirichlet Distribution

The multivariate generalization of the beta distribution is provided by the Dirichlet distribution. Consider a mixture model with $K \geq 2$ components where the random vector $\boldsymbol{\pi} = (\pi_1 \ \ldots \ \pi_K)^{\mathrm{T}}$ exists in the $(K-1)$-dimensional probability simplex Δ_K given by (2.2). The Dirichlet distribution, i.e., the pdf of $\boldsymbol{\pi}$ is given by

$$f(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}, \quad \boldsymbol{\pi} \in \Delta_K, \tag{3.3}$$

with parameter vector $\boldsymbol{\alpha} = (\alpha_1 \ \ldots \ \alpha_K)^{\mathrm{T}} \in \mathbb{R}^K$, $\alpha_k > 0$. Comparing (3.1) and (3.3) shows that for the case $K = 2$, the Dirichlet distribution reduces to the beta distribution, i.e., if $\pi_1 \sim \mathcal{B}(\pi_1; \alpha_1, \alpha_2)$ and $\pi_2 = 1 - \pi_1$, then $\boldsymbol{\pi} = (\pi_1 \ \pi_2)^{\mathrm{T}} \sim \mathcal{D}(\boldsymbol{\pi}; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2)^{\mathrm{T}}$, and vice versa. Similar to the beta distribution, the parameter vector $\boldsymbol{\alpha}$ controls the shape of the distribution, i.e., how concentrated the probability density is in certain areas of the support Δ_K. When $\boldsymbol{\alpha} = c\mathbf{1}_K$, the probability mass concentrates more and more around the center of the support of the distribution as $c \to \infty$, which means the variability of the length of the stick



(a) $\boldsymbol{\alpha} = (1\ \ 1\ \ 1\ \ 1)^{\mathrm{T}}$

(b) $\boldsymbol{\alpha} = (10\ \ 10\ \ 10\ \ 10)^{\mathrm{T}}$

(c) $\boldsymbol{\alpha} = (0.3\ \ 0.3\ \ 0.3\ \ 0.3)^{\mathrm{T}}$

(d) $\boldsymbol{\alpha} = (1\ \ 1\ \ 1\ \ 7)^{\mathrm{T}}$

Figure 3.2: Visualization of realizations of the Dirichlet distribution as the result of breaking a unit-length stick into $K = 4$ pieces, for four different choices of the parameter vector $\boldsymbol{\alpha}$. The stick pieces are represented by different colors. For each parameter vector $\boldsymbol{\alpha}$, five realizations are shown.

pieces decreases. For $c = 1$ the uniform distribution over Δ_K is obtained. Figure 3.2 shows the effect of varying $\boldsymbol{\alpha}$ for $K = 4$ in the stick-breaking representation.

The mean length of the stick pieces is given by

$$\mathrm{E}^{(f(\boldsymbol{\pi}))}\{\boldsymbol{\pi}\} = \frac{\boldsymbol{\alpha}}{\sum_{k=1}^{K} \alpha_k}. \tag{3.4}$$

Note that for $\boldsymbol{\alpha} = c\mathbf{1}_K$, the mean length (3.4) of all stick pieces are equal. The marginal distributions of $\mathcal{D}(\boldsymbol{\pi}; \boldsymbol{\alpha})$ can be shown [43] to be given by

$$\pi_k \sim \mathcal{B}\left(\pi_k; \alpha_k, \sum_{i \neq k} \alpha_k\right). \tag{3.5}$$

Furthermore, because of the neutrality property of the Dirichlet distribution [43] we have

$$\boldsymbol{\pi}_{k:K} | \, \boldsymbol{\pi}_{1:(k-1)} \sim \left(1 - \sum_{i=1}^{k-1} \pi_i\right) \mathcal{D}(\boldsymbol{\pi}_{k:K}; \boldsymbol{\alpha}_{k:K}). \tag{3.6}$$

Using (3.5) and (3.6), the stick-breaking behavior of the Dirichlet distribution can be reformulated in a recursive way using auxiliary variables $v_k \in [0, 1]$ as follows. Assume we have a unit-length stick and want to randomly break it into K parts of length $\pi_1, \ldots, \pi_K$ according to

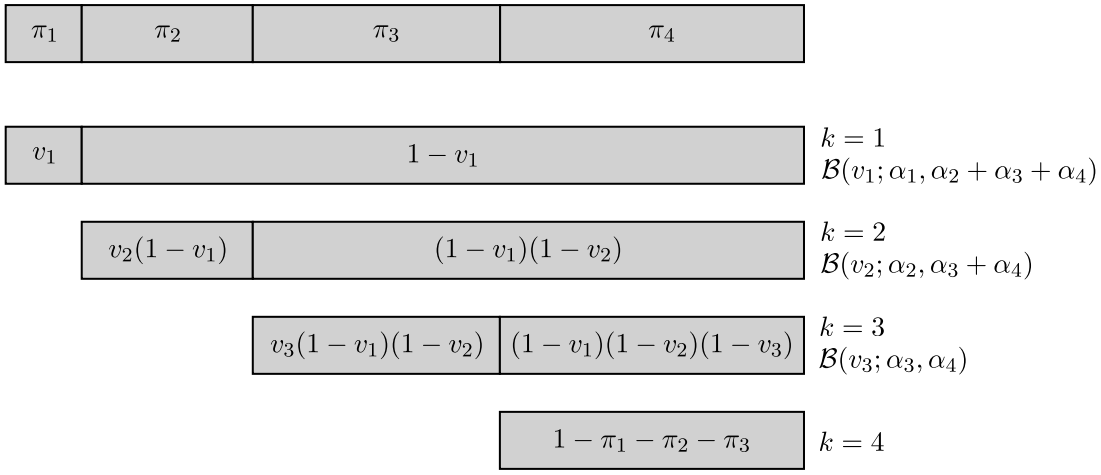


Figure 3.3: Construction of a $\mathcal{D}(\boldsymbol{\pi}; \boldsymbol{\alpha})$-distributed random vector $\boldsymbol{\pi}$ by iteratively breaking a unit-length stick into $K = 4$ pieces. Beta-distributed auxiliary variables v_k are used to break the currently remaining stick in the steps $k = 1, \ldots, K-1$ into two pieces.

(3.3). We will first assume that $K = 4$ and then generalize the overall procedure to any value $K \geq 3$. Figure 3.3 shows a visual representation of the stick-breaking approach of generating a random vector $\boldsymbol{\pi} = (\pi_1 \ \pi_2 \ \pi_3 \ \pi_4)^{\mathrm{T}}$ according to the distribution $\mathcal{D}(\boldsymbol{\pi}; \boldsymbol{\alpha})$. It consists of following steps:

1. First, we break off a piece of length $\pi_1 = v_1$ according to the marginal distribution $v_1 \sim \mathcal{B}\left(v_1; \alpha_1, \sum_{i=2}^{4} \alpha_i\right)$ (see (3.5)). The remaining stick has length $1 - \pi_1 = 1 - v_1$.

2. Using (3.6) the pieces π_2, π_3 and π_4 of the remaining stick are distributed according to $\boldsymbol{\pi}_{2:4} | \, \pi_1 \sim (1 - \pi_1)\mathcal{D}(\boldsymbol{\pi}_{2:4}; \boldsymbol{\alpha}_{2:4})$. Here, we can apply (3.5) to obtain the conditional distribution of π_2 given π_1, i.e., $\pi_2 | \, \pi_1 \sim (1 - \pi_1)\mathcal{B}(v_2; \alpha_2, \alpha_3 + \alpha_4)$, where we use the auxiliary variable v_2. Thus, to obtain π_2 we have to draw v_2 from $\mathcal{B}(v_2; \alpha_2, \alpha_3 + \alpha_4)$ and calculate $\pi_2 = v_2(1 - \pi_1) = v_2(1 - v_1)$. By breaking off the piece of length π_2 the remaining stick has length $1 - \pi_1 - \pi_2 = (1 - v_1)(1 - v_2)$.

3. Similar to the previous step we have $\pi_3, \pi_4 | \, \pi_1, \pi_2 \sim (1 - \pi_1 - \pi_2)\mathcal{D}(\pi_3, \pi_4; \alpha_3, \alpha_4)$ and $\pi_3 | \, \pi_1, \pi_2 \sim (1 - \pi_1 - \pi_2)\mathcal{B}(v_3; \alpha_3, \alpha_4)$. Here, π_3 is obtained by drawing the auxiliary v_3 from $\mathcal{B}(v_3; \alpha_3, \alpha_4)$ and calculating $\pi_3 = v_3(1 - \pi_1 - \pi_2) = v_3(1 - v_1)(1 - v_2)$.

4. The length of the last piece is $\pi_4 = 1 - \pi_1 - \pi_2 - \pi_3 = (1 - v_1)(1 - v_2)(1 - v_3)$.

Note that in each step $k = 1, \ldots, K - 1$ we draw an auxiliary variable v_k from the distribution $\mathcal{B}\left(v_k; \alpha_k, \sum_{i=k+1}^{K} \alpha_i\right)$ and use it to break the currently remaining stick of length[1]

$$1 - \sum_{i=1}^{k-1} \pi_i = \prod_{i=1}^{k-1} (1 - v_i) \tag{3.7}$$

to obtain π_k (see Figure 3.3). For $K \geq 3$ the overall sick-breaking approach can be summarized as follows [43]:

[1] For a mathematical proof of the equality $1 - \sum_{i=1}^{k-1} \pi_i = \prod_{i=1}^{k-1} (1 - v_i)$ see the appendix of [8].

1. Draw $v_1 \sim \mathcal{B}\left(v_1; \alpha_1, \sum_{k=2}^{K} \alpha_k\right)$ and set $\pi_1 = v_1$, which is the first piece of the stick. The remaining piece has length $1 - \pi_1 = 1 - v_1$.

2. For $2 \leq k \leq K-1$, if $k-1$ pieces with lengths $\pi_1, \ldots, \pi_{k-1}$ have been broken off using the auxiliary variables $v_1, \ldots, v_{k-1}$, then $1 - \sum_{i=1}^{k-1} \pi_i = \prod_{i=1}^{k-1}(1 - v_i)$ is the length of the remaining stick. Draw $v_k \sim \mathcal{B}\left(v_k; \alpha_k, \sum_{i=k+1}^{K} \alpha_i\right)$ and set $\pi_k = v_k\left(1 - \sum_{i=1}^{k-1} \pi_i\right) = v_k \prod_{i=1}^{k-1}(1 - v_i)$.

3. The length of the last piece is $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k = \prod_{i=1}^{K-1}(1 - v_i)$.

Next, we will discuss the generation of a sequence of $K = \infty$ mixing proportions $\pi_1, \pi_2, \ldots$ that sum to one, which constitutes an infinite-dimensional generalization of the stick-breaking interpretation of the Dirichlet distribution.

3.2 Dirichlet Process

The DP has been a centerpiece of Bayesian nonparametrics since its introduction in [44] as a random probability measure. In this section, we will consider different representations of the DP, called the stick-breaking process [11], the Blackwell-MacQueen Urn Scheme [12] and the Chinese restaurant process [13], which all have the advantage of not requiring measure theory in their formulations. These representations of the DP will build the basis for the inference technique and simulations considered in Chapters 4 and 5, respectively.

Following [11], we define realizations of a random pdf G to be of the form

$$G(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:\infty}) \triangleq \sum_{k=1}^{\infty} \pi_k \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*_k), \tag{3.8}$$

with random proportions $\pi_k \in [0, 1]$ and random vectors $\boldsymbol{\theta}^*_k \in \mathbb{R}^p$, where the random proportions satisfy $\sum_{k=1}^{\infty} \pi_k = 1$ almost surely. Note that $G(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:\infty})$ is a discrete distribution, i.e., $\boldsymbol{\theta} = \boldsymbol{\theta}^*_k$ with probability π_k or, equivalently, $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}^*_k \mid \boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:\infty}) = \pi_k$ (cf. Section 2.1.3). The DP is a Bayesian nonparametric model capable of generating realizations of the form (3.8) and, therefore, is a distribution over distributions, as (3.8) is a valid probability distribution. We will denote it as $\mathcal{DP}(\alpha, G_0)$, where α is called the concentration parameter and G_0 the base distribution, and write

$$G \sim \mathcal{DP}(G; \alpha, G_0)$$

to express the fact that a random pdf G is distributed according to a DP. The two parameters α and G_0, which are involved in the generation of the proportions π_k and vectors $\boldsymbol{\theta}^*_k$, respectively, will be discussed in the following.

3.2.1 Stick-breaking Process

Following [11], we now present a constructive way of forming (3.8), which is commonly referred to as the stick-breaking process [45] and relates to the stick-breaking representation of the Dirichlet distribution described above. Using a sequence of auxiliary variables $v_k \in [0, 1]$, $k \in \mathbb{N}$, and

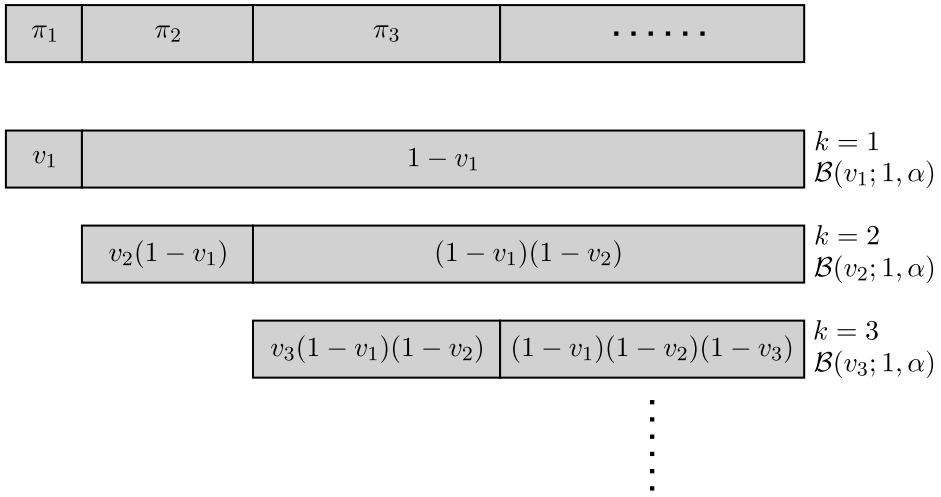


Figure 3.4: Stick-breaking representation of the GEM distribution. Compared to the stick-breaking representation of the Dirichlet distribution in Figure 3.3, the unit-length stick is broken into an infinite number of pieces using $\mathcal{B}(1, \alpha)$-distributed auxiliary variables v_k.

a real-valued parameter $\alpha > 0$, called the concentration parameter, the infinite sequence of (mixing) proportions π_k, $k = 1, 2, \ldots$, is given by (cf. (3.7))

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i) = v_k \left(1 - \sum_{i=1}^{k-1} \pi_i\right), \tag{3.9a}$$

with

$$v_k \overset{\text{i.i.d.}}{\sim} \mathcal{B}(v_k; 1, \alpha). \tag{3.9b}$$

Note that the k-th proportion π_k depends on the first k auxiliary variables $v_1, \ldots, v_k$. Compared to the stick-breaking representation of the Dirichlet distribution, a beta distribution with fixed parameters 1 and α is used to generate the sequence auxiliary variables v_k. It can be shown that $\sum_{k=1}^{\infty} \pi_k = 1$ almost surely [11], i.e., the support of $\boldsymbol{\pi} = (\pi_1 \;\; \pi_2 \;\; \cdots)^{\mathrm{T}}$ is the infinite-dimensional probability simplex. The resulting distribution of the infinite sequence of proportions π_k, or, equivalently, of the infinite-dimensional vector $\boldsymbol{\pi}$, is called the GEM distribution, named after Griffiths, Engen and McCloskey [35]. We denote it by

$$\boldsymbol{\pi} \sim \mathrm{GEM}(\boldsymbol{\pi}; \alpha). \tag{3.10}$$

In Section 3.3, we will show that using (3.10) as a prior for the mixing probabilities $\boldsymbol{\pi}$ of a mixture model represented by (2.7) (with $K = \infty$), is equivalent to using $\mathcal{D}(\boldsymbol{\pi}; \frac{\alpha}{K} \mathbf{1}_K)$ as a prior for $\boldsymbol{\pi}$ and taking the limit $K \to \infty$. This then demonstrates how the Dirichlet distribution can be generalized to the Dirichlet Process. Figure 3.4 shows a representation of the stick-breaking process where a unit-length stick is broken apart into an infinite number of pieces of length π_k, for $k = 1, 2, \ldots$, according to (3.9a) using random auxiliary variables v_k distributed according to (3.9b).

Using (3.2), the mean of the auxiliary variable v_k is given by

$$\mathrm{E}^{(f(v_k))}\{v_k\} = \frac{1}{1+\alpha} > 0. \tag{3.11}$$

To find the mean of the mixing proportions π_k, we start with expression (3.9a), i.e.,

$$\mathrm{E}^{(f(v_1,\dots,v_k))}\{\pi_k\} = \mathrm{E}^{(f(v_1,\dots,v_k))}\left\{ v_k \prod_{i=1}^{k-1}(1-v_i) \right\}.$$

Using the independence between the auxiliary variables v_k (see (3.9b)) and the linearity of expectation yields

$$\begin{aligned}
\mathrm{E}^{(f(v_1,\dots,v_k))}\{\pi_k\} &= \mathrm{E}^{(f(v_k))}\{v_k\} \prod_{i=1}^{k-1}\left(1 - \mathrm{E}^{(f(v_i))}\{v_i\}\right) \\
&= \frac{1}{1+\alpha} \prod_{i=1}^{k-1}\left(1 - \frac{1}{1+\alpha}\right) \\
&= \frac{1}{1+\alpha} \prod_{i=1}^{k-1} \frac{\alpha}{1+\alpha} \\
&= \frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^{k-1},
\end{aligned} \tag{3.12}$$

where (3.11) was used. Because $\alpha/(1+\alpha) < 1$, we conclude from (3.12) that $\mathrm{E}^{(f(v_1,\dots,v_k))}\{\pi_k\}$ decreases with growing k and ultimately tends to zero, i.e., the average mixing proportions are ordered in decreasing size with respect to k. For small values of α, the first few proportions π_k are likely assigned the majority of the stick mass while the rest of the proportions π_k are approximately zero. The reason is the first few terms of the sequence generated by (3.12) dominate for small α. As α grows the average proportions (3.12) become smaller, resulting in the majority of the stick mass being assigned to an increasing number of π_k's. Figure 3.5a shows a plot of (3.12) for $k = 1, \dots, 10$ and illustrates this behavior. In Figure 3.5b we show four different realizations of mixing proportions $\boldsymbol{\pi}$ sampled according to (3.10).

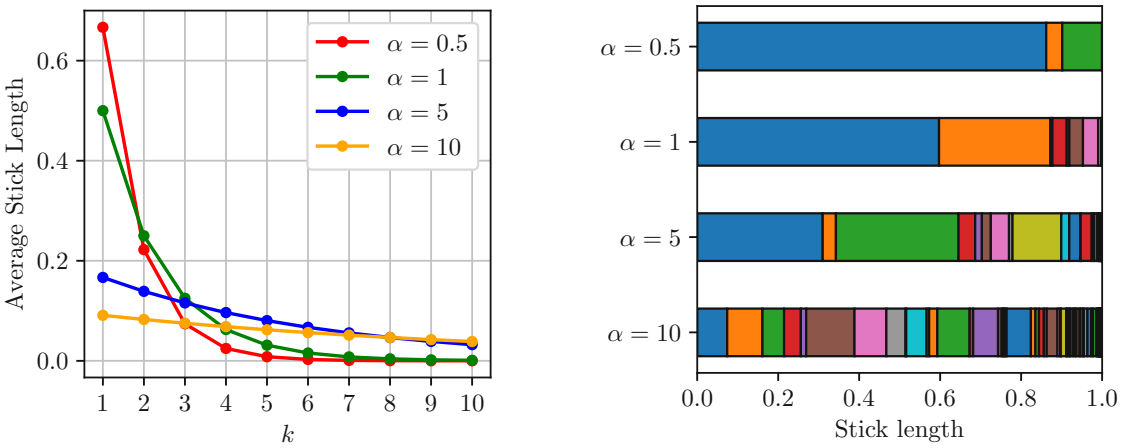
Next, in addition to $\boldsymbol{\pi}$, we consider a sequence of random vectors $\boldsymbol{\theta}_k^* \in \mathbb{R}^p$ that are i.i.d. and individually distributed according to a base distribution G_0, i.e.,

$$\boldsymbol{\theta}_k^* \overset{\text{i.i.d.}}{\sim} G_0(\boldsymbol{\theta}_k^*), \quad k \in \mathbb{N}. \tag{3.13}$$

Pairing together $\boldsymbol{\theta}_k^*$ and the GEM$(\boldsymbol{\pi}; \alpha)$-distributed proportions π_k for $k = 1, 2, \dots$, the stick-breaking construction of (3.8) can be described as

$$\begin{aligned}
&\boldsymbol{\pi} \sim \mathrm{GEM}(\boldsymbol{\pi}; \alpha), \\
&\boldsymbol{\theta}_k^* \overset{\text{i.i.d.}}{\sim} G_0(\boldsymbol{\theta}_k^*), \\
&G(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\theta}_{1:\infty}^*) = \sum_{k=1}^{\infty} \pi_k \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*),
\end{aligned}$$

where each segment π_k of a unit length stick is associated with a random vector $\boldsymbol{\theta}_k^* \in \mathbb{R}^p$. This



(a) (b)

Figure 3.5: (a) Mean of the mixing proportions π_k (average stick lengths) for $k = 1, \ldots, 10$ and $\alpha \in \{0.5, 1, 5, 10\}$ (see (3.12)). For each value of α, a realization of $\boldsymbol{\pi} \sim \mathrm{GEM}(\boldsymbol{\pi}; \alpha)$ is visualized in (b) where the length of each colored stick corresponds to one value of π_k.

construction of G guarantees that $G \sim \mathcal{DP}(G; \alpha, G_0)$ and makes clear that samples from a DP are discrete distributions represented by an infinite sum of weighted Dirac delta functions. Figure 3.6 visualizes realizations of the DP with $p = 1$, i.e., $\theta_k^* \in \mathbb{R}$, and a standard normal base distribution $G_0 = \mathcal{N}(0, 1)$ for four different values of the concentration parameter α. As $\alpha \to \infty$, the DP realizations G approximate the base distribution G_0 by a densely sampled set of discrete values $\boldsymbol{\theta}_k^*$ due to $\boldsymbol{\pi}$ consisting of small, roughly uniform, proportions π_k (cf. the discussion of Figure 3.5).

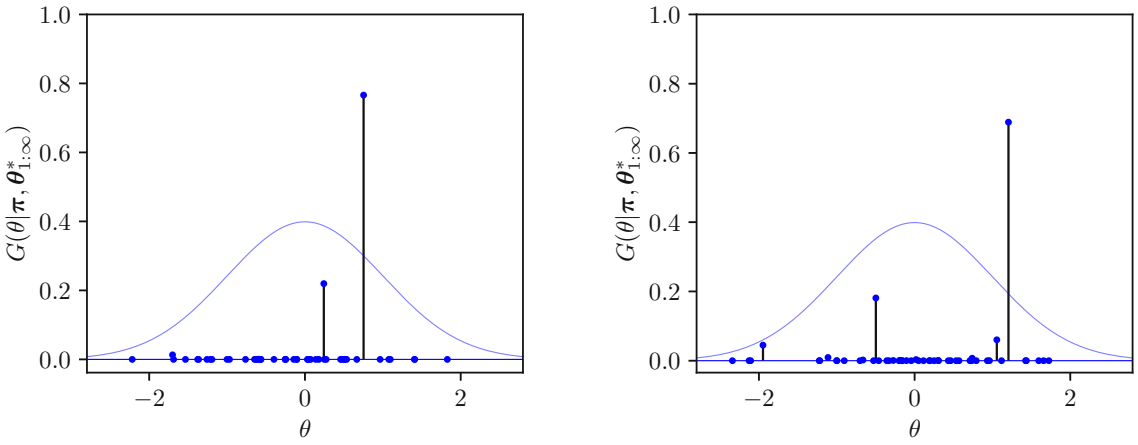
Finally, given a realization G of $\mathcal{DP}(G; G_0, \alpha)$, we can consider N samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$, i.i.d. and individually distributed according to this realization, i.e.,

$$\boldsymbol{\theta}_n | G \overset{\text{i.i.d.}}{\sim} G(\boldsymbol{\theta}_n | \boldsymbol{\pi}, \boldsymbol{\theta}_{1:\infty}^*), \quad n = 1, \ldots, N. \tag{3.14}$$

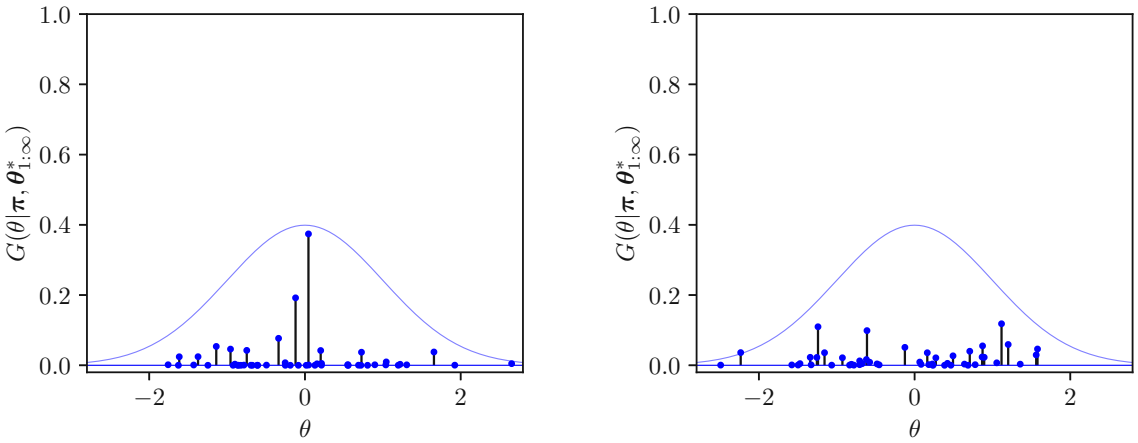
This means that each sample $\boldsymbol{\theta}_n$ is equal to $\boldsymbol{\theta}_k^*$ with probability π_k (note that G consists of a countably infinite number of possible vectors $\boldsymbol{\theta}_k^*$, see (3.8)). In a conventional sampling scheme, this is not feasible since we are required to maintain an infinite number of values for the proportions π_k and vectors $\boldsymbol{\theta}_k^*$ in order to obtain a sample $\boldsymbol{\theta}_n$. However, we can circumvent this requirement and develop a feasible sampling method by using a generalized Pólya urn scheme – discussed next – to directly produce $\boldsymbol{\theta}_n$. Under such an approach, and with a slight abuse of language, we obtain $\boldsymbol{\theta}_{1:N} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_N^{\mathrm{T}})^{\mathrm{T}}$, which we consider as N samples from the Dirichlet process.

3.2.2 Pólya Urn Process

According to (3.14), a realization G of the DP assigns samples $\boldsymbol{\theta}_n$ to distinct values $\boldsymbol{\theta}_k^*$, i.e., $\boldsymbol{\theta}_n = \boldsymbol{\theta}_k^*$ with probability π_k. Therefore, it is possible that multiple samples $\boldsymbol{\theta}_n$ have identical values. For a given N, the samples $\boldsymbol{\theta}_n$ for $n = 1, \ldots, N$ will take on $L \leq N$ distinct values



(a) $\alpha = 0.5$ (b) $\alpha = 1$



(c) $\alpha = 5$ (d) $\alpha = 10$

Figure 3.6: Draws from a Dirichlet process for $p = 1$, base distribution $G_0 = \mathcal{N}(0, 1)$, and $\alpha \in \{0.5, 1, 5, 10\}$ visualized by a stem plot of (3.8). Additionally, the base distribution $G_0(\theta^*_k)$ is shown. The black stem lines can be interpreted as sticks from the stick-breaking process where the height represents the values π_k and the sticks itself are placed at the one dimensional positions θ^*_k. Shown are only the values π_k and θ^*_k for $k = 1, \ldots, 50$.

$\boldsymbol{\theta}'_l \in \{\boldsymbol{\theta}^*_1, \boldsymbol{\theta}^*_2, \ldots\}$, $l = 1, \ldots, L$. Let

$$N_l = \sum_{n=1}^{N} \mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\theta}'_l) \tag{3.15}$$

denote the number of times $\boldsymbol{\theta}'_l$ appears within the sequence $\boldsymbol{\theta}_{1:N}$ (recall the discussion of (2.18)). The probability of a new sample $\boldsymbol{\theta}_{N+1}$ taking one of the values $\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_L$ or a new value $\boldsymbol{\theta}'_{L+1}$ is given by a conditional distribution for $\boldsymbol{\theta}_{N+1}$ given $\boldsymbol{\theta}_{1:N}$. This probability was derived in [12]; it is given by

$$f(\boldsymbol{\theta}_{N+1}|\boldsymbol{\theta}_{1:N}) = \frac{\alpha}{\alpha + N} G_0(\boldsymbol{\theta}_{N+1}) + \frac{1}{\alpha + N} \sum_{l=1}^{L} N_l \delta(\boldsymbol{\theta}_{N+1} - \boldsymbol{\theta}'_l). \tag{3.16}$$

Thus, the DP leads to a closed-form predictive distribution (3.16) which can be evaluated using the numbers $N_1, N_2, \ldots, N_L$ and the L distinct values $\boldsymbol{\theta}'_{1:L} = (\boldsymbol{\theta}'^{\mathrm{T}}_1 \ \cdots \ \boldsymbol{\theta}'^{\mathrm{T}}_L)^{\mathrm{T}}$ appearing within the sequence $\boldsymbol{\theta}_{1:N}$. Note that (3.16) is the sum of a continuous distribution given by the base distribution $G_0(\boldsymbol{\theta})$ weighted by $\frac{\alpha}{\alpha+N}$ and a discrete part consisting of L discrete components at positions $\boldsymbol{\theta}'_l$ with weights $\frac{N_l}{\alpha+N}$. From that it follows that $\boldsymbol{\theta}_{N+1} = \boldsymbol{\theta}'_l$ with probability $\frac{N_l}{\alpha+N}$ or $\boldsymbol{\theta}_{N+1}$ is realized from the base distribution G_0 with probability $\frac{\alpha}{\alpha+N}$, resulting in a new value $\boldsymbol{\theta}'_{L+1}$.

This generative process of successively creating new samples by drawing from (3.16) can be interpreted in terms of the following generalized Pólya urn model [12]. Consider a container (an urn) containing one black ball whose mass is proportional to α and one colored ball for each preceding sample $\boldsymbol{\theta}_n$, where the color uniquely corresponds to the associated $\boldsymbol{\theta}'_l$. Furthermore, interpret the sampling of $\boldsymbol{\theta}_{N+1}$ from $f(\boldsymbol{\theta}_{N+1}|\boldsymbol{\theta}_{1:N})$ in (3.16) as drawing a ball from the urn and putting it back together with a new ball. Each colored ball we draw corresponds to a realization of the discrete part of (3.16). Here, the new ball we add to the urn is of the same color, meaning that we increase the number of balls of that color by one, so that it becomes $N_l + 1$. Drawing the black ball means realizing $\boldsymbol{\theta}_{N+1}$ from the base distribution G_0. Here, the new ball we add to the urn has a new, previously unseen color, corresponding to $\boldsymbol{\theta}'_{L+1}$. The probability of drawing the black ball is assumed to be proportional to its weight α. Thus, the number of uniquely colored balls increases with increasing α. Conversely, for small values of α, it is more likely to draw a colored ball from the urn and return it together with a new ball of the same color. As a consequence, the number of uniquely colored balls is more likely to be small.

The urn scheme can be used to generate samples $\boldsymbol{\theta}_n$ from the DP without explicitly constructing the underlying pdf $G \sim \mathcal{DP}(G; \alpha, G_0)$ and can be summarized as:

1. Realize the first sample $\boldsymbol{\theta}_1$ from the base distribution G_0.
2. For $N = 1, 2, \ldots$, draw the next sample $\boldsymbol{\theta}_{N+1}$ from the predictive pdf in (3.16).

In the next subsection, we will present a different view of this process and emphasize the clustering property of the DP, i.e., the fact that multiple samples $\boldsymbol{\theta}_n$ take identical values $\boldsymbol{\theta}'_l$.

3.2.3 Clustering and Chinese Restaurant Process

Multiple samples $\boldsymbol{\theta}_n$ of the DP taking identical values $\boldsymbol{\theta}'_l$ indicates the existence of a clustering structure within the samples $\boldsymbol{\theta}_n$. We again consider N samples $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$ that take $L \leq N$ distinct values $\boldsymbol{\theta}'_l \in \{\boldsymbol{\theta}^*_1, \boldsymbol{\theta}^*_2, \ldots\}$, $l = 1, \ldots, L$, which means that we can group the $\boldsymbol{\theta}_n$ into L different clusters. Let $z_n \in \mathbb{N}$ be an indicator variable that indicates the cluster associated with the n-th sample, i.e., $\boldsymbol{\theta}_n = \boldsymbol{\theta}'_l$ if and only if $z_n = l$. We can then reformulate (3.15) as

$$N_l = \sum_{n=1}^{N} \mathbb{1}(z_n = l). \tag{3.17}$$

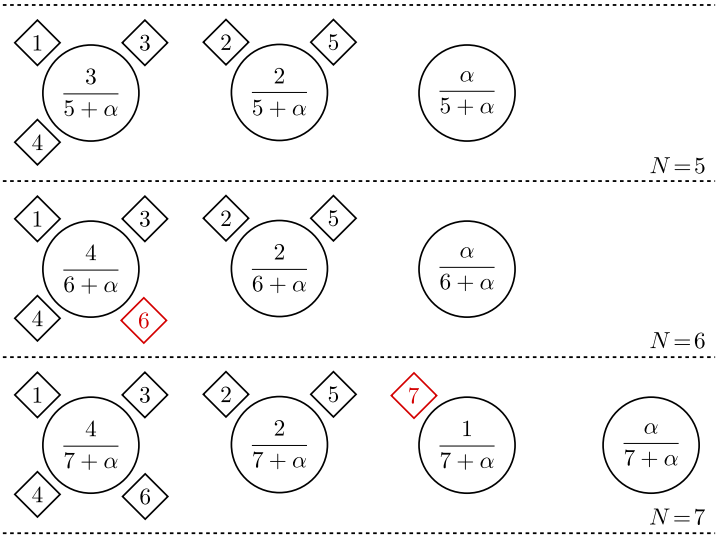


Figure 3.7: Chinese restaurant process with initially $N = 5$ customers sitting at two tables. The sixth customer joins the first table and the seventh customer joins a new table. The probability of a new customer joining a table (see (3.18)) is given by the values within the circles.

Moreover, based on (3.16), we can formulate a predictive pmf for the indicator variable z_{N+1} of the next sample $\boldsymbol{\theta}_{N+1}$ as

$$p(z_{N+1}|\boldsymbol{z}_{1:N}) = \frac{\alpha}{\alpha+N}\mathbb{1}(z_{N+1} = L+1) + \frac{1}{\alpha+N}\sum_{l=1}^{L} N_l \mathbb{1}(z_{N+1} = l). \tag{3.18}$$

The clustering of DP samples $\boldsymbol{\theta}_n$ implies a clustering of the respective indices $n \in \mathbb{N}$ and, thereby, a partitioning of $\mathbb{N}$. This partitioning is described by the indicator variables z_n and called the Chinese restaurant process (CRP) by analogy to the process of seating customers at tables in a restaurant [13]. Each customer (indicator variable z_n) joins an existing table (cluster), i.e., $z_n = l$, with probability $\frac{N_l}{\alpha+N}$ or sits at a new table with probability $\frac{\alpha}{\alpha+N}$. In the latter case a new cluster is created, i.e., $z_n = L+1$. Note that the probability of customer joining a table is proportional to the number N_l of customers already sitting at that table, and the probability of a customer sitting at a new table is proportional to α. The l-th table is associated with the l-th value $\boldsymbol{\theta}'_l$. Figure 3.7 shows an example restaurant with initially $N = 5$ customers sitting at two tables, illustrating the partitioning of five integers into two clusters. Additionally, the process of another customer joining an existing table ($N = 6$) as well as yet another customer joining a new table ($N = 7$) is shown. For a given number of customers N, the expected number of occupied tables (clusters) L can be shown [4] to be given by

$$\bar{L} = \alpha(\Psi(\alpha+N) - \Psi(\alpha)) \tag{3.19}$$

$$\approx \alpha \ln\left(1 + \frac{N}{\alpha}\right), \tag{3.20}$$

where $\Psi(\cdot)$ denotes the digamma function. The approximation in (3.20) is asymptotically exact for $N \to \infty$ and shows that for large N the expected number of tables $\bar{L}$ grows logarithmically with the number of customers N. Figure 3.8 shows the evolution of customers (indexed by

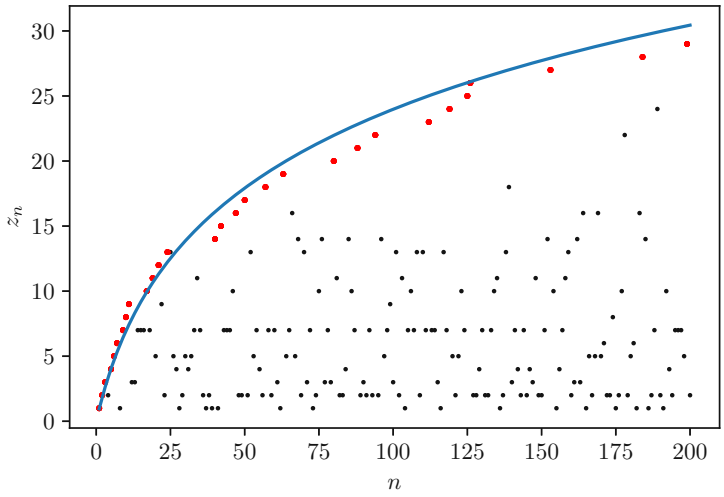


Figure 3.8: Sampling of the indicator variables z_n according to (3.18), which can be interpreted as customers joining tables. Black dots refer to samples of already existing indicator variables (at "time" n) and red dots refer to sampling new clusters. After $N = 200$ samples, there are $L = 29$ clusters. The blue curve shows the approximate expected number of occupied tables as given by (3.20).

n) joining specific tables (indicated by z_n) for $n = 1, \ldots, 200$ and $\alpha = 10$. At $n = 8$, the indicator variable $z_n = 1$ is equal to an already existing value, and at $n = 40$, the indicator variable $z_n = 14$ is equal to a new value. Note that the evolution of the indicator variables z_n corresponding to customers joining a new table (red dots) for increasing N roughly follows $\bar{L}$.

Similar to the generalized Pólya urn scheme, we can use the CRP to sample $\boldsymbol{\theta}_n$ from the DP via a generation of indicator variables z_n according to (3.18), without explicitly constructing the underlying pdf $G \sim \mathcal{DP}(G; \alpha, G_0)$. This CRP-based sampling procedure consists of following steps:

1. Assign the first indicator variable z_1 to an initial cluster by setting $z_1 = 1$.
2. For $N = 1, 2, \ldots$, realize the next indicator variable z_{N+1} from the predictive pmf in (3.18).
3. For each cluster $l \in \{z_1, z_2, \ldots\}$ draw $\boldsymbol{\theta}'_l$ from the base distribution G_0 (see (3.13)).
4. Assign the values $\boldsymbol{\theta}'_l$ to the samples $\boldsymbol{\theta}_n$ by using the indicator variables z_n, i.e., $\boldsymbol{\theta}_n = \boldsymbol{\theta}'_{z_n}$.

Compared to the generalized Pólya urn scheme, where we directly assign the values $\boldsymbol{\theta}'_l$ to the samples $\boldsymbol{\theta}_n$, the CRP describes which cluster l a sample $\boldsymbol{\theta}_n$ belongs to via indicator variables z_n, such that $z_n = l$ and $\boldsymbol{\theta}_n = \boldsymbol{\theta}'_{z_n}$. In the generalized Pólya urn scheme, α influences how likely it is to choose the black ball and therefore add a new color (cluster). In the CRP, α influences how likely it is that a customer decides to sit on a new table, where the number of tables refers to the number of clusters. The same behavior regarding α is exhibited in the stick-breaking process. A large value of α means that there are many small sticks (see Figure 3.5b), which is equivalent to an urn with a black ball of large mass (resulting in many differently colored balls) and a restaurant with customers spread out over many tables. Table 3.1 summarizes the equivalent descriptions of clustering in a DP.

	Clustering of	Cluster labels
Clustering	indicator variables	natural numbers
Generalized Pólya urn scheme	balls	ball colors
CRP	customers	tables

Table 3.1: Equivalences between different descriptions of the clustering property of the DP.

Finally, we address a property of the stick-breaking process that is related to the indicator variables z_n as follows. Imagine we are given a vector $\boldsymbol{\pi}$ of mixing proportions π_k, $k = 1, 2, \ldots$, that was realized from the stick-breaking process (3.9). We can then consider indicator variables z_n that are conditionally i.i.d. and individually distributed according to the categorical distribution

$$z_n|\boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n|\boldsymbol{\pi}). \tag{3.21}$$

Here, the independence of the indicator variables follows from the fact that we condition on $\boldsymbol{\pi}$, which is equivalent to the conditional independence of the samples $\boldsymbol{\theta}_n$ given a realization G from the Dirichlet process (see (3.14)). According to (3.12), the expectation of the mixing proportion π_k is decreasing with increasing k. In other words, on average, the probability of the event $z_n = k$ decreases with increasing k. Given an observed sequence of indicator variables $\boldsymbol{z}_{1:N}$, it is possible to determine the posterior distribution and the posterior expectation of $\boldsymbol{\pi}$, i.e., conditioned on $\boldsymbol{z}_{1:N}$. According to [25], the posterior distribution $f(\boldsymbol{\pi}|\boldsymbol{z}_{1:N})$ has again a stick-breaking representation given by (cf. (3.9))

$$\pi_k|\boldsymbol{z}_{1:N} = v_k \prod_{i=1}^{k-1}(1 - v_i), \tag{3.22a}$$

where

$$v_k|\boldsymbol{z}_{1:N} \overset{\text{i.i.d.}}{\sim} \mathcal{B}(v_k; a_k, b_k), \tag{3.22b}$$

for $k \in \mathbb{N}$, with

$$a_k = 1 + N_k, \tag{3.23a}$$

$$b_k = \alpha + \sum_{i=k+1}^{\infty} N_i, \tag{3.23b}$$

and N_k defined in (3.17). Comparing with the prior $\mathcal{B}(v_k; 1, \alpha)$ in (3.9b), we see that the posterior of v_k in (3.22b) is again a beta distribution but with updated parameters as given by (3.23). Applying (3.2) yields for the posterior mean of the auxiliary variables

$$\mathrm{E}^{(f(v_k|\boldsymbol{z}_{1:N}))}\{v_k\} = \frac{a_k}{a_k + b_k}. \tag{3.24}$$

By proceeding as in (3.12) with obvious modifications, the posterior expectation of the propor-

tion π_k is obtained as

$$\begin{aligned}\mathrm{E}^{(f(v_1,\dots,v_k|\boldsymbol{z}_{1:N}))}\{\pi_k\} &= \mathrm{E}^{(f(v_k|\boldsymbol{z}_{1:N}))}\{v_k\}\prod_{i=1}^{k-1}\left(1-\mathrm{E}^{(f(v_i|\boldsymbol{z}_{1:N}))}\{v_i\}\right)\\ &= \frac{a_k}{a_k+b_k}\prod_{i=1}^{k-1}\frac{b_i}{a_i+b_i}.\end{aligned} \tag{3.25}$$

Note that the prior-to-posterior conversion (3.9) → (3.22) is fully determined by the multiplicities N_k in (3.17), i.e., the numbers of times each value k appears in the sequence $\boldsymbol{z}_{1:N}$.

3.3 From Finite to Infinite Mixture Models

Based on [23] and [46], we now present a way of generalizing finite mixture models to infinite mixture models, which serves as a more formal explanation of the DP as an infinite dimensional generalization of the Dirichlet distribution.

3.3.1 Predictive Distribution for Finite Mixtures

Consider a Bayesian mixture model with K components as given by (2.7) using a symmetric Dirichlet prior on the K mixing proportions. In the symmetric Dirichlet distribution all elements of the parameter vector $\boldsymbol{\alpha}$ have the same value, which we assume to be α/K, i.e.,

$$\boldsymbol{\pi} \sim \mathcal{D}\left(\boldsymbol{\pi}; \frac{\alpha}{K}\mathbf{1}_K\right), \tag{3.26}$$

where according to (3.3)

$$\mathcal{D}\left(\boldsymbol{\pi}; \frac{\alpha}{K}\mathbf{1}_K\right) = \frac{\Gamma(\alpha)}{\Gamma\left(\frac{\alpha}{K}\right)^K}\prod_{k=1}^{K}\pi_k^{\frac{\alpha}{K}-1}. \tag{3.27}$$

Given mixing proportions $\boldsymbol{\pi}$, we draw N indicator variables z_n i.i.d. according to $\mathcal{C}(z_n|\boldsymbol{\pi})$ in (2.5). Thus, the conditional joint distribution of $\boldsymbol{z}_{1:N}$ given $\boldsymbol{\pi}$ becomes

$$p(\boldsymbol{z}_{1:N}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\mathcal{C}(z_n|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{\mathbb{1}(z_n=k)} = \prod_{k=1}^{K}\pi_k^{\sum_{n=1}^{N}\mathbb{1}(z_n=k)} = \prod_{k=1}^{K}\pi_k^{N_k}, \tag{3.28}$$

where we used (3.17) in the last step. Using (3.27) and (3.28), we can now integrate out the mixing proportions $\boldsymbol{\pi}$ to obtain the joint pmf of $\boldsymbol{z}_{1:N}$, i.e.,

$$\begin{aligned}p(\boldsymbol{z}_{1:N}) &= \int_{\Delta_K} f(\boldsymbol{z}_{1:N}|\boldsymbol{\pi})f(\boldsymbol{\pi})\,\mathrm{d}\boldsymbol{\pi}\\ &= \int_{\Delta_K}\left(\prod_{k=1}^{K}\pi_k^{N_k}\right)\frac{\Gamma(\alpha)}{\Gamma\left(\frac{\alpha}{K}\right)^K}\left(\prod_{l=1}^{K}\pi_l^{\frac{\alpha}{K}-1}\right)\mathrm{d}\boldsymbol{\pi}\\ &= \frac{\Gamma(\alpha)}{\Gamma\left(\frac{\alpha}{K}\right)^K}\int_{\Delta_K}\prod_{k=1}^{K}\pi_k^{N_k+\frac{\alpha}{k}-1}\,\mathrm{d}\pi_k.\end{aligned} \tag{3.29}$$

Here, the integration amounts to calculating the normalization constant of a Dirichlet distribution with parameters $\alpha'_k = N_k + \frac{\alpha}{k}$ (see (3.3)). According to (3.3) we have

$$\int_{\Delta_K} \mathcal{D}(\boldsymbol{\pi}; \boldsymbol{\alpha}') \, \mathrm{d}\boldsymbol{\pi} = \frac{\Gamma\left(\sum_{k=1}^K N_k + \frac{\alpha}{K}\right)}{\prod_{k=1}^K \Gamma(N_k + \frac{\alpha}{K})} \int_{\Delta_K} \prod_{k=1}^K \pi_k^{N_k + \frac{\alpha}{k} - 1} \, \mathrm{d}\pi_k = 1$$

and thus

$$\int_{\Delta_K} \prod_{k=1}^K \pi_k^{N_k + \frac{\alpha}{k} - 1} \, \mathrm{d}\pi_k = \frac{\prod_{k=1}^K \Gamma(N_k + \frac{\alpha}{K})}{\Gamma\left(\sum_{k=1}^K N_k + \frac{\alpha}{K}\right)}. \tag{3.30}$$

Inserting (3.30) into (3.29) yields

$$p(\boldsymbol{z}_{1:N}) = \frac{\Gamma(\alpha)}{\Gamma(\frac{\alpha}{K})^K} \frac{\prod_{k=1}^K \Gamma(N_k + \frac{\alpha}{K})}{\Gamma\left(\sum_{k=1}^K N_k + \frac{\alpha}{K}\right)} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^K \frac{\Gamma(N_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})}, \tag{3.31}$$

where we used the fact that $\sum_{k=1}^K N_k = N$.

Based on joint pmf for the indicator variables, we can now derive a predictive pmf for a single indicator variable z_{N+1} similar to (3.18). The conditional pmf for $z_{N+1}|\boldsymbol{z}_{1:N}$ can be expressed as

$$p(z_{N+1}|\boldsymbol{z}_{1:N}) = \frac{f(z_{N+1}, \boldsymbol{z}_{1:N})}{f(\boldsymbol{z}_{1:N})} = \frac{f(\boldsymbol{z}_{1:(N+1)})}{f(\boldsymbol{z}_{1:N})}. \tag{3.32}$$

Using (3.31), we have

$$p(\boldsymbol{z}_{1:(N+1)}) = \frac{\Gamma(\alpha)}{\Gamma(N+1+\alpha)} \prod_{k=1}^K \frac{\Gamma(N'_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})}, \tag{3.33}$$

where $N'_k = N_k + 1$ if $z_{N+1} = k$ and $N'_k = N_k$ else, i.e., for $z_{N+1} = k$ the k-th multiplicity N_k is increased by one. Inserting (3.31) and (3.33) into (3.32) yields

$$p(z_{N+1}|\boldsymbol{z}_{1:N}) = \frac{\frac{\Gamma(\alpha)}{\Gamma(N+1+\alpha)} \prod_{k=1}^K \frac{\Gamma(N'_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})}}{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^K \frac{\Gamma(N_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})}} = \frac{\Gamma(N+\alpha)}{\Gamma(N+1+\alpha)} \prod_{k=1}^K \frac{\Gamma(N'_k + \frac{\alpha}{K})}{\Gamma(N_k + \frac{\alpha}{K})},$$

where

$$\prod_{k=1}^K \frac{\Gamma(N'_k + \frac{\alpha}{K})}{\Gamma(N_k + \frac{\alpha}{K})} = \frac{\Gamma(N_k + 1 + \frac{\alpha}{K})}{\Gamma(N_k + \frac{\alpha}{K})}.$$

Using the property $\Gamma(x+1) = x\Gamma(x)$ gives

$$p(z_{N+1}|\boldsymbol{z}_{1:N}) = \frac{\Gamma(N+\alpha)}{\Gamma(N+\alpha+1)} \frac{\Gamma(N_k + 1 + \frac{\alpha}{K})}{\Gamma(N_k + \frac{\alpha}{K})} = \frac{\Gamma(N+\alpha)}{(N+\alpha)\Gamma(N+\alpha)} \frac{(N_k + \frac{\alpha}{K})\Gamma(N_k + \frac{\alpha}{K})}{\Gamma(N_k + \frac{\alpha}{K})}$$

and the result

$$p(z_{N+1}|\boldsymbol{z}_{1:N}) = \frac{N_k + \frac{\alpha}{K}}{N + \alpha} \tag{3.34}$$

for the predictive distribution $p(z_{N+1}|\boldsymbol{z}_{1:N})$ of the indicator variable $z_{N+1} \in \{1, \ldots, K\}$.

3.3.2 Limit to Infinite Mixtures

Finally, if we take the limit $K \to \infty$ in (3.34) and assume that the indicator variables z_n, $n = 1, \ldots, N$, take on values $l \in \{1, \ldots, L\}$ then

$$p(z_{N+1} = l|z_{1:N}) \to \frac{N_l}{N + \alpha}. \tag{3.35}$$

The limit in (3.35) is the probability that the indicator variable z_{N+1} is equal to one of the previously observed indicators $\boldsymbol{z}_{1:N}$. It can moreover be shown [23] that for $K \to \infty$ the probability that z_{N+1} is different from all the previous z_n for $n = 1, \ldots, N$ becomes

$$p(z_{N+1} \neq z_n \, \forall \, n \leq N|z_{1:N}) \to \frac{\alpha}{N + \alpha}. \tag{3.36}$$

The limit in (3.36) is the probability that the indicator variable z_{N+1} is equal to a new value $L+1$. Note that even when taking the limit to $K = \infty$ components, L can be at most N, which means there is always a strictly positive probability that $z_{N+1} = L+1$. Also note that the sum of both limits (3.35) and (3.36) equals one, i.e.,

$$\frac{\alpha}{N + \alpha} + \sum_{l=1}^{L} \frac{N_l}{N + \alpha} = \frac{\alpha}{N + \alpha} + \frac{N}{N + \alpha} = 1,$$

and thus that the derived predictive distribution for z_{N+1} is a valid probability distribution.

Comparing with the predictive distribution of the CRP in (3.18), the probability given by (3.35) is equal to the probability that a new customer joins an occupied table and the probability given by (3.36) is equal to the probability that a new customer takes a seat at a new table. If we furthermore consider random vectors $\boldsymbol{\theta}_k^*$, $k = 1, 2, \ldots$, i.i.d. according to the base distribution G_0 and create samples $\boldsymbol{\theta}_n$ by setting $\boldsymbol{\theta}_n = \boldsymbol{\theta}'_{z_n}$ with $\boldsymbol{\theta}'_l \in \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots\}$, $l = 1, \ldots, L$, then this generation of the $\boldsymbol{\theta}_n$ is completely equivalent to the DP-based generation of the $\boldsymbol{\theta}_n$ described in Section 3.2.3.

Finally, we can use the samples $\boldsymbol{\theta}_n$ to generate data $\boldsymbol{x}_n$ according to the component distributions $f(\boldsymbol{x}_n|\boldsymbol{\theta}_n)$ of a mixture model (see (2.14)). The resulting observations $\boldsymbol{x}_n$ are then distributed according to a DPM as we will define it in Section 3.4. Thus, by considering (2.7) with a symmetric Dirichlet prior for the mixing proportions $\boldsymbol{\pi}$ according to (3.26) and by taking the limit $K \to \infty$, we arrive at a mixture model with an infinite number of components, more specifically, a DPM.

3.4 Dirichlet Process Mixture Definition

Another way of generalizing a finite mixture model to an infinite mixture model is to use the DP as a prior for the mixing distribution G (see Section 2.1.3). That is, we directly draw G from the DP instead of drawing the mixing proportions $\boldsymbol{\pi}$ and component parameters $\boldsymbol{\theta}_k^*$ separately as in (2.14). In order to model a set of conditionally i.i.d. observations $\boldsymbol{x}_n$, $n = 1, \ldots, N$, we define a DPM according to

$$G \sim \mathcal{DP}(G; \alpha, G_0), \tag{3.37a}$$

$$\boldsymbol{\theta}_n | G \overset{\text{i.i.d.}}{\sim} G(\boldsymbol{\theta}_n | \boldsymbol{\pi}, \boldsymbol{\theta}_{1:\infty}^*), \tag{3.37b}$$

$$\boldsymbol{x}_n | \boldsymbol{\theta}_n \sim f(\boldsymbol{x}_n | \boldsymbol{\theta}_n), \tag{3.37c}$$

for $n = 1, \ldots, N$, where the DP samples $\boldsymbol{\theta}_n$ are hidden parameters that are used to realize the n-th data point $\boldsymbol{x}_n$ via the respective component distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta}_n)$. Each observation $\boldsymbol{x}_n$ is based on a conditionally independently sampled parameter $\boldsymbol{\theta}_n$; note that multiple samples $\boldsymbol{\theta}_n$ can take on the same value because of the discrete nature of G. Therefore, observations $\boldsymbol{x}_n$ based on the same value $\boldsymbol{\theta}_n$ belong to the same cluster, as they are realized from the same conditional distribution [4].

The statistical model of the DPM defined in (3.37) is a Bayesian mixture model similar to the definition of a Bayesian mixture model in (2.14). Given a realization G (see (3.8)) from the DP, i.e.,

$$G(\boldsymbol{\theta}_n | \boldsymbol{\pi}, \boldsymbol{\theta}_{1:\infty}^*) = \sum_{k=1}^{\infty} \pi_k \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_k^*), \tag{3.38}$$

and proceeding as in the derivation of (2.17) in Section 2.1.3, yields the conditional distribution

$$f(\boldsymbol{x}_n | G) = \sum_{k=1}^{\infty} \pi_k f(\boldsymbol{x}_n | \boldsymbol{\theta}_k^*). \tag{3.39}$$

This has exactly the form of a mixture distribution as defined in (2.1), but with a countably infinite number of components.

Using the stick-breaking process, we can define a representation in terms of indicator variables (cf. (2.7)) for the DPM as follows:

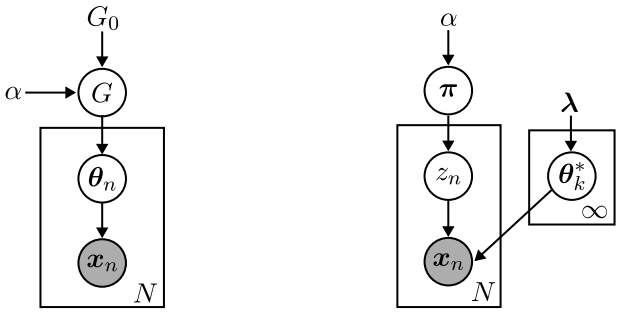
$$\boldsymbol{\pi} \sim \text{GEM}(\boldsymbol{\pi}; \alpha), \tag{3.40a}$$

$$\boldsymbol{\theta}_k^* \overset{\text{i.i.d.}}{\sim} G_0(\boldsymbol{\theta}_k^*; \boldsymbol{\lambda}), \tag{3.40b}$$

$$z_n | \boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n | \boldsymbol{\pi}), \tag{3.40c}$$

$$\boldsymbol{x}_n | z_n, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_\infty^* \sim f(\boldsymbol{x}_n | \boldsymbol{\theta}_{z_n}^*), \tag{3.40d}$$

for $k \in \mathbb{N}$ and $n \in \mathbb{N}$. Here, the mixing proportions $\boldsymbol{\pi}$ are distributed according to the GEM distribution (3.10) with concentration parameter α, and the component parameters $\boldsymbol{\theta}_k^*$, $k = 1, 2, \ldots$, are i.i.d. realizations from the base distribution G_0, which we assume to be parameterized by a (deterministic) hyperparameter $\boldsymbol{\lambda}$. As explained in Section 2.1.2 (see the derivation of (2.9)),



(a) (b)

Figure 3.9: Bayesian network of a DPM model. (a) Model description in terms of the random distribution $G \sim \mathcal{DP}(G; \alpha, G_0)$. The parameters $\boldsymbol{\theta}_{1:N}|G$ are i.i.d. according to G. (b) The mixing proportions $\boldsymbol{\pi} \sim \text{GEM}(\boldsymbol{\pi}; \alpha)$ follow the stick-breaking process and the component parameters $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots$, are i.i.d. according to the base distribution $G_0(\boldsymbol{\theta}_k^*; \boldsymbol{\lambda})$. The indicator variables z_n determine the component that generates $\boldsymbol{x}_n$.

marginalizing the indicator variables z_n reveals a mixture model with the form (2.1), with the difference that the model (3.40) is an infinite component mixture as we sample an infinite number of proportions π_k and component parameters $\boldsymbol{\theta}_k^*$ for $k \in \mathbb{N}$. Observing a dataset $\boldsymbol{x}_{1:N}$ of finite size N of a DPM means that we observe $L \leq N$ clusters in the data, i.e., the indicator variables z_n or the parameters $\boldsymbol{\theta}_n$ take on at most $L \leq N$ different values. Mathematically speaking, the number of clusters L is a random variable and its expected value grows as new data points are observed (see (3.19)).

Finally, we summarize the DPM model descriptions (3.37) and (3.40) in Figures 3.9a and 3.9b, respectively, using a Bayesian network representation [47]. Nodes, which are depicted as circles, represent random quantities and edges represent statistical dependencies. A shaded node means that we observe the respective quantity in the generative process whereas a nonshaded node represents a hidden quantity. Plates are used to compactly represent a repetition of nodes. The hyperparameters α and $\boldsymbol{\lambda}$ are also included but not represented by nodes since they are considered as known quantities that parameterize the distribution of the unknown (random) parameters of the model. Note that we can invoke the CRP (see Section 3.2.3) to generate the sequence of indicator variables $\boldsymbol{z}_{1:N}$ according to (3.18) or the Pólya urn process (see Section 3.2.2) to generate the samples $\boldsymbol{\theta}_{1:N}$ according to (3.16). Using this approach, we are then able to generate $\boldsymbol{x}_n$ according to the component distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta}_n) = f(\boldsymbol{x}|\boldsymbol{\theta}_{z_n}^*)$.

3.5 Exponential Family Dirichlet Process Mixture Model with Conjugate Prior

For the remainder of the thesis, we will consider DPMs where the observed data $\boldsymbol{x}_n$ is drawn from a mixture of EF distributions and the base distribution G_0 is the corresponding conjugate prior (see Section 2.2). This greatly simplifies Bayesian inference. In particular, as shown in [6], it simplifies the derivation of variational inference methods, which we will consider in Chapter 4.

We will represent the DPM in terms of the stick-breaking process (see Section 3.2.1) and hidden indicator variables (see (3.40)). Moreover, we assume the EF to be in canonical form

(see (2.20)) with natural (canonical) parameter $\boldsymbol{\eta}_k^* = \boldsymbol{\eta}_k^*(\boldsymbol{\theta}_k^*)$. This means that the parameter of the component distribution $f(\boldsymbol{x}_n|\boldsymbol{\eta}_k^*)$ is given by $\boldsymbol{\eta}_k^*$ rather than $\boldsymbol{\theta}_k^*$. The auxiliary variables v_k of the stick-breaking process are distributed according to (3.9b) and the mixing proportions π_k are given by (3.9a). Recall that (3.9b) and (3.9a) together are the GEM$(\boldsymbol{\pi}; \alpha)$ distribution. Using (3.40), we can summarize the model for N conditionally independent observations $\boldsymbol{x}_n$ as

$$v_k \overset{\text{i.i.d.}}{\sim} \mathcal{B}(v_k; 1, \alpha), \tag{3.41a}$$

$$\pi_k = \pi_k(v_1, \ldots, v_k) = v_k \prod_{i=1}^{k-1}(1 - v_i), \tag{3.41b}$$

$$\boldsymbol{\eta}_k^* \overset{\text{i.i.d.}}{\sim} G_0(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda}), \tag{3.41c}$$

$$z_n|\boldsymbol{\pi}(v_1, v_2, \ldots) \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n|\boldsymbol{\pi}(v_1, v_2, \ldots)), \tag{3.41d}$$

$$\boldsymbol{x}_n|z_n, \boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*, \ldots \sim f(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*), \tag{3.41e}$$

for $k = 1, 2, \ldots$ and $n = 1, \ldots, N$. Here, $G_0(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda})$ is given by (see (2.26))

$$G_0(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda}) = b(\boldsymbol{\lambda}) \exp\Big(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_k^* - \lambda_2 a(\boldsymbol{\eta}_k^*)\Big) \tag{3.42}$$

and $f(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)$ is given by (see (2.6) and (2.20))

$$f(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*) = \prod_{k=1}^{\infty} \Big(h(\boldsymbol{x}_n) \exp\Big(\boldsymbol{\eta}_k^{*\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_k^*)\Big)\Big)^{\mathbb{1}(z_n = k)} \tag{3.43}$$

$$= h(\boldsymbol{x}_n) \exp\Big(\boldsymbol{\eta}_{z_n}^{*\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_{z_n}^*)\Big). \tag{3.44}$$

The hyperparameter $\boldsymbol{\lambda} = \begin{pmatrix}\boldsymbol{\lambda}_1^{\mathrm{T}} & \lambda_2\end{pmatrix}^{\mathrm{T}}$ consists of a p-dimensional vector $\boldsymbol{\lambda}_1 \in \mathbb{R}^p$, where $p = \dim(\boldsymbol{\eta}_k^*)$, and a scalar $\lambda_2 \in \mathbb{R}$. Note that the auxiliary variables v_k, mixing proportions π_k and the component parameters $\boldsymbol{\eta}_k^*$ are "global" model parameters as explained in Section 2.1.2. In contrast, the indicator variables z_n are "local" model parameters in that each z_n associates the corresponding data point $\boldsymbol{x}_n$ with a certain component k.

The model (3.41) implies the following conditional or unconditional independence relations:

$$v_k \perp\!\!\!\perp v_{k'}, \quad \text{where } k = k' = 1, 2, \ldots \text{ with } k \neq k', \tag{3.45a}$$

$$\boldsymbol{\eta}_k^* \perp\!\!\!\perp \boldsymbol{\eta}_{k'}^*, \quad \text{where } k = k' = 1, 2, \ldots \text{ with } k \neq k', \tag{3.45b}$$

$$\boldsymbol{\eta}_k^* \perp\!\!\!\perp v_1, v_2, \ldots, \quad \text{for all } k = 1, 2, \ldots, \tag{3.45c}$$

$$\boldsymbol{\eta}_k^* \perp\!\!\!\perp z_n \mid v_1, v_2, \ldots, \quad \text{for all } k = 1, 2, \ldots \text{ and } n = 1, \ldots, N, \tag{3.45d}$$

$$z_n \perp\!\!\!\perp z_{n'} \mid v_1, v_2, \ldots, \quad \text{where } n, n' = 1, \ldots, N \text{ with } n \neq n', \tag{3.45e}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp v_1, v_2, \ldots, z_{n'}, x_{n'}, \boldsymbol{\eta}_k^* \mid z_n, \boldsymbol{\eta}_{z_n}^*, \quad \text{where } n, n' = 1, \ldots, N \text{ with } n \neq n' \text{ and } k \neq z_n. \tag{3.45f}$$

Henceforth, we will further abbreviate the observed data $\boldsymbol{x}_{1:N}$ by $\boldsymbol{x} = \begin{pmatrix}\boldsymbol{x}_1^{\mathrm{T}} & \cdots & \boldsymbol{x}_N^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$, the sequence of auxiliary variables $v_1, v_2, \ldots$ by $\boldsymbol{v} = \begin{pmatrix}v_1 & v_2 & \cdots\end{pmatrix}^{\mathrm{T}}$, the sequence of component parameters $\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*, \ldots$ by $\boldsymbol{\eta}^* = \begin{pmatrix}\boldsymbol{\eta}_1^{*\mathrm{T}} & \boldsymbol{\eta}_2^{*\mathrm{T}} & \cdots\end{pmatrix}^{\mathrm{T}}$, and the sequence of indicator variables $z_1, \ldots, z_N$ by $\boldsymbol{z} = \begin{pmatrix}z_1 & \cdots & z_N\end{pmatrix}^{\mathrm{T}}$.

By using the chain rule, the joint pdf of the overall statistical model can be expressed as

$$f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*) = f(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*) p(\boldsymbol{z}|\boldsymbol{v}, \boldsymbol{\eta}^*) f(\boldsymbol{\eta}^*|\boldsymbol{v}) f(\boldsymbol{v}). \tag{3.46}$$

Here, the observations $\boldsymbol{x}$ do not depend on the auxiliary variables $\boldsymbol{v}$ given the indicator variables $\boldsymbol{z}$ and component parameters $\boldsymbol{\eta}^*$, i.e., $f(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*) = f(\boldsymbol{x}|\boldsymbol{\eta}^*, \boldsymbol{z})$, because the knowledge of $\boldsymbol{v}$ is contained in $\boldsymbol{z}$ (see the hierarchical model description (3.41)). Moreover, using the independence assumptions (3.45c) and (3.45d), (3.46) becomes

$$f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*) = f(\boldsymbol{x}|\boldsymbol{\eta}^*, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{v}) f(\boldsymbol{\eta}^*) f(\boldsymbol{v}). \tag{3.47}$$

Each of the four factors in (3.47) can be further simplified by using the independence assumptions (3.45a), (3.45b), (3.45e) and (3.45f), i.e.,

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*) &= \left(\prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})\right)\left(\prod_{n=1}^{N} p(z_n|\boldsymbol{v})\right)\left(\prod_{k=1}^{\infty} f(\boldsymbol{\eta}^*_k)\right)\left(\prod_{k=1}^{\infty} f(v_k)\right) \\ &= \left(\prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n}) p(z_n|\boldsymbol{v})\right)\left(\prod_{k=1}^{\infty} f(\boldsymbol{\eta}^*_k) f(v_k)\right). \end{aligned} \tag{3.48}$$

Finally, inserting the distribution (3.41a) of the auxiliary variable v_k and the conditional distribution (3.41d) of the indicator variable z_n into (3.48) gives

$$f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*) = \left(\prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n}) \mathcal{C}(z_n|\boldsymbol{\pi}(\boldsymbol{v}))\right)\left(\prod_{k=1}^{\infty} G_0(\boldsymbol{\eta}^*_k; \boldsymbol{\lambda}) \mathcal{B}(v_k; 1, \alpha)\right), \tag{3.49}$$

where the likelihood function $f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})$ is given by the EF distribution (3.44) and the prior $f(\boldsymbol{\eta}^*_k) = G_0(\boldsymbol{\eta}^*_k; \boldsymbol{\lambda})$ by the corresponding conjugate prior (3.42). The Bayesian network in Figure 3.10 presents a graphical summary of the model. It visualizes the dependencies among the parameters $\boldsymbol{v}$, $\boldsymbol{z}$ and $\boldsymbol{\eta}^*$ and the observations $\boldsymbol{x}$. The mixture distribution of an individual

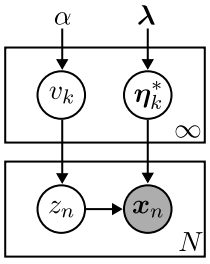


Figure 3.10: Bayesian network representation of (3.49).

observation $\boldsymbol{x}_n$ is given by (cf. (3.39))

$$f(\boldsymbol{x}_n|\boldsymbol{v}, \boldsymbol{\eta}^*) = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{v}_{1:k}) f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_k), \tag{3.50}$$

where the mixing proportions $\pi_k(\boldsymbol{v}_{1:k})$ are determined according to (3.41b).

4 Variational Inference for Exponential Family Dirichlet Process Mixtures

The central object of interest in Bayesian inference is the posterior distribution

$$f(\boldsymbol{w}|\boldsymbol{x}) = \frac{f(\boldsymbol{w},\boldsymbol{x})}{f(\boldsymbol{x})} \tag{4.1}$$

$$= \frac{f(\boldsymbol{w},\boldsymbol{x})}{\int_{\mathbb{R}^P} f(\boldsymbol{w},\boldsymbol{x})\,\mathrm{d}\boldsymbol{w}}, \tag{4.2}$$

because it allows us to estimate the hidden parameters $\boldsymbol{w} = (w_1 \ \ldots \ w_P)^{\mathrm{T}} \in R^P$ given observations $\boldsymbol{x} = \left(\boldsymbol{x}_1^{\mathrm{T}} \ \ldots \ \boldsymbol{x}_N^{\mathrm{T}}\right)^{\mathrm{T}}$ and various Bayesian estimators, such as the MMSE and MAP estimators (see (2.40) and (2.41)), can be obtained from it. The joint pdf $f(\boldsymbol{w},\boldsymbol{x})$ in (4.2) describes the statistical model (see (3.49) for an example) and is assumed to be known. The evidence $f(\boldsymbol{x})$ follows by marginalization of $f(\boldsymbol{w},\boldsymbol{x})$ with respect to the hidden parameters $\boldsymbol{w}$. In statistical models for which a closed-form expression of the posterior is costly to compute or in worst cases intractable, such as mixture models (see Section 2.3), computation of the MMSE and MAP estimators is typically not possible. We then resort to approximate inference methods which approximate the posterior distribution.

In this chapter, we will address an approximate inference technique known as variational inference (VI) [6]. The principle of VI is to approximate the posterior distribution $f(\boldsymbol{w},\boldsymbol{x})$ by a different distribution that is tractable and not so costly to compute. This is done by finding the "best" approximation to the posterior pdf within a family of approximating pdfs where "best" means closest in terms of some divergence measure. Thus, VI methods are a discrete, optimization-based methods where determining the approximate posterior pdf is formulated as an optimization problem. The choice of divergence measure is typically the Kullback-Leibler divergence [42] and the family of approximating pdfs is chosen to exhibit some structure that simplifies the problem at hand.

The following sections are, unless stated otherwise, based on [7] and [6], where [7] was the first paper that demonstrated a VI method for DPMs. We will first present a general formulation of the VI optimization problem and then present a solution to this problem using the mean field (MF) approximation and coordinate ascent variational inference (CAVI). Furthermore, we will apply CAVI to the exponential family DPM model described in Section 3.5. At the end of the chapter, we will discuss possible practical issues that have to be considered when using the CAVI algorithm.

4.1 Variational Inference as an Optimization Problem

Instead of computing the exact posterior pdf (4.2), the basic idea of VI is to approximate it by a pdf $q(\boldsymbol{w})$ with a simple structure (i.e., a structure that makes $q(\boldsymbol{w})$ computationally efficient to compute) called the variational distribution. Optimization of the variational distribution involves the divergence $D(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ which measures the distance of the two distributions $q(\boldsymbol{w})$ and $f(\boldsymbol{w}|\boldsymbol{x})$. VI amounts to minimizing the divergence $D(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ between the variational pdf $q(\boldsymbol{w})$ and the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$. By replacing the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$ with the approximation $q(\boldsymbol{w})$ in Bayesian estimators such as (2.40) and (2.41), we can view VI as a reformulation of Bayesian inference as an optimization problem. In the following we will define the Kullback-Leibler divergence along with a related quantity known as the evidence lower bound. At the end of the section, the optimization problem of VI is mathematically formulated.

4.1.1 Kullback-Leibler Divergence

The Kullback-Leibler divergence (KLD) [42] is a measure of the dissimilarity between two probability distributions $q(\boldsymbol{w})$ and $f(\boldsymbol{w})$. It is given by

$$D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w})) = \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{q(\boldsymbol{w})}{f(\boldsymbol{w})}\right\} = \int_{\mathbb{R}^P} q(\boldsymbol{w}) \ln \frac{q(\boldsymbol{w})}{f(\boldsymbol{w})} \,\mathrm{d}\boldsymbol{w}. \tag{4.3}$$

Equivalent expressions are

$$\begin{aligned} D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w})) &= \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w})\} - \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln f(\boldsymbol{w})\} \\ &= \int_{\mathbb{R}^P} q(\boldsymbol{w}) \ln q(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w} - \int_{\mathbb{R}^P} q(\boldsymbol{w}) \ln f(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w} \end{aligned} \tag{4.4}$$

and

$$D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w})) = -h_q - \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln f(\boldsymbol{w})\}, \tag{4.5}$$

where h_q is the differential entropy corresponding to the pdf $q(\boldsymbol{w})$, i.e.,

$$h_q = -\mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w})\} = -\int_{\mathbb{R}^P} q(\boldsymbol{w}) \ln q(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w}. \tag{4.6}$$

Important properties of the KLD include

$$D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w})) \geq 0, \tag{4.7}$$

$$D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w})) = 0 \quad \text{if and only if} \quad q(\boldsymbol{w}) = f(\boldsymbol{w}), \tag{4.8}$$

and in general $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}))$ does not equal $D_{\mathrm{KL}}(f(\boldsymbol{w}) \,||\, q(\boldsymbol{w}))$, i.e.,

$$D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w})) \neq D_{\mathrm{KL}}(f(\boldsymbol{w}) \,||\, q(\boldsymbol{w})). \tag{4.9}$$

Due to the lack of symmetry (see (4.9)) and because it does not satisfy the triangle equality, the KLD is not a metric in the strict mathematical sense.

Depending on the two different ways $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}))$ and $D_{\mathrm{KL}}(f(\boldsymbol{w}) \,||\, q(\boldsymbol{w}))$ we can use the KLD in an optimization problem we obtain different approximate inference methods [2]. VI methods employ $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ (also called the exclusive or reverse KLD), i.e., the first term corresponds to the approximating pdf $q(\boldsymbol{w})$ and the second term is the true posterior $f(\boldsymbol{w}|\boldsymbol{x})$. This way of expressing the KLD leads to the evidence lower bound, which will be discussed presently. Note that if we want to minimize $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ with respect to $q(\boldsymbol{w})$, we must have $q(\boldsymbol{w}) = 0$ whenever $f(\boldsymbol{w}|\boldsymbol{x}) = 0$ (see (4.3)) in order to prevent the KLD becoming infinite. This behavior is called zero-forcing or mode-seeking. By contrast, to prevent $D_{\mathrm{KL}}(f(\boldsymbol{w}|\boldsymbol{x}) \,||\, q(\boldsymbol{w}))$ (also called the inclusive or forward KLD) from becoming infinite, we must have $q(\boldsymbol{w}) > 0$ whenever $f(\boldsymbol{w}|\boldsymbol{x}) > 0$ and as a result $q(\boldsymbol{w})$ tends to cover all modes of $f(\boldsymbol{w}|\boldsymbol{x})$. This behavior is called zero-avoiding or mode-covering and gives rise to an approximate inference method known as expectation propagation (or moment matching), which we will not consider in this thesis.

4.1.2 Evidence Lower Bound

VI methods seek to minimize $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ such that the variational pdf $q(\boldsymbol{w})$ is as similar as possible to the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$. Since we assume that the posterior $f(\boldsymbol{w}|\boldsymbol{x})$ is intractable, we can not compute the KLD and therefore can not minimize it directly. Starting with the KLD expression (4.4), a further decomposition of $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ yields

$$\begin{aligned} D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})) &= \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w})\} - \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln f(\boldsymbol{w}|\boldsymbol{x})\} \\ &= \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w})\} - \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{f(\boldsymbol{w},\boldsymbol{x})}{f(\boldsymbol{x})}\right\} \\ &= \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w})\} - \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln f(\boldsymbol{w},\boldsymbol{x}) - \ln f(\boldsymbol{x})\} \\ &= \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w}) - \ln f(\boldsymbol{w},\boldsymbol{x})\} + \ln f(\boldsymbol{x}), \end{aligned} \tag{4.10}$$

where we inserted (4.1) for the posterior $f(\boldsymbol{w}|\boldsymbol{x})$ and $\mathrm{E}^{(q(\boldsymbol{w}))}\{\ln f(\boldsymbol{x})\} = \ln f(\boldsymbol{x})$ since the expectation is with respect to $q(\boldsymbol{w})$ and the evidence $f(\boldsymbol{x})$ does not involve the hidden parameters $\boldsymbol{w}$. Equation (4.10) shows that the logarithm of evidence $f(\boldsymbol{x})$, also called the log-evidence, can be expressed as

$$\begin{aligned} \ln f(\boldsymbol{x}) &= \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln f(\boldsymbol{w},\boldsymbol{x}) - \ln q(\boldsymbol{w})\} + D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})) \\ &= \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{f(\boldsymbol{w},\boldsymbol{x})}{q(\boldsymbol{w})}\right\} + D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})) \qquad (4.11) \\ &\geq \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{f(\boldsymbol{w},\boldsymbol{x})}{q(\boldsymbol{w})}\right\}. \qquad (4.12) \end{aligned}$$

In the last step we used the fact that $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})) \geq 0$ (cf. (4.7)) and therefore the right side of (4.12) is a lower-bound on the log-evidence, called the evidence lower bound (ELBO):

$$\begin{aligned} L(q;\boldsymbol{x}) &\triangleq \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{f(\boldsymbol{w},\boldsymbol{x})}{q(\boldsymbol{w})}\right\} \qquad (4.13) \\ &= \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln f(\boldsymbol{w},\boldsymbol{x})\} - \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w})\}. \qquad (4.14) \end{aligned}$$

According to (4.11) the ELBO is equivalently given by

$$L(q;\boldsymbol{x}) = \ln f(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})). \tag{4.15}$$

Note that the lower bound (4.12) is attained if and only if the variational pdf $q(\boldsymbol{w})$ equals the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$ because $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})) = 0$ if and only if $q(\boldsymbol{w}) = f(\boldsymbol{w}|\boldsymbol{x})$. Also note that the ELBO is an approximation of the evidence $f(\boldsymbol{x})$, which provides a basis for selecting a statistical model [6]. It can therefore be used as a criterion in model selection (see [48] for a discussion).

The expression (4.15) for the ELBO leads to an interesting result. Since a closed-form of $f(\boldsymbol{w}|\boldsymbol{x})$ is not available and we can not minimize $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ directly, we can instead maximize the ELBO $L(q;\boldsymbol{x})$ with respect to the variational pdf $q(\boldsymbol{w})$. According to (4.14), the ELBO does not depend on the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$. In order to calculate and maximize the ELBO we instead need the joint pdf $f(\boldsymbol{w},\boldsymbol{x})$ of the statistical model, which we assume to be known. Considering the expression (4.15) and denoting the set of all possible distributions for the vector $\boldsymbol{w}$ as $\mathcal{F}$, maximization of the ELBO is equivalent to minimizing the KLD (cf. [39]), i.e.,

$$\begin{aligned}
q_0(\boldsymbol{w};\boldsymbol{x}) &= \underset{q(\boldsymbol{w})\in\mathcal{F}}{\arg\max}\, L(q;\boldsymbol{x}) && (4.16)\\
&= \underset{q(\boldsymbol{w})\in\mathcal{F}}{\arg\max}\{\ln f(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))\} && (4.17)\\
&= \underset{q(\boldsymbol{w})\in\mathcal{F}}{\arg\max}\{-D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))\} && (4.18)\\
&= \underset{q(\boldsymbol{w})\in\mathcal{F}}{\arg\min}\, D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})) && (4.19)\\
&= f(\boldsymbol{w}|\boldsymbol{x}). && (4.20)
\end{aligned}$$

This is because the evidence $f(\boldsymbol{x})$ corresponds to an additive constant when maximizing with respect to $q(\boldsymbol{w})$ and due to (4.8) the maximum of the ELBO is obtained if and only if $q(\boldsymbol{w}) = f(\boldsymbol{w}|\boldsymbol{x})$. With this approach we can avoid computing the intractable KLD and instead maximize the ELBO, which depends on the joint pdf of the model $f(\boldsymbol{w},\boldsymbol{x})$ and the variational pdf $q(\boldsymbol{w})$. As the principle of VI is to approximate the posterior pdf by a simpler pdf, we will next constrain the set $\mathcal{F}$ to a subset $\mathcal{Q}$ of pdfs with a simple structure. Although we then no longer obtain the true posterior as result (i.e., $f(\boldsymbol{w}|\boldsymbol{x}) \notin \mathcal{Q}$), we gain the advantage of an efficient computation of the maximum of the ELBO.

4.1.3 Constrained Optimization

We will denote the set of approximating pdfs $q(\boldsymbol{w})$ by $\mathcal{Q}$ and refer to it as the variational family. The VI optimization problem is then formulated as

$$q^*(\boldsymbol{w};\boldsymbol{x}) = \underset{q(\boldsymbol{w})\in\mathcal{Q}}{\arg\max}\, L(q;\boldsymbol{x}) = \underset{q(\boldsymbol{w})\in\mathcal{Q}}{\arg\max}\, \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{f(\boldsymbol{w},\boldsymbol{x})}{q(\boldsymbol{w})}\right\}, \tag{4.21}$$

i.e., the pdf $q^*(\boldsymbol{w};\boldsymbol{x}) \in \mathcal{Q}$ is the pdf that maximizes the ELBO. We have shown in (4.16)–(4.20) that this is equivalent to minimizing the divergence measure (see (4.3))

$$D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x})) = \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{q(\boldsymbol{w})}{f(\boldsymbol{w}|\boldsymbol{x})}\right\} = \int_{\mathbb{R}^p} q(\boldsymbol{w}) \ln \frac{q(\boldsymbol{w})}{f(\boldsymbol{w}|\boldsymbol{x})}\, \mathrm{d}\boldsymbol{w}, \tag{4.22}$$

which means finding the pdf $q(\boldsymbol{w}) \in \mathcal{Q}$ that is most similar to the posterior $f(\boldsymbol{w}|\boldsymbol{x})$, with the difference that we now consider a constrained version of $\mathcal{F}$, i.e., $\mathcal{Q} \subset \mathcal{F}$. Note that the variational family is not a model of the observed data, that is, $q(\boldsymbol{w})$ does not depend on the data. Instead, it is the ELBO and the corresponding optimization problem (4.21) that connects the fitted variational density $q^*(\boldsymbol{w};\boldsymbol{x})$ to the model $f(\boldsymbol{w},\boldsymbol{x})$ and the observed data $\boldsymbol{x}$. The complexity of the variational family $\mathcal{Q}$ determines the complexity of the optimization problem, where $\mathcal{Q}$ should be chosen such that (4.21) can be solved with low computational effort while approximating the posterior with sufficient accuracy.

The solution of (4.21) can be obtained by various optimization techniques. In what follows we will consider a popular choice for $\mathcal{Q}$ and an iterative way of solving the maximization problem.

4.2 Mean Field Variational Family

In this thesis we focus on the mean field variational family. This approach assumes that the variational distribution $q(\boldsymbol{w})$ factorizes with respect to some partition of the parameters $\boldsymbol{w}$. We will first define $\mathcal{Q}$ and later present a factorization of $q(\boldsymbol{w})$ for the case of the DPM.

4.2.1 Definition

We define members $q(\boldsymbol{w})$ of the mean field variational family $\mathcal{Q}$ to be of the form

$$q(\boldsymbol{w}) = \prod_{j=1}^{J} q_j(\boldsymbol{w}_j), \tag{4.23}$$

where the P-dimensional parameter vector $\boldsymbol{w}$ is partitioned into J disjoint P_j-dimensional subvectors $\boldsymbol{w}_j$ such that $\sum_{j=1}^{J} P_j = P$. Each subvector $\boldsymbol{w}_j$ contains a unique set of P_j scalar parameters and the union of all these subsets is contained in the entire parameter vector $\boldsymbol{w}$. According to (4.23) the variational distribution $q(\boldsymbol{w})$ factorizes into approximating pdfs $q_j(\boldsymbol{w}_j)$ of the individual subvectors $\boldsymbol{w}_j$. Note that we place no restriction on the functional form of the individual distributions $q_j(\boldsymbol{w}_j)$. Accordingly, denoting the set of all possible pdfs for a P_j-dimensional random vector $\boldsymbol{w}_j$ with $\mathcal{F}_j$, $\mathcal{Q}$ is defined as

$$\mathcal{Q} = \left\{ q(\boldsymbol{w}) \in \mathcal{F} \,\middle|\, q(\boldsymbol{w}) = \prod_{j=1}^{J} q_j(\boldsymbol{w}_j), \text{ with } q_j(\boldsymbol{w}_j) \in \mathcal{F}_j \right\}. \tag{4.24}$$

By applying the mean field variational family in the VI optimization problem (4.21), the resulting approximation $q^*(\boldsymbol{w};\boldsymbol{x})$ for the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$ also factorizes according to (4.23),

i.e.,

$$q^*(\boldsymbol{w};\boldsymbol{x}) = \prod_{j=1}^{J} q_j^*(\boldsymbol{w}_j;\boldsymbol{x}). \tag{4.25}$$

4.2.2 Properties

The factorization of $q(\boldsymbol{w})$ and $q^*(\boldsymbol{w};\boldsymbol{x})$ (see (4.23) and (4.25)) leads to computations involving low-dimensional subvectors $\boldsymbol{w}_j$, which tend to be less complex. Furthermore, we obtain the marginal pdfs $q_j^*(\boldsymbol{w}_j;\boldsymbol{x})$ for $\boldsymbol{w}_j$ that would be obtained by marginalizing out all other parameter vectors $\boldsymbol{w}_i$, where $i \neq j$, from the joint pdf $q^*(\boldsymbol{w};\boldsymbol{x})$, which is another way to save on computational complexity. The marginal pdfs can be used to directly calculate estimates for the parameters $\boldsymbol{w}_j$ by using the MMSE estimator or the MAP estimator defined in (2.40) and (2.41), e.g.,

$$\hat{\boldsymbol{w}}_{j,\mathrm{MMSE}}(\boldsymbol{x}) \approx \mathrm{E}^{(q_j^*(\boldsymbol{w}_j;\boldsymbol{x}))}\{\boldsymbol{w}_j\} = \int_{\mathbb{R}^{P_j}} \boldsymbol{w}_j q_j^*(\boldsymbol{w}_j;\boldsymbol{x})\,\mathrm{d}\boldsymbol{w}_j,$$

where we replaced the true marginal posterior $f(\boldsymbol{w}|\boldsymbol{x})$ with the approximation $q_j^*(\boldsymbol{w}_j;\boldsymbol{x})$. We note that this approach only approximates the MMSE and MAP estimates since we use an approximation of the true posterior.

Although the mean field variational family can describe any marginal density of the hidden parameters $\boldsymbol{w}$, we can not describe correlation between them. In combination with the mode-seeking behavior of the reverse KLD $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$, this often leads to an approximate posterior $q^*(\boldsymbol{w};\boldsymbol{x})$ that under-represents the true posterior, more specifically, the approximated posterior underestimates the second order statistics of the true posterior. An example is given by Figure 4.1 where a two-dimensional Gaussian posterior $f(\boldsymbol{w}|\boldsymbol{x}) = f(w_1, w_2|\boldsymbol{x})$ is approximated by the product of two one-dimensional Gaussians, i.e., $q(\boldsymbol{w}) = q(w_1, w_2) = q(w_1)q(w_2)$, by minimizing $D_{\mathrm{KL}}(q(\boldsymbol{w}) \,||\, f(\boldsymbol{w}|\boldsymbol{x}))$ [2]. Because the reverse KLD penalizes placing probability mass in $q(\boldsymbol{w})$ on areas where the posterior $f(\boldsymbol{w}|\boldsymbol{x})$ has little mass (cf. (4.22)), the variational approximation under-represents the posterior, i.e., it does not expand into areas where the posterior has little mass. Therefore, using the mean field variational family (4.24) allows a trade-off between accuracy (which increases as J decreases) and computational effort (which decreases as J increases).

4.3 Coordinate Ascent Variational Inference

The mean field variational family (4.24) allows us to use a coordinate ascent [2] scheme to approximate the mean field solution (4.25) in an iterative manner. This is called coordinate ascent variational inference (CAVI) [6] and consists of optimizing each variational pdf factor $q_j(\boldsymbol{w}_j)$ individually while keeping the other pdf factors $q_i(\boldsymbol{w}_i)$ with $i \neq j$ fixed. Note that this is in contrast to the general VI optimization problem (4.21) where we perform a joint optimization of all the pdf factors $q_j(\boldsymbol{w}_j)$, $j = 1, \ldots, J$, while defining $\mathcal{Q}$ as the mean field variational family. Thus, using the CAVI algorithm, the ELBO $L(q;\boldsymbol{x})$ is maximized with respect to each pdf

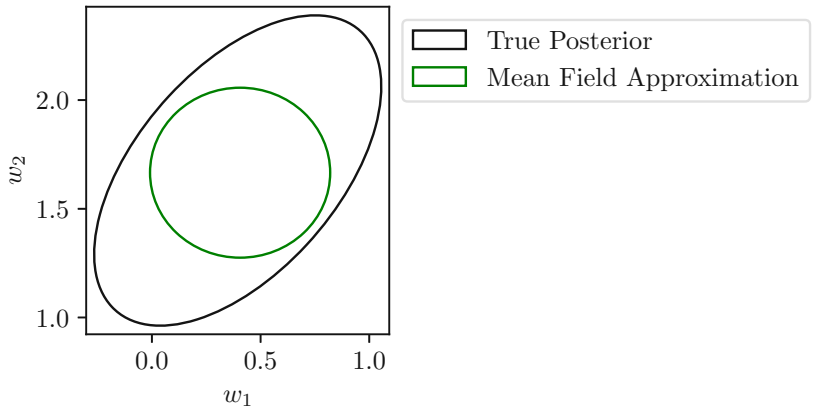


Figure 4.1: Variational approximation $q(\boldsymbol{w}) = q(w_1)q(w_2)$ of a two-dimensional Gaussian posterior $f(w_1, w_2|\boldsymbol{x})$ with a Gaussian variational family having a diagonal covariance matrix corresponding to the mean field approximation. Shown is the solution minimizing the reverse KLD. The Gaussian distributions are visualized with the contours at two standard deviations.

factor $q_j(\boldsymbol{w}_j)$ in turn, instead of jointly, i.e.,

$$q_j^{(\ell)}(\boldsymbol{w}_j; \boldsymbol{x}) = \underset{q_j(\boldsymbol{w}_j)\in\mathcal{F}_j}{\arg\max}\, L(q; \boldsymbol{x}) = \underset{q_j(\boldsymbol{w}_j)\in\mathcal{F}_j}{\arg\max}\, \mathrm{E}^{(q(\boldsymbol{w}))}\left\{\ln \frac{f(\boldsymbol{w}, \boldsymbol{x})}{q(\boldsymbol{w})}\right\}, \tag{4.26}$$

where ℓ is the index of the current iteration. In each iteration ℓ the j-th substep, $j \in \{1, \ldots, J\}$, updates the previous iterate $q_j^{(\ell-1)}(\boldsymbol{w}; \boldsymbol{x})$ according to (4.26). The variational pdf $q(\boldsymbol{w})$ used in (4.26) is given by (4.23), i.e., $q(\boldsymbol{w}) = \prod_{j=1}^{J} q_j(\boldsymbol{w}_j)$. Each pdf factor $q_i(\boldsymbol{w}_i)$ with $i \neq j$ is equal to the result of the most recent respective update, which was either calculated in iteration step $\ell - 1$ or in a previous substep of iteration step ℓ.

The result of the CAVI algorithm depends on the initial choice of the pdfs $q_j^{(0)}(\boldsymbol{w}; \boldsymbol{x})$ before starting with the first iteration $\ell = 1$. After initialization, the factor pdfs $q_j^{(\ell)}(\boldsymbol{w}; \boldsymbol{x})$, $\ell \geq 1$, are updated in each iteration until the algorithm converges, i.e., the change in $L(q; \boldsymbol{x})$ relative to the previous iteration step falls below a predefined threshold. In general, the ELBO is a non-convex objective function, but convergence to a local optimum is guaranteed since the bound is convex with respect to each of the individual factors $q_j^{(\ell)}(\boldsymbol{w}; \boldsymbol{x})$ [42]. We will discuss initialization and convergence of the algorithm in more detail in Section 4.5.

4.3.1 Update Equation for the Variational Factors

We now derive a general solution to the CAVI optimization problem (4.26). We will denote the parameter vector $\boldsymbol{w} \in \mathbb{R}^P$ with the j-th subvector $\boldsymbol{w}_j \in \mathbb{R}^{P_j}$ removed with $\boldsymbol{w}_{\sim j} \in \mathbb{R}^{P_{\sim j}}$, i.e., $\boldsymbol{w}_{\sim j}$ contains $P_{\sim j} \triangleq P - P_j$ scalar parameters that are not included in $\boldsymbol{w}_j$. Using this notation, we define the variational pdf $q(\boldsymbol{w})$ with the j-th pdf factor $q_j(\boldsymbol{w}_j)$ removed as

$$q_{\sim j}(\boldsymbol{w}_{\sim j}) \triangleq \frac{q(\boldsymbol{w})}{q_j(\boldsymbol{w}_j)} = \frac{\prod_{i=1}^{J} q_i(\boldsymbol{w}_i)}{q_j(\boldsymbol{w}_j)} = \prod_{i\neq j} q_i(\boldsymbol{w}_i). \tag{4.27}$$

Furthermore, considering the joint pdf $q(\boldsymbol{w})$ of the parameter vector $\boldsymbol{w} = \left(\boldsymbol{w}_1{}^{\mathrm{T}} \ \dots \ \boldsymbol{w}_J{}^{\mathrm{T}}\right)^{\mathrm{T}}$ and arbitrary functions $g_j(\boldsymbol{w}_j)$, we will make extensive use of the following properties of the expectation:

- Linearity of expectation:

$$\mathrm{E}^{(q(\boldsymbol{w}))}\{g_j(\boldsymbol{w}_j) + g_{\sim j}(\boldsymbol{w}_{\sim j})\} = \mathrm{E}^{(q(\boldsymbol{w}))}\{g_j(\boldsymbol{w}_j)\} + \mathrm{E}^{(q(\boldsymbol{w}))}\{g_{\sim j}(\boldsymbol{w}_{\sim j})\}, \tag{4.28a}$$

$$\mathrm{E}^{(q(\boldsymbol{w}))}\{cg(\boldsymbol{w})\} = c\mathrm{E}^{(q(\boldsymbol{w}))}\{g(\boldsymbol{w})\}. \tag{4.28b}$$

- Marginalization within the expectation operation:

$$\mathrm{E}^{(q(\boldsymbol{w}))}\{g_j(\boldsymbol{w}_j)\} = \mathrm{E}^{(q(\boldsymbol{w}_j))}\{g_j(\boldsymbol{w}_j)\}, \tag{4.29}$$

  or in terms of integrals

$$\begin{aligned}\mathrm{E}^{(q(\boldsymbol{w}))}\{\boldsymbol{w}_j\} &= \int_{\mathbb{R}^P} g_j(\boldsymbol{w}_j) q(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbb{R}^{P_j}} g_j(\boldsymbol{w}_j) \int_{\mathbb{R}^{P_{\sim j}}} q(\boldsymbol{w}_{\sim j}, \boldsymbol{w}_j) \,\mathrm{d}\boldsymbol{w}_{\sim j} \,\mathrm{d}\boldsymbol{w}_j \\ &= \int_{\mathbb{R}^{P_j}} g_j(\boldsymbol{w}_j) q_j(\boldsymbol{w}_j) \,\mathrm{d}\boldsymbol{w}_j \\ &= \mathrm{E}^{(q(\boldsymbol{w}_j))}\{g_j(\boldsymbol{w}_j)\}.\end{aligned}$$

- If $\boldsymbol{w}_j$ and $\boldsymbol{w}_{\sim j}$ are independent, i.e. $\boldsymbol{w}_j \perp\!\!\!\perp \boldsymbol{w}_{\sim j}$ and $q(\boldsymbol{w}) = q_j(\boldsymbol{w}_j) q_{\sim j}(\boldsymbol{w}_{\sim j})$, then

$$\mathrm{E}^{(q(\boldsymbol{w}))}\{g_j(\boldsymbol{w}_j) g_{\sim j}(\boldsymbol{w}_{\sim j})\} = \mathrm{E}^{(q_j(\boldsymbol{w}_j))}\{g_j(\boldsymbol{w}_j)\} \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{g_{\sim j}(\boldsymbol{w}_{\sim j})\}, \tag{4.30}$$

  or in terms of integrals

$$\begin{aligned}\mathrm{E}^{(q(\boldsymbol{w}))}\{g_j(\boldsymbol{w}_j) g_{\sim j}(\boldsymbol{w}_{\sim j})\} &= \int_{\mathbb{R}^P} g_j(\boldsymbol{w}_j) g_{\sim j}(\boldsymbol{w}_{\sim j}) q(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbb{R}^{P_j}} \int_{\mathbb{R}^{P_{\sim j}}} g_j(\boldsymbol{w}_j) g_{\sim j}(\boldsymbol{w}_{\sim j}) q_j(\boldsymbol{w}_j) q_{\sim j}(\boldsymbol{w}_{\sim j}) \,\mathrm{d}\boldsymbol{w}_j \,\mathrm{d}\boldsymbol{w}_{\sim j} \\ &= \int_{\mathbb{R}^{P_j}} g_j(\boldsymbol{w}_j) q_j(\boldsymbol{w}_j) \,\mathrm{d}\boldsymbol{w}_j \int_{\mathbb{R}^{P_{\sim j}}} g_{\sim j}(\boldsymbol{w}_{\sim j}) q_{\sim j}(\boldsymbol{w}_{\sim j}) \,\mathrm{d}\boldsymbol{w}_{\sim j} \\ &= \mathrm{E}^{(q_j(\boldsymbol{w}_j))}\{g_j(\boldsymbol{w}_j)\} \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{g_{\sim j}(\boldsymbol{w}_{\sim j})\}.\end{aligned}$$

- If $q(\boldsymbol{w}) = q_j(\boldsymbol{w}_j) q_{\sim j}(\boldsymbol{w}_{\sim j})$, then we can compute the expectation of $g(\boldsymbol{w})$ iteratively with respect to the marginal pdfs $q_j(\boldsymbol{w}_j)$ and $q_{\sim j}(\boldsymbol{w}_{\sim j})$, i.e.,

$$\mathrm{E}^{(q(\boldsymbol{w}))}\{g(\boldsymbol{w})\} = \mathrm{E}^{(q(\boldsymbol{w}_j))}\left\{\mathrm{E}^{(q(\boldsymbol{w}_{\sim j}))}\{g(\boldsymbol{w})\}\right\}, \tag{4.31}$$

  or in terms of integrals

$$\begin{aligned}\mathrm{E}^{(q(\boldsymbol{w}))}\{\boldsymbol{w}\} &= \int_{\mathbb{R}^P} g(\boldsymbol{w}) q(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbb{R}^P} g(\boldsymbol{w}) q(\boldsymbol{w}) \,\mathrm{d}\boldsymbol{w}\end{aligned}$$

$$= \int_{\mathbb{R}^{P_j}} q(\boldsymbol{w}_j) \Bigg(\underbrace{\int_{\mathbb{R}^{P_{\sim j}}} g(\boldsymbol{w}) q(\boldsymbol{w}_{\sim j}) \, \mathrm{d}\boldsymbol{w}_{\sim j}}_{\mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\boldsymbol{w}\}} \Bigg) \mathrm{d}\boldsymbol{w}_j$$
$$= \mathrm{E}^{(q(\boldsymbol{w}_j))} \Big\{ \mathrm{E}^{(q(\boldsymbol{w}_{\sim j}))} \{ g(\boldsymbol{w}) \} \Big\}.$$

We are now prepared to further develop the optimization problem (4.26) for all $j = 1, \ldots, J$. We start by substituting the factorization $q(\boldsymbol{w}) = q_j(\boldsymbol{w}_j) q_{\sim j}(\boldsymbol{w}_{\sim j})$ (cf. (4.27)) into the definition (4.14) of the ELBO and decompose it into terms that depend on $q_j(\boldsymbol{w}_j)$ or $q_{\sim j}(\boldsymbol{w}_{\sim j})$. The definition (4.14) of the ELBO is given by

$$L(q; \boldsymbol{x}) = \mathrm{E}^{(q(\boldsymbol{w}))} \{ \ln f(\boldsymbol{w}, \boldsymbol{x}) \} - \mathrm{E}^{(q(\boldsymbol{w}))} \{ \ln q(\boldsymbol{w}) \}. \tag{4.32}$$

For the first term in (4.32) we obtain

$$\mathrm{E}^{(q(\boldsymbol{w}))} \{ \ln f(\boldsymbol{w}, \boldsymbol{x}) \} = \mathrm{E}^{(q_j(\boldsymbol{w}_j))} \Big\{ \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))} \{ \ln f(\boldsymbol{w}_j, \boldsymbol{w}_{\sim j}, \boldsymbol{x}) \} \Big\}, \tag{4.33}$$

where we used the expectation property (4.31). The second term in (4.32) can be developed as

$$\mathrm{E}^{(q(\boldsymbol{w}))} \{ \ln q_j(\boldsymbol{w}_j) q_{\sim j}(\boldsymbol{w}_{\sim j}) \} = \mathrm{E}^{(q_j(\boldsymbol{w}_j))} \{ \ln q_j(\boldsymbol{w}_j) \} + \mathrm{E}^{(q_j(\boldsymbol{w}_{\sim j}))} \{ \ln q_{\sim j}(\boldsymbol{w}_{\sim j}) \}, \tag{4.34}$$

by using the expectation properties (4.28) and (4.29). Inserting (4.33) and (4.34) into (4.32) yields

$$L(q; \boldsymbol{x}) = \mathrm{E}^{(q_j(\boldsymbol{w}_j))} \Big\{ \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))} \{ \ln f(\boldsymbol{w}_j, \boldsymbol{w}_{\sim j}, \boldsymbol{x}) \} \Big\}$$
$$- \mathrm{E}^{(q_j(\boldsymbol{w}_j))} \{ \ln q_j(\boldsymbol{w}_j) \} - \mathrm{E}^{(q_j(\boldsymbol{w}_{\sim j}))} \{ \ln q_{\sim j}(\boldsymbol{w}_{\sim j}) \}.$$

Next, using the definition (4.6) of the differential entropy, i.e., $h_q = -\mathrm{E}^{(q(\boldsymbol{w}))} \{ \ln q(\boldsymbol{w}) \}$, we obtain

$$L(q; \boldsymbol{x}) = \mathrm{E}^{(q_j(\boldsymbol{w}_j))} \Big\{ \ln \check{f}(\boldsymbol{w}_j; \boldsymbol{x}) - c \Big\} + h_{q_j} + h_{q_{\sim j}}, \tag{4.35}$$

where we have defined a new distribution $\check{f}(\boldsymbol{w}_j; \boldsymbol{x})$ by the relation

$$\ln \check{f}(\boldsymbol{w}_j; \boldsymbol{x}) \triangleq \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))} \{ \ln f(\boldsymbol{w}_j, \boldsymbol{w}_{\sim j}, \boldsymbol{x}) \} + c. \tag{4.36}$$

Here, c is an additive constant (depending on $\boldsymbol{x}$) that is determined such that (4.36) represents a valid pdf, i.e., $\int_{\mathbb{R}^{P_j}} \check{f}(\boldsymbol{w}_j; \boldsymbol{x}) \, \mathrm{d}\boldsymbol{w}_j = 1$. Using (4.5), the terms depending on $\boldsymbol{w}_j$ in (4.35) can be replaced by the negative KLD between $q_j(\boldsymbol{w}_j)$ and $\check{f}(\boldsymbol{w}_j; \boldsymbol{x})$, i.e.,

$$L(q; \boldsymbol{x}) = h_{q_j} + \mathrm{E}^{(q_j(\boldsymbol{w}_j))} \Big\{ \ln \check{f}(\boldsymbol{w}_j; \boldsymbol{x}) \Big\} + h_{q_{\sim j}} - c$$
$$= -D_{\mathrm{KL}} \Big(q_j(\boldsymbol{w}_j) \,||\, \check{f}(\boldsymbol{w}_j; \boldsymbol{x}) \Big) + h_{q_{\sim j}} - c. \tag{4.37}$$

By inserting (4.37) into (4.26), we observe that the CAVI optimization problem reduces to maximizing the negative KLD (which is equivalent to minimizing the positive KLD) since $h_{q_{\sim j}}$

and c do not depend on $\boldsymbol{w}_j$:

$$\begin{aligned} q_j^{(\ell)}(\boldsymbol{w}_j; \boldsymbol{x}) &= \underset{q_j(\boldsymbol{w}_j)\in\mathcal{F}_j}{\arg\max} \Big\{ -D_{\mathrm{KL}}\Big(q_j(\boldsymbol{w}_j) \,||\, \check{f}(\boldsymbol{w}_j; \boldsymbol{x})\Big) + h_{q_{\sim j}} - c \Big\} \\ &= \arg \min_{q_j(\boldsymbol{w}_j)\in\mathcal{F}_j} D_{\mathrm{KL}}\Big(q_j(\boldsymbol{w}_j) \,||\, \check{f}(\boldsymbol{w}_j; \boldsymbol{x})\Big) \\ &= \check{f}(\boldsymbol{w}_j; \boldsymbol{x}). \end{aligned}$$

Note that $q_j^{(\ell)}(\boldsymbol{w}_j;\, \boldsymbol{x}) = \check{f}(\boldsymbol{w}_j; \boldsymbol{x})$ in the last step because we consider the unconstrained set of pdfs $\mathcal{F}_j$ for the minimization of $D_{\mathrm{KL}}\Big(q_j(\boldsymbol{w}_j) \,||\, \check{f}(\boldsymbol{w}_j; \boldsymbol{x})\Big)$. Recalling (4.36), this shows that the log of the j-th variational pdf factor in the ℓ-th iteration of the CAVI algorithm is given by

$$\begin{aligned} \ln q_j^{(\ell)}(\boldsymbol{w}_j; \boldsymbol{x}) &= \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}, \boldsymbol{x})\} + c \\ &\overset{\mathrm{c}}{=} \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}, \boldsymbol{x})\} \\ &= \int_{\mathbb{R}^{P_{\sim j}}} q_{\sim j}(\boldsymbol{w}_{\sim j}) \ln f(\boldsymbol{w}, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{w}_{\sim j} \\ &= \int_{\mathbb{R}^{P_{\sim j}}} \ln f(\boldsymbol{w}, \boldsymbol{x}) \prod_{i\neq j} q_i(\boldsymbol{w}_i) \,\mathrm{d}\boldsymbol{w}_i, \end{aligned} \tag{4.38}$$

where $\overset{\mathrm{c}}{=}$ denotes equality up to an additive constant c. Equivalently, we have

$$q_j^{(\ell)}(\boldsymbol{w}_j; \boldsymbol{x}) \propto \exp\Big(\mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}, \boldsymbol{x})\}\Big). \tag{4.39}$$

Furthermore, it is possible to express the CAVI update (4.38) in terms of the conditional pdf $f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x})$, which is referred to as the complete conditional. Using the chain rule, we note that

$$f(\boldsymbol{w}, \boldsymbol{x}) = f(\boldsymbol{w}_j, \boldsymbol{w}_{\sim j}, \boldsymbol{x}) = f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x}) f(\boldsymbol{w}_{\sim j}, \boldsymbol{x}),$$

which we insert into (4.38), i.e.,

$$\begin{aligned} \ln q_j^{(\ell)}(\boldsymbol{w};\, \boldsymbol{x}) &\overset{\mathrm{c}}{=} \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x}) f(\boldsymbol{w}_{\sim j}, \boldsymbol{x})\} \\ &= \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x})\} + \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}_{\sim j}, \boldsymbol{x})\}. \end{aligned}$$

Here, $\mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}_{\sim j}, \boldsymbol{x})\}$ does not depend on $\boldsymbol{w}_j$ and can be absorbed into the additive constant. Therefore, the CAVI update (4.38) can be alternatively computed as

$$\begin{aligned} \ln q_j^{(\ell)}(\boldsymbol{w}_j; \boldsymbol{x}) &\overset{\mathrm{c}}{=} \mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x})\} \\ &= \int_{\mathbb{R}^{P_{\sim j}}} q_{\sim j}(\boldsymbol{w}_{\sim j}) \ln f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{w}_{\sim j} \\ &= \int_{\mathbb{R}^{P_{\sim j}}} \ln f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x}) \prod_{i\neq j} q_i(\boldsymbol{w}_i) \,\mathrm{d}\boldsymbol{w}_i, \end{aligned} \tag{4.40}$$

or equivalently

$$q_j^{(\ell)}(\boldsymbol{w}_j; \boldsymbol{x}) \propto \exp\Big(\mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x})\}\Big). \tag{4.41}$$

4.3.2 Coordinate Ascent Variational Inference Algorithm

Equations (4.39) and (4.41) show that the solution of the CAVI optimization problem (4.26) can be determined by calculating the expectation of the logarithm of the joint pdf $f(\boldsymbol{w}, \boldsymbol{x})$ or the complete conditional $f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x})$, where the expectation is with respect to $q_{\sim j}(\boldsymbol{w}_{\sim j}) = \prod_{i \neq j} q_i(\boldsymbol{w}_i)$ in both cases. The solution for the ℓ-th iteration and j-th substep $\ln q_j^{(\ell)}(\boldsymbol{w}; \boldsymbol{x})$ involves all other variational pdf factors $q_i(\boldsymbol{w}_i)$, $i \neq j$, which are given by the results of most recent updates. Convergence is determined by calculating the ELBO $L(q^{(\ell)}(\boldsymbol{w}); \boldsymbol{x})$ at the end of each iteration and monitoring the change relative to the previous iteration $\ell - 1$. When the relative change of the ELBO is below some predefined threshold, convergence is declared. This means the algorithm has reached a local optimum. We denote the solution at the final iteration as $q^*(\boldsymbol{w}; \boldsymbol{x})$. A summary of the CAVI algorithm is provided by Algorithm 1.

Algorithm 1: General formulation of CAVI

Input: Observations $\boldsymbol{x}$, model pdf $f(\boldsymbol{w}, \boldsymbol{x})$ or complete conditionals $f(\boldsymbol{w}_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x})$, mean field factorization of $q(\boldsymbol{w})$

Output: Variational factor pdfs $q_j^*(\boldsymbol{w}_j; \boldsymbol{x})$, for $j = 1, \ldots, J$

Initialize: Variational factor pdfs $q_j^{(0)}(\boldsymbol{w}_j)$, for $j = 1, \ldots, J$ $(\ell = 0)$

while *the ELBO has not converged* **do**

  $\ell = \ell + 1$

  for j *from* 1 *to* J **do**

    Compute $q_j^{(\ell)}(\boldsymbol{w}_j; \boldsymbol{x})$ according to (4.39) or (4.41)

  Compute the ELBO $L(q^{(\ell)}(\boldsymbol{w}); \boldsymbol{x})$ and check convergence

return $q^*(\boldsymbol{w}; \boldsymbol{x}) = \prod_{j=1}^{J} q_j^*(\boldsymbol{w}_j; \boldsymbol{x})$

Note that the concepts explained so far in this chapter can be reformulated for the cases of discrete and mixed continuous-and-discrete parameters $\boldsymbol{w}_j$. This can be done by replacing pdfs with pmfs, and the subsequent modification of expectations from integrals to sums. We will consider the mixed case in the application presented next.

4.4 Coordinate Ascent Variational Inference for Dirichlet Process Mixtures

In this section, we apply the CAVI method to the exponential family DPM model described in Section 3.5, where the joint pdf $f(\boldsymbol{w}, \boldsymbol{x})$ of the statistical model is given by (3.49). We will first define a mean field variational family $\mathcal{Q}$ for the variational approximation $q(\boldsymbol{w})$ and then derive the corresponding update equations for the variational factors $q_j(\boldsymbol{w}_j)$. After that, we proceed with a derivation of the ELBO $L(q; \boldsymbol{x})$ and a final summary of the CAVI algorithm.

4.4.1 Truncated Mean Field Approximation

Following [7], we now define the mean field variational family $\mathcal{Q}$ for the approximation of the posterior distribution $f(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*|\boldsymbol{x})$ of the exponential family DPM model (3.41). The hidden parameters in this model are the scalar auxiliary variables $v_k \in [0, 1]$, $k = 1, 2, \ldots$, the natural

parameter vectors $\boldsymbol{\eta}_k^* \in \mathbb{R}^p$, $k = 1, 2, \ldots$, and the scalar indicator variables $z_n \in \mathbb{N}$, $n = 1, \ldots, N$, i.e., $\boldsymbol{w} = \left(\boldsymbol{v}^{\mathrm{T}} \ \boldsymbol{\eta}^{*\mathrm{T}} \ \boldsymbol{z}^{\mathrm{T}}\right)^{\mathrm{T}}$ with $\boldsymbol{v} = (v_1 \ v_2 \ \cdots)^{\mathrm{T}}$, $\boldsymbol{\eta}^* = \left(\boldsymbol{\eta}_1^{*\mathrm{T}} \ \boldsymbol{\eta}_2^{*\mathrm{T}} \ \cdots\right)^{\mathrm{T}}$ and $\boldsymbol{z} = (z_1 \ \cdots \ z_N)^{\mathrm{T}}$. Note that the number of scalar parameters $P = \dim(\boldsymbol{w})$ is infinite in this model. The (deterministic) hyperparameters are given by the concentration parameter α of the stick-breaking process and the $(p+1)$-dimensional vector $\boldsymbol{\lambda} = \left(\boldsymbol{\lambda}_1{}^{\mathrm{T}} \ \lambda_2\right)^{\mathrm{T}}$ of the conjugate base distribution $G_0(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda})$.

In order to approximate the posterior $f(\boldsymbol{w}|\boldsymbol{x})$ for the infinite-dimensional parameter vector $\boldsymbol{w}$ in closed form, we consider a so-called truncated stick-breaking representation for the DPM, i.e., we truncate the infinite sequence of variables v_k and $\boldsymbol{\eta}_k^*$ at the truncation level $T \in \mathbb{N}$. To ensure the truncation of the stick-breaking process (cf. (3.9)) we set the T-th auxiliary variable to a fixed value $v_T = 1$, i.e., we stop breaking the unit length stick in the T-th iteration of the stick-breaking process. According to (3.9a) this implies that the mixture proportions $\pi_t = v_t \prod_{i=1}^{t-1}(1 - v_t) = 0$ for all $t > T$. As a consequence of the truncation we have $\boldsymbol{v} = (v_1 \ \cdots \ v_{T-1})^{\mathrm{T}}$ and $\boldsymbol{\eta}^* = \left(\boldsymbol{\eta}_1^{*\mathrm{T}} \ \cdots \ \boldsymbol{\eta}_T^{*\mathrm{T}}\right)^{\mathrm{T}}$, where v_T is not included in the parameter vector $\boldsymbol{v}$ because it is a deterministic quantity. Hence, we approximate an infinite mixture model, specifically the DPM, with a finite mixture consisting of T components by employing a truncated stick-breaking process. The dimension of the parameter vector $\boldsymbol{w}$ of the truncated DPM model is given by

$$\begin{aligned} P &= \dim(\boldsymbol{w}) \\ &= \dim(\boldsymbol{v}) + \dim(\boldsymbol{\eta}^*) + \dim(\boldsymbol{z}) \\ &= T - 1 + Tp + N. \end{aligned} \tag{4.42}$$

Based on the truncated parameter vectors $\boldsymbol{v}$ and $\boldsymbol{\eta}^*$, and the vector of indicator variables $\boldsymbol{z}$, we define a fully factorized variational family of variational distributions $q(\boldsymbol{w})$ as

$$q(\boldsymbol{w}) = q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) \triangleq q(\boldsymbol{v}) q(\boldsymbol{\eta}^*) q(\boldsymbol{z}) = \left(\prod_{t=1}^{T-1} q_t(v_t)\right)\left(\prod_{t=1}^{T} q_t(\boldsymbol{\eta}_t^*)\right)\left(\prod_{n=1}^{N} q_n(z_n)\right), \tag{4.43}$$

where $\mathcal{Q}$ is constructed according to (4.24). Note that we do not restrict the functional form of the individual pdfs in the products of (4.43), we only restrict the variational distribution $q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z})$ to be factorized into $J = 2T - 1 + N$ individual pdfs (cf. (4.23)). Consequently, the optimal solution $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{x})$ also factorizes according to (4.43) into $J = 2T - 1 + N$ individual pdfs (cf. (4.25)).

We can summarize the truncated version of (3.41) for N conditionally independent observations $\boldsymbol{x}_n$ as

$$v_t \overset{\text{i.i.d.}}{\sim} \mathcal{B}(v_t; 1, \alpha), \tag{4.44a}$$

$$v_T = 1, \tag{4.44b}$$

$$\pi_t(\boldsymbol{v}_{1:t}) = v_t \prod_{i=1}^{t-1}(1 - v_i), \tag{4.44c}$$

$$\boldsymbol{\eta}_t^* \overset{\text{i.i.d.}}{\sim} G_0(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda}), \tag{4.44d}$$

$$z_n|\boldsymbol{\pi}(\boldsymbol{v}) \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n|\boldsymbol{\pi}(\boldsymbol{v}); v_T), \tag{4.44e}$$

$$\boldsymbol{x}_n|z_n, \boldsymbol{\eta}_1^*, \dots, \boldsymbol{\eta}_T^* \sim f(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*), \tag{4.44f}$$

for $t = 1, \dots, T$ and $n = 1, \dots, N$. For the base distribution $G_0(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda})$ and component distribution $f(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)$ we have

$$G_0(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda}) = b(\boldsymbol{\lambda}) \exp(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)) \tag{4.45}$$

and

$$f(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*) = \prod_{t=1}^{T} \Big(h(\boldsymbol{x}_n) \exp(\boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*)) \Big)^{\mathbb{1}(z_n = t)} \tag{4.46}$$

$$= h(\boldsymbol{x}_n) \exp(\boldsymbol{\eta}_{z_n}^{*\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_{z_n}^*)). \tag{4.47}$$

The conditional or unconditional independence relations are similar to (3.45) given by

$$v_t \perp\!\!\!\perp v_{t'}, \quad \text{where } t = t' = 1, \dots, T-1 \text{ with } t \neq t', \tag{4.48a}$$

$$\boldsymbol{\eta}_t^* \perp\!\!\!\perp \boldsymbol{\eta}_{t'}^*, \quad \text{where } t = t' = 1, \dots, T \text{ with } t \neq t', \tag{4.48b}$$

$$\boldsymbol{\eta}_t^* \perp\!\!\!\perp v_t \text{ and } \boldsymbol{\eta}_t^* \perp\!\!\!\perp v_t \mid \boldsymbol{z}, \quad \text{for all } t = 1, \dots, T, \tag{4.48c}$$

$$\boldsymbol{\eta}_t^* \perp\!\!\!\perp z_n \mid \boldsymbol{v}, \quad \text{for all } t = 1, \dots, T \text{ and } n = 1, \dots, N, \tag{4.48d}$$

$$z_n \perp\!\!\!\perp z_{n'} \mid \boldsymbol{v}, \quad \text{where } n, n' = 1, \dots, N \text{ with } n \neq n', \tag{4.48e}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{v}, z_{n'}, x_{n'}, \boldsymbol{\eta}_t^* \mid z_n, \boldsymbol{\eta}_{z_n}^*, \quad \text{where } n, n' = 1, \dots, N \text{ with } n \neq n' \text{ and } t \neq z_n. \tag{4.48f}$$

We will denote joint pdfs (or pmfs) that involve a truncation with $f^{(T)}$, e.g., $f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*)$ is the joint pdf of the model according to which we assume the observation $\boldsymbol{x}$ to be generated from, while $f^{(T)}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*)$ is the truncated version used for the mean field approximation. Using the independence relations (4.48) and proceeding as in the derivation of (3.49) yields the joint pdf of the truncated stick-breaking model:

$$f^{(T)}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\eta}^*) = \left(\prod_{n=1}^{N} f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*) p^{(T)}(z_n|\boldsymbol{v}) \right) \left(\prod_{t=1}^{T} f(\boldsymbol{\eta}_t^*) \right) \left(\prod_{t=1}^{T-1} f(v_t) \right) \tag{4.49}$$

$$= \left(\prod_{n=1}^{N} f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*) \mathcal{C}(z_n|\boldsymbol{\pi}(\boldsymbol{v}); v_T) \right) \left(\prod_{t=1}^{T} G_0(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda}) \right) \left(\prod_{t=1}^{T-1} \mathcal{B}(v_t; 1, \alpha) \right). \tag{4.50}$$

Figure 4.2 shows the corresponding Bayesian network (see Figure 3.10 for a comparison).

It is important to note, that the stick-breaking construction of the underlying statistical model (3.41), which we will refer to as observation model, is not truncated. We only truncate the stick-breaking process of the approximating model (4.44) in order to be able to computationally approximate the posterior pdf in closed form. This is because the true posterior pdf involves an infinite number of scalar parameters, while the truncated approximation only involves $P = T - 1 + Tp + N$ scalar parameters (see (4.42)). The truncation level T is a parameter that must be chosen initially and its choice trades-off computationally efficiency and approximation

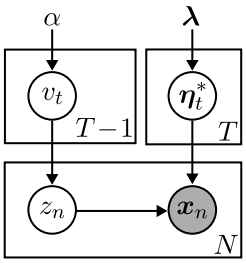


Figure 4.2: Bayesian network for the truncated stick-breaking approximation of the DPM. The corresponding joint pdf is given by (4.50).

accuracy. It is not part of the prior model specification (3.41) and as $T \to \infty$ the truncated stick-breaking approximation (4.44) becomes exact.

4.4.2 Derivation of the Variational Parameters

In what follows, we will evaluate the CAVI update for each factor pdf in the mean field factorization (4.43) according to (4.41). We will observe that the update equations each reduce to a closed-form solution given by the computation of a variational parameter which fully determines the respective variational pdf factor. For the sake of brevity, we will skip the truncation parameter T in the superscript of the pdfs $f^{(T)}(\cdot)$ and pmfs $p^{(T)}(\cdot)$ in the derivation of the variational parameters and note that in the following all pdfs and pmfs are related to the truncated model (4.44).

Variational Parameter for $q_t(v_t)$

We start by considering the variational pdf factor $q(\boldsymbol{v})$ of the auxiliary variable vector $\boldsymbol{v} = (v_1 \ \ldots \ v_{T-1})^{\mathrm{T}}$ and, later on, apply the factorization $q(\boldsymbol{v}) = \prod_{t=1}^{T-1} q(v_t)$ to obtain the pdf factor $q(v_t)$. According to (4.40), for the case of $q(\boldsymbol{v})$, we have to evaluate

$$\ln q^{(\ell)}(\boldsymbol{v}; \boldsymbol{x}) \overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{\eta}^*, \boldsymbol{z}))}\{\ln f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x})\}, \tag{4.51}$$

i.e., the expectation with respect to all other random parameters except $\boldsymbol{v}$ of the log complete conditional of $\boldsymbol{v}$. In order to compute this expectation we will first derive an expression for the complete conditional $f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x})$. Applying Bayes rule we have

$$f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z})}{f(\boldsymbol{x}|\boldsymbol{\eta}^*, \boldsymbol{z})}. \tag{4.52}$$

This can be further simplified by using the independence relations (4.48c) and the fact that the knowledge of $\boldsymbol{v}$ is contained in $\boldsymbol{z}$, i.e., $f(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) = f(\boldsymbol{x}|\boldsymbol{\eta}^*, \boldsymbol{z})$, which results

$$f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\eta}^*, \boldsymbol{z}) f(\boldsymbol{v}|\boldsymbol{z})}{f(\boldsymbol{x}|\boldsymbol{\eta}^*, \boldsymbol{z})} = f(\boldsymbol{v}|\boldsymbol{z}).$$

Here, we again apply Bayes rule and put variables that are constant with respect to $\boldsymbol{v}$ into a proportional factor, i.e.,

$$f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) = \frac{p(\boldsymbol{z}|\boldsymbol{v})f(\boldsymbol{v})}{p(\boldsymbol{z})} \propto p(\boldsymbol{z}|\boldsymbol{v})f(\boldsymbol{v}) = \left(\prod_{n=1}^{N} p(z_n|\boldsymbol{v})\right) \prod_{t=1}^{T-1} f(v_t), \tag{4.53}$$

where the last step is due to the conditional independence among indicator variables z_n (see (4.48e)) and the independence among auxiliary variables v_t (see (4.48a)). According to (4.53), the complete conditional for $\boldsymbol{v}$ statistically depends only on the indicator variables $\boldsymbol{z}$ since we removed all other variables from the condition by the corresponding independence relations.

The conditional pmf $p(z_n|\boldsymbol{v})$ can be expressed as

$$\begin{aligned} p(z_n|\boldsymbol{v}) &= \prod_{t=1}^{T} (\pi_t(\boldsymbol{v}_{1:t}))^{\mathbb{1}(z_n=t)} \\ &= \prod_{t=1}^{T} \left(v_t \prod_{j=1}^{t-1}(1-v_j)\right)^{\mathbb{1}(z_n=t)} \qquad (4.54) \\ &= \prod_{t=1}^{T} v_t^{\mathbb{1}(z_n=t)} \prod_{j=1}^{t-1}(1-v_j)^{\mathbb{1}(z_n=t)}, \qquad (4.55) \end{aligned}$$

where we used the fact that z_n given $\boldsymbol{v}$ is distributed according to a categorical distribution (cf. (4.44e)) with probabilities π_t, $t = 1, \ldots, T$, given by (4.44c). Using the indicator function $\mathbb{1}(z_n > t)$ allows us to aggregate all factors in (4.55) into one product, i.e.,

$$p(z_n|\boldsymbol{v}) = \prod_{t=1}^{T} v_t^{\mathbb{1}(z_n=t)}(1-v_t)^{\mathbb{1}(z_n>t)}$$

and by considering $v_T = 1$ (see (4.44b)) we obtain

$$p(z_n|\boldsymbol{v}) = \prod_{t=1}^{T-1} v_t^{\mathbb{1}(z_n=t)}(1-v_t)^{\mathbb{1}(z_n>t)}. \tag{4.56}$$

We are now able to compute the CAVI update (4.51). Inserting (4.53) for the complete conditional $f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x})$ and using the linearity property (4.28a) of the expectation operator gives

$$\begin{aligned} \ln q^{(\ell)}(\boldsymbol{v}; \boldsymbol{x}) &\overset{c}{=} \mathrm{E}^{(q(\boldsymbol{\eta}^*, \boldsymbol{z}))}\left\{\ln\left(\left(\prod_{n=1}^{N} p(z_n|\boldsymbol{v})\right)\prod_{t=1}^{T-1} f(v_t)\right)\right\} \\ &= \left(\sum_{n=1}^{N} \mathrm{E}^{(q(\boldsymbol{\eta}^*, \boldsymbol{z}))}\{\ln p(z_n|\boldsymbol{v})\}\right) + \sum_{t=1}^{T-1} \mathrm{E}^{(q(\boldsymbol{\eta}^*, \boldsymbol{z}))}\{\ln f(v_t)\}. \end{aligned}$$

If we furthermore consider the marginalization property (4.29), we obtain

$$\ln q^{(\ell)}(\boldsymbol{v}; \boldsymbol{x}) = \left(\sum_{n=1}^{N} \mathrm{E}^{(q_n(z_n))}\{\ln p(z_n|\boldsymbol{v})\}\right) + \sum_{t=1}^{T-1} \ln f(v_t). \tag{4.57}$$

Here, it remains to compute the expectation $\mathrm{E}^{(q_n(z_n))}\{\ln p(z_n|\boldsymbol{v})\}$. Substituting $p(z_n|\boldsymbol{v})$ with (4.56) yields

$$\begin{aligned}\mathrm{E}^{(q_n(z_n))}\{\ln p(z_n|\boldsymbol{v})\} &= \mathrm{E}^{(q_n(z_n))}\left\{\ln \prod_{t=1}^{T-1} v_t^{\mathbb{1}(z_n=t)}(1-v_t)^{\mathbb{1}(z_n>t)}\right\} \\ &= \mathrm{E}^{(q_n(z_n))}\left\{\sum_{t=1}^{T-1} \ln\left(v_t^{\mathbb{1}(z_n=t)}\right) + \ln\left((1-v_t)^{\mathbb{1}(z_n>t)}\right)\right\}.\end{aligned}$$

Again using the linearity property (4.28) results in

$$\begin{aligned}\mathrm{E}^{(q_n(z_n))}\{\ln p(z_n|\boldsymbol{v})\} &= \sum_{t=1}^{T-1} \mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\ln v_t + \mathbb{1}(z_n>t)\ln(1-v_t)\} \\ &= \sum_{t=1}^{T-1} \ln v_t \mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\} + \ln(1-v_t)\mathrm{E}^{(q(z_n))}\{\mathbb{1}(z_n>t)\}. \qquad (4.58)\end{aligned}$$

Furthermore, the expectation $\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\}$ in (4.58) reduces to the probability $q_n(z_n=t)$ (note that $q_n(z_n)$ is a variational pmf since $z_n \in \{1,\dots,T\}$ is a discrete random variable), i.e.,

$$\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\} = \sum_{z_n=1}^{T} \mathbb{1}(z_n=t)q_n(z_n) = q_n(z_n=t). \qquad (4.59)$$

This is due to the indicator function $\mathbb{1}(z_n=t)$ which reduces the sum in (4.59) to the term $q_n(z_n=t)$. We denote the T probabilities $q_n(z_n=t)$, $t=1,\dots,T$, with

$$\phi_{n,t} \triangleq q_n(z_n=t). \qquad (4.60)$$

As $q_n(z_n)$ is the approximate posterior distribution of the indicator variable z_n, the probability $\phi_{n,t}$ is the approximate posterior probability that the n-th data point $\boldsymbol{x}_n$ was generated by the t-th mixture component and thus can be interpreted as a soft cluster assignment. Using (4.60) the expectation $\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n>t)\}$ is given by

$$\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n>t)\} = \sum_{z_n=1}^{T} \mathbb{1}(z_n>t)q_n(z_n) = \sum_{j=t+1}^{T} q_n(z_n=j) = \sum_{j=t+1}^{T} \phi_{n,j}. \qquad (4.61)$$

Inserting the two expectations (4.59) and (4.61) into (4.58) then results

$$\mathrm{E}^{(q_n(z_n))}\{\ln p(z_n|\boldsymbol{v})\} = \sum_{t=1}^{T-1} \phi_{n,t}\ln v_t + \left(\sum_{j=t+1}^{T} \phi_{n,j}\right)\ln(1-v_t), \qquad (4.62)$$

and furthermore substituting (4.62) into (4.57) results

$$\begin{aligned}\ln q^{(\ell)}(\boldsymbol{v};\boldsymbol{x}) &\overset{\mathrm{c}}{=} \left(\sum_{n=1}^{N}\sum_{t=1}^{T-1} \phi_{n,t}\ln v_t + \left(\sum_{j=t+1}^{T} \phi_{n,j}\right)\ln(1-v_t)\right) + \sum_{t=1}^{T-1} \ln f(v_t) \\ &= \sum_{t=1}^{T-1}\left(\sum_{n=1}^{N} \phi_{n,t}\ln v_t + \left(\sum_{j=t+1}^{T} \phi_{n,j}\right)\ln(1-v_t)\right) + \ln f(v_t). \qquad (4.63)\end{aligned}$$

From (4.63) we can conclude that the log variational pdf $\ln q^{(\ell)}(\boldsymbol{v};\boldsymbol{x})$ consists of $T-1$ independent terms that are given by

$$\ln q_t^{(\ell)}(v_t;\boldsymbol{x}) \overset{\text{c}}{=} \left(\sum_{n=1}^{N}\phi_{n,t}\ln v_t + \left(\sum_{j=t+1}^{T}\phi_{n,j}\right)\ln(1-v_t)\right) + \ln f(v_t). \tag{4.64}$$

Note that the factorization $q(\boldsymbol{v}) = \prod_{t=1}^{T-1} q_t(v_t)$ (see (4.43)), which is equivalent to $\ln q(\boldsymbol{v}) = \sum_{t=1}^{T-1}\ln q_t(v_t)$, naturally arises from the independence relations within model.

As a final step, we insert the distribution $f(v_t)$ into (4.64). Recalling that the auxiliary variable v_t is distributed according to $\mathcal{B}(v_t;1,\alpha)$ gives

$$f(v_t) = \mathcal{B}(v_t;1,\alpha) = \frac{\Gamma(1+\alpha)}{\Gamma(1)\Gamma(\alpha)}v_t^{1-1}(1-v_t)^{\alpha-1} = \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)}(1-v_t)^{\alpha-1},$$

where we used (3.1) with $\alpha_1 = 1$ and $\alpha_2 = \alpha$. Equivalently, we have

$$\ln f(v_t) = \ln\frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} + (\alpha-1)\ln(1-v_t). \tag{4.65}$$

Inserting (4.65) into (4.64) then yields

$$\begin{aligned}\ln q_t^{(\ell)}(v_t;\boldsymbol{x}) &\overset{\text{c}}{=} \left(\sum_{n=1}^{N}\phi_{n,t}\ln v_t + \left(\sum_{j=t+1}^{T}\phi_{n,j}\right)\ln(1-v_t)\right) + \ln\frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} + (\alpha-1)\ln(1-v_t)\\ &\overset{\text{c}}{=} \left(\sum_{n=1}^{N}\phi_{n,t}\ln v_t + \left(\sum_{j=t+1}^{T}\phi_{n,j}\right)\ln(1-v_t)\right) + (\alpha-1)\ln(1-v_t),\end{aligned}$$

where $\ln(\Gamma(1+\alpha)/\Gamma(\alpha))$ can be put into the normalization constant c because it is constant with respect to v_t. By exponentiation, we get

$$\begin{aligned}q_t^{(\ell)}(v_t;\boldsymbol{x}) &\propto \exp\left(\ln v_t\sum_{n=1}^{N}\phi_{n,t}\right)\exp\left(\ln(1-v_t)\sum_{n=1}^{N}\sum_{j=t+1}^{T}\phi_{n,j}\right)(1-v_t)^{\alpha-1}\\ &= v_t^{\sum_{n=1}^{N}\phi_{n,t}}(1-v_t)^{\sum_{n=1}^{N}\sum_{j=t+1}^{T}\phi_{n,j}}(1-v_t)^{\alpha-1}\\ &= v_t^{\left(1+\sum_{n=1}^{N}\phi_{n,t}\right)-1}(1-v_t)^{\left(\alpha+\sum_{n=1}^{N}\sum_{j=t+1}^{T}\phi_{n,j}\right)-1}.\end{aligned} \tag{4.66}$$

From the last step it can be observed that the functional form of the variational factor $q_t^{(\ell)}(v_t;\boldsymbol{x})$ is given by a beta distribution. Comparing (4.66) with (3.1) yields

$$q_t^{(\ell)}(v_t;\boldsymbol{x}) = \frac{\Gamma\left(\gamma_{t,1}^{(\ell)}+\gamma_{t,2}^{(\ell)}\right)}{\Gamma\left(\gamma_{t,1}^{(\ell)}\right)\Gamma\left(\gamma_{t,2}^{(\ell)}\right)}v_t^{\gamma_{t,1}^{(\ell)}-1}(1-v_t)^{\gamma_{t,2}^{(\ell)}-1}, \tag{4.67}$$

with

$$\gamma_{t,1}^{(\ell)} = 1 + \sum_{n=1}^{N}\phi_{n,t}, \tag{4.68a}$$

$$\gamma_{t,2}^{(\ell)} = \alpha + \sum_{n=1}^{N} \sum_{j=t+1}^{T} \phi_{n,j}. \tag{4.68b}$$

We conclude that the CAVI update for $q_t^{(\ell)}(v_t; \boldsymbol{x})$ is fully determined by the parameter vector $\boldsymbol{\gamma}_t^{(\ell)} = \begin{pmatrix} \gamma_{t,1}^{(\ell)} & \gamma_{t,2}^{(\ell)} \end{pmatrix}^{\mathrm{T}}$, which depends on the data $\boldsymbol{x}$ through the most recent update for $\phi_{n,t} = q_n(z_n = t)$. We will refer to $\boldsymbol{\gamma}_t^{(\ell)}$ as a (global) variational parameter. It parameterizes the approximate posterior $q_t^{(\ell)}(v_t; \boldsymbol{x}) = \mathcal{B}\left(v_t; \gamma_{t,1}^{(\ell)}, \gamma_{t,2}^{(\ell)}\right)$ of the (global) model parameter v_t. Recall that the prior pdf of v_t is also given by a beta distribution but with parameters 1 and α, i.e., $\mathcal{B}(v_t; 1, \alpha)$. Thus, we interpret the update equations (4.68) as updating the parameters of the prior $\mathcal{B}(v_t; 1, \alpha)$ in each iteration of the CAVI algorithm.

By defining the effective number of samples $\tilde{N}_t$ associated with mixture component t as

$$\tilde{N}_t \triangleq \sum_{n=1}^{N} \phi_{n,t}, \tag{4.69}$$

we can rewrite the update (4.68) as

$$\gamma_{t,1}^{(\ell)} = 1 + \tilde{N}_t, \tag{4.70a}$$

$$\gamma_{t,2}^{(\ell)} = \alpha + \sum_{j=t+1}^{T} \tilde{N}_j, \tag{4.70b}$$

which closely resembles the parameters of the posterior distribution $f(\boldsymbol{\pi}|\boldsymbol{z})$ given by (3.23). We just have to replace the true (unknown) number of samples associated with component t, i.e., $N_t \in \mathbb{N}$, by the effective number of samples $\tilde{N}_t \in \mathbb{R}^+$. These can be computed with the most recent soft cluster assignments $\phi_{n,t}$. Note that $\sum_{t=1}^{T} \phi_{n,t} = \sum_{t=1}^{T} q(z_n = t) = 1$ and therefore

$$\sum_{t=1}^{T} \tilde{N}_t = \sum_{t=1}^{T} \sum_{n=1}^{N} \phi_{n,t} = \sum_{n=1}^{N} \sum_{t=1}^{T} \phi_{n,t} = \sum_{n=1}^{N} 1 = N,$$

i.e., the sum of $\tilde{N}_t$ and the sum of N_t both equal the overall sample size N (cf. (2.12)).

Variational Parameter for $q_t(\boldsymbol{\eta}_t^*)$

We proceed with the derivation of the CAVI update of the variational pdf factor $q_t(\boldsymbol{\eta}_t^*)$ by first deriving the CAVI update for the joint pdf $q(\boldsymbol{\eta}^*)$ of the parameter vector $\boldsymbol{\eta}^* = \begin{pmatrix} \boldsymbol{\eta}_1^{*\mathrm{T}} & \dots & \boldsymbol{\eta}_T^{*\mathrm{T}} \end{pmatrix}^{\mathrm{T}}$. According to (4.40), we have to evaluate

$$\ln q^{(\ell)}(\boldsymbol{\eta}^*; \boldsymbol{x}) \overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{z}))}\{\ln f(\boldsymbol{\eta}^*|\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{x})\}, \tag{4.71}$$

which means we first have to find an expression for the complete conditional of $\boldsymbol{\eta}^*$. Similarly to (4.52), we approach this by using Bayes rule such that the likelihood $f(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z})$ appears in the numerator.

$$f(\boldsymbol{\eta}^*|\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) f(\boldsymbol{\eta}^*|\boldsymbol{v}, \boldsymbol{z})}{f(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{z})}. \tag{4.72}$$

The likelihood can again be simplified as $f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}) = f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})$. A simplification of $f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z})$ with respect to the independence relations (4.48c) and (4.48d) yields $f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z}) = f(\boldsymbol{\eta}^*|\boldsymbol{v}) = f(\boldsymbol{\eta}^*)$ and thus

$$f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z}) f(\boldsymbol{\eta}^*),$$

where we omitted the denominator of (4.72) because it is constant with respect to $\boldsymbol{\eta}^*$. Furthermore, using the independence relations (4.48b) and (4.48f), we have

$$f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x}) \propto \left(\prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*, z_n)\right) \prod_{t=1}^{T} f(\boldsymbol{\eta}_t^*).$$

Here, $f(\boldsymbol{x}_n|\boldsymbol{\eta}^*, z_n) = f(\boldsymbol{x}_n|\boldsymbol{\eta}_1^*,\dots,\boldsymbol{\eta}_T^*, z_n) = f(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)$ (cf. (4.44f)) is given by (4.46) and $f(\boldsymbol{\eta}_t^*)$ is given by (4.45). Inserting both distributions, and omitting the factors involving $b(\boldsymbol{\lambda})$ and $h(\boldsymbol{x}_n)$ since they do not depend on $\boldsymbol{\eta}^*$, results in

$$\begin{aligned} f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x}) &\propto \left(\prod_{n=1}^{N}\prod_{t=1}^{T}\Big(\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))\Big)^{\mathbb{1}(z_n=t)}\right)\prod_{t=1}^{T}\exp(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)) \\ &= \prod_{t=1}^{T}\exp(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)) \prod_{n=1}^{N}\Big(\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))\Big)^{\mathbb{1}(z_n=t)}. \end{aligned} \tag{4.73}$$

Using expression (4.73) of the complete conditional for $\boldsymbol{\eta}^*$ we are now able to further develop the CAVI update (4.71), i.e.,

$$\begin{aligned} \ln q^{(\ell)}(\boldsymbol{\eta}^*;\boldsymbol{x}) &\overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{z}))}\left\{\ln\prod_{t=1}^{T}\exp(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*))\prod_{n=1}^{N}\Big(\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))\Big)^{\mathbb{1}(z_n=t)}\right\} \\ &= \sum_{t=1}^{T}\mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{z}))}\left\{\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N}\mathbb{1}(z_n=t)({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))\right\} \\ &= \sum_{t=1}^{T}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N}\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\}({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*)), \end{aligned}$$

where we used the properties (4.28) and (4.29). Inserting $\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\} = \phi_{n,t}$ (see (4.59)) we obtain

$$\ln q^{(\ell)}(\boldsymbol{\eta}^*;\boldsymbol{x}) = \sum_{t=1}^{T}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N}\phi_{n,t}({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*)). \tag{4.74}$$

According to (4.74) the solution for $\ln q^{(\ell)}(\boldsymbol{\eta}^*;\boldsymbol{x})$ depends on T independent terms that are given by

$$\ln q_t^{(\ell)}(\boldsymbol{\eta}_t^*;\boldsymbol{x}) \overset{\mathrm{c}}{=} \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N}\phi_{n,t}({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*)), \tag{4.75}$$

which is, as explained for the update of the variational pdf of the auxiliary variable v_t, a

consequence of the independence relations within the model. Exponentiation of (4.75) yields

$$\begin{aligned} q^{(\ell)}(\boldsymbol{\eta}_t^*; \boldsymbol{x}) &\propto \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N} \phi_{n,t}(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))\right) \\ &= \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \left(\sum_{n=1}^{N} \phi_{n,t}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}_n)\right)\boldsymbol{\eta}_t^* - \sum_{n=1}^{N}\phi_{n,t} a(\boldsymbol{\eta}_t^*)\right) \\ &= \exp\left(\left(\boldsymbol{\lambda}_1^{\mathrm{T}} + \sum_{n=1}^{N}\phi_{n,t}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}_n)\right)\boldsymbol{\eta}_t^* - \left(\lambda_2 + \sum_{n=1}^{N}\phi_{n,t}\right)a(\boldsymbol{\eta}_t^*)\right). \end{aligned} \tag{4.76}$$

Comparing (4.76) and (4.45) we observe that the functional form of the approximate posterior $q^{(\ell)}(\boldsymbol{\eta}_t^*; \boldsymbol{x})$ is given by the prior distribution $f(\boldsymbol{\eta}_t^*)$, which is due to the conjugacy between the prior and the component distributions of the mixture. Therefore, we can summarize the CAVI update for $q^{(\ell)}(\boldsymbol{\eta}_t^*; \boldsymbol{x})$ as

$$q_t^{(\ell)}(\boldsymbol{\eta}_t^*; \boldsymbol{x}) = b\left(\boldsymbol{\tau}_t^{(\ell)}\right) \exp\left(\boldsymbol{\tau}_{t,1}^{(\ell)\mathrm{T}}\boldsymbol{\eta}_t^* - \tau_{t,2}^{(\ell)} a(\boldsymbol{\eta}_t^*)\right), \tag{4.77}$$

where $\boldsymbol{\tau}_{t,1}^{(\ell)}$ and $\tau_{t,2}^{(\ell)}$ are updated versions of the hyperparameters $\boldsymbol{\lambda}_1$ and λ_2 given by

$$\boldsymbol{\tau}_{t,1}^{(\ell)} = \boldsymbol{\lambda}_1 + \sum_{n=1}^{N}\phi_{n,t}\boldsymbol{t}(\boldsymbol{x}_n), \tag{4.78a}$$

$$\tau_{t,2}^{(\ell)} = \lambda_2 + \sum_{n=1}^{N}\phi_{n,t} = \lambda_2 + \tilde{N}_t, \tag{4.78b}$$

and $b(\boldsymbol{\tau}) \in \mathbb{R}^+$ is a normalization coefficient given by

$$b\left(\boldsymbol{\tau}_t^{(\ell)}\right) = \frac{1}{\int_{\mathbb{R}^p} \exp\left(\boldsymbol{\tau}_{t,1}^{(\ell)\mathrm{T}}\boldsymbol{\eta}^* - \tau_{t,2}^{(\ell)} a(\boldsymbol{\eta}^*)\right) \mathrm{d}\boldsymbol{\eta}^*}.$$

The update equations (4.78) are of similar form compared to the update equations (2.33) for the posterior distribution (2.32), where we assumed a EF likelihood function with conjugate prior. The only difference is that the sufficient statistic $\boldsymbol{t}(\boldsymbol{x}_n)$ in (4.78a) is scaled by the soft cluster assignment $\phi_{n,t}$ of cluster t (cf. (2.33a)) and the effective number of samples $\tilde{N}_t$ associated with component t arises in (4.78b) (instead of the overall number of data points N as does in (2.33b)). We conclude that the approximate posterior pdf $q^{(\ell)}(\boldsymbol{\eta}_t^*; \boldsymbol{x})$ of the global model parameter $\boldsymbol{\eta}_t^*$ is fully determined by the global variational parameter $\boldsymbol{\tau}_t^{(\ell)} = \left(\boldsymbol{\tau}_{t,1}^{(\ell)\mathrm{T}} \;\; \tau_{t,2}^{(\ell)}\right)^{\mathrm{T}}$, which may depend on the data $\boldsymbol{x}$ through the most recent update for $\phi_{n,t} = q_n(z_n = t)$ (cf. (4.78)).

Variational Parameter for $q_n(z_n)$

It remains to find the CAVI update for the variational pmf factor $q_n(z_n)$. Similar to the derivation of the previous updates, we start with the joint factor pmf of the parameter vector $\boldsymbol{z} = (z_1 \;\; \cdots \;\; z_N)^{\mathrm{T}}$. According to (4.40), the update is determined by

$$\ln q^{(\ell)}(\boldsymbol{z}; \boldsymbol{x}) \overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*))}\{\ln p(\boldsymbol{z}|\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{x})\}, \tag{4.79}$$

which involves the complete conditional of $\boldsymbol{z}$. By using Bayes rules we express the complete conditional as follows

$$p(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*)}{f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{\eta}^*)}.$$

Applying the conditional independence relation (3.45d) and $f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}) = f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})$ further yields

$$p(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{v}),$$

where we only kept terms that depend on $\boldsymbol{z}$. The independence relations (4.48e) and (4.48f) allow to factorize this expression with respect to n, i.e.,

$$p(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) \propto \left(\prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*,z_n)\right)\prod_{n=1}^{N} p(z_n|\boldsymbol{v}) = \prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*,z_n)p(z_n|\boldsymbol{v}). \tag{4.80}$$

In (4.80) the likelihood $f(\boldsymbol{x}_n|\boldsymbol{\eta}^*,z_n) = f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})$ is given by (4.46). For the conditional pmf $p(z_n|\boldsymbol{v})$ we use the expression (4.54). Inserting both distributions into (4.80) results in

$$\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) &= \prod_{n=1}^{N}\left(\prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))\Big)^{\mathbb{1}(z_n=t)}\right)\left(\prod_{t=1}^{T}\Big(v_t\prod_{j=1}^{t-1}(1-v_j)\Big)^{\mathbb{1}(z_n=t)}\right)\\
&= \prod_{n=1}^{N}\prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))\Big)^{\mathbb{1}(z_n=t)}\Big(v_t\prod_{j=1}^{t-1}(1-v_j)\Big)^{\mathbb{1}(z_n=t)}\\
&= \prod_{n=1}^{N}\prod_{t=1}^{T}\left(h(\boldsymbol{x}_n)\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))v_t\prod_{j=1}^{t-1}(1-v_j)\right)^{\mathbb{1}(z_n=t)}.
\end{aligned} \tag{4.81}$$

Note that we used (4.54) instead of (4.56) for $p(z_n|\boldsymbol{v})$ because it allowed us to formally join the products with respect to t and the exponent $\mathbb{1}(z_n = t)$ in the first step.

With the expression (4.81) of the complete conditional $p(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x})$ we can now further develop the CAVI update (4.79). By using the properties (4.28) and (4.29) of the expectation operator this leads to the expression

$$\begin{aligned}
\ln q^{(\ell)}(\boldsymbol{z};\boldsymbol{x}) &\overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*))}\left\{\ln\prod_{n=1}^{N}\prod_{t=1}^{T}\left(h(\boldsymbol{x}_n)\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))v_t\prod_{j=1}^{t-1}(1-v_j)\right)^{\mathbb{1}(z_n=t)}\right\}\\
&= \sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{1}(z_n=t)\mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*))}\left\{\ln\left(h(\boldsymbol{x}_n)\exp({\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}_t^*))v_t\prod_{j=1}^{t-1}(1-v_j)\right)\right\}\\
&= \sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{1}(z_n=t)\Big(\ln h(\boldsymbol{x}_n) + \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{{\boldsymbol{\eta}_t^*}^{\mathrm{T}}\right\}\boldsymbol{t}(\boldsymbol{x}_n) - \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}\\
&\qquad + \mathrm{E}^{(q(v_t))}\{\ln v_t\} + \sum_{j=1}^{t-1}\mathrm{E}^{(q(v_t))}\{\ln(1-v_j)\}\Big).
\end{aligned}$$

Here, we introduce the shorthand notation

$$S_{n,t}^{(\ell)} \triangleq \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_t^{*\mathrm{T}}\right\}\boldsymbol{t}(\boldsymbol{x}_n) - \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\} + \mathrm{E}^{(q(v_t))}\{\ln v_t\} + \sum_{j=1}^{t-1}\mathrm{E}^{(q(v_j))}\{\ln(1-v_j)\} \quad (4.82)$$

and obtain

$$\ln q^{(\ell)}(\boldsymbol{z};\boldsymbol{x}) = \sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{1}(z_n=t)\Big(\ln h(\boldsymbol{x}_n) + S_{n,t}^{(\ell)}\Big). \quad (4.83)$$

We observe that the update (4.83) consists of N independent terms given by

$$\ln q_n^{(\ell)}(z_n;\boldsymbol{x}) \overset{\mathrm{c}}{=} \sum_{t=1}^{T}\mathbb{1}(z_n=t)\Big(\ln h(\boldsymbol{x}_n) + S_{n,t}^{(\ell)}\Big). \quad (4.84)$$

By exponentiation, (4.84) can be equivalently written as

$$\begin{aligned} q_n^{(\ell)}(z_n;\boldsymbol{x}) &\propto \exp\left(\sum_{t=1}^{T}\mathbb{1}(z_n=t)\Big(\ln h(\boldsymbol{x}_n) + S_{n,t}^{(\ell)}\Big)\right) \\ &= \prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp\Big(S_{n,t}^{(\ell)}\Big)\Big)^{\mathbb{1}(z_n=t)} \\ &\propto \prod_{t=1}^{T}\exp\Big(S_{n,t}^{(\ell)}\Big)^{\mathbb{1}(z_n=t)}. \end{aligned} \quad (4.85)$$

In the last step we omitted $h(\boldsymbol{x}_n)$ since it is constant with respect to t and thus does not change the shape of the distribution. We conclude that the functional form of (4.85) is given by a categorical distribution (cf. (2.5)) with probabilities $\boldsymbol{\phi}_n^{(\ell)} = \begin{pmatrix}\phi_{n,1}^{(\ell)} & \dots & \phi_{n,T}^{(\ell)}\end{pmatrix}$ that can be obtained by normalizing (4.85), i.e.,

$$q_n^{(\ell)}(z_n;\boldsymbol{x}) = \prod_{t=1}^{T}\phi_{n,t}^{(\ell)\,\mathbb{1}(z_n=t)} \quad (4.86)$$

with

$$\phi_{n,t}^{(\ell)} = \frac{\exp\Big(S_{n,t}^{(\ell)}\Big)}{\sum_{t=1}^{T}\exp\Big(S_{n,t}^{(\ell)}\Big)}. \quad (4.87)$$

Thus, the approximate posterior pmf $q_n^{(\ell)}(z_n;\boldsymbol{x})$ of the local model parameter z_n is given by the distribution $\mathcal{C}\Big(z_n;\boldsymbol{\phi}_n^{(\ell)}\Big)$. Like the prior distribution $\mathcal{C}(z_n|\boldsymbol{\pi})$ of z_n it is a categorical distribution but with updated versions $\phi_{n,t}^{(\ell)}$ of the prior probabilities π_t. The (local) variational parameter $\phi_{n,t}^{(\ell)}$ is the approximate posterior probability that the observation $\boldsymbol{x}_n$ was generated by the t-th mixture component, i.e., $\phi_{n,t}^{(\ell)} \approx f(z_n=t|\boldsymbol{x})$, while π_t is the prior probability for the event $z_n=t$. Compared to a hard assignment, in which each observed data point $\boldsymbol{x}_n$ is associated uniquely with one cluster, $\phi_{n,t}^{(\ell)}$ represents a soft assignment based on an approximate posterior probability. It quantifies the level of uncertainty regarding the optimal assignment.

The CAVI update (4.87) is fully determined by the local variational parameter $\boldsymbol{\phi}_n^{(\ell)}$ which itself can be computed by evaluating the variables $S_{n,t}^{(\ell)}$ via (4.82). First, $S_{n,t}^{(\ell)}$ has to be evaluated for all $t = 1, \dots, T$ and then each $\phi_{n,t}^{(\ell)}$ can be determined via (4.87). The two expectations $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_t^{*\mathrm{T}}\right\}$ and $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}$ in (4.82) depend on the choice of the EF, i.e., on the choice of the base measure $h(\boldsymbol{x}_n)$ and sufficient statistic $\boldsymbol{t}(\boldsymbol{x}_n)$. The remaining expectations $\mathrm{E}^{(q(v_t))}\{\ln v_t\}$ and $\mathrm{E}^{(q(v_t))}\{\ln(1 - v_t)\}$ can be shown [49] to be given by

$$\mathrm{E}^{(q_t(v_t))}\{\ln v_t\} = \Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1} + \gamma_{t,2}), \tag{4.88}$$

$$\mathrm{E}^{(q_t(v_t))}\{\ln(1 - v_t)\} = \Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1} + \gamma_{t,2}). \tag{4.89}$$

Here, $\Psi(\cdot)$ denotes the digamma function, which is defined by the derivative of the logarithm of the gamma function, i.e.,

$$\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma(x).$$

Finally, note that the update (4.87) for the local variational parameter $\boldsymbol{\phi}_n^{(\ell)}$ depends on the global variational parameters $\boldsymbol{\tau}_t$ and $\boldsymbol{\gamma}_t$ (which are given by the most recent respective update) through the expectations in $S_{n,t}^{(\ell)}$. The updates (4.68) and (4.78) of the global variational parameters $\boldsymbol{\tau}_t^{(\ell)}$ and $\boldsymbol{\gamma}_t^{(\ell)}$ only depend on the local variational parameters $\boldsymbol{\phi}_n$ and thus the algorithm alternates between updating the local and global variational parameters.

4.4.3 Derivation of the Evidence Lower Bound

Given the variational pdf $q(\boldsymbol{w}) = q(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z})$ and the observation $\boldsymbol{x}$, we now derive an expression for the ELBO $L(q; \boldsymbol{x})$. To avoid excessive notation, we will abbreviate the joint pdf $q(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z})$ as q in the derivation of the ELBO and thus write $\mathrm{E}^{(q)}\{\cdot\}$ instead of $\mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}*,\boldsymbol{z}))}\{\cdot\}$ for the expectation operator. Also, recall the properties (4.28), (4.29) and (4.30) of expectation, which will be used several times in the following derivation of the ELBO.

We start with the expression (4.14) of the ELBO, i.e.,

$$L(q; \boldsymbol{x}) = \mathrm{E}^{(q)}\left\{\ln f^{(T)}(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z}, \boldsymbol{x})\right\} - \mathrm{E}^{(q)}\{\ln q(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z})\}.$$

Here, the joint pdf $f^{(T)}(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z}, \boldsymbol{x})$ is with respect to the truncated model (4.44), since we used the truncated model for the variational approximation (4.43) (see Section 4.4.1). Inserting the joint pdf (4.49) of the truncated stick-breaking approximation of the DPM model and the corresponding mean field factorization (4.43) of $q(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z})$ yields

$$\begin{aligned} L(q; \boldsymbol{x}) = \mathrm{E}^{(q)}\left\{\left(\sum_{n=1}^{N}\left(\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)\right) + \ln p^{(T)}(z_n|\boldsymbol{v})\right) + \left(\sum_{t=1}^{T} \ln f(\boldsymbol{\eta}_t^*)\right) + \sum_{t=1}^{T-1} \ln f(v_t)\right\} \\ - \mathrm{E}^{(q)}\left\{\left(\sum_{t=1}^{T-1} \ln q_t(v_t)\right) + \left(\sum_{t=1}^{T} \ln q_t(\boldsymbol{\eta}_t^*)\right) + \sum_{n=1}^{N} \ln q_n(z_n)\right\}. \end{aligned}$$

This can be further simplified by using the linearity property (4.28a) of expectation. Conse-

quently, the ELBO splits up into seven distinct expectation values according to

$$\begin{aligned} L(q;\boldsymbol{x}) = &\sum_{n=1}^{N}\Big(\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})\} + \mathrm{E}^{(q)}\{\ln p^{(T)}(z_n|\boldsymbol{v})\} + \mathrm{E}^{(q)}\{\ln q_n(z_n)\}\Big) \\ &+ \sum_{t=1}^{T}\Big(\mathrm{E}^{(q)}\{\ln f(\boldsymbol{\eta}^*_t)\} + \mathrm{E}^{(q)}\{\ln q_t(\boldsymbol{\eta}^*_t)\}\Big) \\ &+ \sum_{t=1}^{T-1}\Big(\mathrm{E}^{(q)}\{\ln f(v_t)\} + \mathrm{E}^{(q)}\{\ln q_t(v_t)\}\Big). \end{aligned} \tag{4.90}$$

In the subsequent analysis, we will evaluate these expectation values in sequential order.

For the first expectation in (4.90) we insert (4.46) for the component distribution $f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})$ and obtain

$$\begin{aligned} \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})\} &= \mathrm{E}^{(q)}\left\{\sum_{t=1}^{T}\mathbb{1}(z_n=t)\Big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}^*_t)\Big)\right\} \\ &= \sum_{t=1}^{T}\mathrm{E}^{(q(z_n,\boldsymbol{\eta}^*_t))}\Big\{\mathbb{1}(z_n=t)\Big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}^*_t)\Big)\Big\} \\ &= \sum_{t=1}^{T}\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\}\mathrm{E}^{(q_t(\boldsymbol{\eta}^*_t))}\Big\{\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}_n) - a(\boldsymbol{\eta}^*_t)\Big\} \\ &= \sum_{t=1}^{T}\phi_{n,t}\Big(\ln h(\boldsymbol{x}_n) + \mathrm{E}^{(q_t(\boldsymbol{\eta}^*_t))}\Big\{\boldsymbol{\eta}_t^{*\mathrm{T}}\Big\}\boldsymbol{t}(\boldsymbol{x}_n) - \mathrm{E}^{(q_t(\boldsymbol{\eta}^*_t))}\{a(\boldsymbol{\eta}^*_t)\}\Big). \end{aligned} \tag{4.91}$$

Here, we made use of the fact that $q(z_n,\boldsymbol{\eta}^*_t) = q_n(z_n)q_t(\boldsymbol{\eta}^*_t)$, which is because of the mean field assumption, and of the previous result (4.59). As mentioned earlier, once the exponential family for the model has been specified, it becomes possible to derive the two expectations $\mathrm{E}^{(q_t(\boldsymbol{\eta}^*_t))}\Big\{\boldsymbol{\eta}_t^{*\mathrm{T}}\Big\}$ and $\mathrm{E}^{(q_t(\boldsymbol{\eta}^*_t))}\{a(\boldsymbol{\eta}^*_t)\}$.

Using (4.56) for $p^{(T)}(z_n|\boldsymbol{v})$, the second expectation in (4.90) is given by

$$\begin{aligned} &\mathrm{E}^{(q)}\{\ln p^{(T)}(z_n|\boldsymbol{v})\} \\ &= \mathrm{E}^{(q)}\left\{\sum_{t=1}^{T-1}(\mathbb{1}(z_n=t)\ln v_t + \mathbb{1}(z_n>t)\ln(1-v_t))\right\} \\ &= \sum_{t=1}^{T-1}\Big(\mathrm{E}^{(q(z_n,v_t))}\{\mathbb{1}(z_n=t)\ln v_t\} + \mathrm{E}^{(q(z_n,v_t))}\{\mathbb{1}(z_n>t)\ln(1-v_t)\}\Big) \\ &= \sum_{t=1}^{T-1}\Big(\mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n=t)\}\mathrm{E}^{(q_t(v_t))}\{\ln v_t\} + \mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n>t)\}\mathrm{E}^{(q_t(v_t))}\{\ln(1-v_t)\}\Big) \\ &= \sum_{t=1}^{T-1}\left(\phi_{n,t}(\Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1}+\gamma_{t,2})) + \left(\sum_{j=t+1}^{T}\phi_{n,j}\right)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1}+\gamma_{t,2}))\right), \end{aligned} \tag{4.92}$$

where we inserted (4.88) for $\mathrm{E}^{(q_t(v_t))}\{\ln v_t\}$ and (4.89) for $\mathrm{E}^{(q_t(v_t))}\{\ln(1-v_t)\}$.

Substituting (4.86) into the third expectation in (4.90) results in

$$\begin{aligned}
\mathrm{E}^{(q)}\{\ln q_n(z_n)\} &= \mathrm{E}^{(q)}\left\{\sum_{t=1}^{T} \mathbb{1}(z_n = t) \ln \phi_{n,t}\right\} \\
&= \sum_{t=1}^{T} \mathrm{E}^{(q_n(z_n))}\{\mathbb{1}(z_n = t)\} \ln \phi_{n,t} \\
&= \sum_{t=1}^{T} \phi_{n,t} \ln \phi_{n,t} \\
&= \boldsymbol{\phi}_n^{\mathrm{T}} \boldsymbol{\phi}_n. \qquad (4.93)
\end{aligned}$$

Inserting the base distribution (4.45), i.e. the prior pdf for $\boldsymbol{\eta}_t^*$, into the fourth expectation in (4.90) yields

$$\begin{aligned}
\mathrm{E}^{(q)}\{\ln f(\boldsymbol{\eta}_t^*)\} &= \mathrm{E}^{(q)}\left\{\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\right\} \\
&= \ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\} - \lambda_2 \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}. \qquad (4.94)
\end{aligned}$$

Furthermore, inserting the corresponding approximate posterior (4.77) of $\boldsymbol{\eta}_t^*$ into the fifth expectation in (4.90) results

$$\begin{aligned}
\mathrm{E}^{(q)}\{\ln q_t(\boldsymbol{\eta}_t^*)\} &= \mathrm{E}^{(q)}\left\{\ln b(\boldsymbol{\tau}_t) + \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\eta}_t^* - \tau_{t,2} a(\boldsymbol{\eta}_t^*)\right\} \\
&= \ln b(\boldsymbol{\tau}_t) + \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\} - \tau_{t,2} \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}. \qquad (4.95)
\end{aligned}$$

The last two expectations in (4.90) consider the beta prior $f(v_t) = \mathcal{B}(v_t; 1, \alpha)$ and the approximate beta posterior $q_t(v_t) = \mathcal{B}(v_t; \gamma_{t,1}, \gamma_{t,2})$. Inserting (4.65) into the sixth expectation term yields

$$\begin{aligned}
\mathrm{E}^{(q)}\{\ln f(v_t)\} &= \mathrm{E}^{(q)}\left\{\ln \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} + (\alpha - 1)\ln(1 - v_t)\right\} \\
&= \ln \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} + (\alpha - 1)\mathrm{E}^{(q_t(v_t))}\{\ln(1 - v_t)\} \\
&= \ln \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} + (\alpha - 1)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1} + \gamma_{t,2})) \qquad (4.96)
\end{aligned}$$

and inserting (4.67) into the last expectation gives

$$\begin{aligned}
&\mathrm{E}^{(q)}\{\ln q_t(v_t)\} \\
&= \mathrm{E}^{(q)}\left\{\ln \frac{\Gamma(\gamma_{t,1} + \gamma_{t,2})}{\Gamma(\gamma_{t,1})\Gamma(\gamma_{t,2})} + (\gamma_{t,1} - 1)\ln v_t + (\gamma_{t,2} - 1)\ln(1 - v_t)\right\} \\
&= \ln \frac{\Gamma(\gamma_{t,1} + \gamma_{t,2})}{\Gamma(\gamma_{t,1})\Gamma(\gamma_{t,2})} + (\gamma_{t,1} - 1)\mathrm{E}^{(q_t(v_t))}\{\ln v_t\} + (\gamma_{t,2} - 1)\mathrm{E}^{(q_t(v_t))}\{\ln(1 - v_t)\} \\
&= \ln \frac{\Gamma(\gamma_{t,1} + \gamma_{t,2})}{\Gamma(\gamma_{t,1})\Gamma(\gamma_{t,2})} + (\gamma_{t,1} - 1)(\Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1} + \gamma_{t,2})) + (\gamma_{t,2} - 1)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1} + \gamma_{t,2})). \\
&\qquad (4.97)
\end{aligned}$$

Here, we again employed (4.88) for $\mathrm{E}^{(q_t(v_t))}\{\ln v_t\}$ and (4.89) for $\mathrm{E}^{(q_t(v_t))}\{\ln(1-v_t)\}$.

We conclude that the ELBO for the truncated exponential family DPM model with joint pdf $f^{(T)}(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z}, \boldsymbol{x})$ (see (4.49)) and variational distribution $q(\boldsymbol{v}, \boldsymbol{\eta}*, \boldsymbol{z})$ (see (4.43), (4.67), (4.77) and (4.86)) is given by (4.90), where the respective expectations are given by (4.91), (4.92), (4.93), (4.94), (4.95), (4.96) and (4.97). Note that we do not have to compute any integrals to evaluate the ELBO. We arrive at a fully closed form solution that depends on the hyperparameters α and $\boldsymbol{\lambda}$, the observation $\boldsymbol{x}$, and the variational parameters $\boldsymbol{\gamma}_{1:(T-1)}$, $\boldsymbol{\tau}_{1:T}$ and $\boldsymbol{\phi}_{1:N}$.

4.4.4 Summary

Finally, we summarize all equations that are necessary to implement CAVI for the DPM using a truncated stick-breaking approximation. Algorithm 2 depicts the overall procedure to compute the solution of the CAVI optimization problem (4.26). The input consists of the observations $\boldsymbol{x}_n$, $n = 1, \ldots, N$, the truncation level T, the hyperparameter α of the stick-breaking process, and the hyperparameter $\boldsymbol{\lambda}$ of the base distribution. We track the number of iterations with the variable ℓ and start with $\ell = 0$, which corresponds to the initialization of the variational parameters, i.e., $\boldsymbol{\phi}_{1:N}^{(0)}$, $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$, and $\boldsymbol{\tau}_{1:T}^{(0)}$.

As we have shown in the above derivation, the updates for the variational parameters of the variational factor distributions are given by the following list of equations:

- Variational parameters $\boldsymbol{\phi}_n^{(\ell)} = \begin{pmatrix} \phi_{n,1}^{(\ell)} & \cdots & \phi_{n,T}^{(\ell)} \end{pmatrix}$ of $q_n^{(\ell)}(z_n; \boldsymbol{x}) = \mathcal{C}\left(z_n; \boldsymbol{\phi}_n^{(\ell)}\right)$:

$$\phi_{n,t}^{(\ell)} = \frac{\exp\left(S_{n,t}^{(\ell)}\right)}{\sum_{i=1}^{T} \exp\left(S_{n,t}^{(\ell)}\right)}, \tag{4.98a}$$

  where

$$\begin{aligned} S_{n,t}^{(\ell)} = {} & \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_t^{*\mathrm{T}}\right\}\boldsymbol{t}(\boldsymbol{x}_n) - \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\} \\ & + \Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1} + \gamma_{t,2}) + \sum_{j=1}^{t-1} \Psi(\gamma_{j,2}) - \Psi(\gamma_{j,1} + \gamma_{j,2}). \end{aligned} \tag{4.98b}$$

- Variational parameters $\boldsymbol{\gamma}_t^{(\ell)} = \begin{pmatrix} \gamma_{t,1}^{(\ell)} & \gamma_{t,2}^{(\ell)} \end{pmatrix}^{\mathrm{T}}$ of $q_t^{(\ell)}(v_t; \boldsymbol{x}) = \mathcal{B}\left(v_t; \gamma_{t,1}^{(\ell)}, \gamma_{t,2}^{(\ell)}\right)$:

$$\gamma_{t,1}^{(\ell)} = 1 + \sum_{n=1}^{N} \phi_{n,t}, \tag{4.99a}$$

$$\gamma_{t,2}^{(\ell)} = \alpha + \sum_{n=1}^{N} \sum_{j=t+1}^{T} \phi_{n,j}. \tag{4.99b}$$

- Variational parameters $\boldsymbol{\tau}_t^{(\ell)} = \begin{pmatrix} \boldsymbol{\tau}_{t,1}^{(\ell)\mathrm{T}} & \tau_{t,2}^{(\ell)} \end{pmatrix}^{\mathrm{T}}$ of $q_t^{(\ell)}(\boldsymbol{\eta}_t^*; \boldsymbol{x}) \propto \exp\left(\boldsymbol{\tau}_{t,1}^{(\ell)\mathrm{T}} \boldsymbol{\eta}_t^* - \tau_{t,2}^{(\ell)} a(\boldsymbol{\eta}_t^*)\right)$:

$$\boldsymbol{\tau}_{t,1}^{(\ell)} = \boldsymbol{\lambda}_1 + \sum_{n=1}^{N} \phi_{n,t} \boldsymbol{t}(\boldsymbol{x}_n), \tag{4.100a}$$

$$\tau_{t,2}^{(\ell)} = \lambda_2 + \sum_{n=1}^{N} \phi_{n,t}. \tag{4.100b}$$

In each iteration ℓ we update the local variational parameters $\boldsymbol{\phi}_{1:N}^{(\ell)}$ using the most recent update of the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}$ and $\boldsymbol{\tau}_{1:T}$, and the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}^{(\ell)}$ and $\boldsymbol{\tau}_{1:T}^{(\ell)}$ using the most recent update of the local variational parameters $\boldsymbol{\phi}_{1:N}$. Note that $\boldsymbol{\tau}_t$ appears in (4.98b) through $q_t(\boldsymbol{\eta}_t^*)$.

At the end of each iteration, the ELBO $\mathcal{L}(q^{(\ell)}; \boldsymbol{x})$ is evaluated using results of the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}^{((\ell))}$ and $\boldsymbol{\tau}_{1:T}^{((\ell))}$ and the local variational parameters $\boldsymbol{\phi}_{1:N}^{((\ell))}$ of the current iteration, and the relative change is monitored to assess convergence. Convergence is declared once the relative change of the ELBO has fallen below some predefined threshold. The ELBO is given by

$$\begin{aligned} L(q; \boldsymbol{x}) = &\sum_{n=1}^{N} \Big(\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n | \boldsymbol{\eta}_{z_n}^*)\} + \mathrm{E}^{(q)}\{\ln p^{(T)}(z_n | \boldsymbol{v})\} + \mathrm{E}^{(q)}\{\ln q_n(z_n)\} \Big) \\ &+ \sum_{t=1}^{T} \Big(\mathrm{E}^{(q)}\{\ln f(\boldsymbol{\eta}_t^*)\} + \mathrm{E}^{(q)}\{\ln q_t(\boldsymbol{\eta}_t^*)\} \Big) \\ &+ \sum_{t=1}^{T-1} \Big(\mathrm{E}^{(q)}\{\ln f(v_t)\} + \mathrm{E}^{(q)}\{\ln q_t(v_t)\} \Big), \end{aligned} \tag{4.101a}$$

with

$$\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n | \boldsymbol{\eta}_{z_n}^*)\} = \sum_{t=1}^{T} \phi_{n,t} \Big(\ln h(\boldsymbol{x}_n) + \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))} \Big\{ \boldsymbol{\eta}_t^{*\mathrm{T}} \Big\} \boldsymbol{t}(\boldsymbol{x}_n) - \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\} \Big), \tag{4.101b}$$

$$\begin{aligned} &\mathrm{E}^{(q)}\{\ln p^{(T)}(z_n | \boldsymbol{v})\} \\ &\quad = \sum_{t=1}^{T-1} \left(\phi_{n,t} (\Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1} + \gamma_{t,2})) + \left(\sum_{j=t+1}^{T} \phi_{n,j} \right) (\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1} + \gamma_{t,2})) \right), \end{aligned} \tag{4.101c}$$

$$\mathrm{E}^{(q)}\{\ln q_n(z_n)\} = \boldsymbol{\phi}_n^{\mathrm{T}} \boldsymbol{\phi}_n, \tag{4.101d}$$

$$\mathrm{E}^{(q)}\{\ln f(\boldsymbol{\eta}_t^*)\} = \ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\} - \lambda_2 \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}, \tag{4.101e}$$

$$\mathrm{E}^{(q)}\{\ln q_t(\boldsymbol{\eta}_t^*)\} = \ln b(\boldsymbol{\tau}_t) + \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\} - \tau_{t,2} \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}, \tag{4.101f}$$

$$\mathrm{E}^{(q)}\{\ln f(v_t)\} = \ln \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} + (\alpha - 1)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1} + \gamma_{t,2})), \tag{4.101g}$$

$$\begin{aligned} &\mathrm{E}^{(q)}\{\ln q_t(v_t)\} \\ &\quad = \ln \frac{\Gamma(\gamma_{t,1} + \gamma_{t,2})}{\Gamma(\gamma_{t,1})\Gamma(\gamma_{t,2})} + (\gamma_{t,1} - 1)(\Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1} + \gamma_{t,2})) + (\gamma_{t,2} - 1)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1} + \gamma_{t,2})). \end{aligned} \tag{4.101h}$$

Once the algorithm has converged, we obtain the variational approximation $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{x})$ of the posterior pdf $f(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z} | \boldsymbol{x})$ using the latest state of the variational parameters at convergence. Note that the final solutions for the update equations (4.98)-(4.100) and the ELBO (4.101) do not require performing numerical integration, which enables the algorithm to remain computationally efficient with increasing sample size N and truncation level T.

Algorithm 2: CAVI for Exponential Family DPM Models with Conjugate Prior

Input: Observations $\boldsymbol{x}$, truncation level T, hyperparameters α and $\boldsymbol{\lambda}$

Output: Variational factor pdfs

- $q_n^*(z_n; \boldsymbol{x})$ for $n = 1, \ldots, N$
- $q_t^*(v_t; \boldsymbol{x})$ for $t = 1, \ldots, T-1$
- $q_t^*(\boldsymbol{\eta}_t^*; \boldsymbol{x})$ for $t = 1, \ldots, T$

Initialize: Variational parameters $\boldsymbol{\phi}_{1:N}^{(0)}$, $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$, and $\boldsymbol{\tau}_{1:T}^{(0)}$;

while *the ELBO has not converged* **do**

  $\ell = \ell + 1$ **for** n *from* 1 *to* N **do**

    for t *from* 1 *to* T **do**

      set $S_{n,t}^{(\ell)}$ according to (4.98b)

    for t *from* 1 *to* T **do**

      set $\phi_{n,t}^{(\ell)}$ according to (4.98a)

  for t *from* 1 *to* $T-1$ **do**

    set $\boldsymbol{\gamma}_t^{(\ell)}$ according to (4.99)

  for t *from* 1 *to* T **do**

    set $\boldsymbol{\tau}_t^{(\ell)}$ according to (4.100)

  Compute the ELBO $\mathcal{L}(q^{(\ell)}; \boldsymbol{x})$ according to (4.101)

return $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{x}) = \left(\prod_{t=1}^{T-1} q_t^*(v_t; \boldsymbol{x})\right)\left(\prod_{t=1}^{T} q_t^*(\boldsymbol{\eta}_t^*; \boldsymbol{x})\right)\left(\prod_{n=1}^{N} q_n^*(z_n; \boldsymbol{x})\right)$

4.5 Practicalities

In concluding this chapter, we will discuss practical considerations that arise when implementing CAVI-based methods for the DPM. Specifically, we will address the following relevant aspects: initializing the algorithm, reordering of component labels to enhance accuracy, and exploring different methods for assessing the convergence of the algorithm.

4.5.1 Initialization

The initialization of the CAVI algorithm involves setting initial values for the variational parameters $\boldsymbol{\phi}_{1:N}^{(0)}$, $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$, and $\boldsymbol{\tau}_{1:T}^{(0)}$ (cf. Algorithm 2), and has an impact on the quality of the approximation of the posterior distribution. Poor choices of initialization may lead to slow convergence or getting stuck in a poor local maxima of the ELBO. Therefore, experimentation with the initialization procedure is often required to obtain good results.

There are three potential approaches for the sequence of the initialization of the variational parameters. Firstly, we can begin by initializing the local variational parameters $\boldsymbol{\phi}_{1:N}^{(0)}$ and subsequently use them to set the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$ (according to (4.99) and (4.100)), or, secondly, we can follow the opposite sequence (using (4.98)). Thirdly, we can initialize the variational parameters independently of one another.

Recalling that $\boldsymbol{\phi}_{1:N}^{(0)}$ represent (soft) cluster assignments, clustering methods [50] such as K-

means, expectation-maximization, and DBSCAN can be employed to initialize these values. If the number of clusters has to be provided as an input to the clustering algorithm, one can use either the truncation level T or use a value $T' < T$ and set $\phi_{n,t}$ with $t > T'$ to 0. On the other hand, if the number of clusters is determined by the clustering algorithm, the truncation level T can be set according to the output of the algorithm. Possibilities that do not use clustering for initialization include choosing T and randomly generate soft or hard cluster assignments or use uniformly distributed cluster assignments (i.e., $\phi_{n,t} = 1/T$ for $t = 1, \dots, T$). Note that a hard cluster assignment means that $\boldsymbol{\phi}_n$ is a vector of all zeros except for a single element $\phi_{n,t}$ set to one. In contrast, a soft cluster assignment can have $\phi_{n,t} \in [0, 1]$ for $t = 1, \dots, T$ with $\sum_{t=1}^{T} \phi_{n,t} = 1$.

An additional option that can be explored involves setting $T = N$ and uniquely assign each observation to its own cluster. The advantage of this approach is that the truncation level T is high enough to capture any number of possible clusters in the observations. However, this comes at the cost of sacrificing computational efficiency because there are more terms in the update equations (4.98), (4.99), and (4.100), and in the ELBO (4.101), which have to be computed.

Similarly to the random initialization of the local variational parameters $\boldsymbol{\phi}_{1:N}^{(0)}$, the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$ can also bet set randomly within a specified range. Furthermore, if we exclude the variational parameters $\boldsymbol{\phi}_{1:N}^{(0)}$ in equation (4.99) and (4.100), it is possible to only utilize the prior knowledge, provided by the hyperparameters α and $\boldsymbol{\lambda}$, to initialize $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$. Alternatively, simple parameter estimators, like sample moments, can be employed to compute initial values for the global variational parameters using the available observations.

Lastly, note that the variational parameters also depend on the truncation level T and the hyperparameters α and $\boldsymbol{\lambda}$ (see (4.98), (4.99) and (4.100)). Hence, the choice of these input values also influences the overall quality of the posterior approximation. The truncation level T can be treated as a variational parameter and optimized by executing the algorithm multiple times for different values of T. We can then select the value for T that yields the highest ELBO. Recall that the ELBO is an approximation of the evidence $f(\boldsymbol{x})$, which provides a basis for selecting a model [6]. Selecting the number of components T of the mixture model based on the ELBO is therefore referred to as model selection. A discussion about model selection in mixture models using the ELBO can be found in [51]. Choosing a value for T may also involve the stick-breaking parameter α. Specifically, we can set T equal to the average number of clusters $\bar{L} = \alpha(\Psi(\alpha + N) - \Psi(\alpha))$ that are expected to be present within the observation $\boldsymbol{x}$ (see (3.19)). Since the maximum possible value of L is given by number of observations N, a starting point for the truncation level T can be chosen such that $\bar{L} < T < N$.

4.5.2 Component Label Reordering

According to [17], the approximation accuracy of CAVI for models formulated in the stick-breaking representation can be improved by ordering the mixture component labels $t = 1, \dots, T$ during the CAVI algorithm. As discussed in Section 3.2.1 the stick-breaking process implies that the average mixing proportions are ordered in decreasing size with respect to the component labels t (see (3.12)). The optimal relabelling of the components t is thus given by the one

that orders the posterior mean of the mixing proportions in decreasing order. The approximate posterior distribution $q_t^{(\ell)}(v_t; \boldsymbol{x})$ of the t-th auxiliary variable v_t is given by $\mathcal{B}\left(v_t; \gamma_{t,1}^{(\ell)}, \gamma_{t,2}^{(\ell)}\right)$, and thus, similar to (3.25), the posterior mean of mixing proportions π_t is given by

$$\mathrm{E}^{(q^{(\ell)}(\boldsymbol{v};\boldsymbol{x}))}\{\pi_t(\boldsymbol{v}_{1:t})\} = \mathrm{E}^{(q_t^{(\ell)}(v_t;\boldsymbol{x}))}\{v_t\} \prod_{j=1}^{t-1} \left(1 - \mathrm{E}^{(q_t^{(\ell)}(v_t;\boldsymbol{x}))}\{v_t\}\right) \tag{4.102}$$

$$= \frac{\gamma_{t,1}^{(\ell)}}{\gamma_{t,1}^{(\ell)} + \gamma_{t,2}^{(\ell)}} \prod_{j=1}^{t-1} \left(\frac{\gamma_{j,2}^{(\ell)}}{\gamma_{j,1}^{(\ell)} + \gamma_{j,2}^{(\ell)}}\right), \tag{4.103}$$

for $t = 1, \ldots, T-1$ and

$$\mathrm{E}^{(q^{(\ell)}(\boldsymbol{v};\boldsymbol{x}))}\{\pi_T(\boldsymbol{v})\} = \prod_{j=1}^{T-1} \frac{\gamma_{j,2}^{(\ell)}}{\gamma_{j,1}^{(\ell)} + \gamma_{j,2}^{(\ell)}}, \tag{4.104}$$

since $v_T = 1$ due to the truncation. Evaluating (4.103) and (4.104) at the end of each iteration ℓ with the current values of the variational parameters $\gamma_{t,1}^{(\ell)}$ and $\gamma_{t,2}^{(\ell)}$, $t = 1, \ldots, T-1$, the optimal reordering of the component labels t is achieved by

$$\mathrm{E}^{(q^{(\ell)}(\boldsymbol{v};\boldsymbol{x}))}\left\{\pi_{\sigma(1)}(v_1)\right\} > \mathrm{E}^{(q^{(\ell)}(\boldsymbol{v};\boldsymbol{x}))}\left\{\pi_{\sigma(2)}(v_1, v_2)\right\} > \cdots > \mathrm{E}^{(q^{(\ell)}(\boldsymbol{v};\boldsymbol{x}))}\left\{\pi_{\sigma(T)}(\boldsymbol{v})\right\}.$$

Here, we made use of a permutation function $\sigma(\cdot)$ which makes sure that the posterior mean of the mixing proportions are ordered in decreasing size with respect to the relabeled component labels $\sigma(1), \ldots, \sigma(T)$.

We can also reorder the component labels according to the posterior mean of the auxiliary variable v_t since the posterior mean of π_t is proportional to the posterior mean of the t-th auxiliary variable v_t (see (4.102)). Thus, we equivalently have

$$\mathrm{E}^{(q_t^{(\ell)}(v_t;\boldsymbol{x}))}\left\{v_{\sigma(1)}\right\} > \mathrm{E}^{(q_t^{(\ell)}(v_t;\boldsymbol{x}))}\left\{v_{\sigma(2)}\right\} > \cdots > \mathrm{E}^{(q_t^{(\ell)}(v_t;\boldsymbol{x}))}\left\{v_{\sigma(T-1)}\right\},$$

which is easier to evaluate in terms of computational complexity compared to using the expectations (4.102) of the mixture proportions for the relabelling.

4.5.3 Convergence

Assessing convergence in the CAVI algorithm is essential to ensure that the algorithm has reached a local optimum. The ELBO indicates how well the CAVI algorithm is approximating the posterior, because maximizing the ELBO is equivalent to minimizing the KLD between the approximate posterior and the true posterior (see (4.16)–(4.20)). As the ELBO should increase with each iteration ℓ, monitoring the convergence of the ELBO indicates how well the algorithm is progressing. Convergence is declared when the relative change of the ELBO falls below a predefined threshold, indicating that the algorithm has reached a local optimum. In the simulations in Chapter 5 we will use the percentage change of the ELBO to assess convergence,

i.e.,

$$\frac{\mathcal{L}(q^{(\ell)};\boldsymbol{x}) - \mathcal{L}(q^{(\ell-1)};\boldsymbol{x})}{\mathcal{L}(q^{(\ell-1)};\boldsymbol{x})} \times 100 < \epsilon, \tag{4.105}$$

where ϵ denotes the predefined convergence threshold.

Alternatively, assessing convergence can be done by tracking the changes in the variational parameters at each iteration. When the variational parameters stabilize or show very small changes between iterations, it suggests that the algorithm has converged. Indeed, convergence of the variational parameters is sufficient to guarantee the convergence of the ELBO, because the change of the ELBO results from the change of the variational parameters (see (4.101)). An advantage of the latter method is, that we do not have to evaluate the computationally burdensome ELBO. The disadvantage is that instead of monitoring a single scalar value we have to monitor the change of all variational parameters. In the case of the truncated stick-breaking DPM model this amounts to

$$\dim(\boldsymbol{\phi}_{1:N}) + \dim(\boldsymbol{\gamma}_{1:(T-1)}) + \dim(\boldsymbol{\tau}_{1:T}) = NT + (T-1)2 + T(p+1) \tag{4.106}$$

scalar values. The set of values that need to be monitored can be reduced by tracking specified metrics of subgroups of variational parameters, e.g., by monitoring the relative change of the quadratic norm of $\boldsymbol{\phi}_{1:N}$, $\boldsymbol{\gamma}_{1:(T-1)}$, and $\boldsymbol{\tau}_{1:T}$.

As mentioned above, computing the ELBO for the entire observation $\boldsymbol{x}$ may be computationally intensive. Nonetheless, it is advantageous to monitor the relative change of a single scalar value that quantifies the quality of the approximation. Another way to do this with less computational burden is by evaluating the approximate predictive performance of the mixture model, which should improve with each iteration. Using a small test dataset $\tilde{\boldsymbol{x}} \notin \boldsymbol{x}$ of dimension $\dim(\tilde{\boldsymbol{x}}) = MJ$, we can track the relative change of the average log predictive as follows. The posterior predictive distribution [1] of a new value $\tilde{\boldsymbol{x}}_j$ under the model (4.44) is given by

$$f(\tilde{\boldsymbol{x}}_j|\boldsymbol{x}) = \int_{\mathbb{R}^{Tp}} \int_{\mathbb{R}^{(T-1)}} f(\tilde{\boldsymbol{x}}_j|\boldsymbol{v},\boldsymbol{\eta}^*) f(\boldsymbol{v},\boldsymbol{\eta}^*|\boldsymbol{x})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{\eta}^* = \mathrm{E}^{(f(\boldsymbol{v},\boldsymbol{\eta}^*|\boldsymbol{x}))}\{f(\tilde{\boldsymbol{x}}_j|\boldsymbol{v},\boldsymbol{\eta}^*)\}.$$

Replacing the conditional distribution $f(\tilde{\boldsymbol{x}}_j|\boldsymbol{v},\boldsymbol{\eta}^*)$ with the mixture distribution

$$f^{(T)}(\tilde{\boldsymbol{x}}_j|\boldsymbol{v},\boldsymbol{\eta}^*) = \sum_{t=1}^{T} \pi_t(\boldsymbol{v}_{1:t}) f(\tilde{\boldsymbol{x}}_j|\boldsymbol{\eta}_t^*)$$

of the truncated model (4.44), and using the approximate posterior pdf

$$q(\boldsymbol{v},\boldsymbol{\eta}^*) = q(\boldsymbol{v})q(\boldsymbol{\eta}^*) = \left(\prod_{t=1}^{T-1} q_t(v_t)\right)\left(\prod_{t=1}^{T} q_t(\boldsymbol{\eta}_t^*)\right)$$

instead of the true posterior pdf $f(\boldsymbol{v},\boldsymbol{\eta}^*|\boldsymbol{x})$, yields an approximation of the posterior predictive,

i.e.,

$$\begin{aligned} f(\tilde{\boldsymbol{x}}_j|\boldsymbol{x}) &\approx \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*))}\left\{\sum_{t=1}^{T}\pi_t(\boldsymbol{v}_{1:t})f(\tilde{\boldsymbol{x}}_j|\boldsymbol{\eta}_t^*)\right\} \\ &= \sum_{t=1}^{T}\mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*))}\{\pi_t(\boldsymbol{v}_{1:t})f(\tilde{\boldsymbol{x}}_j|\boldsymbol{\eta}_t^*)\} \\ &= \sum_{t=1}^{T}\mathrm{E}^{(q(\boldsymbol{v}))}\{\pi_t(\boldsymbol{v}_{1:t})\}\mathrm{E}^{(q(\boldsymbol{\eta}^*))}\{f(\tilde{\boldsymbol{x}}_j|\boldsymbol{\eta}_t^*)\} \\ &= \sum_{t=1}^{T}\mathrm{E}^{(q(v_t))}\{\pi_t(\boldsymbol{v}_{1:t})\}\mathrm{E}^{(q(\boldsymbol{\eta}_t^*))}\{f(\tilde{\boldsymbol{x}}_j|\boldsymbol{\eta}_t^*)\}. \end{aligned}$$

Taking the logarithm and averaging over the test dataset $\tilde{\boldsymbol{x}}_j$, $j = 1, \ldots, J$, results in the average log predictive

$$\frac{1}{J}\sum_{j=1}^{J}\ln f(\tilde{\boldsymbol{x}}_j|\boldsymbol{x}) \approx \frac{1}{J}\sum_{j=1}^{J}\ln\sum_{t=1}^{T}\mathrm{E}^{(q(v_t))}\{\pi_t(\boldsymbol{v}_{1:t})\}\mathrm{E}^{(q(\boldsymbol{\eta}_t^*))}\{f(\tilde{\boldsymbol{x}}_j|\boldsymbol{\eta}_t^*)\}, \tag{4.107}$$

where $\mathrm{E}^{(q(v_t))}\{\pi_t(\boldsymbol{v}_{1:t})\}$ is given by (4.103) and $\mathrm{E}^{(q(\boldsymbol{\eta}_t^*))}\{f(\tilde{\boldsymbol{x}}_j|\boldsymbol{\eta}_t^*)\}$ depends on the choice of the EF. Equation (4.107) can be evaluated at the end of each iteration ℓ using the factors $q_t^{(\ell)}(v_t;\boldsymbol{x})$ and $q_t^{(\ell)}(\boldsymbol{\eta}_t^*;\boldsymbol{x})$, which itself depend on the variational parameters $\boldsymbol{\gamma}_t^{(\ell)}$ and $\boldsymbol{\tau}_t^{(\ell)}$. Note that, unlike the full ELBO, the average log predictive is not guaranteed to monotonically increase across iterations of the CAVI algorithm.

Finally, note that the initial values for the variational parameters, the choice of the hyperparameters and the convergence threshold influence how fast the CAVI algorithm converges. If the algorithm converges too quickly, it is most likely because the threshold is set too large and needs to be reduced. Moreover, it can converge to a local optimum that does not correspond to a good approximation of the posterior. To address this concern in practice, we can run the algorithm multiple times with a different initialization to check for consistency in the results and the quality of the approximation via the ELBO (the higher the ELBO the better the approximation).

5 Gaussian Estimation

In this chapter, we use the coordinate ascent variational inference (CAVI) algorithm for a Gaussian estimation problem. We will adapt the CAVI algorithm for exponential family Dirichlet process mixture (DPM) models (see Algorithm 2) to Gaussian DPM models, where the mean of the mixture components is assumed to be random. Furthermore, we will present simulation results obtained by using the CAVI algorithm and compare them to results from [8], which were obtained by using a Markov chain Monte Carlo (MCMC) method known as the Gibbs sampler.

5.1 Model Definition

Based on [8], we consider a Gaussian model for objects that are indexed by $n = 1, \ldots, N$, where N is the total number of objects. Each object is described by a random feature vector $\boldsymbol{x}_n = (x_{n,1} \ \cdots \ x_{n,M})^{\mathrm{T}} \in \mathbb{R}^M$, which depends on a random local parameter vector $\boldsymbol{\theta}_n = (\theta_{n,1} \ \cdots \ \theta_{n,M})^{\mathrm{T}} \in \mathbb{R}^M$ through the equation

$$\boldsymbol{x}_n = \boldsymbol{\theta}_n + \boldsymbol{u}_n, \quad n = 1, \ldots, N. \tag{5.1}$$

The vectors $\boldsymbol{u}_n \in \mathbb{R}^M$, $n = 1, \ldots, N$, in (5.1) will be called the parameter noise and are assumed to be i.i.d. according to a zero-mean Gaussian distribution with known covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}}$, i.e.,

$$f(\boldsymbol{u}_n) = \mathcal{N}(\boldsymbol{u}_n; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{5.2}$$

From (5.1) and (5.2) it follows that the conditional pdf $f(\boldsymbol{x}_n|\boldsymbol{\theta}_n)$ is given by

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\theta}_n, \boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{5.3}$$

Thus, conditioned on $\boldsymbol{\theta}_n$, the object features $\boldsymbol{x}_n$ are Gaussian distributed with mean $\boldsymbol{\mu}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Note that the mean $\boldsymbol{\mu}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n}$ of (5.3) is assumed to be random while the covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n}$ is assumed to be known.

We assume the local parameters $\boldsymbol{\theta}_n$ to be random and distributed according to a Dirichlet process (DP) as introduced in Section 3.2. In other words, given a realization $G(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:\infty})$ (see (3.8)) of $\mathcal{DP}(G; G_0, \alpha)$, we consider N local parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ i.i.d. and individually distributed according to this realization (see (3.14)), i.e.,

$$G \sim \mathcal{DP}(G; G_0, \alpha) \tag{5.4}$$

$$\boldsymbol{\theta}_n|G \overset{\text{i.i.d.}}{\sim} G(\boldsymbol{\theta}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*_{1:\infty}), \quad n = 1, \ldots, N. \tag{5.5}$$

Recall that $\alpha > 0$ represents the concentration parameter and $G_0(\boldsymbol{\theta}_k^*)$ the base distribution of the DP. We assume a Gaussian base distribution

$$G_0(\boldsymbol{\theta}_k^*) = \mathcal{N}(\boldsymbol{\theta}_k^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}), \tag{5.6}$$

with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$. The random vectors $\boldsymbol{\theta}_k^*$, $k = 1, 2, \ldots$, represent global parameters of the model which are i.i.d. according to (5.6) as explained in the stick-breaking construction of G (see (3.13)). Each local parameter $\boldsymbol{\theta}_n$ is equal to one of the global parameters, i.e., $\boldsymbol{\theta}_n \in \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots\}$. Note that (5.6) is a conjugate prior for likelihood function (5.3), because the conjugate prior for a Gaussian likelihood function with random mean is given by a Gaussian distribution [42].

Rather than directly observing the object features $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, we instead observe an altered version of the object features corrupted by additive noise. We denote the observations by $\boldsymbol{y}_n = (y_{n,1} \ \ldots \ y_{n,M})^{\mathrm{T}} \in \mathbb{R}^M$, $n = 1, \ldots, N$. The observation model is assumed to be given by

$$\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{v}_n, \quad n = 1, \ldots, N, \tag{5.7}$$

where $\boldsymbol{v}_n \in \mathbb{R}^M$ is i.i.d. across n according to the zero-mean Gaussian distribution

$$f(\boldsymbol{v}_n) = \mathcal{N}(\boldsymbol{v}_n; \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \tag{5.8}$$

with known covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}}$. Because of (5.8) and (5.7), the distribution of $\boldsymbol{y}_n$ given $\boldsymbol{x}_n$ is given by

$$f(\boldsymbol{y}_n|\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}), \tag{5.9}$$

i.e., the observations $\boldsymbol{y}_n$ are Gaussian distributed with unknown mean $\boldsymbol{\mu}_{\boldsymbol{y}_n|\boldsymbol{x}_n} = \boldsymbol{x}_n$ and known covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{x}_n} = \boldsymbol{\Sigma}_{\boldsymbol{v}}$. Moreover, inserting (5.1) into (5.7) yields

$$\boldsymbol{y}_n = \boldsymbol{\theta}_n + \boldsymbol{u}_n + \boldsymbol{v}_n, \quad n = 1, \ldots, N, \tag{5.10}$$

which entails that $\boldsymbol{y}_n$ given $\boldsymbol{\theta}_n$ is, similar to (5.3), given by a Gaussian distribution

$$f(\boldsymbol{y}_n|\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\theta}_n, \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}), \tag{5.11}$$

with unknown mean $\boldsymbol{\mu}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n$ but with known covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$.

Note that we assume the following independence assumptions for all $n, n' \in \{1, \ldots, N\}$:

$$\boldsymbol{u}_n \perp\!\!\!\perp \boldsymbol{u}_{n'}, \quad \boldsymbol{v}_n \perp\!\!\!\perp \boldsymbol{v}_{n'}, \quad \text{where } n \neq n', \tag{5.12a}$$

and

$$\boldsymbol{\theta}_n \perp\!\!\!\perp \boldsymbol{u}_{n'}, \quad \boldsymbol{\theta}_n \perp\!\!\!\perp \boldsymbol{v}_{n'}, \quad \boldsymbol{u}_n \perp\!\!\!\perp \boldsymbol{v}_{n'}. \tag{5.12b}$$

In contrast to $\boldsymbol{u}_n$ and $\boldsymbol{v}_n$ being independent across object index n, the local parameters $\boldsymbol{\theta}_n$ are

not independent across n because of the DP prior (cf. (3.16)). However, the vectors $\boldsymbol{\theta}_n$ are conditionally independent, i.e., given a realization G of the DP as defined in (5.5). Thus, we have

$$\boldsymbol{\theta}_n \perp\!\!\!\perp \boldsymbol{\theta}_{n'} \mid G, \quad n \neq n', \tag{5.13}$$

for all $n, n' \in \{1, \ldots, N\}$.

According to (5.3) and (5.11) the conditional distributions of the feature vector $\boldsymbol{x}_n$ and the corresponding noisy observation $\boldsymbol{y}_n$ are both parameterized by the local parameter $\boldsymbol{\theta}_n$, which is generated according to a DP. Relating the conditional distributions (5.3) and (5.11) and the DP-distributed local parameters $\boldsymbol{\theta}_n$ with the model definition (3.37), yields that the statistical behavior of $\boldsymbol{x}_n$ and the statistical behavior of $\boldsymbol{y}_n$ are both characterized by the DPM model. Since the parameters α and G_0 of the DP are identical in both cases, the DPM governing $\boldsymbol{x}_n$ and the DPM governing $\boldsymbol{y}_n$ only differ in terms of their component distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*)$ and $f(\boldsymbol{y}_n|\boldsymbol{\theta}_k^*)$. In the case of $\boldsymbol{x}_n$ the mixture distribution (see (3.39)) is given by

$$f(\boldsymbol{x}_n|G) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*, \boldsymbol{\Sigma}_{\boldsymbol{u}}), \tag{5.14}$$

where the component distribution of the mixture corresponds to $f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*, \boldsymbol{\Sigma}_{\boldsymbol{u}})$ (cf. (5.3)). In the case of $\boldsymbol{y}_n$ it is given by

$$f(\boldsymbol{y}_n|G) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\theta}_k^*, \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}). \tag{5.15}$$

Here, the component distribution of the mixture corresponds to $f(\boldsymbol{y}_n|\boldsymbol{\theta}_k^*) = \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\theta}_k^*, \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})$ (cf. (5.11)).

We conclude that the statistical model for the observation $\boldsymbol{y}_n$ and the statistical model for the object feature $\boldsymbol{x}_n$ are equal to an infinite Gaussian mixture model where the mean of each Gaussian component k is equal to the respective global parameter $\boldsymbol{\theta}_k^*$. The local parameters $\boldsymbol{\theta}_n$ take $L < N$ distinct values $\boldsymbol{\theta}_l' \in \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots\}$, $l = 1, \ldots, L$, and associate each object n with one of the global parameters $\boldsymbol{\theta}_k^*$. All objects n that share the same value $\boldsymbol{\theta}_l'$ build a cluster. In this context the values $\boldsymbol{\theta}_l'$ can be referred to as cluster parameters, where the number of clusters is given by L. Using indicator variables $z_n \in \{1, \ldots, L\}$, we can mathematically formulate the association of $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_l'$ as $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{z_n}'$. For a discussion of the clustering property of the DP we refer to Section 3.2.2 and Section 3.2.3.

In Figure 5.1a we show the Bayesian network of the overall model. Each object n is associated with a set comprising the random variables $\boldsymbol{\theta}_n$, $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$. The hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ parameterize the base distribution $G_0(\boldsymbol{\theta}_k^*)$, which is together with the concentration parameter α responsible for generating the local parameter $\boldsymbol{\theta}_n$ through the DP. The hyperparameter $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ determines the statistical properties of the additive parameter noise (see (5.2)) and the hyperparameter $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ determines the statistical properties of the additive observation noise (see (5.8)). Note that the set of hyperparameters $\{\alpha, \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}\}$ is assumed to be deterministic and the observations $\boldsymbol{y}_n$, $n = 1, \ldots, N$, are assumed to be known. Figure 5.1b

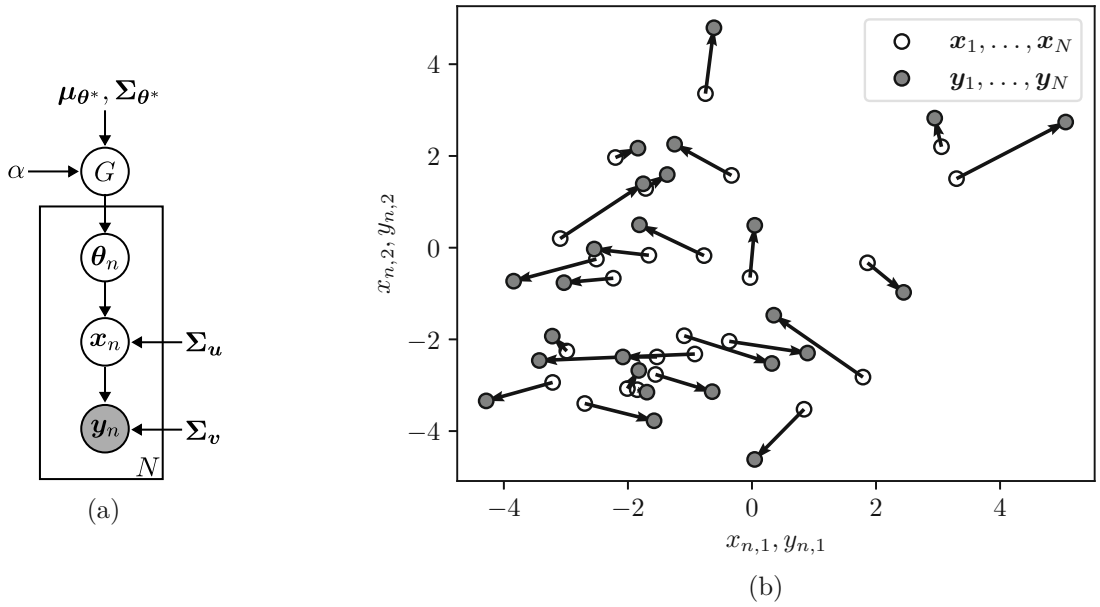


Figure 5.1: (a) Bayesian network of the overall model. The respective distributions for the hierarchical generation of $\boldsymbol{y}_n$ are given by (5.4) (5.5), (5.3) and (5.9). (b) Plot of $N = 25$ observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{25}$ and the corresponding object features $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{25}$. The arrows illustrate the added observation noise $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{25}$. The set of hyperparameters is given by $\{\alpha = 2, \boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = 5\mathbf{I}_2, \boldsymbol{\Sigma}_{\boldsymbol{u}} = \mathbf{I}_2, \boldsymbol{\Sigma}_{\boldsymbol{v}} = \mathbf{I}_2\}$.

shows an example plot of $N = 25$ observations with dimension $M = 2$. In what follows we will denote the vector of all local parameters as $\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\theta}_1^{\mathrm{T}} & \cdots & \boldsymbol{\theta}_N^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$, the vector of all indicator variables as $\boldsymbol{z} = \begin{pmatrix}z_1 & \cdots & z_N\end{pmatrix}^{\mathrm{T}}$, the vector of all objects features as $\boldsymbol{x} = \begin{pmatrix}\boldsymbol{x}_1^{\mathrm{T}} & \cdots & \boldsymbol{x}_N^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$, and the vector of all observations as $\boldsymbol{y} = \begin{pmatrix}\boldsymbol{y}_1^{\mathrm{T}} & \cdots & \boldsymbol{y}_N^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$.

In the remaining sections of this chapter we will discuss estimating the object features $\boldsymbol{x}_n$ given noisy observations $\boldsymbol{y}_n$, $n = 1, \ldots, N$, and estimating the clustering structure underlying the observations. Due to the DP prior imposed on the local parameters $\boldsymbol{\theta}_n$ (see (5.4) and (5.5)), the posterior pdf $f(\boldsymbol{x}_n|\boldsymbol{y})$ can not be calculated in closed form [8]. Consequently, we can not calculate the minimum mean square error (MMSE) estimator (cf. Section 2.3) in closed form, where the MMSE estimator for the Gaussian estimation problem described above is given by

$$\hat{\boldsymbol{x}}_n(\boldsymbol{y}) = \mathrm{E}^{(f(\boldsymbol{x}_n|\boldsymbol{y}))}\{\boldsymbol{x}_n\} = \int_{\mathbb{R}^M} \boldsymbol{x}_n f(\boldsymbol{x}_n|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}_n. \tag{5.16}$$

Thus, we resort to approximate inference, specifically to the CAVI algorithm, as described in Chapter 4, specialized for the Gaussian estimation problem in [8].

5.2 Performance Benchmarks

To assess performance results of CAVI for the above statistical model we will rely on three inference methods formulated in [8] as our benchmark methods. Two of the methods are based on exact posterior inference of the objects features $\boldsymbol{x}$, and one of the methods is based on an approximate inference method called Gibbs sampling [5]. In this section we will briefly summarize all three inference methods. For a more detailed explanation we refer to [8].

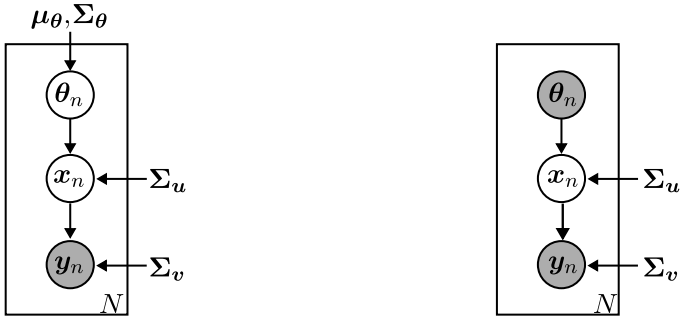


Figure 5.2: (a) Bayesian network for the case where $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$, are i.i.d. and individually distributed according to (5.17). Compared to Figure 5.1a the local parameters are not assumed to be generated by a DP. (b) Bayesian network for the case where $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$, are assumed to be observed and thus known, which is indicated by the respective shaded node.

5.2.1 Theoretical Performance Bounds

We now discuss two simplifying prior assumptions for the local parameter vector $\boldsymbol{\theta}_n$ which allow to derive the MMSE estimator (5.16) in closed form. Instead of choosing a DP prior for $\boldsymbol{\theta}_n$ as discussed in Section 5.1, we consider the following two modifications:

1. The local parameters $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$ are i.i.d. according to

$$f(\boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{5.17}$$

   with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Instead of the independence assumption (5.13) we then have

$$\boldsymbol{\theta}_n \perp\!\!\!\perp \boldsymbol{\theta}_{n'}, \quad \text{where } n, n' = 1, \ldots, N \text{ with } n \neq n'. \tag{5.18}$$

2. The local parameters $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$ are observed and thus known.

Figure 5.2 shows the modified Bayesian network for both models (cf. Figure 5.1a). Despite changing the modeling assumption of $\boldsymbol{\theta}_n$, the dependencies are as described in Section 5.1. Note that in the first case we removed the dependence between the local parameters $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$, by replacing the DP prior (5.5) with (5.17), and thus we assume there is no clustering structure underlying the local parameters and observations. In the second case the local parameters $\boldsymbol{\theta}_n$ are known, and thus any clustering structure underlying these parameters and the observations is known. From (5.1) follows the (conditional) distribution of the object feature $\boldsymbol{x}_n$ and from (5.7) follows the (conditional) distribution of the observation $\boldsymbol{y}_n$.

The simplifying assumption for the modeling of the local parameter $\boldsymbol{\theta}_n$ of both scenarios allow to derive the MMSE estimator (5.16) for the object features $\boldsymbol{x}_n$, $n = 1, \ldots, N$ in closed form, i.e., we do not have to resort to approximate inference methods. Due to the assumptions that $\boldsymbol{x}_n$ is embedded in additive white Gaussian noise (see (5.7)), that $\boldsymbol{x}_n$ is itself Gaussian because it linearly depends on Gaussian distributed variables (see (5.1)), and the independence

assumptions (5.12) and (5.18), the MMSE estimators for both scenarios adhere to the following formula [3]:

$$\hat{\boldsymbol{x}}_n(\boldsymbol{y}_n) = \boldsymbol{\mu}_{\boldsymbol{x}_n} + \boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n}\boldsymbol{\Sigma}_{\boldsymbol{y}_n}^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{y}_n}). \tag{5.19}$$

Here, $\boldsymbol{\mu}_{\boldsymbol{x}_n} = \mathrm{E}^{(f(\boldsymbol{x}_n))}\{\boldsymbol{x}_n\}$ is the mean of the n-th object feature, $\boldsymbol{\mu}_{\boldsymbol{y}_n} = \mathrm{E}^{(f(\boldsymbol{y}_n))}\{\boldsymbol{y}_n\}$ is the mean and $\boldsymbol{\Sigma}_{\boldsymbol{y}_n} = \mathrm{E}^{(f(\boldsymbol{y}_n))}\left\{(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{y}_n})(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{y}_n})^{\mathrm{T}}\right\}$ is the covariance matrix of the n-th observation, and $\boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n} = \mathrm{E}^{(f(\boldsymbol{y}_n,\boldsymbol{x}_n))}\left\{(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{y}_n})(\boldsymbol{x}_n - \boldsymbol{\mu}_{\boldsymbol{x}_n})^{\mathrm{T}}\right\}$ is the cross-covariance matrix of $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$. Note that the estimate $\hat{\boldsymbol{x}}_n = \hat{\boldsymbol{x}}_n(\boldsymbol{y}_n)$ for the n-th object feature $\boldsymbol{x}_n$ only depends on the n-th observation $\boldsymbol{y}_n$ because of the independence relations among the variables discussed above.

Given an estimate $\hat{\boldsymbol{x}}_n(\boldsymbol{y}_n)$, we will use the mean square error (MSE), which is minimized by the MMSE estimator (5.19), to quantify the average estimation accuracy. Using (5.19) it can be shown [3] to be given by

$$\begin{aligned}\mathrm{MSE}_{\min} &= \frac{1}{M}\mathrm{E}^{(f(\boldsymbol{x}_n,\boldsymbol{y}_n))}\left\{(\hat{\boldsymbol{x}}_n(\boldsymbol{y}_n) - \boldsymbol{x}_n)^{\mathrm{T}}(\hat{\boldsymbol{x}}_n(\boldsymbol{y}_n) - \boldsymbol{x}_n)\right\} \\ &= \frac{1}{M}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}_n} - \boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n}\boldsymbol{\Sigma}_{\boldsymbol{y}_n}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n}).\end{aligned} \tag{5.20}$$

For the two models summarized in Figure 5.2, the MMSE estimators for the object features $\boldsymbol{x}_n$ are given as follows (see [8] for a detailed derivation):

1. With $\boldsymbol{\mu}_{x_n} = \boldsymbol{\mu}_{y_n} = \boldsymbol{\mu}_{\boldsymbol{\theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}_n} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}$, $\boldsymbol{\Sigma}_{\boldsymbol{y}_n} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ and the cross-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n} = \boldsymbol{\Sigma}_{\boldsymbol{x}_n} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}}$, the MMSE estimator (5.19) for the case where $\boldsymbol{\theta}_n$ is distributed according to (5.17) is given by

$$\hat{\boldsymbol{x}}_n(\boldsymbol{y}_n) = \boldsymbol{\mu}_{\boldsymbol{\theta}} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}})(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}}). \tag{5.21}$$

   The minimum MSE (5.20) is obtained as

$$\mathrm{MSE}_{\min} = \frac{1}{M}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{v}}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{u}})). \tag{5.22}$$

2. In the second case the local parameter $\boldsymbol{\theta}_n$ is in addition to the observation $\boldsymbol{y}_n$ known. Thus, we have to evaluate

$$\hat{\boldsymbol{x}}_n(\boldsymbol{y}_n, \boldsymbol{\theta}_n) = \boldsymbol{\mu}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n} + \boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n})$$

   and

$$\mathrm{MSE}_{\min} = \frac{1}{M}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n} - \boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n}|\boldsymbol{\theta}_n)$$

   instead of (5.19) and (5.20). With $\boldsymbol{\mu}_{x_n|\boldsymbol{\theta}_n} = \boldsymbol{\mu}_{y_n|\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n$, $\boldsymbol{\Sigma}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}}$, $\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ and the cross-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_n\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{x}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}}$ this results

$$\hat{\boldsymbol{x}}_n(\boldsymbol{y}_n, \boldsymbol{\theta}_n) = \boldsymbol{\theta}_n + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_n). \tag{5.23}$$

The corresponding minimum MSE (5.20) is

$$\mathrm{MSE}_{\min} = \frac{1}{M} \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{v}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}}). \tag{5.24}$$

We conclude that the MMSE estimator (5.21) for the model in Figure 5.2a consists of the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ of the local parameters and a correction term that depends on the observation $\boldsymbol{y}_n$ and the various covariance matrices of the model. In contrast, the MMSE estimator (5.23) for the model in Figure 5.2b can leverage its knowledge of the local parameter $\boldsymbol{\theta}_n$. It replaces the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ with the actual value of $\boldsymbol{\theta}_n$. Finally, note that the MSEs (5.22) and (5.24) do not depend on the observation $\boldsymbol{y}$ in both cases. They only depend on the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ of the local parameters, the covariance matrix of the parameter noise $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ of the observation noise. Since the local parameters $\boldsymbol{\theta}_n$ are known in the second case the MSE is expected to be improved, i.e., lowered, compared to the first case. We will employ the MSE performance (5.22) and (5.24) of both MMSE estimators as theoretical performance bounds when evaluating the simulation results of more sophisticated methods like CAVI and Gibbs sampling in Section 5.5.

5.2.2 Gibbs Sampler

The Gibbs sampler is a Markov chain Monte Carlo (MCMC) method for obtaining samples from a multivariate distribution when direct sampling is difficult. A general discussion of Gibbs sampling and other sampling algorithms can be found in [5] and [22].

In [8] the Gibbs sampler is applied to the statistical model described in Section 5.1 (see Figure 5.1a), where the posterior pdf $f(\boldsymbol{x}_n|\boldsymbol{y})$ of the object feature $\boldsymbol{x}_n$ can not be calculated in closed form. It is shown, that a sequence of $Q \in \mathbb{N}$ samples $\boldsymbol{x}_n^{(q)}$, $q = 1, \ldots, Q$, can be generated from $f(\boldsymbol{x}_n|\boldsymbol{y})$ by sampling $\boldsymbol{\theta}_n^{(q)} = \boldsymbol{\theta}'^{(q)}_{z_n^{(q)}}$ and $\boldsymbol{x}_n^{(q)}$ from the complete conditionals of the joint posterior pdf $f(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. Recall that the local parameters $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$, take $L < N$ distinct values $\boldsymbol{\theta}'_l \in \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots\}$, $l = 1, \ldots, L$, and that $\boldsymbol{\theta}_n = \boldsymbol{\theta}'_{z_n}$ by using indicator variables $z_n \in \{1, \ldots, L\}$. We denote $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1'^{\mathrm{T}} \ \cdots \ \boldsymbol{\theta}_L'^{\mathrm{T}})^{\mathrm{T}}$ as the vector containing all cluster parameters $\boldsymbol{\theta}'_l$, $l = 1, \ldots, L$. The complete conditionals are then given by $p(z_n|\boldsymbol{z}_{\sim n}, \boldsymbol{\theta}'_{\boldsymbol{z}_{\sim n}}, \boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{\theta}'_{\boldsymbol{z}_{\sim n}}$ denotes the parameters associated with all objects $n' \neq n$, $f(\boldsymbol{\theta}'_l|\boldsymbol{\theta}'_{\sim l}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y})$, and $f(\boldsymbol{x}_n|\boldsymbol{\theta}'_{z_n}, \boldsymbol{z}_n, \boldsymbol{y}_n)$. In each iteration q the samples $z_n^{(q)}$, $\boldsymbol{\theta}_l'^{(q)}$ and $\boldsymbol{x}_n^{(q)}$ are generated from the respective complete conditionals by conditioning on the most recent samples and the observations. The sampling within each iteration q is done in the following order:

1. Generate a sample $z_n^{(q)}$ of the indicator variable z_n for all $n = 1, \ldots, N$.
2. Generate a sample $\boldsymbol{\theta}_l'^{(q)}$ of the cluster parameter $\boldsymbol{\theta}'_l$ for all $l = 1, \ldots, L$.
3. Generate a sample $\boldsymbol{x}_n^{(q)}$ of the object feature $\boldsymbol{x}_n$ for all $n = 1, \ldots, N$.

After Q iterations the samples $\boldsymbol{x}_n^{(q)}$, $q = 1, \ldots, Q$ can be regarded as samples from the marginal posterior $f(\boldsymbol{x}_n|\boldsymbol{y})$ by disregarding all other samples. The initial values for $\boldsymbol{\theta}_n^{(0)} = \boldsymbol{\theta}'^{(0)}_{z_n^{(0)}}$ and $\boldsymbol{x}_n^{(0)}$ for $q = 0$, i.e., before starting sampling iteratively from the complete conditionals, are obtained

as follows. The local parameters $\boldsymbol{\theta}_n^{(0)}$ are obtained by sampling from the base distribution, i.e.,

$$\boldsymbol{\theta}_n^{(0)} \sim \mathcal{N}\left(\boldsymbol{\theta}_n^{(0)}; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\right),$$

which means that each object n is uniquely associated with a cluster parameter $\boldsymbol{\theta}_n'^{(0)}$ $(L = N)$ and

$$z_n^{(0)} = n, \quad n = 1, \ldots, N.$$

Inspired by (5.21), the initial value of the object feature $\boldsymbol{x}_n^{(0)}$ is obtained as

$$\hat{\boldsymbol{x}}_n^{(0)}(\boldsymbol{y}_n) = \boldsymbol{\mu}_{\boldsymbol{\theta}^*} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{u}})(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} + \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}^*}), \quad n = 1, \ldots, N,$$

using the observations $\boldsymbol{y}_n$. Note that both CAVI and Gibbs sampling update one variable at a time while keeping others fixed. CAVI updates each variational pdf factor by calculating an expectation of each complete conditional given the other variational pdf factors (cf. (4.40)), whereas the Gibbs sampler samples from each complete conditional given the current values of the other samples. For a more detailed description of the above Gibbs sampling algorithm and a mathematical derivation of the complete conditionals we refer to [8].

Given the sequence of samples of the object feature $\boldsymbol{x}_n^{(q)}$, $q = 1, \ldots, Q$, which approximately represents posterior pdf $f(\boldsymbol{x}_n|\boldsymbol{y})$, we can calculate a Monte Carlo approximation of the MMSE estimator (5.16), i.e.,

$$\hat{\boldsymbol{x}}_n(\boldsymbol{y}) \approx \frac{1}{Q}\sum_{q=1}^{Q} \boldsymbol{x}_n^{(q)}. \tag{5.25}$$

From the law of large numbers it follows that this approximation is accurate for sufficiently large Q and is exact for $Q \to \infty$. The empirical MSE of the estimates $\hat{\boldsymbol{x}}_n = \hat{\boldsymbol{x}}_n(\boldsymbol{y})$ can be evaluated by averaging the squared estimation error over all objects, i.e.,

$$\text{MSE} = \frac{1}{NM}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2.$$

5.3 CAVI for Gaussian Dirichlet Process Mixtures

The goal of this section is to apply the CAVI algorithm from Section 4.3 (see Algorithm 2) to the Gaussian estimation problem described in Section 5.1. This will yield an approximate posterior distribution which allows us to approximate the MMSE estimator of the local parameters $\boldsymbol{\theta}_n$ and subsequently approximate the MMSE estimator (5.23) of the object features $\boldsymbol{x}_n$.

5.3.1 Exponential Family and Stick-breaking Representation

To utilize Algorithm 2, our initial step involves reformulating the DPM model governing the observation $\boldsymbol{y}_n$ (see (5.15)). This reformulation entails representing the DP in terms of the stick-breaking process and employing the exponential family (EF) form for the conditional

distribution $f(\boldsymbol{y}_n|\boldsymbol{\theta}_n)$ (or equivalently $f(\boldsymbol{y}_n|\boldsymbol{\theta}_k^*)$), along with the corresponding conjugate prior for the global parameter $f(\boldsymbol{\theta}_k^*)$ (cf. Section 3.5).

According to (5.11) the pdf of $\boldsymbol{y}_n$ given $\boldsymbol{\theta}_n$ is equal to the Gaussian pdf $\mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\theta}_n, \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})$, where the mean $\boldsymbol{\mu}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\theta}_n$ in the condition is a random parameter and the covariance $\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ is a known hyperparameter, i.e.,

$$f(\boldsymbol{y}_n|\boldsymbol{\theta}_n) = \frac{1}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})}} \exp\Big(-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{\theta}_n)^{\mathrm{T}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \boldsymbol{\theta}_n)\Big). \tag{5.26}$$

Multiplying out the term in the exponent of (5.26) and comparing the resulting expression with the EF distribution (2.19) yields (see [39])

$$f(\boldsymbol{y}_n|\boldsymbol{\theta}_n) = h(\boldsymbol{y}_n) \exp(\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{\theta}_n)\boldsymbol{t}(\boldsymbol{y}_n) - a(\boldsymbol{\eta}(\boldsymbol{\theta}_n))), \tag{5.27}$$

with

$$h(\boldsymbol{y}_n) = \frac{\exp(-\frac{1}{2}\boldsymbol{y}_n^{\mathrm{T}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{y}_n)}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})}}, \tag{5.28}$$

$$\boldsymbol{\eta}(\boldsymbol{\theta}_n) = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\theta}_n, \tag{5.29}$$

$$\boldsymbol{t}(\boldsymbol{y}_n) = \boldsymbol{y}_n, \tag{5.30}$$

$$a(\boldsymbol{\eta}(\boldsymbol{\theta}_n)) = \frac{1}{2}\boldsymbol{\theta}_n^{\mathrm{T}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\theta}_n. \tag{5.31}$$

Note that $\boldsymbol{\eta}(\boldsymbol{\theta}_n)$ and $\boldsymbol{t}(\boldsymbol{y}_n)$ are M-dimensional vectors. In what follows, we consider the parameter of the EF distribution (5.27) to be $\boldsymbol{\eta}_n = \boldsymbol{\eta}(\boldsymbol{\theta}_n)$ rather than $\boldsymbol{\theta}_n$, i.e., we consider the EF to be in canonical form. We will refer to both $\boldsymbol{\theta}_n$ and $\boldsymbol{\eta}_n$ as the local parameters and to both $\boldsymbol{\theta}_k^*$ and $\boldsymbol{\eta}_k^*$ as the global parameters, respectively. The parameter function $\boldsymbol{\eta}(\cdot)$ in (5.29) can be inverted as follows

$$\boldsymbol{\theta}_n = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})\boldsymbol{\eta}_n, \tag{5.32}$$

allowing to convert one representation into the other at any given time using the known covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$. Inserting (5.32) into (5.31) results

$$a(\boldsymbol{\eta}_n) = \frac{1}{2}\boldsymbol{\eta}_n^{\mathrm{T}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})\boldsymbol{\eta}_n = \frac{1}{2}\boldsymbol{\eta}_n^{\mathrm{T}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})\boldsymbol{\eta}_n. \tag{5.33}$$

We conclude that the EF representation of (5.26) is given by

$$f(\boldsymbol{y}_n|\boldsymbol{\eta}_n) = h(\boldsymbol{y}_n) \exp(\boldsymbol{\eta}_n^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{y}_n) - a(\boldsymbol{\eta}_n)), \tag{5.34}$$

with base measure (5.28), sufficient statistic (5.30) and log-partition (5.33).

We continue by reformulating the prior distribution (5.6) of the global parameter $\boldsymbol{\theta}_k^*$. Since we consider the EF to be in canonical form we apply the parameter transformation (5.29) to obtain $\boldsymbol{\eta}_k^* = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\theta}_k^*$. In Section 2.2, we showed that every EF likelihood function has

a conjugate prior. According to (2.26), it is given by

$$f(\boldsymbol{\eta}_k^*) = b(\boldsymbol{\lambda}) \exp(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_k^* - \lambda_2 a(\boldsymbol{\eta}_k^*)).$$

Inserting $a(\boldsymbol{\eta}_k^*) = \frac{1}{2} \boldsymbol{\eta}_k^{*\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}) \boldsymbol{\eta}_k^*$ (cf. (5.33)) yields

$$f(\boldsymbol{\eta}_k^*) = b(\boldsymbol{\lambda}) \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_k^* - \frac{\lambda_2}{2} \boldsymbol{\eta}_k^{*\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}) \boldsymbol{\eta}_k^*\right), \tag{5.35}$$

where $b(\boldsymbol{\lambda}) \in \mathbb{R}^+$ is a normalization constant and $\boldsymbol{\lambda}_1 \in \mathbb{R}^p$ and $\lambda_2 \in \mathbb{R}$ are hyperparameters represented by the vector $\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1^{\mathrm{T}} & \lambda_2 \end{pmatrix}^{\mathrm{T}}$. Inserting $\boldsymbol{\eta}_k^* = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\theta}_k^*$ we equivalently have

$$f(\boldsymbol{\theta}_k^*) = \tilde{b}(\boldsymbol{\lambda}) \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\theta}_k^* - \frac{\lambda_2}{2} \boldsymbol{\theta}_k^{*\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\theta}_k^*\right). \tag{5.36}$$

By an evaluation and comparison of the exponent in (5.36) with the exponent of the Gaussian pdf $\mathcal{N}(\boldsymbol{\theta}_k^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*})$, we obtain the following important relationships (see [39]):

$$\boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \frac{1}{\lambda_2} \boldsymbol{\lambda}_1, \tag{5.37}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \frac{1}{\lambda_2} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}), \tag{5.38}$$

or equivalently for the linearly transformed parameter $\boldsymbol{\eta}_k^* = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\theta}_k^*$ and the Gaussian pdf $\mathcal{N}(\boldsymbol{\eta}_k^*; \boldsymbol{\mu}_{\boldsymbol{\eta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}^*})$:

$$\boldsymbol{\mu}_{\boldsymbol{\eta}^*} = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \frac{1}{\lambda_2} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\lambda}_1, \tag{5.39}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}^*} = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} = \frac{1}{\lambda_2} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}. \tag{5.40}$$

Thus, the conjugate prior (5.35) (and similarly (5.36)) is in fact a Gaussian pdf with mean $\boldsymbol{\mu}_{\boldsymbol{\eta}^*}$ given by (5.39) and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}^*}$ given by (5.40). This is the reason for the choice of a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}_k^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*})$ for the base distribution $G_0(\boldsymbol{\theta}_k^*)$ of the DP (see (5.6)). Note that the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ of the base distribution has to be a multiple of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ of the likelihood function (5.26), which is a restriction that is due to the EF framework. The hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ of the base distribution and the hyperparameters $\boldsymbol{\lambda}_1$ and λ_2 of its equivalent representation (5.36) are related according to the equations (5.37) and (5.38). Given the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$, $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, we can determine $\boldsymbol{\lambda}_1$ and λ_2 as follows:

$$\lambda_2 \mathbf{I}_M = \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}^{-1} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}), \tag{5.41}$$

$$\boldsymbol{\lambda}_1 = \lambda_2 \boldsymbol{\mu}_{\boldsymbol{\theta}^*}. \tag{5.42}$$

It remains to compute the normalization constant $b(\boldsymbol{\lambda})$ of (5.35). As explained above the

distribution of $\boldsymbol{\eta}_k^*$ is given by

$$\mathcal{N}(\boldsymbol{\eta}_k^*; \boldsymbol{\mu}_{\boldsymbol{\eta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}^*}) = \underbrace{\frac{1}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{\eta}^*})}}}_{\triangleq c} \underbrace{\exp\left(-\frac{1}{2}(\boldsymbol{\eta}_k^* - \boldsymbol{\mu}_{\boldsymbol{\eta}^*})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}^*}^{-1} (\boldsymbol{\eta}_k^* - \boldsymbol{\mu}_{\boldsymbol{\eta}^*})\right)}_{\triangleq d},$$

where we temporarily defined the two factors c and d. Replacing $\boldsymbol{\mu}_{\boldsymbol{\eta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}^*}$ with (5.39) and (5.40), and using $\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ yields

$$c = \frac{1}{\sqrt{(2\pi)^M \det\left(\frac{1}{\lambda_2}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\right)}} = \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1})}} \tag{5.43}$$

and

$$\begin{aligned}
d &= \exp\left(-\frac{1}{2}\left(\boldsymbol{\eta}_k^* - \frac{1}{\lambda_2}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\lambda}_1\right)^{\mathrm{T}} \left(\frac{1}{\lambda_2}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\right)^{-1} \left(\boldsymbol{\eta}_k^* - \frac{1}{\lambda_2}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\lambda}_1\right)\right) \\
&= \exp\left(-\frac{\lambda_2}{2}\left(\boldsymbol{\eta}_k^* - \frac{1}{\lambda_2}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\lambda}_1\right)^{\mathrm{T}} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\eta}_k^* - \frac{1}{\lambda_2}\boldsymbol{\lambda}_1\right)\right) \\
&= \exp\left(-\frac{\lambda_2}{2}\left(\boldsymbol{\eta}_k^{*\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\eta}_k^* - \frac{1}{\lambda_2}\boldsymbol{\eta}_k^{*\mathrm{T}}\boldsymbol{\lambda}_1 - \frac{1}{\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_k^* + \frac{1}{\lambda_2^2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\lambda}_1\right)\right) \\
&= \exp\left(-\frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\lambda}_1\right) \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_k^* - \frac{\lambda_2}{2}\boldsymbol{\eta}_k^{*\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\eta}_k^*\right) \\
&= \exp\left(-\frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\lambda}_1\right) \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_k^* - \lambda_2 a(\boldsymbol{\eta}_k^*)\right),
\end{aligned} \tag{5.44}$$

where we used (5.33) in the last step. When we compare the two factors (5.43) and (5.44) with (5.35), we obtain

$$\begin{aligned}
b(\boldsymbol{\lambda}) &= \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1})}} \exp\left(-\frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\lambda}_1\right) \\
&= \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det((\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1})}} \exp\left(-\frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\lambda}_1\right),
\end{aligned} \tag{5.45}$$

i.e., we are able to compute the normalization constant $b(\boldsymbol{\lambda})$ in closed form and do not have to perform numerical integration as described in (2.27).

We are now ready to summarize the DPM model for N conditionally independent observations $\boldsymbol{y}_n$ as (cf. (3.41))

$$v_k \overset{\text{i.i.d.}}{\sim} \mathcal{B}(v_k; 1, \alpha), \tag{5.46a}$$

$$\pi_k(v_1, \ldots, v_k) = v_k \prod_{i=1}^{k-1}(1 - v_i), \tag{5.46b}$$

$$\boldsymbol{\eta}_k^* \overset{\text{i.i.d.}}{\sim} G_0(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda}), \tag{5.46c}$$

$$z_n|\boldsymbol{\pi}(v_1, v_2, \ldots) \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n|\boldsymbol{\pi}(v_1, v_2, \ldots)), \tag{5.46d}$$

$$\boldsymbol{y}_n|z_n, \boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*, \ldots \sim f(\boldsymbol{y}_n|\boldsymbol{\eta}_{z_n}^*), \tag{5.46e}$$

for $k = 1, 2, \ldots$ and $n = 1, \ldots, N$. Here, $v_1, v_2, \ldots$ are the auxiliary variables of the stick-breaking process and $z_1, \ldots, z_N$ are indicator variables. Note that our notation does not distinguish between scalar auxiliary variables v_k and the observation noise vector $\boldsymbol{v}_n$. It will be clear from the context if we mean either of the two quantities. According to (5.35) the base distribution $G_0(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda})$ is given by

$$G_0(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda}) = b(\boldsymbol{\lambda}) \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_k^* - \frac{\lambda_2}{2} \boldsymbol{\eta}_k^{*\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}) \boldsymbol{\eta}_k^*\right),$$

where $\boldsymbol{\lambda}_1$ and λ_2 are determined by (5.41) and (5.42), respectively. The conditional pdf $f(\boldsymbol{y}_n | \boldsymbol{\eta}_{z_n}^*)$ is in EF form as specified in (5.34), i.e.,

$$f(\boldsymbol{y}_n | \boldsymbol{\eta}_{z_n}^*) = h(\boldsymbol{y}_n) \exp(\boldsymbol{\eta}_{z_n}^{*\mathrm{T}} \boldsymbol{t}(\boldsymbol{y}_n) - a(\boldsymbol{\eta}_{z_n}^*))$$

with

$$\begin{aligned} h(\boldsymbol{y}_n) &= \frac{\exp(-\frac{1}{2} \boldsymbol{y}_n^{\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{y}_n)}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})}}, \\ \boldsymbol{t}(\boldsymbol{y}_n) &= \boldsymbol{y}_n, \\ a(\boldsymbol{\eta}_{z_n}^*) &= \frac{1}{2} \boldsymbol{\eta}_{z_n}^{*\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}) \boldsymbol{\eta}_{z_n}^*. \end{aligned}$$

The n-th indicator variable z_n allows us to determine the n-th local parameter $\boldsymbol{\eta}_n$ by the relation $\boldsymbol{\eta}_n = \boldsymbol{\eta}_{z_n}^*$, which means that $f(\boldsymbol{y}_n | \boldsymbol{\eta}_{z_n}^*)$ is equal to $f(\boldsymbol{y}_n | \boldsymbol{\eta}_n)$. Recall that we refer to both $\boldsymbol{\eta}_n$ and $\boldsymbol{\theta}_n$ as the local parameter because they are connected by the deterministic relation $\boldsymbol{\theta}_n = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}) \boldsymbol{\eta}_n$ (see (5.32)).

5.3.2 CAVI Algorithm

We proceed with applying the CAVI algorithm (see Algorithm 2) to the reformulation (5.46) of the model described in Section 5.1. The output of the CAVI algorithm is the approximate posterior distribution $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{y})$ of the auxiliary variables $\boldsymbol{v} = (v_1 \ \cdots \ v_{T-1})^{\mathrm{T}}$, global parameters $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^{*\mathrm{T}} \ \cdots \ \boldsymbol{\eta}_T^{*\mathrm{T}})^{\mathrm{T}}$ and indicator variables $\boldsymbol{z} = (z_1 \ \cdots \ z_N)^{\mathrm{T}}$. This approximation of the posterior pdf involves a truncated mean field approximation and entails truncation of (5.46) using the truncated stick-breaking model (4.44). The truncation level is given by the truncation parameter $T \in \mathbb{N}$, where the component label of the truncated DPM is $t \in \{1, \ldots, T\}$. For details of the truncated mean field approximation, we refer to Section 4.4.1. Note that the observation $\boldsymbol{y}$ is still assumed to be generated as described in Section 5.1, i.e., we do not truncate the model according to which the observations $\boldsymbol{y}_n$ are generated.

With the specification of the EF, it is possible to calculate the expectations $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\}$ and $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}$ that we were unable to further elaborate on in the derivations of Algorithm 2 in Section 4.4. According to (4.77) the approximate posterior pdf $q_t(\boldsymbol{\eta}_t^*)$ maintains the functional form of the conjugate prior (5.35), but it is parameterized by $\boldsymbol{\tau}_{t,1}$ and $\tau_{t,2}$, which are updated versions of the hyperparameters $\boldsymbol{\lambda}_1$ and λ_2. Thus, the posterior mean $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\}$ of $\boldsymbol{\eta}_t^*$ is

given by (5.39) with $\boldsymbol{\lambda}_1$ and λ_2 replaced by $\boldsymbol{\tau}_{t,1}$ and $\tau_{t,2}$, i.e.,

$$\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\} = \frac{1}{\tau_{t,2}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\tau}_{t,1}. \tag{5.47}$$

Note that this is the approximate MMSE estimate for $\boldsymbol{\eta}_t^*$ since it is the mean of $\boldsymbol{\eta}_t^*$ with respect to the approximate posterior $q_t(\boldsymbol{\eta}_t^*)$. Similarly, using (5.40) with $\boldsymbol{\lambda}_1$ and λ_2 replaced by $\boldsymbol{\tau}_{t,1}$ and $\tau_{t,2}$, we obtain

$$\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\left(\boldsymbol{\eta}_t^* - \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}_t^* - \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\}\right)\right\} = \frac{1}{\tau_{t,2}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}, \tag{5.48}$$

i.e., the approximate posterior covariance matrix of $\boldsymbol{\eta}_t^*$. The posterior expectations (5.47) and (5.48) allow us to compute the last missing expectation $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}$. For the sake of brevity, we will use the abbreviations $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*} = \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\}$, $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_t^*} = \mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\left(\boldsymbol{\eta}_t^* - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}_t^* - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}\right)\right\}$ and $\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ in the calculation of $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}$. Replacing the log-partition function $a(\boldsymbol{\eta}_t^*)$ in $\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}$ with (5.33) yields

$$\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\} = \frac{1}{2}\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_k^{*\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\eta}_k^*\right\}. \tag{5.49}$$

Here, it remains to compute the expectation of the quadratic form $\boldsymbol{\eta}_k^{*\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\eta}_k^*$ with respect to the approximate posterior $q_t(\boldsymbol{\eta}_t^*)$. This can be shown [52] to be given by

$$\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_k^{*\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\eta}_k^*\right\} = \mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_t^*}\right) + \tilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\tilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}.$$

Replacing $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_t^*}$ with (5.47) and (5.48) yields

$$\begin{aligned}
\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_k^{*\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\eta}_k^*\right\} &= \mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\frac{1}{\tau_{t,2}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\right) + \frac{1}{\tau_{t,2}^2}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\tau}_{t,1} \\
&= \mathrm{tr}\left(\frac{1}{\tau_{t,2}}\mathbf{I}_M\right) + \frac{1}{\tau_{t,2}^2}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\tau}_{t,1} \\
&= \frac{1}{\tau_{t,2}^2}\left(M\tau_{t,2} + \boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\tau}_{t,1}\right).
\end{aligned} \tag{5.50}$$

By inserting (5.50) into (5.49) we obtain the final result:

$$\begin{aligned}
\mathrm{E}^{(q_t(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\} &= \frac{1}{2\tau_{t,2}^2}\left(M\tau_{t,2} + \boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n}^{-1}\boldsymbol{\tau}_{t,1}\right) \\
&= \frac{1}{2\tau_{t,2}^2}\left(M\tau_{t,2} + \boldsymbol{\tau}_{t,1}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\tau}_{t,1}\right).
\end{aligned} \tag{5.51}$$

We are now able to adapt the more general formulation of CAVI for DPMs in Algorithm 2 to our specific case of a Gaussian model. The corresponding update equations are given in (4.98), (4.99) and (4.100). The solution for the ELBO is given in (4.101). We start by inserting the two missing expectations (5.47) and (5.51), and the sufficient statistic (5.30), into the update equations. This yields the following list of equations:

- Variational parameters $\boldsymbol{\phi}_n^{(\ell)} = \begin{pmatrix} \phi_{n,1}^{(\ell)} & \dots & \phi_{n,T}^{(\ell)} \end{pmatrix}$ of $q_n^{(\ell)}(z_n; \boldsymbol{y}) = \mathcal{C}\left(z_n; \boldsymbol{\phi}_n^{(\ell)}\right)$:

$$\phi_{n,t}^{(\ell)} = \frac{\exp\left(S_{n,t}^{(\ell)}\right)}{\sum_{i=1}^{T} \exp\left(S_{n,t}^{(\ell)}\right)}, \tag{5.52a}$$

where

$$\begin{aligned} S_{n,t}^{(\ell)} = {} & \frac{1}{\tau_{t,2}} \boldsymbol{\tau}_{t,1}^{\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{y}_n - \frac{1}{2\tau_{t,2}^2} \left(M \tau_{t,2} + \boldsymbol{\tau}_{t,1}^{\mathrm{T}} (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{\tau}_{t,1} \right) \\ & + \Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1} + \gamma_{t,2}) + \sum_{j=1}^{t-1} \Psi(\gamma_{j,2}) - \Psi(\gamma_{j,1} + \gamma_{j,2}) \end{aligned} \tag{5.52b}$$

- Variational parameters $\boldsymbol{\gamma}_t^{(\ell)} = \begin{pmatrix} \gamma_{t,1}^{(\ell)} & \gamma_{t,2}^{(\ell)} \end{pmatrix}^{\mathrm{T}}$ of $q_t^{(\ell)}(v_t; \boldsymbol{y}) = \mathcal{B}\left(v_t; \gamma_{t,1}^{(\ell)}, \gamma_{t,2}^{(\ell)}\right)$:

$$\gamma_{t,1}^{(\ell)} = 1 + \sum_{n=1}^{N} \phi_{n,t}, \tag{5.53a}$$

$$\gamma_{t,2}^{(\ell)} = \alpha + \sum_{n=1}^{N} \sum_{j=t+1}^{T} \phi_{n,j}. \tag{5.53b}$$

- Variational parameters $\boldsymbol{\tau}_t^{(\ell)} = \begin{pmatrix} \boldsymbol{\tau}_{t,1}^{(\ell)\mathrm{T}} & \tau_{t,2}^{(\ell)} \end{pmatrix}^{\mathrm{T}}$ of $q_t^{(\ell)}(\boldsymbol{\eta}_t^*; \boldsymbol{y}) \propto \exp\left(\boldsymbol{\tau}_{t,1}^{(\ell)\mathrm{T}} \boldsymbol{\eta}_t^* - \tau_{t,2}^{(\ell)} a(\boldsymbol{\eta}_t^*) \right)$:

$$\boldsymbol{\tau}_{t,1}^{(\ell)} = \boldsymbol{\lambda}_1 + \sum_{n=1}^{N} \phi_{n,t} \boldsymbol{y}_n, \tag{5.54a}$$

$$\tau_{t,2}^{(\ell)} = \lambda_2 + \sum_{n=1}^{N} \phi_{n,t}. \tag{5.54b}$$

Using the base measure (5.28), the sufficient statistic (5.30), the normalization constant (5.45) and the two expectations (5.47) and (5.51) in (4.101), the ELBO is given by

$$\begin{aligned} L(q; \boldsymbol{y}) = {} & \sum_{n=1}^{N} \left(\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{y}_n | \boldsymbol{\eta}_{z_n}^*)\} + \mathrm{E}^{(q)}\{\ln p^{(T)}(z_n | \boldsymbol{v})\} + \mathrm{E}^{(q)}\{\ln q_n(z_n)\} \right) \\ & + \sum_{t=1}^{T} \left(\mathrm{E}^{(q)}\{\ln f(\boldsymbol{\eta}_t^*)\} + \mathrm{E}^{(q)}\{\ln q_t(\boldsymbol{\eta}_t^*)\} \right) \\ & + \sum_{t=1}^{T-1} \left(\mathrm{E}^{(q)}\{\ln f(v_t)\} + \mathrm{E}^{(q)}\{\ln q_t(v_t)\} \right), \end{aligned} \tag{5.55a}$$

with

$$\begin{aligned} & \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{y}_n | \boldsymbol{\eta}_{z_n}^*)\} \\ & = \sum_{t=1}^{T} \phi_{n,t} \left(\ln \frac{\exp\left(-\frac{1}{2} \boldsymbol{y}_n^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_n\right)}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma})}} + \frac{1}{\tau_{t,2}} \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_n - \frac{1}{2\tau_{t,2}^2} \left(M \tau_{t,2} + \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1} \right) \right), \end{aligned} \tag{5.55b}$$

$$\mathrm{E}^{(q)}\{\ln p^{(T)}(z_n|\boldsymbol{v})\} = \sum_{t=1}^{T-1}\left(\phi_{n,t}(\Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1}+\gamma_{t,2})) + \left(\sum_{i=t+1}^{T}\phi_{n,i}\right)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1}+\gamma_{t,2}))\right), \tag{5.55c}$$

$$\mathrm{E}^{(q)}\{\ln q_n(z_n)\} = \boldsymbol{\phi}_n^{\mathrm{T}}\boldsymbol{\phi}_n, \tag{5.55d}$$

$$\mathrm{E}^{(q)}\{\ln f(\boldsymbol{\eta}_t^*)\} = \ln \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} - \frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1 + \frac{1}{\tau_{t,2}}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} - \frac{\lambda_2}{2\tau_{t,2}^2}\left(M\tau_{t,2} + \boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1}\right), \tag{5.55e}$$

$$\mathrm{E}^{(q)}\{\ln q_t(\boldsymbol{\eta}_t^*)\} = \ln \frac{1}{\sqrt{\left(\frac{2\pi}{\tau_{t,2}}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} - \frac{1}{2\tau_{t,2}}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} + \frac{1}{\tau_{t,2}}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} - \frac{1}{2\tau_{t,2}}\left(M\tau_{t,2} + \boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1}\right), \tag{5.55f}$$

$$\mathrm{E}^{(q)}\{\ln f(v_t)\} = \ln \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} + (\alpha-1)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1}+\gamma_{t,2})), \tag{5.55g}$$

$$\mathrm{E}^{(q)}\{\ln q_t(v_t)\} = \ln \frac{\Gamma(\gamma_{t,1}+\gamma_{t,2})}{\Gamma(\gamma_{t,1})\Gamma(\gamma_{t,2})} + (\gamma_{t,1}-1)(\Psi(\gamma_{t,1}) - \Psi(\gamma_{t,1}+\gamma_{t,2})) + (\gamma_{t,2}-1)(\Psi(\gamma_{t,2}) - \Psi(\gamma_{t,1}+\gamma_{t,2})), \tag{5.55h}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{y}_n|\boldsymbol{\theta}_n} = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}$ because of brevity.

Finally, the CAVI algorithm for a DPM with Gaussian components, where the mean of each component is assumed to be random and the covariance matrix of each component is assumed to be identical to the covariance matrix of the other components and known, can be summarized as shown in Algorithm 3. In each iteration ℓ we update each variational parameter with the most recent updates of the other variational parameters. At the end of each iteration, the current state of the ELBO is evaluated and the relative change is monitored to assess convergence. Different ways to assess convergence and the initialization of the CAVI algorithm have been discussed in Section 4.5. The output of the algorithm is given by the approximate posterior pdf $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{y})$, which is a local optimum for the CAVI optimization problem (4.26).

5.4 Estimation of Model Parameters

Using the approximate posterior pdf $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{y})$, i.e., the output of Algorithm 3, we can estimate the mixing proportions π_t, $t = 1, \ldots, T$, the global parameters $\boldsymbol{\theta}_t^*$, $t = 1, \ldots, T$, and the indicator variables z_n, $n = 1, \ldots, N$ of the Gaussian DPM model associated with the observations $\boldsymbol{y}_n$. Moreover, using these estimates enables to approximate the MMSE estimate (5.23) of the object features $\boldsymbol{x}_n$.

We start with the approximate MMSE estimate $\hat{\pi}_t = \hat{\pi}_t(\boldsymbol{y})$ of the mixing proportions π_t. In

Algorithm 3: CAVI for Gaussian DPM Models

Input: Observations $\boldsymbol{y}$, truncation level T, set of hyperparameters $\{\alpha, \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}\}$

Output: Variational factor pdfs

- $q_n^*(z_n; \boldsymbol{y})$ for $n = 1, \ldots, N$
- $q_t^*(v_t; \boldsymbol{y})$ for $t = 1, \ldots, T-1$
- $q_t^*(\boldsymbol{\eta}_t^*; \boldsymbol{y})$ for $t = 1, \ldots, T$

Initialize: Variational parameters $\boldsymbol{\phi}_{1:N}^{(0)}$, $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$;

while *the ELBO has not converged* **do**

  $\ell = \ell + 1$ **for** n *from* 1 *to* N **do**

    for t *from* 1 *to* T **do**

      set $S_{n,t}^{(\ell)}$ according to (5.52b)

    for t *from* 1 *to* T **do**

      set $\phi_{n,t}^{(\ell)}$ according to (5.52a)

  for t *from* 1 *to* $T-1$ **do**

    set $\boldsymbol{\gamma}_t^{(\ell)}$ according to (5.53)

  for t *from* 1 *to* T **do**

    set $\boldsymbol{\tau}_t^{(\ell)}$ according to (5.54)

  Compute the ELBO $\mathcal{L}(q^{(\ell)}; \boldsymbol{y})$ according to (5.55)

return $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{y}) = \Big(\prod_{t=1}^{T-1} q_t^*(v_t; \boldsymbol{y})\Big)\Big(\prod_{t=1}^{T} q_t^*(\boldsymbol{\eta}_t^*; \boldsymbol{y})\Big)\Big(\prod_{n=1}^{N} q_n^*(z_n; \boldsymbol{y})\Big)$

(4.103) and (4.104) we already calculated the posterior mean of the mixing proportions π_t with respect to the approximate posterior pdf $q^*(\boldsymbol{v}; \boldsymbol{y})$. Thus, we already have a formula for the approximate MMSE estimates of the mixing proportions π_t that is given by

$$\hat{\pi}_t = \mathrm{E}^{(q^*(\boldsymbol{v};\boldsymbol{x}))}\{\pi_t(\boldsymbol{v}_{1:t})\} = \frac{\gamma_{t,1}}{\gamma_{t,1}+\gamma_{t,2}} \prod_{j=1}^{t-1} \left(\frac{\gamma_{j,2}}{\gamma_{j,1}+\gamma_{j,2}}\right), \quad t = 1, \ldots, T-1, \tag{5.56a}$$

$$\hat{\pi}_T = \mathrm{E}^{(q^*(\boldsymbol{v};\boldsymbol{x}))}\Big\{\pi_T\Big(\boldsymbol{v}_{1:(T-1)}\Big)\Big\} = \prod_{j=1}^{T-1} \left(\frac{\gamma_{j,2}}{\gamma_{j,1}+\gamma_{j,2}}\right). \tag{5.56b}$$

We proceed with the approximate MMSE estimates $\hat{\boldsymbol{\theta}}_t^* = \hat{\boldsymbol{\theta}}_t^*(\boldsymbol{y})$ and $\hat{\boldsymbol{\eta}}_t^* = \hat{\boldsymbol{\eta}}_t^*(\boldsymbol{y})$ of the global parameters $\boldsymbol{\theta}_t^*$ and $\boldsymbol{\eta}_t^*$. According to (5.47), the posterior mean of the global parameters $\boldsymbol{\eta}_t^*$ with respect to the approximate posterior pdf $q_t^*(\boldsymbol{\eta}_t^*; \boldsymbol{y})$ is given by

$$\hat{\boldsymbol{\eta}}_t^* = \mathrm{E}^{(q_t^*(\boldsymbol{\eta}_t^*;\boldsymbol{y}))}\{\boldsymbol{\eta}_t^*\} = \frac{1}{\tau_{t,2}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{\tau}_{t,1}, \quad t = 1, \ldots, T.$$

Using the deterministic transformation (5.32) we obtain

$$\hat{\boldsymbol{\theta}}_t^* = (\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})\hat{\boldsymbol{\eta}}_t^* = \frac{1}{\tau_{t,2}}\boldsymbol{\tau}_{t,1}, \quad t = 1, \ldots, T, \tag{5.57}$$

for the approximate MMSE estimates of the global parameters $\boldsymbol{\theta}_t^*$.

Next, we estimate the indicator variables z_n and the clustering structure of the observations $\boldsymbol{y}_n$. In order to obtain a hard clustering, we choose the approximate maximum a-posteriori (MAP) estimate for the estimation of the indicator variables z_n. The approximate posterior pmf of the n-th indicator variable z_n is given by the Categorical distribution $q_n^*(z_n; \boldsymbol{y}) = \mathcal{C}(z_n; \boldsymbol{\phi}_n)$. Calculating the mode of this distribution, i.e., the value t for which $q_n^*(z_n = t; \boldsymbol{y})$ takes it maximum, yields the approximate MAP estimate $\hat{z}_n = \hat{z}_n(\boldsymbol{y})$. The location of the maximum of the Categorical distribution $\mathcal{C}(z_n; \boldsymbol{\phi}_n)$ is given by the location of the maximum entry of the vector $\boldsymbol{\phi}_n^{\mathrm{T}} = (\phi_{n,1} \ \ldots \ \phi_{n,T})$ (cf. (2.5)). Thus, we obtain

$$\hat{z}_n = \arg\max_t \phi_{n,t}, \quad n = 1, \ldots, N, \tag{5.58}$$

for the approximate MAP estimates of the indicator variables. Each estimate $\hat{z}_n$ represents a hard assignment of the n-th object to one of the $t \in \{1, \ldots, T\}$ components of the truncated DPM. Applying (2.10), we can count the number of objects $\hat{N}_t$ that we associate with the t-th mixture component, i.e.,

$$\hat{N}_t = \sum_{n=1}^{N} \mathbb{1}(\hat{z}_n = t),$$

and determine the number of cluster as

$$\hat{L} = \sum_{t=1}^{T} \mathbb{1}\Big(\hat{N}_t > 0\Big). \tag{5.59}$$

Note that the estimated number of clusters $\hat{L}$ is likely to be smaller than the truncation level T meaning there can be components t with no associated object n ($\hat{N}_t = 0$).

Combining the estimates $\hat{z}_n$ of the indicator variables and the estimates $\hat{\boldsymbol{\theta}}_t^*$ of the global parameters, we can obtain estimates $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\boldsymbol{y})$ for the local parameters $\boldsymbol{\theta}_n$ as follows

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{\hat{z}_n}^*. \tag{5.60}$$

Here, each object n is assigned a global parameter $\hat{\boldsymbol{\theta}}_t^*$ through the local parameter $\hat{\boldsymbol{\theta}}_n$ by using the indicator variable $\hat{z}_n$. These estimates (5.60) can be used to approximate the MMSE estimator (5.23) of the object features $\boldsymbol{x}_n$. As explained in Section 5.2.1, the MSE of this estimator represents an upper performance bound, because it assumes the local parameters $\boldsymbol{\theta}_n$ to be known. Thus, the lower the error $e_n = \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n$ between our estimate $\hat{\boldsymbol{\theta}}_n$ and the true local parameter $\boldsymbol{\theta}_n$, the better will be our approximation of the MMSE estimator (5.23). Inserting $\hat{\boldsymbol{\theta}}_n$ into (5.23) results

$$\hat{\boldsymbol{x}}_n = \hat{\boldsymbol{\theta}}_n + \boldsymbol{\Sigma}_{\boldsymbol{u}}(\boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}(\boldsymbol{y}_n - \hat{\boldsymbol{\theta}}_n), \tag{5.61}$$

i.e., an estimate $\hat{\boldsymbol{x}}_n = \hat{\boldsymbol{x}}_n(\boldsymbol{y})$ for the n-th object feature $\boldsymbol{x}_n$. Note that while equation (5.61) may imply that the estimate $\hat{\boldsymbol{x}}_n$ solely depends on the n-th observation $\boldsymbol{y}_n$, we have to consider that the estimate $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\boldsymbol{y})$ depends on all observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ through the estimates $\hat{\boldsymbol{\theta}}_t^*$ and $\hat{z}_n$. Also note that the accuracy of all estimates depends on the input values chosen for the

CAVI algorithm (see Section 4.5 for a discussion).

5.5 Simulation Results

In the final section of this chapter, we present simulation results that are obtained from the application of the CAVI algorithm, specifically from Algorithm 3, in conjunction with the corresponding approximate MMSE/MAP estimators derived in Section 5.4. We will discuss the simulation procedure, examine clustering results, analyze the behavior of the ELBO for different types of initialization, and compare MSE performance results with those obtained in [8].

5.5.1 Simulation Setup

In order to be able to compare our simulation results to the simulation results of [8], we will adopt the following model parameters. Each object n is associated with a feature vector $\boldsymbol{x}_n = (x_{n,1} \;\; x_{n,2})^{\mathrm{T}} \in \mathbb{R}^2$, a local parameter vector $\boldsymbol{\theta}_n = (\theta_{n,1} \;\; \theta_{n,2})^{\mathrm{T}} \in \mathbb{R}^2$, and an observation vector $\boldsymbol{y}_n = (y_{n,1} \;\; y_{n,2})^{\mathrm{T}} \in \mathbb{R}^2$, i.e., we consider a two-dimensional model ($M = 2$). The local parameter $\boldsymbol{\theta}_n$ is generated from a DP with Gaussian base distribution $G_0(\boldsymbol{\theta}_k^*) = \mathcal{N}(\boldsymbol{\theta}_k^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*})$, where $\boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = 5\mathbf{I}_2$. In terms of the concentration parameter α, we will examine the three different values $\alpha \in \{0.5, 1, 5\}$. As explained in Section 5.1, the feature vector $\boldsymbol{x}_n$ is statistically related to local parameter $\boldsymbol{\theta}_n$ according to (5.1), where we assume that the distribution of the parameter noise $\boldsymbol{u}_n$ is given by a zero-mean Gaussian distribution (see (5.2)) with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \mathbf{I}_2$. Similarly, the observation $\boldsymbol{y}_n$ is statistically related to the feature vector $\boldsymbol{x}_n$ according to (5.7), where we assume that the distribution of the observation noise $\boldsymbol{v}_n$ is given by a zero-mean Gaussian distribution (see (5.8)) with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \mathbf{I}_2$. The total number of objects N is selected within the range of $N = 1, \ldots, 50$. A summary of all model parameters is provided by Table 5.1.

Parameter	Value
M	2
N	$\{1, \ldots, 50\}$
$\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$	$\mathbf{0}$
$\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$	$5\mathbf{I}_2$
$\boldsymbol{\Sigma}_{\boldsymbol{u}}$	$\mathbf{I}_2$
$\boldsymbol{\Sigma}_{\boldsymbol{v}}$	$\mathbf{I}_2$
α	$\{0.5, 1, 5\}$

Table 5.1: Model parameters for the simulations.

The subsequent simulation results are obtained by using Algorithm 3 and the approximate MMSE/MAP estimators (5.56) – (5.61). Our implementation of the algorithm and the estimators can be found in [53]. We used the programming language Python in combination with the package NumPy for numerical calculations. Observations $\boldsymbol{y}_n$, $n = 1, \ldots, N$, are generated with the parameters in Table 5.1, i.e., we use synthetic data for our simulations. Recall that we assume that the n-th observation $\boldsymbol{y}_n$ is distributed according to a Gaussian DPM (see (5.15)). For the generation of the local parameters $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$, we used the Chinese restaurant

process (cf. Section 3.2.3), which also provides cluster assignment variables z_n, $n = 1, \ldots, N$. The local parameters $\boldsymbol{\theta}_n$ are then used to generate the observations $\boldsymbol{y}_n$ by adding parameter noise $\boldsymbol{u}_n$ and observation noise $\boldsymbol{v}_n$ as explained above. The observations $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{y}_N^{\mathrm{T}})^{\mathrm{T}}$ and values for the truncation parameter T and hyperparameters $\{\alpha, \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}\}$, are then used as input for the CAVI algorithm. Initial values $\boldsymbol{\phi}_{1:N}^{(0)}$, $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$ of the variational parameters are set according to a specified initialization method (see Section 4.5 for a discussion about initialization). The percentage change (4.105) of the ELBO is monitored to assess convergence. When the percentage change of the ELBO falls below a prespecified convergence threshold ϵ the algorithm stops and outputs the current values of the variational parameters $\boldsymbol{\phi}_{1:N}$, $\boldsymbol{\gamma}_{1:(T-1)}$ and $\boldsymbol{\tau}_{1:T}$ of the approximate posterior distribution

$$q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}; \boldsymbol{y}) = \left(\prod_{t=1}^{T-1} q_t(v_t; \boldsymbol{y})\right)\left(\prod_{t=1}^{T} q_t(\boldsymbol{\eta}_t^*; \boldsymbol{y})\right)\left(\prod_{n=1}^{N} q_n(z_n; \boldsymbol{y})\right).$$

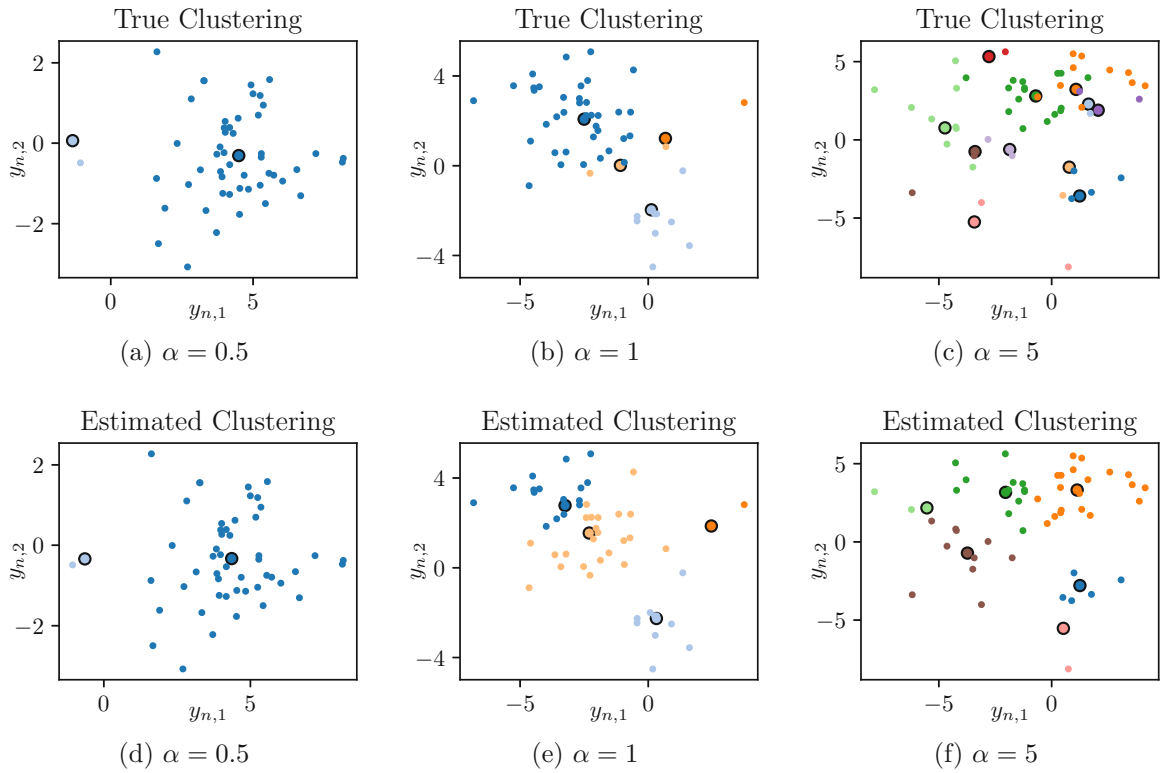
The approximate MMSE/MAP estimators from Section 5.4 are applied in a post-processing step, which results a posterior estimate of the clustering structure and the object features.

5.5.2 Estimation of the Cluster Assignments and Parameters

We first want to show simulation results for the estimation of the clustering structure underlying the observation $\boldsymbol{y}$ in the case of $N = 50$. Such clustering results are subsequently used to estimate the object features $\boldsymbol{x}_n$. For the input of the CAVI algorithm we choose the truncation parameter $T = N = 50$ and assume the set $\{\alpha, \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}\}$ of hyperparameters to be known, i.e., to be equal to the parameters in Table 5.1 of the model governing the observations. We initialize the local variational parameter $\boldsymbol{\phi}_{1:N}^{(0)}$ by uniquely associating each object n with one component t of the truncated DPM, i.e., $\phi_{n,t}^{(0)}$ equals one for $n = t$ and zero else. The initialization for the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$ is obtained from the initial value of the local variational parameter $\boldsymbol{\phi}_{1:N}^{(0)}$ by using (5.53) and (5.54). The convergence threshold is set to $\epsilon = 0.01\,\%$. In the post-processing, we estimate the T mixing proportions $\hat{\pi}_1, \ldots, \hat{\pi}_T$ by applying (5.56), the T global parameters (component means) $\hat{\boldsymbol{\theta}}_1^*, \ldots, \hat{\boldsymbol{\theta}}_T^*$ by applying (5.57) and the N indicator variables $\hat{z}_1, \ldots, \hat{z}_N$ by applying (5.58). The estimated number of clusters $\hat{L}$ is obtained by (5.59).

Figure 5.3 shows three different simulation results corresponding to three different values of the concentration parameter $\alpha \in \{0.5, 1, 5\}$. For each value of α, we compare the true clustering structure of $N = 50$ observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{50}$ to the estimated clustering structure obtained by using the CAVI algorithm. Figures 5.3a to 5.3c show the true number of clusters $L < N$ (different colors) in the observation $\boldsymbol{y}$ and the corresponding cluster means $\boldsymbol{\theta}_l' \in \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots\}$, $l = 1, \ldots, L$, (big dots). Out of the T component estimates $\hat{\boldsymbol{\theta}}_t^*$, $t = 1, \ldots, T$, only $\hat{L}$ are shown in Figures 5.3d to 5.3f because no object has been associated to one of the other $T - \hat{L}$ components of the truncated DPM by the approximate MAP estimator. Again the clustering structure is illustrated by the different colors.

The expected number of clusters $\bar{L}$ (see (3.19)) are 2.94, 5.84 and 12.46 for α equal to 0.5, 1.5 and 5, respectively. We conclude that the higher the concentration parameter α the more small



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 5$

(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 5$

Figure 5.3: True and estimated clustering structure of the observation $\boldsymbol{y}$. Big dots refer to the value of cluster means and small dots refer to the values of the observations. Each cluster is represented by a unique color. (a)-(c) True clustering properties. There are $L = 2$, $L = 4$ and $L = 11$ clusters present for α equal to 0.5, 1.5 and 5, respectively. (d)-(f) Estimated clustering results. The estimated numbers of clusters $\hat{L}$ are 2, 4 and 6 for α equal to 0.5, 1.5 and 5, respectively.

and densely packed clusters are present within the observations and the more the estimated clustering structure deviates from the true clustering structure.

5.5.3 Behavior of the ELBO for Different Initialization Types

Next, we want to show how the initialization $\boldsymbol{\phi}_{1:N}^{(0)}$, $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$ of the variational parameters impacts the ELBO and the number of iterations to reach convergence. We set the concentration parameter equal to $\alpha = 5$, the truncation parameter equal to $T = 30$, the number of objects equal to $N = 50$ and the convergence threshold equal to $\epsilon = 0.01\,\%$. The set of hyperparameters $\{\alpha, \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{u}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}\}$ for the input of the CAVI algorithm is again assumed to be known. In the simulation we run a fixed number of $\ell = 1, \ldots, 80$ CAVI iterations and save the ELBO for each iteration, alongside an index indicating the iteration at which convergence was achieved. For the initialization of the variational parameters we choose eight different approaches that can be summarized as follows:

- "One cluster" means that we set $\phi_{n,t}^{(0)} = 1$ for a chosen t and all $n = 1, \ldots, N$, and zero else, i.e., we associate all observations n with one specific component t of the truncated model. Then we run the CAVI algorithm for each possible association $t \in \{1, \ldots, T\}$ and choose the result with the highest ELBO.

- For the initialization type "True", we use the true cluster indicators z_n (which we can access because we generated the observation $\boldsymbol{y}$ by applying the CRP), i.e., $\phi_{n,z_n}^{(0)} = 1$ for all $n = 1, \ldots, N$ and zero else. This initialization type serves as a reference.

- In the case of "DBSCAN", we use the hard cluster assignments of the output of the DBSCAN algorithm [50] to initialize the variational parameter $\boldsymbol{\phi}_{1:N}^{(0)}$. Note that this algorithm requires two input parameters, which are the neighborhood radius and the minimum number of neighbors. We set the neighborhood radius to 0.3 and the number of neighbors to three.

- "Unique" means we uniquely associate each observation $\boldsymbol{y}_n$ with a component t of the truncated model, i.e., $\phi_{n,t}^{(0)} = 1$ for $n = t$ and zero else. Note that compared to the other initialization types, where we choose $T = 30$, we here choose $T = N$. The reason is we want to uniquely assign each observation n to a component t of the truncated model.

- The initialization type "Uniform" corresponds to uniform soft cluster assignment, i.e., $\phi_{n,t}^{(0)} = 1/T$ for all $n = 1, \ldots, N$ and $t = 1, \ldots, T$.

- In the case of "KMeans", we use the hard cluster assignments of the output of the KMeans algorithm to initialize the variational parameter $\boldsymbol{\phi}_{1:N}^{(0)}$. We choose 12 clusters (equal to the expected number of clusters $\bar{L}$) for the input of the KMeans algorithm, with randomly selected observations $\boldsymbol{y}_n$ serving as the initial centroids.

- For the initialization type "Random" we generate ten different hard cluster assignments from a discrete uniform distribution for the initialization of $\boldsymbol{\phi}_{1:N}^{(0)}$ and run the CAVI algorithm for reach random generation. Then, we check the ELBO at convergence for each CAVI run and choose the output with the highest ELBO.

- "Global" means we initialize the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$ with the knowledge of the hyperparameters α and λ, i.e., $\gamma_{t,1}^{(0)} = 1$, $\gamma_{t,2}^{(0)} = \alpha$, $\boldsymbol{\tau}_{t,1}^{(\ell)} = \boldsymbol{\lambda}_1$ and $\tau_{t,2}^{(\ell)} = \lambda_2$, for all $t = 1, \ldots, T$ (cf. (5.53) and (5.54)).

For all initialization types except "Global" we initialize the local variational parameters $\boldsymbol{\phi}_{1:N}^{(0)}$ and use (5.53) and (5.54) to obtain initial values for the global variational parameters $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$. Recall that $\phi_{n,t}^{(0)}$ represents a soft assignment, i.e., $\phi_{n,t}^{(0)}$ is the approximate posterior probability of the event $z_n = t$ (see the discussion of (4.60)). For the initialization type "Global" we reverse this procedure and first set $\boldsymbol{\gamma}_{1:(T-1)}^{(0)}$ and $\boldsymbol{\tau}_{1:T}^{(0)}$, and then compute $\boldsymbol{\phi}_{1:N}^{(0)}$ by using (5.52). For each initialization type we average the resulting ELBO and iteration of convergence over 500 simulation runs and determine the 95 % confidence interval. The results are shown in Figure 5.4.

We observe that ELBO approaches the same maximum for each initialization type in our simulation scenario, but varies in terms of the speed of convergence. The fastest convergence is obtained by the "True" initialization type, where the CAVI algorithm converges after 10 iterations on average. The slowest convergence is obtained by the "Uniform" and "One Cluster" initialization types. Here, the CAVI algorithm converges after approximately 23 iterations on average. For the "Global" initialization type, the CAVI algorithm converges after 19 iterations on average. Similar convergence behavior is obtained by the initialization types "DBSCAN",

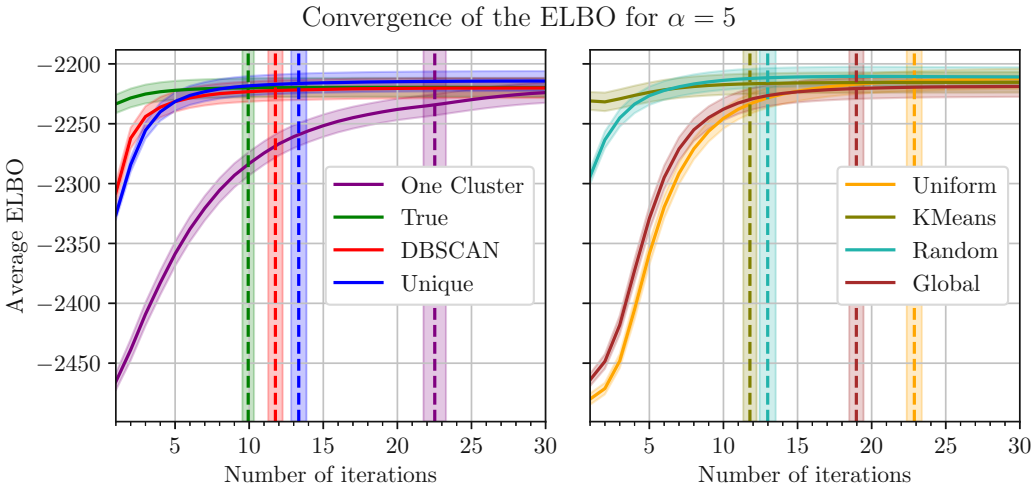


Figure 5.4: ELBO (solid lines) and iteration of convergence (dashed lines) averaged over 500 simulation runs for eight different types of initialization. The plot includes the 95 % confidence interval (shaded regions).

"Unique", "KMeans" and "Random", where the average iterations needed for convergence are 12, 13, 12 and 13, respectively. Note that using the initialization types "One Cluster" and "Random" needs the most computational effort, since we run the CAVI algorithm multiple times and take the best run. The initialization types using hard clustering algorithms, i.e., "KMeans" and "DBSCAN", deliver good performance results in terms of the ELBO, but come with additional overhead as additional parameters are needed which may have to be adapted for varying model parameters. The "True" initialization type is only included for reference since it assumes the indicator variables z_n to be known. Using "Global" and "Uniform" results in the second and third-slowest convergence speed after "One Cluster".

We conclude that using the "Unique" initialization type for our simulation scenario gives a good trade-off in terms of convergence speed and computational effort. On one hand, the computational costs for running CAVI with the initialization type "Unique" are higher compared to "True", "Global", "KMeans" and "DBSCAN" due to $T = N$. On the other hand, the computational costs are lower than for "One Cluster" and "Random" since we do not have to run the CAVI algorithm multiple times. For higher values of N (and correspondingly large T when using "Unique") we can resort to one of the other initialization types, where the truncation parameter T can be set arbitrarily and that deliver a similar performance in terms of the ELBO.

5.5.4 Estimation of the Object Features

Finally, we show simulation results for the estimation of the objects features $\boldsymbol{x}_n$, $n = 1, \ldots, N$, given observations $\boldsymbol{y}_n$, $n = 1, \ldots, N$, and compare them to the results obtained in [8]. Note that we do not assume to know the clustering structure of $\boldsymbol{y}$. The input to the CAVI algorithm (see Algorithm 3) only consists of the "raw" observations $\boldsymbol{y}$, the truncation level T, and the set of hyperparameters $\{\alpha, \boldsymbol{\mu}_{\theta^*}, \boldsymbol{\Sigma}_{\theta^*}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_v\}$. The parameters used for generating the observation $\boldsymbol{y}$ are given by Table 5.1 and the hyperparameters for the CAVI input are still assumed to be known, i.e., equal to the values in Table 5.1. For the initialization type we use "Unique" as explained above, which means the truncation parameter T is equal to the number of objects

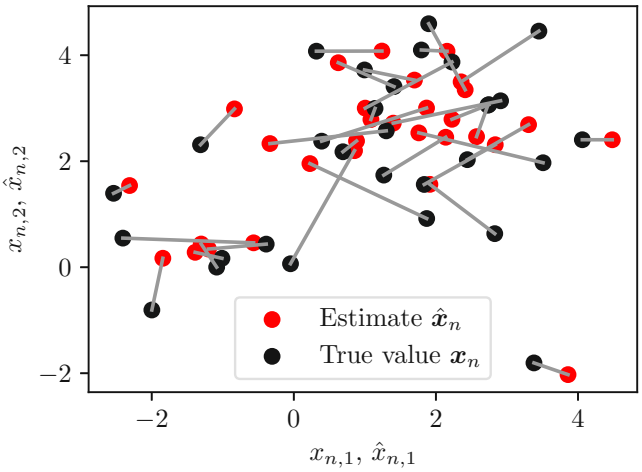


Figure 5.5: True object features $\boldsymbol{x}_n$ (black dots) compared to the estimates $\hat{\boldsymbol{x}}_n$ (red dots). The gray lines correspond to the estimation errors $\boldsymbol{e}_n = \hat{\boldsymbol{x}}_n - \boldsymbol{x}_n$.

N. This is also the initialization type used for the initialization $z_n^{(0)}$ of the indicator variables in the Gibbs sampler (see Section 5.2.2). The convergence threshold is set to $\epsilon = 0.001\,\%$.

In order to quantify the performance of CAVI and the approximate MMSE estimator (5.61) for the object features $\boldsymbol{x}_n$, we calculate the empirical MSE by averaging the squared estimation error over all objects $n = 1, \ldots, N$ and multiple simulation runs $j = 1, \ldots, J$, i.e.,

$$\mathrm{MSE} \triangleq \frac{1}{JNM} \sum_{j=1}^{J} \sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_{n,j} - \boldsymbol{x}_{n,j}\|^2 . \tag{5.62}$$

Here, $\boldsymbol{x}_{n,j}$ corresponds to the true feature of the n-th object in the j-th simulation run and $\hat{\boldsymbol{x}}_{n,j} = \hat{\boldsymbol{x}}_n(\boldsymbol{y}^{(j)})$ is the respective estimate. The estimate $\hat{\boldsymbol{x}}_{n,j}$ is obtained by using the variational parameters from the CAVI output (see Algorithm 3) in the approximate MMSE/MAP estimators (5.57), (5.58), (5.60) and (5.61), which depend on the observation $\boldsymbol{y}^{(j)}$ of the j-th simulation run. Figure 5.5 illustrates the result of a single simulation run in the case of $\alpha = 0.5$ and $N = 50$. It shows the true object features $\boldsymbol{x}_n$ and the estimates $\hat{\boldsymbol{x}}_n$ in conjunction with the estimation error $\boldsymbol{e}_n = \hat{\boldsymbol{x}}_n - \boldsymbol{x}_n$. The MSE of a single simulation run $J = 1$ is obtained by adding the squared length of the gray lines, i.e., the squared error, and dividing it by the number of objects N and the dimension M. For the following results we additionally average the MSE over $J = 1000$ simulation runs (see (5.62)).

Figure 5.6 presents the empirical MSE achieved by using CAVI, the two theoretical performance bounds (5.22) and (5.24), and the empirical MSE achieved by using Gibbs sampling as described in Section 5.2.2 with $Q = 1000$ samples (kindly provided by the author of [8]). Inserting $M = 2$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = 5\mathbf{I}_2$, $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \mathbf{I}_2$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \mathbf{I}_2$ (see Table 5.1) into (5.22) yields

$$\mathrm{MSE}_{\min} = \frac{1}{2} \operatorname{tr}(6\mathbf{I}_2(7\mathbf{I}_2)^{-1}) = \frac{6}{7} \tag{5.63}$$

for the MMSE estimator (5.21) which assumes no underlying clustering structure in the observation $\boldsymbol{y}$, i.e., where the local parameters $\boldsymbol{\theta}_n$ are assumed to be i.i.d. and individually distributed according to (5.17). It serves as a benchmark which indicates if it is worthwhile using the more

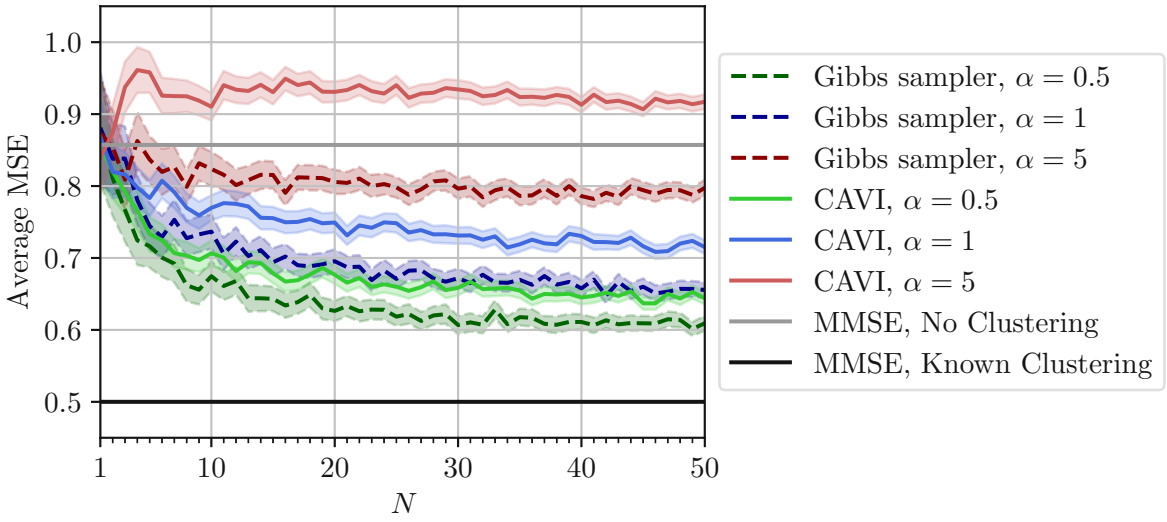


Figure 5.6: Average MSE achieved by CAVI compared to the benchmarks discussed in Section 5.2. The results are shown for three different values of the concentration parameter α. Shaded regions visualize the 95 % confidence interval.

computationally involved CAVI algorithm. Inserting $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \sigma_u^2 \mathbf{I}_M$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sigma_v^2 \mathbf{I}_M$ into (5.24) yields

$$\mathrm{MSE}_{\mathrm{min}} = \frac{1}{2} \operatorname{tr}((2\mathbf{I}_2)^{-1}) = \frac{1}{2} \tag{5.64}$$

for the MMSE estimator (5.23) where the underlying clustering structure in the observation $\boldsymbol{y}$ is known, i.e., where the local parameters $\boldsymbol{\theta}_n$ are assumed to be known. It serves as a benchmark which indicates how well we could perform in terms of MSE when we know the clustering structure. Note that, in line with (5.22) and (5.24), both MSEs (5.63) and (5.64) do not depend on the number of objects N. In the case of CAVI and the Gibbs sampler, the MSE depends on N and the concentration parameter α.

We observe that in terms of the number of objects N and the concentration parameter α, the overall MSE results of the CAVI algorithm are similar to the results of the Gibbs sampler. In the case of $\alpha = 5$, the CAVI algorithm is not able to perform better than the closed form MMSE estimator (5.21) using no clustering, as the clustering structure can not be estimated with sufficiently high accuracy by the approximate MMSE/MAP estimators. Recall that the higher the value of α is, the more small and densely packed clusters are within the observation $\boldsymbol{y}$ and the more difficult it is for the CAVI algorithm to estimate the clustering structure (see Figure 5.3). In the case of $\alpha = 0.5$ and $\alpha = 1$, the CAVI algorithm performs better than the closed form MMSE estimator (5.21) using no clustering, similar to the results of the Gibbs sampler. Here, the empirical MSE decreases as the number of objects N increases. This is due to the fact that more observations $\boldsymbol{y}_n$ are available to the CAVI algorithm and the approximate MMSE/MAP estimators. Consequently, more observations can be associated with the same cluster, i.e., share the same cluster parameters, which results in an improved estimation accuracy. The concentration parameter α determines how strong the MSE decreases with N. For a small value of α the clusters are more concentrated, i.e., there are few distinct clusters among the observations, and thus it is easier to estimate the clustering structure. This is beneficial for the

estimation of the object features $\boldsymbol{x}_n$, since the approximate MMSE estimator (5.61) relies on an accurate estimate of the underlying clustering structure represented by the estimates $\hat{\boldsymbol{\theta}}_n$ of the local parameters.

The improved MSE performance compared to the MMSE estimator (5.21) assuming no underlying clustering structure in the observation $\boldsymbol{y}$ can be captured with a performance metric referred to as the clustering gain (CG) [8], [22]. The CG compares the MSE of an estimator using no clustering, denoted as $\mathrm{MSE}^{(\text{no clustering})}$, to the MSE of an estimator that takes account for the underlying clustering structure in the observations, denoted as $\mathrm{MSE}^{(\text{clustering})}$. It is defined as

$$\mathrm{CG} \triangleq 10 \log_{10}\left(\frac{\mathrm{MSE}^{(\text{no clustering})}}{\mathrm{MSE}^{(\text{clustering})}}\right). \tag{5.65}$$

Table 5.2 shows the CG obtained by using CAVI and Gibbs sampling for $N = 50$ observations. According to (5.63) we have $\mathrm{MSE}^{(\text{no clustering})} = 6/7$. The MSE values for the case of CAVI

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 5$
CG achieved by CAVI	1.243 dB	0.787 dB	−0.294 dB
CG achieved Gibbs sampler	1.483 dB	1.164 dB	0.313 dB

Table 5.2: Clustering gain for $N = 50$ objects.

and Gibbs sampling ($\mathrm{MSE}^{(\text{clustering})}$) can be read from Figure 5.6. The CG is then obtained by inserting the MSE values into (5.65). Comparing the CG obtained by CAVI to the CG obtained by Gibbs sampling, we obtain the following results. For $\alpha = 0.5$ the CG of CAVI is 16.2 % less than that achieved by Gibbs sampling, for $\alpha = 1$ the CG of CAVI is 32.4 % less than that achieved by Gibbs sampling and for $\alpha = 5$ the CG of CAVI is 193.93 % less than that achieved by Gibbs sampling. We conclude that all MSE and CG results obtained by the CAVI algorithm are lower than the results obtained by Gibbs sampling, because the MSE obtained by CAVI is bigger than the MSE obtained by Gibbs sampling for all $N = 1, \ldots, 50$ (see Figure 5.6). This was to be expected, since VI methods are known to suffer from oversimplified posterior approximations compared to MC sampling methods. However, the loss in CG in the case of small α is not too high considering that we save in the required computational effort by using the CAVI algorithm.

The runtime on our simulation machine (a mid-range laptop) of our implementation [53] of Algorithm 3 for the above scenario is shown in Figure 5.7 as function of N. It includes the average runtime, the minimum runtime and the maximum runtime of $J = 1000$ simulation runs for each $N = 1, \ldots, 50$. For $N = 50$, the average is approximately 20 ms with a 95 % confidence interval of ± 0.62 ms. Unfortunately, we are not aware of the runtime of the Gibbs sampler implementation used in [8]. Nonetheless, considering that VI methods are known to be more computationally efficient than MCMC sampling methods [6], [7] we expect the runtime of the CAVI algorithm to be significantly less than that of the Gibbs sampler for the aforementioned scenario. In [54], the runtime of the CAVI algorithm in a simple simulation was reported to be 300 times faster than that of the Gibbs sampler.

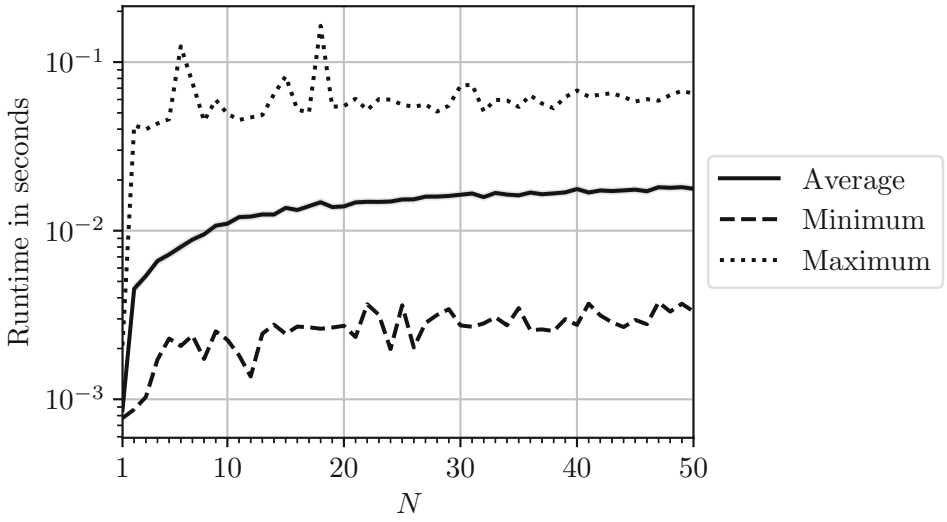


Figure 5.7: Average, maximum and minimum runtime in seconds on our simulation machine of our implementation of the CAVI algorithm (Algorithm 3). The 95 % confidence interval is within ±0.62 ms for all $N = 1, \ldots, 50$.

Up until now, we have assumed that the set of hyperparameters that we use for the input of the CAVI algorithm is equal to the parameters of the model responsible for generating the observations, i.e., that we perfectly know the values of the hyperparameters. As a final simulation result we want to demonstrate the effect of deviating from the true values (see Table 5.1) in the case of $\alpha = 0.5$. We denote the hyperparameters that we choose as an input for the CAVI algorithm as $\{\tilde{\alpha}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}^*}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}^*}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}\}$. For the mean $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}^*}$ of the base distribution we use the median of the observations $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{y}_N^{\mathrm{T}})^{\mathrm{T}}$, i.e.,

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}^*} = \mathrm{median}(\boldsymbol{y}),$$

where $\mathrm{median}(\boldsymbol{y})$ returns the middle value separating the greater and lesser halves of the observations for each of the two dimensions. For the covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}^*}$ of the base distribution we compute the empirical variance of the observations in each dimension and take the maximum, i.e.,

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}^*} = \left(\max_j \frac{1}{N} \sum_{n=1}^{N} (y_{n,j} - \bar{y}_j)^2 \right) \mathbf{I}_2,$$

where

$$\bar{y}_j = \frac{1}{N} \sum_{n=1}^{N} y_{n,j}. \tag{5.66}$$

In terms of the concentration parameter $\tilde{\alpha}$, we modify the true value $\alpha = 0.5$ with a multiplicative factor. We will evaluate $\tilde{\alpha} = 100\alpha = 50$ and $\tilde{\alpha} = \alpha/100 = 0.005$. For the covariance matrix of the parameter noise $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}^*}$ and the covariance matrix of the observation noise $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$, we choose the values $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} = \frac{1}{2}\boldsymbol{\Sigma}_{\boldsymbol{u}} = \frac{1}{2}\mathbf{I}_2$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} = 2\boldsymbol{\Sigma}_{\boldsymbol{u}} = 2\mathbf{I}_2$. We then run the same simulation as for Figure 5.6, i.e., the true hyperparameters of the model responsible for

generating the observation $\boldsymbol{y}$ is given by Table 5.1.

Figure 5.8 shows the simulation results. We observe that a precise knowledge of the concen-

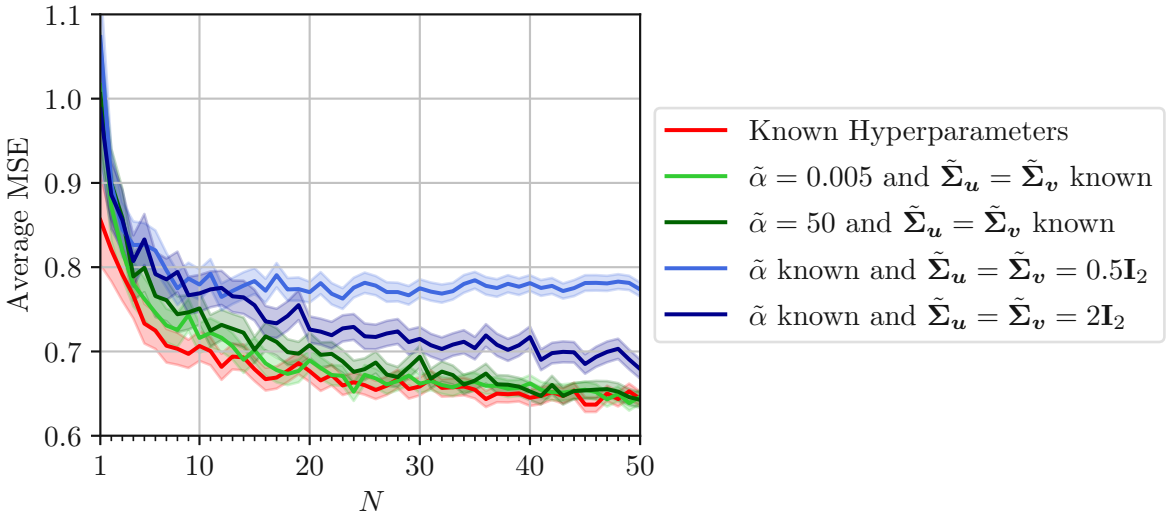


Figure 5.8: Average MSE achieved by CAVI in the case of unknown hyperparameters compared to the case of known hyperparameters. Shaded regions visualize the 95 % confidence interval.

tration parameter is not crucial, since the results in terms of MSE are similar to the case where $\tilde{\alpha}$ is known. Deviating from the true values of the covariance matrices $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$ has more impact on the performance results. Here, we observe a noticeable degradation in terms of the MSE. Note that we still perform better than the MMSE estimator that uses no clustering, i.e., all MSE results are below $6/7 \approx 0.86$ (see (5.63)) for a sufficiently high number of observations N.

6 Conclusions

In this thesis, we discussed VI, which is a methodology for approximating the posterior distribution for models — such as the DPM model — where the posterior distribution is intractable and thus one has to resort to approximate inference techniques. We presented a detailed derivation of the CAVI algorithm for DPM models with EF component distributions and specialized the CAVI algorithm to the Gaussian estimation problem considered in [8]. We also presented simulation results for the proposed CAVI algorithm and compared them to the results from [8], which were obtained using an MCMC method known as the Gibbs sampler.

We first provided an introduction to BMMs and the EF, and subsequently we discussed the DP and its clustering properties. This led to the definition of a mixture model with an infinite number of components, specifically a DPM model. Moreover, we defined a DPM model with component distributions from the EF and with a conjugate prior, using the stick-breaking representation of the DP. Based on [7], we then presented a detailed derivation of the CAVI algorithm for that model. This algorithm relies on a truncated mean field approximation of the posterior pdf and entails truncating the stick-breaking representation of the DP. We also addressed practical considerations in the application of CAVI-based methods to DPM models.

Furthermore, we adapted the CAVI algorithm for DPM models with component distributions from the EF to the case of DPM models with Gaussian component distributions, and performed simulations for the Gaussian estimation problem proposed in [8]. In this Gaussian estimation problem, the object features are modeled by a DPM of Gaussians, i.e., a mixture model with an infinite number of Gaussian components, and the goal is to estimate the object features from noisy observations. In view of the high computational complexity of the MCMC method used in [8], we developed a more efficient method based on the CAVI algorithm for a DPM of Gaussians. A drawback of the CAVI algorithm is that it can lead to oversimplified posterior approximations, and thus we performed simulations to assess the estimation accuracy of our method and compare it to that of the method used in [8] as well as to two theoretical performance bounds established in [8]. Each of these theoretical performance bounds is based on a simplifying model assumption: on one hand, that the object features do not possess a clustering structure (in contrast to the DP prior), and on the other hand that the clustering structure is known.

Our simulation results demonstrated that the CAVI method can achieve a clustering gain, corresponding to an improved performance compared to the performance bound assuming no clustering. Moreover, we found that the difference between the clustering gains achieved by our CAVI method and the Gibbs sampler of [8] depends on the concentration parameter α of the DP. For high α (i.e., low concentration), the CAVI algorithm struggles to estimate the cluster structure underlying the observations with sufficient accuracy and performs less well than the Gibbs sampler in [8]. This is the price paid for the significantly better computational efficiency of the CAVI algorithm. In this regard, we showed that our implementation of the

CAVI algorithm for a DPM of Gaussians [53] has a runtime of approximately 20 ms when using the simulation parameters listed in Table 5.1. Unfortunately, we are not aware of the runtime of the Gibbs sampler implementation used in [8], but considering that VI methods are known to be more computationally efficient than MCMC methods [6], [7], we expect the runtime of the CAVI algorithm to be significantly less than that of the Gibbs sampler.

Finally, we presented simulation results demonstrating the effects of deviating in the implementation of the CAVI algorithm from the true model hyperparameters used to generate the observations. We observed that accurate knowledge of the concentration parameter of the DPM model is less critical than accurate knowledge of the covariance matrix of the noise model.

Given the recent and ongoing evolution of the VI methodology, there is scope for further research aiming to develop VI methods for the considered estimation problem with increased clustering gain (estimation accuracy) and computational efficiency. Possible research directions include "accurate VI," which employs variational models beyond the mean field approximation, and "amortized VI," which jointly predicts local parameters with a parameterized function of the observations instead of optimizing a local variational parameter for each observation [14]. Furthermore, there is potential for exploring different model assumptions and model parameters and their impact on the clustering gain. For instance, one could consider treating the parameters of the DP as random variables with a specified prior distribution and estimating them from the observed data. Another strategy could be to collapse the model as demonstrated in [17], thereby eliminating parameters that are not essential for estimating the object features and potentially improving the estimation.

Bibliography

[1] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, NY, USA, 2013.

[2] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, London, UK, 2023.

[3] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, NJ, USA, 1993.

[4] Y. W. Teh, "Dirichlet Process," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., Springer, MA, USA, 2010, pp. 280–287.

[5] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An Introduction to MCMC for Machine Learning," *Machine Learning*, vol. 50, no. 1, pp. 5–43, 2003.

[6] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2018.

[7] D. M. Blei and M. I. Jordan, "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[8] E. Šauša, "Advanced Bayesian Estimation in Hierarchical Gaussian Models: Dirichlet Process Mixtures and Clustering Gain," Master's thesis, TU Wien, 2024.

[9] S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert, Eds., *Handbook of Mixture Analysis*. Chapman and Hall/CRC, FL, USA, 2018.

[10] S. Ghosal and A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, Cambridge, UK, 2017.

[11] J. Sethuraman, "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[12] D. Blackwell and J. B. MacQueen, "Ferguson Distributions Via Polya Urn Schemes," *The Annals of Statistics*, vol. 1, no. 2, pp. 353–355, 1973.

[13] J. Pitman, *Combinatorial Stochastic Processes*, J. Picard, Ed. Springer, Berlin, Germany, 2006.

[14] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in Variational Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 2008–2026, 2019.

[15] K. Kurihara, M. Welling, and N. Vlassis, "Accelerated Variational Dirichlet Process Mixtures," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19, MIT Press, 2006.

[16] K.-L. Lim, "Variational Inference of Dirichlet Process Mixture using Stochastic Gradient Ascent," in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, 2020, pp. 33–42.

[17] K. Kurihara, M. Welling, and Y. W. Teh, "Collapsed Variational Dirichlet Process Mixture Models," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, Morgan Kaufmann Publishers, 2007, pp. 2796–2801.

[18] C. Wang and D. M. Blei, "Truncation-free Stochastic Variational Inference for Bayesian Nonparametric Models," in *Neural Information Processing Systems*, Curran Associates, 2012, pp. 413–421.

[19] M. C. Hughes and E. B. Sudderth, "Memoized Online Variational Inference for Dirichlet Process Mixture Models," in *Advances in Neural Information Processing Systems*, C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26, Curran Associates, 2013.

[20] V. Huynh, D. Q. Phung, and S. Venkatesh, "Streaming Variational Inference for Dirichlet Process Mixtures," in *Asian Conference on Machine Learning*, vol. 45, PMLR, 2016, pp. 237–252.

[21] D. Lin, "Online Learning of Nonparametric Mixture Models via Sequential Variational Approximation," in *Advances in Neural Information Processing Systems*, C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26, Curran Associates, 2013.

[22] B. Kreidl, "Bayesian Nonparametric Inference in State-Space Models with an Application to Extended Target Tracking," Master's thesis, TU Wien, 2021.

[23] R. M. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[24] P. J. Green and S. Richardson, "Modelling Heterogeneity with and without the Dirichlet Process," *Scandinavian Journal of Statistics*, vol. 28, pp. 355–375, 2 2001.

[25] H. Ishwaran and L. F. James, "Some Further Developments for Stick-Breaking Priors: Finite and Infinite Clustering and Classification," *Sankhyā: The Indian Journal of Statistics (2003-2007)*, vol. 65, no. 3, pp. 577–592, 2003.

[26] S. Jain and R. M. Neal, "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, vol. 13, no. 1, pp. 158–182, 2004.

[27] S. G. Walker, "Sampling the Dirichlet Mixture Model with Slices," *Communications in Statistics - Simulation and Computation*, vol. 36, no. 1, pp. 45–54, 2006.

[28] I. Porteous, A. Ihler, P. Smyth, and M. Welling, "Gibbs Sampling for (Coupled) Infinite Mixture Models in the Stick Breaking Representation," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 2006, pp. 385–392.

[29] T. J. Bucco, "Extended Multi-target Tracking Using Probabilistic Data Association and Bayesian Nonparametric Inference," Master's thesis, TU Wien, 2020.

[30] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*. Springer, NY, USA, 2006, p. 494.

[31] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley, NY, USA, 2000.

[32] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut" — Interactive Foreground Extraction using Iterated Graph Cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[34] S. Rattanasiri, D. Böhning, P. Rojanavipart, and S. Athipanyakom, "A Mixture Model Application in Disease Mapping of Malaria," *Southeast Asian Journal of Tropical Medicine and Public Health*, vol. 35, pp. 38–47, 2004.

[35] W. J. Ewens, "Population Genetics Theory — The Past and the Future," in *Mathematical and Statistical Developments of Evolutionary Theory*, S. Lessard, Ed. Springer, 1990, pp. 177–227.

[36] S. Frühwirth-Schnatter, G. Malsiner-Walli, and B. Grün, "Generalized Mixtures of Finite Mixtures and Telescoping Sampling," *Bayesian Analysis*, vol. 16, no. 4, pp. 1279–1307, 2021.

[37] J. W. Miller and M. T. Harrison, "Mixture Models With a Prior on the Number of Components," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 340–356, 2017.

[38] F. Nielsen and V. Garcia, "Statistical exponential families: A digest with flash cards," *arXiv*, 2011. arXiv: `0911.4863`.

[39] F. Hlawatsch, *Bayesian Machine Learning*, Lecture notes, Course 389.207, Institute of Telecommunications, TU Wien, 2022.

[40] C. P. Robert, *The Bayesian Choice*. Springer, NY, USA, 2007, p. 606.

[41] J.-M. Marin, K. Mengersen, and C. P. Robert, "Bayesian Modelling and Inference on Mixtures of Distributions," in *Bayesian Thinking*, D. K. Dey and C. R. Rao, Eds., vol. 25, Elsevier, 2005, pp. 459–507.

[42] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, NY, USA, 2006.

[43] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet Distribution and Related Processes," Department of Electrical Engineering, University of Washington, Tech. Rep. UWEETR-2010-0006, 2010.

[44] T. S. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[45] H. Ishwaran and L. F. James, "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

[46] C. E. Rasmussen, "The Infinite Gaussian Mixture Model," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12, MIT Press, 1999, pp. 554–560.

[47] E. B. Sudderth, "Graphical Models for Visual Object Recognition and Tracking," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[48] B.-E. Chérief-Abdellatif, "Consistency of ELBO maximization for model selection," in *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, F. Ruiz, C. Zhang, D. Liang, and T. Bui, Eds., vol. 96, PMLR, 2019, pp. 11–31.

[49] H. Tamae, K. Irie, and T. Kubokawa, "A score-adjusted approach to closed-form estimators for the gamma and beta distributions," *Japanese Journal of Statistics and Data Science*, vol. 3, pp. 543–561, 2020.

[50] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC, FL, USA, 2013.

[51] B.-E. Chérief-Abdellatif and P. Alquier, "Consistency of Variational Bayes Inference for Estimation and Model Selection in Mixtures," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 2995–3035, 2018.

[52] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley, NJ, USA, 2012.

[53] T. Lipovec and W. Wiedner, *vi-gaussian-dpm*, GitHub repository, 2023. [Online]. Available: `https://github.com/lipovec-t/vi-gaussian-dpm` (visited on 11/21/2023).

[54] M.-N. Tran, T.-N. Nguyen, and V.-H. Dao, "A practical tutorial on Variational Bayes," in *arXiv*, 2021. arXiv: `2103.01327`.