institute of telecommunications

TU WIEN

MASTER'S THESIS

Variational Inference for Bayesian Mixture Models with a Random Number of Components

for obtaining the academic degree

Diplom-Ingenieur

in the Masters's degree program

Telecommunications

submitted by

Wilfried Wiedner

matriculation number: 00073804

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

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

Vienna, January 3, 2024

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, January 3, 2024

Wilfried Wiedner

Abstract

Finite mixture models, i.e., mixture models with a fixed number of components, have a long tradition in statistical modeling and are a well-established tool to explore structures in complex data. In scenarios where the number of components is unknown, choosing an appropriate number of components is a crucial modeling decision that can be challenging. In order to formalize this modeling decision in a Bayesian fashion, we investigate the *mixture of finite mixtures* (MFM) model. The MFM model extends the traditional finite mixture model to a Bayesian mixture model with a random number of components. Using the MFM model, it is possible to group data into meaningful subpopulations and estimate the model parameters without specifying the number of components a priori. We discuss equivalent representations of the MFM model such as the stick-breaking representation and relevant distributions such as the exchangeable partition probability function. For Bayesian inference of the model parameters, we propose a computationally efficient coordinate-ascent variational inference (CAVI) algorithm for MFM models and provide detailed derivations of the corresponding update equations. Subsequently, we focus on mixtures consisting of multivariate Gaussian component distributions with unknown means and known covariance matrices, resulting in a novel CAVI algorithm for the static mixture of finite Gaussian mixtures (MFGM) model. We evaluate the clustering performance of our CAVI algorithm using synthetic data generated according to a finite Gaussian mixture and observe high accuracy for suitably chosen hyperparameters. Furthermore, we apply an existing CAVI algorithm for Dirichlet process mixture (DPM) models, which are frequently used in scenarios with an unknown (but finite) number of components, to our data. A comparison reveals that the proposed CAVI algorithm for static MFGMs outperforms the CAVI algorithm for DPMs, especially for large datasets.

Contents

List of Abbreviations

MFM	mixture of finite mixtures
MFGM	mixture of finite Gaussian mixtures
BMM	Bayesian mixture model
DPM	Dirichlet process mixture
VB	variational Bayes
VI	variational inference
CAVI	coordinate-ascent variational inference
KLD	Kullback-Leibler divergence
ELBO	evidence lower bound
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
RJMCMC	reversible jump Markov chain Monte Carlo
EPPF	exchangeable partition probability function
pdf	probability density function
pmf	probability mass function
cdf	cumulative distribution function
i.i.d.	independent and identically distributed
MMSE	minimum mean square error
MAP	maximum a-posteriori

Chapter 1

Introduction

1.1 Background

Mixture models, i.e., models based on a linear combination of distributions, find applications in various fields such as model-based clustering and classification [1], density estimation [2], image and signal processing [3], biology and bioinformatics [4], and text mining and natural language processing [5]. Mixture models posit that the observed data arise from a mixture of several subpopulations commonly called *components*, each characterized by its own probability distribution. Traditionally, mixture models have assumed that the number of components is fixed. The inception of finite mixture modeling, a landmark in statistical modeling, dates back nearly 130 years to a first significant publication where the number of components was fixed to two [6]. However, it soon became clear that choosing an appropriate number of components is a critical modeling decision. The frequentist statistical viewpoint does not allow to infer the number of components jointly with the model parameters [7] and requires a separate model selection task using heuristic algorithms.

Bayesian mixture models (BMMs) extend the traditional concept of a mixture model by incorporating Bayesian principles, i.e., the parameters of the mixture are modeled as random unknown quantities that are distributed according to known prior distributions, and Bayes' theorem is used to update beliefs about these parameters according to observed data. This allows for the incorporation of prior knowledge and uncertainty into the modeling process. Within the Bayesian framework, a natural way to deal with an unknown number of components is to treat it as a random parameter with a (discrete) prior. In the literature, this type of BMM is commonly referred to as the *mixture of finite mixtures* (MFM) model. The MFM model was introduced in [8] and further studied in [9]–[11].

Unfortunately, practical statistical inference, in particular inference about the number of components in the MFM model, is difficult analytically and computationally. A notable advance was achieved in [9], where the reversible jump Markov chain Monte Carlo (RJMCMC) method [12] was used for the MFM model. This method enables sampling all the model parameters including the number of components from the joint posterior distribution, i.e., the joint distribution of the model parameters given the observed data. However, the RJMCMC method is difficult to use since designing good reversible jump moves is often nontrivial, especially in high-dimensional parameter spaces [13].

Due to these challenges, infinite mixture models, i.e., mixture models with an infinite number of components such as the *Dirichlet process mixture* (DPM) model, have become popular. Since the Dirichlet process has several computationally tractable representations, such as a representation using the partition distribution — which is a member of the family of Gibbs partition distributions, cf. Section 3.2 — and the stick-breaking representation, a variety of efficient inference algorithms for DPM models have been proposed [14]–[16]. For DPMs, the expected number of clusters, i.e., the number of components used in the process of generating the observed data, grows logarithmically with the number of observations [17]. Thus, when the underlying distribution of the observed data has a fixed and finite number of components, the usefulness of DPM models for clustering is questionable. Indeed, in [18], the authors proved inconsistency of DPM models when estimating the number of clusters in a dataset distributed according to a simple univariate mixture with a fixed and finite number of Gaussian components. More specifically, it was shown in [18] that when a DPM model is applied to data from a finite mixture, the posterior distribution of the number of clusters does not concentrate at the true number of components.

1.2 State of the Art and Motivation

Despite the inconsistency issue with DPMs discussed in [18], inference algorithms for DPM models play a major role in state-of-the-art inference algorithms for MFM models [19]. It turned out that for a special case of MFM models (called static MFMs, see Section 3.1), there are MFM counterparts for all the computationally tractable representations of the Dirichlet process mentioned above. Exploiting the fact that static MFM models have a partition distribution of Gibbs form as well, split-merge samplers for DPMs [14],[15] were adapted to static MFM models in [19]. Partitions are sampled from the corresponding posterior distribution, and thereby the number of clusters is inferred.

In [20], an algorithm of a somewhat different nature, called *telescope sampling*, was proposed. Unlike the split-merge sampler in [19], the telescope sampler does not require knowledge of the marginal likelihood, nor is it restricted to the static MFM model. Indeed, the telescope sampler is known as one of the most generic and easily implemented inference algorithms for MFM models with arbitrary component distributions [20].

All state-of-the-art inference algorithms for the MFM model use Monte Carlo sampling techniques — in particular, of the MCMC type — which are prone to be computationally expensive, especially in scenarios with high-dimensional parameter spaces or a large number of observations. This fact motivates the main contribution of this thesis, which is a computationally efficient approximate inference algorithm for the static MFM model based on variational inference (VI).

1.3 Thesis Outline

The rest of this thesis is organized as follows. In Chapter 2, we introduce the (finite) BMM. We also present various equivalent representations and address the challenge of estimating the model parameters in a BMM. In the first part of Chapter 3, we present the generalized MFM model and two special cases, the static and dynamic MFM models. Next, we derive various relevant distributions such as the exchangeable partition probability function (EPPF) and the prior distribution of the number of clusters, and we discuss connections between static and dynamic MFM models as well as the DPM model. The second part of Chapter 3 is focused on the static MFM model and presents equivalent representations including the stick-breaking representation.

In Chapter 4, we derive a novel coordinate-ascent variational inference (CAVI) algorithm for the static MFM model using the stick-breaking representation. This algorithm is furthermore specialized to Gaussian component distributions with unknown means and known covariance matrices in Chapter 5. Subsequently, we evaluate the clustering performance of our algorithm using synthetic data from a finite Gaussian mixture and the well-known Old Faithful geyser dataset.

Finally, Chapter 6 provides a summary of the thesis, reviews the most important results of our numerical simulation, and suggests potential areas for future research.

1.4 Notation

The notation used in this thesis is summarized in Table 1.1. Note that our notation does not distinguish between a random quantity and its realization.

$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
x	scalar
$\boldsymbol{x}$	vector
$\boldsymbol{X}$	matrix
$\mathbf{1}_K$	all-one vector of length K
$\mathbf{I}_M$	identity matrix of size $M \times M$
$\cdot^{\mathrm{T}}$	vector/matrix transpose
$\cdot^{-1}$	inverse
$\det(\cdot)$	determinant of a matrix
$\mathrm{trace}(\cdot)$	trace of a matrix
$\mathbb{1}(\cdot)$	indicator function; equals one if argument is true and zero else
$\delta(\cdot)$	Dirac delta function
$\ln(\cdot)$	natural logarithm
$\Gamma(\cdot)$	gamma function
$\Psi(\cdot)$	digamma function
$f(\cdot)$	probability density function (pdf)
$f(\cdot\|\cdot)$	conditional pdf
$p(\cdot)$	probability mass function (pmf)
$p(\cdot\|\cdot)$	conditional pmf
$q(\cdot)$	variational approximation of a (conditional) pdf/pmf
$G(\cdot)$	discrete mixing distribution
$\mathrm{E}^{(f(\cdot))}\{\cdot\}$	expectation with respect to the pdf $f(\cdot)$
$\mathcal{D}(\cdot;\boldsymbol{\beta})$	Dirichlet distribution with parameter vector $\boldsymbol{\beta}$
$\mathcal{E}(\cdot;\alpha)$	exponential distribution with rate parameter α
$\mathcal{G}(\cdot;\widetilde{\alpha},\alpha)$	gamma distribution with shape parameter $\widetilde{\alpha}$ and rate parameter α
$\mathcal{C}(\cdot;\boldsymbol{\pi})$	categorical distribution with event probabilities $\boldsymbol{\pi}$
$\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$	multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$x \perp\!\!\!\perp y$	random variables x and y are independent
$x \perp\!\!\!\perp y \,\|\, z$	random variables x and y are conditionally independent given random variable z
$x \sim f(x)$	random variable x is distributed according to the pdf $f(x)$
$x_1, x_2 \overset{\text{i.i.d.}}{\sim} f(x)$	random variables x_1 and x_2 are independent and identically distributed according to the pdf $f(x)$

Table 1.1: Summary of notation.

Chapter 2

Bayesian Mixture Models

In this chapter, we introduce the (finite) Bayesian mixture model (BMM). Our aim is to familiarize the reader with the notation and terminology surrounding the theory of BMMs and to build a solid basis for the remainder of the thesis.

In the fields of statistical modeling and data analysis, the BMM is a powerful approach to understanding complex data distributions and uncovering latent (i.e., not observed) structures within them. The key concept of mixture models [21],[22] is that observations are assumed to originate from multiple underlying components with each component characterized by its own statistical distribution defined by parameters. The Bayesian variant of these models, i.e., BMM, draws its distinction by treating the component parameters as random variables. By specifying prior distributions over the component parameters, additional global parameters of the mixture model (e.g., the mixture weights), and even the number of components, existing knowledge or assumptions about the data can be incorporated in a Bayesian fashion. As new observations become available, the posterior distributions are updated, refining the information about the components and parameters of the model.

2.1 Basic Formulation

Let $\boldsymbol{x}_n \in \mathbb{R}^M$, for $n = 1, \dots, N$, be conditionally independent random vectors. The basic (finite) mixture model assumes that observations $\boldsymbol{x}_n$ arise from a density $f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*)$ defined as a convex combination of $K \in \mathbb{N}$ component distributions, each of specified parametric form $f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*)$, i.e.,

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

We assume that the functional form of the component distributions $f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*)$ is completely known, but for the parameters $\boldsymbol{\theta}_k^*$. The vector $\boldsymbol{\theta}^* = \left(\boldsymbol{\theta}_1^{*\mathrm{T}} \ \cdots \ \boldsymbol{\theta}_K^{*\mathrm{T}}\right)^{\mathrm{T}}$ contains the random component parameters $\boldsymbol{\theta}_k^* \in \mathbb{R}^p$, for $k = 1, \dots, K$, often referred to as *mixand parameters*. We assume the component parameters to be i.i.d. according to the prior density $f(\boldsymbol{\theta}_k^*)$. The vector $\boldsymbol{\pi}$ is composed of K random weights $\pi_k \in [0,1]$, which we call the *mixture weights*, and lies in the $(K-1)$-dimensional probability simplex

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

Because of the constraints on the mixture weights $\boldsymbol{\pi}$ specified by (2.2), it is guaranteed that (2.1) is a valid probability distribution as long as every component distribution is also a valid probability distribution. Mathematically, the conditional distribution $f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*)$ given by (2.1), called the *mixture distribution*, can be interpreted as a linear combination of its individual components $f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*)$, where the mixture weights π_k, for $k = 1, \dots, K$, are the (random) coefficients. By adjusting their respective values, one can control the influence of each component $f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*)$ on the mixture distribution $f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*)$. Therefore, we put a prior on the mixture weights, i.e., we assume that the random vector $\boldsymbol{\pi}$ is distributed according to the prior density $f(\boldsymbol{\pi})$ with support given by the simplex Δ_K defined in (2.2). Furthermore, statistical independence between $\boldsymbol{\theta}_k^*$ and $\boldsymbol{\pi}$ is assumed, i.e., $\boldsymbol{\theta}_k^* \perp\!\!\!\perp \boldsymbol{\pi}$ for all $k = 1, \dots, K$. The BMM for N observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, as described above, can be summarized as follows:

$$\boldsymbol{\pi} \sim f(\boldsymbol{\pi}), \tag{2.3a}$$

$$\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^* \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}_k^*), \tag{2.3b}$$

$$\boldsymbol{x}_1, \dots, \boldsymbol{x}_N | \boldsymbol{\pi}, \boldsymbol{\theta}^* \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*). \tag{2.3c}$$

2.2 Likelihood Function

In the case of N conditionally independent observations from (2.1), the joint distribution of the observations given $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*$ has the form

$$f(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}^*) = \prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*), \tag{2.4}$$

where $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ are arranged in the vector $\boldsymbol{x}$ according to $\boldsymbol{x} = (\boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$. The joint distribution given by (2.4) is referred to as the *likelihood function*. Due to its product-of-sums form, evaluating the likelihood function for the BMM can be computationally expensive; particularly when dealing with a large number of observations or components. A full expansion of the likelihood function involves a sum of K^N terms and thus may require an infeasible amount of computational resources to perform conventional Bayesian inference methods.

2.3 Latent Indicator Variables

We next provide an alternative representation of the BMM defined in (2.3). Suppose we are given N observations $\boldsymbol{x}_n$, independently drawn from the mixture distribution $f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*)$ given by (2.1). Furthermore, let $z_1, \ldots, z_N$ be discrete random variables, which are assumed to be conditionally independent given the mixture weights $\boldsymbol{\pi}$ and take on values from $\{1, \ldots, K\}$ with probability π_k, i.e., $P\{z_n = k\} = \pi_k$. Thus, each z_n given $\boldsymbol{\pi}$ is distributed according to the categorical distribution

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

We can characterize the generation procedure of an observation $\boldsymbol{x}_n$ from the mixture distribution $f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*)$ in two steps: first, z_n is drawn from (2.5) and secondly, given z_n, $\boldsymbol{x}_n$ is drawn from the component distribution corresponding to z_n, i.e.,

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n) = f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}). \tag{2.6}$$

Therefore, each observation $\boldsymbol{x}_n$ is assigned its own variable z_n indicating the component of the mixture distribution that is responsible for generating the observation. For this reason, we refer to $z_1, \ldots, z_N$ as the *indicator variables*. We assume that the component parameter $\boldsymbol{\theta}^*_k$ and the indicator variable z_n are conditionally independent given the mixture weights $\boldsymbol{\pi}$, i.e., $\boldsymbol{\theta}^*_k \perp\!\!\!\perp z_n \,|\, \boldsymbol{\pi}$ for all $k = 1, \ldots, K$ and $n = 1, \ldots, N$. Including the indicator variables $z_1, \ldots, z_N$ in the BMM given by (2.3) results in a hierarchical model, which is summarized by

$$\boldsymbol{\pi} \sim f(\boldsymbol{\pi}), \tag{2.7a}$$

$$z_1, \ldots, z_N|\boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} p(z_n|\boldsymbol{\pi}) = \mathcal{C}(z_n; \boldsymbol{\pi}), \tag{2.7b}$$

$$\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_K \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}^*_k), \tag{2.7c}$$

$$\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n \sim f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n) = f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}) \quad \text{independently for} \quad n = 1, \ldots, N, \tag{2.7d}$$

where we summarize the independence/conditional independence assumptions by

$$\boldsymbol{\theta}^*_k \perp\!\!\!\perp \boldsymbol{\pi} \quad \text{for all} \quad k = 1, \ldots, K, \tag{2.8a}$$

$$\boldsymbol{\theta}^*_k \perp\!\!\!\perp z_n \,|\, \boldsymbol{\pi} \quad \text{for all} \quad k = 1, \ldots, K \quad \text{and} \quad n = 1, \ldots, N, \tag{2.8b}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{\pi}, \boldsymbol{x}_{n'}, z_{n'}, \boldsymbol{\theta}^*_k \,|\, z_n, \boldsymbol{\theta}^*_{z_n} \quad \text{for all} \quad n' \neq n = 1, \ldots, N \quad \text{and} \quad k \neq z_n. \tag{2.8c}$$

A graphical model representation [23] of the BMM in (2.7) is shown in Figure 2.1. We note that the mixture weights $\boldsymbol{\pi}$ and the component parameters $\boldsymbol{\theta}^*$ are global parameter vectors influencing the entire observation vector $\boldsymbol{x}$, whereas the vector $\boldsymbol{z} = (z_1 \;\cdots\; z_N)^{\mathrm{T}}$ containing the

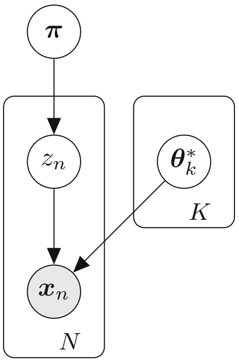


Figure 2.1: Bayesian network representing the BMM with indicator variables in (2.7). White nodes indicate the variables that are latent, while grey nodes indicate the ones that are observed. Plates indicate the repetition of nodes or groups of nodes.

indicator variables is a local parameter vector since its nth element z_n influences only the nth observation $\boldsymbol{x}_n$. It can be shown [24] that marginalizing out the indicator variables from the joint conditional distribution $f(\boldsymbol{x}_n, z_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*)$ yields

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

Thus, the hierarchical model in (2.7) again gives a mixture distribution in the form of (2.1), but with the indicator variable z_n replacing k as the component index.

Indicator variables are particularly useful when the goal is to group observations into meaningful subpopulations. This unsupervised learning task is commonly known as *clustering*. We emphasize the fundamental distinction between the (known) number of components K in the mixture distribution (2.1) and the number of clusters $L \leq K$ in the observations $\boldsymbol{x}$, i.e., the number of mixture components responsible for generating the observations $\boldsymbol{x}$. Using the indicator variables $\boldsymbol{z}$, or more practically, estimates of the indicator variables, the number N_k of observations $\boldsymbol{x}_n$ generated by component k of the mixture distribution is given by

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

With (2.10), the number of clusters L in the observations $\boldsymbol{x}$ given $N_1, \ldots, N_K$ and $\boldsymbol{z}$, i.e., the number of mixture components assigned at least one observation by $\boldsymbol{z}$, can be determined by

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

Furthermore, the number of observations N can be recovered from (2.10) via summation with respect to the component index k, i.e.,

$$N = \sum_{k=1}^{K} N_k. \tag{2.12}$$

Obviously, the number of clusters L is upper bounded by the number of components K of the mixture model. The number of clusters L may be smaller than K, due to, e.g., components associated with small mixture weights π_k, too few observations (i.e., $N < K$), or not well-separated components.

2.4 Latent Mixing Distribution

We now consider an important representation [25, Chapter 1] of the BMM in (2.3) where we can represent the unknown mixture weights $\boldsymbol{\pi}$ and component parameters $\boldsymbol{\theta}^*$ with a probability distribution $G(\cdot)$. Let $\boldsymbol{\theta}_n$, for $n = 1, \ldots, N$, be a random vector assigned to the nth observation $\boldsymbol{x}_n$. The $\boldsymbol{\theta}_n$ are random samples from a discrete probability distribution $G(\cdot)$ which places probability mass π_k at the support point $\boldsymbol{\theta}_k^*$, i.e.,

$$P\{\boldsymbol{\theta}_n = \boldsymbol{\theta}_k^*\} = G(\boldsymbol{\theta}_n) = \pi_k.$$

Therefore, the mixture weights $\boldsymbol{\pi}$ and the component parameters $\boldsymbol{\theta}_k^*$ are subsumed into the discrete probability distribution $G(\cdot)$ consisting of K points of support $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*$ and the corresponding probability masses π_k, for $k = 1, \ldots, K$, as follows:

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

In literature, (2.13) is referred to as the *mixing distribution*. Since $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_k^*$ are latent random parameters, the mixing distribution $G(\cdot)$ is latent and random as well. Thus, a realization of $G(\cdot)$ can be determined by sampling from the corresponding prior distributions $f(\boldsymbol{\pi})$ and $f(\boldsymbol{\theta}_k^*)$. Note that the conditional distribution of an observation $\boldsymbol{x}_n$ given a realization of the mixing distribution $G(\cdot)$ according to (2.13) equals the mixture distribution defined in (2.1) [25], i.e.,

$$f(\boldsymbol{x}_n|G) = f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*). \tag{2.14}$$

As in the previously considered representation (cf. (2.7)), including the mixing distribution $G(\cdot)$ in the BMM in (2.3) also leads to a hierarchical model. Considering N conditionally independent observations $\boldsymbol{x}_n$, it is summarized according to

$$\boldsymbol{\pi} \sim f(\boldsymbol{\pi}), \tag{2.15a}$$

$$\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^* \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}_k^*), \tag{2.15b}$$

$$G(\boldsymbol{\theta}_n) = \sum_{k=1}^{K} \pi_k \delta(\boldsymbol{\theta}_n - \boldsymbol{\theta}_k^*), \tag{2.15c}$$

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N|G \overset{\text{i.i.d.}}{\sim} G(\boldsymbol{\theta}_n), \tag{2.15d}$$

$$\boldsymbol{x}_n|\boldsymbol{\theta}_n \sim f(\boldsymbol{x}_n|\boldsymbol{\theta}_n) \quad \text{independently for} \quad n = 1, \ldots, N. \tag{2.15e}$$

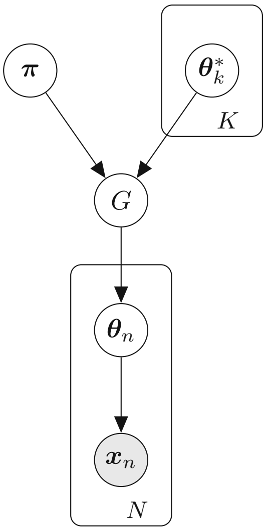


Figure 2.2: Bayesian network representing the BMM with the mixing distribution in (2.15).

A graphical model representation of the BMM in (2.15) is shown in Figure 2.2.

We note that the BMM using latent indicator variables (2.7) or the BMM using the latent mixing distribution (2.15) can be used interchangeably, since we arrive at the same conditional distribution $f(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta}^*)$ for the observations $\boldsymbol{x}_n$ (cf. (2.9) and (2.14)). The choice between the two hierarchical models is motivated by the use case. When only the clustering results are required, i.e., the cluster assignment for each observation, it is preferable to use the latent indicator variable model since the cluster assignment for each observation $\boldsymbol{x}_n$ is directly given by the estimate of the corresponding indicator variable z_n. When estimates of the local parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ are required, the latent mixing distribution representation is preferable. This representation provides us with the cluster parameters, or more precisely, the component parameters $\boldsymbol{\theta}_k^*$ of the non-empty components, directly. Hence, each parameter $\boldsymbol{\theta}_n$, for $n = 1, \dots, N$, takes on values from $\{\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_L^*\}$, where $L \leq K$. Clustering the observations $\boldsymbol{x}$ using this representation requires an additional post processing step in which we assign the observations $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$ to the same cluster, if and only if $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n'}$. The number of observations generated by component k of the BMM in (2.15) is given by

$$N_k = \sum_{n=1}^{N} \mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\theta}_k^*). \tag{2.16}$$

The number of clusters L in the observations $\boldsymbol{x}$ is again determined according to (2.11).

Choosing an appropriate number of components K in advance can be quite challenging. There are several strategies to tackle this problem from a modeling perspective. One is to fix K at $+\infty$, which leads to *Bayesian nonparametric* (BNP) mixture models, i.e., BMMs using an infinite-dimensional parameter space and with an infinite number of components. The most prominent example of a BNP mixture model is the *Dirichlet process mixture* (DPM) model [26, Chapter 3]. A second, and maybe the most intuitive approach from a Bayesian perspective, is to treat the number of components K as an unknown and random parameter with (a discrete) prior $p(K)$. In consequence, K must now be inferred jointly with the other parameters $\boldsymbol{\theta}^*$, $\boldsymbol{\pi}$ and $\boldsymbol{z}$. In literature, such kinds of BMM are referred to as *mixtures of finite mixtures* (MFMs) [20],[19]. The MFM model will be discussed in detail in Chapter 3.

2.5 Bayesian Estimation in Mixture Models

In this section, we address the challenge of estimating the latent parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*$ in a BMM using the basic formulation in (2.3). We arrange the latent parameters in the vector $\boldsymbol{w} \in \mathbb{R}^P$ according to $\boldsymbol{w} = \left(\boldsymbol{\pi}^{\mathrm{T}} \;\; \boldsymbol{\theta}^{*\mathrm{T}}\right)^{\mathrm{T}}$ where $P = K + Kp$ is the number of scalar parameters in the BMM. In Bayesian estimation, the conditional distribution of the latent parameters given the observations $f(\boldsymbol{w}|\boldsymbol{x})$ is of special importance. It is referred to as the *posterior pdf* or the *posterior*. Let $\hat{\boldsymbol{w}}(\boldsymbol{x})$ be an estimator, i.e., a (vector) function of the observations $\boldsymbol{x}$ and let the cost function $C(\boldsymbol{e})$ be a scalar-valued, nonnegative function of the P-dimensional estimation error $\boldsymbol{e} = \hat{\boldsymbol{w}} - \boldsymbol{w}$. An estimator $\hat{\boldsymbol{w}}_{\mathrm{B}}(\boldsymbol{x})$ is said to be a Bayesian estimator, if it minimizes the mean cost [27] of the posterior distribution $f(\boldsymbol{w}|\boldsymbol{x})$, i.e.,

$$\hat{\boldsymbol{w}}_{\mathrm{B}}(\boldsymbol{x}) = \arg\min_{\hat{\boldsymbol{w}}} \int_{\mathbb{R}^P} C(\hat{\boldsymbol{w}} - \boldsymbol{w}) f(\boldsymbol{w}|\boldsymbol{x})\, d\boldsymbol{w} = \arg\min_{\hat{\boldsymbol{w}}} \; \mathrm{E}^{(f(\boldsymbol{w}|\boldsymbol{x}))}\{C(\hat{\boldsymbol{w}} - \boldsymbol{w})\}. \tag{2.17}$$

Various Bayesian estimators can be obtained from (2.17) by choosing different cost functions. In what follows, we will consider two special Bayesian estimators that will be used later on to estimate the parameters of a BMM.

2.5.1 Minimum Mean Square Error (MMSE) Estimator

The MMSE estimator is one of the most important Bayesian estimators. It is obtained when the cost function in (2.17) is specified as the squared estimation error, i.e., $C(\hat{\boldsymbol{w}} - \boldsymbol{w}) = \|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 = \|\boldsymbol{e}\|_2^2$. Working out (2.17) using the squared estimation error as cost function yields

$$\hat{\boldsymbol{w}}_{\mathrm{MMSE}}(\boldsymbol{x}) = \int_{\mathbb{R}^P} \boldsymbol{w} f(\boldsymbol{w}|\boldsymbol{x})\, d\boldsymbol{w} = \mathrm{E}^{(f(\boldsymbol{w}|\boldsymbol{x}))}\{\boldsymbol{w}\}. \tag{2.18}$$

Hence, the MMSE estimator equals the conditional expectation of $\boldsymbol{w}$ with respect to the posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$, also known as the *posterior expectation* or the *posterior mean*.

2.5.2 Maximum A-Posteriori (MAP) Estimator

A cost function which assigns zero cost to the case where $\hat{\boldsymbol{w}} = \boldsymbol{w}$ and evaluates to one otherwise leads to the MAP estimator. It is given by

$$\hat{\boldsymbol{w}}_{\mathrm{MAP}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{w}} f(\boldsymbol{w}|\boldsymbol{x}). \tag{2.19}$$

Hence, the MAP estimator equals the global maximum of the posterior $f(\boldsymbol{w}|\boldsymbol{x})$. Using Bayes' law, the posterior can be expressed as

$$f(\boldsymbol{w}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{w})f(\boldsymbol{w})}{f(\boldsymbol{x})}. \tag{2.20}$$

Therefore, the MAP estimator can alternatively be expressed as

$$\hat{\boldsymbol{w}}_{\mathrm{MAP}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{w}} f(\boldsymbol{x}|\boldsymbol{w})f(\boldsymbol{w}), \tag{2.21}$$

where the *evidence* $f(\boldsymbol{x})$ in (2.20) is a constant with respect to the maximization in (2.19) and thus has been omitted. Equation (2.21) shows that both the observations $\boldsymbol{x}$ (via the likelihood function $f(\boldsymbol{x}|\boldsymbol{w})$) and the prior $f(\boldsymbol{w})$ influence the estimation result. If a weak-informative prior (e.g., $f(\boldsymbol{w})$ is a flat function) is chosen, it will weakly (if at all) influence the position of the maximum in (2.21), hence [27]

$$\hat{\boldsymbol{w}}_{\mathrm{MAP}}(\boldsymbol{x}) \approx \arg\max_{\boldsymbol{w}} f(\boldsymbol{x}|\boldsymbol{w}). \tag{2.22}$$

We conclude that the posterior $f(\boldsymbol{w}|\boldsymbol{x})$ is a key component of Bayesian estimation. For the BMM in (2.3), we obtain an expression for the posterior from (2.20):

$$f(\boldsymbol{w}|\boldsymbol{x}) = f(\boldsymbol{\pi}, \boldsymbol{\theta}^*|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}^*)f(\boldsymbol{\pi}, \boldsymbol{\theta}^*)}{f(\boldsymbol{x})}. \tag{2.23}$$

Due to the statistical independence of the latent parameters (cf. Section 2.1), the joint distribution of the latent model parameters can be factorized as

$$f(\boldsymbol{\pi}, \boldsymbol{\theta}^*) = f(\boldsymbol{\pi}) \prod_{k=1}^{K} f(\boldsymbol{\theta}_k^*). \tag{2.24}$$

Inserting the likelihood function (2.4) and the joint prior pdf (2.24) into (2.23) yields

$$f(\boldsymbol{\pi}, \boldsymbol{\theta}^*|\boldsymbol{x}) = \frac{\left(\prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_n|\boldsymbol{\theta}_k^*)\right) f(\boldsymbol{\pi}) \prod_{k=1}^{K} f(\boldsymbol{\theta}_k^*)}{f(\boldsymbol{x})}. \tag{2.25}$$

This expression shows that the calculation of the corresponding posterior is generally associated with a high computational cost since the likelihood function (2.4) is involved. In order to apply Bayesian estimators efficiently in the BMM framework, approximation methods are used to obtain a tractable approximation of the posterior $f(\boldsymbol{w}|\boldsymbol{x})$. One prominent example is variational inference (VI) [28] — in particular, coordinate-ascent variational inference (CAVI) — where the posterior $f(\boldsymbol{w}|\boldsymbol{x})$ is approximated by a computationally more tractable distribution $q(\boldsymbol{w})$ which is referred to as a *variational distribution*. Once it has been determined, an approximation of the MMSE and MAP estimator is given by exchanging $f(\boldsymbol{w}|\boldsymbol{x})$ with $q(\boldsymbol{w})$ in (2.18) and (2.19), respectively. CAVI for BMMs — in particular, CAVI for MFMs — will be discussed in Chapter 4. Note that a well-established alternative to VI is given by Monte Carlo (MC) sampling methods which are not addressed in this thesis.

Chapter 3

Mixture of Finite Mixtures Model

In this chapter, we provide an introduction to the mixture of finite mixtures (MFM) model, which is a BMM with an unknown number of components K that is distributed according to some specified prior pmf $p(K)$. The first four sections of this chapter are based on [20] where different approaches to the MFM model regarding the specification of the prior on the number of components and the mixture weights are combined to construct a *generalized MFM* model. In Section 3.5, we present representations for the special case of static MFMs based on [19].

3.1 Model Formulation

The generalized MFM model for N conditionally independent observations $\boldsymbol{x} = \left(\boldsymbol{x}_1^{\mathrm{T}} \quad \cdots \quad \boldsymbol{x}_N^{\mathrm{T}}\right)^{\mathrm{T}}$ builds on the BMM using the latent indicator variables discussed in Section 2.3 and is therefore defined in a hierarchical way:

$$K \sim p(K), \tag{3.1a}$$

$$\boldsymbol{\pi}|K;\beta(\cdot) \sim f(\boldsymbol{\pi}|K;\beta(\cdot)) = \mathcal{D}(\boldsymbol{\pi} = (\pi_1 \quad \cdots \quad \pi_K)^{\mathrm{T}}; \boldsymbol{\beta} = \beta(K)\mathbf{1}_K), \tag{3.1b}$$

$$z_1, \ldots, z_N|\boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} p(z_n|\boldsymbol{\pi}) = \mathcal{C}(z_n;\boldsymbol{\pi}), \tag{3.1c}$$

$$\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*|K \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}_k^*), \tag{3.1d}$$

$$\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n \sim f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n) = f\left(\boldsymbol{x}_n|\boldsymbol{\theta}_{z_n}^*\right) \quad \text{independently for} \quad n = 1, \ldots, N, \tag{3.1e}$$

with conditional independence assumptions summarized by

$$\boldsymbol{\theta}_k^* \perp\!\!\!\perp \boldsymbol{\pi} \,|\, K \quad \text{for all} \quad k = 1, \ldots, K, \tag{3.2a}$$

$$\boldsymbol{\theta}_k^* \perp\!\!\!\perp z_n \,|\, \boldsymbol{\pi}, K \quad \text{for all} \quad k = 1, \ldots, K \text{ and } n = 1, \ldots, N, \tag{3.2b}$$

$$z_n \perp\!\!\!\perp K \,|\, \boldsymbol{\pi} \quad \text{for all} \quad n = 1, \ldots, N, \tag{3.2c}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{\pi}, \boldsymbol{x}_{n'}, z_{n'}, \boldsymbol{\theta}_k^* \,|\, z_n, \boldsymbol{\theta}_{z_n}^* \quad \text{for all} \quad n' \neq n = 1, \ldots, N \quad \text{and} \quad k \neq z_n, \tag{3.2d}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp K \,|\, \boldsymbol{\theta}^*, z_n. \tag{3.2e}$$

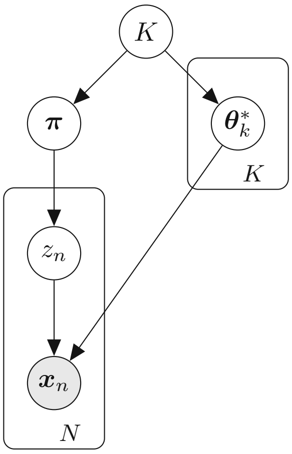


Figure 3.1: Bayesian network representing the generalized MFM model given in (3.1).

The corresponding Bayesian network is shown in Figure 3.1.

To have consistency for the number of components, it is necessary that the prior pmf $p(K)$ satisfies $p(K) > 0$ for all $K \in \mathbb{N}$ [10]. Furthermore, $p(K)$ has to exclude zero in order to be a proper prior in a mixture context, which is not fulfilled by most of the discrete probability distributions. A convenient way to achieve this condition for many discrete distributions is to define a translated prior $K - 1 \sim p_t(\cdot)$. The prior $p(K)$ is then obtained by evaluating the translated prior at $K - 1$, i.e., $p(K) = p_t(K - 1)$. Recall the crucial distinction between the number of components K and the number of clusters L. For a given K, the number of clusters is determined according to (2.10) and (2.11). For cluster labeling purposes, $l \in \{1, \dots, L\}$ is used as a subscript later on. We note that due to the prior $p(K)$ on the number of components K, the number of clusters L is a priori random as well.

The meaning of the term *mixture of finite mixtures* becomes clear when one examines the corresponding joint conditional distribution $f(\boldsymbol{x}|\boldsymbol{\pi}', \boldsymbol{\theta}^{*\prime})$. It can be expressed as

$$f(\boldsymbol{x}|\boldsymbol{\pi}', \boldsymbol{\theta}^{*\prime}) = \sum_{K=1}^{\infty} p(K) \prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, \boldsymbol{\pi}, K), \tag{3.3}$$

which is a countably infinite mixture of finite mixtures with K components. The parameter vectors $\boldsymbol{\pi}'$ and $\boldsymbol{\theta}^{*\prime}$ are super vectors, which contain the mixture weights $\boldsymbol{\pi}$ and component parameters $\boldsymbol{\theta}^*$ of each finite mixture combined in (3.3). The finite mixture distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, \boldsymbol{\pi}, K)$ depends on the random number of components K through

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, \boldsymbol{\pi}, K) = \sum_{z_n=1}^{K} \pi_{z_n} f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n, K). \tag{3.4}$$

Because of the conditional independence assumptions given by (3.2e) and (3.2d) we have

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n, K) = f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n) = f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}).$$

Therefore, the mixture distribution in (3.4) can be expressed as

$$f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, \boldsymbol{\pi}, K) = \sum_{z_n=1}^{K} \pi_{z_n} f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}). \tag{3.5}$$

By inserting (3.5) into (3.3) we obtain

$$f(\boldsymbol{x}|\boldsymbol{\pi}', \boldsymbol{\theta}^{*\prime}) = \sum_{K=1}^{\infty} p(K) \prod_{n=1}^{N} \sum_{z_n=1}^{K} \pi_{z_n} f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}), \tag{3.6}$$

i.e., a weighted sum of finite mixtures with different numbers of components, each represented by its likelihood $\prod_{n=1}^{N} \sum_{z_n=1}^{K} \pi_{z_n} f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n})$. The corresponding weights are given by the prior pmf on the number of components $p(K)$ evaluated at $K \in \{1, 2, \ldots\}$.

In the generalized MFM model in (3.1), a symmetric Dirichlet prior on the mixture weights $\boldsymbol{\pi}$ is assumed (cf. (3.1b)). It is defined for each $K \in \mathbb{N}$ according to the corresponding hyperparameter $\boldsymbol{\beta} = \beta(K)\mathbf{1}_K$, where $\mathbf{1}_K$ is the all-one vector of length K and the scalar $\beta(K) \geq 0$ is given by some deterministic function of K. This includes the special cases where $\beta(K) = \kappa/K$ referred to as the *dynamic MFM* model [11] and $\beta(K) = \beta$ referred to as the *static MFM* model [19],[9]. A priori, for both cases the mean of the mixture weights given K is given by

$$\mathrm{E}^{(f(\boldsymbol{\pi}|K;\beta(\cdot)))}\{\boldsymbol{\pi}\} = \frac{1}{K}\mathbf{1}_K \tag{3.7}$$

and the variance for each individual weight π_k, for $k = 1, \ldots, K$, is given by

$$\mathrm{var}\{\pi_k\} = \frac{K-1}{K\beta(K)+1}. \tag{3.8}$$

From (3.8) it can be concluded, that the variance decreases with increasing $\beta(K)$ which leads to more balanced mixture weights, i.e., mixture weights taking on similar values, and vice versa. Hence, the types of finite mixtures which are combined in (3.6) vary with $\beta(K)$ since the dependency of the hyperparameter $\beta(K)$ on K leads to a combination of finite mixtures favouring different distributions of the component sizes (cf. (2.10)) $N_1, \ldots, N_K$.

For a dynamic MFM with $\beta(K) = \kappa/K$, finite mixtures with balanced component sizes for small K are mixed with sparse finite mixture (SFM) models [29] for moderate K and, as K goes to infinity, with DPMs where the component sizes are extremely unbalanced. As stated in [29], the concept of SFM modeling is based on specifying an overfitting finite mixture model with too many components K and assuming heterogeneity for all available variables a priori. Sparse solutions with regard to the number of mixture components and with regard to heterogeneity of component locations are induced by specifying suitable shrinkage priors on, respectively, the mixture weights and the component parameters.

In contrast to the dynamic MFM model, finite mixtures of a similar type are combined for static MFMs, where $\beta(K) = \beta$ which obviously limits the MFM model in its flexibility.

3.2 Exchangeable Partition Probability Function and Distribution of the Cluster Sizes

Let $\mathcal{C} = \{\mathcal{C}_1, \cdots, \mathcal{C}_L\}$ be a random partition of the N observations $\boldsymbol{x} = (\boldsymbol{x}_1^{\mathrm{T}} \; \cdots \; \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$ induced by the MFM model in (3.1) through the indicator variables $\boldsymbol{z} = (z_1 \; \cdots \; z_N)^{\mathrm{T}}$, where $\mathcal{C}_l = \{n : z_n = l\}$, for $l = 1, \dots, L$. In other words, cluster $\mathcal{C}_l$ contains all observations generated by the component $f(\boldsymbol{x}_n|\boldsymbol{\theta}_l^*)$. Further, we denote the corresponding cluster sizes by N_l', i.e., $N_l' = \mathrm{card}(\mathcal{C}_l)$, for $l = 1, \dots, L$. Here, $\mathrm{card}(\mathcal{C}_l)$ denotes the number of observations assigned to cluster $\mathcal{C}_l$. Note that all proofs in this section are based on [20]. For the sake of comprehensibility, intermediate steps have been added where it made sense.

Theorem 3.1: For a generalized MFM with $\boldsymbol{\pi}|K;\beta(\cdot) \sim \mathcal{D}(\boldsymbol{\pi};\beta(K)\mathbf{1}_K)$ and prior pmf $p(K)$, the set partition $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_L\}$ is distributed according to

$$p(\mathcal{C};N,\beta(\cdot)) = \sum_{K=L}^{\infty} p(K)p(\mathcal{C}|K;N,\beta(\cdot)), \tag{3.9}$$

where

$$p(\mathcal{C}|K;N,\beta(\cdot)) = \frac{V_{N,L}^{(K,\beta(K))}}{\Gamma(\beta(K))^L} \prod_{l=1}^{L} \Gamma(N_l' + \beta(K)) \tag{3.10}$$

and

$$V_{N,L}^{(K,\beta(K))} = \frac{\Gamma(K\beta(K))K!}{\Gamma(K\beta(K)+N)(K-L)!}. \tag{3.11}$$

Proof: Let us first consider the conditional pmf $p(\boldsymbol{z}|K;\beta(\cdot))$ for a fixed number of components K. It can be obtained by marginalizing out the mixture weights $\boldsymbol{\pi}$ from the joint conditional distribution $f(\boldsymbol{z},\boldsymbol{\pi}|K;N,\beta(\cdot))$, i.e.,

$$p(\boldsymbol{z}|K;N,\beta(\cdot)) = \int_{\mathbb{R}^K} f(\boldsymbol{z},\boldsymbol{\pi}|K;N,\beta(\cdot))\, d\boldsymbol{\pi} = \int_{\mathbb{R}^K} p(\boldsymbol{z}|\boldsymbol{\pi}) f(\boldsymbol{\pi}|K;\beta(\cdot))\, d\boldsymbol{\pi}, \tag{3.12}$$

where we used the conditional independence assumption (3.2c) in the last step. It follows from (3.1c) that the indicator variables $\boldsymbol{z}$ given the mixture weights $\boldsymbol{\pi}$ are i.i.d. categoricals and therefore

$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{\mathbb{1}(z_n=k)} = \prod_{k=1}^{K} \pi_k^{\sum_{n=1}^{N}\mathbb{1}(z_n=k)} = \prod_{k=1}^{K} \pi_k^{N_k}, \tag{3.13}$$

where we used (2.10) in the last step. Since a symmetric Dirichlet prior is assumed for the mixture weights $\boldsymbol{\pi}$ (cf. (3.1b)), $f(\boldsymbol{\pi}|K;\beta(\cdot))$ is given by

$$f(\boldsymbol{\pi}|K;\beta(\cdot)) = c(\boldsymbol{\beta}) \prod_{k=1}^{K} \pi_k^{\beta(K)-1}, \tag{3.14}$$

where

$$c(\boldsymbol{\beta}) = \frac{\Gamma\left(\sum_{k=1}^{K} \beta(K)\right)}{\prod_{k=1}^{K} \Gamma(\beta(K))} \tag{3.15}$$

is the normalization constant. Multiplying (3.13) by (3.14) yields

$$p(\boldsymbol{z}|\boldsymbol{\pi}) f(\boldsymbol{\pi}|K;\beta(\cdot)) = c(\boldsymbol{\beta}) \prod_{k=1}^{K} \pi_k^{\beta(K)+N_k-1} = \frac{c(\boldsymbol{\beta})}{c(\tilde{\boldsymbol{\beta}})} c(\tilde{\boldsymbol{\beta}}) \prod_{k=1}^{K} \pi_k^{\tilde{\beta}(k)-1} = \frac{c(\boldsymbol{\beta})}{c(\tilde{\boldsymbol{\beta}})} \mathcal{D}(\boldsymbol{\pi}; \tilde{\beta}(K)\mathbf{1}_K). \tag{3.16}$$

Note that the product of $p(\boldsymbol{z}|\boldsymbol{\pi})$ and the prior $f(\boldsymbol{\pi}|K;\beta(\cdot))$ in (3.16) takes on, up to a proportionality constant, the same functional form as $f(\boldsymbol{\pi}|K,\beta(K))$, although with a different hyperparameter $\tilde{\beta}(K) = \beta(K) + N_k$. This is a consequence of choosing the Dirichlet distribution as the prior $f(\boldsymbol{\pi}|K;\beta(\cdot))$ for the mixture weights, which is conjugate to the categorical distribution $p(z_n|\boldsymbol{\pi})$ of the indicator variables. Due to the specific form of (3.16), the integral in (3.12) simplifies to an integral over a Dirichlet distribution, which equals one because the Dirichlet distribution is a valid pdf. We therefore obtain

$$p(\boldsymbol{z}|K;N,\beta(\cdot)) = c(\boldsymbol{\beta}) \frac{1}{c(\tilde{\boldsymbol{\beta}})} \int_{\mathbb{R}^K} \mathcal{D}(\boldsymbol{\pi}; \tilde{\beta}(K)\mathbf{1}_K)\, d\boldsymbol{\pi} = c(\boldsymbol{\beta}) \frac{1}{c(\tilde{\boldsymbol{\beta}})}. \tag{3.17}$$

Inserting (3.15) and substituting back $\tilde{\beta}(K) = \beta(K) + N_k$ leads to

$$p(\boldsymbol{z}|K;N,\beta(\cdot)) = \frac{\Gamma\left(\sum_{k=1}^{K} \beta(K)\right)}{\prod_{k=1}^{K} \Gamma(\beta(K))} \frac{\prod_{k=1}^{K} \Gamma(\beta(K)+N_k)}{\Gamma\left(\sum_{k=1}^{K} (\beta(K)+N_k)\right)} = \frac{\Gamma(K\beta(K))}{\Gamma(K\beta(K)+N)} \prod_{k=1}^{K} \frac{\Gamma(\beta(K)+N_k)}{\Gamma(\beta(K))}, \tag{3.18}$$

where we used (2.12) in the last step. We next reorder the components such that the $L \leq K$ occupied components appear first. Then, the indicator variables $\boldsymbol{z} = (z_1 \ \cdots \ z_N)^{\mathrm{T}}$ define a set partition $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}$ of the observation indices $n \in \{1, \ldots, N\}$. The number of assignment vectors $\boldsymbol{z}$ that lead to the same partition $\mathcal{C}$ is given by

$$\binom{K}{L} L! = \frac{K!}{(K-L)!}. \tag{3.19}$$

Here, $\binom{K}{L}$ denotes the number of possibilities to choose L clusters labeled by $l \in \{1, \ldots, L\}$ among K components, and L factorial accounts for the possibilities to relabel these L clusters. Note that the product in (3.18) can be reduced to L factors since $\frac{\Gamma(\beta(K)+N_k)}{\Gamma(\beta(K))} = 1$ for $N_k = 0$. Thus, (3.18) can be reformulated according to

$$p(\boldsymbol{z}|K;N,\beta(\cdot)) = \frac{\Gamma(K\beta(K))}{\Gamma(K\beta(K)+N)} \prod_{l=1}^{L} \frac{\Gamma(\beta(K)+N_l')}{\Gamma(\beta(K))} \tag{3.20}$$

We assume that each of the assignment vectors $\boldsymbol{z}$ leading to the same partition $\mathcal{C}$ is equally likely, i.e., we assume a uniform distribution over the number of outcomes given in (3.19):

$$p(\boldsymbol{z}|\mathcal{C},K) = \frac{(K-L)!}{K!}. \tag{3.21}$$

Using Bayes' law, the pmf of the set partition $\mathcal{C}$ given the number of components K can be expressed as

$$p(\mathcal{C}|K;N,\beta(\cdot)) = \frac{p(\mathcal{C}|\boldsymbol{z},K)p(\boldsymbol{z}|K;N,\beta(\cdot))}{p(\boldsymbol{z}|\mathcal{C},K)}. \tag{3.22}$$

Let $\mathcal{C}(\boldsymbol{z})$ be a function which obtains the partition $\mathcal{C}$ defined by $\boldsymbol{z}$. Thus, the conditional pmf $p(\mathcal{C}|\boldsymbol{z},K)$ in (3.22) is given by $p(\mathcal{C}|\boldsymbol{z},K) = \mathbb{1}(\mathcal{C} = \mathcal{C}(\boldsymbol{z}))$. We assume that $\mathcal{C}$ obeys $\boldsymbol{z}$, i.e., $\mathcal{C} = \mathcal{C}(\boldsymbol{z})$, which leads to

$$p(\mathcal{C}|\boldsymbol{z},K) = 1. \tag{3.23}$$

Inserting (3.20), (3.21), and (3.23) into (3.22) yields

$$p(\mathcal{C}|K;N,\beta(\cdot)) = \frac{\Gamma(K\beta(K))K!}{\Gamma(K\beta(K)+N)(K-L)!}\prod_{l=1}^{L}\frac{\Gamma(\beta(K)+N'_l)}{\Gamma(\beta(K))} = \frac{V_{N,L}^{(K,\beta(K))}}{\Gamma(\beta(K))^L}\prod_{l=1}^{L}\Gamma(\beta(K)+N'_l). \tag{3.24}$$

The pmf $p(\mathcal{C};N,\beta(\cdot))$ of the set partition $\mathcal{C} = \{\mathcal{C}_1,\ldots,\mathcal{C}_L\}$ given by (3.9) is obtained from (3.24) by marginalizing out K:

$$p(\mathcal{C};N,\beta(\cdot)) = \sum_{K=L}^{\infty} p(K)p(\mathcal{C}|K;N,\beta(\cdot)).$$

Here, the summation starts at $K = L$ since $p(\mathcal{C}|K;N,\beta(\cdot)) = 0$ for $K < L$ due to the binomial coefficient $\binom{K}{L}$ being involved. □

Theorem 3.2: For a generalized MFM with $\boldsymbol{\pi}|K;\beta(\cdot) \sim \mathcal{D}(\boldsymbol{\pi};\beta(K)\mathbf{1}_K)$ and prior pmf $p(K)$, the labeled cluster sizes $N'_1,\ldots,N'_L$ are jointly distributed according to

$$p(N'_1,\ldots,N'_L;N,\beta(\cdot)) = \frac{N!}{L!}\sum_{K=L}^{\infty} p(K)\frac{V_{N,L}^{(K,\beta(k))}}{\Gamma(\beta(K))^L}\prod_{l=1}^{L}\frac{\Gamma(N_l+\beta(K))}{\Gamma(N_l+1)}, \tag{3.25}$$

where the factor $V_{N,L}^{(K,\beta(k))}$ is given by (3.11).

We note that the pmfs $p(\mathcal{C};N,\beta(\cdot))$ and $p(N'_1,\ldots,N'_L;N,\beta(\cdot))$ are commonly referred to as *prior distributions* in the sense that they are available before the values of $\boldsymbol{x}_n$ are observed as long as the number of observations N is known.

Proof: We now consider the sizes $N'_1,\ldots,N'_L$, $N'_l \in \mathbb{N}$, of the $L \leq K$ clusters, i.e., non-empty components, labeled according to $l \in \{1,\ldots,L\}$. Recall that the number of possibilities to choose L clusters from K components is given by

$$\binom{K}{L} = \frac{K!}{L!(K-L)!}. \tag{3.26}$$

Furthermore, there are

$$\frac{N!}{N'_1!N'_2!\cdots N'_L!} = \frac{N!}{\prod_{l=1}^{L} N'_l!} \tag{3.27}$$

different ways to partition N observations into L clusters with sizes $N'_1, \ldots, N'_L$. Thus, the number of assignment vectors $\boldsymbol{z}$ that lead to the same cluster sizes $N'_1, \ldots, N'_L$ is given by the product of (3.26) and (3.27). Assuming that each of those vectors is equally likely leads to

$$p(\boldsymbol{z}|N'_1, \ldots, N'_L, K) = \frac{L!(K-L)! \prod_{l=1}^{L} N'_l!}{K!N!}. \tag{3.28}$$

Since $\Gamma(y+1) = y\Gamma(y) = y(y-1)! = y!$ for $y \in \mathbb{N}$, (3.28) can be further developed as

$$p(\boldsymbol{z}|N'_1, \ldots, N'_L, K) = \frac{L!(K-L)! \prod_{l=1}^{L} \Gamma(N'_l + 1)}{K!N!}. \tag{3.29}$$

Using Bayes' law, the joint pmf of the labeled cluster sizes $N'_1, \ldots, N'_L$ given the number of components K can be expressed as

$$p(N'_1, \ldots, N'_L|K; N, \beta(\cdot)) = \frac{p(N'_1, \ldots, N'_L|\boldsymbol{z},K) p(\boldsymbol{z}|K; N, \beta(\cdot))}{p(\boldsymbol{z}|N'_1, \ldots, N'_L, K)}, \tag{3.30}$$

Let the cluster sizes $N'_1, \ldots, N'_L$ be arranged in the vector $\boldsymbol{n}' = (N'_1 \ \cdots \ N'_L)^{\mathrm{T}}$ and let $\boldsymbol{n}'(\boldsymbol{z})$ be a vector function which obtains the cluster sizes $\boldsymbol{n}'$ defined by $\boldsymbol{z}$. Thus, the conditional pmf $p(N'_1, \ldots, N'_L|\boldsymbol{z},K)$ in (3.30) is given by $p(N'_1, \ldots, N'_L|\boldsymbol{z},K) = \mathbb{1}(\boldsymbol{n}' = \boldsymbol{n}'(\boldsymbol{z}))$. We assume that $\boldsymbol{n}'$ obeys $\boldsymbol{z}$, i.e., $\boldsymbol{n}' = \boldsymbol{n}'(\boldsymbol{z})$, which leads to

$$p(N'_1, \ldots, N'_L|\boldsymbol{z},K) = 1. \tag{3.31}$$

Inserting (3.20), (3.29), and (3.31) into (3.30) yields

$$\begin{aligned} p(N'_1, \ldots, N'_L|K; N, \beta(\cdot)) &= \frac{N!}{L! \prod_{l=1}^{L} \Gamma(N'_l + 1)} \frac{\Gamma(K\beta(K))K!}{\Gamma(K\beta(K) + N)(K-L)!} \prod_{l=1}^{L} \frac{\Gamma(\beta(K) + N'_l)}{\Gamma(\beta(K))} \\ &= \frac{N!}{L!} \frac{V_{N,L}^{(K,\beta(K))}}{\Gamma(\beta(K))^L} \prod_{l=1}^{L} \frac{\Gamma(\beta(K) + N'_l)}{\Gamma(N'_l + 1)}. \end{aligned} \tag{3.32}$$

Finally, by marginalizing out K from (3.32) we obtain

$$p(N'_1, \ldots, N'_L; N, \beta(\cdot)) = \frac{N!}{L!} \sum_{K=L}^{\infty} p(K) \frac{V_{N,L}^{(K,\beta(k))}}{\Gamma(\beta(K))^L} \prod_{l=1}^{L} \frac{\Gamma(N'_l + \beta(K))}{\Gamma(N'_l + 1)},$$

which is the joint distribution of the labeled cluster sizes in (3.25). Note that the summation starts at $K = L$ since the binomial coefficient $\binom{K}{L}$ is involved in (3.32). □

The pmf $p(\mathcal{C}; N, \beta(\cdot))$ in (3.9) is symmetric in its arguments, i.e., it remains the same for every permutation of its arguments — the cluster sizes $N'_1, \ldots, N'_L$. It defines an exchangeable random partition $\mathcal{C}$ of the N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ under the generalized MFM model in (3.1). In other words, the distribution of the partition $\mathcal{C}$ is invariant under permutations of $\{1, \ldots, N\}$. Due to the symmetry with respect to the cluster sizes $N'_1, \ldots, N'_L$, the distribution $p(\mathcal{C}; N, \beta(\cdot))$ is an *exchangeable partition probability function* (EPPF) [30].

For a static MFM, where $\beta(K) = \beta$, the EPPF is given by

$$p_{\text{stat}}(\mathcal{C}; N, \beta) = V_{N,L}^{(\beta)} \prod_{l=1}^{L} \frac{\Gamma(N_l' + \beta)}{\Gamma(\beta)}, \tag{3.33}$$

where

$$V_{N,L}^{(\beta)} = \sum_{K=L}^{\infty} p(K) \frac{\Gamma(K\beta)K!}{\Gamma(K\beta + N)(K-L)!}. \tag{3.34}$$

Proposition 3.3: The coefficients $V_{N,L}^{(\beta)}$ in (3.34) satisfy the recursion

$$V_{N,L}^{(\beta)} = (N + L\beta)V_{N+1,L}^{(\beta)} + \beta V_{N+1,L+1}^{(\beta)} \tag{3.35}$$

for $L = 1, \dots, N-1$.

Proof: As in [20], we use the identity $\Gamma(z) = \Gamma(z+1)/z$, and obtain

$$\Gamma(K\beta + N) = \frac{\Gamma(K\beta + N + 1)}{K\beta + N} = \frac{\Gamma(K\beta + N + 1)}{(N + L\beta) + \beta(K - L)}. \tag{3.36}$$

Inserting (3.36) into (3.34) yields

$$\begin{aligned} V_{N,L}^{(\beta)} &= (N + L\beta) \sum_{K=L}^{\infty} p(K) \frac{\Gamma(K\beta)K!}{\Gamma(K\beta + N + 1)(K-L)!} + \beta \sum_{K=L}^{\infty} p(K) \frac{\Gamma(K\beta)K!}{\Gamma(K\beta + N + 1)(K-L-1)!} \\ &= (N + L\beta)V_{N+1,L}^{(\beta)} + \beta V_{N+1,L+1}^{(\beta)}, \end{aligned}$$

which is the recursion given in (3.35). □

Due to the special product form of $p_{\text{stat}}(\mathcal{C}|N, \beta)$ in (3.33) and the recursion given by (3.35), the exchangeable random partition $\mathcal{C}$ for the static MFM model is of *Gibbs form*, i.e., the EPPF $p_{\text{stat}}(\mathcal{C}; N, \beta)$ is a member of the family of Gibbs partition distributions. An exchangeable random partition $\mathcal{C}$ is said to be of Gibbs form if for some sets of nonnegative weights $\{W_l\}$ and $\{V_{N,L}\}$ the EPPF of $\mathcal{C}$ satisfies

$$p(\mathcal{C}) = V_{N,L} \prod_{l=1}^{L} W_{N_l'}$$

for all $1 \leq L \leq N$ and for all compositions $\{N_1', \dots, N_L'\}$ of N [31].

This fact is exploited in [19], where the proposed MCMC algorithms for doing posterior inference of static MFM models are based on sampling techniques for the class of BNP mixtures with exchangeable random partitions of Gibbs form. In particular, the direct application of inference algorithms for DPMs to static MFMs is shown. For a DPM with concentration parameter κ, the EPPF is given by the Ewens distribution, which is of Gibbs form as well:

$$p_{\text{DPM}}(\mathcal{C}; N, \kappa) = \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} \prod_{l=1}^{L} \Gamma(N_l'). \tag{3.37}$$

We note that for a generalized MFM, where the hyperparameter $\beta(K)$ depends on K, an EPPF beyond the family of Gibbs partition distributions is obtained. The EPPF $p_{\text{dyn}}(\mathcal{C}; N, \kappa)$ of dynamic MFM models can be expressed explicitly in relation to (3.37).

Theorem 3.4: For a dynamic MFM with $\beta(K) = \kappa/K$, the corresponding EPPF $p_{\text{dyn}}(\mathcal{C}; N, \kappa)$ can be expressed as

$$p_{\text{dyn}}(\mathcal{C}; N, \kappa) = p_{\text{DPM}}(\mathcal{C}; N, \kappa) \sum_{K=L}^{\infty} p(K) R_{\boldsymbol{n}', L}^{(K,\kappa)}, \tag{3.38}$$

where

$$R_{\boldsymbol{n}', L}^{(K,\kappa)} = \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})(K - l + 1)}{\Gamma(1 + \frac{\kappa}{K})\Gamma(N_l')K}. \tag{3.39}$$

In (3.38), $p_{\text{DPM}}(\mathcal{C}; N, \kappa)$ is the EPPF for a DPM given by (3.37) and $\boldsymbol{n}'$ is the vector of induced cluster sizes, i.e., $\boldsymbol{n}' = (N_1' \;\; \cdots \;\; N_L')^{\mathrm{T}}$.

Proof: Inserting $\beta(K) = \kappa/K$ in (3.10) and (3.11), we obtain

$$\begin{aligned} p_{\text{dyn}}(\mathcal{C}|K; N, \kappa) &= \frac{V_{N,L}^{(K,\kappa)}}{\Gamma(\frac{\kappa}{K})^L} \prod_{l=1}^{L} \Gamma\left(N_l' + \frac{\kappa}{K}\right) \quad \text{and} \\ V_{N,L}^{(K,\kappa)} &= \frac{\Gamma(\kappa)K!}{\Gamma(\kappa + N)(K - L)!}. \end{aligned} \tag{3.40}$$

Therefore, the EPPF (cf. (3.9)) for a dynamic MFM is given by

$$p_{\text{dyn}}(\mathcal{C}; N, \kappa) = \frac{\Gamma(\kappa)}{\Gamma(\kappa + N)} \sum_{K=L}^{\infty} p(K) \frac{K!}{(K - L)!} \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})}{\Gamma(\frac{\kappa}{K})} \tag{3.41}$$

Using $\Gamma(\frac{\kappa}{K}) = \frac{K}{\kappa}\Gamma(1 + \frac{\kappa}{K})$, the factor $\frac{K!}{(K-L)!} \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})}{\Gamma(\frac{\kappa}{K})}$ in (3.41) can be reformulated as

$$\begin{aligned} \frac{K!}{(K - L)!} \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})}{\Gamma(\frac{\kappa}{K})} &= \kappa^L \frac{K!}{K^L(K - L)!} \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})}{\Gamma(1 + \frac{\kappa}{K})} \\ &= \kappa^L \prod_{l=1}^{L} \frac{K - l + 1}{K} \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})}{\Gamma(1 + \frac{\kappa}{K})} \\ &= \kappa^L \prod_{l=1}^{L} \Gamma(N_l') \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})(K - l + 1)}{\Gamma(1 + \frac{\kappa}{K})\Gamma(N_l')K}, \end{aligned} \tag{3.42}$$

where we multiplied by $\prod_{l=1}^{L} \frac{\Gamma(N_l')}{\Gamma(N_l')}$ in the last step. Inserting (3.42) into (3.41) leads to

$$\begin{aligned} p_{\text{dyn}}(\mathcal{C}; N, \kappa) &= \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} \prod_{l=1}^{L} \Gamma(N_l') \sum_{K=L}^{\infty} p(K) \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})(K - l + 1)}{\Gamma(1 + \frac{\kappa}{K})\Gamma(N_l')K} \\ &= p_{\text{DPM}}(\mathcal{C}; N, \kappa) \sum_{K=L}^{\infty} p(K) \prod_{l=1}^{L} \frac{\Gamma(N_l' + \frac{\kappa}{K})(K - l + 1)}{\Gamma(1 + \frac{\kappa}{K})\Gamma(N_l')K}, \end{aligned}$$

which proves the expression in (3.38). □

The joint distribution $p_{\text{dyn}}(N'_1, \dots, N'_L; N, \kappa)$ can be derived as follows. From (3.32) with $\beta(K) = \kappa/K$ and (3.40) we obtain the conditional joint pmf of the labeled cluster sizes for a fixed K, which is given by

$$p_{\text{dyn}}(N'_1, \dots, N'_L | K; N, \kappa) = \frac{N!}{L!} \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} \prod_{l=1}^{L} \frac{\Gamma(N'_l + \frac{\kappa}{K})(K - l + 1)}{\Gamma(1 + \frac{\kappa}{K})\Gamma(N'_l + 1)K}. \tag{3.43}$$

Marginalizing out K yields

$$p_{\text{dyn}}(N'_1, \dots, N'_L; N, \kappa) = \frac{N!}{L!} \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} \sum_{K=L}^{\infty} p(K) \prod_{l=1}^{L} \frac{\Gamma(N'_l + \frac{\kappa}{K})(K - l + 1)}{\Gamma(1 + \frac{\kappa}{K})\Gamma(N'_l + 1)K}, \tag{3.44}$$

which is the joint pmf of the labeled cluster sizes for a dynamic MFM model.

We conclude from Theorem 3.4 that dynamic MFMs can be viewed as a generalization of the Dirichlet process prior beyond the class of Gibbs-type priors. It can easily be seen from (3.39) that $\lim_{K\to\infty} R^{(K,\kappa)}_{\boldsymbol{n}',L} = 1$. Furthermore, specifying a prior pmf which places all probability mass on $K = \infty$ leads to $p(K = \infty) = 1$. Therefore, we obtain $p_{\text{dyn}}(\mathcal{C}; N, \kappa) = p_{\text{DPM}}(\mathcal{C}; N, \kappa)$ from (3.38).

3.3 Prior Distribution of the Number of Clusters

In this section, we derive the prior pmf $p(L; N, \beta(\cdot))$ of the number of clusters L, exploiting the distributions introduced in the previous section. Recall that due to the number of components being random with prior $p(K)$, the number of clusters L is a priori random as well. Although both priors are closely related to each other, $p(L; N, \beta(\cdot))$ does not necessarily match $p(K)$ for a finite number of observations, i.e., $p(L; N, \beta(\cdot)) \neq p(K)$ for $1 \leq N < \infty$.

Theorem 3.5: For a generalized MFM with $\boldsymbol{\pi}|K; \beta(\cdot) \sim \mathcal{D}(\boldsymbol{\pi}; \beta(K)\mathbf{1}_K)$ and prior $p(K)$, the prior pmf of the number of clusters L parametrized by the number of observations N and the hyperparameter $\beta(\cdot)$ is given by

$$p(L; N, \beta(\cdot)) = \frac{N!}{L!} \sum_{K=L}^{\infty} p(K) \frac{V^{(K,\beta(K))}_{N,L}}{\Gamma(\beta(K))^L} C^{(K,\beta(K))}_{N,L}, \tag{3.45}$$

where $V^{(K,\beta(K))}_{N,L}$ is given by (3.11) for each K and $C^{(K,\beta(K))}_{N,L}$ is obtained via summation over the labeled cluster sizes $N'_1, \dots, N'_L$:

$$C^{(K,\beta(K))}_{N,L} = \sum_{\mathcal{N}'_L} \prod_{l=1}^{L} \frac{\Gamma(N'_l + \beta(K))}{\Gamma(N'_l + 1)}, \tag{3.46}$$

where

$$\mathcal{N}'_L := \{N'_1, \dots, N'_L : N'_1, \dots, N'_L > 0 \text{ and } N'_1 + \dots + N'_L = N\}. \tag{3.47}$$

Proof: This proof is based on [20]. Recall the joint conditional distribution of the cluster sizes $p(N'_1,\dots,N'_L|K;N,\beta(\cdot))$ for a fixed K given by (3.32). By aggregating $p(N'_1,\dots,N'_L|K;N,\beta(\cdot))$ over $\mathcal{N}'_L$ given in (3.47), we obtain the prior pmf of the number of clusters $p(L|K;N,\beta(\cdot))$ given K:

$$\begin{aligned} p(L|K;N,\beta(\cdot)) &= \sum_{\mathcal{N}'_L} p(N'_1,\dots,N'_L|K;N,\beta(\cdot)) \\ &= \frac{N!}{L!}\frac{V_{N,L}^{(K,\beta(K))}}{\Gamma(\beta(K))^L}\sum_{\mathcal{N}'_L}\prod_{l=1}^{L}\frac{\Gamma(\beta(K)+N'_l)}{\Gamma(N'_l+1)} \\ &= \frac{N!}{L!}\frac{V_{N,L}^{(K,\beta(K))}}{\Gamma(\beta(K))^L}C_{N,L}^{(K,\beta(K))}. \end{aligned} \tag{3.48}$$

By marginalizing out K from (3.48), we obtain

$$p(L;N,\beta(K)) = \frac{N!}{L!}\sum_{K=L}^{\infty} p(K)\frac{V_{N,L}^{(K,\beta(K))}}{\Gamma(\beta(K))^L}C_{N,L}^{(K,\beta(K))},$$

which is the prior pmf of the number of clusters given in (3.45). □

We note that $C_{N,L}^{(K,\beta(K))}$ can be computed recursively with relatively low computational cost; see Algorithm 1 in [20] for further details.

For static MFM models, where $\beta(K)=\beta$, the prior on the number of clusters $p_{\text{stat}}(L;N,\beta)$ can directly be derived from Theorem 3.5 and is given by

$$p_{\text{stat}}(L;N,\beta) = \frac{N!}{L!}\frac{V_{N,L}^{(\beta)}}{\Gamma(\beta)^L}C_{N,L}^{(\beta)}, \tag{3.49}$$

where $V_{N,L}^{(\beta)}$ is given by (3.34) and $C_{N,L}^{(K,\beta(K))}$ becomes $C_{N,L}^{(\beta)}$ which is independent of K due to $\beta(K)=\beta$ (cf. (3.46)).

Finally, for a dynamic MFM model with $\beta(K)=\kappa/K$, the prior pmf of the number of clusters $p_{\text{dyn}}(L|N,\kappa)$ is given by

$$p_{\text{dyn}}(L;N,\kappa) = \frac{N!}{L!}\frac{\kappa^L\Gamma(\kappa)}{\Gamma(\kappa+N)}\sum_{K=L}^{\infty}p(K)C_{N,L}^{(K,\kappa)}\prod_{l=1}^{L}\frac{K-l+1}{\Gamma(1+\frac{\kappa}{K})K}, \tag{3.50}$$

where $C_{N,L}^{(K,\kappa)}$ is obtained from (3.46) by inserting $\beta(K)=\kappa/K$. Starting with the joint pmf of the labeled cluster sizes for a dynamic MFM given by (3.43), (3.50) can be proved in a similar manner as Theorem 3.5.

In Section 3.2 we derived the EPPF of the DPM model from the EPPF of the dynamic MFM model by placing all probability mass of $p(K)$ on infinity. With $\lim_{K\to\infty}\prod_{l=1}^{L}\frac{K-l+1}{\Gamma\left(1+\frac{\kappa}{K}\right)K}=1$ and $p(K=\infty)=1$, we obtain from (3.50) the prior pmf of the number of clusters for a DPM, i.e.,

$$p_{\text{DPM}}(L;N,\kappa) = \frac{N!}{L!}\frac{\kappa^L\Gamma(\kappa)}{\Gamma(\kappa+N)}C_{N,L}^{(\infty)}. \tag{3.51}$$

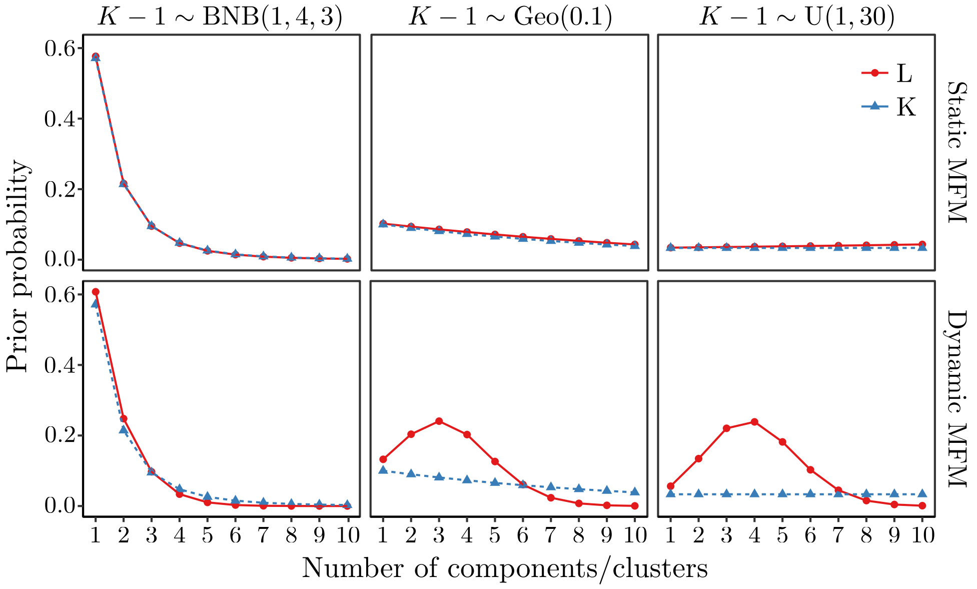


Figure 3.2: Prior pmfs $p(K)$ of the number of components K (blue) and induced prior pmfs $p_{\text{stat}}(L; N, \beta)$ with $\beta = 1$ (top) and $p_{\text{dyn}}(L; N, \kappa)$ with $\kappa = 1$ (bottom) of the number of clusters L (red) for $N = 82$. (Adapted from [20].)

Note that $C_{N,L}^{(K,\kappa)}$ becomes $C_{N,L}^{(\infty)}$ which is independent of the concentration parameter κ, since $\lim_{K\to\infty} \kappa/K = 0$.

The prior pmfs $p(L; N, \beta(K))$ of the number of clusters L for a static MFM (cf. (3.49)) with $\beta = 1$ and for a dynamic MFM (cf. (3.50)) with $\kappa = 1$ are shown in Figure 3.2. In the figure, three different distributions are considered as priors $p(K)$ of the number of components K: the translated (cf. section 3.1) beta-negative-binomial (BNB) distribution $K-1 \sim \mathrm{BNB}(1,3,4)$ with $\mathrm{E}^{(p(K))}\{K\} = 2$ suggested in [20], the translated geometric distribution $K-1 \sim \mathrm{Geo}(0.1)$ with $\mathrm{E}^{(p(K))}\{K\} = 10$ suggested in [19], and the uniform distribution $K \sim \mathcal{U}(1,30)$ with $\mathrm{E}^{(p(K))}\{K\} = 15.5$ suggested in [9]. In the case of the static MFM model, the priors of K and L roughly coincide for all three choices of $p(K)$ for values of K and L between one and ten. On the other hand, for the dynamic MFM this only holds for the BNB prior, which has a small mean value. For the prior pmfs $p(K)$ with larger mean values, a substantial difference between $p(K)$ and $p_{\text{dyn}}(L; N, \kappa)$ can be observed as probability mass is shifted towards smaller values of L.

3.4 Comparing Static and Dynamic MFMs and DPMs

In this section, we provide a more in-depth comparison of the prior pmf of the number of clusters L and the joint pmf of the cluster sizes $N'_1, \ldots, N'_L$ for static and dynamic MFM models as well as the DPM model.

First, we discuss the influence of the hyperparameters β and κ. Therefore, Figure 3.3 shows

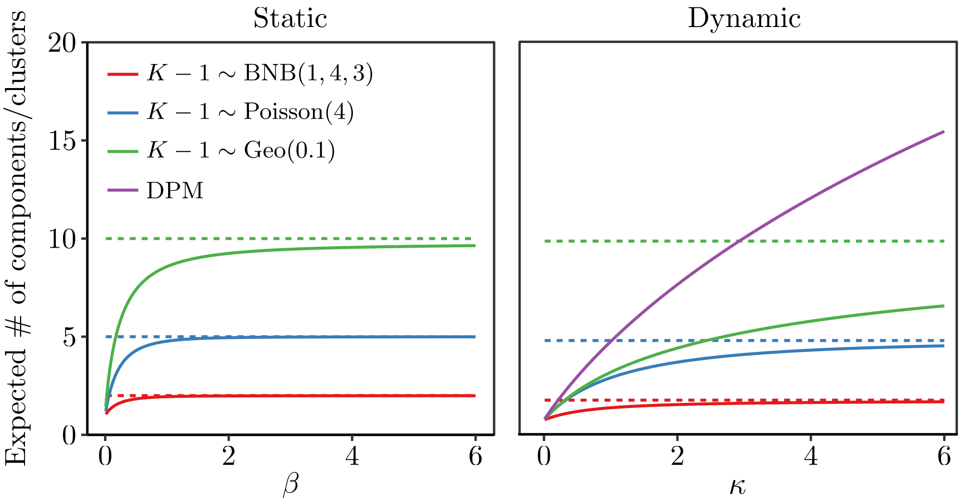


Figure 3.3: Prior expectations of the number of clusters for a static MFM (left) and a dynamic MFM in comparison to a DPM (right) as a function of β or κ under the priors on the number of components $K-1 \sim \text{BNB}(1,4,3)$, $K-1 \sim \text{Poisson}(4)$, $K-1 \sim \text{Geo}(0.1)$ for $N=100$. The corresponding prior expectations of K are illustrated as dashed horizontal lines. (Adapted from [20].)

the prior expectations of the number of clusters $\text{E}^{(p_{\text{stat}}(L;N,\beta))}\{L\}$ for a static MFM model as a function of β using several different priors $p(K)$ and $N=100$ observations. Additionally, Figure 3.3 shows the expectiation $\text{E}^{(p_{\text{dyn}}(L;N,\kappa))}\{L\}$ for a dynamic MFM model as a function of κ using several different priors $p(K)$, including the $p(K)$ which results in the DPM model, and $N=100$ observations. For comparison, the corresponding prior expectations of the number of components $\text{E}^{(p(K))}\{K\}$ are shown in Figure 3.3 as well. We observe that the gap between the prior expectation of K and L decreases with increasing hyperparameter β or κ for both types of MFMs. While $\text{E}^{(p_{\text{stat}}(L;N,\beta))}\{L\}$ approaches $\text{E}^{(p(K))}\{K\}$ rather quickly, there is a substantial gap left between $\text{E}^{(p_{\text{dyn}}(L;N,\kappa))}\{L\}$ and $\text{E}^{(p(K))}\{K\}$, even for larger values of κ. This is a direct consequence of choosing $\beta(K)=\kappa/K$, which prevents a quick growth of the number of clusters as the number of components increases. We conclude that for a static MFM, the prior $p(K)$ has a strong influence on the prior on the number of clusters for nearly all values of β. In contrast, for the dynamic MFM, the influence of $p(K)$ is reduced over an extended range of κ values.

For drawing comparisons between mixture models, the joint pmf $p(N'_1,\dots,N'_L|L;N,\beta(\cdot))$ of the labeled cluster sizes for a given number of clusters, also referred to as *conditional EPPF*, is very useful [32]. It can be expressed through the joint distribution of the labeled cluster sizes $p(N'_1,\dots,N'_L|N,\beta(K))$ and the prior distribution of the number of clusters $p(L;N,\beta(\cdot))$:

$$p(N'_1,\dots,N'_L|L;N,\beta(\cdot)) = \frac{p(N'_1,\dots,N'_L;N,\beta(\cdot))}{p(L;N,\beta(\cdot))}. \tag{3.52}$$

For a DPM, $p(N'_1,\dots,N'_L;N,\beta(\cdot))$ can be obtained from (3.44) by placing all prior mass at

$K = \infty$:

$$p_{\mathrm{DPM}}(N'_1, \ldots, N'_L; N, \kappa) = \frac{N!}{L!} \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} \prod_{l=1}^{L} \frac{\Gamma(N'_l)}{\Gamma(N'_l + 1)} = \frac{N!}{L!} \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} \prod_{l=1}^{L} \frac{1}{N'_l}. \tag{3.53}$$

Inserting (3.53) and (3.51) into (3.52) obtains the conditional EPPF for a DPM, i.e.,

$$p_{\mathrm{DPM}}(N'_1, \ldots, N'_L | L; N) = \frac{\frac{N!}{L!} \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} \prod_{l=1}^{L} \frac{1}{N'_l}}{\frac{N!}{L!} \frac{\kappa^L \Gamma(\kappa)}{\Gamma(\kappa + N)} C_{N,L}^{(\infty)}} = \frac{1}{C_{N,L}^{(\infty)}} \prod_{l=1}^{L} \frac{1}{N'_l}, \tag{3.54}$$

which is inversely proportional to the cluster sizes $N'_1, \ldots, N'_L$ and thus favouring a partition structure with many small clusters and a few large ones. Furthermore, the conditional EPPF in (3.54) is independent of the concentration parameter κ. Hence, it can not be made more flexible.

For a static MFM, $p(N'_1, \ldots, N'_L; N, \beta(\cdot))$ is obtained from (3.25) with $\beta(K) = \beta$:

$$p_{\mathrm{stat}}(N'_1, \ldots, N'_L; N, \beta) = \frac{N!}{L!} \frac{V_{N,L}^{(\beta)}}{\Gamma(\beta)^L} \prod_{l=1}^{L} \frac{\Gamma(N'_l + \beta)}{\Gamma(N'_l + 1)}, \tag{3.55}$$

where $V_{N,L}^{(\beta)}$ is given by (3.34). Inserting (3.55) and (3.49) into (3.52) obtains the conditional EPPF for a static MFM, i.e.,

$$p_{\mathrm{stat}}(N'_1, \ldots, N'_L | L; N, \beta) = \frac{1}{C_{N,L}^{(\beta)}} \prod_{l=1}^{L} \frac{\Gamma(N'_l + \beta)}{\Gamma(N'_l + 1)}. \tag{3.56}$$

In contrast to the DPM, the conditional EPPF of a static MFM depends on the hyperparameter (in this case, β), which leads to a more flexible partitioning structure: For $\beta = 1$, the uniform distribution over all partitions of N observations into L clusters is obtained. For values of $\beta > 1$ partitions where the clusters are of generally equal size are favored, whereas values of $\beta < 1$ favor partitions consisting of only a few large clusters and many small ones. We note that this behaviour is directly related to the symmetric Dirichlet prior on the mixture weights $\boldsymbol{\pi}$ (cf. Section 3.1).

Finally, the conditional EPPF for a dynamic MFM is obtained by inserting (3.44) and (3.50) into (3.52), i.e.,

$$p_{\mathrm{dyn}}(N'_1, \ldots, N'_L | L; N, \kappa) = \frac{\sum_{K=L}^{\infty} p(K) \prod_{l=1}^{L} \frac{\Gamma(N'_l + \frac{\kappa}{K})(K - l + 1)}{\Gamma(1 + \frac{\kappa}{K})\Gamma(N'_l + 1)K}}{\sum_{K=L}^{\infty} p(K) C_{N,L}^{(K,\kappa)} \prod_{l=1}^{L} \frac{K - l + 1}{\Gamma(1 + \frac{\kappa}{K})K}}. \tag{3.57}$$

We conclude that, of the three conditional EPPFs (3.54), (3.56), and (3.57), the one for the dynamic MFM model (i.e., (3.57)) is the most flexible since it depends on both the hyperparameter κ and the prior pmf $p(K)$.

3.5 Equivalent Representations of Static MFMs

In this section, we present several equivalent representations of the static MFM model based on [19]. One of these representations, namely the stick-breaking representation, is investigated in more depth since it builds the basis of our VI algorithm developed in Chapter 4.

3.5.1 Representation Using Latent Indicator Variables

We start with the representation using latent indicator variables which can directly be derived from the generalized MFM model given in (3.1) by using a fixed value for the hyperparameter $\beta(K)$, i.e., $\beta(K) = \beta$. The static MFM model using latent indicator variables is defined as follows:

$$
\begin{aligned}
&K \sim p(K), && \text{(3.58a)}\\
&\boldsymbol{\pi}|K;\beta \sim f(\boldsymbol{\pi}|K;\beta) = \mathcal{D}\big(\boldsymbol{\pi} = (\pi_1 \;\; \cdots \;\; \pi_K)^{\mathrm{T}}; \boldsymbol{\beta} = \beta \mathbf{1}_K\big), && \text{(3.58b)}\\
&z_1,\ldots,z_N|\boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} p(z_n|\boldsymbol{\pi}) = \mathcal{C}(z_n;\boldsymbol{\pi}), && \text{(3.58c)}\\
&\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_K^*|K \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}_k^*), && \text{(3.58d)}\\
&\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n \sim f(\boldsymbol{x}_n|\boldsymbol{\theta}^*, z_n) = f\big(\boldsymbol{x}_n|\boldsymbol{\theta}_{z_n}^*\big) \quad \text{independently for} \quad n = 1,\ldots,N, && \text{(3.58e)}
\end{aligned}
$$

with conditional independence assumptions as in (3.2).

3.5.2 Representation Using the EPPF

Recall the random partition $\mathcal{C} = \{\mathcal{C}_1,\ldots,\mathcal{C}_L\}$ of the N observations $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N$ induced by an MFM model through the latent indicator variables $\boldsymbol{z} = (z_1 \;\; \cdots \;\; z_N)^{\mathrm{T}}$. Exploiting the corresponding EPPF $p_{\text{stat}}(\mathcal{C};N,\beta)$ for a static MFM given by (3.33), we find the following equivalent representation to the model given in (3.58):

$$
\begin{aligned}
&\mathcal{C} \sim p_{\text{stat}}(\mathcal{C};N,\beta),\\
&\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_L^*|\mathcal{C} \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}_l^*),\\
&\boldsymbol{x}_n|\boldsymbol{\theta}^*,\mathcal{C} \sim f(\boldsymbol{x}_n|\boldsymbol{\theta}_l^*) \quad \text{for} \quad n \in \mathcal{C}_l \quad \text{and} \quad n = 1\ldots,N,
\end{aligned}
\tag{3.59}
$$

where the L cluster parameters $\boldsymbol{\theta}_l^*$ are arranged according to $\boldsymbol{\theta}^* = \big(\boldsymbol{\theta}_l^{*\mathrm{T}} \;\; \cdots \;\; \boldsymbol{\theta}_L^{*\mathrm{T}}\big)^{\mathrm{T}}$. Once a partition $\mathcal{C}$ is determined, one does not have to deal with component labels or empty components, which makes this representation of the static MFM particularly useful for doing inference.

The representation of the static MFM in (3.59) enables the formulation of a restaurant process, which falls under the general category of urn schemes considered in [33]. Given the partition $\mathcal{C} = \{\mathcal{C}_1,\ldots,\mathcal{C}_L\}$ of N observations, a new observation is placed in an existing cluster $\mathcal{C}_l$, for

$l = 1, \dots, L$, with probability proportional to $N'_l + \beta$ or is placed in a new cluster $\mathcal{C}_{L+1}$ with probability proportional to $\beta V^{(\beta)}_{N+1,L+1}/V^{(\beta)}_{N+1,L}$. The corresponding partition generating process can be formulated as follows:

- Assign the first observation $\boldsymbol{x}_1$ to cluster one.
- For $n = 2, 3, \dots,$ assign observation $\boldsymbol{x}_n$ to
  - an existing cluster $\mathcal{C}_l \in \{\mathcal{C}_1, \dots, \mathcal{C}_L\}$ with probability proportional to $N'_l + \beta$
  - a new cluster $\mathcal{C}_{L+1} \notin \{\mathcal{C}_1, \dots, \mathcal{C}_L\}$ with probability proportional to $\frac{V^{(\beta)}_{n,L+1}}{V^{(\beta)}_{n,L}}\beta$.

This process is closely related to the Chinese restaurant process for DPMs [34], where the nth observation is either placed in an existing cluster with probability proportional to N'_l or is placed in a new cluster with probability proportional to the corresponding concentration parameter κ.

3.5.3 Latent Mixing Distribution

Recall the latent mixing distribution $G(\cdot)$ introduced in Section 2.4. With K, $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*_1, \dots, \boldsymbol{\theta}^*_K$ as in (3.58), we define realizations of $G(\cdot)$ to be of the form (cf. (2.13))

$$G(\boldsymbol{\theta}_n) = \sum_{k=1}^{K} \pi_k \delta(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*_k). \tag{3.60}$$

Furthermore, we denote the distribution of $G(\cdot)$ by $\mathcal{M}(G; p(K), f(\boldsymbol{\theta}^*_k), \beta)$. As already mentioned, BMMs with latent indicator variables and the latent mixing distribution can be used interchangeably. Since the static MFM model defined in (3.58) includes latent indicator variables, we have the following equivalent representation:

$$\begin{aligned}
&G(\cdot) \sim \mathcal{M}(G; p(K), f(\boldsymbol{\theta}^*_k), \beta),\\
&\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N | G \overset{\text{i.i.d.}}{\sim} G(\boldsymbol{\theta}_n),\\
&\boldsymbol{x}_n | \boldsymbol{\theta}_n \sim f(\boldsymbol{x}_n | \boldsymbol{\theta}_n) \quad \text{independently for} \quad n = 1, \dots, N.
\end{aligned}$$

According to (3.60), $\boldsymbol{\theta}_n$ is equal to $\boldsymbol{\theta}^*_k$ with probability π_k. Hence, we observe $L \leq N$ distinct values $\boldsymbol{\theta}^*_1, \dots, \boldsymbol{\theta}^*_L$ for N samples $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ from $G(\cdot)$. The next sample $\boldsymbol{\theta}_{N+1}$ either belongs to the set of existing cluster parameters $\boldsymbol{\theta}_n \in \{\boldsymbol{\theta}^*_1, \dots, \boldsymbol{\theta}^*_L\}$ or takes on a new value from a new cluster parameter $\boldsymbol{\theta}^*_{L+1}$ sampled from the prior pdf $f(\boldsymbol{\theta}^*_k)$. The corresponding conditional distribution of $\boldsymbol{\theta}_{N+1}$ given $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$, i.e., the predictive distribution, can be formulated as follows:

$$f(\boldsymbol{\theta}_{N+1} | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N) \propto \frac{V^{(\beta)}_{N+1,L+1}}{V^{(\beta)}_{N+1,L}} \beta f(\boldsymbol{\theta}_{N+1}) + \sum_{l=1}^{L} (N'_l + \beta) \delta(\boldsymbol{\theta}_{N+1} - \boldsymbol{\theta}^*_l), \tag{3.61}$$

where N'_l is given by

$$N'_l = \sum_{n=1}^{N} \mathbb{1}(\boldsymbol{\theta}_n = \boldsymbol{\theta}^*_l).$$

Note the close relation of (3.61) to the partition generating process in Section 3.5.2.

For comparison, the predictive distribution for a DPM [16] can be expressed as

$$f(\boldsymbol{\theta}_{N+1}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \propto \kappa f(\boldsymbol{\theta}_{N+1}) + \sum_{l=1}^{L}(N'_l + \beta)\delta(\boldsymbol{\theta}_{N+1} - \boldsymbol{\theta}^*_l),$$

where the mixing distribution $G(\cdot)$ is distributed according to the Dirichlet process with the prior distribution of the component parameters $f(\boldsymbol{\theta}^*_k)$ and concentration parameter κ, i.e., $G(\cdot) \sim \mathrm{DP}(G; f(\boldsymbol{\theta}^*_k), \kappa)$.

3.5.4 Stick-Breaking Representation

Recall that the mixture weights $\boldsymbol{\pi} = (\pi_1 \ \cdots \ \pi_K)^{\mathrm{T}}$ exist in the $(K-1)$-dimensional probability simplex Δ_K defined in (2.2). Due to the constraint that $\sum_{k=1}^{K} \pi_k = 1$, a realization of $\boldsymbol{\pi}$ can conceptually be generated by breaking off random portions from a unit-length stick. This so-called *stick-breaking analogy* is commonly used to represent various kinds of priors for mixture models, such as the Dirichlet process [35] or the beta process [36] and enables the development of efficient inference algorithms.

In a certain special case — namely, when $p(K) = \mathrm{Poisson}(K-1; \alpha)$ and $\beta = 1$ — the static MFM model given in (3.58) also has an interesting representation that can be described using the stick-breaking analogy [19], which permits the development of efficient inference algorithms for static MFM models. The underlying procedure can be described as follows: Take a unit-length stick and break off pieces whose sizes are i.i.d. according to an exponential distribution with rate parameter $\alpha > 0$ until the stick is entirely depleted, i.e., all the pieces sum up to one. We summarize the MFM model for N conditionally independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ using the stick-breaking analogy as

$$v_1, v_2, \ldots \overset{\text{i.i.d.}}{\sim} \mathcal{E}(v_k; \alpha) \tag{3.62a}$$

$$\widetilde{K} = \min_j \left\{ j : \sum_{k=1}^{j} v_k \geq 1 \right\}, \tag{3.62b}$$

$$\widetilde{\pi}_k = v_k \quad \text{for} \quad k = 1, \ldots, \widetilde{K} - 1, \tag{3.62c}$$

$$\widetilde{\pi}_{\widetilde{K}} = 1 - \sum_{k=1}^{\widetilde{K}-1} \widetilde{\pi}_k, \tag{3.62d}$$

$$\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_{\widetilde{K}} | \widetilde{K} \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\theta}^*_k), \tag{3.62e}$$

$$z_1, \ldots, z_N | \widetilde{\boldsymbol{\pi}} \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n; \widetilde{\boldsymbol{\pi}}), \tag{3.62f}$$

$$\boldsymbol{x}_n \sim f(\boldsymbol{x}_n | \boldsymbol{\theta}^*_{z_n}) \quad \text{independently for} \quad n = 1, \ldots, N, \tag{3.62g}$$

where $\widetilde{\boldsymbol{\pi}} = (\widetilde{\pi}_1 \ \cdots \ \widetilde{\pi}_{\widetilde{K}})^{\mathrm{T}}$. Note that the random variables v_k, for $k = 1, \ldots, \widetilde{K} - 1$, in this model definition should not be confused with the coefficients V introduced in Section 3.2.

Proposition 3.6: The stick lengths $\widetilde{\boldsymbol{\pi}}$ and the mixture weights $\boldsymbol{\pi}$ in the static MFM model (3.58) have the same distribution, i.e., the symmetric Dirichlet distribution with hyperparameter $\boldsymbol{\beta} = \beta \mathbf{1}_K$, when $p(K) = \text{Poisson}(K-1; \alpha)$ and $\beta = 1$.

Since the stick-breaking representation restricts the prior pmf $p(K)$ to the Poisson family of distributions and the hyperparameter $\boldsymbol{\beta} = \mathbf{1}_K$, it in turn may restrict the static MFM model in terms of flexibility. However, the induced priors on the number of components K and the mixture weights $\boldsymbol{\pi}$ are still commonly used or are favorable in certain scenarios. For example, the choice of a Poisson prior on the number of components K with rate parameter $\alpha = 1$, i.e., $p(K) = \text{Poisson}(K-1; \alpha = 1)$, is proposed and justified in [37]. To investigate the influence of the parameter α on the stick-breaking process, let us first consider the distribution of the stick lengths v_k, for $k = 1, \ldots, \widetilde{K} - 1$, given by $\mathcal{E}(v_k; \alpha)$ (cf. (3.62a)). In contrast to the beta distribution, which is used in the stick-breaking representation of DPMs, the exponential distribution is supported on the interval $[0, \infty)$. Depending on the parameter α, a considerable amount of probability mass can lie in the interval $(1, \infty)$ and therefore, the MFM model may a priori be restricted to a single or small number of components. From Figure 3.4, it is demonstrated that the probability of observing a larger number of components grows with growing rate parameter α. Since $K - 1 \sim \text{Poisson}\,(K-1; \alpha)$, we have $\mathrm{E}\{K-1\} = \alpha$. Therefore, the expected value of the total number of components K is given by $\mathrm{E}\{K\} = \alpha + 1$.

According to the static MFM model in (3.58), we assume a symmetric Dirichlet prior for the mixture weights $\boldsymbol{\pi}$, i.e.,

$$f(\boldsymbol{\pi} | K; \beta) = \mathcal{D}(\boldsymbol{\pi}; \beta \mathbf{1}_K). \tag{3.63}$$

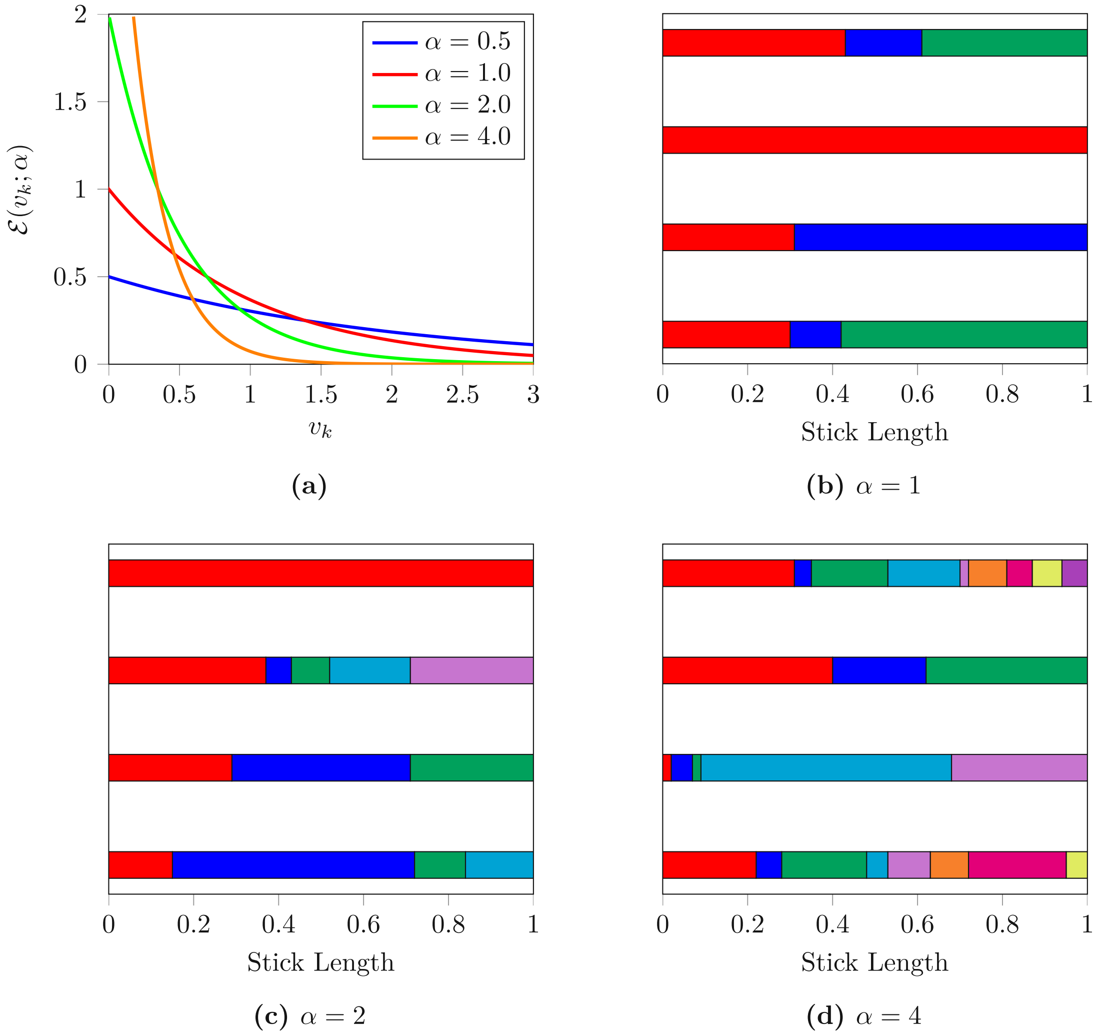
The Dirichlet distribution in (3.63) is given by

$$f(\boldsymbol{\pi} | K; \beta) = \mathcal{D}(\boldsymbol{\pi}; \beta \mathbf{1}_K) = \frac{\Gamma\left(\sum_{k=1}^{K} \beta\right)}{\prod_{k=1}^{K} \Gamma(\beta)} \prod_{k=1}^{K} \pi_k^{\beta - 1}, \tag{3.64}$$

where the random vector $\boldsymbol{\pi} = (\pi_1 \ \cdots \ \pi_K)^{\mathrm{T}}$ with mean vector

$$\mathrm{E}^{(f(\boldsymbol{\pi}|K;\beta))}\{\boldsymbol{\pi}\} = \frac{\beta \mathbf{1}_K}{\sum_{k=1}^{K} \beta} = \frac{1}{K} \mathbf{1}_K \tag{3.65}$$

exists in the $(K-1)$-dimensional probability simplex Δ_K given by (2.2). As stated in Proposition 3.6, the stick-breaking representation of the static MFM model restricts the prior on



(a)

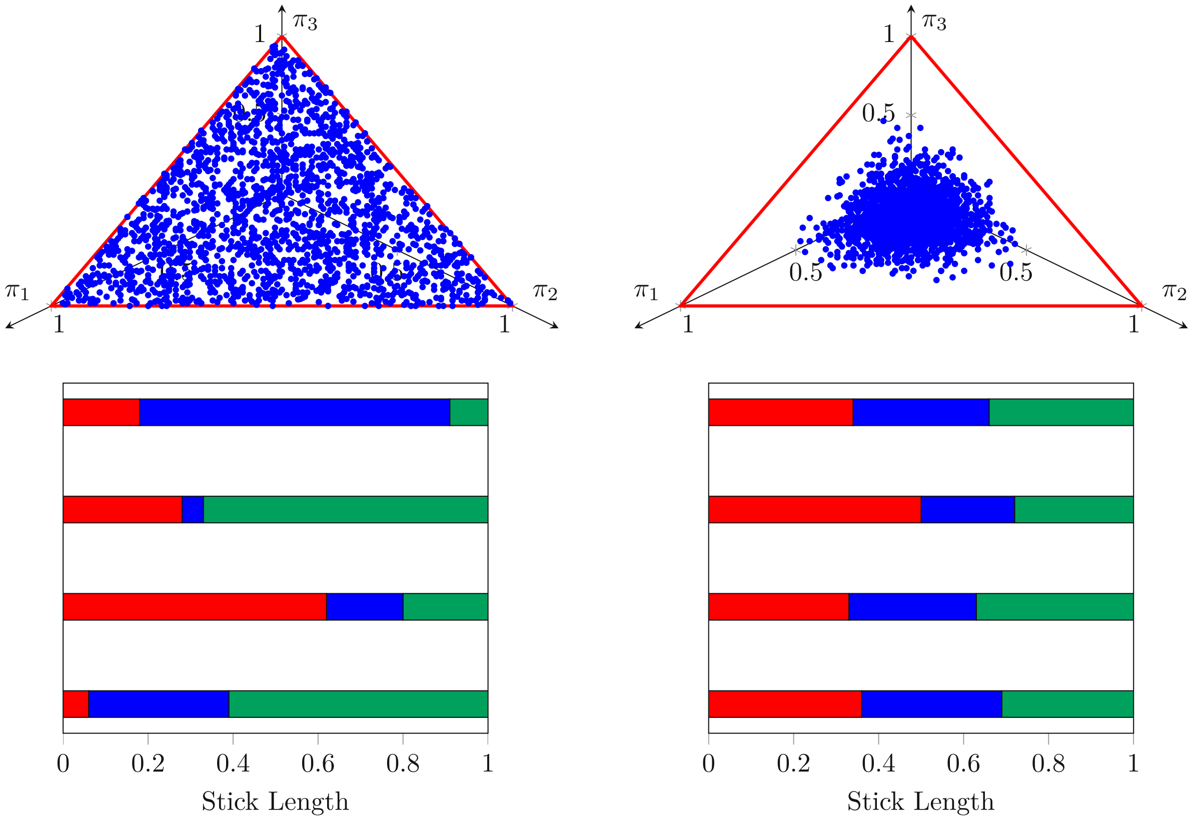
(b) $\alpha = 1$

(c) $\alpha = 2$

(d) $\alpha = 4$

Figure 3.4: (a) Exponential distribution for four different rate parameters α. The smaller α, the more probability mass lies in the interval $(1, \infty)$. (b)-(d) Visualization of the stick-breaking process. In each plot, four realizations of the mixture weights $\widetilde{\boldsymbol{\pi}}$ according to (3.62a)–(3.62d) are shown. Larger values of α means that more components are expected.

the mixture weights to the case where $\beta = 1$. Figure 3.5 shows realizations of the symmetric Dirichlet distribution for the case where $K = 3$ and $\beta \in \{1, 10\}$. We observe that none of the individual mixture weights is dominating, meaning that the symmetric Dirichlet prior favors the components of the mixture model equally without introducing any additional bias towards specific components. This gives rise to the exchangeability property of the symmetric Dirichlet distribution. With growing β, the probability mass is more and more concentrated at the center of the support region, i.e., the mean of the symmetric Dirichlet distribution given by (3.65), resulting in a single peak at the center for $\beta \to \infty$. For the special case where $\beta = 1$, the Dirichlet distribution is uniform over its support, which is also known as the flat Dirichlet distribution. Therefore, the static MFM model obtained using the stick-breaking representation



(a) $\beta = 1$ **(b)** $\beta = 10$

Figure 3.5: Visualization of realizations of the symmetric Dirichlet distribution $\mathcal{D}(\boldsymbol{\pi}; \beta\mathbf{1}_K)$ for $K = 3$. Top: 2000 realizations of the random vector $\boldsymbol{\pi} = (\pi_1 \quad \pi_2 \quad \pi_3)^{\mathrm{T}}$ per plot illustrated as blue dots existing in the $(K-1)$-dimensional probability simplex Δ_K indicated by the red triangle. Bottom: Four realizations of $\boldsymbol{\pi}$ per plot illustrated as breaking a unit-length stick into $K = 3$ pieces.

(3.62) uses a weakly-informative prior on the mixture weights $\boldsymbol{\pi}$. Choosing this particular prior is especially useful in scenarios where there is only weak prior information about the mixture weights available.

Chapter 4

Variational Inference for Static Mixtures of Finite Mixtures

4.1 Introduction

Consider a model with latent random variables $\boldsymbol{w} = (w_1 \ \cdots \ w_P)^{\mathrm{T}} \in \mathbb{R}^P$ and observations $\boldsymbol{x} = (\boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{x}_n \in \mathbb{R}^M$. The posterior pdf $f(\boldsymbol{w}|\boldsymbol{x})$ is of special importance for Bayesian estimation, since it summarizes all the information that $\boldsymbol{x}$ contains about $\boldsymbol{w}$. Thus, various estimators can be obtained from it. The posterior pdf can be written as

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

where $f(\boldsymbol{w}, \boldsymbol{x})$ is the joint pdf of $\boldsymbol{w}$ and $\boldsymbol{x}$ and $f(\boldsymbol{x})$ is called the *evidence*. The evidence can be calculated by marginalizing out all of the latent variables from the joint density $f(\boldsymbol{w}, \boldsymbol{x})$, i.e.,

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

For many models, this evidence integral is unavailable in closed form or requires exponential time to compute [28]. One way to circumvent this problem is to use sampling techniques such as Markov chain Monte Carlo (MCMC) to approximate the posterior pdf. An alternative to sampling techniques is given by variational inference (VI). VI methods are computationally efficient and thus usually much faster than sampling techniques. Although the accuracy of VI methods is lower compared with sampling techniques, they can be favorable in high-dimensional scenarios, settings with large datasets or very complex models. When approximating a posterior density using VI, the resulting methodology is called variational Bayes (VB).

The objective of VB is to find the best approximation to the posterior distribution from a computationally more tractable class of distributions of the latent variables. The best approximation is selected from the class of approximating distributions by minimizing a discrepancy called a *divergence* between the posterior distribution of interest and the member of this class of approximating distributions. The most popular choices for the discrepancy and the approximating class of distributions are the Kullback–Leibler divergence (KLD) and the class of product distributions called the mean-field family of distributions, respectively. This combination is popularly known as mean-field VI, originating from mean-field theory in physics [38]. Mean-field VI has percolated through a wide variety of disciplines, including statistical mechanics, electrical engineering, information theory, neuroscience, cognitive sciences, and deep neural networks. While computing the KLD is intractable for a large class of distributions, reframing the minimization problem for maximizing the evidence lower bound (ELBO) leads to efficient algorithms. In particular, for (conditionally) conjugate exponential family models, the optimal distribution for mean-field VI can be computed by the iteration of closed-form updates. These updates form a coordinate-ascent algorithm known as coordinate-ascent variational inference (CAVI) [39].

Let $\mathcal{F}$ denote the mean-field family of distributions over the latent variables, i.e.,

$$\mathcal{F} := \left\{ q(\boldsymbol{w}) = \prod_{j=1}^{P} q_j(w_j), \text{with } q_j(w_j) \in \mathcal{F}_j \right\}. \tag{4.3}$$

Here, $\mathcal{F}_j$ is a subset of all possible probability distributions for a random variable w_j. Each distribution $q(\boldsymbol{w}) \in \mathcal{F}$ is called a *variational* distribution, which approximates the posterior $f(\boldsymbol{w}|\boldsymbol{x})$, and the marginal pdf $q_j(w_j) \in \mathcal{F}_j$ is referred to as *variational factor* distribution. We search for the optimal $q(\boldsymbol{w})$, which maximizes the ELBO, i.e.,

$$q^*(\boldsymbol{w}) = \underset{q(\boldsymbol{w}) \in \mathcal{F}}{\arg\max}\, \mathcal{L}(q; \boldsymbol{x}). \tag{4.4}$$

The ELBO is defined as

$$\mathcal{L}(q; \boldsymbol{x}) = \mathrm{E}^{(q(\boldsymbol{w}))}\left\{ \ln \frac{f(\boldsymbol{w}, \boldsymbol{x})}{q(\boldsymbol{w})} \right\} = \int_{\boldsymbol{w} \in \mathbb{R}^P} q(\boldsymbol{w}) \ln \frac{f(\boldsymbol{w}, \boldsymbol{x})}{q(\boldsymbol{w})}\, d\boldsymbol{w}, \tag{4.5}$$

i.e., the expected value obtained by taking the expectation with respect to the variational distribution $q(\boldsymbol{w})$ of the logarithm of the joint pdf $f(\boldsymbol{w}, \boldsymbol{x})$ divided by the variational distribution $q(\boldsymbol{w})$ itself. To solve the optimization problem (4.4), the CAVI algorithm iteratively optimizes each $q_j(w_j)$ while keeping $q_i(w_i)$ for $i \neq j$ fixed. Thus, in the ℓth iteration step of the iterative algorithm, the jth substep (with $j \in \{1, \ldots, P\}$) updates the previous iterate $q_j^{(\ell-1)}(w_j)$ by solving the optimization problem

$$q_j^{(\ell)}(w_j) = \underset{q_j(w_j) \in \mathcal{F}_j}{\arg\max}\, \mathcal{L}(q; \boldsymbol{x}). \tag{4.6}$$

Here, the variational distribution $q^{(\ell,j)}(\boldsymbol{w})$ used in $\mathcal{L}(q^{(\ell,j)}; \boldsymbol{x})$ is, according to (4.3), given by

$$q^{(\ell,j)}(\boldsymbol{w}) = \left(\prod_{i=1}^{j} q_i^{(\ell)}(w_i) \right) \prod_{i=j+1}^{P} q_i^{(\ell-1)}(w_i),$$

where all currently fixed variational factor distributions $q_i(w_i)$ (for $i \neq j$) are equal to the results of the most recent respective updates either calculated in the previous iteration step $\ell - 1$ or in a substep of the current iteration step ℓ [40]. As stated in [28], the solution of the optimization problem (4.6) is proportional to the exponentiated expected log of the complete conditional of w_j, i.e.,

$$q_j^{(\ell)}(w_j) \propto \exp\big(\mathrm{E}^{(q_{\sim j}(\boldsymbol{w}_{\sim j}))}\{\ln f(w_j|\boldsymbol{w}_{\sim j}, \boldsymbol{x})\}\big), \tag{4.7}$$

where $\propto$ denotes equality up to a constant normalization factor and we denote by $\boldsymbol{w}_{\sim j}$ the vector of latent variables $\boldsymbol{w}$ with the jth variable removed, i.e., $\boldsymbol{w}_{\sim j} = (w_1 \ \cdots \ w_{j-1} \ w_{j+1} \ \cdots \ w_P)^{\mathrm{T}}$. The expectation in (4.7) is with respect to the currently fixed variational density $q_{\sim j}(\boldsymbol{w}_{\sim j}) = \prod_{i \neq j} q_i(w_i)$. With $q_j^*(w_j)$ given by (4.7) for each substep $j = 1, \ldots, P$ of the very last iteration step, i.e., when the ELBO converges, the resulting optimal variational distribution $q^*(\boldsymbol{w})$ can be written as

$$q^*(\boldsymbol{w}) = \prod_{j=1}^{P} q_j^*(w_j), \tag{4.8}$$

and thus is a member of $\mathcal{F}$. In summary, the CAVI algorithm described above is presented as Algorithm 1.

Algorithm 1: General formulation of CAVI

```
Input: Observations $\boldsymbol{x}$ and number of factor distributions $P$
Output: Variational factor distributions $q_j^*(w_j)$, for $j = 1, \ldots, P$
Initialize: Variational factor distributions $q_j^{(0)}(w_j)$, for $j = 1, \ldots, P$;
while the ELBO has not converged do
    $\ell = \ell + 1$
    for $j$ from 1 to $P$ do
        update $q_j^{(\ell)}(w_j)$ according to (4.7)
    Compute the ELBO $\mathcal{L}(q^{(\ell)}; \boldsymbol{x})$ according to (4.5)
return $q^*(\boldsymbol{w}) = \prod_{j=1}^{P} q_j^*(w_j)$
```

4.2 Conjugate Exponential Family Model for Static Mixtures of Finite Mixtures

For the remainder of the thesis, we will consider the static MFM model in its stick-breaking representation given in (3.62). We restrict ourselves to the case where the components are exponential family distributions and $f(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda})$ is the corresponding conjugate prior. For N conditionally independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ given $\boldsymbol{\eta}_{z_n}^*$, the conjugate exponential family MFM model can be summarized as

$$v_1, v_2, \ldots \overset{\text{i.i.d.}}{\sim} \mathcal{E}(v_k; \alpha) \tag{4.9a}$$

$$K = \min_j \left\{ j : \sum_{k=1}^{j} v_k \geq 1 \right\}, \tag{4.9b}$$

$$\pi_k = v_k \quad \text{for} \quad k = 1, \dots, K-1, \tag{4.9c}$$

$$\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k, \tag{4.9d}$$

$$\boldsymbol{\eta}_1^*, \dots, \boldsymbol{\eta}_K^* | K \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda}), \tag{4.9e}$$

$$z_1, \dots, z_N | \boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n; \boldsymbol{\pi}), \tag{4.9f}$$

$$\boldsymbol{x}_n \sim f(\boldsymbol{x}_n | \boldsymbol{\eta}_{z_n}^*) \quad \text{independently for } n = 1, \dots, N. \tag{4.9g}$$

The corresponding (conditional) independence relations are given by

$$v_k \perp\!\!\!\perp v_{k'} \quad \text{for all} \quad k \neq k' = 1, 2, \dots, \tag{4.10a}$$

$$\boldsymbol{\eta}_k^* \perp\!\!\!\perp \boldsymbol{\eta}_{k'}^* \,|\, K \quad \text{for all} \quad k \neq k' = 1, \dots, K, \tag{4.10b}$$

$$\boldsymbol{\eta}_k^* \perp\!\!\!\perp \boldsymbol{v} \,|\, K \quad \text{for all} \quad k = 1, \dots, K, \tag{4.10c}$$

$$\boldsymbol{\eta}_k^* \perp\!\!\!\perp z_n \,|\, K, \boldsymbol{v} \quad \text{for all} \quad k = 1, \dots, K \quad \text{and} \quad n = 1, \dots, N, \tag{4.10d}$$

$$z_n \perp\!\!\!\perp K \,|\, \boldsymbol{v} \quad \text{for all} \quad n = 1, \dots, N, \tag{4.10e}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{v}, \boldsymbol{x}_{n'}, z_{n'}, \boldsymbol{\eta}_{k'}^* \,|\, z_n, \boldsymbol{\eta}_{z_n}^* \qquad n \neq n' = 1, \dots, N \quad \text{and} \quad k \neq z_n, \tag{4.10f}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp K \,|\, \boldsymbol{\eta}^*, \boldsymbol{v}, \tag{4.10g}$$

where $\boldsymbol{v} = (v_1 \ \cdots \ v_K)^{\mathrm{T}}$ and $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^{*\mathrm{T}} \ \cdots \ \boldsymbol{\eta}_K^{*\mathrm{T}})^{\mathrm{T}}$.

In the conjugate exponential family MFM model given in (4.9), the component distributions $f(\boldsymbol{x}_n | \boldsymbol{\eta}_{z_n}^*)$ are of natural exponential family form, i.e.,

$$f(\boldsymbol{x}_n | \boldsymbol{\eta}_{z_n}^*) = f(\boldsymbol{x}_n | z_n, \boldsymbol{\eta}^*) = \prod_{k=1}^{K} \Big(h(\boldsymbol{x}_n) \exp(\boldsymbol{\eta}_k^{*\mathrm{T}} \boldsymbol{x}_n - a(\boldsymbol{\eta}_k^*)) \Big)^{\mathbb{1}(z_n = k)}, \tag{4.11}$$

where $a(\cdot)$ is the log-partition function, and we assume for simplicity that $\boldsymbol{x}_n$ is the sufficient statistic for the natural component parameter $\boldsymbol{\eta}_k^*$. The prior distribution of the natural component parameters $f(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda})$ is a member of the corresponding conjugate family

$$f(\boldsymbol{\eta}_k^*; \boldsymbol{\lambda}) = b(\boldsymbol{\lambda}) \exp(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_k^* - \lambda_2 a(\boldsymbol{\eta}_k^*)), \tag{4.12}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^{\mathrm{T}} \quad \lambda_2)^{\mathrm{T}}$ is the hyperparameter. It consists of a vector $\boldsymbol{\lambda}_1 \in \mathbb{R}^M$ and a scalar $\lambda_2 \in \mathbb{R}$. The normalization constant $b(\boldsymbol{\lambda})$ is given by

$$b(\boldsymbol{\lambda}) = \left(\int_{\boldsymbol{\eta}_k^* \in \mathbb{R}^M} \exp(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_k^* - \lambda_2 a(\boldsymbol{\eta}_k^*)) \, d\boldsymbol{\eta}_k^* \right)^{-1} \tag{4.13}$$

and depends on the hyperparameter $\boldsymbol{\lambda}$ [40]. Furthermore, we assume $\boldsymbol{\lambda}$ as well as α to be deterministic. A graphical summary of the considered model is presented in Figure 4.1.

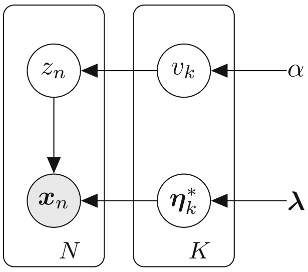


Figure 4.1: Bayesian network representing the conjugate exponential family MFM model given in (4.9).

4.3 CAVI Algorithm

In this section, we derive the CAVI algorithm for static MFMs being represented via the stick-breaking analogy (4.9). Although the VI framework has been applied to various kinds of mixture models such as finite mixtures or DPMs for example, there is, at least to our best knowlege, no literature about VI in context to MFMs. Due to the close relation to the stick-breaking representation of DPMs, our algorithm is inspired by the VB method derived for DPMs in [16].

4.3.1 Truncated Mean-Field Approximation

We now define the mean-field variational family $\mathcal{F}$ for approximating the posterior distribution $f(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}|\boldsymbol{x})$. In the considered representation (4.9), the latent variables are the auxiliary variables $\boldsymbol{v} = (v_1 \;\; \cdots \;\; v_K)^{\mathrm{T}}$, the natural component parameters $\boldsymbol{\eta}^* = \left(\boldsymbol{\eta}_1^{*\mathrm{T}} \;\; \cdots \;\; \boldsymbol{\eta}_K^{*\mathrm{T}}\right)^{\mathrm{T}}$ with $\boldsymbol{\eta}_k^* \in \mathbb{R}^M$ and the indicator variables $\boldsymbol{z} = (z_1 \;\; \cdots \;\; z_N)^{\mathrm{T}}$ and thus, $\boldsymbol{w} = \left(\boldsymbol{v}^{\mathrm{T}} \;\; \boldsymbol{\eta}^{*\mathrm{T}} \;\; \boldsymbol{z}^{\mathrm{T}}\right)^{\mathrm{T}}$. Hence, the variational distribution under the mean-field assumption (cf. (4.3)) is given by

$$q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) = q_{\boldsymbol{\gamma}}(\boldsymbol{v}) q_{\boldsymbol{\tau}}(\boldsymbol{\eta}^*) q_{\boldsymbol{\phi}}(\boldsymbol{z}) = \left(\prod_{k=1}^{K} q_{\boldsymbol{\gamma}_k}(v_k)\right)\left(\prod_{k=1}^{K} q_{\boldsymbol{\tau}_k}(\boldsymbol{\eta}_k^*)\right)\left(\prod_{n=1}^{N} q_{\boldsymbol{\phi}_n}(z_n)\right). \tag{4.14}$$

Due to the random number of components K involved in the first two products in (4.14), we heavily doubt the existence of a closed-form solution for the corresponding CAVI updates. To circumvent this problem, K is exchanged with a deterministic parameter T, which can be freely set. In the context of DPMs, T is referred to as *truncation parameter*. This leads to the truncated mean-field approximation for the static MFM model:

$$q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) = q_{\boldsymbol{\gamma}}(\boldsymbol{v}) q_{\boldsymbol{\tau}}(\boldsymbol{\eta}^*) q_{\boldsymbol{\phi}}(\boldsymbol{z}) = \left(\prod_{t=1}^{T} q_{\boldsymbol{\gamma}_t}(v_t)\right)\left(\prod_{t=1}^{T} q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)\right)\left(\prod_{n=1}^{N} q_{\boldsymbol{\phi}_n}(z_n)\right), \tag{4.15}$$

where $\boldsymbol{v} = (v_1 \;\; \cdots \;\; v_T)^{\mathrm{T}}$ and $\boldsymbol{\eta}^* = \left(\boldsymbol{\eta}_1^{*\mathrm{T}} \;\; \cdots \;\; \boldsymbol{\eta}_T^{*\mathrm{T}}\right)^{\mathrm{T}}$. Note that the subscripts in (4.15) depict the variational parameters, i.e., the hyperparameters of the corresponding variational distributions $q(\cdot)$.

We will show that the iterative optimization process (4.7) leads to closed-form updates of the variational parameters vectors $\boldsymbol{\gamma}_t = (\gamma_{t,1} \quad \gamma_{t,2})^{\mathrm{T}}$, $\boldsymbol{\tau}_t = (\boldsymbol{\tau}_{t,1}^{\mathrm{T}} \;\; \tau_{t,2})^{\mathrm{T}}$ and $\boldsymbol{\phi}_n = (\phi_{n,1} \quad \cdots \quad \phi_{n,T})^{\mathrm{T}}$ of the corresponding variational factor distributions. Note that applying the truncation parameter T is a major simplification. However, we only consider the variational distribution $q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z})$ to be truncated; there is no such constraint on the full model. Thus, the algorithm will still obtain an approximation of the full stick-breaking representation described by (4.9). The approximating model can be summarized as

$$v_t \overset{\text{i.i.d.}}{\sim} \mathcal{E}(v_t;\alpha) \quad \text{for} \quad t = 1,\ldots,T, \tag{4.16a}$$

$$\pi_t = v_t \quad \text{for} \quad t = 1,\ldots,T, \tag{4.16b}$$

$$\boldsymbol{\eta}_1^*,\ldots,\boldsymbol{\eta}_T^* \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda}), \tag{4.16c}$$

$$z_1,\ldots,z_N|\boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n;\boldsymbol{\pi}), \tag{4.16d}$$

$$\boldsymbol{x}_n \sim f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*) \quad \text{independently for } n = 1,\ldots,N, \tag{4.16e}$$

with

$$f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*) = \prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*))\Big)^{\mathbb{1}(z_n=t)} \tag{4.17}$$

and

$$f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda}) = b(\boldsymbol{\lambda})\exp(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)). \tag{4.18}$$

Here, we denote by $f^{(T)}(\cdot)$ the truncated version of the corresponding pdf $f(\cdot)$ in (4.9). Model (4.16) implies (conditional) independencies which we summarize by

$$v_t \perp\!\!\!\perp v_{t'} \quad \text{for all} \quad t \neq t' = 1,\ldots,T, \tag{4.19a}$$

$$\boldsymbol{\eta}_t^* \perp\!\!\!\perp \boldsymbol{\eta}_{t'}^* \quad \text{for all} \quad t \neq t' = 1,\ldots,T, \tag{4.19b}$$

$$\boldsymbol{\eta}_t^* \perp\!\!\!\perp \boldsymbol{v} \quad \text{for all} \quad t = 1,\ldots,T, \tag{4.19c}$$

$$\boldsymbol{\eta}_t^* \perp\!\!\!\perp z_n \,|\, \boldsymbol{v} \quad \text{for all} \quad t = 1,\ldots,T \quad \text{and} \quad n = 1,\ldots,N, \tag{4.19d}$$

$$\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{v}, \boldsymbol{x}_{n'}, z_{n'}, \boldsymbol{\eta}_t^* \,|\, z_n, \boldsymbol{\eta}_{z_n}^* \qquad n \neq n' = 1,\ldots,N \quad \text{and} \quad t \neq z_n. \tag{4.19e}$$

4.3.2 Derivation of the CAVI Updates

We next formulate the CAVI algorithm for static MFMs by deriving the updates for the variational factor distributions $q_{\boldsymbol{\gamma}_t}(v_t)$, $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)$ and $q_{\boldsymbol{\phi}_n}(z_n)$ in (4.15) using the approximating model given in (4.16). For the sake of brevity, the iteration index ℓ as well as the superscript (T) is omitted throughout the derivations.

As a preparatory step, we work out the expectation $\mathrm{E}^{(q_{\boldsymbol{\phi}_n}(z_n))}\{\mathbb{1}(z_n = t)\}$ which will show up frequently throughout the derivation of the CAVI updates and the ELBO. Since the indicator variable z_n is a discrete random variable and, according to (4.17), $z_n \in \{1, \ldots, T\}$, we have

$$\mathrm{E}^{(q_{\boldsymbol{\phi}_n}(z_n))}\{\mathbb{1}(z_n = t)\} = \sum_{z_n=1}^{T} \mathbb{1}(z_n = t) q_{\boldsymbol{\phi}_n}(z_n). \tag{4.20}$$

Because of the indicator function, the sum in (4.20) reduces to a single term, namely the one where $z_n = t$. With $\phi_{n,t} := q_{\boldsymbol{\phi}_n}(z_n = t)$, we obtain

$$\mathrm{E}^{(q_{\boldsymbol{\phi}_n}(z_n))}\{\mathbb{1}(z_n = t)\} = q_{\boldsymbol{\phi}_n}(z_n = t) = \phi_{n,t}. \tag{4.21}$$

CAVI Update for $q_{\boldsymbol{\gamma}_t}(v_t)$

Applying (4.7) to $q_{\boldsymbol{\gamma}_t}(v_t)$, the updated variational factor pdfs of the auxiliary variables are given by

$$q_{\boldsymbol{\gamma}_t}(v_t) \propto \exp\Big(\mathrm{E}^{(q(\boldsymbol{v}_{\sim t}, \boldsymbol{\eta}^*, \boldsymbol{z}))}\{\ln f(v_t | \boldsymbol{v}_{\sim t}, \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x})\}\Big). \tag{4.22}$$

For convenience, we work with the log of (4.22), i.e.,

$$\ln q_{\boldsymbol{\gamma}_t}(v_t) \overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v}_{\sim t}, \boldsymbol{\eta}^*, \boldsymbol{z}))}\{\ln f(v_t | \boldsymbol{v}_{\sim t}, \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x})\}, \tag{4.23}$$

where the symbol $\overset{\mathrm{c}}{=}$ denotes equality up to an additive constant. Due to the statistical independence of the auxiliary variables (cf. (4.19a)), the complete conditional can be written as

$$f(v_t | \boldsymbol{v}_{\sim t}, \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) = f(v_t | \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}). \tag{4.24}$$

First, let us consider the joint conditional pdf $f(\boldsymbol{v} | \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x})$. Applying Bayes' law yields

$$f(\boldsymbol{v} | \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) = \frac{f(\boldsymbol{x} | \boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) f(\boldsymbol{v} | \boldsymbol{\eta}^*, \boldsymbol{z})}{f(\boldsymbol{x} | \boldsymbol{\eta}^*, \boldsymbol{z})}. \tag{4.25}$$

To simplify this expression, we exploit independencies among the variables under the model (4.16). As implied in (4.19e), the observations $\boldsymbol{x}$ are conditionally independent of $\boldsymbol{v}$ given $\boldsymbol{\eta}^*$ and $\boldsymbol{z}$, so

$$f(\boldsymbol{x} | \boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) = f(\boldsymbol{x} | \boldsymbol{\eta}^*, \boldsymbol{z}). \tag{4.26}$$

The auxiliary variables $\boldsymbol{v}$ are statistically independent of $\boldsymbol{\eta}^*$ (cf. (4.19c)) and thus, we have

$$f(\boldsymbol{v} | \boldsymbol{\eta}^*, \boldsymbol{z}) = f(\boldsymbol{v} | \boldsymbol{z}). \tag{4.27}$$

Inserting (4.26) and (4.27) into (4.25) leads to

$$f(\boldsymbol{v} | \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) = \frac{f(\boldsymbol{x} | \boldsymbol{\eta}^*, \boldsymbol{z}) f(\boldsymbol{v} | \boldsymbol{z})}{f(\boldsymbol{x} | \boldsymbol{\eta}^*, \boldsymbol{z})} = f(\boldsymbol{v} | \boldsymbol{z}). \tag{4.28}$$

Since we do not know $f(\boldsymbol{v} | \boldsymbol{z})$, we apply Bayes' law and obtain for (4.28)

$$f(\boldsymbol{v} | \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) = f(\boldsymbol{v} | \boldsymbol{z}) = \frac{p(\boldsymbol{z} | \boldsymbol{v}) f(\boldsymbol{v}; \alpha)}{p(\boldsymbol{z})}. \tag{4.29}$$

Keeping in mind that we search for the variational factor pdf $q_{\gamma_t}(v_t)$, the denominator $p(\boldsymbol{z})$ in (4.29) is a normalization constant, since it does not depend on v_t. Thus, $p(\boldsymbol{z})$ can be omitted in the optimization process and (4.29) becomes

$$f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) \propto p(\boldsymbol{z}|\boldsymbol{v}) f(\boldsymbol{v}; \alpha). \tag{4.30}$$

In our model (cf. (4.16a)) the auxiliary variables $v_1, \dots, v_T$ are i.i.d. according to the exponential distribution

$$f(v_t; \alpha) = \alpha e^{-\alpha v_t}. \tag{4.31}$$

Thus, the random vector $\boldsymbol{v}$ is distributed according to the joint pdf

$$f(\boldsymbol{v}; \alpha) = \prod_{t=1}^{T} f(v_t; \alpha) = \prod_{t=1}^{T} \alpha e^{-\alpha v_t} \tag{4.32}$$

According to (4.16d), the indicator variables z_n, for $n = 1, \dots, N$, given $\boldsymbol{v}$, are independent and identically distributed according to the categorical distribution

$$p(z_n|\boldsymbol{v}) = \prod_{t=1}^{T} v_t^{\mathbb{1}(z_n=t)}. \tag{4.33}$$

Thus, the random vector $\boldsymbol{z}$ is distributed according to the joint pmf

$$p(\boldsymbol{z}|\boldsymbol{v}) = \prod_{n=1}^{N} p(z_n|\boldsymbol{v}) \tag{4.34}$$

$$= \prod_{n=1}^{N} \prod_{t=1}^{T} v_t^{\mathbb{1}(z_n=t)}. \tag{4.35}$$

Inserting (4.32) and (4.35) into (4.30) yields

$$f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) \propto \left(\prod_{n=1}^{N} \prod_{t=1}^{T} v_t^{\mathbb{1}(z_n=t)} \right) \prod_{t=1}^{T} \alpha e^{-\alpha v_t} = \prod_{t=1}^{T} \alpha e^{-\alpha v_t} \prod_{n=1}^{N} v_t^{\mathbb{1}(z_n=t)}. \tag{4.36}$$

According to (4.24), the complete conditional of v_t is the marginal pdf of $f(\boldsymbol{v}|\boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x})$. Using the statistical independence of the auxiliary variables $\boldsymbol{v}$ once again, it can easily be retrieved from (4.36) by omitting the product with respect to t:

$$f(v_t|\boldsymbol{v}_{\sim t}, \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) \propto \alpha e^{-\alpha v_t} \prod_{n=1}^{N} v_t^{\mathbb{1}(z_n=t)}. \tag{4.37}$$

Next, we apply the log to (4.37), which yields

$$\begin{aligned} \ln f(v_t|\boldsymbol{v}_{\sim t}, \boldsymbol{\eta}^*, \boldsymbol{z}, \boldsymbol{x}) &\overset{\text{c}}{=} \ln\left(\alpha e^{-\alpha v_t} \prod_{n=1}^{N} v_t^{\mathbb{1}(z_n=t)} \right) = \ln \alpha + \ln e^{-\alpha v_t} + \sum_{n=1}^{N} \ln v_t^{\mathbb{1}(z_n=t)} \\ &= \ln \alpha - \alpha v_t + \sum_{n=1}^{N} \mathbb{1}(z_n = t) \ln v_t. \end{aligned} \tag{4.38}$$

Inserting (4.38) into (4.23) leads to

$$\begin{aligned}\ln q_{\gamma_t}(v_t) &\overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v}_{\sim t},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\ln\alpha - \alpha v_t + \sum_{n=1}^{N} \mathbb{1}(z_n = t)\ln v_t\right\}\\ &= \mathrm{E}^{(q(\boldsymbol{v}_{\sim t},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln\alpha - \alpha v_t\} + \mathrm{E}^{(q(\boldsymbol{v}_{\sim t},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{n=1}^{N} \mathbb{1}(z_n = t)\ln v_t\right\}\\ &= \ln\alpha - \alpha v_t + \mathrm{E}^{(q(\boldsymbol{v}_{\sim t},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{n=1}^{N} \mathbb{1}(z_n = t)\ln v_t\right\}. \qquad (4.39)\end{aligned}$$

In the last step, we used the fact that $\ln\alpha$ and αv_t are constants with respect to the expectation, i.e.,

$$\begin{aligned}\mathrm{E}^{(q(\boldsymbol{v}_{\sim t},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln\alpha - \alpha v_t\} &= \sum_{\boldsymbol{z}\in\mathbb{N}^N} \int_{\boldsymbol{\eta}^*\in\mathbb{R}^{TM}} \int_{\boldsymbol{v}_{\sim t}\in\mathbb{R}^{T-1}} q(\boldsymbol{v}_{\sim t})q(\boldsymbol{\eta}^*)q(\boldsymbol{z})(\ln\alpha - \alpha v_t)\, d\boldsymbol{v}_{\sim t}\, d\boldsymbol{\eta}^*\\ &= (\ln\alpha - \alpha v_t) \sum_{\boldsymbol{z}\in\mathbb{N}^N} \int_{\boldsymbol{\eta}^*\in\mathbb{R}^{TM}} \underbrace{\left(\int_{\boldsymbol{v}_{\sim t}\in\mathbb{R}^{T-1}} q(\boldsymbol{v}_{\sim t})\, d\boldsymbol{v}_{\sim t}\right)}_{1} q(\boldsymbol{\eta}^*)q(\boldsymbol{z})\, d\boldsymbol{\eta}^*\\ &= (\ln\alpha - \alpha v_t) \sum_{\boldsymbol{z}\in\mathbb{N}^N} \underbrace{\left(\int_{\boldsymbol{\eta}^*\in\mathbb{R}^{TM}} q(\boldsymbol{\eta}^*)\, d\boldsymbol{\eta}^*\right)}_{1} q(\boldsymbol{z})\\ &= (\ln\alpha - \alpha v_t) \sum_{\boldsymbol{z}\in\mathbb{N}^N} q(\boldsymbol{z})\\ &= \ln\alpha - \alpha v_t.\end{aligned}$$

It remains to work out the last term in (4.39), i.e., $\mathrm{E}^{(q(\boldsymbol{v}_{\sim t},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{n=1}^{N} \mathbb{1}(z_n = t)\ln v_t\right\}$. Since $\boldsymbol{v}_{\sim t}$ and $\boldsymbol{\eta}^*$ are not present in the sum, the expectation is only with respect to $q(\boldsymbol{z})$. Thus, we obtain

$$\begin{aligned}\mathrm{E}^{(q(\boldsymbol{v}_{\sim t},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{n=1}^{N} \mathbb{1}(z_n = t)\ln v_t\right\} &= \mathrm{E}^{(q(\boldsymbol{z}))}\left\{\sum_{n=1}^{N} \mathbb{1}(z_n = t)\ln v_t\right\}\\ &= \sum_{n=1}^{N} \mathrm{E}^{(q_{\phi_n}(z_n))}\{\mathbb{1}(z_n = t)\}\ln v_t\\ &= \sum_{n=1}^{N} \phi_{n,t}\ln v_t, \qquad (4.40)\end{aligned}$$

where we used (4.21) in the last step. Inserting (4.40) into (4.39) yields

$$\ln q_{\gamma_t}(v_t) \overset{\mathrm{c}}{=} \ln\alpha - \alpha v_t + \sum_{n=1}^{N} \phi_{n,t}\ln v_t. \qquad (4.41)$$

To find the updated variational factor pdf $q_{\gamma_t}(v_t)$, we omit the additive constant $\ln\alpha$ and develop (4.41) further as

$$q_{\gamma_t}(v_t) \propto \exp\left(-\alpha v_t + \sum_{n=1}^{N} \phi_{n,t}\ln v_t\right)$$

$$
\begin{aligned}
&= \exp(-\alpha v_t) \exp(\ln v_t)^{\sum_{n=1}^{N} \phi_{n,t}} \\
&= e^{-\alpha v_t} v_t^{\sum_{n=1}^{N} \phi_{n,t}}.
\end{aligned} \tag{4.42}
$$

Let us consider a gamma distribution with shape parameter $\tilde{\gamma}_{t,1}$ and rate parameter $\gamma_{t,2}$, i.e.,

$$
f(v_t; \tilde{\gamma}_{t,1}, \gamma_{t,2}) = \mathcal{G}(v_t; \tilde{\gamma}_{t,1}, \gamma_{t,2}) = \frac{1}{\Gamma(\tilde{\gamma}_{t,1})} \gamma_{t,2}^{\tilde{\gamma}_{t,1}} v_t^{\tilde{\gamma}_{t,1}-1} e^{-\gamma_{t,2} v_t}, \tag{4.43}
$$

where $\Gamma(\cdot)$ is the gamma function. Omitting the constant factor $\frac{1}{\Gamma(\tilde{\gamma}_{t,1})} \gamma_{t,2}^{\tilde{\gamma}_{t,1}}$ in (4.43) leads to

$$
f(v_t; \tilde{\gamma}_{t,1}, \gamma_{t,2}) \propto e^{-\gamma_{t,2} v_t} v_t^{\tilde{\gamma}_{t,1}-1}. \tag{4.44}
$$

A comparison of (4.44) and (4.42) results in

$$
\tilde{\gamma}_{t,1} = 1 + \sum_{n=1}^{N} \phi_{n,t} \tag{4.45}
$$

and

$$
\gamma_{t,2} = \alpha. \tag{4.46}
$$

We conclude that the variational factor pdf $q_{\boldsymbol{\gamma}_t}(v_t)$ is a gamma distribution with variational parameters $\tilde{\gamma}_{t,1}$ and $\gamma_{t,2}$, i.e.,

$$
q_{\boldsymbol{\gamma}_t}(v_t) = \frac{1}{\Gamma(\tilde{\gamma}_{t,1})} \gamma_{t,2}^{\tilde{\gamma}_{t,1}} v_t^{\tilde{\gamma}_{t,1}-1} e^{-\gamma_{t,2} v_t}. \tag{4.47}
$$

At first glance, it seems that the variational factor pdf $q_{\boldsymbol{\gamma}_t}(v_t)$, i.e., the approximated marginal posterior, does not have the exact same functional form as the corresponding prior $f(v_t; \alpha)$ given by (4.31). In fact, the exponential distribution is a special case of the gamma distribution, i.e., $\mathcal{E}(v_t; \alpha) = \mathcal{G}(v_t; \tilde{\alpha} = 1, \alpha)$. This relation can easily verified by setting $\tilde{\gamma}_{t,1} = 1$ and $\gamma_{t,2} = \alpha$ in (4.47) and comparing the result with (4.31). Thus, we can equivalently use the gamma prior $\mathcal{G}(v_t; \tilde{\alpha} = 1, \alpha)$ instead of the exponential distribution $\mathcal{E}(v_t; \alpha)$ in (4.16a). Note that, according to (4.45) and (4.46), the CAVI algorithm updates the shape parameter ($\tilde{\alpha} = 1$) of the corresponding prior pdf only, while the rate parameter α is fixed.

In the approximating model (4.16), we did not explicitly restrict the mixture weight π_T to be of the remaining portion of a unit-length stick (cf. (4.16b)), whereas in the full model (4.9) we have $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$. We therefore normalize the variational factor pdfs $q_{\boldsymbol{\gamma}_t}(v_t)$, for $t = 1, \ldots, T$, such that the auxiliary variables v_t — and thus the mixture weights π_t as well — sum up to one with respect to their posterior mean, i.e.,

$$
\sum_{t=1}^{T} \mathrm{E}^{(q_{\boldsymbol{\gamma}_t}(v_t))}\{v_t\} = \sum_{t=1}^{T} \frac{\gamma_{t,1}}{\gamma_{t,2}} = 1.
$$

Here, the normalized variational parameter $\gamma_{t,1}$, for $t = 1, \ldots, T$ is given by

$$
\gamma_{t,1} = \frac{\tilde{\gamma}_{t,1}}{\sum_{t=1}^{T} \frac{\tilde{\gamma}_{t,1}}{\gamma_{t,2}}} \tag{4.48}
$$

with $\tilde{\gamma}_{t,1}$ according to (4.45).

CAVI Update for $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)$

Applying the log of (4.7) to $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)$, the CAVI solution is given by

$$\begin{aligned} \ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*) &\overset{c}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*_{\sim t},\boldsymbol{z}))}\{\ln f(\boldsymbol{\eta}_t^*|\boldsymbol{\eta}^*_{\sim t},\boldsymbol{v},\boldsymbol{z},\boldsymbol{x},\alpha,\boldsymbol{\lambda})\} \\ &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*_{\sim t},\boldsymbol{z}))}\{\ln f(\boldsymbol{\eta}_t^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x},\alpha,\boldsymbol{\lambda})\}, \end{aligned} \tag{4.49}$$

where (4.19b) has been exploited in the last step. To find an expression for the complete conditional $f(\boldsymbol{\eta}_t^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x})$, our starting point is the joint conditional pdf $f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x})$. By means of Bayes' law, we have

$$f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z})f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z})}{f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{z})}. \tag{4.50}$$

Next, we exploit independencies among the variables under the approximating model (4.16): According to (4.16c), the component parameters $\boldsymbol{\eta}^*$ follow the family of distributions $f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda})$ given by (4.18). They are independent of the auxiliary variables $\boldsymbol{v}$ (cf. (4.19c)) and conditionally independent of the indicator variables $\boldsymbol{z}$ given $\boldsymbol{v}$ (cf. (4.19d)). Thus, we have

$$f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z}) = f(\boldsymbol{\eta}^*;\boldsymbol{\lambda}). \tag{4.51}$$

Inserting (4.51) and (4.26) into (4.50) and omitting the normalization constant $f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{z})$ yields

$$f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})f(\boldsymbol{\eta}^*;\boldsymbol{\lambda}). \tag{4.52}$$

In our truncated model the component parameters $\boldsymbol{\eta}_t^*$, for $t = 1,\ldots,T$, are independent and identically distributed according to the family of distributions

$$f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda}) = b(\boldsymbol{\lambda})\exp\big(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big). \tag{4.53}$$

Thus, the random vector $\boldsymbol{\eta}^*$ is distributed according to the joint pdf

$$f(\boldsymbol{\eta}^*;\boldsymbol{\lambda}) = \prod_{t=1}^{T} f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda}) = \prod_{t=1}^{T} b(\boldsymbol{\lambda})\exp\big(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big). \tag{4.54}$$

The vector $\boldsymbol{z}$ is a local parameter vector in the sense that its nth element z_n influences the nth observation $\boldsymbol{x}_n$ only, i.e.,

$$f(\boldsymbol{x}_n|\boldsymbol{\eta}^*,\boldsymbol{z}) = f(\boldsymbol{x}_n|\boldsymbol{\eta}^*,z_n). \tag{4.55}$$

Furthermore, we assume that, given the component parameters $\boldsymbol{\eta}^*$ and the indicator variable z_n, an observation $\boldsymbol{x}_n$ is conditionally independent of an observation $\boldsymbol{x}_{n'}$ and, due to locality of $\boldsymbol{z}$, as well conditionally independent of an indicator variable $z_{n'}$ for $n \neq n' = 1,\ldots,N$, see (4.19e). Thus, the joint conditional density $f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})$ factorizes as

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z}) &= \prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\eta}^*,z_n) && (4.56) \\ &= \prod_{n=1}^{N}\prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp\big(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\Big)^{\mathbb{1}(z_n=t)}, && (4.57) \end{aligned}$$

where we used (4.17) in the last step. Next, we insert (4.54) and (4.57) into (4.52) which leads to

$$\begin{aligned} f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x},\alpha,\boldsymbol{\lambda}) &\propto \left(\prod_{n=1}^{N}\prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp\big(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\Big)^{\mathbb{1}(z_n=t)}\right)\prod_{t=1}^{T} b(\boldsymbol{\lambda})\exp\big(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big) \\ &= \prod_{t=1}^{T} b(\boldsymbol{\lambda})\exp\big(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big)\prod_{n=1}^{N}\Big(h(\boldsymbol{x}_n)\exp\big(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\Big)^{\mathbb{1}(z_n=t)}. \end{aligned} \tag{4.58}$$

The desired complete conditional $f(\boldsymbol{\eta}_t^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x})$ can be obtained from the joint density $f(\boldsymbol{\eta}^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x})$ by marginalizing out all of the component parameters $\boldsymbol{\eta}_{\sim t}^*$. Due to the statistical independence, this marginalization is done by omitting the product with respect to t in (4.58), i.e.,

$$f(\boldsymbol{\eta}_t^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x}) \propto b(\boldsymbol{\lambda})\exp\big(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big)\prod_{n=1}^{N}\Big(h(\boldsymbol{x}_n)\exp\big(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\Big)^{\mathbb{1}(z_n=t)}. \tag{4.59}$$

To solve the update equation (4.49), we next apply the log to the complete conditional given by (4.59):

$$\begin{aligned} \ln f(\boldsymbol{\eta}_t^*|\boldsymbol{v},\boldsymbol{z},\boldsymbol{x}) &\overset{\mathrm{c}}{=} \ln\Big(b(\boldsymbol{\lambda})\exp\big(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big)\Big) + \ln\left(\prod_{n=1}^{N}\Big(h(\boldsymbol{x}_n)\exp\big(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\Big)^{\mathbb{1}(z_n=t)}\right) \\ &= \ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N}\mathbb{1}(z_n=t)\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big). \end{aligned} \tag{4.60}$$

Inserting (4.60) into (4.49) gives

$$\begin{aligned} \ln q_{\boldsymbol{\tau}_t}^*(\boldsymbol{\eta}_t^*) &\overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}_{\sim t}^*,\boldsymbol{z}))}\left\{\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N}\mathbb{1}(z_n=t)\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\right\} \\ &= \underbrace{\mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}_{\sim t}^*,\boldsymbol{z}))}\left\{\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\right\}}_{A} \\ &\quad + \underbrace{\mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}_{\sim t}^*,\boldsymbol{z}))}\left\{\sum_{n=1}^{N}\mathbb{1}(z_n=t)\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\right\}}_{B}. \end{aligned} \tag{4.61}$$

We will now develop A and B. For A, we have

$$\begin{aligned} A &= \sum_{\boldsymbol{z}\in\mathbb{N}^N}\int_{\boldsymbol{\eta}_{\sim t}^*\in\mathbb{R}^{(T-1)M}}\int_{\boldsymbol{v}\in\mathbb{R}^T} q(\boldsymbol{v})q(\boldsymbol{\eta}_{\sim t}^*)q(\boldsymbol{z})\big(\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big)\,d\boldsymbol{v}\,d\boldsymbol{\eta}_{\sim t}^* \\ &= \big(\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big)\sum_{\boldsymbol{z}\in\mathbb{N}^N}\int_{\boldsymbol{\eta}_{\sim t}^*\in\mathbb{R}^{(T-1)M}}\underbrace{\left(\int_{\boldsymbol{v}\in\mathbb{R}^T} q(\boldsymbol{v})\,d\boldsymbol{v}\right)}_{1} q(\boldsymbol{\eta}_{\sim t}^*)q(\boldsymbol{z})\,d\boldsymbol{\eta}_{\sim t}^* \\ &= \big(\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big)\sum_{\boldsymbol{z}\in\mathbb{N}^N}\underbrace{\left(\int_{\boldsymbol{\eta}_{\sim t}^*\in\mathbb{R}^{(T-1)M}} q(\boldsymbol{\eta}_{\sim t}^*)\,d\boldsymbol{\eta}_{\sim t}^*\right)}_{1} q(\boldsymbol{z}) \\ &= \big(\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*)\big)\sum_{\boldsymbol{z}\in\mathbb{N}^N} q(\boldsymbol{z}) \end{aligned}$$

$$= \ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*). \tag{4.62}$$

With (4.21), we obtain for B

$$\begin{aligned} B &= \sum_{n=1}^{N} \mathrm{E}^{(q_{\phi_n}(z_n))}\{\mathbb{1}(z_n = t)\}\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big) \\ &= \sum_{n=1}^{N} \phi_{n,t}\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big). \end{aligned} \tag{4.63}$$

Inserting the expressions for A and B into (4.61) leads to

$$\ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*) \overset{\mathrm{c}}{=} \ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N} \phi_{n,t}\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big), \tag{4.64}$$

or equivalently

$$\begin{aligned} q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*) &\propto \exp\left(\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\eta}_t^* - \lambda_2 a(\boldsymbol{\eta}_t^*) + \sum_{n=1}^{N} \phi_{n,t}\left(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\right)\right) \\ &= b(\boldsymbol{\lambda})\exp\left(\sum_{n=1}^{N} \phi_{n,t} h(\boldsymbol{x}_n)\right)\exp\left(\left(\boldsymbol{\lambda}_1^{\mathrm{T}} + \sum_{n=1}^{N} \phi_{n,t} \boldsymbol{x}_n{}^{\mathrm{T}}\right)\boldsymbol{\eta}_t^* - \left(\lambda_2 + \sum_{n=1}^{N} \phi_{n,t}\right) a(\boldsymbol{\eta}_t^*)\right) \\ &\propto \exp\left(\left(\boldsymbol{\lambda}_1^{\mathrm{T}} + \sum_{n=1}^{N} \phi_{n,t} \boldsymbol{x}_n{}^{\mathrm{T}}\right)\boldsymbol{\eta}_t^* - \left(\lambda_2 + \sum_{n=1}^{N} \phi_{n,t}\right) a(\boldsymbol{\eta}_t^*)\right), \end{aligned} \tag{4.65}$$

where the constant factor $b(\boldsymbol{\lambda})\exp\left(\sum_{n=1}^{N} \phi_{n,t} h(\boldsymbol{x}_n)\right)$ has been omitted in the last step. Via comparing (4.65) and (4.53), we conclude that the variational factor pdf $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)$ takes the same functional form as the prior pdf $f(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda})$ given by (4.53) with variational parameter $\boldsymbol{\tau}_t = (\boldsymbol{\tau}_{t,1}^{\mathrm{T}} \;\; \tau_{t,2})^{\mathrm{T}}$, i.e.,

$$q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*) = b_t(\boldsymbol{\tau}_t)\exp\big(\boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\eta}_t^* - \tau_{t,2} a(\boldsymbol{\eta}_t^*)\big), \tag{4.66}$$

where $b_t(\boldsymbol{\tau}_t) = 1/\int_{\boldsymbol{\eta}_t^* \in \mathbb{R}^M} \exp\big(\boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\eta}_t^* - \tau_{t,2} a(\boldsymbol{\eta}_t^*)\big)\, d\boldsymbol{\eta}_t^*$. The CAVI update for the variational parameters is given by

$$\boldsymbol{\tau}_{t,1} = \boldsymbol{\lambda}_1 + \sum_{n=1}^{N} \phi_{n,t} \boldsymbol{x}_n, \tag{4.67}$$

$$\tau_{t,2} = \lambda_2 + \sum_{n=1}^{N} \phi_{n,t}. \tag{4.68}$$

CAVI Update for $q_{\boldsymbol{\phi}_n}(z_n)$

Finally, applying (4.7) to $q_{\boldsymbol{\phi}_n}(z_n)$, the updated variational factor distributions of the indicator variables z_n are given by

$$q_{\boldsymbol{\phi}_n}(z_n) \propto \exp\left(\mathrm{E}^{(q(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}_{\sim n}))}\{\ln f(z_n | \boldsymbol{z}_{\sim n}, \boldsymbol{\eta}^*, \boldsymbol{v}, \boldsymbol{x})\}\right). \tag{4.69}$$

Since it is more convenient to work with the log of (4.69), we consider

$$\ln q_{\phi_n}(z_n) \overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}_{\sim n}))}\{\ln f(z_n|\boldsymbol{z}_{\sim n},\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x})\}. \tag{4.70}$$

Our first task is to find an expression for the complete conditional $f(z_n|\boldsymbol{z}_{\sim n},\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x})$. We start with a simplification of the complete conditional, namely

$$f(z_n|\boldsymbol{z}_{\sim n},\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) = f(z_n|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}), \tag{4.71}$$

which is due to the conditional independence of the indicator variables $z_1,\ldots,z_N$ given $\boldsymbol{v}$ (or equivalently given $\boldsymbol{\pi}$) assumed in (4.16d). Next, we consider the joint conditional pdf $f(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x})$ and use Bayes' law, which leads to

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

The indicator variables $\boldsymbol{z}$ are conditionally independent of $\boldsymbol{\eta}^*$ given $\boldsymbol{v}$ (cf. (4.19d)). Thus, we obtain

$$p(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*) = p(\boldsymbol{z}|\boldsymbol{v}). \tag{4.73}$$

Inserting (4.26) and (4.73) into (4.72) yields

$$f(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{v}), \tag{4.74}$$

where we omitted the constant $f(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{\eta}^*)$ since it does not affect the optimization process. According to (4.35) and (4.57), the joint conditional pdf $f(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*)$ in (4.74) can be further developed as

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) &\propto \left(\prod_{n=1}^{N}\prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*))\Big)^{\mathbb{1}(z_n=t)}\right)\prod_{n=1}^{N}\prod_{t=1}^{T} v_t^{\mathbb{1}(z_n=t)} \\ &= \prod_{n=1}^{N}\prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*))v_t\Big)^{\mathbb{1}(z_n=t)}. \end{aligned} \tag{4.75}$$

The desired complete conditional $f(z_n|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x})$ can be obtained from the joint conditional pdf $f(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x})$ by marginalizing out all of the indicator variables except the nth, i.e., marginalizing out $\boldsymbol{z}_{\sim n}$. Due to the conditional independence of the indicator variables $z_1,\ldots,z_N$ given $\boldsymbol{v}$, this is simply done by omitting the product with respect to n in (4.75):

$$f(z_n|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) \propto \prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*))v_t\Big)^{\mathbb{1}(z_n=t)}. \tag{4.76}$$

Applying the log to (4.76) yields

$$\begin{aligned} \ln f(z_n|\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{x}) &\overset{\mathrm{c}}{=} \ln\left(\prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp(\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*))v_t\Big)^{\mathbb{1}(z_n=t)}\right) \\ &= \sum_{t=1}^{T}\mathbb{1}(z_n=t)\big(\ln h(\boldsymbol{x}_n) + \ln v_t + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big). \end{aligned} \tag{4.77}$$

Next, we insert (4.77) into the update equation (4.70), which leads to

$$\begin{aligned}\ln q_{\boldsymbol{\phi}_n}(z_n) &\overset{\mathrm{c}}{=} \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}_{\sim n}))}\left\{\sum_{t=1}^{T} \mathbb{1}(z_n = t)\big(\ln h(\boldsymbol{x}_n) + \ln v_t + {\boldsymbol{\eta}_t^*}^{\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\right\} \\ &= \sum_{t=1}^{T} \mathbb{1}(z_n = t)\big(\ln h(\boldsymbol{x}_n) + \mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} + \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{{\boldsymbol{\eta}_t^*}^{\mathrm{T}}\}\boldsymbol{x}_n - \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}\big).\end{aligned} \tag{4.78}$$

Furthermore, exponentiating (4.78), we obtain the desired variational factor pmf $q_{\boldsymbol{\phi}_n}(z_n)$:

$$\begin{aligned}q_{\boldsymbol{\phi}_n}(z_n) &\propto \exp\left(\sum_{t=1}^{T} \mathbb{1}(z_n = t)\big(\ln h(\boldsymbol{x}_n) + \mathrm{E}^{\left(q_{\gamma_t}\right)}\{\ln v_t\} + \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{{\boldsymbol{\eta}_t^*}^{\mathrm{T}}\}\boldsymbol{x}_n - \mathrm{E}^{\left(q_{\tau_t}\right)}\{a(\boldsymbol{\eta}_t^*)\}\big)\right) \\ &= \prod_{t=1}^{T}\Big(h(\boldsymbol{x}_n)\exp\Big(\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} + \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{{\boldsymbol{\eta}_t^*}^{\mathrm{T}}\}\boldsymbol{x}_n - \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}\Big)\Big)^{\mathbb{1}(z_n=t)}.\end{aligned} \tag{4.79}$$

Equation (4.79) already indicates, that $q_{\boldsymbol{\phi}_n}(z_n)$ takes the same functional form as the prior pdf $f(z_n|\boldsymbol{v})$. Thus, let us consider the categorical distribution

$$p(z_n; \boldsymbol{\phi}_n) = \prod_{t=1}^{T} \phi_{n,t}^{\mathbb{1}(z_n=t)}, \tag{4.80}$$

where $\boldsymbol{\phi}_n = (\phi_{n,1} \ \cdots \ \phi_{n,T})^{\mathrm{T}}$ is the updated variational parameter vector of interest. For convenience, we introduce the shorthand notation

$$S_{n,t} := \mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} + \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{{\boldsymbol{\eta}_t^*}^{\mathrm{T}}\}\boldsymbol{x}_n - \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}. \tag{4.81}$$

Inserting (4.81) into (4.79) gives us

$$q_{\boldsymbol{\phi}_n}(z_n) \propto \prod_{t=1}^{T}(h(\boldsymbol{x}_n)\exp(S_{n,t}))^{\mathbb{1}(z_n=t)} \propto \prod_{t=1}^{T}\exp(S_{n,t})^{\mathbb{1}(z_n=t)}, \tag{4.82}$$

where the factor $h(\boldsymbol{x}_n)$ has been omitted, since it constant with respect to t and thus will cancel out in the normalization step which will be mentioned shortly. Comparing (4.82) with the categorical distribution given by (4.80), we conclude that the updated variational parameter $\phi_{n,t}$ is proportional to $\exp(S_{n,t})$. To turn (4.82) into a valid categorical distribution, we normalize $\exp(S_{n,t})$, for all $t = 1, \ldots, T$, to have $\sum_{t=1}^{T}\exp(S_{n,t}) = 1$. Thus, the CAVI solution is given by

$$q_{\boldsymbol{\phi}_n}(z_n) = \mathcal{C}(z_n; \boldsymbol{\phi}_n) = \prod_{t=1}^{T} \phi_{n,t}^{\mathbb{1}(z_n=t)}, \tag{4.83}$$

where the updated variational parameter is the (posterior) probability of an observation $\boldsymbol{x}_n$ being assigned to the tth mixture component, i.e.,

$$\phi_{n,t} = \frac{\exp(S_{n,t})}{\sum_{i=1}^{T}\exp(S_{n,i})}, \tag{4.84}$$

which is often referred to as the *responsibility* of component t for observation $\boldsymbol{x}_n$.

4.3.3 Derivation of the Evidence Lower Bound

In terms of assessing convergence, the ELBO is a common choice since it can easily be computed and stored throughout the CAVI algorithm. Typically, the algorithm is terminated once the change in the ELBO has fallen below some small threshold.

According to (4.5), the ELBO is defined as

$$\begin{aligned}\mathcal{L}(q;\boldsymbol{x}) &= \mathrm{E}^{(q(\boldsymbol{w}))}\Big\{\ln f^{(T)}(\boldsymbol{w},\boldsymbol{x})\Big\} - \mathrm{E}^{(q(\boldsymbol{w}))}\{\ln q(\boldsymbol{w})\}\\ &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\Big\{\ln f^{(T)}(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},\boldsymbol{x})\Big\} - \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z})\}. \end{aligned} \tag{4.85}$$

For convenience only, we will use $\mathrm{E}^{(q)}\{\cdot\}$ instead of $\mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\cdot\}$ below. Let us first consider the joint distribution $f^{(T)}(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},\boldsymbol{x})$. Applying the chain rule to the joint distribution $f^{(T)}(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},\boldsymbol{x})$ leads to

$$\begin{aligned}f^{(T)}(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},\boldsymbol{x}) &= f^{(T)}(\boldsymbol{z},\boldsymbol{x}|\boldsymbol{v},\boldsymbol{\eta}^*)f^{(T)}(\boldsymbol{v},\boldsymbol{\eta}^*)\\ &= f^{(T)}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{v},\boldsymbol{\eta}^*)p^{(T)}(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{\eta}^*)f^{(T)}(\boldsymbol{v}|\boldsymbol{\eta}^*)f^{(T)}(\boldsymbol{\eta}^*)\\ &= f^{(T)}(\boldsymbol{x}|\boldsymbol{\eta}^*,\boldsymbol{z})p^{(T)}(\boldsymbol{z}|\boldsymbol{v})f^{(T)}(\boldsymbol{v};\alpha)f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda}),\end{aligned} \tag{4.86}$$

where we used (4.26), (4.73) and, due to the statistical independence of $\boldsymbol{v}$ and $\boldsymbol{\eta}^*$, $f^{(T)}(\boldsymbol{v}|\boldsymbol{\eta}^*) = f^{(T)}(\boldsymbol{v};\alpha)$ in the last step. Furthermore, we appended the hyperparameter $\boldsymbol{\lambda}$ to the argument of $f^{(T)}(\boldsymbol{\eta}^*)$ for the sake of completeness. Using (4.56) and (4.34), we can rewrite (4.86) as

$$f^{(T)}(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z},\boldsymbol{x}) = f^{(T)}(\boldsymbol{v};\alpha)f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda})\prod_{n=1}^{N} p^{(T)}(z_n|\boldsymbol{v})f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*,z_n). \tag{4.87}$$

Note that, given z_n, $\boldsymbol{x}_n$ is independent of $\boldsymbol{\eta}^*_{t'}$ for $t' \neq z_n$ (cf. (4.19e)) and therefore

$$f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*,z_n) = f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n}).$$

Inserting (4.87) into (4.85) gives us

$$\begin{aligned}\mathcal{L}(q;\boldsymbol{x}) &= \mathrm{E}^{(q)}\Bigg\{\ln f^{(T)}(\boldsymbol{v};\alpha) + \ln f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda}) + \sum_{n=1}^{N}\Big(\ln p^{(T)}(z_n|\boldsymbol{v}) + \ln f^{(T)}(x_n|\boldsymbol{\eta}^*,z_n)\Big)\Bigg\}\\ &\quad - \mathrm{E}^{(q)}\{\ln q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z})\}\\ &= \mathrm{E}^{(q)}\Bigg\{\ln f^{(T)}(\boldsymbol{v};\alpha) + \ln f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda}) + \sum_{n=1}^{N}\Big(\ln p^{(T)}(z_n|\boldsymbol{v}) + \ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})\Big)\Bigg\}\\ &\quad - \mathrm{E}^{(q)}\Bigg\{\sum_{t=1}^{T}(\ln q_{\gamma_t}(v_t) + \ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)) + \sum_{n=1}^{N}\ln q_{\boldsymbol{\phi}_n}(z_n)\Bigg\},\end{aligned} \tag{4.88}$$

where we used (4.15) in the last step. Exploiting the linearity of the expectation, we can rewrite

(4.88) as

$$\begin{aligned}\mathcal{L}(q;\boldsymbol{x}) = {} & \underbrace{\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{v};\alpha)\}}_{A} + \underbrace{\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda})\}}_{B} \\ & + \sum_{n=1}^{N}\Big(\underbrace{\mathrm{E}^{(q)}\{\ln p^{(T)}(z_n|\boldsymbol{v})\}}_{C} + \underbrace{\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})\}}_{D}\Big) \\ & - \sum_{t=1}^{T}\Big(\underbrace{\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\gamma}_t}(v_t)\}}_{E} + \underbrace{\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\}}_{F}\Big) - \sum_{n=1}^{N}\underbrace{\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\phi}_n}(z_n)\}}_{G}.\end{aligned} \tag{4.89}$$

We will now develop the individual expectations in (4.89) denoted by letter A to letter G. With (4.32), we obtain for A

$$A = \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln f^{(T)}(\boldsymbol{v};\alpha)\} = \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{t=1}^{T} -\alpha v_t \ln \alpha\right\} = -\sum_{t=1}^{T}\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{v_t\}\alpha\ln\alpha. \tag{4.90}$$

Since the variational factor pdf $q_{\boldsymbol{\gamma}_t}(v_t)$ is a gamma distribution with shape parameter $\gamma_{t,1}$ and rate parameter $\gamma_{t,2}$ the expectation $\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{v_t\}$ in (4.90) is given by

$$\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{v_t\} = \frac{\gamma_{t,1}}{\gamma_{t,2}}. \tag{4.91}$$

Because of (4.54), B can be expressed as

$$\begin{aligned} B &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda})\} \\ &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{t=1}^{T}\Big(\ln b(\boldsymbol{\lambda}) + \boldsymbol{\lambda}_1{}^{\mathrm{T}}\boldsymbol{\eta}^*_t + \lambda_2 a(\boldsymbol{\eta}^*_t)\Big)\right\} \\ &= T\ln b(\boldsymbol{\lambda}) + \sum_{t=1}^{T}\Big(\boldsymbol{\lambda}_1{}^{\mathrm{T}}\mathrm{E}^{\left(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\right)}\{\boldsymbol{\eta}^*_t\} + \lambda_2\mathrm{E}^{\left(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\right)}\{a(\boldsymbol{\eta}^*_t)\}\Big).\end{aligned} \tag{4.92}$$

Although we know the functional form of the prior pdf $f(\boldsymbol{\eta}^*_t;\boldsymbol{\lambda})$ and derived the variational factor pdf $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)$ to take the same form as well, we can not further develop (4.92), since we have not chosen a specific distribution for the corresponding likelihood $f(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})$.

For C, we obtain

$$\begin{aligned} C &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln p^{(T)}(z_n|\boldsymbol{v})\} \\ &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{t=1}^{T}\mathbb{1}(z_n = t)\ln v_t\right\} \\ &= \sum_{t=1}^{T}\mathrm{E}^{(q(v_t,z_n))}\{\mathbb{1}(z_n = t)\ln v_t\},\end{aligned} \tag{4.93}$$

where (4.33) has been applied in the first step. Note that the indicator variables $z_1,\ldots,z_N$ and the auxiliary variables $v_1,\ldots,v_T$ are statistically independent under the mean-field assumption

(4.15) due to the fully factorized form. Thus, the expectation of the product in (4.93) is equal to the product of expectations and C can be further developed as

$$C = \sum_{t=1}^{T} \mathrm{E}^{(q_{\phi_n}(z_n))}\{\mathbb{1}(z_n = t)\}\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} = \sum_{t=1}^{T} \phi_{n,t}\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\}, \tag{4.94}$$

where (4.21) has been applied.

Using (4.17), we obtain for D

$$\begin{aligned} D &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})\} \\ &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{t=1}^{T} \mathbb{1}(z_n = t)\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\right\} \\ &= \sum_{t=1}^{T} \mathrm{E}^{(q(\boldsymbol{\eta}_t^*,z_n))}\Big\{\mathbb{1}(z_n = t)\big(\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\big)\Big\} \\ &= \sum_{t=1}^{T} \mathrm{E}^{(q_{\phi_n}(z_n))}\{\mathbb{1}(z_n = t)\}\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\Big\{\ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{x}_n - a(\boldsymbol{\eta}_t^*)\Big\} \\ &= \sum_{t=1}^{T} \phi_{n,t}\big(\ln h(\boldsymbol{x}_n) + \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{\boldsymbol{\eta}_t^{*\mathrm{T}}\}\boldsymbol{x}_n - \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}\big), \end{aligned} \tag{4.95}$$

where the last step is due to (4.21). For similar reasons as mentioned after the derivation of B, we can not further develop (4.95) at the moment and refer to Chapter 5.

According to (4.99), the expectation denoted by the letter E is given by

$$\begin{aligned} E &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln q_{\gamma_t}(v_t)\} \\ &= \mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{(\gamma_{t,1} - 1)\ln v_t - \gamma_{t,2}v_t - (\ln\Gamma(\gamma_{t,1}) - \gamma_{t,1}\ln\gamma_{t,2})\} \\ &= (\gamma_{t,1} - 1)\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} - \gamma_{t,2}\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{v_t\} - \ln\Gamma(\gamma_{t,1}) + \gamma_{t,1}\ln\gamma_{t,2}. \end{aligned} \tag{4.96}$$

With (4.66), we obtain for F

$$\begin{aligned} F &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)\} \\ &= \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\Big\{\ln b_t(\boldsymbol{\tau}_t) + \boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\eta}_t^* - \tau_{t,2}a(\boldsymbol{\eta}_t^*)\Big\} \\ &= \ln b_t(\boldsymbol{\tau}_t) + \boldsymbol{\tau}_{t,1}^{\mathrm{T}}\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{\boldsymbol{\eta}_t^*\} - \tau_{t,2}\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}. \end{aligned} \tag{4.97}$$

Finally, for G we have

$$\begin{aligned} G &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\{\ln q_{\boldsymbol{\phi}_n}(z_n)\} \\ &= \mathrm{E}^{(q(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}))}\left\{\sum_{t=1}^{T} \mathbb{1}(z_n = t)\ln\phi_{n,t}\right\} \\ &= \sum_{t=1}^{T} \mathrm{E}^{(q_{\phi_n}(z_n))}\{\mathbb{1}(z_n = t)\}\ln\phi_{n,t} \end{aligned}$$

$$\begin{aligned} &= \sum_{t=1}^{T} \phi_{n,t} \ln \phi_{n,t} \\ &= \boldsymbol{\phi}_n^{\mathrm{T}} \ln \boldsymbol{\phi}_n, \end{aligned} \tag{4.98}$$

where we used (4.83) and (4.21).

Let us now consider the expectation $\mathrm{E}^{(q_{\gamma_t}(v_t))}\{\ln v_t\}$, which has shown up several times. As shown previously, the variational factor pdf $q_{\boldsymbol{\gamma}_t}(v_t)$ is given by

$$q_{\boldsymbol{\gamma}_t}(v_t) = \frac{1}{\Gamma(\gamma_{t,1})} \gamma_{t,2}^{\gamma_{t,1}} v_t^{\gamma_{t,1}-1} e^{-\gamma_{t,2} v_t}. \tag{4.99}$$

We next express the gamma distribution (4.99) using its exponential family representation, i.e.,

$$q_{\boldsymbol{\gamma}_t}(v_t) = \breve{h}(v_t) \exp\Big(\breve{\boldsymbol{\eta}}^{\mathrm{T}} \breve{\boldsymbol{t}}(v_t) - \breve{a}(\breve{\boldsymbol{\eta}})\Big). \tag{4.100}$$

Exponentiating the log of (4.99) leads to

$$\begin{aligned} q_{\boldsymbol{\gamma}_t}(v_t) &= \exp\bigg(\ln\bigg(\frac{1}{\Gamma(\gamma_{t,1})} \gamma_{t,2}^{\gamma_{t,1}} v_t^{\gamma_{t,1}-1} e^{-\gamma_{t,2} v_t} \bigg)\bigg) \\ &= \exp(-\ln\Gamma(\gamma_{t,1}) + \gamma_{t,1} \ln \gamma_{t,2} + (\gamma_{t,1} - 1) \ln v_t - \gamma_{t,2} v_t) \\ &= \exp((\gamma_{t,1} - 1) \ln v_t - \gamma_{t,2} v_t - (\ln\Gamma(\gamma_{t,1}) - \gamma_{t,1} \ln \gamma_{t,2})). \end{aligned} \tag{4.101}$$

Comparing (4.101) with (4.100), we can see that the base measure $\breve{h}(v_t)$ is given by

$$\breve{h}(v_t) = 1$$

and the natural parameter $\breve{\boldsymbol{\eta}}$ is

$$\breve{\boldsymbol{\eta}} = (\breve{\eta}_1 \quad \breve{\eta}_2)^{\mathrm{T}} = (\gamma_{t,1} - 1 \quad -\gamma_{t,2})^{\mathrm{T}}, \tag{4.102}$$

where the reverse substitution is given by

$$\boldsymbol{\gamma}_t = (\gamma_{t,1} \quad \gamma_{t,2})^{\mathrm{T}} = (\breve{\eta}_1 + 1 \quad -\breve{\eta}_2)^{\mathrm{T}}. \tag{4.103}$$

Furthermore, for the sufficient statistic $\breve{\boldsymbol{t}}(v_t)$ we obtain

$$\breve{\boldsymbol{t}}(v_t) = (\ln v_t \quad v_t)^{\mathrm{T}} \tag{4.104}$$

and the log-partition function $\breve{a}(\breve{\boldsymbol{\eta}})$ is given by

$$\begin{aligned} \breve{a}(\breve{\boldsymbol{\eta}}) &= \ln\Gamma(\gamma_{t,1}) - \gamma_{t,1} \ln \gamma_{t,2} \\ &= \ln\Gamma(\breve{\eta}_1 + 1) - (\breve{\eta}_1 + 1) \ln(-\breve{\eta}_2), \end{aligned} \tag{4.105}$$

where we used (4.103) in the last step. Note that $\ln v_t$ is part of the sufficient statistic given by (4.104). To obtain the desired expression for $\mathrm{E}^{(q_{\boldsymbol{\gamma}_t}(v_t))}\{\ln v_t\}$, we use the well known fact (see [40] for a detailed proof) that the expectation of the sufficient statistic $\breve{\boldsymbol{t}}(v_t)$ is given by the gradient of the log-partition function $\breve{a}(\breve{\boldsymbol{\eta}})$, i.e., $\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\left\{\breve{\boldsymbol{t}}(v_t)\right\} = \nabla\breve{a}(\breve{\boldsymbol{\eta}})$, where $\nabla = (\partial/\partial\breve{\eta}_1 \quad \partial/\partial\breve{\eta}_2)^{\mathrm{T}}$. Thus, we have

$$\begin{aligned}
\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} &= \frac{\partial}{\partial\breve{\eta}_1}\breve{a}(\breve{\boldsymbol{\eta}}) \\
&= \frac{\partial}{\partial\breve{\eta}_1}\left(\ln\Gamma(\breve{\eta}_1+1) - (\breve{\eta}_1+1)\ln(-\breve{\eta}_2)\right) \\
&= \frac{\Gamma'(\breve{\eta}_1+1)}{\Gamma(\breve{\eta}_1+1)} - \ln(-\breve{\eta}_2) \\
&= \Psi(\breve{\eta}_1+1) - \ln(-\breve{\eta}_2), \qquad (4.106)
\end{aligned}$$

where we used (4.105) in the second step and $\Psi(\cdot)$ is the digamma function. Inserting (4.103) into (4.106) gives

$$\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} = \Psi(\gamma_{t,1}) - \ln\gamma_{t,2}. \qquad (4.107)$$

4.3.4 Summary

In summary, according to (4.45), (4.46), (4.67), (4.68), and (4.84), the CAVI algorithm for static MFMs being described via the stick-breaking analogy (4.9) leads to the following update rules for the variational parameters:

$$\gamma_{t,1} = \frac{\tilde{\gamma}_{t,1}}{\sum_{t=1}^{T}\frac{\tilde{\gamma}_{t,1}}{\gamma_{t,2}}}, \qquad (4.108\text{a})$$

$$\gamma_{t,2} = \alpha, \qquad (4.108\text{b})$$

$$\boldsymbol{\tau}_{t,1} = \boldsymbol{\lambda}_1 + \sum_{n=1}^{N}\phi_{n,t}\boldsymbol{x}_n, \qquad (4.108\text{c})$$

$$\tau_{t,2} = \lambda_2 + \sum_{n=1}^{N}\phi_{n,t}, \qquad (4.108\text{d})$$

$$\phi_{n,t} = \frac{\exp(S_{n,t})}{\sum_{i=1}^{T}\exp(S_{n,i})}, \qquad (4.108\text{e})$$

where

$$\tilde{\gamma}_{t,1} = 1 + \sum_{n=1}^{N}\phi_{n,t}, \qquad (4.108\text{f})$$

$$S_{n,t} := \mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\} + \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{\boldsymbol{\eta}_t^{*\mathrm{T}}\}\boldsymbol{x}_n - \mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}, \qquad (4.108\text{g})$$

for $t = 1,\ldots,T$ and $n = 1,\ldots,N$.

We assess convergence of the ELBO to terminate the algorithm. The ELBO is given by

$$\begin{aligned}
\mathcal{L}(q;\boldsymbol{x}) = {} & \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{v};\alpha)\} + \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda})\} \\
& + \sum_{n=1}^{N}\left(\mathrm{E}^{(q)}\{\ln f^{(T)}(z_n|\boldsymbol{v})\} + \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)\}\right)
\end{aligned}$$

$$-\sum_{t=1}^{T}\Big(\mathrm{E}^{(q)}\{\ln q_{\gamma_t}(v_t)\}+\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)\}\Big)-\sum_{n=1}^{N}\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\phi}_n}(z_n)\}. \tag{4.109}$$

Recalling (4.90), (4.92), (4.94), (4.95), (4.96), (4.97), and (4.98), the individual expectations in (4.109) are

$$\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{v};\alpha)\}=-\sum_{t=1}^{T}\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{v_t\}\alpha\ln\alpha, \tag{4.110a}$$

$$\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{\eta}^*;\boldsymbol{\lambda})\}=T\ln b(\boldsymbol{\lambda})+\sum_{t=1}^{T}\Big(\boldsymbol{\lambda}_1{}^{\mathrm{T}}\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{\boldsymbol{\eta}_t^*\}+\lambda_2\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}\Big), \tag{4.110b}$$

$$\mathrm{E}^{(q)}\{\ln p^{(T)}(z_n|\boldsymbol{v})\}=\sum_{t=1}^{T}\phi_{n,t}\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\}, \tag{4.110c}$$

$$\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)\}=\sum_{t=1}^{T}\phi_{n,t}\big(\ln h(\boldsymbol{x}_n)+\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{\boldsymbol{\eta}_t^{*\mathrm{T}}\}\boldsymbol{x}_n-\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}\big), \tag{4.110d}$$

$$\mathrm{E}^{(q)}\{\ln q_{\gamma_t}(v_t)\}=(\gamma_{t,1}-1)\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\}-\gamma_{t,2}\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{v_t\}-\ln\Gamma(\gamma_{t,1})+\gamma_{t,1}\ln\gamma_{t,2}, \tag{4.110e}$$

$$\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)\}=\ln b_t(\boldsymbol{\tau}_t)+\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{\boldsymbol{\eta}_t^*\}-\tau_{t,2}\mathrm{E}^{\left(q_{\tau_t}(\boldsymbol{\eta}_t^*)\right)}\{a(\boldsymbol{\eta}_t^*)\}, \tag{4.110f}$$

$$\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\phi}_n}(z_n)\}=\boldsymbol{\phi}_n^{\mathrm{T}}\ln\boldsymbol{\phi}_n, \tag{4.110g}$$

and, according to (4.91) and (4.107), we have

$$\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{v_t\}=\frac{\gamma_{t,1}}{\gamma_{t,2}},$$
$$\mathrm{E}^{\left(q_{\gamma_t}(v_t)\right)}\{\ln v_t\}=\Psi(\gamma_{t,1})-\ln\gamma_{t,2}.$$

Finally, in Algorithm 2 we present all the necessary steps for our CAVI algorithm for static MFMs.

Algorithm 2: CAVI for static MFM models

```
Input: Observations $\boldsymbol{x}$, truncation level $T$, hyperparameters $\alpha$ and $\boldsymbol{\lambda}$
Output: Variational factor distributions $q^*_{\boldsymbol{\gamma}_t}(v_t)$, $q^*_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)$, for $t = 1, \dots, T$ and $q^*_{\boldsymbol{\phi}_n}(z_n)$,
        for $n = 1, \dots, N$
Initialize: Variational parameters $\boldsymbol{\gamma}_t^{(0)}, \boldsymbol{\tau}_t^{(0)}$, for $t = 1, \dots T$ and $\boldsymbol{\phi}_n^{(0)}$, for $n = 1, \dots, N$
while the ELBO has not converged do
    $\ell = \ell + 1$
    for $n$ from 1 to $N$ do
        for $t$ from 1 to $T$ do
            compute $S_{n,t}$ according to (4.108g)
        for $t$ from 1 to $T$ do
            set $\phi_{n,t}^{(\ell)}$ according to (4.108e)
    for $t$ from 1 to $T$ do
        compute $\tilde{\gamma}_{t,1}$ and $\gamma_{t,2}$ according to (4.108f) and (4.108b)
    for $t$ from 1 to $T$ do
        set $\boldsymbol{\gamma}_t^{(\ell)}$ according to (4.108a) and (4.108b)
        set $\boldsymbol{\tau}_t^{(\ell)}$ according to (4.108c) and (4.108d)
    Compute the ELBO $\mathcal{L}(q^{(\ell)}; \boldsymbol{x})$ according to (4.109)–(4.110g)
return $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) = \left(\prod_{t=1}^T q^*_{\boldsymbol{\gamma}_t}(v_t)\right)\left(\prod_{t=1}^T q^*_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\right)\left(\prod_{n=1}^N q^*_{\boldsymbol{\phi}_n}(z_n)\right)$
```

Chapter 5

Variational Inference for Static Mixtures of Finite Gaussian Mixtures

In this chapter, we test the clustering capability of the CAVI algorithm from Section 4.3. In order to be able to apply our proposed algorithm to any given set of observations, we need to specify the explicit functional form of the component distributions $f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n})$.

5.1 Conjugate Exponential Family Model for Static Mixtures of Finite Gaussian Mixtures

We now describe the static mixture of finite Gaussian mixtures (MFGM) model which arises by specifying the component distributions $f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n})$ as Gaussians with unknown mean $\boldsymbol{\theta}^*_{z_n}$ and known covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_M$, i.e.,

$$f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\theta}^*_{z_n})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\theta}^*_{z_n})\right) \tag{5.1}$$

with $\boldsymbol{x}_n \in \mathbb{R}^M$, $\boldsymbol{\theta}^*_{z_n} \in \mathbb{R}^M$, and $\mathbf{I}_M$ denoting the identity matrix of size $M \times M$. Recall that our CAVI algorithm is based on component distributions of canonical exponential family form. According to (4.17), we have

$$f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n}) = h(\boldsymbol{x}_n)\exp(\boldsymbol{\eta}^{*\mathrm{T}}_{z_n}\boldsymbol{x}_n - a(\boldsymbol{\eta}^*_{z_n})). \tag{5.2}$$

We will next show that the multivariate Gaussian distribution in (5.1) is a member of the exponential family and determine $h(\boldsymbol{x}_n)$, the natural parameter $\boldsymbol{\eta}^*_{z_n}$, and the log-partition $a(\boldsymbol{\eta}^*_{z_n})$. Expression (5.1) can be further developed as

$$f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}) = \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n\right)\exp\left(\boldsymbol{\theta}^{*\mathrm{T}}_{z_n}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n - \frac{1}{2}\boldsymbol{\theta}^{*\mathrm{T}}_{z_n}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^*_{z_n}\right).$$

By comparing with (5.2), we find that this expression is of exponential family form with

$$h(\boldsymbol{x}_n) = \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n\right), \tag{5.3}$$

$$\boldsymbol{\eta}^*_{z_n} = \boldsymbol{\eta}^*(\boldsymbol{\theta}^*_{z_n}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^*_{z_n}, \tag{5.4}$$

$$a(\boldsymbol{\eta}^*(\boldsymbol{\theta}^*_{z_n})) = \frac{1}{2}\boldsymbol{\theta}^{*\mathrm{T}}_{z_n}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^*_{z_n}. \tag{5.5}$$

Note that the component parameters $\boldsymbol{\theta}^*_{z_n}$, i.e., the means of the Gaussian component distributions, can be obtained from the natural component parameters $\boldsymbol{\eta}^*_{z_n}$ by inverting the function $\boldsymbol{\eta}^*(\cdot)$ given in (5.4):

$$\boldsymbol{\theta}^*_{z_n} = \boldsymbol{\Sigma}\boldsymbol{\eta}^*_{z_n}. \tag{5.6}$$

Inserting (5.6) into (5.5) yields

$$a(\boldsymbol{\eta}^*_{z_n}) = \frac{1}{2}\boldsymbol{\eta}^{*\mathrm{T}}_{z_n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\eta}^*_{z_n} = \frac{1}{2}\boldsymbol{\eta}^{*\mathrm{T}}_{z_n}\boldsymbol{\Sigma}\boldsymbol{\eta}^*_{z_n}. \tag{5.7}$$

With (5.3) and (5.7) the desired canonical exponential family form of the Gaussian components $f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})$ is given by

$$f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n}) = \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n\right)\exp\left(\boldsymbol{\eta}^{*\mathrm{T}}_{z_n}\boldsymbol{x}_n - \frac{1}{2}\boldsymbol{\eta}^{*\mathrm{T}}_{z_n}\boldsymbol{\Sigma}\boldsymbol{\eta}^*_{z_n}\right). \tag{5.8}$$

The complete specification of the component distributions in terms of their exponential family representation in canonical form now enables the complete specification of the corresponding conjugate prior.

5.1.1 Conjugate Prior

According to (4.18), a conjugate prior to the likelihood of the form in (5.2) is given by

$$f(\boldsymbol{\eta}^*_t;\boldsymbol{\lambda}) = b(\boldsymbol{\lambda})\exp(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}^*_t - \lambda_2 a(\boldsymbol{\eta}^*_t)), \tag{5.9}$$

where $b(\boldsymbol{\lambda}) \in \mathbb{R}^+$ is a normalization constant and the hyperparameters $\boldsymbol{\lambda}_1 \in \mathbb{R}^M$ and $\lambda_2 \in \mathbb{R}$ are arranged in the vector $\boldsymbol{\lambda}$ according to $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^{\mathrm{T}} \quad \lambda_2)^{\mathrm{T}}$. Inserting (5.7) with $z_n = t \in \{1, \ldots, T\}$ (cf. (4.16)) into (5.9) leads to

$$f(\boldsymbol{\eta}^*_t;\boldsymbol{\lambda}) = b(\boldsymbol{\lambda})\exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}^*_t - \frac{\lambda_2}{2}\boldsymbol{\eta}^{*\mathrm{T}}_t\boldsymbol{\Sigma}\boldsymbol{\eta}^*_t\right). \tag{5.10}$$

Equivalently, we have

$$f(\boldsymbol{\theta}^*_t;\boldsymbol{\lambda}) = c(\boldsymbol{\lambda})\exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^*_t - \frac{\lambda_2}{2}\boldsymbol{\theta}^{*\mathrm{T}}_t\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^*_t\right), \tag{5.11}$$

which is obtained from (5.10) by inserting (5.4). Since the likelihood is Gaussian with known covariance matrix (cf. (5.1)), the conjugate prior is Gaussian as well. Therefore, we choose the prior distribution of the component parameters $\boldsymbol{\theta}_t^*$, i.e., the component means, to be Gaussian with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ according to

$$f(\boldsymbol{\theta}_t^*) = \mathcal{N}(\boldsymbol{\theta}_t^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}) = \frac{1}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*})}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_t^* - \boldsymbol{\mu}_{\boldsymbol{\theta}^*})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}^{-1} (\boldsymbol{\theta}_t^* - \boldsymbol{\mu}_{\boldsymbol{\theta}^*})\right). \tag{5.12}$$

Note that the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ do not contain the subscript t because of the component means $\boldsymbol{\theta}_t^*$ being identically distributed a priori. Working out the exponent in (5.12) and comparing the result with (5.11), we find the following important relationships [40]:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \frac{1}{\lambda_2} \boldsymbol{\lambda}_1, \tag{5.13a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \frac{1}{\lambda_2} \boldsymbol{\Sigma}. \tag{5.13b}$$

By applying the linear transformation $\boldsymbol{\eta}_t^* = \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}_t^*$ (cf. (5.4)) to the Gaussian pdf in (5.12) we obtain that $\boldsymbol{\eta}_t^* \sim \mathcal{N}(\boldsymbol{\eta}_t^*; \boldsymbol{\mu}_{\boldsymbol{\eta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}^*})$, where

$$\boldsymbol{\mu}_{\boldsymbol{\eta}^*} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \frac{1}{\lambda_2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\lambda}_1, \tag{5.14a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}^*} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \boldsymbol{\Sigma}^{-1} = \frac{1}{\lambda_2} \boldsymbol{\Sigma}^{-1}. \tag{5.14b}$$

Note that the prior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ is, due to the EF framework, restricted to be a multiple of the covariance matrix $\boldsymbol{\Sigma}$ of the corresponding likelihood function $f^{(T)}(\boldsymbol{x}_n | \boldsymbol{\theta}_{z_n}^*)$ and thus can not be chosen arbitrarily (cf. (5.13b)). Furthermore, we emphasize that the hyperparameter vector $\boldsymbol{\lambda}$ is the same for both representations (5.10) and (5.11), whereas the normalization constants $b(\boldsymbol{\lambda})$ and $c(\boldsymbol{\lambda})$ are different.

Having the same hyperparameters $\boldsymbol{\lambda}$ for both, the representation using the component parameters $\boldsymbol{\theta}_t^*$ and the natural component parameters $\boldsymbol{\eta}_t^*$ leads to a convenient way of determining the conjugate prior $f(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda})$ in its desired representation given in (5.10): First, the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ of $f(\boldsymbol{\theta}_t^*) = \mathcal{N}(\boldsymbol{\theta}_t^*; \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*})$ are specified. Given $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \sigma_{\boldsymbol{\theta}^*}^2 \mathbf{I}_M$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_M$, the hyperparameters $\boldsymbol{\lambda}_1$ and λ_2 of the equivalent representation (5.11) can be obtained from (5.13a) and (5.13b) according to

$$\lambda_2 = \frac{\sigma^2}{\sigma_{\boldsymbol{\theta}^*}^2}, \tag{5.15}$$

$$\boldsymbol{\lambda}_1 = \lambda_2 \boldsymbol{\mu}_{\boldsymbol{\theta}^*}. \tag{5.16}$$

Once $\boldsymbol{\lambda}_1$ and λ_2 have been determined, they can directly be applied to $f(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda})$ given in (5.10).

Regarding its complete specification, it remains to elaborate the normalization constant $b(\boldsymbol{\lambda})$. As mentioned above, the prior distribution of the natural component parameters $\boldsymbol{\eta}_t^*$ is Gaussian and thus given by

$$f(\boldsymbol{\eta}_t^*) = \mathcal{N}(\boldsymbol{\eta}_t^*; \boldsymbol{\mu}_{\boldsymbol{\eta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}^*}) = \frac{1}{\sqrt{(2\pi)^M \det(\boldsymbol{\Sigma}_{\boldsymbol{\eta}^*})}} \exp\Big(-\frac{1}{2}(\boldsymbol{\eta}_t^* - \boldsymbol{\mu}_{\boldsymbol{\eta}^*})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}^*}{}^{-1} (\boldsymbol{\eta}_t^* - \boldsymbol{\mu}_{\boldsymbol{\eta}^*})\Big). \tag{5.17}$$

Inserting the expressions (5.14a) and (5.14b) leads to

$$f(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda}) = \frac{1}{\sqrt{(2\pi)^M \det(\frac{1}{\lambda_2}\boldsymbol{\Sigma}^{-1})}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\eta}_t^* - \frac{1}{\lambda_2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1\right)^{\mathrm{T}} \left(\frac{1}{\lambda_2}\boldsymbol{\Sigma}^{-1}\right)^{-1} \left(\boldsymbol{\eta}_t^* - \frac{1}{\lambda_2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1\right)\right),$$

which can be further developed as

$$\begin{aligned} f(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda}) &= \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} \exp\left(-\frac{\lambda_2}{2}\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}_t^* + \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1\right) \\ &= \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} \exp\left(-\frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1\right) \exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \frac{\lambda_2}{2}\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}_t^*\right). \end{aligned} \tag{5.18}$$

A comparison of (5.18) and (5.10) yields that

$$b(\boldsymbol{\lambda}) = \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} \exp\left(-\frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1\right). \tag{5.19}$$

We conclude that the normalization constant $b(\boldsymbol{\lambda})$ can be calculated in closed form. Thus, we do not rely on numerical integration according to (4.13). Consequently, the computational complexity of the CAVI algorithm is massively reduced, overall leading to shorter runtimes.

5.1.2 Truncated MFGM Model

Recall the approximating model given in (4.16). The results for the Gaussian case just obtained enable a more precise specification:

$$v_t \overset{\text{i.i.d.}}{\sim} \mathcal{E}(v_t; \alpha) \quad \text{for} \quad t = 1, \dots, T, \tag{5.20a}$$

$$\pi_t = v_t \quad \text{for} \quad t = 1, \dots, T, \tag{5.20b}$$

$$\boldsymbol{\eta}_1^*, \dots, \boldsymbol{\eta}_T^* \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{\eta}_t^*; \boldsymbol{\lambda}), \tag{5.20c}$$

$$z_1, \dots, z_N | \boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n; \boldsymbol{\pi}), \tag{5.20d}$$

$$\boldsymbol{x}_n \sim f^{(T)}(\boldsymbol{x}_n | \boldsymbol{\eta}_{z_n}^*) \quad \text{independently for } n = 1, \dots, N. \tag{5.20e}$$

Here, the components $f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)$ are canonical exponential family distributions, i.e.,

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

where, according to (5.3) and (5.7),

$$h(\boldsymbol{x}_n) = \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}} \exp\Big(-\frac{1}{2}\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n\Big),$$

$$a(\boldsymbol{\eta}_t^*) = \frac{1}{2}\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}_t^*. \tag{5.21}$$

The corresponding conjugate prior pdf $f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda})$ is given by

$$f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda}) = b(\boldsymbol{\lambda})\exp\left(\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\eta}_t^* - \frac{\lambda_2}{2}\boldsymbol{\eta}_t^{*\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}_t^*\right) \tag{5.22}$$

with

$$b(\boldsymbol{\lambda}) = \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}^{-1})}}\exp\left(-\frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1\right), \tag{5.23}$$

$$\lambda_2 = \frac{\sigma^2}{\sigma^2_{\boldsymbol{\theta}^*}}, \tag{5.24}$$

$$\boldsymbol{\lambda}_1 = \lambda_2\boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \tag{5.25}$$

and σ^2, $\sigma^2_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ as in Section 5.1.1. Note that the full model given in (4.9) can specified in a similar manner.

5.2 CAVI Algorithm

In this section, we adapt the general CAVI algorithm for static MFMs in the conjugate EF framework (described in Section 4.3.4 and summarized by Algorithm 2) to the Gaussian case considered in Section 5.1.

5.2.1 Expectations and Quantities Involved in the CAVI Algorithm

The CAVI algorithm for static MFGMs provides us with an approximation $q^*(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z})$ of the posterior pdf $f(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z}|\boldsymbol{x})$, where $\boldsymbol{v} = (v_1 \ \cdots \ v_T)^{\mathrm{T}}$, $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^{*\mathrm{T}} \ \cdots \ \boldsymbol{\eta}_T^{*\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{z} = (z_1 \ \cdots \ z_N)^{\mathrm{T}}$ are the latent model parameters. Due to the truncated mean-field approximation, the variational distribution $q^*(\boldsymbol{v},\boldsymbol{\eta}^*,\boldsymbol{z})$ is factorized according to (4.15), where the truncation parameter $T \in \mathbb{N}$ can in general be freely set and acts as an input parameter for the CAVI algorithm. Further inputs are the observations $\boldsymbol{x} = (\boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$, the hyperparameter α of the prior distribution of the auxiliary variables v_t (cf. (5.20a)) and the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ of the prior pdf $f(\boldsymbol{\theta}_t^*)$.

According to (4.66), the variational factor pdf $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)$ is given by

$$q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*) = b_t(\boldsymbol{\tau}_t)\exp(\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\eta}_t^* - \tau_{t,2}a(\boldsymbol{\eta}_t^*)),$$

and thus has the same functional form as the corresponding prior pdf $f(\boldsymbol{\eta}_t^*;\boldsymbol{\lambda})$ given in (5.22), but with hyperparameters $\boldsymbol{\tau}_{t,1}$ and $\tau_{t,2}$ instead of $\boldsymbol{\lambda}_1$ and λ_2. In consequence, the normalization constant $b_t(\boldsymbol{\tau}_t)$ can directly be obtained from (5.23) by exchanging $\boldsymbol{\lambda}$ with $\boldsymbol{\tau}_t = \left(\boldsymbol{\tau}_{t,1}^{\mathrm{T}} \;\; \tau_{t,2}\right)^{\mathrm{T}}$:

$$b_t(\boldsymbol{\tau}_t) = \frac{1}{\sqrt{\left(\frac{2\pi}{\tau_{t,2}}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} \exp\left(- \frac{1}{2\tau_{t,2}} \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1} \right). \tag{5.26}$$

Similarly, from (5.14) we obtain

$$\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*} = \frac{1}{\tau_{t,2}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1}, \tag{5.27a}$$

$$\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_t^*} = \frac{1}{\tau_{t,2}} \boldsymbol{\Sigma}^{-1}. \tag{5.27b}$$

Note that $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}^*}$ are referred to as approximated posterior mean and approximated posterior covariance matrix of the tth natural component parameter $\boldsymbol{\eta}_t^*$, since the variational factor pdf $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)$ approximates the true marginal posterior $f(\boldsymbol{\eta}_t^*|\boldsymbol{x})$. Furthermore, $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}$ is equal to the approximated MMSE estimate $\hat{\boldsymbol{\eta}}_t^*$ of $\boldsymbol{\eta}_t^*$ (cf. Section 2.5).

We next work out the expectations $\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\}$ and $\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}$, which we have not been able to further develop in Section 4.3. Obviously, $\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\}$ is the approximated posterior mean of $\boldsymbol{\eta}_t^*$ and thus given by (5.27a), i.e.,

$$\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\{\boldsymbol{\eta}_t^*\} = \frac{1}{\tau_{t,2}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1}. \tag{5.28}$$

Applying expression (5.21) to the approximated posterior expectation $\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\}$ of the log-partition function $a(\boldsymbol{\eta}_t^*)$ leads to

$$\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\} = \frac{1}{2} \mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\eta}_t^*\right\}. \tag{5.29}$$

It can be shown [41] that the expectation of the quadratic form $\boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\eta}_t^*$ with respect to the variational factor pdf $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)$ is given by

$$\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\eta}_t^*\right\} = \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*}^{\mathrm{T}} \boldsymbol{\Sigma} \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\eta}_t^*} + \mathrm{trace}(\boldsymbol{\Sigma} \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_t^*}). \tag{5.30}$$

With (5.27a) and (5.27b), (5.30) can further be developed as

$$\begin{aligned}
\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\left\{\boldsymbol{\eta}_t^{*\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\eta}_t^*\right\} &= \frac{1}{\tau_{t,2}^2} \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1} + \mathrm{trace}\left(\frac{1}{\tau_{t,2}} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \right), \\
&= \frac{1}{\tau_{t,2}^2} \boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1} + \mathrm{trace}\left(\frac{1}{\tau_{t,2}} \mathbf{I}_M \right), \\
&= \frac{1}{\tau_{t,2}^2} (\boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1} + \tau_{t,2} M).
\end{aligned}$$

Inserting this expression into (5.29) yields the final result for the approximate posterior expectation of the log-partition function $a(\boldsymbol{\eta}_t^*)$

$$\mathrm{E}^{(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*))}\{a(\boldsymbol{\eta}_t^*)\} = \frac{1}{2\tau_{t,2}^2} (\boldsymbol{\tau}_{t,1}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\tau}_{t,1} + \tau_{t,2} M). \tag{5.31}$$

5.2.2 Summary

Due to the results obtained above, we are now ready to adapt the general CAVI algorithm for static MFMs described in Section 4.3.4 to the special case of static MFGMs. Inserting the two missing expectations (5.28) and (5.31), the updates of the variational parameters $\boldsymbol{\gamma}_t$, $\boldsymbol{\tau}_t$ and $\boldsymbol{\phi}_n$ of the corresponding variational factor pdfs $q_{\boldsymbol{\gamma}_t}(v_t) = \mathcal{G}(v_t; \gamma_{t,1}, \gamma_{t,2})$, $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*) = b_t(\boldsymbol{\tau}_t)\exp(\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\eta}_t^* - \tau_{t,2}a(\boldsymbol{\eta}_t^*))$ and the variational factor pmf $q_{\boldsymbol{\phi}_n}(z_n) = \mathcal{C}(z_n; \boldsymbol{\phi}_n)$, for $t = 1, \ldots, T$ and $n = 1, \ldots, N$ are given by

$$\gamma_{t,1} = \frac{\tilde{\gamma}_{t,1}}{\sum_{t=1}^{T} \frac{\tilde{\gamma}_{t,1}}{\gamma_{t,2}}}, \tag{5.32a}$$

$$\gamma_{t,2} = \alpha, \tag{5.32b}$$

$$\boldsymbol{\tau}_{t,1} = \boldsymbol{\lambda}_1 + \sum_{n=1}^{N} \phi_{n,t}\boldsymbol{x}_n, \tag{5.32c}$$

$$\tau_{t,2} = \lambda_2 + \sum_{n=1}^{N} \phi_{n,t}, \tag{5.32d}$$

$$\phi_{n,t} = \frac{\exp(S_{n,t})}{\sum_{i=1}^{T} \exp(S_{n,i})}, \tag{5.32e}$$

where

$$\tilde{\gamma}_{t,1} = 1 + \sum_{n=1}^{N} \phi_{n,t}, \tag{5.32f}$$

$$S_{n,t} = \Psi(\gamma_{t,1}) - \ln \gamma_{t,2} + \frac{1}{\tau_{t,2}}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n - \frac{1}{2\tau_{t,2}^2}\big(\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} + \tau_{t,2}M\big). \tag{5.32g}$$

It remains to adapt the ELBO, which is used to assess the convergence of the CAVI algorithm for static MFGMs. According to (4.109), it is given by

$$\begin{aligned}\mathcal{L}(q; \boldsymbol{x}) = {} & \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{v}; \alpha)\} + \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{\eta}^*; \boldsymbol{\lambda})\} \\ & + \sum_{n=1}^{N}\Big(\mathrm{E}^{(q)}\{\ln p^{(T)}(z_n|\boldsymbol{v})\} + \mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}_{z_n}^*)\}\Big) \\ & - \sum_{t=1}^{T}\Big(\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\gamma}_t}(v_t)\} + \mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}_t^*)\}\Big) - \sum_{n=1}^{N} \mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\phi}_n}(z_n)\}.\end{aligned} \tag{5.33a}$$

Using the base measure (5.3), the missing normalization constants (5.23) and (5.26), and the missing expectations (5.28) and (5.31), the individual terms in (5.33a) can be expressed as

$$\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{v}; \alpha)\} = -\sum_{t=1}^{T} \frac{\gamma_{t,1}}{\gamma_{t,2}}\alpha \ln \alpha, \tag{5.33b}$$

$$\begin{aligned}\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{\eta}^*; \boldsymbol{\lambda})\} = {} & T \ln \frac{1}{\sqrt{\left(\frac{2\pi}{\lambda_2}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} - \frac{1}{2\lambda_2}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_1 \\ & + \sum_{t=1}^{T}\left(\frac{1}{\tau_{t,2}}\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} + \frac{\lambda_2}{2\tau_{t,2}^2}\Big(\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} + \tau_{t,2}M\Big)\right),\end{aligned} \tag{5.33c}$$

$$\mathrm{E}^{(q)}\{\ln p^{(T)}(z_n|\boldsymbol{v})\} = \sum_{t=1}^{T} \phi_{n,t}(\Psi(\gamma_{t,1}) - \ln \gamma_{t,2}), \tag{5.33d}$$

$$\begin{aligned}&\mathrm{E}^{(q)}\{\ln f^{(T)}(\boldsymbol{x}_n|\boldsymbol{\eta}^*_{z_n})\}\\ &\quad= \sum_{t=1}^{T} \phi_{n,t}\left(\ln \frac{\exp(-\frac{1}{2}\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n)}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}} + \frac{1}{\tau_{t,2}}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n - \frac{1}{2\tau_{t,2}^2}\left(\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} + \tau_{t,2}M\right)\right),\end{aligned} \tag{5.33e}$$

$$\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\gamma}_t}(v_t)\} = (\gamma_{t,1} - 1)(\Psi(\gamma_{t,1}) - \ln \gamma_{t,2}) - \gamma_{t,2}\frac{\gamma_{t,1}}{\gamma_{t,2}} - \ln \Gamma(\gamma_{t,1}) + \gamma_{t,1}\ln \gamma_{t,2}, \tag{5.33f}$$

$$\begin{aligned}\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\} = \ln \frac{1}{\sqrt{\left(\frac{2\pi}{\tau_{t,2}}\right)^M \det(\boldsymbol{\Sigma}^{-1})}} &- \frac{1}{2\tau_{t,2}}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1}\\ &+ \frac{1}{\tau_{t,2}}\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} - \frac{1}{2\tau_{t,2}}(\boldsymbol{\tau}_{t,1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1} + \tau_{t,2}M),\end{aligned} \tag{5.33g}$$

$$\mathrm{E}^{(q)}\{\ln q_{\boldsymbol{\phi}_n}(z_n)\} = \boldsymbol{\phi}_n^{\mathrm{T}} \ln \boldsymbol{\phi}_n. \tag{5.33h}$$

Finally, in Algorithm 3 we present all the necessary steps for the CAVI algorithm for static MFGMs. At the end of each iteration, the ELBO is evaluated and the algorithm is terminated, if the relative change of the ELBO with respect to the previous iteration falls below some small threshold ε, i.e., if

$$\frac{|\mathcal{L}(q^{(\ell)};\boldsymbol{x}) - \mathcal{L}(q^{(\ell-1)};\boldsymbol{x})|}{|\mathcal{L}(q^{(\ell-1)};\boldsymbol{x})|} < \varepsilon. \tag{5.34}$$

The output of the algorithm is further processed to estimate the latent model parameters $\boldsymbol{v} = (v_1 \;\; \cdots \;\; v_T)^{\mathrm{T}}$, $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^{*\mathrm{T}} \;\; \cdots \;\; \boldsymbol{\eta}_T^{*\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{z} = (z_1 \;\; \cdots \;\; z_N)^{\mathrm{T}}$.

5.2.3 Initialization

Taking a closer look at the update equations given by (5.32), we see that the update procedure cycles through two coupled stages. In the first stage, we use the current global hyperparameters $\boldsymbol{\gamma}_t$ and $\boldsymbol{\tau}_t$ of the corresponding variational factor pdfs $q_{\boldsymbol{\gamma}_t}(v_t)$ and $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)$ to evaluate (5.32g) and hence update the responsibilities $\mathrm{E}^{(q_{\boldsymbol{\phi}_n}(z_n))}\{\mathbb{1}(z_n = t)\} = \phi_{n,t}$ using (5.32e). In the following stage, we keep the responsibilities fixed and use them to implicitly update $q_{\boldsymbol{\gamma}_t}(v_t)$ and $q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)$ by updating the corresponding variational parameters $\boldsymbol{\gamma}_t$ and $\boldsymbol{\tau}_t$ according to (5.32a)–(5.32d).

Based on this two-stage update procedure, we suggest two different ways to initialize our CAVI algorithm. First, one could initialize the variational parameters $\boldsymbol{\gamma}_t$ and $\boldsymbol{\tau}_t$ and use them to calculate the responsibilities $\phi_{n,t}$ using the respective update equation. Obviously, we use the hyperparameters of the corresponding prior pdfs as initial values for the variational parameters, i.e., $\boldsymbol{\gamma}_t^{(0)} = (1 \quad \alpha)^{\mathrm{T}}$ and $\boldsymbol{\tau}_t^{(0)} = (\boldsymbol{\lambda}_1^{\mathrm{T}} \quad \lambda_2)^{\mathrm{T}}$, for $t = 1, \ldots, T$. Since v_t and $\boldsymbol{\eta}^*_t$ are global model parameters, this initialization type will be referred to as *global*.

The second option is to invert this procedure, i.e., initializing the responsibilities $\phi_{n,t}$ and calculating the variational parameters $\boldsymbol{\gamma}_t$ and $\boldsymbol{\tau}_t$ according to their respective update equations.

Algorithm 3: CAVI for static MFGM models

```
Input: Observations $\boldsymbol{x}$, truncation level $T$, threshold $\varepsilon$, hyperparameters
       $\{\alpha, \boldsymbol{\Sigma} = \sigma^2\mathbf{I}_M, \boldsymbol{\mu}_{\boldsymbol{\theta}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \sigma^2_{\boldsymbol{\theta}^*}\mathbf{I}_M\}$
Output: Variational factor distributions $q^*_{\boldsymbol{\gamma}_t}(v_t)$, $q^*_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)$, for $t = 1, \ldots, T$ and $q^*_{\boldsymbol{\phi}_n}(z_n)$,
        for $n = 1, \ldots, N$
compute $\boldsymbol{\lambda}$ according to (5.24) and (5.25)
Initialize: Variational parameters $\boldsymbol{\gamma}_t^{(0)}, \boldsymbol{\tau}_t^{(0)}$, for $t = 1, \ldots T$ and $\boldsymbol{\phi}_n^{(0)}$, for $n = 1, \ldots, N$
while the ELBO has not converged do
    $\ell = \ell + 1$
    for $n$ from 1 to $N$ do
        for $t$ from 1 to $T$ do
            compute $S_{n,t}$ according to (5.32g)
        for $t$ from 1 to $T$ do
            set $\phi_{n,t}^{(\ell)}$ according to (5.32e)
    for $t$ from 1 to $T$ do
        compute $\tilde{\gamma}_{t,1}$ and $\gamma_{t,2}$ according to (5.32f) and (5.32b)
    for $t$ from 1 to $T$ do
        set $\boldsymbol{\gamma}_t^{(\ell)}$ according to (5.32a) and (5.32b)
        set $\boldsymbol{\tau}_t^{(\ell)}$ according to (5.32c) and (5.32d)
    Compute the ELBO $\mathcal{L}(q^{(\ell)}; \boldsymbol{x})$ according to (5.33) and check convergence according
     to (5.34)
return $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) = \left(\prod_{t=1}^T q^*_{\boldsymbol{\gamma}_t}(v_t)\right)\left(\prod_{t=1}^T q^*_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\right)\left(\prod_{n=1}^N q^*_{\boldsymbol{\phi}_n}(z_n)\right)$
```

In what follows, we describe two approaches of initializing the responsibilities $\phi_{n,t}$ that are used for the simulations in Section 5.4.

The first one is to assign each observation $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ to a separate component. In consequence, the truncation level T can not be chosen freely since it has to be equal to the number of observations N. The initial responsibilities are given by

$$\phi_{n,t}^{(0)} = \begin{cases} 1 & t = n, \\ 0 & \text{else}. \end{cases} \tag{5.35}$$

Thus, each initial variational parameter vector $\boldsymbol{\phi}_n^{(0)} = \begin{pmatrix} \phi_{n,1}^{(0)} & \cdots & \phi_{n,t}^{(0)} & \cdots & \phi_{n,T}^{(0)} \end{pmatrix}^{\mathrm{T}}$, for $n = 1, \ldots, N$, contains $N - 1$ elements which are equal to zero and one element (for index $t = n$) which is equal to one. This initialization type will be referred to as *unique*.

In the second approach, we preset the truncation level T and assign each observation $\boldsymbol{x}_n$, for $n = 1, \ldots, N$, randomly to one of the T components. Thus, each initial variational parameter vector $\boldsymbol{\phi}_n^{(0)} = \begin{pmatrix} \phi_{n,1}^{(0)} & \cdots & \phi_{n,t'}^{(0)} & \cdots & \phi_{n,T}^{(0)} \end{pmatrix}^{\mathrm{T}}$, for $n = 1, \ldots, N$, contains $T - 1$ elements which are equal to zero and one element (with randomly chosen index $t' \in \{1, 2, \ldots, T\}$) which is equal to one. As stated in [28], CAVI only guarantees convergence to a local maximum of the ELBO, which can be sensitive to initialization. In other words, poor initializations can lead to low local maxima. Therefore, we run the CAVI algorithm multiple times, i.e., we perform multiple permutations, each time with a new realization of the random component assignment described above. As a final result, we take the output of the run (permutation) with the highest ELBO. This initialization type will be referred to as *permute*.

5.3 Estimation of the Latent Model Parameters

In this section we describe the estimation of the auxiliary variables $\boldsymbol{v}$ (or equivalently the mixture weights $\boldsymbol{\pi}$), the component parameters $\boldsymbol{\theta}^*$, and the indicator variables $\boldsymbol{z}$ using the output $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z})$ of Algorithm 3. Recall that the variational distribution $q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z})$ is an approximation of the posterior distribution $f(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}|\boldsymbol{x})$ and is factorized according to

$$q^*(\boldsymbol{v}, \boldsymbol{\eta}^*, \boldsymbol{z}) = \left(\prod_{t=1}^{T} q^*_{\boldsymbol{\gamma}_t}(v_t)\right)\left(\prod_{t=1}^{T} q^*_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\right)\left(\prod_{n=1}^{N} q^*_{\boldsymbol{\phi}_n}(z_n)\right).$$

Therefore, the variational factor distributions $q^*_{\boldsymbol{\gamma}_t}(v_t)$, $q^*_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)$ and $q^*_{\boldsymbol{\phi}_n}(z_n)$ approximate the corresponding marginal posterior distributions $f(v_t|\boldsymbol{x})$, $f(\boldsymbol{\eta}^*_t|\boldsymbol{x})$ and $f(z_n|\boldsymbol{x})$. Hence, they can directly be applied to the Bayesian estimators introduced in Section 2.5.

We start with the approximate MMSE estimate $\hat{\boldsymbol{\theta}}^*_t$ of the tth component parameter $\boldsymbol{\theta}^*_t$, i.e., the mean of the tth component. According to (5.28), the approximate posterior mean of the tth natural component parameter $\boldsymbol{\eta}^*_t$ is given by

$$\hat{\boldsymbol{\eta}}^*_t = \mathrm{E}^{\left(q_{\boldsymbol{\tau}_t}(\boldsymbol{\eta}^*_t)\right)}\{\boldsymbol{\eta}^*_t\} = \frac{1}{\tau_{t,2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}_{t,1}.$$

Applying the deterministic transformation $\boldsymbol{\theta}^*_t = \boldsymbol{\Sigma}\boldsymbol{\eta}^*_t$ (cf. (5.6)) yields the desired approximate MMSE estimate

$$\hat{\boldsymbol{\theta}}^*_t = \frac{1}{\tau_{t,2}}\boldsymbol{\tau}_{t,1} \tag{5.36}$$

of the tth component parameter $\boldsymbol{\theta}^*_t$.

Due to the relation $\pi_t = v_t$, for $t = 1, \ldots, T$ (cf. (4.16b)), the approximate MMSE estimate $\hat{\pi}_t$ of the tth mixture weight π_t is equal to the approximate posterior mean of the tth auxiliary variable v_t. According to (4.91), it is given by

$$\hat{\pi}_t = \mathrm{E}^{\left(q_{\boldsymbol{\gamma}_t}(v_t)\right)}\{v_t\} = \frac{\gamma_{t,1}}{\gamma_{t,2}}. \tag{5.37}$$

Finally, we estimate the indicator variables $z_1, \ldots, z_N$. According to (4.83), the variational factor pmf $q_{\boldsymbol{\phi}_n}(z_n)$ is a categorical distribution with variational parameter $\boldsymbol{\phi}_n = (\phi_{n,1} \ \cdots \ \phi_{n,T})^{\mathrm{T}}$.

Recall that the $\phi_{n,t}$ are responsibilities, i.e., the probabilities that an observation $\boldsymbol{x}_n$ is assigned to the tth mixture component. Hence, we choose the estimate $\hat{z}_n$ of the nth indicator variable z_n to be the component with the largest responsibility. In other words, the estimate $\hat{z}_n$ is equal to the mode of the approximated posterior pdf $q_{\boldsymbol{\phi}_n}(z_n)$, i.e., the approximate MAP estimate of z_n:

$$\hat{z}_n = \arg\max_t \phi_{n,t}. \tag{5.38}$$

Using the estimates $\hat{z}_1, \ldots, \hat{z}_n$, the number of clusters in the observations $\boldsymbol{x}$ (cf. (2.11)) can be estimated according to

$$\hat{L} = \sum_{t=1}^{T} \mathbb{1}(\hat{N}_t > 0), \tag{5.39}$$

where $\hat{N}_t = \sum_{n=1}^{N} \mathbb{1}(\hat{z}_n = t)$ is the (estimated) number of observations assigned to the tth mixture component.

5.4 Performance Evaluation

5.4.1 Data Generative Model

Until further notice, we consider the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^M$ to be drawn from a finite mixture of multivariate Gaussian distributions with $K = 8$ equally sized components [20] of dimension $M = 2$. The data generative model for N conditionally independent observations $\boldsymbol{x} = (\boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$ can be summarized as follows:

$$\boldsymbol{\pi} = \frac{1}{8}\mathbf{1}_8, \tag{5.40a}$$

$$z_1, \ldots, z_N \overset{\text{i.i.d.}}{\sim} \mathcal{C}(z_n; \boldsymbol{\pi}), \tag{5.40b}$$

$$\boldsymbol{x}_n | \boldsymbol{\theta}^*, z_n \sim f(\boldsymbol{x}_n | \boldsymbol{\theta}^*_{z_n}) \quad \text{independently for} \quad n = 1, \ldots, N. \tag{5.40c}$$

Here, the component distributions $f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n})$ are Gaussians, i.e., $f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{z_n}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_2$ and component means

$$\boldsymbol{\theta}^*_{z_n} \in \left\{ \begin{pmatrix} -6.0 \\ -2.5 \end{pmatrix}, \begin{pmatrix} -6.0 \\ 2.5 \end{pmatrix}, \begin{pmatrix} 6.0 \\ -2.5 \end{pmatrix}, \begin{pmatrix} 6.0 \\ 2.5 \end{pmatrix}, \begin{pmatrix} -2.0 \\ -2.5 \end{pmatrix}, \begin{pmatrix} -2.0 \\ 2.5 \end{pmatrix}, \begin{pmatrix} 2.0 \\ -2.5 \end{pmatrix}, \begin{pmatrix} 2.0 \\ 2.5 \end{pmatrix} \right\}. \tag{5.41}$$

In what follows, a realization $\boldsymbol{x} = (\boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$ of (5.40) will be referred to as *synthetic dataset*. For illustration, an example dataset with $N = 300$ is shown in Figure 5.1.

We emphasize that the true component assignments $z_1, \ldots, z_N \in \{1, \ldots, 8\}$ (cf.(5.40b)) and the corresponding component means $\boldsymbol{\theta}^*_{z_n}$ as in (5.41) are used for performance evaluation purposes, but the CAVI algorithm itself is provided with the raw synthetic dataset, see Figure 5.1. In a post processing step, the cluster (i.e., the non-empty components) assignments and the

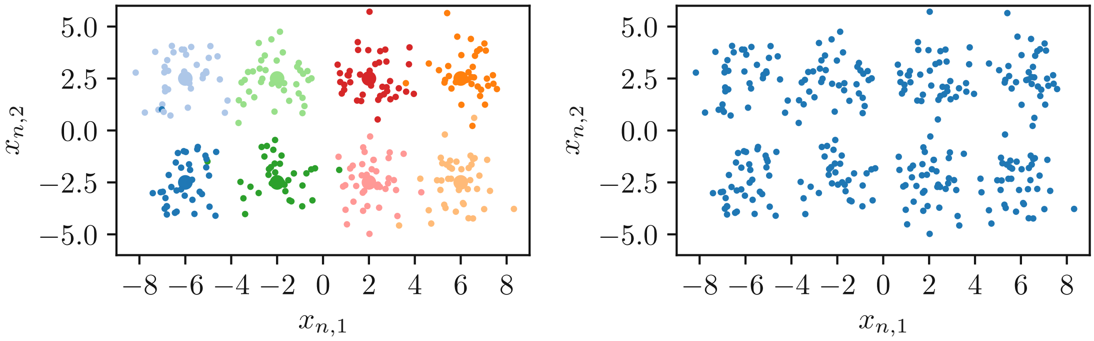


Figure 5.1: Synthetic dataset $\boldsymbol{x}$ with eight clusters generated according to (5.40) for $N = 300$. Left: Labeled observations $\boldsymbol{x}_n = (x_{n,1} \quad x_{n,2})^{\mathrm{T}}$ for $n = 1, \ldots, 300$, illustrated as colored dots and cluster means illustrated as large colored dots. Right: Raw synthetic dataset $\boldsymbol{x}$ as seen at the input of the CAVI algorithm.

corresponding cluster means are estimated according to (5.38) and (5.36). Note that empty components, i.e., components for which $\hat{N}_t = 0$, are thrown away.

5.4.2 Simulation Results

We next perform simulations to evaluate the clustering performance and estimation accuracy of the number of clusters of our CAVI algorithm for static MFGMs, which is summarized in Algorithm 3. The presented results are averaged over 200 MC runs of the algorithm and the corresponding post processing step, where in each MC run a new synthetic dataset $\boldsymbol{x}$ is drawn from (5.40). The maximum number of iterations is set to 50, i.e., $\ell_{\max} = 50$ and the algorithm is terminated according to (5.34) with $\varepsilon = 10^{-10}$. Regarding the specification of the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ of the Gaussian prior pdf $f(\boldsymbol{\theta}_t^*)$ (cf. (5.12)), we take a data driven approach inspired by [29]. The mean $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ is set to be the median of the observations, i.e.,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}^*} = \mathrm{median}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N), \tag{5.42}$$

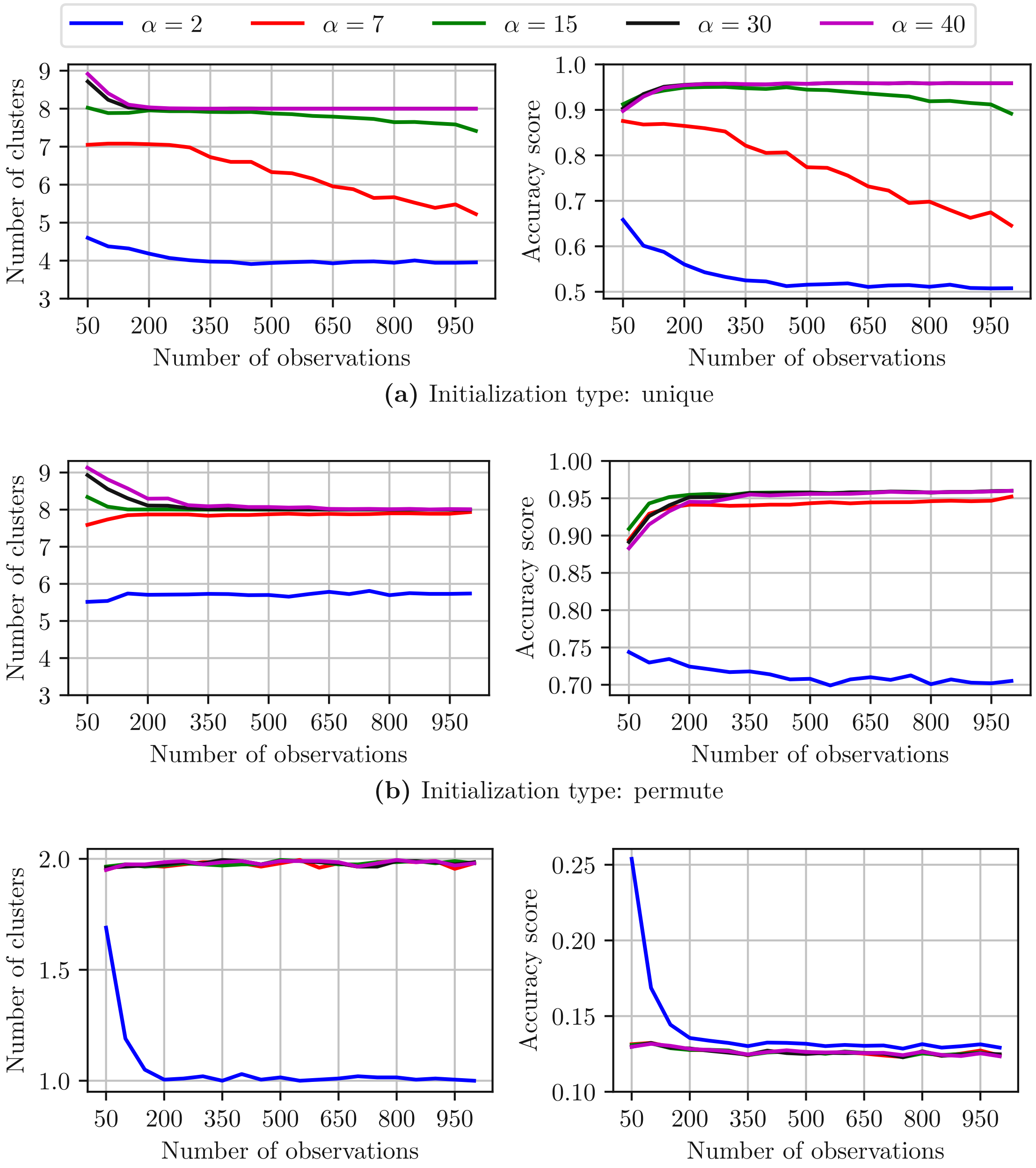
and the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \sigma^2_{\boldsymbol{\theta}^*}\mathbf{I}_2$ are chosen according to

$$\sigma^2_{\boldsymbol{\theta}^*} = \max(\mathrm{var}(x_{1,1}, \ldots, x_{N,1}), \mathrm{var}(x_{1,2}, \ldots, x_{N,2})). \tag{5.43}$$

With $\boldsymbol{\Sigma} = \mathbf{I}_2$ chosen to be the same as in the data generative model, there are only two input parameters of the CAVI algorithm, which we are in control of, namely the truncation parameter T and the hyperparameter α.

Influence of the Hyperparemater $\boldsymbol{\alpha}$

We now investigate the influence of the hyperparameter α on the clustering performance of the CAVI algorithm using the three initialization types unique, permute, and global described in



(a) Initialization type: unique

(b) Initialization type: permute

(c) Initialization type: global

Figure 5.2: Clustering performance of the CAVI algorithm for static MFGMs. Estimated number of clusters (first column) and accuracy score (second column) for $N = 50, 100, 150, \dots, 1000$ and different choices of the hyperparameter α. Each subfigure corresponds to a different initialization approach of the CAVI algorithm according to Section 5.2.3.

Section 5.2.3. In case of initialization type permute, the number of permutations is set to 10 and the truncation parameter T is set to 20 for both, permute and global. For a rough overview of the performance, the estimated number of clusters $\hat{L}$ and the accuracy score are shown in Figure 5.2. With $\hat{z}_n$ according to (5.38), the accuracy score is given by

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(\hat{z}_n = z_n).$$

The accuracy score indicates the number of correctly assigned observations in relation to the total number of observations. From Figure 5.2c, it follows that initializing the variational parameters of the CAVI algorithm according to the global procedure leads to a very poor performance. Therefore, the initialization type global is not considered further in our discussions.

In general, we observe an increasing estimated number of clusters with increasing hyperparameter α. Recall that the prior on the number of components is given by $p(K) = \text{Poisson}(K-1; \alpha)$. Thus, the a priori expected number of components is given by $\alpha + 1$. Since the ground truth regarding the number of components is known to be equal to eight, an obvious choice for the hyperparameter α is given by $\alpha = 7$. But for both initialization types, unique and permute, we find the values of α leading to the best overall clustering performance to be larger than 7, namely $\alpha = 30$ for initialization type unique and $\alpha = 15$ for initialization type permute. Although the number of clusters is overestimated for $N < 150$, the estimated number of clusters tends towards the true number of clusters as the size of the synthetic dataset grows, i.e., with increasing number of observations N. The corresponding accuracy scores are no smaller than 0.9, i.e., at least 90 % of the observations are assigned to the correct cluster. For smaller values of α, the initialization types demonstrate significantly different behaviour from one another. Whereas the estimated number of clusters grows with an increasing number of observations in the case of permute, it decreases in the case of unique.

Comparison of Initialization Types Unique and Permute

Figure 5.3 shows a comparison of initialization type unique with $\alpha = 30$ and permute with $\alpha = 15$, i.e., the hyperparameter for which the CAVI algorithm performs best for each initialization type. Again, the estimated number of clusters and the accuracy score are used to quantify the clustering performance. For a lower number of observations N, permute outperforms unique, which is indicated by a slightly lower estimated number of clusters and a slightly higher accuracy score. As soon as N reaches 200, the performance difference is negligible. Note that in both cases, the 95 % confidence interval decreases with increasing N, suggesting an increasing level of information content in the observations.

The average computation times of a CAVI run and the corresponding post processing step for a Python implementation (single core) executed on a system with an Intel Core i9-13900H CPU and 16 GB RAM are shown in Table 5.1. Recall that the truncation parameter T is set to be equal to the number of observations N in case of initialization type unique and thus, the number of variational parameter vectors $\boldsymbol{\gamma}_t$ and $\boldsymbol{\tau}_t$ to be computed and the dimension of the variational parameter vectors $\boldsymbol{\phi}_n \in \mathbb{R}^T$, for $n = 1, \ldots, N$, grow linearly with N. For permute on

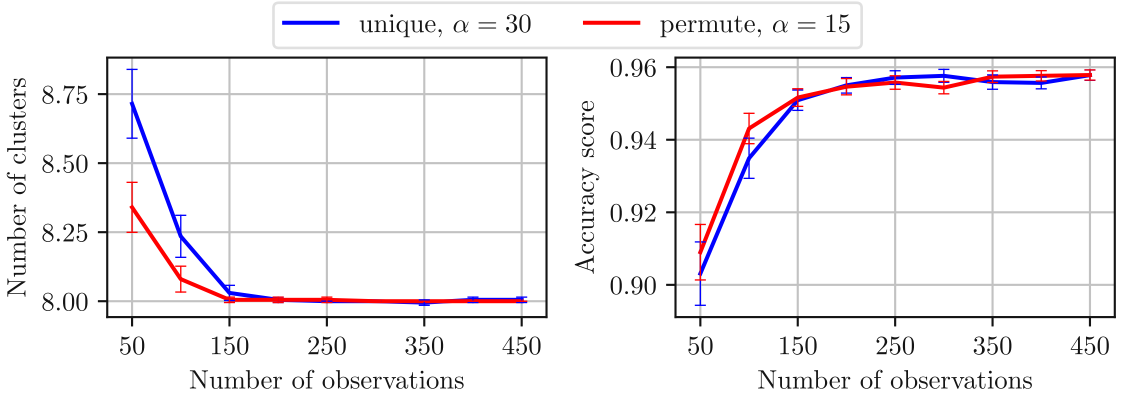


Figure 5.3: Estimated number of clusters and accuracy score with 95 % confidence interval for initialization type unique with $\alpha = 30$ and initialization type permute with $\alpha = 15$.

N	50	150	200	350	500	1000
Unique	13.68	40.23	64.45	150.50	357.11	1226.15
Permute	77.80	98.61	102.29	173.62	200.37	300.80

Table 5.1: Joint computation time of the CAVI algorithm and the post processing step (in ms) for initialization type permute with $T = 20$, $\alpha = 15$, and 10 permutations per MC run and initialization type unique with $\alpha = 30$ averaged over 200 MC runs.

the other hand, the truncation parameter T is fixed at some predefined number but the CAVI algorithm is repeated several times (number of permutations) in each MC run. This results in a larger computation time for a small to moderate number of observations N compared with unique. As soon as N gets large, the computational complexity using initialization type unique grows significantly and we observe smaller overall runtimes of permute compared with unique.

Finally, some specific clustering results for $N = 50$ and $N = 300$ using the CAVI algorithm for static MFGMs with initialization type permute are shown in Figure 5.4. Here, the presented results arise from a single MC run, i.e., from a specific synthetic dataset depicted as ground truth. For $N = 50$, the level of information in the observations tends to be low, since the population of each cluster in the corresponding artificial dataset is rather sparse. This may lead to the number of clusters being estimated incorrectly. As an example, let us consider the artificial dataset shown in Figure 5.4a. The upper lightblue observation and the most right orange observation are located far from their corresponding cluster means. Furthermore, the distances to observations assigned to different clusters are large as well. In consequence, the estimator assigns each observation to a new cluster and therefore the number of clusters is overestimated. The opposite behaviour can be observed in Figure 5.4b. Consisting of a single observation located far from the corresponding cluster mean but nearby the red cluster, the orange cluster is not present in the estimation result. Figure 5.4c shows a specific case with higher information content in the observations in the sense that the number of clusters is estimated correctly. For datasets with a larger number of observations and thus a higher level of information content, we over all observe a higher estimation accuracy. A specific example for $N = 300$ is presented in Figure 5.4d.

Influence of the Truncation Parameter T

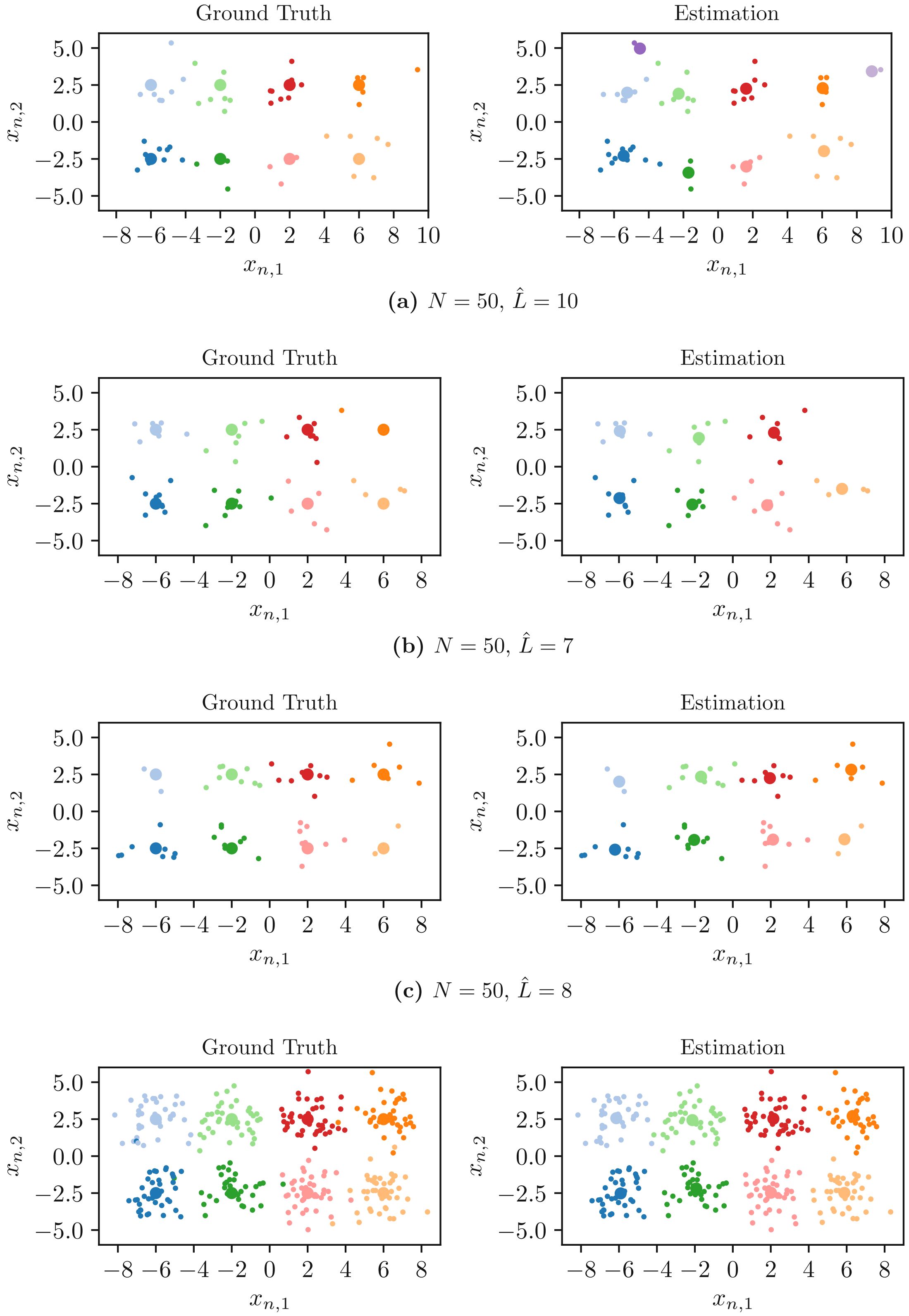
We now investigate the influence of the truncation parameter T on the clustering performance of the CAVI algorithm for static MFGMs with initialization type permute. Recall that the truncation parameter T depicts the number of components in the approximating model given in (4.16). Since our data generative model given in (5.40) consists of $K = 8$ components with equal mixture weights $\pi_k = \frac{1}{8}$, for $k = 1, \ldots, 8$, it is very likely that, for $N \geq 50$, each synthetic dataset drawn from (5.40) has eight clusters. Therefore, we do not consider the truncation parameter to be smaller than eight in our discussion.

Figure 5.5 shows the estimated number of clusters and the accuracy score averaged over 200 MC runs of the CAVI algorithm and the corresponding post processing step using initialization type permute with 10 permutations. In general, we observe an increasing estimated number of clusters with increasing truncation parameter T. If the number of observations is small, the number of clusters is overestimated for $T = 15$ and $T = 20$. Interestingly, this is not the case for $T = 50$. However, if large enough, the particular choice of T does not change the estimated number of clusters for $N \geq 200$. In other words, the influence of T is suppressed for a higher level of information content in the observations. Qualitatively, the same statement holds for the accuracy score.

We emphasize that choosing appropriate values for α and T when using initialization type permute should be viewed as a joint problem. Since the computational complexity increases with increasing truncation parameter T, keeping T as small as possible is desirable. The associated decrease in clustering performance can — at least partly — be compensated for by increasing the hyperparameter α. For illustration, we consider the cases where $T = 9$ and $T = 20$ for $\alpha = 15$, which have already been shown in Figure 5.5. By increasing α from 15 to 27 for $T = 9$, the clustering performance can be raised close to the level of $T = 20$, see Figure 5.6.

Evidence Lower Bound

Figure 5.7 shows the ELBO computed at each iteration of the CAVI algorithm using initialization types unique and permute for $N = 50$ and $N = 500$ observations. The specific iteration at which the algorithm is terminated, i.e., the condition (5.34) is fulfilled, is indicated by vertical dashed lines. In order to enable a fair comparison, the hyperparameter α is set to 30 in both cases,



(a) $N = 50,\ \hat{L} = 10$

(b) $N = 50,\ \hat{L} = 7$

(c) $N = 50,\ \hat{L} = 8$

(d) $N = 300,\ \hat{L} = 8$

Figure 5.4: Clustering performance using initialization type permute, $T = 20$, $\alpha = 15$, and 10 permutations. Left: Synthetic datasets for $N = 50$ and $N = 300$ generated according to (5.40) and labeled according to the ground truth. Right: MMSE estimated cluster means $\hat{\boldsymbol{\theta}}_l^*$, for $l = 1, \ldots, \hat{L}$ and observations $\boldsymbol{x}_n = (x_{n,1} \quad x_{n,2})^{\mathrm{T}}$ labeled using the MAP estimated indicator variables $\hat{z}_n$.

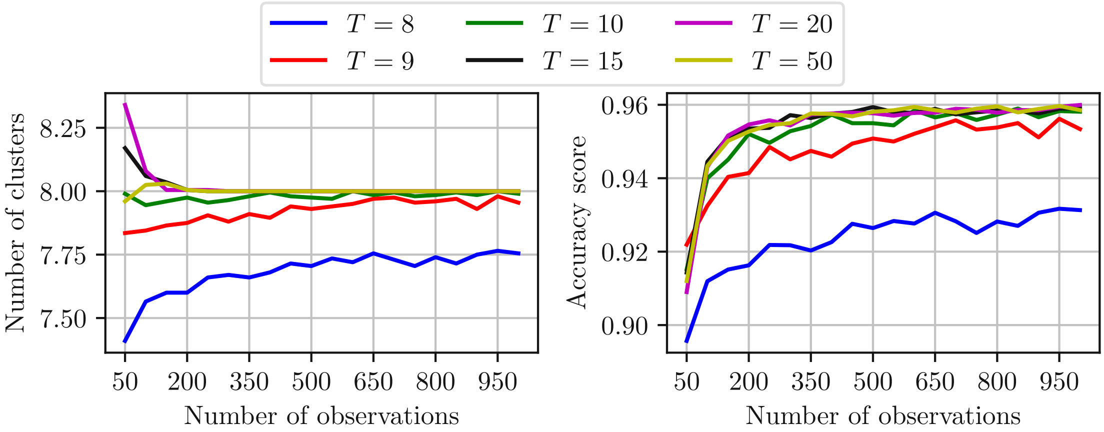


Figure 5.5: Clustering performance of the CAVI algorithm for static MFGMs with initialization type permute and $\alpha = 15$. Estimated number of clusters (left) and accuracy score (right) for $N = 50, 100, 150, \ldots, 1000$ and different choices of the truncation parameter T.

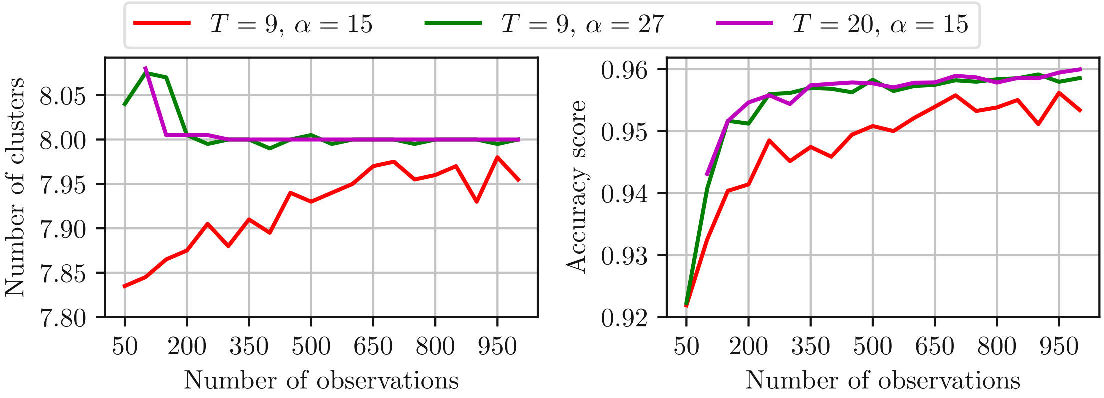


Figure 5.6: Improvement of the clustering performance for the specific choice of $T = 9$ by increasing the hyperparameter α.

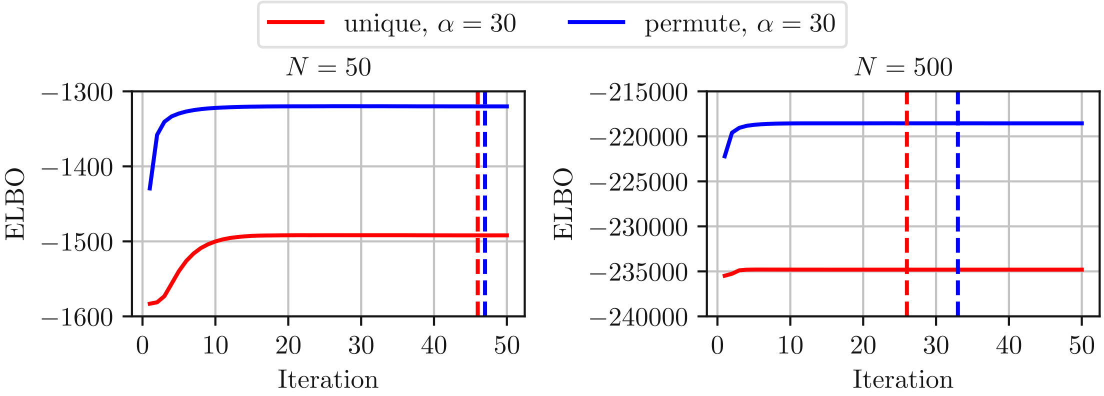


Figure 5.7: Evolution of the ELBO for initialization type permute with $T = 20$ and 10 permutations and initialization type unique. Dashed lines indicate the termination of the algorithm according to (5.34) with $\varepsilon = 10^{-10}$.

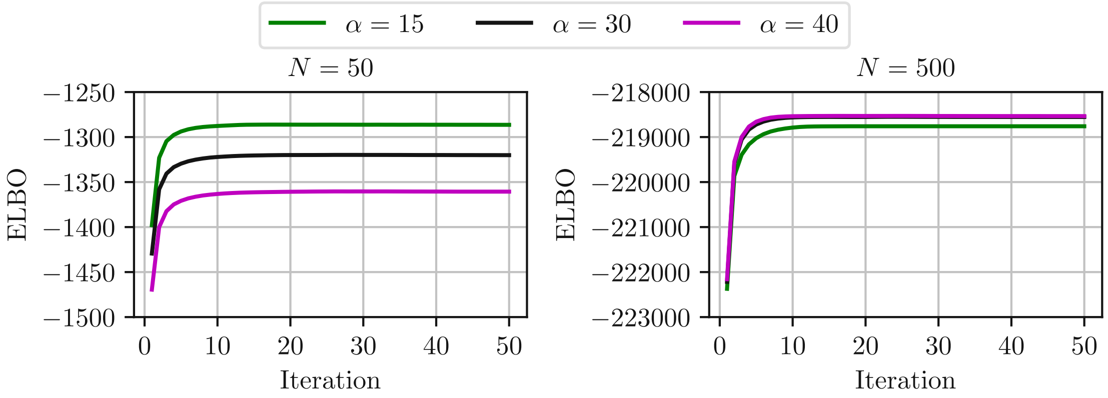


Figure 5.8: Evolution of the ELBO for initialization type permute with $T = 20$ and 10 permutations for different choices of the hyperparameter α.

	$N = 50$		$N = 500$	
	number of clusters	accuracy score	number of clusters	accuracy score
$\alpha = 15$	8.34	0.909	8.00	0.958
$\alpha = 30$	8.93	0.892	8.01	0.958
$\alpha = 40$	9.13	0.883	8.07	0.956

Table 5.2: Estimated number of clusters and accuracy score for initialization type permute with $T = 20$ and 10 permutations for different choices of the hyperparameter α.

unique and permute. For both initialization types, the CAVI algorithm converges faster for $N = 500$, which is again related to the higher level of information content in the observations. Furthermore, the ELBO using initialization type permute is larger for both cases, $N = 50$ and $N = 500$. Especially for $N = 500$, a significant gap between unique and permute is evident, even though the clustering performance is basically the same. Thus, it can not be inferred from the ELBO which initialization type is most suitable.

In Figure 5.8, we present ELBO curves for $N = 50$ and $N = 500$ using initialization type permute only, but with different choices of the hyperparameter α. The corresponding clustering results, i.e., the evaluated performance metrics, are given in Table 5.2. Note that these results are part of those already presented in Figure 5.2b. We observe that the particular choice of $\alpha = 15$ leads to the best clustering performance in both cases, $N = 50$ and $N = 500$, whereas $\alpha = 40$ performs worst. Since, for $N = 50$, the choice of $\alpha = 15$ results in the largest ELBO as well, one might tend to use the ELBO as decision basis for hyperparameter tuning. However, for $N = 500$, it is evident from Figure 5.8 that $\alpha = 15$ has the smallest ELBO. From this, we conclude that the ELBO, in general, cannot be used as a sole basis for hyperparameter tuning. This statement becomes even more evident by investigating the ELBO curves shown in Figure 5.9 and the corresponding clustering results given in Table 5.3 for initialization type unique. Here, for $N = 50$, the choice of the hyperparameter $\alpha = 15$ for the best clustering

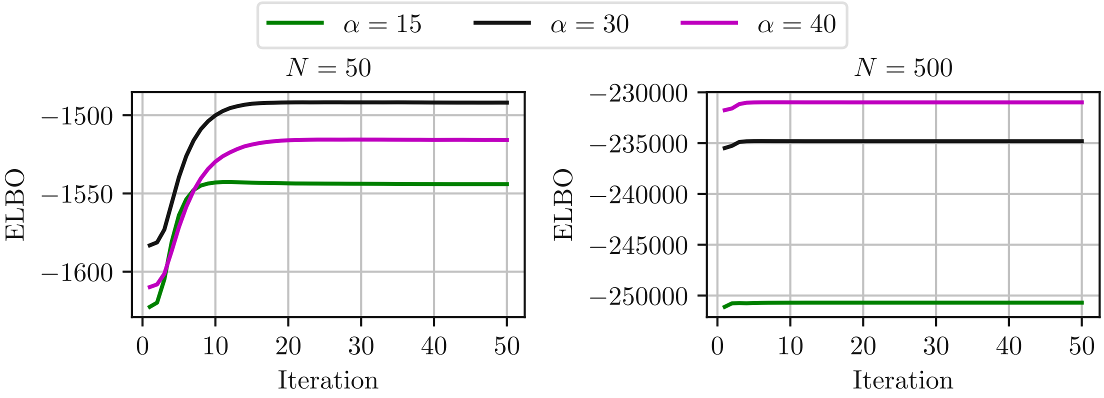


Figure 5.9: Evolution of the ELBO for initialization type unique for different choices of the hyperparameter α.

	$N = 50$		$N = 500$	
	number of clusters	accuracy score	number of clusters	accuracy score
$\alpha = 15$	8.03	0.912	7.88	0.944
$\alpha = 30$	8.72	0.903	8.00	0.957
$\alpha = 40$	8.92	0.898	8.00	0.957

Table 5.3: Estimated number of clusters and accuracy score for initialization type unique for different choices of the hyperparameter α.

performance does not coincide with that for the largest ELBO, i.e., $\alpha = 30$.

Comparison with the CAVI Algorithm for DPMs

We now compare the clustering performance of our CAVI algorithm for static MFGMs with the CAVI algorithm for DPMs from [16] specialized to the case of Gaussian component distributions with unknown means. Recall that, for a DPM model, the expected number of clusters grows logarithmically with the number of observations. More specifically, it is given by [17]

$$\bar{L} \approx \kappa \ln\left(1 + \frac{N}{\kappa}\right).$$

Thus, we expect the estimated number of clusters to grow with the number of observations when the CAVI algorithm for DPMs is applied to synthetic data from the finite Gaussian mixture model in (5.40). For simulation, the value of the concentration parameter κ is chosen such that $\bar{L} = 8$ for $N = 300$, i.e., $\kappa = 1.51$. For our CAVI algorithm, we use initialization type permute with 10 permutations, $\alpha = 15$ and $T = 20$. From Figure 5.10 it is demonstrated that the number of clusters in the observations is not reliably inferred by the CAVI algorithm for DPMs since, as expected, $\hat{L}$ grows with an increasing number of observations N. Correspondingly, the accuracy score goes down. More generally, the DPM model is not able to exploit the increasing information content in the data as the number of observations grows which can be inferred from the growing confidence intervals. On the other hand, we observe decreasing and even vanishing

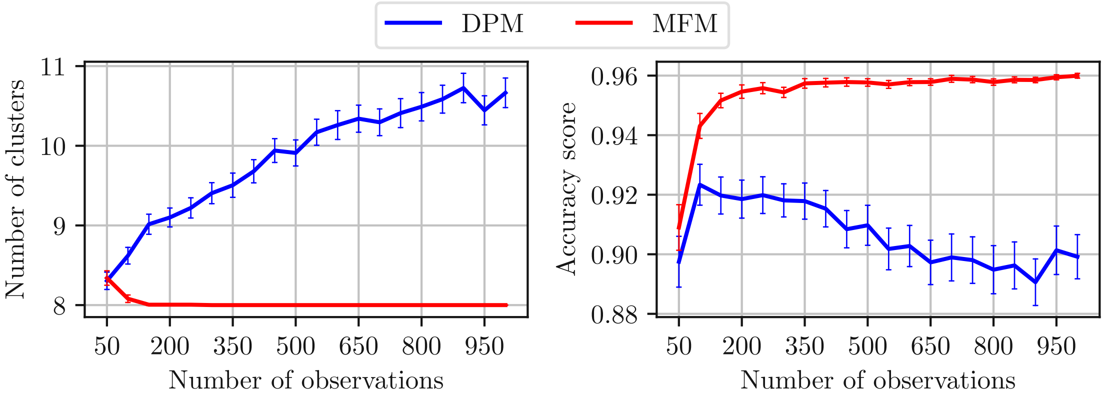


Figure 5.10: Comparison of the clustering performance of the CAVI algorithm for the DPM of Gaussians and the CAVI algorithm for the static MFGM. Estimated number of clusters (left) and accuracy score (right) with 95 % confidence intervals.

confidence intervals when our CAVI algorithm for static MFGMs is applied. We can conclude that, at least within the specified finite Gaussian mixture scenario, our novel CAVI algorithm for static MFGMs effectively resolves the inconsistency issue related to the inferred number of clusters associated with DPMs.

5.4.3 Old Faithful Geyser Dataset

Finally, we apply our CAVI algorithm for static MFGMs to the Old Faithful geyser dataset[1]. It consists of $N = 272$ observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, where each observation $\boldsymbol{x}_n = (x_{n,1} \quad x_{n,2})^{\mathrm{T}} \in \mathbb{R}^2$ contains the duration of a single eruption and the waiting time until the next eruption of the Old Faithful geyser, which is located in Yellowstone National Park, Wyoming, USA.

Note that the waiting times are roughly an order of magnitudes larger than the eruption durations. To avoid bias towards the waiting times in the estimation result, the CAVI algorithm is fed with standardized data. Standardization means that both, the eruption durations and the waiting times have zero mean and a standard deviation of one, i.e.,

$$\mu_1 = \frac{1}{N}\sum_{n=1}^{N} x_{n,1} = 0, \qquad \sigma_1 = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{n,1} - \mu_1)^2} = 1,$$
$$\mu_2 = \frac{1}{N}\sum_{n=1}^{N} x_{n,2} = 0, \qquad \sigma_2 = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{n,2} - \mu_2)^2} = 1.$$

The raw Old Faithful geyser dataset and the corresponding standardized version are shown in Figure 5.11.

Figure 5.12 presents the clustering result for the Old Faithful geyser dataset using the CAVI

[1] Source: `https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat`

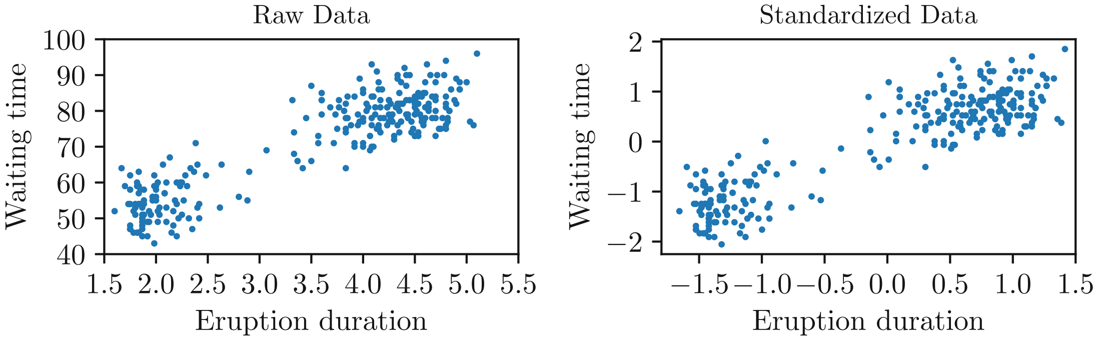


Figure 5.11: Raw Old Faithful geyser dataset (left) with eruption duration and waiting time until the next eruption in minutes and standardized Old Faithful geyser dataset (right).

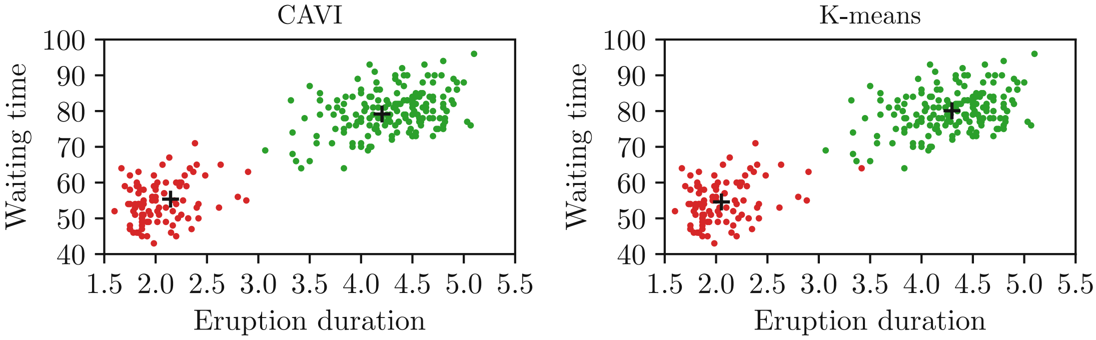


Figure 5.12: Clustered data from Old Faithful geyser using the CAVI algorithm for static MFGMs (left) and the K-means algorithm (right). Cluster means are indicated by black crosses.

algorithm for static MFGMs with initialization type permute, $\alpha = 8$, $T = 10$, and 10 permutations. The hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ are chosen according to (5.42) and (5.43). In addition, a clustering result using the K-means algorithm kmeans2[2] with $k = 2$ and random initialization of the centroids is shown for comparison. We see that the clustering results of both algorithms are the same, except the assignment of a single observation.

We note that the CAVI algorithm for static MFGMs forms two clusters regardless of the particular choice of α, as soon as $\alpha > 2$. Each of the clusters can be interpreted as a separate series of eruptions. The red cluster contains short eruptions lasting no longer than three minutes and is associated with shorter waiting times between these eruptions. The green cluster contains long lasting eruptions associated with longer waiting times. Both, the eruption duration and the time interval between eruptions have bimodal distributions. For example, the mean waiting times are given by 55 minutes following a short eruption and 80 minutes following a long eruption.

[2] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.vq.kmeans2.html`

Chapter 6

Conclusion

In this work, we investigated the mixture of finite mixtures (MFM) model based on [19] and [20]. The MFM model is a Bayesian mixture model in which the number of components is modeled as a random parameter with a specified prior. This model is therefore particularly useful in mixture scenarios with an unknown but finite number of components. We presented a generalized MFM model together with two special cases, the static and dynamic MFM models. For both cases, we derived relevant probability distributions such as the exchangeable partition probability function (EPPF), the distribution of the cluster sizes, and the prior distribution of the number of clusters. These distributions were used to compare the clustering behavior between the static and the dynamic MFM models as well as the Dirichlet process mixture (DPM) model. We furthermore discussed equivalent representations for the static MFM model, including the representation using the EPPF and the stick-breaking representation.

Next, we proposed a novel coordinate-ascent variational inference (CAVI) algorithm for estimating the parameters of the static MFM model using the stick-breaking representation. Our CAVI algorithm is suited to component distributions belonging to the exponential family of distributions and the corresponding conjugate prior distribution of the component parameters. Subsequently, we restricted the component distributions to be Gaussians with unknown means and known covariance matrices, resulting in a novel CAVI algorithm for the static mixture of finite Gaussian mixtures (MFGM) model. Finally, we evaluated the clustering performance of our CAVI algorithm using various hyperparameters and initializations for synthetic data corresponding to a mixture of eight equally weighted Gaussian components. We observed that, for suitable choices of the hyperparameter α and the truncation parameter T, at least $90\,\%$ of the observations are assigned to the correct cluster. Furthermore, the number of clusters in the observed data is correctly inferred if the number of observations is sufficiently large. In addition, to compare the clustering performance of our proposed MFGM-based CAVI algorithm with that of an established baseline method, we also performed simulations using the CAVI algorithm for

DPM models from [16] specialized to Gaussian components. This comparison revealed that our MFGM-based CAVI algorithm estimates the number of clusters and the cluster assignments more accurately than the DPM-based CAVI algorithm when applied to our synthetic dataset.

The clustering results for static MFGM models summarized above are based on the assumption that the component distributions are perfectly specified, i.e., the component distributions in the MFM model match the component distributions that generated the data. In many practical applications, this assumption does not hold since the component distributions for real-world data do not conform to a convenient parametric form. In such cases, it seems that, at least for static MFMs, the posterior distribution of the number of components does not concentrate at the true number of components. In particular, it was shown in [42] that the posterior distribution of the number of components converges to zero, i.e., $p(K|\boldsymbol{x}) \to 0$, for any finite number of components $K \in \mathbb{N}$, as the number of observations N approaches infinity. This result was also demonstrated in [42] by simulations using the split-merge sampler for static MFM models from [19]. This issue does not seem to be of practical relevance in clustering tasks since the number of components K is typically not of interest and thus not inferred. For example, the application of our CAVI algorithm for static MFGMs to the Old Faithful geyser dataset resulted in an estimated number of clusters $\hat{L} = 2$, which is a reasonable result. However, in future work, it could still be interesting to investigate whether our CAVI algorithm for static MFGMs exhibits a similar behavior regarding the posterior distribution of the number of components as that described in [42].

The proposed CAVI algorithm is based on the assumption that the variational distribution is a member of the truncated mean-field family of distributions. The truncation parameter T is a hyperparameter that restricts the underlying static MFM model to at most T components. Our CAVI algorithm could be improved by removing the truncation assumption. In [43], a truncation-free stochastic variational inference (VI) algorithm for Bayesian nonparametric models was proposed, which, slightly adapted, might be applicable to static MFM models as well.

Finally, to exploit the full flexibility of the MFM model, advanced VI methods [44] could be used in future work. This involves developing VI algorithms that do not rely on the stick-breaking representation of the static MFM. More generally, such VI algorithms could serve as a basis for developing efficient inference methods for dynamic MFMs.

Appendix

Proof of Proposition 3.6: Derivations and verifications in this proof are based on [45, Chapter 5]. Let us consider a time dependent counting process, where the random number $N(t)$ counts the total number of arrivals up to the time $t \in [0, \infty)$. In the context of the stick-breaking analogy, $N(t)$ counts the number of pieces broken off from the stick and thus, breaking off a stick is referred to as arrival. Instead of time, t indicates the location on the stick. We denote the interarrival time, i.e., the time elapsed between two consecutive arrivals indexed by $k-1$ and k, $k \in \mathbb{N}$, by the i.i.d. exponential random variable Y_k with rate parameter $\alpha > 0$, i.e.,

$$f_{Y_k}(t) = \alpha e^{-\alpha t} \quad \text{for} \quad k = 1, 2, \dots \tag{A.1}$$

The corresponding cdf is given by

$$\mathrm{P}\{Y_k \leq t\} = F_{Y_k}(t) = \begin{cases} 1 - e^{-\alpha t} & t \geq 0, \\ 0 & \text{else.} \end{cases} \tag{A.2}$$

As it can be seen from Figure A.1, it follows that the arrival times T_k, for $k = 1, 2, \dots$, i.e., the locations of the breakpoints on the stick, are given by

$$T_k = \sum_{i=1}^{k} Y_i. \tag{A.3}$$

More specifically, T_k is a sum of k random variables $Y_1, \dots, Y_k$ i.i.d. according to (A.1). Thus, T_k is distributed according to a gamma distribution with shape parameter k and rate parameter α, i.e.,

$$f_{T_k}(t) = \frac{1}{(k-1)!} \alpha^k t^{k-1} e^{-\alpha t}, \tag{A.4}$$

where $(k-1)! = \Gamma(k)$ since $k \in \mathbb{N}$. We next verify (A.4) by using mathematical induction. For the case where $k = 1$ it is trivial to show that (A.1) and (A.4) are the same and, thus, $T_1 = Y_1$. According to (A.3), the time of the next arrival T_{k+1} is given by $T_{k+1} = T_k + Y_{k+1}$. Since T_k and Y_{k+1} are statistically independent, the density $f_{T_{k+1}}(t)$ can be calculated by convolving the products of the respective pdfs $f_{T_k}(t)$ and $f_{Y_{k+1}}(t)$:

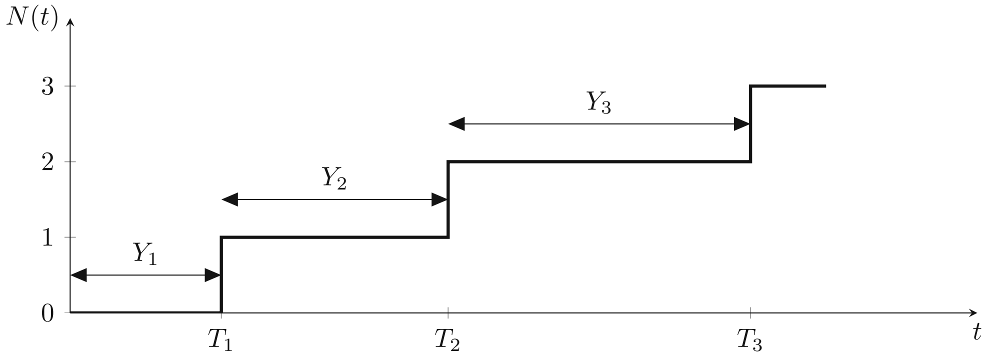


Figure A.1: A graphical depiction of the counting process $N(t)$ with arrival times T_k and interarrival times Y_k for three arrivals.

$$f_{T_{k+1}}(t) = \int_0^\infty f_{Y_{k+1}}(t-s) f_{T_k}(s)\, ds.$$

Inserting (A.1) and (A.4) yields

$$f_{T_{k+1}}(t) = \int_0^t \alpha e^{-\alpha(t-s)} \frac{1}{(k-1)!} \alpha^k s^{k-1} e^{-\alpha s}\, ds = \frac{1}{(k-1)!} \alpha^{k+1} e^{-\alpha t} \int_0^t s^{k-1}\, ds = \frac{1}{k!} \alpha^{k+1} t^k e^{-\alpha t}. \tag{A.5}$$

Now by replacing $k+1$ with k in (A.5), the resulting expression is the same as (A.4) which proves that $T_k = \sum_{i=1}^k Y_i$ has a gamma distribution.

We next derive the probability that the total number of arrivals up to time t is equal to k, i.e., $\mathrm{P}\{N(t)=k\}$. A convenient way to compute $\mathrm{P}\{N(t)=k\}$ is to condition on the time of the kth arrival T_k. As shown previously, T_k has a gamma distribution and therefore is a continuous random variable. Applying the law of total probability to our problem yields

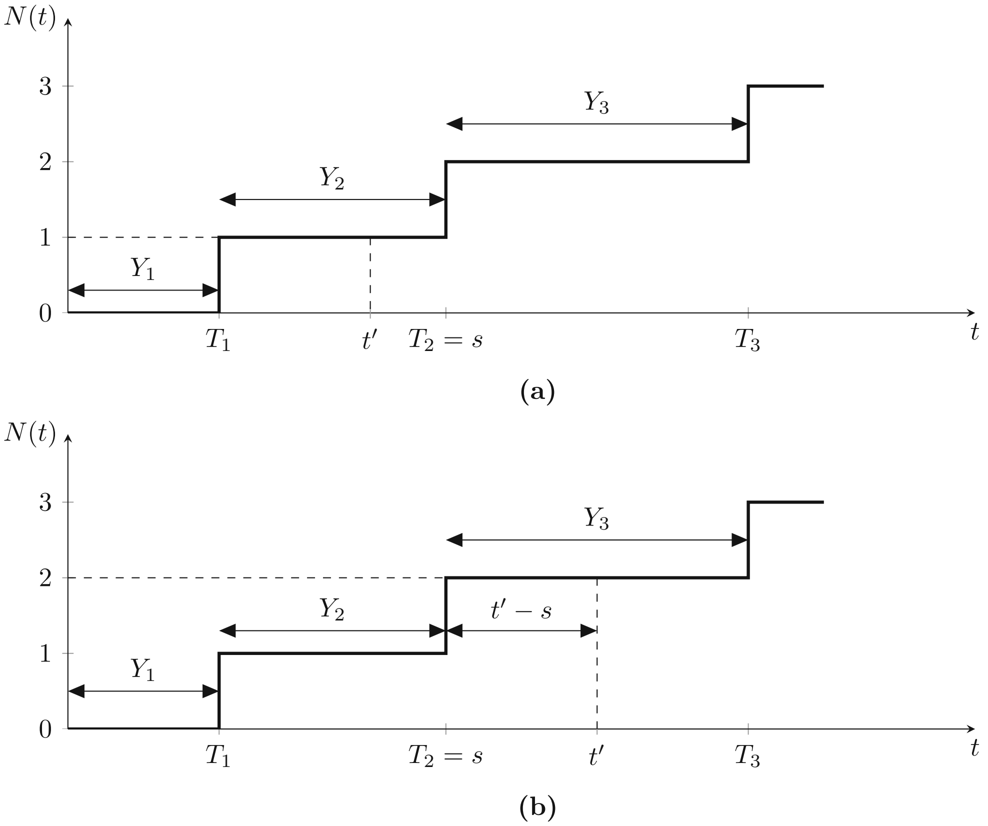
$$\mathrm{P}\{N(t)=k\} = \int_0^\infty \mathrm{P}\{N(t)=k|T_k=s\} f_{T_k}(s)\, ds = \int_0^t \mathrm{P}\{N(t)=k|T_k=s\} f_{T_k}(s)\, ds, \tag{A.6}$$

where the second step is due to $\mathrm{P}\{N(t)=k|T_k=s\}=0$ for $s>t$. An example of this particular case is illustrated in Figure A.2a. Let us now focus on $\mathrm{P}\{N(t)=k|T_k=s\}$ for $0<s<t$: a total number of k arrivals up to time t given that the time of the kth arrival equals s implies that there is no arrival observed in the interval $t-s$. In other words, the next interarrival time Y_{k+1} has to be larger than $t-s$ (an example of such case is shown in Figure A.2b). Thus, we have

$$\mathrm{P}\{N(t)=k|T_k=s\} = \mathrm{P}\{Y_{k+1} > t-s|T_k=s\}.$$

Since T_k and Y_{k+1} are statistically independent, the events $Y_{k+1} > t-s$ and $T_k = s$ are independent as well. Hence,

$$\mathrm{P}\{N(t)=k|T_k=s\} = \mathrm{P}\{Y_{k+1} > t-s\} = 1 - F_{Y_{k+1}}(t-s) = e^{-\alpha(t-s)}, \tag{A.7}$$



(a)

(b)

Figure A.2: A graphical depiction of the event $\{N(t') = 2 | T_2 = s\}$. (a) The case where $s > t'$. The probability $\mathrm{P}\{N(t') = 2 | T_2 = s\}$ is zero since the time of the second arrival T_2 is always larger than t'. (b) The case where $s < t'$. The event $\{N(t') = 2 | T_2 = s\}$ is the same as the event $\{Y_3 > t' - s | T_2 = s\}$.

where (A.2) is used. Inserting (A.7) and (A.4) into (A.6) yields

$$\mathrm{P}\{N(t) = k\} = \int_0^t e^{-\alpha(t-s)} \frac{1}{(k-1)!} \alpha^k s^{k-1} e^{-\alpha s}\, ds = \frac{1}{(k-1)!} \alpha^k e^{-\alpha t} \int_0^t s^{k-1}\, ds = \frac{(\alpha t)^k}{k!} e^{-\alpha t}. \tag{A.8}$$

We conclude that $N(t)$ is a Poisson random variable with rate αt. The underlying counting process is a Poisson process with rate α.

Recall the stick-breaking representation of the static MFM model given in (3.62). It follows from (3.62a)–(3.62d), that we break off $\widetilde{K} - 1$ i.i.d. exponential pieces from a unit-length stick, i.e., $N(t = 1) = \widetilde{K} - 1$, while the remaining portion of the original unit-length stick is considered the $\widetilde{K}$th piece. An example for $\widetilde{K} = 4$ pieces is shown in Figure A.3.

According to (A.8), the probability of breaking off $\widetilde{K} - 1$ i.i.d. exponential pieces from the stick up to length $t = 1$ is given by

$$\mathrm{P}\{N(t = 1) = \widetilde{K} - 1\} = \frac{\alpha^{\widetilde{K}-1}}{(\widetilde{K} - 1)!} e^{-\alpha}. \tag{A.9}$$

Therefore, the stick-breaking representation (3.62) induces through (A.9) the translated (cf.

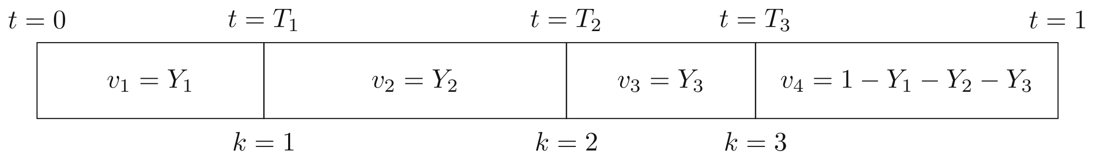
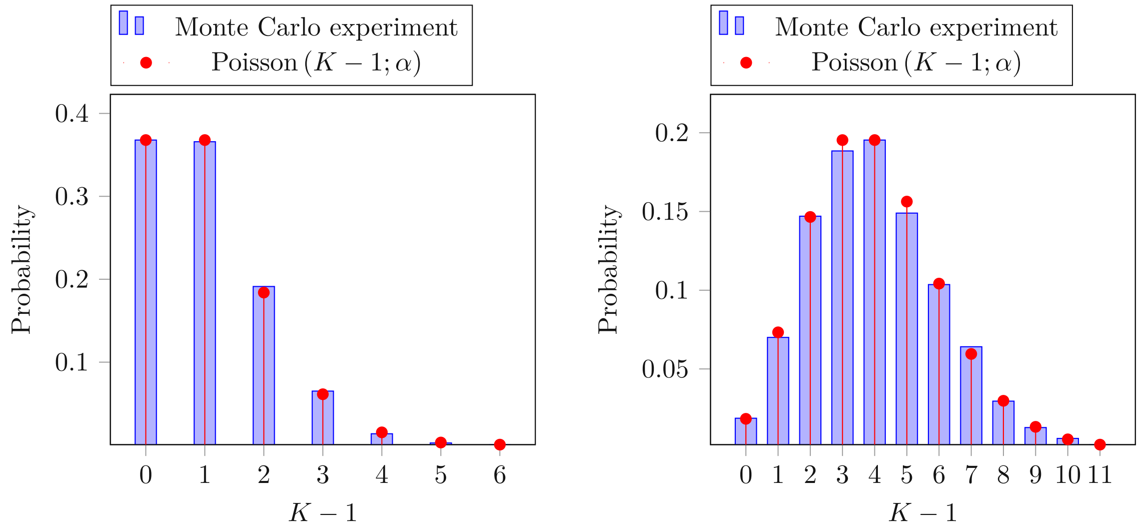


Figure A.3: Breaking a unit-length stick into $\widetilde{K} = 4$ pieces. From the perspective of the counting process, we have $N(t = 1) = \widetilde{K} - 1 = 3$, since the length of the last piece v_4 occurs outside of the counting process and is deterministically set according to v_1, v_2 and v_3.



(a) $\alpha = 1$

(b) $\alpha = 4$

Figure A.4: Illustration of the translated Poisson prior on the number of components induced by the stick-breaking representation of the static MFM for rate parameters $\alpha = 1$ in (a) and $\alpha = 4$ in (b). The histogram in blue is based on realizations of $\widetilde{K}-1$ according to (3.62a)–(3.62b). The corresponding Poisson pmf (A.10) is shown in red.

Section 3.1) Poisson prior $K - 1 \sim p_t(\cdot) = \text{Poisson}(\cdot; \alpha)$ on the number of components K in the static MFM model given in (3.58). The prior pmf $p(K)$ is obtained by evaluating the translated prior pmf at $K - 1$, i.e.,

$$p(K) = p_t(K - 1) = \text{Poisson}(K - 1; \alpha). \tag{A.10}$$

In addition, we provide experimental evidence using Monte Carlo approximation, which is illustrated in Figure A.4. The histogram in blue is based on 10000 realizations of $\widetilde{K} - 1$ according to (3.62a)–(3.62b). It can be observed, that the histogram matches the corresponding Poisson distribution pretty well.

It remains to show that the stick-breaking representation of the static MFM model induces a symmetric Dirichlet prior with hyperparameter $\boldsymbol{\beta} = \mathbf{1}_K$ on the mixture weights $\boldsymbol{\pi}$ in the ”original” representation of the static MFM model given in (3.58). Based on [46], we start by deriving the Dirichlet distribution from scratch. Let x_k, for $k = 1, \ldots, K$ and $K \geq 2$, be independent random variables each distributed according to the gamma distribution $\mathcal{G}(x_k; \beta_k, \alpha)$ with shape and rate parameters β_k and α, respectively. Thus, the joint pdf $f(x_1, \ldots, x_K)$ is given by

$$f(x_1, \ldots, x_K) = \prod_{k=1}^{K} f(x_k) = \prod_{k=1}^{K} \frac{1}{\Gamma(\beta_k)} \alpha^{\beta_k} x_k^{\beta_k - 1} e^{-\alpha x_k}, \tag{A.11}$$

where we inserted the expression for the gamma distribution for $f(x_k)$. We next perform a multivariate variable transformation and compute the resulting joint pdf $f(y_1, \ldots, y_{K-1})$. Therefore, let

$$y_k = \frac{x_k}{x_1 + x_2 + \cdots + x_K} \quad \text{for } k = 1, \ldots, K-1, \tag{A.12}$$

$$y_K = x_1 + x_2 + \cdots + x_K. \tag{A.13}$$

The inverse transformation is given by

$$x_k = y_k y_K \quad \text{for } k = 1, \ldots, K-1, \tag{A.14}$$

$$x_K = y_K \left(1 - \sum_{k=1}^{K-1} y_k\right), \tag{A.15}$$

and for the corresponding Jacobian we obtain

$$\boldsymbol{J} = \begin{pmatrix} \partial x_1/\partial y_1 & \ldots & \partial x_1/\partial y_K \\ \vdots & \ddots & \vdots \\ \partial x_K/\partial y_1 & \ldots & \partial x_K/\partial y_K \end{pmatrix} = \begin{pmatrix} y_K & 0 & \ldots & 0 & y_1 \\ 0 & y_K & \ldots & 0 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & y_K & y_{K-1} \\ -y_K & -y_K & \ldots & -y_K & 1 - y_1 - \cdots - y_{K-1} \end{pmatrix}.$$

Inserting (A.14) and (A.15) into (A.11) and multiplying with $\det(\boldsymbol{J}) = y_K^{K-1}$, we obtain the joint pdf $f(y_1, \ldots, y_K)$

$$\begin{aligned} f(y_1, \ldots, y_K) &= \det(\boldsymbol{J}) \prod_{k=1}^{K} \frac{1}{\Gamma(\beta_k)} \alpha^{\beta_k} x_k^{\beta_k - 1} e^{-\alpha x_k} \\ &= y_K^{K-1} \left(\prod_{k=1}^{K-1} \frac{\alpha^{\beta_k}}{\Gamma(\beta_k)} (y_k y_K)^{\beta_k - 1} e^{-\alpha y_k y_K} \right) \\ &\quad \times \frac{\alpha^{\beta_K}}{\Gamma(\beta_K)} \left(y_K \left(1 - \sum_{k=1}^{K-1} y_k \right) \right)^{\beta_K - 1} e^{-\alpha y_K \left(1 - \sum_{k=1}^{K-1} y_k\right)} \\ &= \underbrace{\frac{1}{\prod_{k=1}^{K} \Gamma(\beta_k)} \left(\prod_{k=1}^{K-1} y_k^{\beta_k - 1} \right) \left(1 - \sum_{k=1}^{K-1} y_k \right)^{\beta_K - 1}}_{\mathrm{C}} \alpha^b y_K^{b-1} e^{-\alpha y_K}, \end{aligned} \tag{A.16}$$

where

$$b = \sum_{k=1}^{K} \beta_k. \tag{A.17}$$

Marginalizing out y_K from (A.16) yields

$$f(y_1, \dots, y_{K-1}) = C \int_0^\infty \alpha^b y_K^{b-1} e^{-\alpha y_K} \, dy_K = C\,\Gamma(b) \int_0^\infty \underbrace{\frac{1}{\Gamma(b)} \alpha^b y_K^{b-1} e^{-\alpha y_K}}_{A(y_K; b, \alpha)} \, dy_K, \tag{A.18}$$

where we multiplied by $\Gamma(b)/\Gamma(b)$ in the second step. We note that $A(y_K; b, \alpha)$ is again a gamma distribution. Hence, $\int_0^\infty A(y_K; b, \alpha)\, dy_K = 1$ and (A.18) becomes

$$f(y_1, \dots, y_{K-1}) = C\,\Gamma(b). \tag{A.19}$$

Substituting back for b (see (A.17)) and C (see (A.16)) into (A.19), we obtain

$$f(y_1, \dots, y_{K-1}) = \frac{\Gamma\left(\sum_{k=1}^K \beta_k\right)}{\prod_{k=1}^K \Gamma(\beta_k)} \left(\prod_{k=1}^{K-1} y_k^{\beta_k - 1}\right) \left(1 - \sum_{k=1}^{K-1} y_k\right)^{\beta_K - 1}, \tag{A.20}$$

which is a Dirichlet distribution with parameter vector $\boldsymbol{\beta} = (\beta_1 \;\; \cdots \;\; \beta_K)^{\mathrm{T}} \in \mathbb{R}^K$, where $\beta_k > 0$, $y_k \geq 0$ and $\sum_{k=1}^{K-1} y_k < 1$. Alternatively, the Dirichlet pdf is commonly defined as

$$\mathcal{D}(\boldsymbol{y}; \boldsymbol{\beta}) = \frac{\Gamma\left(\sum_{k=1}^K \beta_k\right)}{\prod_{k=1}^K \Gamma(\beta_k)} \prod_{k=1}^K y_k^{\beta_k - 1}, \tag{A.21}$$

where $\beta_k > 0$, $y_k \geq 0$ and $\sum_{k=1}^K y_k = 1$. This implies that $y_K = 1 - \sum_{k=1}^{K-1} y_k$ and therefore, (A.20) and (A.21) can be used interchangeably. We emphasize that, although each individual y_k, for $k = 1, \dots, K-1$, clearly depends on y_K (cf. (A.12)–(A.13)), the joint pdf $f(y_1, \dots, y_{K-1})$ given by (A.20) is independent of y_K. In other words, a Dirichlet distribution can be constructed using $K-1$ properly normalized random variables $y_1, \dots, y_{K-1}$ such that $\sum_{k=1}^{K-1} y_k < 1$ and the shape parameters $\beta_1, \dots, \beta_K$ of the K underlying independent gamma distributions.

It follows from (3.62b)–(3.62d) that $\sum_{k=1}^{\widetilde{K}-1} \widetilde{\pi}_k < 1$ and $\widetilde{\pi}_{\widetilde{K}} = 1 - \sum_{k=1}^{\widetilde{K}-1} \widetilde{\pi}_k$. Since the exponential distribution is a special case of the gamma distribution, i.e., $\mathcal{E}(\cdot; \alpha) = \mathcal{G}(\cdot; 1, \alpha)$, it follows from (3.62a) and (3.62b) that the construction of the mixture weights $\widetilde{\boldsymbol{\pi}}$ is based on i.i.d. gamma variables $v_k \overset{\text{i.i.d.}}{\sim} \mathcal{G}(v_k; 1, \alpha)$, for $k = 1, \dots, \widetilde{K}$. Therefore, the joint distribution of the mixture weights $f(\widetilde{\pi}_1, \dots, \widetilde{\pi}_{\widetilde{K}-1})$ is given by (A.20) with $\beta_k = 1$ for $k = 1, \dots, \widetilde{K}$, $y_k = \widetilde{\pi}_k$ and $K = \widetilde{K}$. We conclude that the stick-breaking representation of the static MFM model (3.62) induces a symmetric Dirichlet prior with hyperparameter $\boldsymbol{\beta} = \mathbf{1}_K$ on the mixture weights in the "original" static MFM model given in (3.58).

Bibliography

[1] G. Celeux, “Mixture models for classification”, in *Advances in Data Analysis*, Berlin, Germany: Springer, 2007, pp. 3–14.

[2] K. Roeder and L. A. Wasserman, “Practical Bayesian density estimation using mixtures of normals”, *Journal of the American Statistical Association*, vol. 92, pp. 894–902, 1997.

[3] M. Niknejad, H. Rabbani, M. Babaie-Zadeh, and C. Jutten, “Image interpolation using Gaussian Mixture Models with spatially constrained patch clustering”, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1613–1617.

[4] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, “Model-based clustering and data transformations for gene expression data”, *Bioinformatics*, vol. 17, no. 10, pp. 977–987, Oct. 2001.

[5] A. Bietti and L. Chizat, *Inference in Dirichlet process mixtures with applications to text document clustering*, Jan. 2014. [Online]. Available: `https://alberto.bietti.me/files/dpmixtures.pdf`, (visited on 2023-05-04).

[6] K. Pearson, “Contributions to the mathematical theory of evolution”, *Philosophic Transactions of the Royal Society of London. A*, vol. 185, pp. 71–100, 1894.

[7] S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert, Eds., *Handbook of Mixture Analysis*. FL, USA: CRC Press, 2018.

[8] A. Nobile, “Bayesian Analysis of Finite Mixture Distributions”, Ph.D. dissertation, Carnegie Mellon University, 1994.

[9] S. Richardson and P. J. Green, “On Bayesian analysis of mixtures with an unknown number of components”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 59, no. 4, pp. 731–792, 1997.

[10] A. Nobile, “On the posterior distribution of the number of components in a finite mixture”, *The Annals of Statistics*, vol. 32, no. 5, pp. 2044–2073, 2004.

[11] P. McCullagh and J. Yang, “How many clusters?”, *Bayesian Analysis*, vol. 3, no. 1, pp. 101–120, 2008.

[12] P. J. Green, “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”, *Biometrika*, vol. 82, no. 4, pp. 711–732, Dec. 1995.

[13] P. Dellaportas and I. Papageorgiou, “Multivariate mixtures of normals with unknown number of components”, *Statistics and Computing*, vol. 16, no. 1, pp. 57–68, 2006.

[14] 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.

[15] S. Jain and R. M. Neal, “Splitting and merging components of a nonconjugate Dirichlet process mixture model”, *Bayesian Analysis*, vol. 2, no. 3, pp. 445–472, 2007.

[16] D. M. Blei and M. I. Jordan, “Variational inference for Dirichlet process mixtures”, *Bayesian Analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[17] Y. W. Teh, “Dirichlet Process”, in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2010, pp. 280–287.

[18] J. W. Miller and M. T. Harrison, “A simple example of Dirichlet process mixture inconsistency for the number of components”, in *Advances in Neural Information Processing Systems*, vol. 26, MIT Press, 2013.

[19] 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, 2018.

[20] 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.

[21] P. J. Green, *Introduction to finite mixtures*, 2018. arXiv: `1705.01505 [stat.ME]`.

[22] V. Melnykov and R. Maitra, “Finite mixture models and model-based clustering”, *Statistics Surveys*, vol. 4, pp. 80–116, 2010.

[23] L. Dietz, “Directed factor graph notation for generative models”, Max Planck Institute for Informatics, Saarbrücken, Germany, Tech. Rep., 2010.

[24] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*. NY, USA: Springer, 2006.

[25] B. G. Lindsay, *Mixture Models: Theory, Geometry and Applications*. Hayward, USA: Institute of Mathematical Statistics, 1995.

[26] B. Kreidl, “Bayesian Nonparametric Inference in State-Space Models with an Application to Extended Target Tracking”, M.S. thesis, TU Wien, 2021.

[27] F. Hlawatsch, *Parameter Estimation Methods*, Lecture Notes, Course 389.119, Institute of Telecommunications, TU Wien, Mar. 2020.

[28] 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, Apr. 2017.

[29] G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün, “Model-based clustering based on sparse finite Gaussian mixtures”, *Statistics and Computing*, vol. 26, no. 1, pp. 303–324, 2016.

[30] J. Pitman, “Exchangeable and partially exchangeable random partitions”, *Probability Theory and Related Fields*, vol. 102, no. 2, pp. 145–158, 1995.

[31] A. Gnedin and J. Pitman, “Exchangeable Gibbs partitions and Stirling triangles”, *Journal of Mathematical Sciences*, vol. 138, no. 3, pp. 5674–5685, 2006.

[32] P. J. Green and S. Richardson, “Modelling heterogeneity with and without the Dirichlet process”, *Scandinavian Journal of Statistics*, vol. 28, no. 2, pp. 355–375, 2001.

[33] J. Pitman, “Some developments of the Blackwell-Macqueen urn scheme”, *Lecture Notes-Monograph Series*, vol. 30, pp. 245–267, 1996.

[34] P. Orbanz, *Lecture Notes on Bayesian Nonparametrics*, May 2014. [Online]. Available: `http://www.gatsby.ucl.ac.uk/~porbanz/papers/porbanz_BNP_draft.pdf`, (visited on 2023-10-09).

[35] J. Sethuraman, “A constructive definition of Dirichlet priors”, *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[36] J. Paisley, A. Zaas, C. Woods, G. Ginsburg, and L. Carin, “A stick-breaking construction of the beta process”, in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, Aug. 2010, pp. 847–854.

[37] A. Nobile, *Bayesian finite mixtures: A note on prior specification and posterior computation*, Technical Report 05-3, Department of Statistics, University of Glasgow, 2005.

[38] P. Weiss, “L’hypothèse du champ moléculaire et la propriété ferromagnétique”, *J. Phys. Theor. Appl.*, vol. 6, no. 1, pp. 661–690, 1907.

[39] S. Plummer, D. Pati, and A. Bhattacharya, “Dynamics of coordinate ascent variational inference: A case study in 2D ising models.”, *Entropy*, vol. 22, no. 11, 2020.

[40] F. Hlawatsch, *Bayesian Machine Learning*, Lecture notes, Course 389.207, Institute of Telecommunications, TU Wien, 2022.

[41] J. Soch, *The Book of Statistical Proofs*, Proof 131: Expectation of a quadratic form, 2020. [Online]. Available: `https://statproofbook.github.io/P/mean-qf.html`, (visited on 2023-10-20).

[42] D. Cai, T. Campbell, and T. Broderick, “Finite mixture models do not reliably learn the number of components”, in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[43] C. Wang and D. M. Blei, “Truncation-free stochastic variational inference for Bayesian nonparametric models”, *Advances in Neural Information Processing Systems*, vol. 1, pp. 413–421, Jan. 2012.

[44] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, “Advances in variational inference”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2008–2026, 2017.

[45] S. M. Ross, *Introduction to Probability Models*, Twelfth edition. London, UK: Elsevier Academic Press, 2019.

[46] J. Lin, “On the Dirichlet Distribution”, M.S. thesis, Queen’s University, Sep. 2016.