Uncertainty in radar emitter classification and clustering

Guillaume Revillon

université PARIS-SACLAY

THALES

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

NNT : 2019SACLS098

Thèse de doctorat

Uncertainty in radar emitter classification and clustering

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

Ecole doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Traitement du signal et des images

Thèse présentée et soutenue à Gif-sur-Yvette, le 18 avril 2019, par

GUILLAUME REVILLON

Composition du Jury :

Arthur Tenenhaus Professeur, CentraleSupélec (L2S)	Président
Nathalie Peyrard Directrice de Recherches, INRA Toulouse	Rapporteur
Charles Bouveyron Professeur, Université Côte d'Azur (Laboratoire J.A. Dieudonné)	Rapporteur
Paul Honeine Professeur, Université de Rouen Normandie (LITIS)	Examinateur
Cyrille Enderli Ingénieur, Thales DMS	Examinateur
Charles Soussen Professeur, CentraleSupélec (L2S)	Directeur de thèse
Ali Mohammad-Djafari Directeur de Recherche (CNRS)	Invité
Jean-François Grandin Ingénieur, Thales DMS	Invité

Acknowledgements

I would like to thank Technical Department of Thales Defense Mission Systems (DMS) and Inverse Problems Group (GPI) from Laboratory of Signals and Systems (L2S) for giving me the opportunity to realise a PhD.

Into the Technical Department, I mostly thank my advisors Cyrille Enderli and Jean-François Grandin for introducing me to Electronic Warfare domain and guiding me during these 3 years. I have really appreciated their motivation and their scientific open-mindedness.

Into the Inverse Problems Group, I would also like to truly thank my academic supervisors Ali Mohammad-Djafari and Charles Soussen for their concrete and wise advice, their full educational support and their time throughout the past 3 years.

As for corrections related to this manuscript, I am extremely grateful to my referees Nathalie Peyrard and Charles Bouveyron for their careful proofreading and their relevant remarks.

At last, I would like to lovely thank my beloved family and friends for their emotional support that has helped me to achieve this work.

Contents

List of Abbreviations

In this thesis, several abbreviations are used and they are summarized here in order to facilitate reading and examination.

The abbreviations related to operational context are :

- EW : Electronic Warfare
- ESM : Electronic Support Measures
- ELINT : Electronic Intelligence
- SNR : Signal-to-Noise Ratio

The abbreviations related to state-of-the-art algorithms are :

- DA : Discriminant Analysis
- DBSCAN : Density-Based Spatial Clustering of Applications with Noise
- GMM : Gaussian Mixture Model
- KM : k-means algorithm
- KNN or k-nn : k-nearest neighbors algorithm
- LDA : Linear Discriminant Analysis
- NN : Neural Networks
- RdF : Random Forests
- SC : Spectral Clustering
- SVM : Support Vector Machines

The abbreviations related to radar signal pattern are :

- PDW : Pulse Description Word
- TOA : Time-Of-Arrival
- RF : Radio Frequency
- A : Amplitude
- PRI : Pulse Repetition Interval

- PW : Pulse Width

The abbreviations related to Bayesian theory are :

- MAP : Maximum A Posteriori
- VB : Variational Bayes
- VBA : Variational Bayes Approximation
- VBE : Variational Bayes Expectation
- VBM : Variational Bayes Maximization

List of Figures

List of Tables

List of Algorithms

Introduction

In Electronic Warfare (EW) [Sch86], radar signal identification is a crucial component of Electronic Support Measures (ESM) systems [Rog85]. ESM functions enable surveillance of enemy forces such as movements of enemy planes and warning of imminent attack such as launches of rockets. By providing information about the presence of threats, classification of radar signal has a self protection role ensuring that countermeasures against enemies are well-chosen by ESM systems [Wil82]. Furthermore, Electronic Intelligence (ELINT) functions focus on the interception and the analysis of unknown radar signals to update and improve EW databases. Then clustering of radar signals can play a significant role by detecting unknown signal waveforms and supporting ESM functions. Through its classification and clustering aspects, identification of radar signals is a supreme asset for decision making in military tactical situations. Depending on the information available in databases, the identification process can be distinguished into Source Emission Identification, also known as Radar Emitter Classification (REC), which concerns the classification of types of emission sources and Specific Emitter Identification (SEI) which focuses on recognition of copies of electromagnetic emission sources which are of the same type [Dud16]. REC practically relies on statistical analysis of radar signal patterns from distinct emitters [SL02, HZWT09, PJR13, YWY+13, LJLC16, ZWCZ16, Che17, Sun18] whereas SEI aims to extract distinctive features in the process of signal processing to identify even a single copy of an emission source [GBB+03, KO04, DK13, CLH14, Shi14, DK15b, Dud16]. Generally, the main difficulties for SEI result from the lack of a precise and detailed description of a source emission model in databases [DK15a]. Indeed, information required for SEI can be exhaustive and are not always available in databases provided by operational entities. Therefore, this work only focuses on REC to meet operational constraints.

A radar signal [Ric05] is conceived as a pulse-to-pulse modulation pattern in order to perform a specific role such as surveillance, missile guidance or short range tracking. The ability of a radar to perform such a role relies on its capacity to measure range and velocity of its targets. As a pulse-to-pulse modulation pattern, a radar signal pattern is decomposed into a relevant arrangement of sequences of pulses where each pulse is defined by continuous features and each sequence is characterized by categorical features. The continuous features of a pulse mainly refer to its time of interception, its radio frequency, its duration and its amplitude whereas the categorical features of a sequence refer to modulations of the continuous features. Then, a radar signal pattern is chosen as the combination of continuous and categorical features that minimizes ambiguities related to range measurements and velocity measurements. Depending on its expected function and the military context, radar signal patterns can be either simple or extremely complex. As an example, Multi-Function Radars (MFR) emitters [But98, BWW15], widely used in surveillance and tracking, are able to adapt their emitted patterns to a specific tactical situation they are operating in. Indeed, their emitted waveforms are designed to fit with characteristics of intercepted targets. Therefore in presence of multiple targets, their patterns can become extremely complex and provide to ESM systems a real challenge in terms of

identification. As for the military context, it deeply has an effect on the choice of patterns for radar emitters. Indeed, the continuous and categorical features related to radar emitter patterns mostly remain unchanged in a peace context and radar emitters can be mainly identified through their continuous features such as their pulse frequencies and pulse durations which refer to their spectral signature. On the contrary in a war context, radar emitters likely change their spectral signature to avoid being identified by enemy ESM systems which listed their continuous features during the previous peace context. Nonetheless, some categorical and continuous features remain identical since they characterize the way radar emitters operate in the electromagnetic environment. As an example, temporal evolution of radar emitter amplitudes completely characterizes scanning behaviours of radar emitters regardless of the military context.

Most of the time, ESM systems receive mixtures of signals from different radar emitters in the electromagnetic environment. Before identifying radar emitters, ESM systems have to isolate each radar signal from the received mixtures of radar signals. To this end, deinterleaving methods [Mar89, MP92, MK94] are deployed as source separation algorithms to transform the homogeneous signal into a set of heterogeneous signals. Nonetheless, deinterleaving techniques cannot always manage to group all the pulses that belong to a radar emitter which results in a partial observation of its pattern. Furthermore, EW sensor deficiency and low Signal-to-Noise Ratio (SNR) values in sensors can also cause measurement errors [KP16] that disable detection of modulations related to radar signal patterns. When measurements are known to be erroneous, considering them as missing measurements can also be a more reliable approach than using them or discarding them. These material constraints introduce outliers in continuous radar data and missing components in both continuous and categorical radar data. At last, military databases are filled by human beings and may also be imperfect by gathering outliers and missing data due to human errors.

In statistical words, a radar signal pattern is described by continuous and categorical data which can be partially missing and erroneous. Depending on the complexity of radar signal patterns, the classification and clustering procedure should take into consideration any type of data and model a dependence structure to handle outliers and missing data. Classification and clustering problems are closely connected with pattern recognition [Bis06] where many general algorithms [HW79, EKS+96, Bre01] have been developed and used in various fields [SEKX98, Jai10]. However, most algorithms cannot handle missing data and imputation methods [TCS+01] are required to generate data to use them. Hence, the main objective of this work is to define a classification and clustering framework that handles both outliers and missing values. Here, an approach based on mixture models is preferred since mixture models provide a mathematically based, flexible and meaningful framework for the wide variety of classification and clustering requirements [BCG00]. More precisely, a scale mixture of Normal distributions [AM74] is updated to handle outliers and missing data issues for any types of data. Exact inference in that Bayesian approach is unfortunately intractable, therefore a Variational Bayesian (VB) inference [WMR+96] is used to find approximate posterior distributions of parameters and to provide a lower bound on the model log evidence used as a criterion for selecting the number of clusters.

Outline of the thesis is as follows. In Chapter 1, classification and clustering methods from state of the art are first presented according to their degree of supervision. After detailing supervised and unsupervised learning methods dedicated to classification and clustering, the selected approach is introduced through its theoretical aspects by defining mixture models as a flexible probabilistic framework that can handle both classification and clustering applications. However, estimation of parameters can turn out to be a cumbersome task and an approximation method

is proposed to overcome this issue. Once theoretical aspects of mixture models have been presented, mixture models for continuous data are studied in Chapter 2 where generalizations of standard Gaussian mixture models are developed to handle outliers and missing data issues. The resulting model is performed on realistic simulated data obtained through an experimental protocol reproducing faults from real acquisition systems. Chapter 3 is dedicated to the extension of the model for mixed data composed of continuous and categorical features. After presenting categorical features of radar emitters by defining different types of modulations, a dependence structure between mixed data is investigated in order to develop a mixture model that handles both continuous and categorical data even in presence of outliers and missing values. The resulting model is performed on simulated data issued from a real-world database gathering various radar emitter patterns. Then Chapter 4 focuses on integration of temporal evolution radar data into the mixture model framework to take into consideration temporal evolution of radar emitter amplitudes that significantly reveal radar emitter scanning behaviours. To this end, the temporal evolution of radar emitter amplitudes is assumed to be either parabolic or piecewise parabolic and resulting models are evaluated on real operational cases to exhibit their performance. At last, an overall conclusion and work perspectives are given to conclude this thesis.

Chapter 1

State of the art and selected approach

Classification is used mostly as a supervised learning method to achieve a predictive goal [VP98] by extrinsically adding unlabelled groups of data to reference classes. As for clustering, it is used for unsupervised learning to achieve a descriptive goal by discovering new groups of interest in data *via* an intrinsic assessment [RM05]. The main goal of this work is to develop a framework that handles both classification and clustering for mixed-type data gathering outliers and missing values. Then in this chapter, general state-of-the-art algorithms are first detailed in Section 1.1 before introducing the selected approach based on mixture models in Section 1.2.

Contents

1.1 State of the art

According to classification and clustering methods, three families of methods can be distinguished. Partitioning methods focus on relocating observations by moving them from one cluster to another conditionally to an initial partitioning. These partitioned-based methods generally require an a priori number of clusters set by the user. Then, density-based methods assume that each cluster is distributed according to a specific probability distribution [BR93] involving that the resulting marginal distribution of the data follows a mixture of distributions related to clusters. These density-based methods are designed for discovering clusters of arbitrary shape and for identifying their distribution parameters. At last, model-based methods aim to optimize the fit between data and chosen mathematical models by finding characteristic descriptions for each class or cluster [RM05]. Classification and clustering methods from these three families are successively presented in subsections 1.1.1 and 1.1.2.

1.1.1 Supervised learning

General state-of-the-art classification algorithms are introduced in this subsection.

Discriminant Analysis

Discriminant Analysis (DA) is a generalization of Fisher's linear discriminant [Fis36], a method used to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. DA works when the measurements made on independent variables for each observation are continuous quantities. DA is used on labelled data to learn model parameters and then DA can perform classification. Moreover, the analysis is quite sensitive to outliers and the size of the smallest group must be larger than the number of predictor variables. In the case where there are more than two classes, the analysis used in the derivation of the Fisher discriminant can be extended to find a subspace which appears to contain all of the class variability.

DA approaches the problem by assuming that the conditional probability density functions $p(\boldsymbol{x}|y=0)$ and $p(\boldsymbol{x}|y=1)$ are both normally distributed with mean and covariance parameters $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Under this assumption, a classification rule is built by computing the log of likelihoods ratio and testing if it is higher than some threshold ϵ. The obtained classification rule is given by

$$(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0) + \ln\boldsymbol{\Sigma}_0 - (\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - \ln\boldsymbol{\Sigma}_1 > \epsilon\ . \tag{1.1}$$

The resulting classifier is referred to Quadratic Discriminant Analysis (QDA) since the classification rule (1.1) is quadratic according to data $\boldsymbol{x}$. The classification rule (1.1) can be linearly relaxed by assuming homoscedasticity ($\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$) and becomes

$$\boldsymbol{a}\cdot\boldsymbol{x} + \boldsymbol{b} > 0 \tag{1.2}$$

with

$$\begin{aligned}\boldsymbol{a} &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}\ ,\\ \boldsymbol{b} &= \frac{1}{2}\left(\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_1 - \epsilon\right)\ .\end{aligned}$$

In that case, the resulting classifier is referred to Linear Discriminant Analysis (LDA). Performance of LDA and QDA classifiers are illustrated on Figure 1.1.

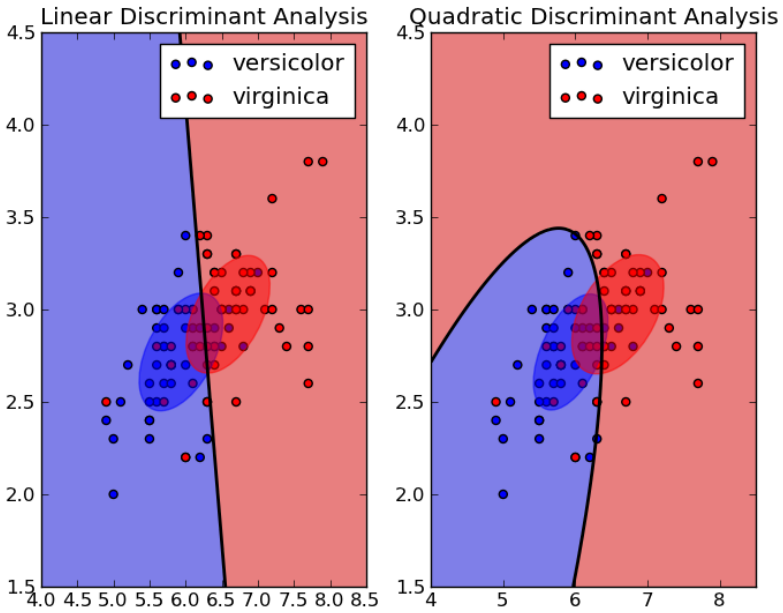


Figure 1.1: Linear Discriminant Analysis and Quadratic Discriminant Analysis performed on Iris dataset [Lic13]

Logistic regression

Logistic regression [Cox58] is a powerful statistical way of modeling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. The predictive variables can be from any type ranges from continuous to categorical. Logistic regression can be extended to multinomial logistic regression [J.88] which is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.

Mathematically, the outcome variable y is modeled conditionally to the explanatory variable $\boldsymbol{x}$ as follows

$$y = \begin{cases} 1 \text{ if } \boldsymbol{\beta} \cdot \boldsymbol{x} + \epsilon > 0 \\ 0 \text{ otherwise } . \end{cases} \tag{1.3}$$

where ϵ follows a standard logistic distribution. The classification rule (1.3) can be interpreted in probabilistic way such that

$$y \sim \begin{cases} p(y = 0|\boldsymbol{x}) = \dfrac{1}{1 + e^{-\boldsymbol{\beta} \cdot \boldsymbol{x}}} \\ p(y = 1|\boldsymbol{x}) = \dfrac{e^{-\boldsymbol{\beta} \cdot \boldsymbol{x}}}{1 + e^{-\boldsymbol{\beta} \cdot \boldsymbol{x}}} \end{cases} \tag{1.4}$$

k-nearest neighbors

k-nearest neighbors (k-NN) algorithm belongs to the family of instance-based learning algorithms [AKA91] which are non-parametric general algorithms that classify a new unlabelled observation

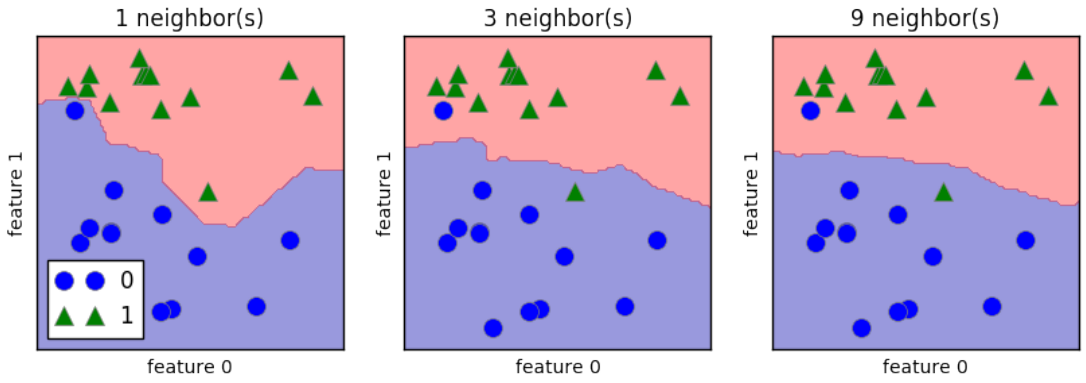


Figure 1.2: Impact of the choice of the number of neighbors on the classification rule

according to similar labelled observations in the training set. The k-NN algorithm [Alt92] particularly assigns a new observation to the class of training observations gathering its k-nearest neighbors. The k-nearest neighbors of the unlabelled observation result from a simple search procedure based on selection of the k nearest training samples measured from a given distance metric. The distance metric is generally designed to meet the data structure in order to create the most relevant neighbors according to the type of data. The k-NN algorithm can induce complex classifiers from a relatively small training set and can be really effective for large training datasets [Rok10]. Nevertheless, it can be sensitive to outliers and can not handle multimodal classes. Furthermore, it requires a value for k and can have a high computation cost since for a new observation its distance to all training samples has to be computed. In the case of mixed-type data, finding an appropriate and meaningful distance can be complex. The k-NN algorithm is detailed in procedure 1.1 and Figure 1.2 shows the impact of the choice of k on the classification rule. Indeed the choice of a larger k involves a smoother classification rule leading to a simpler modeling of data.

Procedure 1.1 k-NN algorithm

Input: Training data $(\boldsymbol{x_n})_{n=1}^{N}$ with labels $\boldsymbol{z} = (z_1, \ldots, z_N)$, number of neighbors k, a distance measure d and unlabelled observation $\boldsymbol{x}^*$
Output: Label z^* of the observation $\boldsymbol{x}^*$
  for $n = 1$ **to** N **do**
    Compute distance $d(\boldsymbol{x}_n, \boldsymbol{x}^*)$
  end for
  Compute set I containing labels z_n for the k smallest distances $d(\boldsymbol{x}_n, \boldsymbol{x}^*)$
  Compute $z^* = \text{mode}(I)$ to find the majority label in I
  return label z^*

Decision trees and random forests

Decision trees [Bre17] are decision support tools that use tree-like graphs or models of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules. Decision trees are simple to understand and interpret since they can be generated from experts' rules based on mixed data. However, they are not robust to outliers since a small

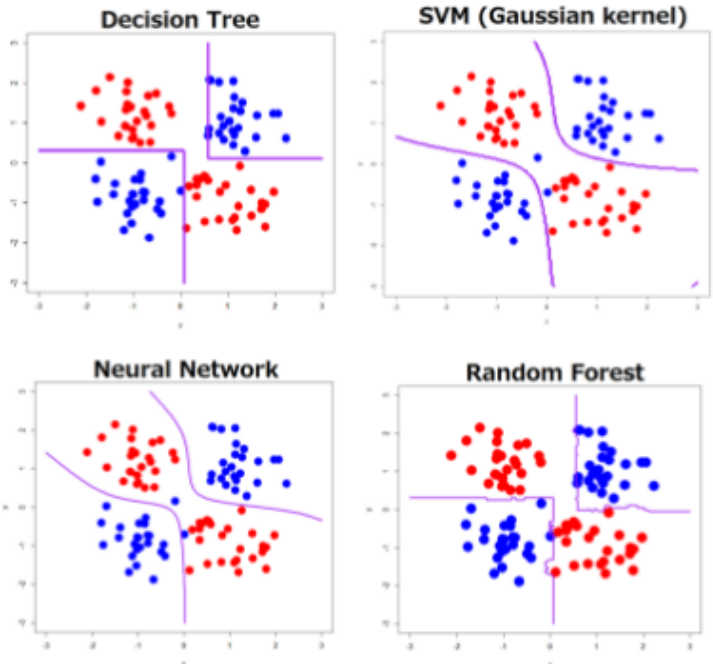


Figure 1.3: A classification rule differently learned by four standard algorithms

change in the data can lead to a large change in the structure of the optimal decision tree. This can be remedied by replacing a single decision tree with a random forest of decision trees, but a random forest is not as easy to interpret as a single decision tree.

Random Forests [Bre01] are an ensemble learning method [Rok10] for classification and regression that operates by constructing a multitude of decision trees at training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests use boosting and bagging techniques [Sch90, Fre95, FS97] in order to correct for decision trees' habit of overfitting to their training set and having high variances [FHT+00]. Indeed as an aggregation of multiple decision trees randomly trained on different feature sets of the training dataset [Ho98], random forests reduce the correlation between trees by avoiding over-focusing on features that appear highly significant in the training set but reveal less relevant in the test set. The ability of random forests to learn smoother decision rules than classical decision trees is visible on Figure 1.3. As predictive tools, random forests can not provide a description of features' relationships in datasets leading to infer on missing data.

Support vector machines

Support vector machines (SVM) [CV95] are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of binary examples, the SVM method maps that set into a transformed feature space [BGV92] where data are linearly separable and generates an hyperplane to separate those points into two distinct groups. SVM are scalable algorithms and can perform classification on large and sparse datasets. However, SVM methods focus on maximizing the margin between the two groups, they can not provide information about clusters' structure used to infer on missing data. An example of a decision rule learned by a SVM algorithm is presented on Figure 1.3.

Neural networks

Neural Networks (NN) [MP43] were popularized by [Ros58] with networks called perceptrons making predictions based on a linear predictor function combining a set of weights with feature vectors. Perceptrons' limitations had been stated by [MP69] before [RHW85] introduced internal representation that enables a non linear mapping of data and ensures a better representation of the problem by adding hidden units in neural networks architecture. Thanks to the back-propagation procedure which distributes pattern recognition errors throughout the network, [RHW88] generalized the learning rule for multilayer networks [HSW89] demonstrated that standard multilayer feed-forward networks are capable of approximating any measurable function f to any desired degree of accuracy. Indeed, feed-forward networks define a mapping $y = g(x;\theta)$ and learn values of parameter θ in order to find the best function approximation. Their architectures are straightforward in a sense that information flows through the function being evaluated from x, through the intermediate computations used to define g, and finally to the output y.

Mathematically, a feed-forward neural network with K layers is a function from a subset $X_0 \in \mathbb{R}^n$ to a subset $X_K \in \mathbb{R}^p$ recursively defined by :

$$X_k = g_k(W_k X_{k-1} + b_k), k \in \{1, \dots, K\} \tag{1.5}$$

where W_k and b_k are the k^{th} layer weights and bias. X_0 and X_K are the input and output layers whereas $(X_k)_{2:(K-1)}$ are hidden layers. The mapping g results in the composition of the $(g_k)_{1:K}$ called activation functions.

Therefore, NN show strong results in classification [LBBH98, KSH12] since they can extract features and learn classification rules (Figure 1.3) for a given architecture. However, NN are not descriptive models and can not provide explainable and relevant information about structure of classes.

1.1.2 Unsupervised learning

General state-of-the-art clustering algorithms are introduced in this subsection.

k-means

k-means algorithm [HW79] partitions observations into K clusters in which each observation belongs to the cluster with the nearest mean defined as the cluster whose mean has the least squared Euclidean distance (Figure 1.4). The k-means algorithm is described in procedure 1.2. The k-means algorithm is easily scalable and can be applied to large datasets without extra computational costs. Nonetheless, a key limitation is its cluster model which is based on isotropic clusters that are separable so that the mean converges towards the cluster center. The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment. In addition, this algorithm is sensitive to noisy data and outliers and is limited to numeric attributes since it uses euclidean distance as metric. At last, it requires a value for the number of clusters k which is not trivial when no prior knowledge is available

The k-prototypes algorithm was presented by [Hua98] to handle categorical data by defining the k-modes algorithm which uses a simple matching dissimilarity measure to deal with categorical objects, replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimise the clustering cost function. Then, the

k-prototypes algorithm results from integrating the k-means and k-modes algorithms to enable clustering of mixed-type data.

As for outliers and noise handling, [KR87] proposed the k-medoids method which differs from the k-means mainly in its representation of the different clusters. Each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the cluster. Hence, the k-medoids method is more robust than the k-means algorithm in the presence of noise and outliers since a medoid is less influenced by outliers or other extreme values than a mean.

At last, the kernel k-means method was introduced as an extension of the k-means method by mapping the input data points non-linearly into a higher-dimensional feature space *via* a kernel function [DGK04]. The kernel k-means method enables discovering clusters with no arbitrary shape by relaxing the assumption on isotropic clusters.

Both presented updates of the k-means algorithm are more complex in nature and have a larger time complexity than the standard k-means algorithm. Moreover both methods still require the user to specify the a priori number of clusters K.

Procedure 1.2 K-means algorithm

Input: Unlabeled dataset $\boldsymbol{x} \in \mathbb{R}^{d*N}$ and number of clusters K
Output: Partition of $\boldsymbol{z}$ and cluster centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^d$
Initialise cluster centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ randomly
repeat
Assign each data point to its closest cluster centroid :

$$\forall n \in \{1, \ldots, N\}, z_n = \arg\min_k ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||$$

Update each cluster center by computing the mean of all points assigned to it :

$$\forall k \in \{1, \ldots, K\}, \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \mathbb{I}_{z_n=k} \boldsymbol{x}_n}{\sum_{n=1}^{N} \mathbb{I}_{z_n=k}}$$

until convergence
return labels $\boldsymbol{z}$ and cluster centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$

Density-based spatial clustering of applications with noise

Density-based spatial clustering of applications with noise (DBSCAN) [EKS+96] is a density-based clustering algorithm which groups together points with many nearby neighbors and marks as outliers points that lie alone in low-density regions. The mean idea behind DBSCAN is to continue growing a cluster as long as the density in the neighborhood exceeds a given threshold ϵ under the constraint that the neighborhood has to contain at least a minimum number of data points $minPts$. An abstract algorithm for the DBSCAN algorithm is proposed in procedure 1.3. Advantages of the DBSCAN algorithm lie in its ability to discover clusters of arbitrary shapes even for large spatial databases. Indeed, DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature since it can better learn the underlying structure of data (Figure 1.5). In addition to its robustness to outliers, DBSCAN does not

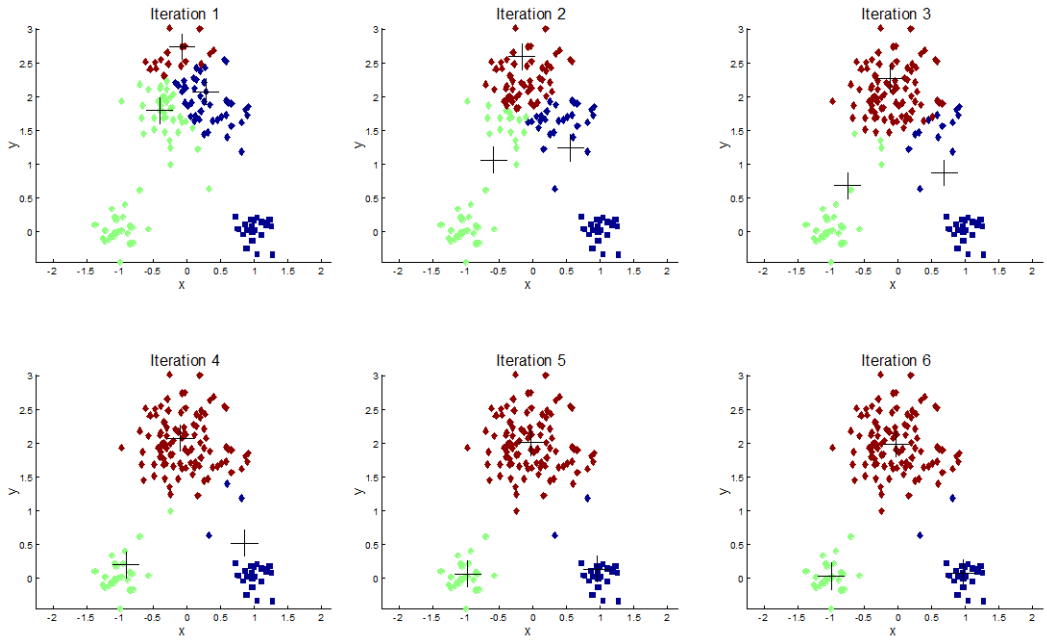


Figure 1.4: Clusters centers are moving as and when iterations of the k-means algorithm are progressing.

require one to specify the number of clusters in the data a priori but DBSCAN cannot cluster data sets well with large differences in densities.

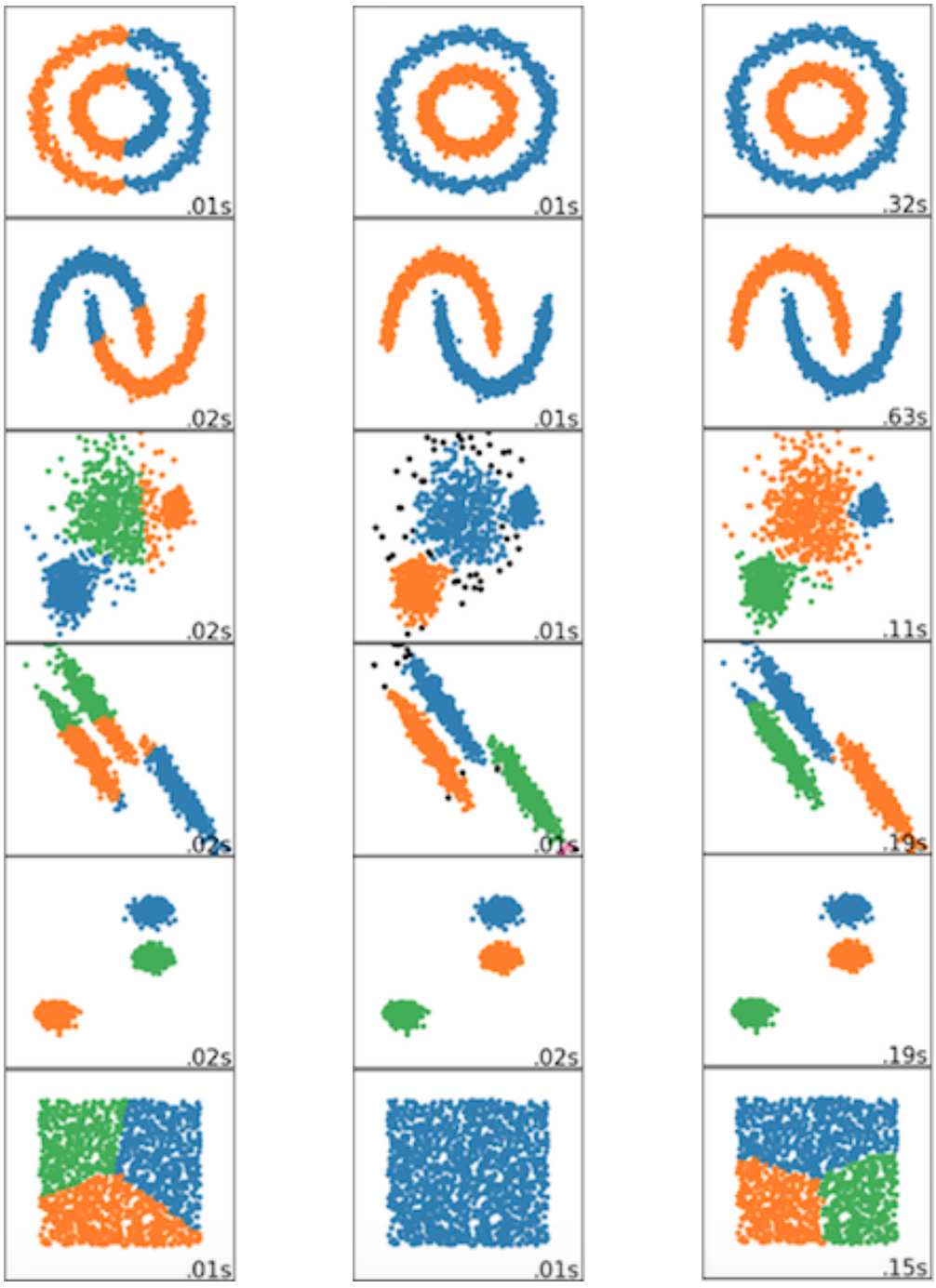
Procedure 1.3 Abstract Algorithm for DBSCAN

Input: Unlabeled dataset, threshold ϵ and minimum number of data points $minPts$
Output: Estimated partition of the dataset
Find the points in the ϵ neighborhood of every point, and identify the core points with more than $minPts$ neighbors
Find the connected components of core points on the neighbor graph, ignoring all non-core points
Assign each non-core point to a nearby cluster if the cluster is an ϵ neighbor, otherwise assign it to noise
return the estimated partition of the dataset

Spectral clustering

The main idea behind spectral clustering techniques [VL07] lies in the use of the spectrum of a given similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions (Figure 1.5). The similarity matrix is provided as a symmetric matrix whose each element represents a measure of the similarity between data points. The general approach to spectral clustering is to use a standard clustering method on relevant eigenvectors of a Laplacian matrix of the similarity matrix. For computational efficiency, these eigenvectors are often computed as the eigenvectors corresponding to the largest several eigenvalues of a function of the Laplacian. When the relevant eigenvectors are processed through a k-means algorithm, the spectral clustering can be reformulated as a weighted kernel k-means problem [DGK04] where the kernel function is assimilated to the dimensionality reduction step leading to creation of the relevant eigenvectors. Moreover, spectral clustering can also be related to DBSCAN clustering

[HMDH18] since optimal spectral clusters can correspond to density-connected components obtained by an asymmetric neighbor graph with edges removed when source points are not dense. Limitations of the spectral clustering lie in its computational cost and the choice of the similarity when data do not have a trivial structure.



(a) K-means clustering (b) DBSCAN clustering (c) Spectral clustering

Figure 1.5: Clustering performance of k-means algorithm in Figure (a), DBSCAN in Figure (b) and spectral clustering in Figure (c) for various datasets by using Scikit-learn library [PVG+11].

1.2 Selected approach : mixture models

Here, an approach based on mixture models is preferred since mixture models provide a mathematical-based, flexible and meaningful framework for the wide variety of classification and clustering requirements [BCG00]. Indeed mixture models, as generative models, enable the creation of a latent space where each latent variable can model a constraint of the problem of interest. Moreover, mixture models incorporate every degree of supervision since they handle both unsupervised or supervised classification problems. Finally, number of classes can be selected using criteria built on the model likelihood.

This section can be considered as a general theoretical framework used as a building block in the following chapters.

1.2.1 Definition

Mixture modeling [JJ94] is a natural framework for classification and clustering. It can be formalized as :

$$p(\boldsymbol{x}_j|\boldsymbol{\Theta}, \mathcal{K}) = \sum_{k \in \mathcal{K}} a_k \psi_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k) \ , \tag{1.6}$$

where $\boldsymbol{x}_j \in \mathcal{X}$ is an observation variable on the observation space $\mathcal{X}$, $\mathcal{K} = \{1, \ldots, K\}$ is the set of clusters and $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, with $\boldsymbol{a} = [a_1, \ldots, a_K]'$, stands for parameters. Each probability distribution ψ_k stands for the k^{th} component distribution with a weight a_k where $a_k \geq 0$ and $\sum_k a_k = 1$.

Assuming a dataset $\boldsymbol{x} \in \mathcal{X}^J$ of i.i.d observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$, the log likelihood function is given by

$$\log p(\boldsymbol{x}|\boldsymbol{\Theta}, \mathcal{K}) = \sum_{j \in \mathcal{J}} \log \sum_{k \in \mathcal{K}} a_k \psi_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k) \ , \tag{1.7}$$

where $\mathcal{J} = \{1, \ldots, J\}$.

According to the degree of supervision, three problems can be distinguished : supervised classification, semi-supervised classification and unsupervised classification known as clustering. Supervised classification consists in parameters estimation of K known classes through a set of training data. Semi-supervised requires estimation of parameters of K unknown clusters whereas clustering proceeds to estimation of both parameters and number of clusters K.

Hence, parameters estimation is required to proceed to these three techniques. Unfortunately, classical Maximum Likelihood estimation turns out to be a complex problem since maximizing the log likelihood function (1.7) requires to deal with the summation over k that appears inside the logarithm and leads to a non closed form solution [Bis06].

One way of solving that estimation problem is to consider it as an incomplete data problem where only $\boldsymbol{x}$ is observed and where the complete data are composed of $\boldsymbol{x}$ and latent variables $\boldsymbol{h}$ such as labels of observations. The likelihood function for the complete dataset simply takes the form $\log p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\Theta})$ and maximization of this complete-data log likelihood function should be straightforward. However since only the incomplete data $\boldsymbol{x}$ are given in practice, the complete data likelihood cannot be used and its expected value under the posterior distribution

of the latent variable $p(\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{\Theta})$ is considered. An elegant and powerful method for solving that issue is called the expectation-maximization algorithm (EM) [DLR77] and consists in performing an expectation (E) step, which creates a function for the expectation of the complete log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

The EM algorithm iterative scheme can be formalized as follows :

- E step : Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $\boldsymbol{h}$ given $\boldsymbol{x}$ under the current estimate of the parameters $\boldsymbol{\Theta}_t$

$$\mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}_t) = \mathbb{E}_{\boldsymbol{h}|\boldsymbol{x},\boldsymbol{\Theta}_t}\left[\log p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\Theta})\right] \ ,$$

- M step: Find the parameters that maximize this quantity

$$\boldsymbol{\Theta}_{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}_t) \ .$$

Nonetheless for some models, it can be infeasible to evaluate the posterior distribution or to compute expectations with respect to this distribution since the dimensionality of the latent space is too high to work with directly or because the posterior distribution has a highly complex form for which expectations are not analytically tractable [Bis06]. Hence approximation schemes are needed and rely on stochastic or deterministic approximations. Stochastic techniques such as Markov chain Monte Carlo have enabled the widespread use of Bayesian methods across many domains. They generally have the property that given infinite computational resource, they can generate exact results, and the approximation arises from the use of a finite amount of processor time. In practice, sampling methods can be computationally demanding, often limiting their use to small-scale problems. Also, it can be difficult to know whether a sampling scheme is generating independent samples from the required distribution. Deterministic approximation schemes, some of which scale well to large applications, are based on analytical approximations to the posterior distribution by assuming that it factorizes in a particular way or that it has a specific parametric form. As such, they can never generate exact results, and so their strengths and weaknesses are complementary to those of sampling methods [Bis06]. In this study, a deterministic approximation method known as Variational Bayes is developed for parameter estimation.

1.2.2 Latent variables

A mixture can be formalized as a latent model since the component label associated to each data point is unobserved. To this end, a categorical variable $z_j \in \mathcal{K}$ can be considered to describe the index of the component distribution generating the observation variable $\boldsymbol{x}_j$. Then, the mixture distribution (1.6) is expressed as

$$p(\boldsymbol{x}_j|\boldsymbol{\Theta}, \mathcal{K}) = \sum_{z_j \in \mathcal{K}} p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) p(z_j|\boldsymbol{\Theta}, \mathcal{K}) \tag{1.8}$$

where

$$p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{k\in\mathcal{K}} \psi_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k)^{\delta^k_{z_j}} ,$$
$$p(z_j|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{k\in\mathcal{K}} a_k^{\delta^k_{z_j}} .$$

and $\delta^k_{z_j}$ denotes the Kronecker symbol which is 1 if $z_j = k$ and 0 otherwise. The latent representation (1.8) can also be viewed in a hierarchical way as

$$\boldsymbol{x}_j|z_j = k \sim \psi_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k) \tag{1.9}$$
$$z_j \sim \text{Categorical}(\boldsymbol{a}) \tag{1.10}$$

Then the joint distribution is

$$p(\boldsymbol{x}_j, z_j|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{k\in\mathcal{K}} [a_k\psi_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k)]^{\delta^k_{z_j}} .$$

Depending on the target problem, other latent variables can be introduced to model the data. $\boldsymbol{h}$ will refer to latent variables in the following parts.

1.2.3 Bayesian framework

Assuming a dataset $\boldsymbol{x} \in \mathcal{X}^J$ of i.i.d observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$ and independent latent data $\boldsymbol{h} = \{\boldsymbol{h}_j\}_{j=1}^J$, likelihood functions can be expressed as

$$p(\boldsymbol{x}|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}} p(\boldsymbol{x}_j|\boldsymbol{\Theta}) , \tag{1.11}$$
$$p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}} p(\boldsymbol{x}_j, \boldsymbol{h}_j|\boldsymbol{\Theta}) , \tag{1.12}$$

where $p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K})$ is called the complete likelihood since it represents the joint distribution of the observed and latent data and $\mathcal{J} = \{1, \ldots, J\}$. That Bayesian framework imposes to specify a prior distribution for the parameters $\boldsymbol{\Theta}$

$$p(\boldsymbol{\Theta}|\mathcal{K}) = p(\boldsymbol{a}) \prod_{k\in\mathcal{K}} p(\boldsymbol{\theta}_k) .$$

Eventually, the posterior distribution of interest is obtained as

$$p(\boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{x}, \mathcal{K}) = \frac{p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K})p(\boldsymbol{\Theta}|\mathcal{K})}{p(\boldsymbol{x}|\mathcal{K})} \tag{1.13}$$

with the marginal distribution of data given by

$$p(\boldsymbol{x}|\mathcal{K}) = \int p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K})p(\boldsymbol{\Theta}|\mathcal{K})\partial\boldsymbol{h}\partial\boldsymbol{\Theta} .$$

1.2.4 Inference

The Variational Bayesian inference was introduced in [WMR+96] as an ensemble learning method for the mixtures of experts in order to avoid over-fitting and noise level under-estimation problems of traditional maximum likelihood inference. In [Att99], the Variational Bayesian inference was generalized for different types of mixture distributions and took the name Variational Bayes (VB). VB can be viewed as a Bayesian generalization of the Expectation-Maximization (EM) algorithm [DLR77] combined with a Mean Field Approach [OS01]. It consists in approximating the intractable posterior distribution $p(\boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{x}, \mathcal{K})$ by a tractable one $q(\boldsymbol{h}, \boldsymbol{\Theta})$ whose parameters are chosen *via* a variational principle to minimize the Kullback-Leibler (KL) divergence

$$KL\left[q||p\right] = \int q(\boldsymbol{h}, \boldsymbol{\Theta}) \log\left(\frac{q(\boldsymbol{h}, \boldsymbol{\Theta})}{p(\boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{x}, \mathcal{K})}\right) d\boldsymbol{h} d\boldsymbol{\Theta} \ .$$

Noting that $p(\boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{x}, \mathcal{K}) = \frac{p(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})}{p(\boldsymbol{x}|\mathcal{K})}$, the KL divergence can be written as

$$KL\left[q||p\right] = \log p(\boldsymbol{x}|\mathcal{K}) - \mathcal{L}(q|\mathcal{K}) \ .$$

$\mathcal{L}(q|\mathcal{K})$ is considered as a lower bound for the log evidence $\log p(\boldsymbol{x}|\mathcal{K})$ and can be expressed as

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h}, \boldsymbol{\Theta})\right] \ , \tag{1.14}$$

where $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}[\cdot]$ denotes the expectation with respect to $q(\boldsymbol{h}, \boldsymbol{\Theta})$. Then, minimizing the KL divergence is equivalent to maximizing $\mathcal{L}(q|\mathcal{K})$. Assuming that $q(\boldsymbol{h}, \boldsymbol{\Theta})$ can be factorized over the latent variables $\boldsymbol{h}$ and the parameters $\boldsymbol{\Theta}$, a free-form maximization with respect to $q(\boldsymbol{h})$ and $q(\boldsymbol{\Theta})$ leads to the following update rules :

$$\begin{aligned}
&\textbf{VBE-step} : q(\boldsymbol{h}) \propto \exp\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K})\right]\right), \\
&\textbf{VBM-step} : q(\boldsymbol{\Theta}) \propto \exp\left(\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{\Theta}, \boldsymbol{x}, \boldsymbol{h}|\mathcal{K})\right]\right) \ .
\end{aligned}$$

The expectations $\mathbb{E}_{\boldsymbol{h}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{\Theta}}[\cdot]$ are respectively taken with respect to the variational posteriors $q(\boldsymbol{h})$ and $q(\boldsymbol{\Theta})$. Thereafter, the algorithm iteratively updates the variational posteriors by increasing the bound $\mathcal{L}(q|\mathcal{K})$.

1.2.5 Classification and clustering

According to the degree of supervision, three problems can be distinguished : supervised classification, semi-supervised classification and unsupervised classification known as clustering.

The supervised classification problem is decomposed into a training step and a prediction step. The training step consists in estimating parameters $\boldsymbol{\Theta}$ given the number of classes K and a set of training data $\boldsymbol{x}$ with known labels $\boldsymbol{z}$. Then, the prediction step results in associating label z^* of a new sample $\boldsymbol{x}^*$ to its class k^* chosen as the Maximum A Posteriori (MAP) solution

$$k^* = \arg\max_{k \in \mathcal{K}} p(z^* = k|\boldsymbol{x}^*, \boldsymbol{\Theta}, \mathcal{K})$$

given the previous estimated parameters $\boldsymbol{\Theta}$.

In the semi-supervised classification, only the number of classes K is known and both labels $\boldsymbol{z}$ of the dataset $\boldsymbol{x}$ and parameters $\boldsymbol{\Theta}$ have to be determined. As for the prediction step, the MAP criterion is retained for affecting observations to classes such that

$$k^* = \arg\max_{k \in \mathcal{K}} p(z = k|\boldsymbol{x}, \boldsymbol{\Theta}, \mathcal{K}) \ .$$

Given a set of data $\boldsymbol{x}$, the clustering problem aims to determine the number of clusters $\tilde{K}$, labels $\boldsymbol{z}$ of data and parameters $\boldsymbol{\Theta}$. Selecting the appropriate $\tilde{K}$ seems like a model selection issue and is usually based on a maximized likelihood criterion given by

$$\tilde{K} = \arg\max_{K} \log p(\boldsymbol{x}|K) \ . \tag{1.15}$$

where

$$p(\boldsymbol{x}|K) = \int p(\boldsymbol{x}, \boldsymbol{\Theta}|K) d\boldsymbol{\Theta} \tag{1.16}$$

Unfortunately, (1.16) is intractable and many penalized likelihood criteria such as AIC [Aka98], BIC [S+78] and ICL [BCG00] had been proposed. The lower bound (1.14) for (1.16), found in subsection 1.2.4, is preferred to other criteria since it does not depend on asymptotical assumptions and does not require Maximum Likelihood estimates.

Then according to an a priori range of numbers of clusters $\{K_{\min}, \ldots, K_{\max}\}$, the semi-supervised classification is performed for each $K \in \{K_{\min}, \ldots, K_{\max}\}$ and both $\boldsymbol{z}^K$ and $\boldsymbol{\Theta}^K$ are estimated. Finally, the number of classes $\tilde{K}$ in (1.15) is chosen as the maximizer of the lower bound $\mathcal{L}(q|\mathcal{K})$:

$$\tilde{K} = \arg\max_{K} \ \mathcal{L}(q|\mathcal{K}) \ . \tag{1.17}$$

After determining $\tilde{K}$, only $\boldsymbol{z}^{\tilde{K}}$ and $\boldsymbol{\Theta}^{\tilde{K}}$ are kept as estimated labels and parameters.

1.3 Conclusion

In this chapter, state-of-the-art classification and clustering algorithms have been presented. Some of them are dedicated to create boundaries to separate data into heterogeneous clusters such as LDA, SVM or NN whereas others focus on learning underlying structure of data to build them such as k-NN or k-means algorithms. However both types of algorithms do not provide an internal framework that infers on missing data and copes with any degree of supervision. Therefore, an approach based on mixture models is proposed and developed through its theoretical aspects. As hierarchical graphical models, mixture models provide a flexible framework to handle classification and clustering issues by introducing a latent space where each latent variable focuses on a specific constraint. However, the resulting model is not tractable and model learning is processed through an approximation method known as Variational Bayes Approximation. Eventually whatever degree of supervision is required, the number of classes K and parameters can be estimated to perform classification and clustering tasks. Next chapters deal with implementations of such models with different types of data.

Chapter 2

Continuous data

Radar emitter patterns are partly described by continuous features that can be partially observed and approximately measured due to a noisy electromagnetic environment and sensor deficiencies. This chapter focuses on the development of a model that handles outliers and missing values to enable classification and clustering of radar emitters. First, continuous features of a radar emitter pattern are presented in Section 2.1 before introducing an experimental protocol developed to acquire realistic data since real military data are often classified. Then, the proposed model is explained in Section 2.2 where latent variables are introduced to model outliers and missing values. Inference procedure is processed through a Variational Bayesian Approximation in Section 2.3. Finally, evaluation of the model is proposed through two experiments and performance of the method are detailed in Section 2.4.

Contents

2.1 Data

In this section, typical continuous radar features are first presented before introducing an acquisition system designed to generate realistic radar data which naturally embed outliers.

2.1.1 Continuous radar features

Continuous features of a radar emitter (Figure 2.1) are traditionally extracted from its Pulse Description Words (PDW). Each PDW gathers information related to a given pulse in the radar signal pattern such as

- its Time of Arrival (TOA) which is the time in μs at which the pulse is detected,
- its Amplitude (A) which the average measured amplitude of the pulse,
- its Radio Frequency (RF) which is the average measured frequency of the pulse in GHz,
- its Pulse Width (PW) which is the pulse duration in μs,
- its Pulse Repetition Interval (PRI) which is the difference in μs between its TOA and the TOA of the previous pulse in the radar signal pattern.

The TOA of a pulse can be taken as the instant that a threshold is crossed. In presence of low signal-to-noise ratio (SNR), this measurement may be not precise and retaining the TOA of the first 3 dB is preferable [DH82]. Moreover, the TOA is not an invariant feature since TOA sequence depends on the first observed TOA. Therefore, the PRI is retained as an invariant feature since it is the difference between times of arrival of two successive pulses. Then, the RF of a pulse can be either fixed or modulated pulse-to-pulse. The RF is said to be frequency agile if it is randomly modulated pulse-to-pulse within fixed bounds and frequency hopping if it has systematic variations. Finally, the PW is the pulse duration chosen to ensure that a radar emits sufficient energy such that reflected pulses are always detectable by its receiver. The amount of energy that can be delivered to a distant target is the product of two things; the output power of the transmitter, and the duration of the transmission. Therefore, pulse width constrains the maximum detection range of a target. Depending on sensor sensitivity, the PW may not be reliable and considering PW as missing data can be preferable.

2.1.2 Realistic data acquisition

In this subsection, an experimental protocol is introduced to acquire unclassified realistic data from different radar emitters. This protocol consists in an acquisition step followed by a feature extraction step.

The acquisition system is composed of two Software Defined Radio (SDR) platforms based on Ettus USRP E312 (Emitter) and B200 (Receiver) boards, linked to a laptop to record the data. As in [SEG+16], this setup was chosen because it allows quick development and experimentation tasks on radiofrequencies from 70 MHz to 6 GHz, it is quite cheap and is available off-the-shelf. Radar waveforms, emitted by the URSP E312 board, are generated from bin files coded from a database gathering more than 40 typical radar waveforms with agile and hopping frequencies and jittered and staggered PRI. The developed system is presented in Figure 2.2. In order to meet hardware constraints, RF range was mapped to a 4MHz bandwidth and patterns of TOA and PW were slightly modified but their dynamic was preserved.

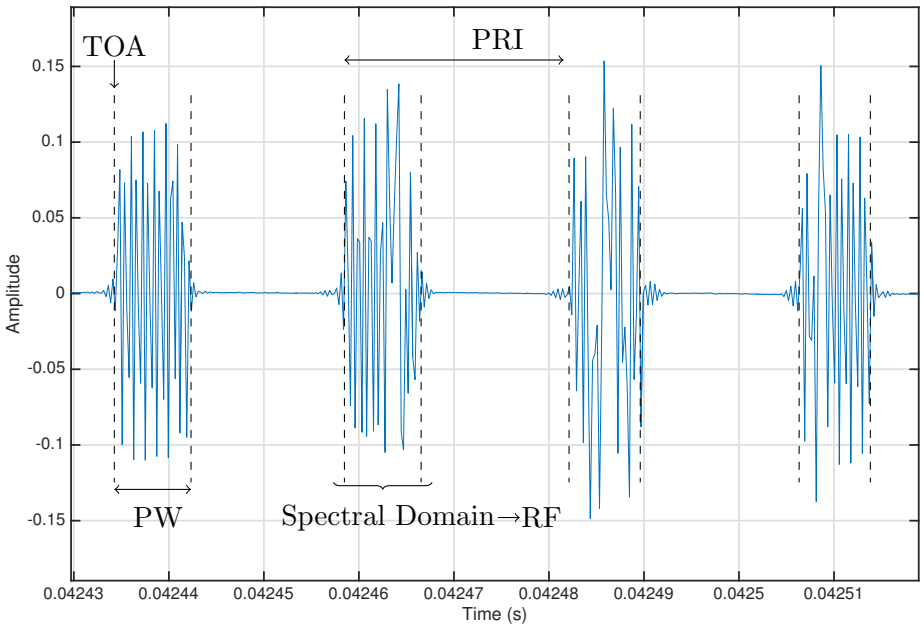


Figure 2.1: Four pulses from a radar emitter whose related amplitudes and times are shown on the vertical axis and the horizontal axis. Then, the continuous features PRI and RF are obtained by delimiting each pulse according to their TOA and PW.

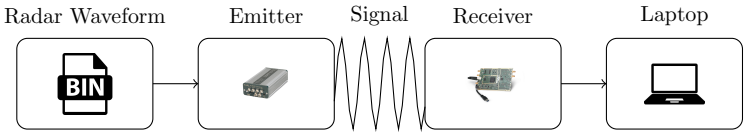


Figure 2.2: Diagram of the acquisition system

Then, a threshold algorithm, provided by [DH82], is used to detect pulses in the recorded signal $s_r(t)$ of duration T. Each pulse at TOA_t is characterised by a triplet $(\mathrm{PRI}_t, \mathrm{RF}_t, \mathrm{PW}_t)_t$ where PRI_t is the difference between TOA_t and TOA_{t-1} and the RF_t feature is estimated with a Fast Fourier Transform (FFT) algorithm. Figure 2.3 shows parameters measurement on real data. For a given recorded signal s gathering n_s pulses, the following PDW matrix is obtained

$$PDW = \begin{pmatrix} RF_1 & PW_1 & PRI_1 \\ \vdots & \vdots & \vdots \\ RF_m & PW_m & PRI_m \\ \vdots & \vdots & \vdots \\ RF_{n_s} & PW_{n_s} & PRI_{n_s} \end{pmatrix} \tag{2.1}$$

where $m \in \{2, \ldots, n_s - 1\}$ is the index of pulses in recording.

SDR platforms are imperfect [FLP+07] and their defects can introduce outliers due to measurement errors. Hardware imperfections are visible on Figure 2.4 where the third pulse is cut into two pulses which leads to the formation of PRI and PW outliers. Furthermore since experiments take place in real outside conditions, other signals and reflections can disturb the acquisition [DH82].

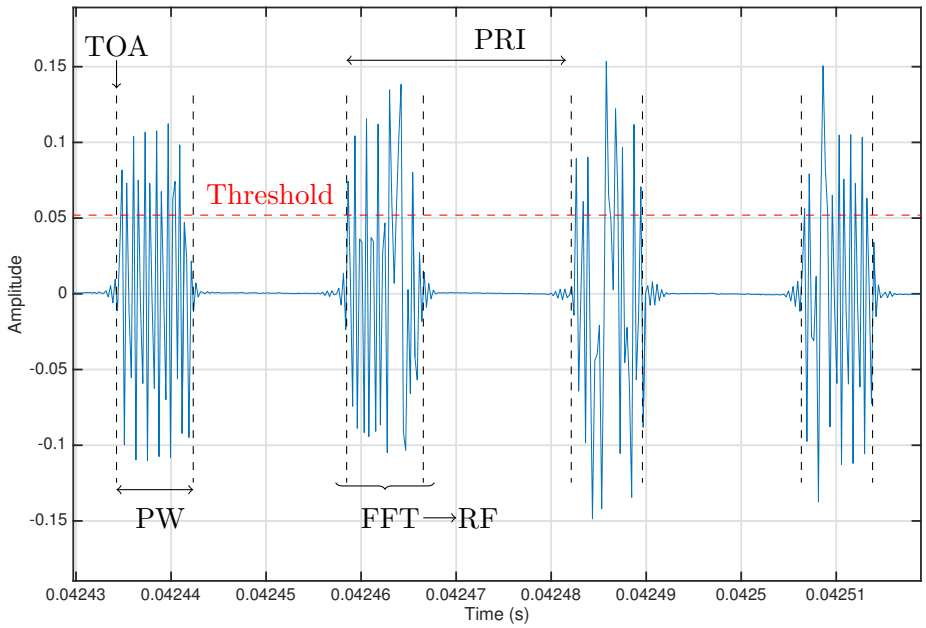


Figure 2.3: Acquired pulses from a radar emitter where the three features (PRI,PW,RF) are shown on the figure.

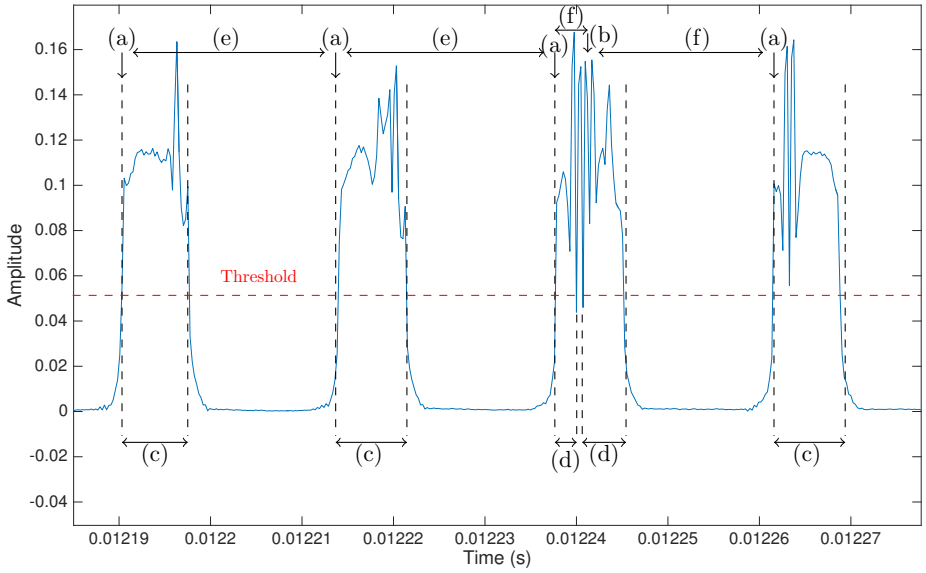


Figure 2.4: Outliers formation during primary parameters measurement on real data. (a), (c) and (e) are respectively exact TOA, PW and PRI. (b), (d) and (f) are outliers for TOA, PW and PRI.

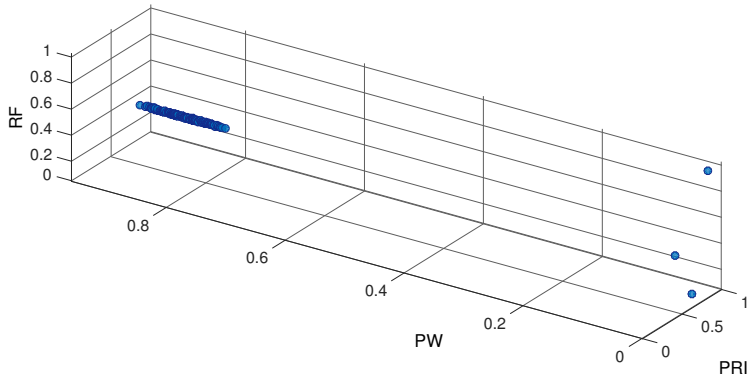


Figure 2.5: Presence of outliers in observations of a radar emitter.

Finally for each signal s_j gathering n_j pulses of the J recorded signals $(s_j)_{j=1}^{J}$, a matrix PDW_j is created from (2.1) and an observation vector $\boldsymbol{x}_j$ is defined according to (2.2) such that

$$\boldsymbol{x}_j = (\bar{\mathrm{RF}}_j, \bar{\mathrm{PW}}_j, \bar{\mathrm{PRI}}_j) \tag{2.2}$$

where $\bar{\mathrm{RF}}_j$ is the average value of RF, $\bar{\mathrm{PW}}_j$ is the average value of PW and $\bar{\mathrm{PRI}}_j$ is the average value of PRI defined in (2.3), (2.4) and (2.5).

$$\bar{\mathrm{RF}}_j = \frac{1}{n_j} \sum_{m=1}^{n_j} RF_m \ , \tag{2.3}$$

$$\bar{\mathrm{PW}}_j = \frac{1}{n_j} \sum_{m=1}^{n_j} PW_m \ , \tag{2.4}$$

$$\bar{\mathrm{PRI}}_j = \frac{1}{n_j} \sum_{m=1}^{n_j} PRI_m \ . \tag{2.5}$$

Once all observation vectors $(\boldsymbol{x}_j)_{j=1}^{J}$ have been constructed, they are normalized to meet constraints of machine learning algorithms. Figure 2.5 shows the distribution of 150 normalized observation vectors of a radar emitter, where three outliers are visible.

2.2 Model

In this section, K emitters defined by continuous data are considered. Therefore, the main objective is to develop a mixture model which can build K distinct clusters even in presence of outliers and missing values. First, state-of-the-art approaches are reviewed. Then, the standard Gaussian mixture model is presented and two varieties of this model are introduced to handle outliers and missing values. At last, the proposed mixture model is developed into a Bayesian framework.

2.2.1 State of the art

Radar emitter classification relies on statistical analysis of Pulse Description Words (PDW) of a radar signal that gather its basic measurable parameters such as Radio Frequency (RF), Amplitude, Pulse Width (PW) or Pulse Repetition Interval (PRI). In terms of classification and clustering of emission sources from different types, many approaches based on data fusion and machine learning have been developed and traditionally proceed to feature extraction, dimensionality reduction and classification or clustering. For example, [SL02, PJR13, LJLC16, Sun18] propose various neural classification approaches based on the PDW structure of observed signals whereas [YWY+13] introduce a hybrid radar emitter recognition method based on rough k-means and relevance vector machine and [Che17] develop an efficient classification method using weighted-xgboost model for complex radar signals in large datasets. As regards the clustering problem, [HZWT09] develop a dynamic clustering algorithm that uses designed distances and dynamic cluster centers and does not require fixing the number of classes which depends on the input data, [ZWCZ16] also introduce a clustering framework composed of local processing and multi-sensor fusion processing and use a Minimum Description Length criterion to update dynamically the number of clusters rather than setup in advance. These practical approaches mostly result from more general algorithms such as Random Forests [Bre01], Neural Networks [Ros58], Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) [EKS+96] and k-means algorithm [HW79] which are also considered as state-of-the-art algorithms since they are used in various fields [SEKX98, Jai10]. However, these practical and general algorithms can not handle missing data and imputation methods [TCS+01] are required to generate data to use them. Hence, an approach based on mixture models is preferred since mixture models provide a mathematically based, flexible and meaningful framework for the wide variety of classification and clustering requirements [BCG00]. More precisely, a scale mixture of Normal distributions [AM74] is updated to handle outliers and missing data issues. On the one hand, this model is robust to outliers by accounting for the uncertainties of variances and covariances since the associated marginal distributions are heavy-tailed [AV07]. On the other hand, dependencies between features can easily be modelled through a multivariate Gaussian distribution in order to infer on missing values by benefiting from attractive Gaussian properties.

2.2.2 Standard Gaussian mixture models

Gaussian mixture models [QR78, JJ94] (GMM) are the most well-known mixture models for continuous data and have been widely used for decades. As a natural framework for classification and clustering, a GMM can be formalized as :

$$\forall j \in \mathcal{J},\ p(\boldsymbol{x}_j|\boldsymbol{\Theta}, \mathcal{K}) = \sum_{k \in \mathcal{K}} a_k \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\ , \tag{2.6}$$

where $\boldsymbol{x}_j \in \mathbb{R}^d$ is an observation vector, $\mathcal{K} = \{1, \ldots, K\}$ is a finite and known set of clusters

and $\boldsymbol{\Theta} = (\boldsymbol{a}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k\in\mathcal{K}})$, with $\boldsymbol{a} = [a_1, \ldots, a_K]'$, stands for parameters. Moreover, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are respectively the mean and the covariance matrix of the k^{th} component distribution with a weight a_k where $a_k \geq 0$ and $\sum_{k\in\mathcal{K}} a_k = 1$.

The GMM can be formalized as a latent model since the component label associated to each data point is unobserved. To this end, a categorical variable $z_j \in \mathcal{K}$ can be considered to describe the index of the component distribution generating the observation variable $\boldsymbol{x}_j$. Then, the mixture distribution (1.6) is expressed as

$$p(\boldsymbol{x}_j|\boldsymbol{\Theta}, \mathcal{K}) = \sum_{z_j\in\mathcal{K}} p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) p(z_j|\boldsymbol{\Theta}, \mathcal{K}) \ , \tag{2.7}$$

where

$$p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\delta^k_{z_j}} \ , \tag{2.8}$$

$$p(z_j|\boldsymbol{\Theta}, \mathcal{K}) = \mathcal{C}at(z_j|\boldsymbol{a}) = \prod_{k\in\mathcal{K}} a_k^{\delta^k_{z_j}} \tag{2.9}$$

and $\delta^k_{z_j}$ denotes the Kronecker symbol which is 1 if $z_j = k$ and 0 otherwise.

Assuming a dataset $\boldsymbol{x} = (\boldsymbol{x}_j)_{j\in\mathcal{J}}$ of i.i.d observations and independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$, the complete likelihood is obtained as follows

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \left[a_k \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{\delta^k_{z_j}} \ . \tag{2.10}$$

At last, the Bayesian framework imposes to specify priors for the parameters $\boldsymbol{\Theta}$. The resulting conjugate priors are

$$\begin{cases} p(\mathbf{a}|\mathcal{K}) = \mathcal{D}(\mathbf{a}|\kappa_0) \\ p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \eta_0^{-1}\boldsymbol{\Sigma}_k) \\ p(\boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{IW}(\boldsymbol{\Sigma}_k|\gamma_0, \boldsymbol{\Sigma}_0) \ . \end{cases} \tag{2.11}$$

where the Dirichlet and the Inverse Wishart distributions are defined as follows :

$$\mathcal{D}(\mathbf{a}|\boldsymbol{\kappa}) = c_{\mathcal{D}}(\boldsymbol{\kappa}) \prod_{k\in\mathcal{K}} a_k^{\kappa_k - 1} \ ,$$

$$\mathcal{IW}(\boldsymbol{\Sigma}|\gamma, \mathbf{S}) = c_{\mathcal{IW}}(\gamma, \mathbf{S}) |\boldsymbol{\Sigma}|^{-\frac{\gamma+d+1}{2}} \exp\left(-\frac{1}{2} tr(\mathbf{S}\boldsymbol{\Sigma}^{-1})\right) \ ,$$

where $c_{\mathcal{D}}(\boldsymbol{\kappa})$ and $c_{\mathcal{IW}}(\gamma, \mathbf{S})$ are normalizing constants such that

$$c_{\mathcal{D}}(\boldsymbol{\kappa}) = \frac{\Gamma\left(\sum_{k\in\mathcal{K}} \kappa_k\right)}{\prod_{k\in\mathcal{K}} \Gamma(\kappa_k)}, \ c_{\mathcal{IW}}(\gamma, \mathbf{S}) = \frac{|\mathbf{S}|^{\frac{\gamma}{2}}}{2^{\frac{d\gamma}{2}} \Gamma_d(\frac{\gamma}{2})} \ .$$

The standard Gaussian model is shown on Figure 2.6.

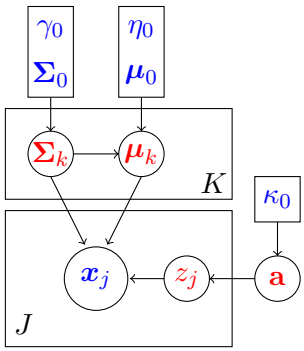


Figure 2.6: Graphical representation of the standard Gaussian mixture model. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate represents the independent identically distributed observations $\boldsymbol{x}_j$ and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red.

2.2.3 Gaussian mixture models with missing data

As weighted sums of Gaussian distributions, GMMs benefit from attractive Gaussian properties that enable modeling dependencies between features to infer on missing data. Indeed, missing values can be handled by decomposing the features vector $\boldsymbol{x}_j \in \mathbb{R}^d$ into observed features $\boldsymbol{x}_j^{\text{obs}} \in \mathbb{R}^{d_j^{\text{obs}}}$ and missing features modeled by a latent variable $\boldsymbol{x}_j^{\text{miss}} \in \mathbb{R}^{d_j^{\text{miss}}}$ such that $1 \leq d_j^{\text{obs}} \leq d$ and $d_j^{\text{miss}} = d - d_{\text{obs}}$. Reminding that conditionally to its index cluster the features vector $\boldsymbol{x}_j$ is Gaussian distributed as

$$\boldsymbol{x}_j = \begin{pmatrix} \boldsymbol{x}_j^{\text{miss}} \\ \boldsymbol{x}_j^{\text{obs}} \end{pmatrix} | z_j = k \sim \mathcal{N}\left(\boldsymbol{\mu}_k = \begin{pmatrix} \boldsymbol{\mu}_k^{\text{miss}} \\ \boldsymbol{\mu}_k^{\text{obs}} \end{pmatrix}, \boldsymbol{\Sigma}_k = \begin{pmatrix} \boldsymbol{\Sigma}_k^{\text{miss}} & \boldsymbol{\Sigma}_k^{\text{cov}} \\ \boldsymbol{\Sigma}_k^{\text{cov}'} & \boldsymbol{\Sigma}_k^{\text{obs}} \end{pmatrix} \right) ,$$

the latent variable $\boldsymbol{x}_j^{\text{miss}}$ can be expressed as a Gaussian distributed variable such that

$$\boldsymbol{x}_j^{\text{miss}} | \boldsymbol{x}_j^{obs}, z_j = k \sim \mathcal{N}\left(\boldsymbol{x}_j^{\text{miss}} | \boldsymbol{\mu}_{jk}^{\boldsymbol{x}^{\text{miss}}}, \boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{miss}}} \right) \tag{2.12}$$

where

$$\boldsymbol{\mu}_{jk}^{\boldsymbol{x}^{\text{miss}}} = \boldsymbol{\mu}_k^{\text{miss}} + \boldsymbol{\Sigma}_k^{\text{cov}} \boldsymbol{\Sigma}_k^{\text{obs}-1} (\boldsymbol{x}_j^{\text{obs}} - \boldsymbol{\mu}_k^{\text{obs}}) , \tag{2.13}$$

$$\boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{miss}}} = \boldsymbol{\Sigma}_k^{\text{miss}} - \boldsymbol{\Sigma}_k^{\text{cov}} \boldsymbol{\Sigma}_k^{\text{obs}-1} \boldsymbol{\Sigma}_k^{\text{cov}'} . \tag{2.14}$$

Then, a marginal distribution for $\boldsymbol{x}_j^{\text{obs}}$ is obtained such that $\boldsymbol{x}_j^{\text{obs}} \sim \mathcal{N}\left(\boldsymbol{x}_j^{\text{obs}} | \boldsymbol{\mu}_k^{\text{obs}}, \boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{obs}}} \right)$ with

$$\boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{obs}}} = \left(\boldsymbol{\Sigma}_k^{\text{obs}-1} + 2 \times \boldsymbol{\Sigma}_k^{\text{obs}-1} \boldsymbol{\Sigma}_k^{\text{cov}'} \left(\boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{miss}}} \right)^{-1} \boldsymbol{\Sigma}_k^{\text{cov}} \boldsymbol{\Sigma}_k^{\text{obs}-1} \right)^{-1} .$$

Eventually, a Gaussian mixture model handling missing data is obtained by integrating missing data distributions (2.12) into the complete likelihood (2.10) such that

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{z} | \boldsymbol{\Theta}, \mathcal{K}) &= \prod_{j \in \mathcal{J}} \prod_{k \in \mathcal{K}} \left[a_k \mathcal{N}(\boldsymbol{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]^{\delta_{z_j}^k} \\ &= \prod_{j \in \mathcal{J}} \prod_{k \in \mathcal{K}} \left[a_k \mathcal{N}\left(\boldsymbol{x}_j^{\text{miss}} | \boldsymbol{\mu}_{jk}^{\boldsymbol{x}^{\text{miss}}}, \boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{miss}}} \right) \mathcal{N}\left(\boldsymbol{x}_j^{\text{obs}} | \boldsymbol{\mu}_k^{\text{obs}}, \boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{obs}}} \right) \right]^{\delta_{z_j}^k} . \end{aligned} \tag{2.15}$$

Parameters are a priori distributed according to (2.11). A graphical representation of the model is exhibited on Figure 2.7.

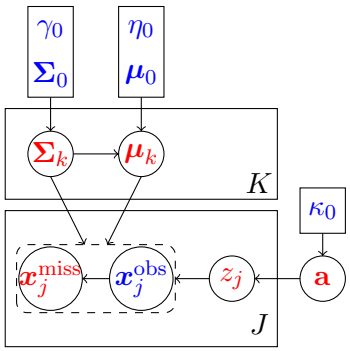


Figure 2.7: Graphical representation of the Gaussian mixture model handling missing data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $\boldsymbol{x}_j$ decomposed into observable data $\boldsymbol{x}_j^{\text{obs}}$ and missing data $\boldsymbol{x}_j^{\text{miss}}$ and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red.

2.2.4 Gaussian mixture models with outliers

A major limitation of GMMs is their lack of robustness to outliers that can lead to over-estimate the number of clusters since they use additional components to capture the tails of the distributions [SB05]. Nonetheless, outlier values in $\boldsymbol{x}_j$ can be handled by introducing a latent variable u_j to scale each mixture component covariance matrix $\boldsymbol{\Sigma}_k$. That family of mixture models is known as scale mixtures of Normal distributions [AM74] and benefits from heavy-tailed marginal distributions accounting for the uncertainties of variances and covariances [AV07]. Introducing the latent positive variable u_j into (2.8), the following scale component distribution is obtained

$$p(\boldsymbol{x}_j|u_j, z_j, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, u_j^{-1}\boldsymbol{\Sigma}_k)^{\delta_{z_j}^k} , \tag{2.16}$$

and the joint distribution of $(\boldsymbol{x}_j, u_j)$ is derived from (2.16) such that

$$p(\boldsymbol{x}_j, u_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{k\in\mathcal{K}} \left[\mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, u_j^{-1}\boldsymbol{\Sigma}_k)p_k(u_j)\right]^{\delta_{z_j}^k} \tag{2.17}$$

where $p_k(u_j)$ is the prior distribution of u_j conditionally to $z_j = k$.

Conditionally to the choice of a prior distribution for u_j, the marginal distribution $p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) = \int_0^\infty p(\boldsymbol{x}_j, u_j|z_j, \boldsymbol{\Theta}, \mathcal{K})\partial u_j$ of $\boldsymbol{x}_j$ over u_j can take different forms [WS00]. [SB05, AV07, SZKL17, NW14] mainly propose using a Gamma distribution parametrized by a deterministic parameter ν_k such that the joint distribution of $(\boldsymbol{x}_j, u_j)$ from (2.17) becomes

$$p(\boldsymbol{x}_j, u_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{k\in\mathcal{K}} \left[\mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, u_j^{-1}\boldsymbol{\Sigma}_k)\mathcal{G}\left(u_j|\frac{\nu_k}{2}, \frac{\nu_k}{2}\right)\right]^{\delta_{z_j}^k} \tag{2.18}$$

Then the resulting marginal distribution $p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K})$ follows a Student-t distribution obtained

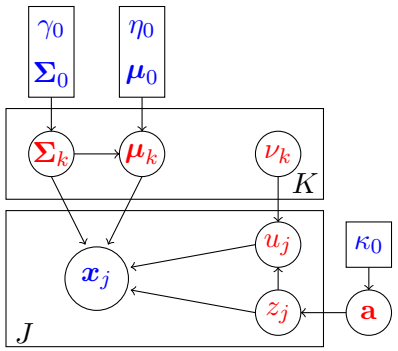


Figure 2.8: Graphical representation of the Gaussian mixture model handling outliers. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $\boldsymbol{x}_j$, the scale variables u_j and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red.

by

$$\begin{aligned} p(\boldsymbol{x}_j|z_j,\boldsymbol{\Theta},\mathcal{K}) &= \int_0^{+\infty} p(\boldsymbol{x}_j,u_j|z_j,\boldsymbol{\Theta},\mathcal{K})\partial u_j \\ &= \int_0^{+\infty} \mathcal{N}\left(\boldsymbol{x}_j|\boldsymbol{\mu}_k,u_j^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{G}\left(u_j|\frac{\nu_k}{2},\frac{\nu_k}{2}\right)\partial u_j \\ &= \frac{\Gamma(\frac{d+\nu_k}{2})}{\Gamma(\frac{\nu_k}{2})(\nu_k\pi)^{\frac{d}{2}}} \times |\boldsymbol{\Sigma}_k^{-1}|^{\frac{1}{2}} \times \left[1+\frac{1}{\nu_k}(\boldsymbol{x}_j-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_k)\right]^{-\frac{d+\nu_k}{2}} \\ &= \mathcal{T}(\boldsymbol{x}_j|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\nu_k) \end{aligned} \tag{2.19}$$

where d is the dimension of the feature space and ν_k is the degree of freedom of the Student-t distribution. Eventually, a Gaussian mixture model handling outliers is obtained by integrating (2.18) into the complete likelihood (2.10) such that

$$p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{\Theta},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\left[a_k\mathcal{N}\left(\boldsymbol{x}_j|\boldsymbol{\mu}_k,u_j^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{G}\left(u_j|\frac{\nu_k}{2},\frac{\nu_k}{2}\right)\right]^{\delta_{z_j}^k} .$$

where $\boldsymbol{u}=(u_j)_{j\in\mathcal{J}}$ are the scale latent variables related to continuous data $\boldsymbol{x}=(\boldsymbol{x}_j)_{j\in\mathcal{J}}$. As for prior distributions, ν_k does not require a prior distribution since it is deterministic and other parameters are a priori distributed according to (2.11). Then, a graphical representation of the model is shown on Figure 2.8.

The degree of freedom variable ν_k has been considered as a deterministic variable updated *via* an optimization argument during the maximization step of the VB inference [PM00] and [SB05, AV07, SZKL17, NW14] did not assume any prior distribution for ν_k since there do not exist any known conjugate priors for ν. For the sake of keeping conjugacy between prior and posterior distributions and adopting a full Bayesian treatment, a Gamma distribution $\mathcal{G}(u_j|\alpha_k,\beta_k)$ with shape and rate parameters (α_k,β_k) is chosen for $p_k(u_j)=p(u_j|z=k)=\mathcal{G}(u_j|\alpha_k,\beta_k)$ such that the joint distribution of $(\boldsymbol{x}_j,u_j)$ from (2.17) becomes

$$p(\boldsymbol{x}_j,u_j|z_j,\boldsymbol{\Theta},\mathcal{K}) = \prod_{k\in\mathcal{K}}\left[\mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k,u_j^{-1}\boldsymbol{\Sigma}_k)\mathcal{G}\left(u_j|\alpha_k,\beta_k\right)\right]^{\delta_{z_j}^k} \tag{2.20}$$

As in (2.19), the resulting marginal distribution $p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K})$ is also a Student-t distribution [DLGMD17] which is obtained as follows

$$\begin{aligned} p(\boldsymbol{x}_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) &= \int_0^{+\infty} p(\boldsymbol{x}_j, u_j|z_j, \boldsymbol{\Theta}, \mathcal{K}) \partial u_j \\ &= \int_0^{+\infty} \mathcal{N}\left(\boldsymbol{x}_j|\boldsymbol{\mu}_k, u_j^{-1}\boldsymbol{\Sigma}_k\right) \mathcal{G}\left(u_j|\alpha_k, \beta_k\right) du \\ &= \frac{\Gamma(\alpha_k + \frac{d}{2})}{\Gamma(\alpha_k)(2\beta_k\pi)^{\frac{d}{2}}} \times |\boldsymbol{\Sigma}_k^{-1}|^{\frac{1}{2}} \times \left[1 + \frac{1}{2\beta_k}(\boldsymbol{x}_j - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_k)\right]^{-(\alpha_k + \frac{d}{2})} \\ &= \mathcal{T}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k, \beta_k) \end{aligned}$$

Then, a Gaussian mixture model handling outliers is obtained by integrating (2.20) into the complete likelihood (2.10) such that

$$p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \left[a_k \mathcal{N}\left(\boldsymbol{x}_j|\boldsymbol{\mu}_k, u_j^{-1}\boldsymbol{\Sigma}_k\right) \mathcal{G}\left(u_j|\alpha_k, \beta_k\right)\right]^{\delta_{z_j}^k} . \tag{2.21}$$

In a full Bayesian treatment, both α_k and β_k require a prior distribution. Hence, a new conjugate prior distribution is introduced to avoid a non closed-form posterior distribution for (α_k, β_k). This prior distribution is defined below :

$$p(\alpha_k, \beta_k|p_0, q_0, s_0, r_0) \propto \frac{p_0^{\alpha_k - 1} e^{-q_0\beta_k} \beta_k^{s_0\alpha_k}}{\Gamma(\alpha_k)^{r_0}} \mathbb{I}_{\{\alpha_k > 0\}} \mathbb{I}_{\{\beta_k > 0\}} \tag{2.22}$$

where $p_0, q_0, s_0, r_0 > 0$. The previous expression can be reformulated as :

$$p(\alpha_k, \beta_k|p_0, q_0, s_0, r_0) = p(\beta_k|\alpha_k, s_0, q_0) p(\alpha_k|p_0, q_0, s_0, r_0)$$

with

$$\begin{aligned} p(\beta_k|\alpha_k, s_0, q_0) &= \mathcal{G}(\beta_k|s_0\alpha_k + 1, q_0) , \\ p(\alpha_k|p_0, q_0, s_0, r_0) &= \frac{1}{M_0} \frac{p_0^{\alpha_k - 1}\Gamma(s_0\alpha_k + 1)}{q_0^{s_0\alpha_k + 1}\Gamma(\alpha_k)^{r_0}} \mathbb{I}_{\{\alpha_k > 0\}} \end{aligned}$$

where

$$M_0 = \int \frac{p_0^{\alpha_k - 1}\Gamma(s_0\alpha_k + 1)}{q_0^{s_0\alpha_k + 1}\Gamma(\alpha_k)^{r_0}} \mathbb{I}_{\{\alpha_k > 0\}} \partial\alpha_k .$$

The normalization constant M_0 is intractable and a Laplace approximation method is derived to estimate it. As for other parameters, they are a priori distributed according to (2.11). At last, the resulting mixture model is shown on Figure 2.9.

2.2.5 Proposed mixture model

Varieties of the standard GMM have been introduced in (2.15) and (2.21) to handle outliers or missing values. The proposed model results from combining these two varieties to enable the handling of both outliers and missing values. Assuming a dataset $\boldsymbol{x} = (\boldsymbol{x}_j)_{j\in\mathcal{J}}$ of i.i.d observations decomposed into observed features $\boldsymbol{x}^{\text{obs}} = (\boldsymbol{x}_j^{\text{obs}})_{j\in\mathcal{J}}$ and missing features $\boldsymbol{x}^{\text{miss}} = (\boldsymbol{x}_j^{\text{miss}})_{j\in\mathcal{J}}$ and independent latent variables $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$ and $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$, the proposed model

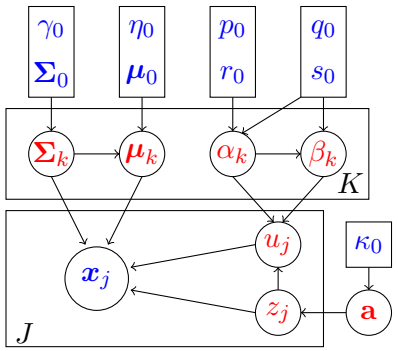


Figure 2.9: Graphical representation of the Gaussian mixture model handling outliers with hyper-parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $\boldsymbol{x}_j$, the scale variables u_j and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red.

results in

$$\begin{aligned} p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K}) &= \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \left[a_k \mathcal{N}\left(\begin{pmatrix} \boldsymbol{x}_j^{\text{miss}} \\ \boldsymbol{x}_j^{\text{obs}} \end{pmatrix} |\boldsymbol{\mu}_k, u_j^{-1}\boldsymbol{\Sigma}_k \right) \mathcal{G}(u_j|\alpha_k, \beta_k) \right]^{\delta_{z_j}^k} \\ &= \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \left[a_k \mathcal{N}\left(\boldsymbol{x}_j^{\text{miss}} | \boldsymbol{\mu}_{jk}^{\boldsymbol{x}^{\text{miss}}}, u_j^{-1}\boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{miss}}} \right) \mathcal{N}\left(\boldsymbol{x}_j^{\text{obs}} | \boldsymbol{\mu}_k^{\text{obs}}, u_j^{-1}\boldsymbol{\Sigma}_k^{\boldsymbol{x}^{\text{obs}}} \right) \mathcal{G}(u_j|\alpha_k, \beta_k) \right]^{\delta_{z_j}^k} \end{aligned} \tag{2.23}$$

where $\boldsymbol{h} = (\boldsymbol{x}^{\text{miss}}, \boldsymbol{u}, \boldsymbol{z})$ is the set of latent variables and $\boldsymbol{\Theta} = (\boldsymbol{a} = (a_k)_{k\in\mathcal{K}}, \boldsymbol{\mu} = (\boldsymbol{\mu}_k)_{k\in\mathcal{K}}, \boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_k)_{k\in\mathcal{K}}, \boldsymbol{\alpha} = (\alpha_k)_{k\in\mathcal{K}}, \boldsymbol{\beta} = (\beta_k)_{k\in\mathcal{K}})$ is the set of parameters. Finally, the required prior distribution for $\boldsymbol{\Theta}$ is decomposed as follows

$$p(\boldsymbol{\Theta}|\mathcal{K}) = p(\boldsymbol{a})p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K})p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathcal{K})p(\boldsymbol{\Sigma}|\mathcal{K})$$

where prior distributions for $\boldsymbol{a}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given in (2.11) and the prior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a product of distributions defined (2.22) such that

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \prod_{k\in\mathcal{K}} p(\alpha_k, \beta_k|p_0, q_0, s_0, r_0) \ . \tag{2.24}$$

A graphical representation of the proposed model is shown on Figure 2.10.

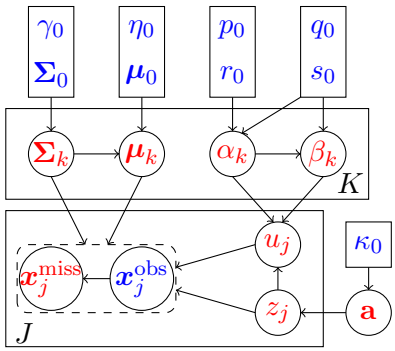


Figure 2.10: Graphical representation of the proposed model. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $\boldsymbol{x}_j$, the scale variables u_j and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red.

2.3 Inference

Direct inference on the proposed model is not trivial since the posterior distribution of latent missing data and parameters is intractable. Therefore, the Variational Bayes (VB) procedure is processed to estimate parameters of the mixture model defined in (2.23). Variational posterior distributions are obtained from the VB Expectation (VBE) and VB Maximization (VBM) steps. These variational posterior distributions are similarly obtained from classical posterior related in [SB05, AV07] and [MP04]. In addition to standard results, missing values are incorporated as latent variables in posterior calculations and a posterior distribution for missing data is proposed. At last, a lower bound on the log evidence is defined to master the convergence of the VB procedure.

2.3.1 Variational posterior distributions

Recalling that the VB procedure consists in approximating the intractable posterior distribution $p(\boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{x}^{\text{obs}}, \mathcal{K})$ by a tractable factorized distribution $q(\boldsymbol{h}, \boldsymbol{\Theta}) = q(\boldsymbol{h})q(\boldsymbol{\Theta})$ that maximizes

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h}, \boldsymbol{\Theta})\right] \ , \tag{2.25}$$

where $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}[\cdot]$ denotes the expectation with respect to $q(\boldsymbol{h}, \boldsymbol{\Theta})$, variational posterior distributions are obtained by performing a free-form maximization through the following update rules :

$$\begin{aligned}
&\textbf{VBE-step} : q(\boldsymbol{h}) \propto \exp\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K})\right]\right) , \\
&\textbf{VBM-step} : q(\boldsymbol{\Theta}) \propto \exp\left(\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})\right]\right) \ .
\end{aligned}$$

According to a conditional factorization of $q(\boldsymbol{h})$ and $q(\boldsymbol{\Theta})$ given by

$$\begin{aligned}
q(\boldsymbol{h}) &= q(\boldsymbol{x}^{\text{miss}}|\boldsymbol{u}, \boldsymbol{z})q(\boldsymbol{u}|\boldsymbol{z})q(\boldsymbol{z}) \ , \\
q(\boldsymbol{\Theta}) &= q(\boldsymbol{a})q(\boldsymbol{\mu}, \boldsymbol{\Sigma})q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \ ,
\end{aligned}$$

the following conjugate variational posterior distributions are obtained from the VB procedure

$$\begin{cases} q(\boldsymbol{x}^{\text{miss}}|\boldsymbol{u},\boldsymbol{z}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{x}_j^{\text{miss}}|\tilde{\boldsymbol{\mu}}_{jk}^{\boldsymbol{x}^{\text{miss}}}, u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}}\right)^{\delta_{z_j}^k} , \\ q(\boldsymbol{u}|\boldsymbol{z}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{G}(\tilde{\alpha}_{jk},\tilde{\beta}_{jk})^{\delta_{z_j}^k} , \\ q(\boldsymbol{z}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j) , \\ q(\boldsymbol{a}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{k}}) , \\ q(\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{k\in\mathcal{K}} p(\alpha_k,\beta_k|\tilde{p}_k,\tilde{q}_k,\tilde{s}_k,\tilde{r}_k) , \\ q(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{k\in\mathcal{K}} \mathcal{N}(\boldsymbol{\mu}_k|\tilde{\boldsymbol{\mu}}_k,\tilde{\eta}_k^{-1}\boldsymbol{\Sigma}_k)\mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k,\tilde{\boldsymbol{\Sigma}}_k) . \end{cases} \tag{2.26}$$

where the variational posterior distributions of $(\boldsymbol{\alpha},\boldsymbol{\beta})$ are defined in (2.22). Their respective parameters are estimated during the VBE and VBM steps by developing expectations $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h}|\boldsymbol{\Theta},\mathcal{K})\right]$ and $\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right]$.

2.3.2 VBE-step

The VBE-step consists in deriving the following expectation

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h}|\boldsymbol{\Theta},\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \Bigg(& \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - \frac{1}{2}\Bigg(d(\log 2\pi - \log u_j) + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] \\ & + u_j\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{x}_j-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_k)\right]\Bigg) + \mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_k\right] \\ & + (\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right]-1)\log u_j - \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_k)\right] - u_j\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_k\right]\Bigg) \end{aligned} \tag{2.27}$$

where $\forall(j,k)\in\mathcal{J}\times\mathcal{K}$:

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{x}_j-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_k)\right] = (\boldsymbol{x}_j-\tilde{\boldsymbol{\mu}}_k)^T\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{x}_j-\tilde{\boldsymbol{\mu}}_k) + \frac{d}{\tilde{\eta}_k} \tag{2.28}$$

is obtained from properties of the variational distribution $q(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})$ in (2.26). Hence continuous data $\boldsymbol{x}$ are distributed a posteriori according to a product of normal distributions conditionally to latent variables $\boldsymbol{u}$ and labels $\boldsymbol{z}$ such that

$$\boldsymbol{x}|\boldsymbol{u},\boldsymbol{z} \sim \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{x}_j|\tilde{\boldsymbol{\mu}}_k, u_j^{-1}\tilde{\gamma}_k^{-1}\tilde{\boldsymbol{\Sigma}}_k\right)^{\delta_{z_j}^k}$$

Mean parameters $(\tilde{\boldsymbol{\mu}}_k)_{k\in\mathcal{K}}$ and variance parameters $(\tilde{\gamma}_k^{-1}\tilde{\boldsymbol{\Sigma}}_k)_{k\in\mathcal{K}}$ of these normal distributions are obtained from (2.28). By decomposing $\boldsymbol{x}$ into $(\boldsymbol{x}^{\text{miss}},\boldsymbol{x}^{\text{obs}})$ and by exploiting properties of the multivariate normal distribution, the following variational posterior distribution is obtained for missing values $\boldsymbol{x}^{\text{miss}}$:

$$q(\boldsymbol{x}^{\text{miss}}|\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{x}_j^{\text{miss}}|\tilde{\boldsymbol{\mu}}_{jk}^{\boldsymbol{x}^{\text{miss}}}, u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}}\right)^{\delta_{z_j}^k}$$

with $\forall(j,k) \in \mathcal{J} \times \mathcal{K}$:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{jk}^{\boldsymbol{x}^{\text{miss}}} &= \tilde{\boldsymbol{\mu}}_k^{\text{miss}} + \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}} \tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}} (\boldsymbol{x}_j^{\text{obs}} - \tilde{\boldsymbol{\mu}}_k^{\text{obs}}) \,, \\
\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}} &= \frac{\tilde{\boldsymbol{\Sigma}}_k^{\text{miss}} - \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}} \tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}} \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}'}}{\tilde{\gamma}_k} \,.
\end{aligned}$$

Then by marginalising over $\boldsymbol{x}^{\text{miss}}$ in (2.27), the expectation (2.27) becomes

$$\begin{aligned}
\int \mathbb{E}_{\boldsymbol{\Theta}} \left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z} | \boldsymbol{\Theta}, \mathcal{K})\right] \partial \boldsymbol{x}^{\text{miss}} = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \delta_{z_j}^k \Bigg(& \mathbb{E}_{\boldsymbol{\Theta}} \left[\log a_k\right] + \mathbb{E}_{\boldsymbol{\Theta}} \left[\log |\boldsymbol{\Sigma}_k|\right] \\
& - \frac{1}{2} \Bigg(d_j^{\text{obs}} (\log 2\pi - \log u_j) - \log |\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}}| \\
& + u_j \left(\left(\boldsymbol{x}_j^{\text{obs}} - \tilde{\boldsymbol{\mu}}_k^{\text{obs}} \right)^T \tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_j^{\text{obs}} - 1} \left(\boldsymbol{x}_j^{\text{obs}} - \tilde{\boldsymbol{\mu}}_k^{\text{obs}} \right) + \frac{d}{\tilde{\eta}_k} \right) \Bigg) \\
& + \mathbb{E}_{\boldsymbol{\Theta}} \left[\alpha_k\right] \mathbb{E}_{\boldsymbol{\Theta}} \left[\log \beta_k\right] + \left(\mathbb{E}_{\boldsymbol{\Theta}} \left[\alpha_k\right] - 1\right) \log u_j \\
& - \mathbb{E}_{\boldsymbol{\Theta}} \left[\log \Gamma(\alpha_k)\right] - u_j \mathbb{E}_{\boldsymbol{\Theta}} \left[\beta_k\right] \Bigg) \Bigg)
\end{aligned} \tag{2.29}$$

with $\forall k \in \mathcal{K}$,

$$\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{obs}}} = \frac{\left(\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}} + 2 \times \tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}} \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}'} \left(\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_j^{\text{miss}}} \right)^{-1} \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}} \tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}} \right)^{-1}}{\tilde{\gamma}_k} \,.$$

Conditionally to $\boldsymbol{z}$, the scale latent variables $\boldsymbol{u}$ are distributed according to a product of Gamma distribution whose parameters are obtained by aggregating terms related to $\boldsymbol{u}$ such that

$$q(\boldsymbol{u} | \boldsymbol{z}, \mathcal{K}) = \prod_{j \in \mathcal{J}} \prod_{k \in \mathcal{K}} \mathcal{G}\left(u_j | \tilde{\alpha}_{jk}, \tilde{\beta}_{jk}\right)^{\delta_{z_j}^k}$$

with $\forall(j,k) \in \mathcal{J} \times \mathcal{K}$:

$$\begin{aligned}
\tilde{\alpha}_{jk} &= \mathbb{E}_{\boldsymbol{\Theta}} \left[\alpha_k\right] + \frac{d_j^{\text{obs}}}{2} \,, \\
\tilde{\beta}_{jk} &= \mathbb{E}_{\boldsymbol{\Theta}} \left[\beta_k\right] + \frac{1}{2} \left(\left(\boldsymbol{x}_j^{\text{obs}} - \tilde{\boldsymbol{\mu}}_k^{\text{obs}} \right)^T \tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_j^{\text{obs}} - 1} \left(\boldsymbol{x}^{\text{obs}} - \tilde{\boldsymbol{\mu}}_k^{\text{obs}} \right) + \frac{d}{\tilde{\eta}_k} \right) \,.
\end{aligned}$$

Finally, variational posterior categorical distributions are obtained for labels $\boldsymbol{z}$ by marginalising over $\boldsymbol{u}$ in (2.29) such that

$$\begin{aligned}
\int \mathbb{E}_{\boldsymbol{\Theta}} \left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z} | \boldsymbol{\Theta}, \mathcal{K})\right] \partial \boldsymbol{x}^{\text{miss}} \partial \boldsymbol{u} = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \delta_{z_j}^k \Bigg(& \mathbb{E}_{\boldsymbol{\Theta}} \left[\log a_k\right] + \log \Gamma(\tilde{\alpha}_{jk}) - \tilde{\alpha}_{jk} \log \tilde{\beta}_{jk} \\
& + \mathbb{E}_{\boldsymbol{\Theta}} \left[\alpha_k\right] \mathbb{E}_{\boldsymbol{\Theta}} \left[\log \beta_k\right] - \mathbb{E}_{\boldsymbol{\Theta}} \left[\log \Gamma(\alpha_k)\right] \\
& - \frac{1}{2} \left(d_j^{\text{obs}} \log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}} \left[\log |\boldsymbol{\Sigma}_k|\right] - \log |\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}}| \right) \Bigg) \\
= & \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \delta_{z_j}^k \log \rho_{jk}
\end{aligned} \tag{2.30}$$

where $\forall j \in \mathcal{J},\ k \in \mathcal{K}$,

$$\begin{aligned}\log \rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - \frac{1}{2}\left(d_j^{\text{obs}} \log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log |\boldsymbol{\Sigma}_k|\right] - \log |\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}}|\right) \\ + \mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right] \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \beta_k\right] - \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \Gamma(\alpha_k)\right] + \log \Gamma(\tilde{\alpha}_{jk}) - \tilde{\alpha}_{jk} \log \tilde{\beta}_{jk}\ .\end{aligned} \tag{2.31}$$

Hence the variational categorical distributions are deduced from (2.30) and are given by

$$q(\boldsymbol{z}|\mathcal{K}) = \prod_{j \in \mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j)$$

where probabilities $(\tilde{\boldsymbol{r}}_j)_{j \in \mathcal{J}}$ are obtained from (2.31) such that $\forall j \in \mathcal{J},\ k \in \mathcal{K}$,

$$\tilde{r}_{jk} = \frac{\rho_{jk}}{\sum\limits_{k \in \mathcal{K}} \rho_{jk}}\ .$$

2.3.3 VBM-step

The VBM-step consists in deriving the following expectation

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K})\right] + p(\boldsymbol{\Theta}|\mathcal{K}) \\ &= \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \Bigg(\log a_k + -\frac{1}{2}\Bigg(\log |\boldsymbol{\Sigma}_k| \\ &+ d(\log 2\pi - \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right]) + \mathbb{E}_{\boldsymbol{h}}\left[u_j (\boldsymbol{x}_j - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_k)\right] \Bigg) \\ &+ \alpha_k \log \beta_k + (\alpha_k - 1)\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] - \log \Gamma(\alpha_k) - \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \beta_k \Bigg) \\ &+ \sum_{k \in \mathcal{K}} (\kappa_{0_k} - 1) \log a_k + \log c_{\mathcal{D}}(\boldsymbol{\kappa}_0) - \frac{1}{2}\Bigg((\gamma_0 + d + 1) \log |\boldsymbol{\Sigma}_k| \\ &+ \text{Trace}\left(\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_k^{-1}\right) \Bigg) + c_{\mathcal{IW}}(\gamma_0, \boldsymbol{\Sigma}_0) + \frac{1}{2}\Bigg(d(\log \eta_{0_k} - \log 2\pi) \\ &- \log |\boldsymbol{\Sigma}_k| - \eta_{0_k} \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{0_k}\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{0_k}\right) \Bigg) - \log M_0 \\ &+ (\alpha_k - 1) \log p_0 - r_0 \log \Gamma(\alpha_k) + s_0 \alpha_k \log \beta_k - q_0 \beta_k\ ,\end{aligned} \tag{2.32}$$

where $\forall (j,k) \in \mathcal{J} \times \mathcal{K}$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h}}\left[u_j (\boldsymbol{x}_j - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_k)\right] &= \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \left(\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right] - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right] - \boldsymbol{\mu}_k\right) \\ &+ \text{Trace}\left(\mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right] \boldsymbol{\Sigma}_k^{-1}\right)\end{aligned} \tag{2.33}$$

is obtained from properties of the variational distribution $q(\boldsymbol{h}|\mathcal{K})$ with

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h}}\left[u_j\right] &= \frac{\tilde{\alpha}_{jk}}{\tilde{\beta}_{jk}}\ , \\ \mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right] &= \begin{pmatrix} \tilde{\boldsymbol{\mu}}_k^{\boldsymbol{x}_j^{\text{miss}}} \\ \boldsymbol{x}_j^{\text{obs}} \end{pmatrix}\ , \\ \mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right] &= \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}} & \mathbf{0}^{d_j^{\text{miss}} \times d_j^{\text{obs}}} \\ \mathbf{0}^{d_j^{\text{obs}} \times d_j^{\text{miss}}} & \mathbf{0}^{d_j^{\text{obs}} \times d_j^{\text{obs}}} \end{pmatrix}\ .\end{aligned}$$

By factorizing terms related to $\boldsymbol{a}$ in (2.32), the following Dirichlet distribution is obtained

$$q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}})$$

where

$$\forall k \in \mathcal{K},\ \tilde{\kappa}_k = \kappa_{0_k} + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \ .$$

Then, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are a posteriori distributed according to the distribution defined in (2.22) such that

$$q(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \prod_{k \in \mathcal{K}} p(\alpha_k, \beta_k|\tilde{p}_k, \tilde{q}_k, \tilde{s}_k, \tilde{r}_k) \ ,$$

where

$$\begin{aligned}
\tilde{p}_k &= p_0 \exp\left(\sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right]\right) \ , \\
\tilde{q}_k &= q_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \ , \\
\tilde{s}_k &= s_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \ , \\
\tilde{r}_k &= r_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \ .
\end{aligned}$$

By aggregating and factorizing terms related to each $\boldsymbol{\mu}_k$ in (2.32), a Normal distribution is obtained for each $\boldsymbol{\mu}_k$ such that

$$q(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathcal{K}) = \prod_{k \in \mathcal{K}} \mathcal{N}\left(\boldsymbol{\mu}_k|\tilde{\boldsymbol{\mu}}_k, \tilde{\eta}_k^{-1}\boldsymbol{\Sigma}_k\right)$$

where $\forall k \in \mathcal{K}$,

$$\begin{aligned}
\tilde{\eta}_k &= \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] + \eta_{0_k} \ , \\
\tilde{\boldsymbol{\mu}}_k &= \frac{\sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right] + \eta_{0_k}\boldsymbol{\mu}_{0_k}}{\tilde{\eta}_k} \ .
\end{aligned}$$

Eventually, variance parameters $\boldsymbol{\Sigma}$ are a posteriori distributed according to Inverse Wishart distributions given by

$$q(\boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k \in \mathcal{K}} \mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k, \tilde{\boldsymbol{\Sigma}}_k)$$

where

$$\begin{aligned}
\tilde{\gamma}_k &= \gamma_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \ , \\
\tilde{\boldsymbol{\Sigma}}_k &= \boldsymbol{\Sigma}_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \left(\mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right] \mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right]^T + \mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right]\right) + \eta_{0_k}\boldsymbol{\mu}_{0_k}\boldsymbol{\mu}_{0_k}^T - \tilde{\eta}_k\tilde{\boldsymbol{\mu}}_k\tilde{\boldsymbol{\mu}}_k^T \ .
\end{aligned}$$

2.3.4 Lower bound

The lower bound $\mathcal{L}(q|\mathcal{K})$ on the log evidence (2.25) is decomposed into the free energy $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right]$ and the entropy of the approximate posterior $q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})$ given by $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right]$. The free energy can be developed as

$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h}|\boldsymbol{\Theta},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right]$$

where

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h}|\boldsymbol{\Theta},\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\Bigg(&\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - \frac{1}{2}\Bigg(d(\log 2\pi - \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right])\\ &+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[u_j(\boldsymbol{x}_j-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_k)\right]\Bigg)\\ &+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_k\right] + (\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right]-1)\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right]\\ &- \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_k)\right] - \mathbb{E}_{\boldsymbol{h}}\left[u_j\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_k\right]\Bigg)\end{aligned}$$

with

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[u_j(\boldsymbol{x}_j-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_k)\right] = \mathbb{E}_{\boldsymbol{h}}\left[u_j\right]\Bigg(&(\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right]-\tilde{\boldsymbol{\mu}}_k)^T\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}(\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right]-\tilde{\boldsymbol{\mu}}_k)\\ &\frac{d}{\tilde{\eta}_k}\Bigg) + \text{Trace}\left(\mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_j\right]\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}\right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right] = &\sum_{k\in\mathcal{K}}(\kappa_{0_k}-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log c_{\mathcal{D}}(\boldsymbol{\kappa}_0) - \frac{1}{2}\Bigg((\gamma_0+d+1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right]\\ &+ \text{Trace}\left(\boldsymbol{\Sigma}_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Sigma}_k^{-1}\right]\right)\Bigg) + c_{\mathcal{IW}}(\gamma_0,\boldsymbol{\Sigma}_0) + \frac{1}{2}\Bigg(d(\log\eta_{0_k}-\log 2\pi)\\ &- \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] - \eta_{0_k}\mathbb{E}_{\boldsymbol{\Theta}}\left[\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_{0_k}\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_{0_k}\right)\right]\Bigg) - \log M_0\\ &+ (\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right]-1)\log p_0 - r_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_k)\right] + s_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_k\right]\\ &- q_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_k\right]\end{aligned}$$

with

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_{0_k}\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_{0_k}\right)\right] = \left(\tilde{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_{0_k}\right)^T\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}\left(\tilde{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_{0_k}\right) + \frac{d}{\tilde{\eta}_k}\,.$$

As for the entropy term, the following decompositon is obtained

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{u},\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\\ &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u},\boldsymbol{z},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{u}|\boldsymbol{z},\mathcal{K})\right]\\ &+ \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u},\boldsymbol{z},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\frac{1}{2}\Big(d_j^{\text{miss}}(\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] \\
&\quad - \log 2\pi - 1) - \log|\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_j}|\Big)\,, \\
\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{u}|\boldsymbol{z},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\Big(\tilde{\alpha}_{jk}\log\tilde{\beta}_{jk} - \log\Gamma(\tilde{\alpha}_{jk}) \\
&\quad + (\tilde{\alpha}_{jk}-1)\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] - \tilde{\beta}_{jk}\mathbb{E}_{\boldsymbol{h}}\left[u_j\right]\Big)\,, \\
\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\tilde{r}_{jk}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right] &= \sum_{k\in\mathcal{K}}(\tilde{\kappa}_k - 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log c_{\mathcal{D}}(\tilde{\boldsymbol{\kappa}}_k) - \frac{1}{2}\Big((\tilde{\gamma}_k + d + 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] \\
&\quad + \text{Trace}\left(\tilde{\boldsymbol{\Sigma}}_k\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Sigma}_k^{-1}\right]\right)\Big) + c_{\mathcal{IW}}(\tilde{\gamma}_k,\tilde{\boldsymbol{\Sigma}}_k) + \frac{1}{2}\Big(d(\log\tilde{\eta}_k - \log 2\pi) \\
&\quad - \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] - \tilde{\eta}_k\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k)\right]\Big) - \log M_k \\
&\quad + (\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right] - 1)\log\tilde{p}_k - \tilde{r}_k\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_k)\right] + \tilde{s}_k\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_k\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_k\right] \\
&\quad - \tilde{q}_k\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_k\right]
\end{aligned}$$

with

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k)\right] = \frac{d}{\tilde{\eta}_k}\,.$$

2.3.5 Expectations from variational distributions

Expectations developed in variational calculations are derived from properties of variational posterior distributions and are obtained as follows. Categorical distribution properties lead to

$$\begin{aligned}
&\forall j\in\mathcal{J},\ \forall k\in\mathcal{K}: \\
&\qquad\qquad \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] = \tilde{r}_{jk}\,.
\end{aligned}$$

Dirichlet distribution properties lead to

$$\begin{aligned}
&\forall k\in\mathcal{K}: \\
&\qquad \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k\in\mathcal{K}}\tilde{\kappa}_k\right)\,,
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function. Gamma distribution properties lead to

$$\begin{aligned}
&\forall j\in\mathcal{J},\ \forall k\in\mathcal{K}: \\
&\qquad\qquad \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] = \frac{\tilde{\alpha}_{jk}}{\tilde{\beta}_{jk}}\,, \\
&\qquad\qquad \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] = \psi(\tilde{\alpha}_{jk}) - \log\tilde{\beta}_{jk}\,.
\end{aligned}$$

Normal distribution properties lead to

$$\begin{aligned}
&\forall k \in \mathcal{K} : \\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_k\right] = \tilde{\boldsymbol{\mu}}_k \ , \\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\right] = \mathbb{V}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_k\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_k\right] \mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_k\right]^T \\
&\qquad\qquad = \tilde{\eta}_k^{-1} \boldsymbol{\Sigma}_k + \tilde{\boldsymbol{\mu}}_k \tilde{\boldsymbol{\mu}}_k^T \ ,
\end{aligned}$$

Inverse Wishart distribution properties lead to

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Sigma}_k^{-1}\right] = \tilde{\gamma}_k \tilde{\boldsymbol{\Sigma}}_k^{-1} \ , \\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\log |\boldsymbol{\Sigma}_k|\right] = \log |\tilde{\boldsymbol{\Sigma}}_k| - \sum_{i=1}^{d} \psi\left(\frac{\tilde{\gamma}_k + 1 - i}{2}\right) - d \log 2 \ .
\end{aligned}$$

Posterior expectations of β_k are derived from the posterior Gamma distribution (2.26) properties and can easily be computed conditionally to α_k

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}}[\beta_k] = \frac{\tilde{s}_k \mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] + 1}{\tilde{q}_k} \ , \\
&\mathbb{E}_{\boldsymbol{\Theta}}[\log \beta_k] = \mathbb{E}_{\boldsymbol{\Theta}}[\psi\left(\tilde{s}_k \alpha_k + 1\right)] - \log \tilde{q}_k \ .
\end{aligned}$$

However, expectations depending on α_k are intractable

$$\mathbb{E}_{\boldsymbol{\Theta}}[\psi\left(\tilde{s}_k \alpha_k + 1\right)] = \int \psi\left(\tilde{s}_k \alpha_k + 1\right) p(\alpha_k | \tilde{p}_k, \tilde{r}_k) d\alpha_k \ , \tag{2.34}$$

$$\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] = \int \alpha_k p(\alpha_k | \tilde{p}_k, \tilde{r}_k) d\alpha_k \ , \tag{2.35}$$

$$\mathbb{E}_{\boldsymbol{\Theta}}[\log \Gamma(\alpha_k)] = \int \log \Gamma(\alpha_k) p(\alpha_k | \tilde{p}_k, \tilde{r}_k) d\alpha_k \ . \tag{2.36}$$

Since lower bound calculation is required as a stop criterion, expectations (2.34), (2.35) and (2.36) have to be approximated. A deterministic method [TK86] based on Laplace approximation is then applied. This method consists in approximating integrals of a smooth function times the posterior $h(\alpha)p(\alpha|p,q,s,r)$ with an approximation proportional to a normal density in θ such that

$$\mathbb{E}[h(\alpha] = h(\alpha_0) p(\alpha_0 | p, q, s, r) (2\pi)^{d_\alpha/2} | - u''(\alpha_0)|^{1/2} \ ,$$

where d_α is the dimension of α, $u(\alpha) = \log\left(h(\alpha)p(\alpha|p,q,r,s)\right)$ and α_0 is the point at which $u(\alpha)$ is maximized.

In the case of unnormalized density $q(\alpha|p,q,r,s)$, Laplace's method can be applied separately to hq and q to evaluate the numerator and denominator here :

$$\mathbb{E}[h(\alpha] = \frac{\int h(\alpha) q(\alpha | p, q, s, r) d\alpha}{\int q(\alpha | p, q, s, r) d\alpha} \ .$$

2.4 Experiments

In this section, the proposed method is performed on the set of acquired data. For comparison, a standard neural network (NN), the k-nearest neighbours (k-NN) algorithm, Random Forests

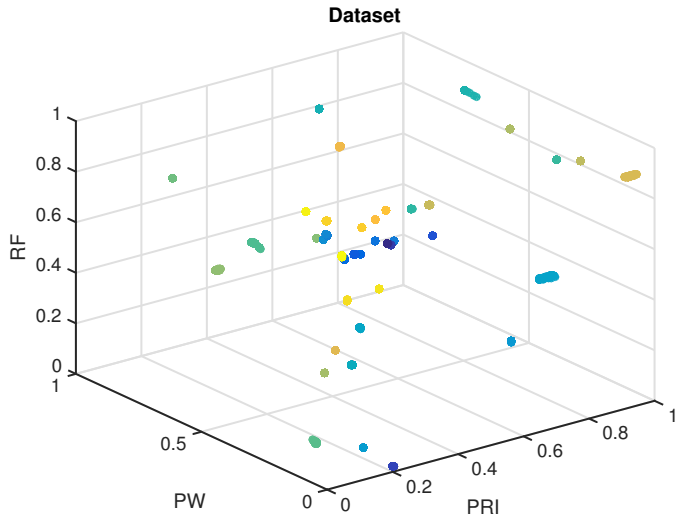


Figure 2.11: Dataset gathering 6300 observations from 42 radar emitters. Some clusters are completely separable whereas some others share features and can not be linearly separated.

(RdF) the k-means algorithm and the DBSCAN are also evaluated. Two experiments are carried out to evaluate classification and clustering performance with respect to a range of percentages of missing values. First, characteristics for realistic data acquisition and imputation methods for missing data are detailed. Then, both experiments are described with their error measure and their performance are shown to exhibit the effectiveness of the proposed model.

2.4.1 Data

Real data are acquired from the system detailed in Section 2.1. For each recording, the sampling frequency and the observation time T are respectively chosen as 4.17 MHz and 20 ms. The database exactly gathers 42 different radars waveforms and 150 observations are recorded for each waveform. Outliers and missing values are naturally embedded in observations due to material defects and real conditions detailed in Section 2.1. The dataset is shown in Figure 2.11. However, extra missing values are added to evaluate limits of the proposed approach. Missing information are introduced by randomly deleting coordinates of $(\boldsymbol{x}_j)_{j=1}^{150}$ for each of the 42 radar emitters. Percentages of deletion range from 5% to 40%. Nevertheless, comparison algorithms do not handle datasets including missing values. Discarding observations that contain missing values can be a restrictive solution, therefore imputation methods have been developed [GLSGFV10]. In this chapter, two classical imputation methods, based on statistical analysis and machine learning, are performed. First, the mean imputation consists in filling a missing component of an observation by the average of observed values of that component. This method has the obvious disadvantage that it under represents the variability and also ignores correlations between observations [Sch97]. Then, imputation can be processed through a K-nearest neighbours method [HTS+01] in order to replace missing values of an observation with a weighted mean of the k nearest completed observations where the weights are inversely proportional to the distances from the neighbours. Since replacements are influenced only by the most similar cases, the KNN method is more robust with respect to the amount and type of missing data [TCS+01]. These imputation methods are compared with the proposed approach in terms of classification, clustering and reconstruction performance. For the comparison of reconstruction performance, mean-squared errors between original data and previous imputation methods are compared with

Table 2.1: Initialisation of hyper-parameters values for classification on continuous data

κ_0	η_0	γ_0	p_0	r_0	q_0	s_0
0.5	10^{-4}	4	0.9	1	1	1

the mean-squared error between original data and the variational posterior marginal mean of missing data given by

$$\begin{aligned} \forall j \in \mathcal{J},\ \mathbb{E}_q[\boldsymbol{x}_j^{\text{miss}}] &= \mathbb{E}_q(\int\int q(\boldsymbol{x}_j^{\text{miss}}, u_j, z_j) du_j dz_j) \\ &= \sum_{k\in\mathcal{K}} \tilde{r}_{jk} \tilde{\boldsymbol{\mu}}_{jk}^{\boldsymbol{x}^{\text{miss}}} \quad . \end{aligned} \tag{2.37}$$

2.4.2 Classification experiment

The classification experiment evaluates the ability of each algorithm to assign unlabeled data to one of the K classes trained by a set of labeled data. The classification task is decomposed into a training step and a prediction step defined in procedures 2.1 and 2.2. The training step consists in estimating variational parameters of $q(\boldsymbol{\Theta})$ defined in (2.26) given a set of training data with known labels. As for the prediction step, it results in associating new data to the class that maximizes their posterior probabilities. Since comparison algorithms do not handle datasets including missing values, a complete dataset is used to enable their training. During the prediction step, incomplete observations are either discarded and gathered in a reject class or completed thanks to the mean and KNN imputation methods. Standard configurations provided by Matlab are chosen for the RnF, the NN and the KNN algorithm. The proposed model and comparisons algorithms are trained on 70% of the initial database without extra missing values and tested on the remaining 30% of the database whose elements are randomly deleted according to different proportions of missing values. The RnF gathers 50 trees. The NN is composed of one hidden layer of 70 neurons and a softmax output layer and is trained with a cross-entropy loss. An accuracy metric is chosen for the classification experiment and observations belonging to the reject class are considered as misclassification errors. For each experiment, hyper-parameters are initialised as in Table 2.1 and 100 simulations are performed to take into account randomness of data deletion.

Procedure 2.1 Classification procedure on continuous data : Training step

Input: Training set $\boldsymbol{x}^{\text{train}}$ and associated labels $\boldsymbol{z}^{\text{train}}$
Output: Learned parameters $\tilde{\boldsymbol{\Theta}}_{\text{train}}$
  Initialise κ_0, γ_0, η_0 , $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, p_0, r_0, s_0 and q_0
  for iter = 1 **to** itermax **do**
    Update $\tilde{\alpha}_{jk}$, $\tilde{\beta}_{jk}$, $\tilde{\boldsymbol{\mu}}_{jk}^{\boldsymbol{x}^{\text{miss}}}$, $\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}^{\text{miss}}}$
    Update $\tilde{\kappa}_k$, $\tilde{\eta}_k$, $\tilde{\gamma}_k$, $\tilde{p}_k$, $\tilde{r}_r$, $\tilde{s}_k$, $\tilde{q}_k$, $\tilde{\boldsymbol{\mu}}_k$, $\tilde{\boldsymbol{\Sigma}}_k$
    Calculate the lower bound $\mathcal{L}$
    if $\mathcal{L}_{\text{iter}} - \mathcal{L}_{\text{iter}-1} \leq \text{tol} \times \mathcal{L}_{\text{iter}-1}$ **then**
      return $\tilde{\boldsymbol{\Theta}}_{\text{train}} = \left(\tilde{\kappa}_k, \tilde{\eta}_k, \tilde{\gamma}_k, \tilde{p}_k, \tilde{r}_r, \tilde{s}_k, \tilde{q}_k, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k\right)_{k\in\mathcal{K}}$
    end if
  end for

Procedure 2.2 Classification procedure on continuous data : Prediction step

Input: Unlabelled dataset $\boldsymbol{x}^{\text{pred}}$ and learned parameters $\tilde{\boldsymbol{\Theta}}^{\text{train}}$
Output: Predicted labels $\tilde{\boldsymbol{z}}^{\text{pred}}$

Update $\tilde{\alpha}_{jk}$, $\tilde{\beta}_{jk}$, $\tilde{\boldsymbol{\mu}}_{jk}^{\boldsymbol{x}^{\text{miss}}}$, $\tilde{\boldsymbol{\Sigma}}_{k}^{\boldsymbol{x}^{\text{miss}}}$, $\tilde{r}_{jk}$

return $\tilde{\boldsymbol{z}}^{\text{pred}}$ such that each $\tilde{z}_j^{\text{pred}} = \arg\max_{k \in \mathcal{K}} \tilde{r}_{jk}$

For the classification experiment, results are shown in Figure 2.12. Without missing data, both algorithms perfectly classify the 42 radar emitters. When the proportion of missing values increases, the proposed model outperforms comparisons algorithms and achieves an accuracy of 85% for 40% of deleted values whereas the accuracy of NN and KNN is lower than 50% with or without missing data imputation. As for the RnF, it outperforms both NN and KNN by achieving accuracies of 67% and 72% with standard imputation methods for 40% of deleted values. This higher performance of the proposed model reveals that the proposed method embeds a more efficient inference method than other imputation methods. That result is confirmed on Figure 2.12 where comparison algorithms are applied on data reconstructed by the proposed model. Indeed when the proposed inference is chosen, performance of NN and KNN increase up to 80% for 40% of deleted values and the RnF has almost the same performance than the proposed model. The Figure 2.12 also reveals that the proposed approach is more robust to missing data since it has a lower variance than other algorithms and imputation methods. Finally, this efficiency is shown on Figure 2.13 where the proposed model exhibits a lower mean-squared error for missing data imputation than the mean and KNN imputation methods. Effectiveness of the proposed model can be explained by the fact that missing data imputation methods can create outliers that deteriorate performance of classification algorithms whereas the inference on missing data and labels prediction are jointly estimated in the proposed model. Indeed, embedding the inference procedure into the model framework allows properties of the model, such as outliers handling, to counterbalance drawbacks of imputation methods such as outlier creation.

Concerning the computational burden of the proposed approach, Figure 2.14 shows the evolution of computing times taken by model learning of the proposed model and comparison algorithms according to different numbers of observations. Considering that the learning of the proposed model is done offline and that its code can be drastically optimized since it is only developed under Matlab, the computational burden of the proposed approach is acceptable. Indeed the proposed model is ten times slower than the RnF but shares similar computing times with the NN when the number of observations increases. Moreover, once the model learning has been performed offline, predictions can be done online in real time.

2.4.3 Clustering experiment

The clustering experiment is composed of two experiments that aim to exhibit the clustering ability of each algorithm according to an a priori number of clusters $K \in \{K_{\min}, \ldots, K_{\max}\}$. As developed in the previous chapter, the clustering algorithm is decomposed into two parts. First, a semi-supervised classification is performed for each K ranges from $K_{\min}$ to $K_{\max}$ to estimate variational parameters of $q(\boldsymbol{\Theta}, \boldsymbol{h})$ in (2.26) and labels of data in a mixture of K components. Then, the value of K that maximizes the lower bound (2.25) is retained as the posterior number of clusters as well as its associated parameters. According to the dataset visualised in Figure 2.11, $K_{\min}$ and $K_{\max}$ are set to 12 and 72 in order to evaluate the impact of the a priori number of clusters on data clustering. Parameters of DBSCAN are set to Minpts = 4 and eps = 8e-3

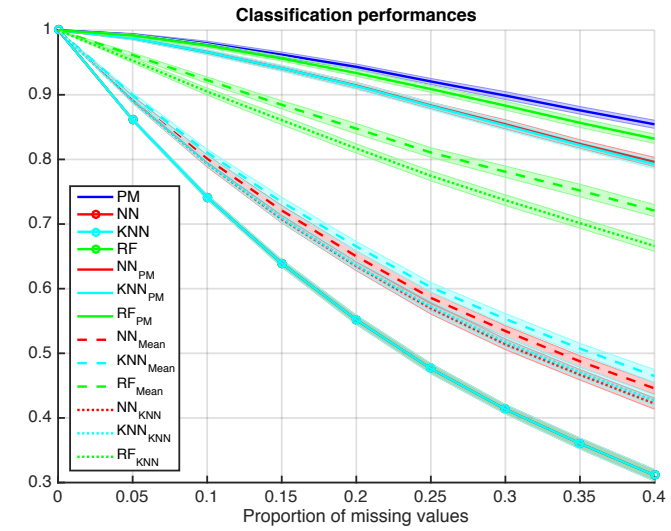


Figure 2.12: Classification performance are presented for the proposed model (PM) in blue, the NN in red, the RnF in green and the KNN in cyan. The solid lines represent the average accuracies with discarded observations for the NN, the RnF and the KNN, the dashed lines stands for the average accuracies with mean imputation for the NN, the RnF and the KNN whereas the doted lines shows average accuracies with KNN imputation. Shaded error regions represent standard deviations of accuracies.

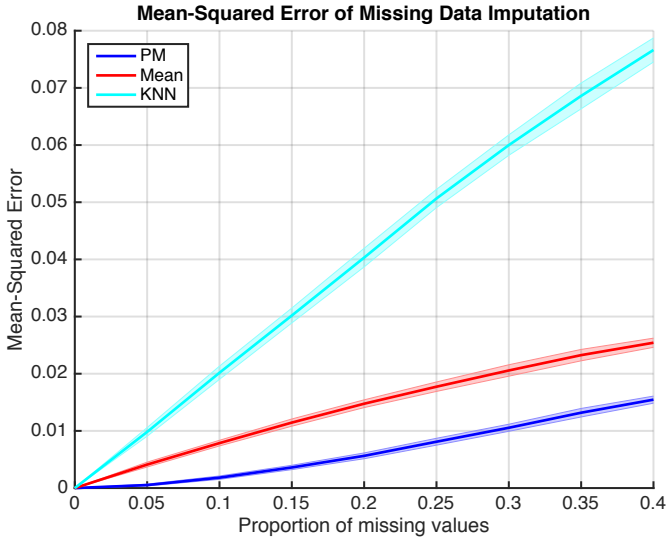


Figure 2.13: Mean-squared errors of missing data imputation methods are presented in blue for the proposed model, in red for the NN and in cyan for the KNN. Solid lines are average mean-squared errors and shaded error regions represent standard deviations of mean-squared errors.

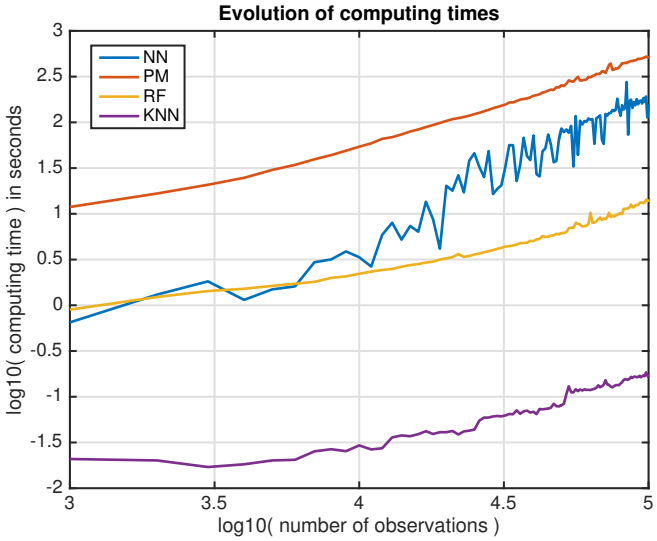


Figure 2.14: Evolution of computing times taken by model learning for Random Forests (RF), K nearest neighbors algorithm (KNN), Neural Network (NN) and the proposed model (PM).

Procedure 2.3 Semi-supervised classification procedure on continuous data

Input: Unlabelled dataset $\boldsymbol{x}$ and number of classes K
Output: Labels $\tilde{\boldsymbol{z}}$ and parameters $\tilde{\boldsymbol{\Theta}}$

Initialise κ_0, γ_0, η_0 , $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, p_0, r_0, s_0 and q_0
for iter $= 1$ **to** itermax **do**
  Update $\tilde{\alpha}_{jk}$, $\tilde{\beta}_{jk}$, $\tilde{\boldsymbol{\mu}}_{jk}^{\boldsymbol{x}^{\text{miss}}}$, $\tilde{\boldsymbol{\Sigma}}_{k}^{\boldsymbol{x}^{\text{miss}}}$, $\tilde{r}_{jk}$
  Update $\tilde{\kappa}_k$, $\tilde{\eta}_k$, $\tilde{\gamma}_k$, $\tilde{p}_k$, $\tilde{r}_r$, $\tilde{s}_k$, $\tilde{q}_k$, $\tilde{\boldsymbol{\mu}}_k$, $\tilde{\boldsymbol{\Sigma}}_k$
  Calculate the lower bound $\mathcal{L}$
  if $\mathcal{L}_{\text{iter}} - \mathcal{L}_{\text{iter}-1} \leq \text{tol} \times \mathcal{L}_{\text{iter}-1}$ **then**
    return $\tilde{\boldsymbol{\Theta}} = \left(\tilde{\kappa}_k, \tilde{\eta}_k, \tilde{\gamma}_k, \tilde{p}_k, \tilde{r}_r, \tilde{s}_k, \tilde{q}_k, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k\right)_{k \in \mathcal{K}}$ and $\tilde{\boldsymbol{z}}$ such that each $\tilde{z}_j = \arg\max_{k \in \mathcal{K}} \tilde{r}_{jk}$
  end if
end for

Procedure 2.4 Clustering procedure on continuous data

Input: Unlabelled dataset $\boldsymbol{x}$ and a priori range of numbers of clusters $K \in \{K_{\min}, \ldots, K_{\max}\}$
Output: Labels $\tilde{\boldsymbol{z}}$, parameters $\tilde{\boldsymbol{\Theta}}$ and optimal number of clusters $\tilde{K}$

for $K = K_{\min}$ **to** $K_{\max}$ **do**
  Perform semi-supervised classification with K classes
  Stock labels $\tilde{\boldsymbol{z}}^K$, parameters $\tilde{\boldsymbol{\Theta}}^K$ and $\mathcal{L}^K$
end for
return $\tilde{\boldsymbol{z}}^{\tilde{K}}$ and $\tilde{\boldsymbol{\Theta}}^{\tilde{K}}$ such that $\tilde{K} = \arg\max_K \mathcal{L}^K$

Table 2.2: Initialisation of hyper-parameters values for clustering on continuous data

κ_0	η_0	γ_0	p_0	r_0	q_0	s_0
0.5	100	4	1	1	1	1

by using an heuristic proposed in the original paper [EKS+96]. A supervised initialisation is retained for the proposed model due to its sensitivity to initialisation. It consists in initialising prior component means $\boldsymbol{\mu}_0$ from results of a k-means algorithm and prior component covariance matrices $\boldsymbol{\Sigma}_0$ from diagonal matrices whose diagonal elements are variances of observed features. Since comparison algorithms do not handle observations with missing values and do not provide a clustering result for them, missing data are either discarded and gathered in a reject class or completed thanks to the mean and KNN imputation methods before running these algorithms. For each experiment, hyper-parameters are initialised as in Table 2.2 and 100 simulations are performed to take into account randomness of data deletion.

The first clustering experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters $K \in \{K_{\min}, \ldots, K_{\max}\}$. Performance are evaluated through the Adjusted Rand Index (ARI) [HA85] that compares estimated partitions of data with the ground-truth. Results of the first experiment on realistic data are shown in Figures 2.15 and 2.16. Without the presence of missing values, performance of DBSCAN and the proposed model are similar with an ARI of 97% (Figure 2.15) whereas the k-means algorithm ARI reaches 95% (Figure 2.16). When the proportion of missing values increases, the proposed model outperforms both DBSCAN and k-means and achieves an ARI of 87% for 40% of deleted values whereas the ARI of comparison algorithms is lower than 30% with standard missing data imputation. This higher performance reveals that the proposed method embeds a more efficient inference method than other imputation methods. That result is confirmed on both Figure 2.15 and Figure 2.16 where DBSCAN and k-means are applied on data reconstructed by the proposed model. Indeed, performance of both algorithms increase up to 77% and 69% for 40% of deleted values when the proposed inference is chosen.

The second experiment tests the ability of each algorithm to find the true number of clusters $\tilde{K}$ among $\{K_{\min}, \ldots, K_{\max}\}$. The lower bound (2.25) and the average Silhouette score [KR09] are criteria used to select the optimal number of clusters for the proposed model and the k-means algorithm. Indeed, the ARI can not be used since it requires the ground-truth and DBSCAN automatically selects a number of clusters for a given dataset. Results of the second experiment on realistic data are visible on Figures 2.18 and 2.17. Figure 2.18 shows the evolution of the number of clusters estimated by DBSCAN according to different proportions of missing values and imputation methods. Since DBSCAN automatically estimates the number of clusters and manages outliers by creating new clusters, results on Figure 2.18 can be used to evaluate performance of imputations methods. For mean and k-NN imputation methods, DBSCAN estimates a number of clusters greater than 140 as proportion of missing values is equal or greater than 5%. When DSBCAN is run on the posterior reconstruction (2.37), the estimated number of clusters stays under 50 until 20% of missing values and reaches 70% for 40% of missing values. These performance indicate that the proposed approach creates less outliers than other imputation methods by providing a more robust inference on missing data since DBSCAN localizes less outliers in the posterior reconstruction (2.37) than in standard imputation methods. Figure 2.17 presents numbers of clusters selected by the lower bound and average Silhouette scores for the proposed model and k-means algorithm according to different proportions of missing values and

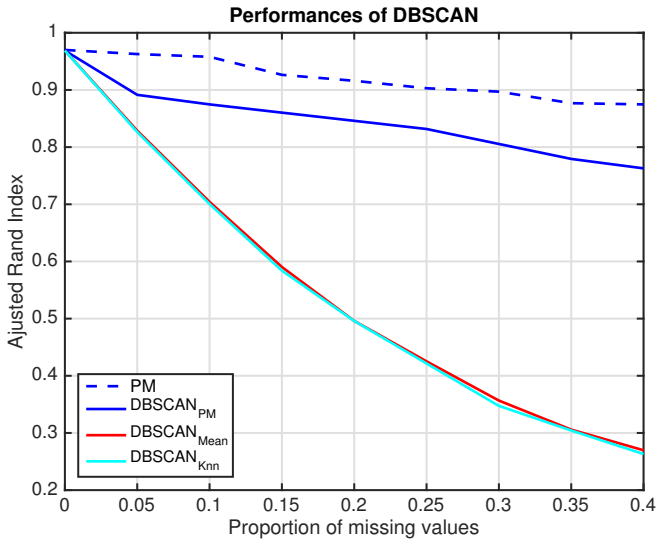


Figure 2.15: Performance of the proposed model compared with DBSCAN for K = 42 according to different proportions of missing values and imputation methods on realistic data.

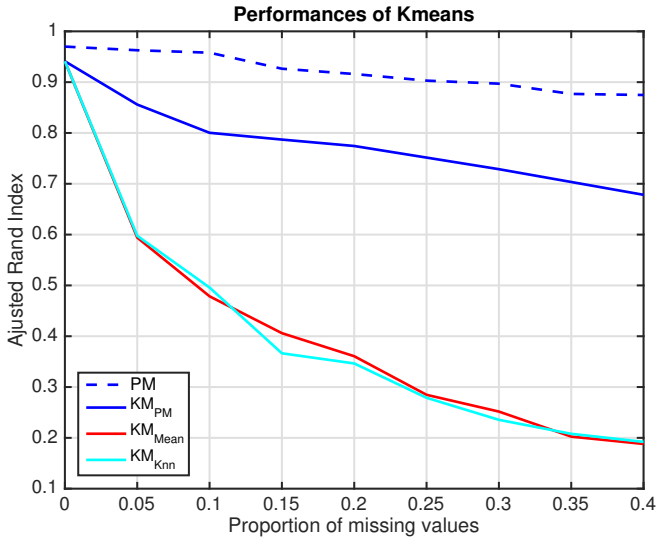


Figure 2.16: Performance of the proposed model compared with k-means algorithm for K = 42 according to different proportions of missing values and imputation methods on realistic data.

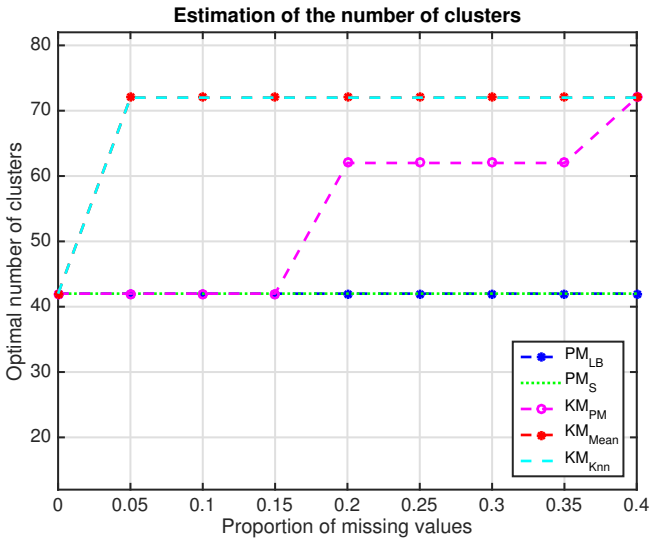


Figure 2.17: Estimation of the number of clusters using the lower bound (LB) and the silhouette score (S) for the proposed model and only the silhouette score (S) for the k-means algorithm.

imputation methods. Without missing data, the correct number of clusters (K=42) is selected by the two criteria for the k-means algorithm and the proposed model. In presence of missing values, the average Silhouette score always selects $K = 72$ when the k-means algorithm is run on data completed by standard imputation methods. When, the k-means algorithm performs clustering on the posterior reconstruction (2.37), the average Silhouette score correctly selects $K = 42$ until 15% of missing values and chooses $K \in \{62, 72\}$ when the proportion of missing values is greater than 20%. Eventually when the proposed model does clustering, the two criteria select the correct number of clusters $K = 42$ for every proportion of missing values. These results show two main advantages of the proposed model. As previously, the proposed model provides a more robust inference on missing data since the average Silhouette score chooses more representative number of clusters when the k-means algorithm is run on the posterior reconstruction (2.37) than on data completed by standard imputation methods. Furthermore, since the lower bound criterion also selects the correct number of clusters as the average Silhouette score, it can be used as a valid criterion for selecting the optimal number of clusters and does not require extra computational costs as the Silhouette score since it is computed during the model parameters estimation. Finally, the proposed approach provides a more robust inference on missing data and a criterion for selecting the optimal number of clusters without extra computations.

Figure 2.19 shows the evolution of computing times taken by model learning of the proposed model and comparison algorithms according to different numbers of clusters and observations. As the model learning in Subsection 2.4.2, clustering is only performed offline to extract information from radar signals recorded during operational missions. Even if the proposed method is ten times slower than the k-means algorithm, the computational burden of the proposed approach is still acceptable and meets operational requirements.

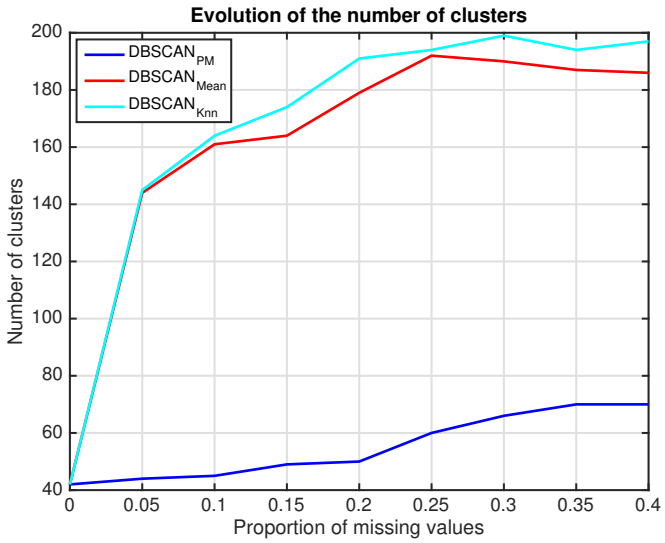


Figure 2.18: Estimation of the number of clusters by DBSCAN according to mean imputation, k-NN imputation and posterior reconstruction of the proposed model.

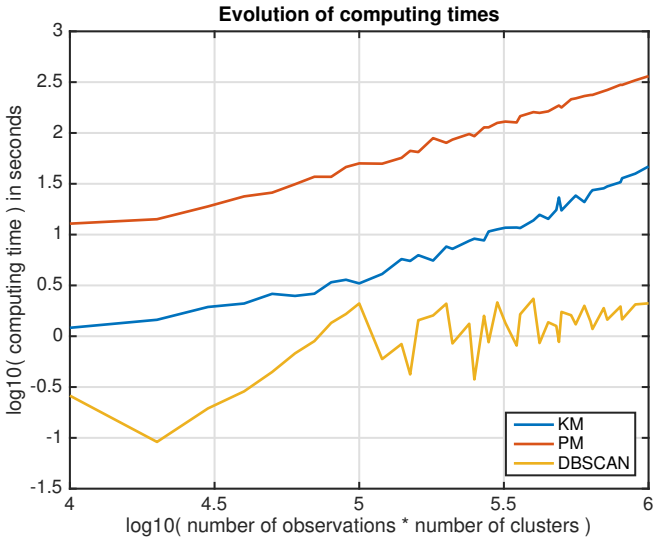


Figure 2.19: Evolution of computing times for DBSCAN, k-means algorithm (KM) and the proposed model (PM).

2.5 Conclusion

In this section, we propose a mixture model to classify and cluster radar emitters. Radar signals are often partially observed due to imperfect conditions of acquisition and deficient hardwares. Therefore to account for missing data and outliers, a scale mixture of Normal distributions, known for its robustness to outliers and its flexible framework for classification and clustering, is chosen. Moreover, thanks to the introduction of latent variables, the proposed model can infer on missing data. Since the posterior distribution is intractable, model learning is processed through a Variational Bayes inference where a variational posterior distribution is proposed for missing values. Experiments on various real data showed that the proposed approach handles both outliers and missing values and can outperform standard algorithms in classification and clustering tasks. Indeed the main advantage of our approach is that it allows properties of the model, such as outliers handling, to counterbalance drawbacks of imputation methods by embedding the inference procedure into the model framework.

Chapter 3

Mixed data

Continuous data describing radar emitter waveforms such as the Carrier Frequency, the Pulse Width and the Pulse Repetition Interval have been previously taken into account in order to cluster radar emitters. Nonetheless, these continuous features are frequently modulated to enhance functions of the radar emitters. Therefore, these modulations can be exploited as categorical features to cluster radar emitters. According to types of modulations, a dependence structure can be established to model conditional relations between continuous and categorical features. This dependence structure is then included into the previous mixture model to take advantage of specific patterns related to each radar emitter. This chapter contains four sections which focus on the integration of mixed data to enhance classification and clustering performance. Section 3.1 presents assumptions on continuous and categorical features of radar emitters. Section 3.2 introduces the dependence structure of mixed data and the proposed model. Section 3.3 details the inference procedure for the estimation of parameters related to the proposed model. At last in Section 3.4, various experiments are carried out to exhibit performance of the proposed approach.

Contents

3.1 Data

In this chapter, data consist of J pulses gathering J continuous features $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ and J categorical features $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ from K distinct emitters. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$ the j^{th} observation vector of mixed variables where

- $\boldsymbol{x}_{qj} \in \mathbb{R}^d$ is a vector of d continuous radar features such as the Radio Frequency, the Pulse Width and the Pulse Repetition Interval,
- $\boldsymbol{x}_{cj} = \left(x_{cj}^0, \ldots, x_{cj}^{q-1}\right) \in \mathcal{C}_q$ is a vector of q categorical radar modulations such as intrapulse modulations, pulse-to-pulse modulations or scanning types.

Radar features and distributions related to continuous and categorical data are presented in the following subsections.

3.1.1 Assumptions on continuous data

In this subsection, continuous radar features are first recalled from the previous chapter. Then, the distribution of continuous data is presented.

Radar Features

As in the previous chapter, continuous features of a radar emitter are extracted from its Pulse Description Words (PDW). Each PDW gathers the radio frequency (RF), the pulse width (PW) and the pulse repetition interval (PRI) of a given pulse in the radar signal pattern. These continuous features are exhibited on Figure 3.1.

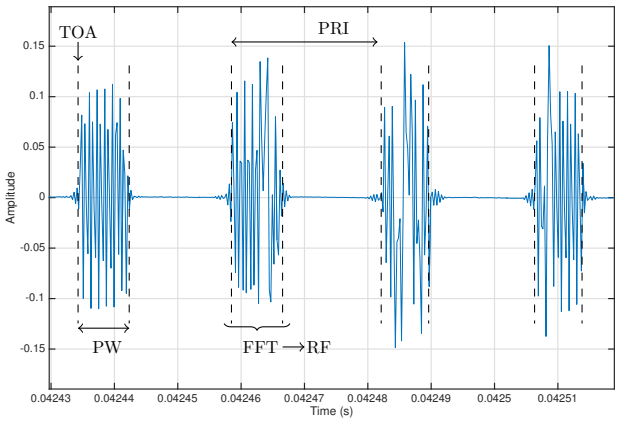


Figure 3.1: Acquired pulses from a radar emitter where the three features (PRI,PW,RF) are shown on the figure.

Distribution of continuous features

The continuous features of the j^{th} observation are modeled through $\boldsymbol{x}_{qj} \in \mathbb{R}^d$ which is a vector of d continuous variable distributed according to a multivariate normal distribution with mean and variance parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Indeed through its properties, the multivariate distribution can enhance the dependence structure between continuous variables in order to handle outliers and missing values.

Outliers Outliers for continuous data $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ are handled as the previous chapter by introducing the scale latent variables $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$ such that

$$\forall j \in \mathcal{J}, \qquad \begin{aligned} &\boldsymbol{x}_{qj}|u_j \sim \mathcal{N}(\boldsymbol{x}_{qj}|\boldsymbol{\mu}, u_j^{-1}\boldsymbol{\Sigma}) \ , \\ &u_j \sim \mathcal{G}(u_j|\alpha,\beta) \ , \end{aligned}$$

where each u_j follows a Gamma distribution with shape and rate parameters $(\alpha,\beta) \in \mathbb{R}^{*+}\times\mathbb{R}^{*+}$.

Missing values Since continuous data $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ can be partially observed, they are decomposed into observed features $\boldsymbol{x}_q^{\text{obs}} = (\boldsymbol{x}_{qj}^{\text{obs}})_{j\in\mathcal{J}}$ and missing features $\boldsymbol{x}_q^{\text{miss}} = (\boldsymbol{x}_{qj}^{\text{miss}})_{j\in\mathcal{J}}$ such that

$$\forall j \in \mathcal{J}, \quad \boldsymbol{x}_{qj} = \begin{pmatrix} \boldsymbol{x}_{qj}^{\text{miss}} \\ \boldsymbol{x}_{qj}^{\text{obs}} \end{pmatrix} \text{ with } (\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{qj}^{\text{obs}}) \in \mathbb{R}^{d_j^{\text{miss}}} \times \mathbb{R}^{d_j^{\text{obs}}} \text{ and } d_j^{\text{miss}} + d_j^{\text{obs}} = d \ ,$$

where $\mathbb{R}^{d_j^{\text{miss}}}$ and $\mathbb{R}^{d_j^{\text{obs}}}$, are disjoint subsets of $\mathbb{R}^d$ embedding missing features $\boldsymbol{x}_{qj}^{\text{miss}}$ and observed features $\boldsymbol{x}_{qj}^{\text{obs}}$. Then, properties of the multivariate normal distribution leads to obtain two normal distributions for observed and missing features such that

$$\forall j \in \mathcal{J}, \qquad \begin{aligned} &\boldsymbol{x}_{qj}^{\text{miss}} \sim \mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{miss}}|\boldsymbol{\mu}_j^{\boldsymbol{x}_q^{\text{miss}}}, \boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{miss}}}\right) \ , \\ &\boldsymbol{x}_{qj}^{\text{obs}} \sim \mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{obs}}|\boldsymbol{\mu}_j^{\boldsymbol{x}_q^{\text{obs}}}, \boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{obs}}}\right) \ , \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}_j^{\boldsymbol{x}_q^{\text{miss}}} &= \boldsymbol{\mu}^{\text{miss}} + \boldsymbol{\Sigma}^{\text{cov}}{\boldsymbol{\Sigma}^{\text{obs}}}^{-1}\left(\boldsymbol{x}_{qj}^{\text{obs}} - \boldsymbol{\mu}^{\text{obs}}\right) \ , \\ \boldsymbol{\mu}_j^{\boldsymbol{x}_q^{\text{obs}}} &= \boldsymbol{\mu}^{\text{obs}} \ , \\ \boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{miss}}} &= \boldsymbol{\Sigma}^{\text{miss}} - \boldsymbol{\Sigma}^{\text{cov}}{\boldsymbol{\Sigma}^{\text{obs}}}^{-1}\boldsymbol{\Sigma}^{\text{cov'}} \ , \\ \boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{obs}}} &= \left({\boldsymbol{\Sigma}^{\text{obs}}}^{-1} + 2\times{\boldsymbol{\Sigma}^{\text{obs}}}^{-1}\boldsymbol{\Sigma}^{\text{cov'}}\left(\boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{miss}}}\right)^{-1}\boldsymbol{\Sigma}^{\text{cov}}{\boldsymbol{\Sigma}^{\text{obs}}}^{-1}\right)^{-1} \ , \end{aligned}$$

and parameters $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ are decomposed as follows

$$\begin{aligned} \boldsymbol{\mu} &= \begin{pmatrix} \boldsymbol{\mu}^{\text{miss}} \\ \boldsymbol{\mu}^{\text{obs}} \end{pmatrix} \ , \\ \boldsymbol{\Sigma} &= \begin{pmatrix} \boldsymbol{\Sigma}^{\text{miss}} & \boldsymbol{\Sigma}^{\text{cov}} \\ \boldsymbol{\Sigma}^{\text{cov'}} & \boldsymbol{\Sigma}^{\text{obs}} \end{pmatrix} \ . \end{aligned}$$

3.1.2 Assumptions on categorical data

In this subsection, categorical radar features are first introduced. Then, the distribution of categorical data is presented.

Radar Features

Categorical radar features are mainly related to continuous radar features since they describe modulations of a radar emitter pattern. Depending on the nature of continuous features, different types of categorical features can be taken into consideration. Considering the continuous features RF, PRI and PW of a radar emitter, categorical features are modulations related to these

quantities. Then as regards a pattern of pulses, modulations of RF, PRI and PW are called pulse-to-pulse modulations when they are applied on a group of pulses whereas modulations of RF are called intrapulse modulations when they are applied on single pulses. These two families of modulations are used by radar emitters to achieve a common goal. Indeed, both families of modulations aim to obtain a higher resolution of targets by reducing ambiguities related to target range and target velocity. Furthermore, intrapulse modulations also focus on avoiding identification of radar emitters since they enable the minimization of these ambiguities even in presence of noise. Considering amplitudes of a radar emitter, the related categorical feature is the scanning behaviour of the radar emitter. These three types of categorical features are presented below.

Pulse-to-pulse modulations Pulse-to-pulse modulations consist in modulating parameters of a group of pulses to minimize ambiguities related to range and velocity. They are mostly applied on RF and PRI parameters through deterministic and random patterns. These various patterns are defined below for parameter values $(p_i)_{1\leq i\leq n}$ of a group of n pulses and are also visible in Figure 3.2.

Constant modulation When there is no modulation, all values $(p_i)_{1\leq i\leq n}$ are identical such that

$$\forall i \in \{1, \ldots, n\},\ p_i = v$$

where v is a constant.

Slide modulation When parameter values $(p_i)_{1\leq i\leq n}$ are sliding, they are linearly modulated around a nominal v value such that

$$\forall i \in \{1, \ldots, n\},\ p_i = a \times i + v$$

where a and v are the slope and the intercept of the linear function.

Dwell and Switch modulation When parameter values $(p_i)_{1\leq i\leq n}$ are piecewise constant, the emission of the n pulses is known to be dwelled or switched. Hence, for J disjoint subsets V_j forming a partition of $\{1, \ldots, n\}$, parameter values p_i are dwelled or switched if they are constant on each V_j such that

$$\forall i \in \{1, \ldots, n\},\ p_i = \sum_{j=1}^{J} v_j \mathbb{I}_{i\in V_j}$$

where v_j is the value of the piecewise V_j.

Stagger modulation The emission of the n pulses is staggered when parameter values $(p_i)_{1\leq i\leq n}$ are distributed according a sequence of q moments $(v_1, \ldots, v_q)$ such that

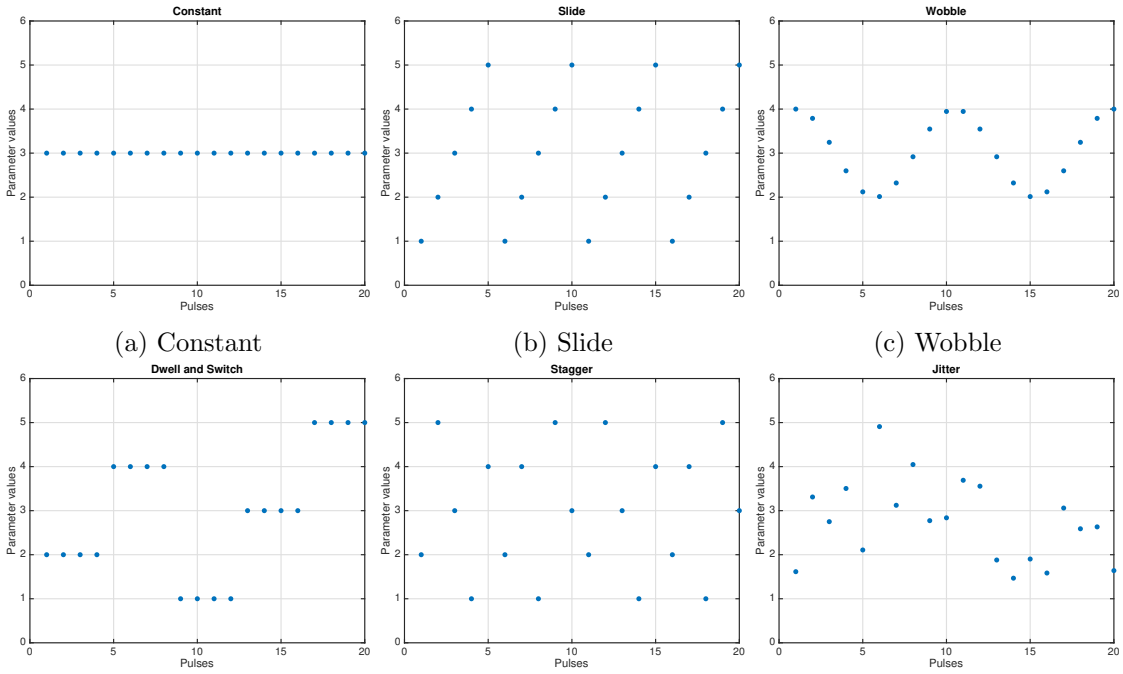
$$\forall i \in \{1, \ldots, n\},\ p_i = v_{m(i)}$$

where $m : \{1, \ldots, n\} \to \{1, \ldots, q\}$ is a surjective application associating moments to parameters values. If m produces a periodic sequence of the q moments, the stagger is regular.

Wobble modulation The emission of the n pulses is wobulated when parameter values $(p_i)_{1\leq i\leq n}$ are repetitively modulated through a periodic pattern such that

$$\forall i \in \{1, \ldots, n\},\ p_i = f(p_i)$$

where f is a periodic pattern usually chosen as a sinus wave or a triangular wave.



(a) Constant (b) Slide (c) Wobble

(d) Dwell and Switch (e) Stagger (f) Jitter

Figure 3.2: Examples of different modulations of parameter values. Figure (a) shows no modulation of parameter values. Figure (b) introduces sliding values on sequences of 5 pulses. Figure (c) exhibits a sinusoidal wobulated emission. Figure (d) presents a dwelled emission shaped with 5 piecewises. Figure (e) shows a staggered emission composed of 5 moments distributed according to a sequence mapped on 10 pulses. At last, Figure (f) presents a jittered emission where parameter values are normally distributed.

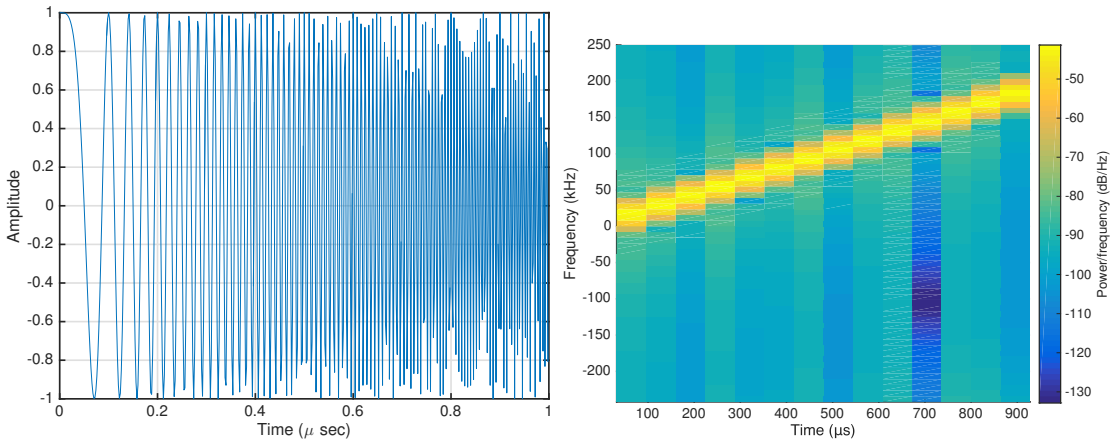
Jitter modulation The emission of n pulses is jittered when parameter values $(p_i)_{1\leq i\leq n}$ are randomly generated around a nominal value v. Jittered emissions are commonly Gaussian such that

$$\forall i \in \{1, \ldots, n\},\ p_i = v + \epsilon$$

where v is the nominal value and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise.

Intrapulse modulations To avoid identification of operating radars by ESM systems, radar designers have developed Low Probability of Intercept (LPI) waveforms. Theses waveforms are either frequency-modulated or phased-modulated in order to improve resolution for the radar emitter at the expense of a suboptimal signal-to-noise ratio (SNR) [LM04]. In other words, theses pulse modulations enable the maximization of the target range and the range resolution of radars. On the contrary, ESM resolution is less accurate since LPI signals are embedded in much noise and the identification task can be compromised. Intrapulse modulations are presented below.

Frequency Modulation Signal By spreading energy over a modulation bandwidth, Frequency Modulation (FM) signals provide a better range resolution than constant frequency signals. A Linear FM (LFM) signal, also known as a chirp, is obtained by swepting linearly the frequency band during the pulse duration. However, the LFM can involve relatively high autocorrelation sidelobes. Therefore to counter that drawback, Nonlinear FM (NLFM) [Cos84] signals are used to provide a more accurate spectrum which is shaped by deviating the constant



(a) Time domain representation of a Linear Frequency Modulation signal.

(b) Frequency domain representation of a Linear Frequency Modulation signal.

Figure 3.3: Linear Frequency Modulation on a pulse. The Figure (a), respectively Figure (b), shows a time domain representation, respectively a frequency domain representation, of a linear frequency modulation signal.

Table 3.1: All known binary Barker codes

Code length	Code
2	11 or 10
3	110
4	1110 or 1101
5	11101
7	1110010
11	11100010010
13	1111100110101

rate of frequency change. That non linear variation results in spending more time at frequencies that need to be enhanced and in avoiding high autocorrelation sidelobes.

Phase Modulation Signal Phase coding is one of the first methods for pulse compression. The concept rests on dividing a pulse of duration T into M bits of identical duration $t_b = \frac{T}{M}$ and assigning a different phase value to each bit. The main advantage of phase coding over frequency modulation is low peak side lobe level [LM04]. Barker codes [Bar53] are the most famous family of phase codes gathering 13 known binary sequences which were reported by [Bar53] and [Tur63] and are given in Table 3.1. Other polyphase codes such as the Frank code [FZH62] and the Zadoff code [Z+63] are widely used for pulse compression. Figure 3.4 exhibits an example of pulse compression with a Zadoff code.

Scanning types A radar emitter can truly differ from another one through its scanning pattern. While searching for targets across the environment, its beam steering can behave differently depending on its antenna shape, its composition and its mission. The most common scanning types are presented below and illustrated on Figure 3.5.

Circular Scan A circular scanning radar is a constant rotational scanning radar that provides accurate target range and azimuth information. It uses an antenna system that continuously

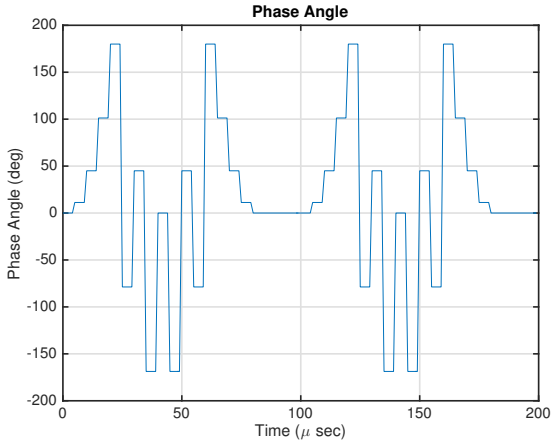


Figure 3.4: Phase coding generated from a Zadoff code.

scans 360° in azimuth making it ideal for the roles of early warning and initial target acquisition.

Sector Scan A sector scanning radar is a radar which scans unidirectionally or bidirectionally in a delimited sector. For example, the helical scan is a unidirectional scan pattern that enables a "pencil" beam to search a 360° pattern. As a bidirectional scan, the raster scan uses a thin beam to cover a rectangular area by scanning in azimuth and elevation.

Track-While-Scan (TWS) A track-while-scan (TWS) system generates two or more distinct radar beams that enable a radar to track multiple targets while scanning for others.

Electronic Scan An electronic scanning radar provides a computer-controlled scanning in which radar beams are electronically steered to point in different directions without moving the antenna.

Distribution of categorical features

The categorical features of the j^{th} observation are modeled through $\boldsymbol{x}_{cj} = \left(x_{cj}^0, \ldots, x_{cj}^{q-1}\right) \in \mathcal{C}_q$ which is a vector of q categorical variables where $\mathcal{C}_q = \mathcal{C}_0 \times \ldots \times \mathcal{C}_{q-1}$ is the tensor gathering each space $\mathcal{C}_i = \left\{m_1^i, \ldots, m_{|\mathcal{C}_i|}^i\right\}$ of events that x_{cj}^i can take $\forall i \in \{0, \ldots, q-1\}$.

As continuous data $\boldsymbol{x}_q$, categorical data $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j \in \mathcal{J}}$ can be partially observed. Hence $\boldsymbol{x}_c$ are decomposed into observed features $\boldsymbol{x}_c^{\text{obs}} = (\boldsymbol{x}_{cj}^{\text{obs}})_{j \in \mathcal{J}}$ and missing features $\boldsymbol{x}_c^{\text{miss}} = (\boldsymbol{x}_{cj}^{\text{miss}})_{j \in \mathcal{J}}$ such that

$$\forall j \in \mathcal{J}, \qquad \boldsymbol{x}_{cj} = \begin{pmatrix} \boldsymbol{x}_{cj}^{\text{miss}} \\ \boldsymbol{x}_{cj}^{\text{obs}} \end{pmatrix} \text{ with } (\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) \in \mathcal{C}_{q_j^{\text{miss}}} \times \mathcal{C}_{q_j^{\text{obs}}} \text{ and } q_j^{\text{miss}} + q_j^{\text{obs}} = q \; .$$

where $\mathcal{C}_{q_j^{\text{miss}}}$ and $\mathcal{C}_{q_j^{\text{obs}}}$, are disjoint subsets of $\mathcal{C}_q$ embedding missing features $\boldsymbol{x}_{cj}^{\text{miss}}$ and observed features $\boldsymbol{x}_{cj}^{\text{obs}}$.

If each categorical feature x_{cj}^i of the j^{th} observation $\boldsymbol{x}_{cj}$ is assumed to be independent from other features and distributed according to a categorical distribution, a dependence structure between features cannot be modeled and inference on missing categorical features cannot be handled. Hence, a multivariate categorical distribution integrating a dependence structure for

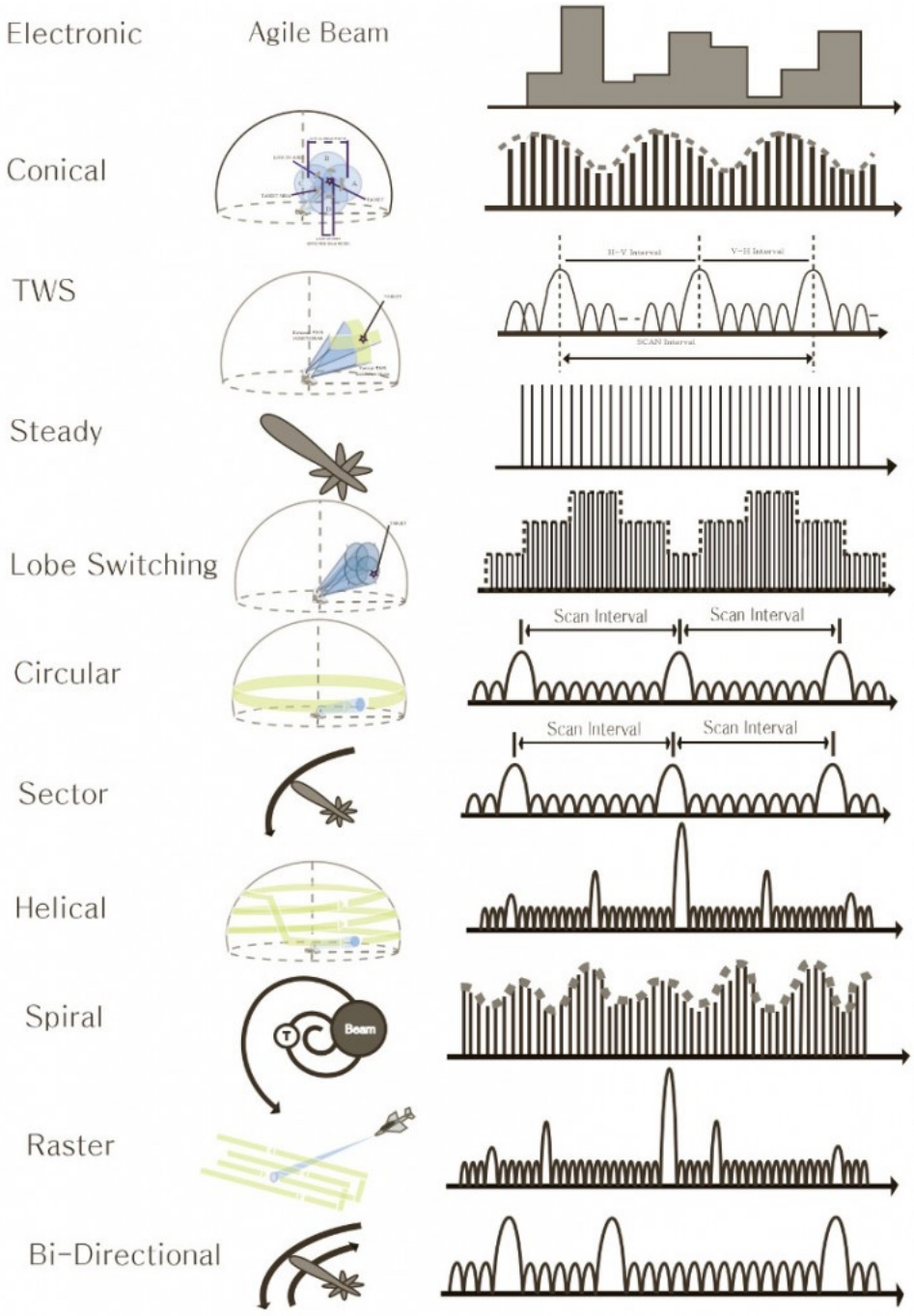


Figure 3.5: Different scanning types from [AWT18]. Sketches on the left illustrate scanning behaviours over the scanning period. Graphics on the right represent the evolution of pulse amplitudes over the scanning period.

$\boldsymbol{x}_{cj}$ is proposed to tackle that issue. As detailed below, $\boldsymbol{x}_{cj}$ follows a multivariate categorical distribution $\mathcal{MC}(\boldsymbol{x}_{cj}|\boldsymbol{\pi})$ if

$$p(\boldsymbol{x}_{cj}) = \prod_{\boldsymbol{c}\in\boldsymbol{C}_q} \pi_{\boldsymbol{c}}^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \tag{3.1}$$

where $\forall \boldsymbol{c} = (c^0, \ldots, c^{q-1}) \in \mathcal{C}_q = \mathcal{C}_0 \times \ldots \times \mathcal{C}_{q-1}$:

$$\sum_{\boldsymbol{c}\in\mathcal{C}_q} \pi_{\boldsymbol{c}} = 1 \ ,$$

$$\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}} = \begin{cases} 1 \text{ if } x_{cj}^0 = c^0, \ldots, x_{cj}^{q-1} = c^{q-1} \\ 0 \text{ otherwise} \end{cases} .$$

Noting that the dependence structure between categorical features is modeled through Kronecker symbols $(\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}$, this dependence structure can be exploited to handle missing features such that the multivariate categorical distribution in (3.1) can be written as

$$p(\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) = \prod_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}\in\boldsymbol{C}_{q_j^{\text{miss}}}\times\boldsymbol{C}_{q_j^{\text{obs}}}} \pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}^{\delta_{\boldsymbol{x}_{cj}^{\text{miss}},\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}}$$

where

$$\delta_{\boldsymbol{x}_{cj}^{\text{miss}},\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}} = \left.\begin{cases} 1 \text{ if } x_{cj}^{\text{miss}} = \boldsymbol{c}^{\text{miss}} \text{ and } x_{cj}^{\text{obs}} = \boldsymbol{c}^{\text{obs}} \\ 0 \text{ otherwise} \end{cases}\right\} = \delta_{\boldsymbol{x}_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}} \times \delta_{\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}} \ . \tag{3.2}$$

By using the previous equality (3.2), the multivariate categorical distribution in (3.1) becomes

$$p(\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) = \prod_{\boldsymbol{c}^{\text{miss}}\in\boldsymbol{C}_{q_j^{\text{miss}}}} \left(\prod_{\boldsymbol{c}^{\text{obs}}\in\boldsymbol{C}_{q_j^{\text{obs}}}} \pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}^{\delta_{\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}}} \right)^{\delta_{\boldsymbol{x}_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}}} \tag{3.3}$$

Therefore, a marginal distribution for observed features $\boldsymbol{x}_{cj}^{\text{obs}}$ and a conditional distribution for missing features $\boldsymbol{x}_{cj}^{\text{miss}}$ are obtained from (3.3) such that

$$p(\boldsymbol{x}_{cj}^{\text{obs}}) = \sum_{\boldsymbol{x}_{cj}^{\text{miss}}\in\boldsymbol{C}_{q_j^{\text{miss}}}} p(\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) = \prod_{\boldsymbol{c}^{\text{obs}}\in\boldsymbol{C}_{q_j^{\text{obs}}}} \left(\sum_{\boldsymbol{c}^{\text{miss}}\in\boldsymbol{C}_{q_j^{\text{miss}}}} \pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}} \right)^{\delta_{\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}}} ,$$

$$p(\boldsymbol{x}_{cj}^{\text{miss}}|\boldsymbol{x}_{cj}^{\text{obs}}) = \frac{p(\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}})}{p(\boldsymbol{x}_{cj}^{\text{obs}})} = \prod_{\boldsymbol{c}^{\text{miss}}\in\boldsymbol{C}_{q_j^{\text{miss}}}} \left(\prod_{\boldsymbol{c}^{\text{obs}}\in\boldsymbol{C}_{q_j^{\text{obs}}}} \left(\frac{\pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}}{\sum_{\boldsymbol{c}^{\text{miss}}\in\boldsymbol{C}_{q_j^{\text{miss}}}} \pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}} \right)^{\delta_{\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}}} \right)^{\delta_{\boldsymbol{x}_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}}} . \tag{3.4}$$

Then, a multivariate categorical distribution for missing features $\boldsymbol{x}_{cj}^{\text{miss}}$ conditionally to observed features $\boldsymbol{x}_{cj}^{\text{obs}}$ is deduced from (3.4) where $\forall(\boldsymbol{c}^{\text{miss}}, \boldsymbol{c}^{\text{obs}}) \in \boldsymbol{C}_{q_j^{\text{miss}}} \times \boldsymbol{C}_{q_j^{\text{obs}}}$:

$$p(\boldsymbol{x}_{cj}^{\text{miss}} = \boldsymbol{c}^{\text{miss}}|\boldsymbol{x}_{cj}^{\text{obs}} = \boldsymbol{c}^{\text{obs}}) = \frac{\pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}}{\sum_{\boldsymbol{c}^{\text{miss}}\in\boldsymbol{C}_{q_j^{\text{miss}}}} \pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}}$$

with $\pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}$ the joint probability $\pi_{\boldsymbol{c}}$ defined in (3.1) for $\boldsymbol{c} = (\boldsymbol{c}^{\text{miss}}, \boldsymbol{c}^{\text{obs}})$. Outliers are not considered for categorical data since in our case only reliable categorical variables are filled in databases and unreliable ones are processed as missing data.

3.2 Model

In this section, K emitters presenting mixed data are considered. Therefore, the main objective is to develop a mixture model which can build K distinct clusters even in presence of outliers and missing values. Before introducing a mixture model that handles mixed data, state-of-the-art mixture models for mixed data are reviewed. Then, component distributions for mixed data are defined and the mixture model is developed into a Bayesian framework.

3.2.1 State of the art

Two families of models emerge from finite mixture models fitting mixed-type data :

- The location mixture model [LK96] that assumes that continuous variables follow a multivariate Gaussian distribution conditionally on both component and categorical variables.
- The underlying variables mixture model [Eve88] that analyzes data sets with continuous and ordinal variables. It assumes that each discrete variable arises from a latent continuous variable and that all continuous variables (observed and unobserved) follow a Gaussian mixture model.

These two families are first detailed before introducing the retained approach.

Location Mixture Model

[LK96] introduced a location mixture model by assuming that the continuous variables are distributed as a finite mixture of Gaussians conditionally on the categorical variables. In other words, a Gaussian mixture exists for the continuous variables and its component mean vectors depend on the specific combination of categories modeled by the categorical variables. As pointed out by [WB99] the mixture of location models is not identifiable without imposing some constraints on the mean parameters of the Gaussian distributions. This is due to the indeterminacy of class memberships at each location. Even if all within component dependences are taken into account, we note that each combination of categories identifies a set of clusters. It follows that the total number of clusters can be unnecessarily large. A more parsimonious model is given by [HJ99], according to which the variables are decomposed into conditionally independent blocks containing a set of continuous variables or one categorical variable. This generally works as well as the within-independence assumption is realistic, and we know that there are cases where it is not. However, the local independence assumption represents a strong limitation, since it could lead to a solution with too many clusters, as shown by [VM02]. By relaxing this assumption, a simpler solution with a lower number of groups can be obtained yielding a better classification. An even better classification can be reached by assuming different dependences in each group.

Underlying Variables mixture model

[Eve88] and [EM90] proposed a model according to which both the continuous and the categorical ordinal variables follow a homoscedastic Gaussian mixture model. However, as regards the ordinal variables, the mixture variables are only partially observed through their ordinal counterparts. In other words, the ordinal variables are modeled following the Underlying Response Variable

(UVR) approach. This satisfies the two main requirements: dealing with ordinal data properly and modeling dependences between ordinal and continuous variables. It is interesting to note that this model can be rewritten in terms of copulas [MBV17]. The main drawback of this model is that, in practice, it cannot be estimated through a full maximum likelihood approach, due to the presence of multidimensional integrals making the estimation time consuming. In sight of this, [Mor12] proposed a model-based clustering for mixed binary and continuous variables: each binary attribute is generated by dichotomizing a latent continuous variable, while the scores of the latent variables are estimated from the binary data. The estimated scores of the latent variables and the observed continuous data follow a multivariate Gaussian mixture model. Thus the estimation is carried out in two steps where the scores for binary data are firstly estimated before estimating the parameters of the mixture model. Eventually, [RR17] proposed a model with no local independence or conditionally independent blocks assumption where the dependences between variables can be easily measured by adopting the URV approach for the ordinal variables and assuming that each component of the mixture follows a multivariate normal distribution such that the corresponding covariance matrices capture all the dependences regardless the nature of variables.

Retained approach

In this work, the location mixture model approach is retained since it better models relations between continuous and categorical radar features. Indeed, a radar pattern is mostly designed by first choosing a pattern of modulation features (categorical variables) to achieve a specific goal and then choosing continuous features (continuous variables) that meet constraints related to the chosen pattern and the tactical environment. Hence, continuous radar features are mainly chosen conditionally to categorical radar features and the location mixture model naturally responds to that dependence structure by assuming that continuous variables are normally distributed conditionally to categorical variables. Moreover, the local independence assumption proposed by [HJ99] is not retained in order to take advantage of the dependence structure between continuous and categorical data to infer on missing data.

3.2.2 Assumptions on mixed data

In this subsection, a joint distribution for mixed data is introduced to model the dependence structure between continuous and categorical data. Then, outliers and missing values are tackled by taking advantage of the joint distribution.

Distribution of mixed data

Considering that the retained approach focuses on conditioning continuous data $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ according to categorical data $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$, the following joint distribution is introduced

$$\forall j \in \mathcal{J},\ p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}) = \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj} | \boldsymbol{\mu}_{\boldsymbol{c}}, \boldsymbol{\Sigma}\right)\right)^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}} \tag{3.5}$$

where continuous variables $\boldsymbol{x}_{qj}$ are normally distributed according to categorical variables $\boldsymbol{x}_{cj}$ with means $(\boldsymbol{\mu}_{\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}$ and variance $\boldsymbol{\Sigma}$. As for categorical variables $\boldsymbol{x}_{cj}$, they are jointly distributed according to the multivariate categorical distribution defined in (3.1) and parametrized by $\boldsymbol{\pi} = (\pi_{\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}$. Indeed, these conditional and marginal distributions can be obtained from (3.5) as

follows :

$$\forall j \in \mathcal{J}, \quad \begin{aligned} p(\boldsymbol{x}_{cj}) &= \int p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}) \partial \boldsymbol{x}_{qj} = \prod_{\boldsymbol{c} \in \mathcal{C}_q} \pi_{\boldsymbol{c}}^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} = \mathcal{MC}(\boldsymbol{x}_{cj}|\boldsymbol{\pi}) \ , \\ p(\boldsymbol{x}_{qj}|\boldsymbol{x}_{cj}) &= \frac{p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})}{p(\boldsymbol{x}_{cj})} = \prod_{\boldsymbol{c} \in \mathcal{C}_q} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{\boldsymbol{c}}, \boldsymbol{\Sigma}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \ . \end{aligned} \tag{3.6}$$

Outliers

As developed in Section 3.1, outliers are only considered for continuous data $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j \in \mathcal{J}}$ and they are handled as in subsection 3.1.1 by introducing scale latent variables $\boldsymbol{u} = (u_j)_{j \in \mathcal{J}}$. Nonetheless, the latent variables $\boldsymbol{u}$ are introduced conditionally to categorical data $\boldsymbol{x}_c$ due to the dependence structure established in (3.5) and (3.6) such that

$$\forall j \in \mathcal{J}, \quad \begin{aligned} \boldsymbol{x}_{qj}|u_j, \boldsymbol{x}_{cj} &\sim \prod_{\boldsymbol{c} \in \mathcal{C}_q} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \ , \\ u_j|\boldsymbol{x}_{cj} &\sim \prod_{\boldsymbol{c} \in \mathcal{C}_q} \mathcal{G}\left(u_j|\alpha_{\boldsymbol{c}}, \beta_{\boldsymbol{c}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \ , \end{aligned}$$

where each u_j follows conditionally to categorical data $\boldsymbol{x}_{cj}$ a Gamma distribution with rate and shape parameters $(\alpha_{\boldsymbol{c}}, \beta_{\boldsymbol{c}}) \in \mathbb{R}^{*+} \times \mathbb{R}^{*+}$.

Missing Data

Both quantitative and categorical data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j \in \mathcal{J}}$ can be partially observed. Hence $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j \in \mathcal{J}}$ are decomposed into observed features $(\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj}^{\text{obs}})_{j \in \mathcal{J}}$ and missing features $(\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{miss}})_{j \in \mathcal{J}}$ such that

$$\forall j \in \mathcal{J}, \quad \begin{aligned} \boldsymbol{x}_{qj} &= \begin{pmatrix} \boldsymbol{x}_{qj}^{\text{miss}} \\ \boldsymbol{x}_{qj}^{\text{obs}} \end{pmatrix} \text{ with } (\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{qj}^{\text{obs}}) \in \mathbb{R}^{d_j^{\text{miss}}} \times \mathbb{R}^{d_j^{\text{obs}}} \text{ and } d_j^{\text{miss}} + d_j^{\text{obs}} = d \ , \\ \boldsymbol{x}_{cj} &= \begin{pmatrix} \boldsymbol{x}_{cj}^{\text{miss}} \\ \boldsymbol{x}_{cj}^{\text{obs}} \end{pmatrix} \text{with } (\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) \in \mathcal{C}_{q_j^{\text{miss}}} \times \mathcal{C}_{q_j^{\text{obs}}} \text{ and } q_j^{\text{miss}} + q_j^{\text{obs}} = q \ . \end{aligned}$$

where $(\mathbb{R}^{d_j^{\text{miss}}}, \mathcal{C}_{q_j^{\text{miss}}})$ and $(\mathbb{R}^{d_j^{\text{obs}}}, \mathcal{C}_{q_j^{\text{obs}}})$, are disjoint subsets of $(\mathbb{R}^d, \mathcal{C}_q)$ embedding missing features $(\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{miss}})$ and observed features $(\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj}^{\text{obs}})$.

Missing continuous data $\boldsymbol{x}_q^{\text{miss}} = (\boldsymbol{x}_{qj}^{\text{miss}})_{j \in \mathcal{J}}$ are nearly handled as in subsection 3.1.1 by taking advantage of properties of the multivariate normal distribution to obtain a distribution for missing values. The only difference with the subsection 3.1.1 lies in the fact that continuous data are also distributed conditionally to categorical data $\boldsymbol{x}_c$ due to the dependence structure established in (3.5) and (3.6). Hence, the following distributions are obtained

$$\forall j \in \mathcal{J}, \quad \begin{aligned} \boldsymbol{x}_{qj}^{\text{miss}}|\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj} &\sim \prod_{\boldsymbol{c} \in \mathcal{C}} \mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{miss}}|\boldsymbol{\mu}_{j\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}}, \boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{miss}}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \ , \\ \boldsymbol{x}_{qj}^{\text{obs}}|\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj} &\sim \prod_{\boldsymbol{c} \in \mathcal{C}} \mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{obs}}|\boldsymbol{\mu}_{j\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{obs}}}, \boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{obs}}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \ , \end{aligned}$$

where $\forall j \in \mathcal{J},\ \forall \boldsymbol{c} \in \mathcal{C}_q$:

$$\begin{aligned}
\boldsymbol{\mu}_{j\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}} &= \boldsymbol{\mu}_{\boldsymbol{c}}^{\text{miss}} + \boldsymbol{\Sigma}^{\text{cov}}\boldsymbol{\Sigma}^{\text{obs}^{-1}}\left(\boldsymbol{x}_{qj}^{\text{obs}} - \boldsymbol{\mu}_{\boldsymbol{c}}^{\text{obs}}\right) , \\
\boldsymbol{\mu}_{j\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{obs}}} &= \boldsymbol{\mu}_{\boldsymbol{c}}^{\text{obs}} , \\
\boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{miss}}} &= \boldsymbol{\Sigma}^{\text{miss}} - \boldsymbol{\Sigma}^{\text{cov}}\boldsymbol{\Sigma}^{\text{obs}^{-1}}\boldsymbol{\Sigma}^{\text{cov'}} , \\
\boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{obs}}} &= \left(\boldsymbol{\Sigma}^{\text{obs}^{-1}} + 2 \times \boldsymbol{\Sigma}^{\text{obs}^{-1}}\boldsymbol{\Sigma}^{\text{cov'}}\left(\boldsymbol{\Sigma}^{\boldsymbol{x}_q^{\text{miss}}}\right)^{-1}\boldsymbol{\Sigma}^{\text{cov}}\boldsymbol{\Sigma}^{\text{obs}^{-1}}\right)^{-1} .
\end{aligned}$$

Regarding missing categorical data, they are handled as in subsection 3.1.2 such that missing features $\boldsymbol{x}_{cj}^{\text{miss}}$ follow a multivariate categorical distribution conditionally to observed features $\boldsymbol{x}_{cj}^{\text{obs}}$ where $\forall(\boldsymbol{c}^{\text{miss}}, \boldsymbol{c}^{\text{obs}}) \in \boldsymbol{C}_{q_j^{\text{miss}}} \times \boldsymbol{C}_{q_j^{\text{obs}}}$:

$$p(\boldsymbol{x}_{cj}^{\text{miss}} = \boldsymbol{c}^{\text{miss}} | \boldsymbol{x}_{cj}^{\text{obs}} = \boldsymbol{c}^{\text{obs}}) = \frac{\pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}}{\sum\limits_{\boldsymbol{c}^{\text{miss}} \in \boldsymbol{C}_{q_j^{\text{miss}}}} \pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}}$$

with $\pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}$ the joint probability $\pi_{\boldsymbol{c}}$ defined in (3.5) and (3.6) for $\boldsymbol{c} = (\boldsymbol{c}^{\text{miss}}, \boldsymbol{c}^{\text{obs}})$.

3.2.3 Proposed model

In this subsection, the retained approach is developed into a Bayesian framework where the proposed mixture model handles mixed-type data. Component distributions of clusters are first introduced before detailing the proposed model and its Bayesian framework.

Component Distribution

Assuming independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$ for continuous and categorical observations $\boldsymbol{x} = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ and according to assumptions on mixed data defined in subsection 3.2.2, a component distribution for each cluster $k \in \mathcal{K}$ is obtained as follows

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ p(\boldsymbol{x}_j | u_j, z_j = k) = p(\boldsymbol{x}_{qj} | u_j, \boldsymbol{x}_{cj}, z_j = k) p(\boldsymbol{x}_{cj} | z_j = k) ,$$

where

$$\begin{aligned}
p(\boldsymbol{x}_{qj} | u_j, \boldsymbol{x}_{cj}, z_j = k) &= \prod_{\boldsymbol{c}\in\mathcal{C}_q} \mathcal{N}\left(\boldsymbol{x}_{qj} | \boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} , \\
p(\boldsymbol{x}_{cj} | z_j = k) &= \prod_{\boldsymbol{c}\in\mathcal{C}_q} \pi_{k\boldsymbol{c}}^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} .
\end{aligned}$$

Finally, the complete component distribution for each cluster $k \in \mathcal{K}$ results in

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ p(\boldsymbol{x}_j, u_j | z_j = k) = \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj} | \boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right) \mathcal{G}(u_j | \alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} , \quad (3.7)$$

with

- $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$ the scale latent variables handling outliers for quantitative data $\boldsymbol{x}_q$ and distributed according to a Gamma distribution with shape and rate parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})_{(k,\boldsymbol{c})\in\mathcal{K}\times\mathcal{C}_q}$,
- $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = ((\boldsymbol{\mu}_{k\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}, \boldsymbol{\Sigma}_k)_{k\in\mathcal{K}}$ the mean and the variance parameters of quantitative data $\boldsymbol{x}_q$ for each cluster,
- $\boldsymbol{\pi} = (\boldsymbol{\pi}_k)_{k\in\mathcal{K}}$ the weights of the multivariate Categorical distribution of categorical data $\boldsymbol{x}_c$ for each cluster.

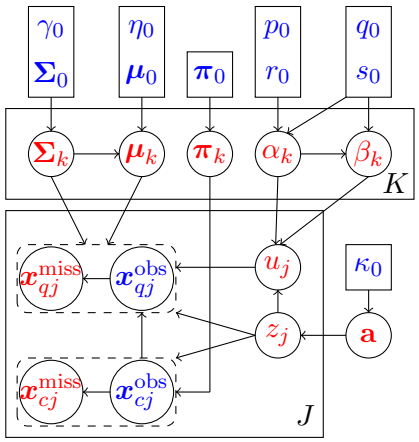


Figure 3.6: Graphical representation of the proposed model integrating mixed data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ decomposed into quantitative data $(\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ and categorical data $(\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$, the scale variables u_j and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red

Mixture model

Recalling that $p(z_j = k) = a_k$ where $\boldsymbol{a} = (a_k)_{k\in\mathcal{K}}$ are the weights related to component distributions, the mixture model is obtained from (3.7) such that $\forall j \in \mathcal{J}$,

$$p(\boldsymbol{x}_j, u_j|\boldsymbol{\Theta}) = \sum_{k\in\mathcal{K}} a_k \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{k\boldsymbol{c}}\mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{G}(u_j|\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \tag{3.8}$$

where $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the set of parameters.

Bayesian framework

As in previous chapters, a Bayesian framework is used to estimate parameters $\boldsymbol{\Theta}$. Assuming a dataset $\boldsymbol{x} = (\boldsymbol{x}_q, \boldsymbol{x}_c)$ of i.i.d observations $(\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$, independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$ and scale latent variables $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$, the complete likelihood associated to (3.8) is defined by

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\left(a_k \prod_{\boldsymbol{c}\in\mathcal{C}_q}\left(\pi_{k\boldsymbol{c}}\mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{G}(u_j|\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}}\right)^{\delta_{z_j}^k}$$

Eventually, the prior distribution required for $\boldsymbol{\Theta}$ is chosen as

$$p(\boldsymbol{\Theta}|\mathcal{K}) = p(\boldsymbol{a}|\mathcal{K})p(\boldsymbol{\pi}|\mathcal{K})p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K})p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K})$$

where

$$\begin{cases} p(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\boldsymbol{\kappa}_0)\ , \\ p(\boldsymbol{\pi}|\mathcal{K}) = \displaystyle\prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{\pi}_k|\boldsymbol{\pi}_0)\ , \\ p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K}) = \displaystyle\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}_q} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}, \eta_{0_{k\boldsymbol{c}}}^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{IW}(\boldsymbol{\Sigma}_k|\gamma_0, \boldsymbol{\Sigma}_0)\ , \\ p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \displaystyle\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}_q} p(\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}}|p_0, q_0, s_0, r_0)\ . \end{cases}$$

and $p(\cdot,\cdot|p,q,s,r)$ is the prior distribution defined in the previous chapter such that

$$\forall(\alpha,\beta)\in\mathbb{R}^{*+}\times\mathbb{R}^{*+},\ p(\alpha,\beta|p,q,s,r)=\frac{1}{M}\frac{p^{\alpha-1}e^{-q\beta}\beta^{s\alpha}}{\Gamma(\alpha)^r} \tag{3.9}$$

where $p,q,s,r>0$ and $M=\int\frac{p^{\alpha-1}e^{-q\beta}\beta^{s\alpha}}{\Gamma(\alpha)^r}\mathbb{I}_{\alpha>0}\mathbb{I}_{\beta>0}\partial\beta\partial\alpha$. Graphical representation of the proposed model is shown in Figure 3.6.

3.3 Inference

Direct inference on the proposed model is not trivial since distributions of latent missing data and parameters may not be defined when both continuous and categorical features are missing. To overcome that issue, latent data and parameters are assumed to be independent a posteriori and their posterior distributions can be defined while keeping dependencies between parameters of these distributions. Therefore, the Variational Bayes (VB) procedure is processed to estimate parameters of the mixture model defined in (3.8). Variational posterior distributions are obtained from the VB Expectation (VBE) and VB Maximization (VBM) steps and a Lower Bound on the log evidence is defined to master the convergence of the VB procedure.

3.3.1 Variational posterior distributions

As previously, a factorized posterior distribution $q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\boldsymbol{\Theta}|\mathcal{K})=q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z}|\mathcal{K})q(\boldsymbol{\Theta}|\mathcal{K})$ is chosen as an approximation of the intractable posterior joint distribution $p(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{x}_q^{\text{obs}},\boldsymbol{x}_c^{\text{obs}},\mathcal{K})$ such that latent variables $\boldsymbol{h}=(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z})$ and parameters $\boldsymbol{\Theta}$ are a posteriori independent and

$$\begin{aligned}
q(\boldsymbol{h}|\mathcal{K})&=q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z},\mathcal{K})q(\boldsymbol{z}|\mathcal{K})\ ,\\
q(\boldsymbol{\Theta}|\mathcal{K})&=q(\boldsymbol{a}|\mathcal{K})q(\boldsymbol{\pi}|\mathcal{K})q(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{K})q(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})\ .
\end{aligned}$$

According to VB assumptions, the following conjugate variational posterior distributions are obtained from the VB procedure

$$\begin{cases}
q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})=\prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}}\mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{miss}}|\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}},u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k}\ ,\\
q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})=\prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}}\mathcal{G}\left(u_j|\tilde{\alpha}_{jk\boldsymbol{c}},\tilde{\beta}_{jk\boldsymbol{c}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k}\ ,\\
q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z},\mathcal{K})=\prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\mathcal{MC}\left(\boldsymbol{x}_{cj}^{\text{miss}}|\tilde{\boldsymbol{r}}_{jk}^{\boldsymbol{x}_c^{\text{miss}}}\right)^{\delta_{z_j}^k}\\
q(\boldsymbol{z}|\mathcal{K})=\prod_{j\in\mathcal{J}}\mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j)\ ,\\
q(\boldsymbol{a}|\mathcal{K})=\mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}})\ ,\\
q(\boldsymbol{\pi}|\mathcal{K})=\prod_{k\in\mathcal{K}}\mathcal{D}(\boldsymbol{\pi}|\tilde{\boldsymbol{\pi}}_k)\ ,\\
q(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})=\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}_q}\mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}},\tilde{\eta}_{k\boldsymbol{c}}^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k,\tilde{\boldsymbol{\Sigma}}_k)\ ,\\
q(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{K})=\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}_q}p(\alpha_{k\boldsymbol{c}},\beta_{k\boldsymbol{c}}|\tilde{p}_k,\tilde{q}_k,\tilde{s}_k,\tilde{r}_k)\ ,
\end{cases} \tag{3.10}$$

where the variational posterior distributions of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are defined in (3.9). Their respective parameters are estimated during the VBE and VBM-steps by developing expectations $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right]$ and $\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right]$.

3.3.2 VBE-step

The VBE-step consists in deriving the following expectation

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^{k}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \sum_{c\in\mathcal{C}_q}\delta_{x_{cj}}^{\boldsymbol{c}}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log \pi_{k\boldsymbol{c}}\right] - \frac{1}{2}\Bigg(d(\log 2\pi - \log u_j) \\ + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] + u_j\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{x}_{qj} - \boldsymbol{\mu}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{qj} - \boldsymbol{\mu}_{k\boldsymbol{c}})\right]\Bigg) + \mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right] \\ + \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right] - 1\right)\log u_j - \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_{k\boldsymbol{c}})\right] - u_j\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_{k\boldsymbol{c}}\right]\Bigg)\Bigg)\end{aligned} \tag{3.11}$$

where $\forall(j, k, \boldsymbol{c}) \in \mathcal{J}\times\mathcal{K}\times\mathcal{C}_q$:

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{x}_{qj} - \boldsymbol{\mu}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{qj} - \boldsymbol{\mu}_{k\boldsymbol{c}})\right] = (\boldsymbol{x}_{qj} - \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})^T\,\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{x}_{qj} - \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}) + \frac{d}{\tilde{\eta}_{k\boldsymbol{c}}} \tag{3.12}$$

is obtained from properties of the variational distribution $q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K})$ in (3.10). Hence quantitative data $\boldsymbol{x}_q$ are distributed a posteriori according to a product of normal distributions conditionally to categorical data $\boldsymbol{x}_c$, latent variables $\boldsymbol{u}$ and labels $\boldsymbol{z}$ such that

$$\boldsymbol{x}_q|\boldsymbol{u}, \boldsymbol{x}_c, \boldsymbol{z} \sim \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}}\mathcal{N}\left(\boldsymbol{x}_{qj}|\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}, u_j^{-1}\tilde{\gamma}_k^{-1}\tilde{\boldsymbol{\Sigma}}_k\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^{k}}$$

Mean parameters $(\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})_{(k,\boldsymbol{c})\in\mathcal{K}\times\mathcal{C}_q}$ and variance parameters $(\tilde{\gamma}_k^{-1}\tilde{\boldsymbol{\Sigma}}_k)_{k\in\mathcal{K}}$ of these normal distributions are obtained from (3.12). By decomposing $\boldsymbol{x}_q$ into $(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{x}_q^{\text{obs}})$ and by exploiting properties of the multivariate normal distribution, the following variational posterior distribution is obtained for missing values $\boldsymbol{x}_q^{\text{miss}}$:

$$q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}}\mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{miss}}|\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}}, u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^{k}}$$

with $\forall(j, k, \boldsymbol{c}) \in \mathcal{J}\times\mathcal{K}\times\mathcal{C}_q$:

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}} &= \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^{\text{miss}} + \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}}\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}}(\boldsymbol{x}_{qj}^{\text{obs}} - \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^{\text{obs}})\,, \\ \tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}} &= \frac{\tilde{\boldsymbol{\Sigma}}_k^{\text{miss}} - \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}}\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}}\tilde{\boldsymbol{\Sigma}}_k^{\text{cov}'}}{\tilde{\gamma}_k}\,.\end{aligned}$$

Then by marginalising over $\boldsymbol{x}_q^{\text{miss}}$ in (3.11), the expectation (3.11) becomes

$$\begin{aligned}\int \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{\Theta},\mathcal{K})\right]\partial\boldsymbol{x}_q^{\text{miss}} = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\delta_{z_j}^k\Bigg(&\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \sum_{c\in\mathcal{C}_q}\delta_{x_{cj}}^{\boldsymbol{c}}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\pi_{k\boldsymbol{c}}\right]\\ &-\frac{1}{2}\Bigg(d_j^{\text{obs}}(\log 2\pi-\log u_j)+\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right]-\log|\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}|\\ &+u_j\left(\left(\boldsymbol{x}_{qj}^{\text{obs}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^{\text{obs}}\right)^T\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_{qj}^{\text{obs}}-1}\left(\boldsymbol{x}_{qj}^{\text{obs}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^{\text{obs}}\right)+\frac{d}{\tilde{\eta}_{k\boldsymbol{c}}}\right)\right)\\ &+\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right]+\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]-1\right)\log u_j\\ &-\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_{k\boldsymbol{c}})\right]-u_j\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_{k\boldsymbol{c}}\right]\Bigg)\Bigg)\end{aligned} \tag{3.13}$$

with $\forall k\in\mathcal{K}$,

$$\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_{qj}^{\text{obs}}} = \frac{\left(\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}}+2\times\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}}\tilde{\boldsymbol{\Sigma}}_k^{\text{cov}'}\left(\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}\right)^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\text{cov}}\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}}\right)^{-1}}{\tilde{\gamma}_k}\,.$$

Conditionally to $\boldsymbol{x}_c$ and $\boldsymbol{z}$, the scale latent variables $\boldsymbol{u}$ are distributed according to a product of Gamma distribution whose parameters are obtained by aggregating terms related to $\boldsymbol{u}$ such that

$$q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}}\mathcal{G}\left(u_j|\tilde{\alpha}_{jk\boldsymbol{c}},\tilde{\beta}_{jk\boldsymbol{c}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k}$$

with $\forall(j,k,\boldsymbol{c})\in\mathcal{J}\times\mathcal{K}\times\mathcal{C}_q$:

$$\begin{aligned}\tilde{\alpha}_{jk\boldsymbol{c}} &= \mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]+\frac{d_j^{\text{obs}}}{2}\,,\\ \tilde{\beta}_{jk\boldsymbol{c}} &= \mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_{k\boldsymbol{c}}\right]+\frac{1}{2}\left(\left(\boldsymbol{x}_{qj}^{\text{obs}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^{\text{obs}}\right)^T\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_{qj}^{\text{obs}}-1}\left(\boldsymbol{x}_{qj}^{\text{obs}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^{\text{obs}}\right)+\frac{d}{\tilde{\eta}_{k\boldsymbol{c}}}\right)\,.\end{aligned}$$

Then by marginalising over $\boldsymbol{u}$ in (3.13), the expectation (3.13) becomes

$$\begin{aligned}\int \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{\Theta},\mathcal{K})\right]\partial\boldsymbol{x}_q^{\text{miss}}\partial\boldsymbol{u} &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\delta_{z_j}^k\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right]+\sum_{c\in\mathcal{C}_q}\delta_{x_{cj}}^{\boldsymbol{c}}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\pi_{k\boldsymbol{c}}\right]\\ &\quad-\frac{1}{2}\left(d_j^{\text{obs}}\log 2\pi+\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right]-\log|\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}|\right)\\ &\quad+\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right]-\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_{k\boldsymbol{c}})\right]\\ &\quad+\log\Gamma(\tilde{\alpha}_{jk\boldsymbol{c}})-\tilde{\alpha}_{jk\boldsymbol{c}}\log\tilde{\beta}_{jk\boldsymbol{c}}\Bigg)\Bigg)\\ &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\delta_{z_j}^k\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right]+\sum_{c\in\mathcal{C}_q}\delta_{x_{cj}}^{\boldsymbol{c}}\log\rho_{k\boldsymbol{c}}^{\boldsymbol{x}_{cj}}\right)\end{aligned} \tag{3.14}$$

where $\forall(j,k,\boldsymbol{c})\in\mathcal{J}\times\mathcal{K}\times\mathcal{C}_q$,

$$\begin{aligned}\log\rho_{k\boldsymbol{c}}^{\boldsymbol{x}_{cj}} &= \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\pi_{k\boldsymbol{c}}\right]-\frac{1}{2}\left(d_j^{\text{obs}}\log 2\pi+\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right]-\log|\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}|\right)+\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right]\\ &\quad-\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_{k\boldsymbol{c}})\right]+\log\Gamma(\tilde{\alpha}_{jk\boldsymbol{c}})-\tilde{\alpha}_{jk\boldsymbol{c}}\log\tilde{\beta}_{jk\boldsymbol{c}}\,.\end{aligned} \tag{3.15}$$

By decomposing each $\boldsymbol{x}_{cj} \in \mathcal{C}_q$ into $(\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) \in \mathcal{C}_{q_j^{\text{miss}}} \times \mathcal{C}_{q_j^{\text{obs}}}$ and $\boldsymbol{c} \in \mathcal{C}_q$ into $(\boldsymbol{c}^{\text{miss}}, \boldsymbol{c}^{\text{obs}}) \in \mathcal{C}_{q_j^{\text{miss}}} \times \mathcal{C}_{q_j^{\text{obs}}}$, the marginalised expectation (3.14) can be developed as

$$\begin{aligned}\int \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right] \partial \boldsymbol{x}_q^{\text{miss}} \partial \boldsymbol{u} &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] \\ &\quad + \sum_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \delta_{x_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}} \sum_{\boldsymbol{c}^{\text{obs}}\in\mathcal{C}_{q_j^{\text{obs}}}} \delta_{x_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}} \log \rho_{k\boldsymbol{c}^{\text{miss}}\boldsymbol{c}^{\text{obs}}}^{\boldsymbol{x}_{cj}} \Bigg) \\ &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \sum_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \delta_{x_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}} \log \rho_{k\boldsymbol{c}^{miss}}^{\boldsymbol{x}_{cj}^{\text{miss}}} \Bigg)\end{aligned} \tag{3.16}$$

where $\forall (j, k, \boldsymbol{c}^{\text{miss}}) \in \mathcal{J} \times \mathcal{K} \times \mathcal{C}_{q_j^{\text{miss}}}$:

$$\log \rho_{k\boldsymbol{c}^{miss}}^{\boldsymbol{x}_{cj}^{\text{miss}}} = \sum_{\boldsymbol{c}^{\text{obs}}\in\mathcal{C}_{q_j^{\text{obs}}}} \delta_{x_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}} \log \rho_{k\boldsymbol{c}^{\text{miss}}\boldsymbol{c}^{\text{obs}}}^{\boldsymbol{x}_{cj}} \tag{3.17}$$

Hence, a multivariate categorical distribution is deduced for $\boldsymbol{x}_c^{\text{miss}}$ from (3.16) conditionally to labels $\boldsymbol{z}$ such that

$$q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{MC}\left(\boldsymbol{x}_{cj}^{\text{miss}}|\tilde{\boldsymbol{r}}_{jk}^{\boldsymbol{x}_c^{\text{miss}}}\right)^{\delta_{z_j}^k}$$

and their parameters $(\tilde{\boldsymbol{r}}_{jk}^{\boldsymbol{x}_c^{\text{miss}}})_{(j,k)\in\mathcal{J}\times\mathcal{K}}$ are obtained from (3.17) where $\forall (j, k, \boldsymbol{c}) \in \mathcal{J} \times \mathcal{K} \times \mathcal{C}_{q_j^{\text{miss}}}$

$$\tilde{r}_{jk\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_c^{\text{miss}}} = \frac{\rho_{k\boldsymbol{c}^{miss}}^{\boldsymbol{x}_{cj}^{\text{miss}}}}{\sum\limits_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \rho_{k\boldsymbol{c}^{miss}}^{\boldsymbol{x}_{cj}^{\text{miss}}}} \ .$$

Finally, variational posterior categorical distributions are obtained for labels $\boldsymbol{z}$ by marginalising over $\boldsymbol{x}_c^{\text{miss}}$ in (3.16) such that

$$\begin{aligned}\int \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right] \partial \boldsymbol{x}_q^{\text{miss}} \partial \boldsymbol{u} \partial \boldsymbol{x}_c^{\text{miss}} &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log \sum_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \rho_{k\boldsymbol{c}^{miss}}^{\boldsymbol{x}_{cj}^{\text{miss}}} \Bigg) \\ &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \log \rho_{jk}\end{aligned} \tag{3.18}$$

where $\forall j \in \mathcal{J},\ k \in \mathcal{K}$,

$$\log \rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log \sum_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \rho_{k\boldsymbol{c}^{miss}}^{\boldsymbol{x}_{cj}^{\text{miss}}} \ . \tag{3.19}$$

Hence the variational categorical distributions are deduced from (3.18) and are given by

$$q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j)$$

where probabilities $(\tilde{\boldsymbol{r}}_j)_{j\in\mathcal{J}}$ are obtained from (3.19) such that $\forall j \in \mathcal{J},\ k \in \mathcal{K}$,

$$\tilde{r}_{jk} = \frac{\rho_{jk}}{\sum\limits_{k\in\mathcal{K}} \rho_{jk}} \ .$$

3.3.3 VBM-step

The VBM-step consists in deriving the following expectation

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z},\boldsymbol{\Theta}|\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{\Theta},\mathcal{K})\right] + p(\boldsymbol{\Theta}|\mathcal{K}) \\
&= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\Bigg(\log a_k + \sum_{c\in\mathcal{C}_q}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{x_{cj}}^c\right]\Bigg(\log\pi_{k\boldsymbol{c}} - \frac{1}{2}\Bigg(\log|\boldsymbol{\Sigma}_k| \\
&\quad + d(\log 2\pi - \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right]) + \mathbb{E}_{\boldsymbol{h}}\left[u_j(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})\right]\Bigg) \\
&\quad + \alpha_{k\boldsymbol{c}}\log\beta_{k\boldsymbol{c}} + (\alpha_{k\boldsymbol{c}}-1)\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] - \log\Gamma(\alpha_{k\boldsymbol{c}}) - \mathbb{E}_{\boldsymbol{h}}\left[u_j\right]\beta_{k\boldsymbol{c}}\Bigg)\Bigg) \\
&\quad + \sum_{k\in\mathcal{K}}(\kappa_{0_k}-1)\log a_k + \log c_{\mathcal{D}}(\boldsymbol{\kappa}_0) - \frac{1}{2}\Bigg((\gamma_0+d+1)\log|\boldsymbol{\Sigma}_k| \\
&\quad + \mathrm{Trace}\left(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_k^{-1}\right)\Bigg) + c_{\mathcal{IW}}(\gamma_0,\boldsymbol{\Sigma}_0) + \sum_{\boldsymbol{c}\in\mathcal{C}_q}\frac{1}{2}\Bigg(d(\log\eta_{0_{k\boldsymbol{c}}} - \log 2\pi) \\
&\quad - \log|\boldsymbol{\Sigma}_k| - \eta_{0_{k\boldsymbol{c}}}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right)\Bigg) - \log M_0 \\
&\quad + (\alpha_{k\boldsymbol{c}}-1)\log p_0 - r_0\log\Gamma(\alpha_{k\boldsymbol{c}}) + s_0\alpha_{k\boldsymbol{c}}\log\beta_{k\boldsymbol{c}} - q_0\beta_{k\boldsymbol{c}} \\
&\quad + (\pi_{0_{k\boldsymbol{c}}}-1)\log\pi_{k\boldsymbol{c}} + \log c_{\mathcal{D}}(\boldsymbol{\pi}_{0_k})\ ,
\end{aligned} \tag{3.20}$$

where $\forall(j,k,\boldsymbol{c}) \in \mathcal{J}\times\mathcal{K}\times\mathcal{C}_q$:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[u_j(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[u_j\right](\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right]-\boldsymbol{\mu}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right]-\boldsymbol{\mu}_{k\boldsymbol{c}}) \\
&\quad + \mathrm{Trace}\left(\mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right]\boldsymbol{\Sigma}_k^{-1}\right)
\end{aligned} \tag{3.21}$$

is obtained from properties of the variational distribution $q(\boldsymbol{h}|\mathcal{K})$ with

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[u_j\right] &= \frac{\tilde{\alpha}_{jk\boldsymbol{c}}}{\tilde{\beta}_{jk\boldsymbol{c}}}\ , \\
\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right] &= \begin{pmatrix}\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_q^{\mathrm{miss}}} \\ \boldsymbol{x}_{qj}^{\mathrm{obs}}\end{pmatrix}\ , \\
\mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right] &= \begin{pmatrix}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\mathrm{miss}}} & \mathbf{0}^{d_j^{\mathrm{miss}}\times d_j^{\mathrm{obs}}} \\ \mathbf{0}^{d_j^{\mathrm{obs}}\times d_j^{\mathrm{miss}}} & \mathbf{0}^{d_j^{\mathrm{obs}}\times d_j^{\mathrm{obs}}}\end{pmatrix}\ .
\end{aligned}$$

By factorizing terms related to $\boldsymbol{a}$ in (3.20), the following Dirichlet distribution is obtained

$$q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}})$$

where

$$\forall k\in\mathcal{K},\ \tilde{\kappa}_k = \kappa_{0_k} + \sum_{j\in\mathcal{J}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\ .$$

Like the variational distribution of $\boldsymbol{a}$, variational posterior distributions of $\boldsymbol{\pi}$ are obtained by factorizing terms related to $\boldsymbol{\pi}$ in (3.20) and are given by

$$q(\boldsymbol{\pi}|\mathcal{K}) = \prod_{k\in\mathcal{K}}\mathcal{D}(\boldsymbol{\pi}_k|\tilde{\boldsymbol{\pi}}_k)$$

where

$$\forall (k, \boldsymbol{c}) \in \mathcal{K} \times \mathcal{C}_q,\ \tilde{\pi}_{k\boldsymbol{c}} = \pi_{0_{k\boldsymbol{c}}} + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \ .$$

with $\mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right]$ is obtained by decomposing categorical features into observed and missing features such that

$$\mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] = \delta_{\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}} \times \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}}\right] \ .$$

Then, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are a posteriori distributed according to the distribution defined in (3.9) such that

$$q(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \prod_{k \in \mathcal{K}} \prod_{\boldsymbol{c} \in \mathcal{C}_q} p(\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}}|\tilde{p}_{k\boldsymbol{c}}, \tilde{q}_{k\boldsymbol{c}}, \tilde{s}_{k\boldsymbol{c}}, \tilde{r}_{k\boldsymbol{c}}) \ ,$$

where

$$\begin{aligned}
\tilde{p}_{k\boldsymbol{c}} &= p_0 \exp\left(\sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right]\right) \ , \\
\tilde{q}_{k\boldsymbol{c}} &= q_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \ , \\
\tilde{s}_{k\boldsymbol{c}} &= s_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \ , \\
\tilde{r}_{k\boldsymbol{c}} &= r_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \ .
\end{aligned}$$

By aggregating and factorizing terms related to each $\boldsymbol{\mu}_{k\boldsymbol{c}}$ in (3.20), a Normal distribution is obtained for each $\boldsymbol{\mu}_{k\boldsymbol{c}}$ such that

$$q(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathcal{K}) = \prod_{k \in \mathcal{K}} \prod_{\boldsymbol{c} \in \mathcal{C}_q} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}, \tilde{\eta}_{k\boldsymbol{c}}^{-1}\boldsymbol{\Sigma}_k\right)$$

where $\forall k \in \mathcal{K}$ and $\forall \boldsymbol{c} \in \mathcal{C}_q$

$$\begin{aligned}
\tilde{\eta}_{k\boldsymbol{c}} &= \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] + \eta_{0_{k\boldsymbol{c}}} \ , \\
\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}} &= \frac{\sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right] + \eta_{0_{k\boldsymbol{c}}}\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}}{\tilde{\eta}_{k\boldsymbol{c}}} \ .
\end{aligned}$$

Eventually, variance parameters $\boldsymbol{\Sigma}$ are a posteriori distributed according to Inverse Wishart distributions given by

$$q(\boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k \in \mathcal{K}} \mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k, \tilde{\boldsymbol{\Sigma}}_k)$$

where

$$\begin{aligned}
\tilde{\gamma}_k &= \gamma_0 + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \ , \\
\tilde{\boldsymbol{\Sigma}}_k &= \boldsymbol{\Sigma}_0 + \sum_{j \in \mathcal{J}} \sum_{\boldsymbol{c} \in \mathcal{C}_q} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right] \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] \mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right] \mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right]^T + \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right] \\
&\quad + \sum_{\boldsymbol{c} \in \mathcal{C}_q} \eta_{0_{k\boldsymbol{c}}}\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}^T - \tilde{\eta}_{k\boldsymbol{c}}\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^T \ .
\end{aligned}$$

3.3.4 Lower Bound

Recalling that the Lower Bound on the log evidence is given by

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right] \tag{3.22}$$

where $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right]$ is the free energy and $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right]$ is the entropy of the approximate posterior $q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})$. The free energy can be developed as

$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{h}|\boldsymbol{\Theta},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right]$$

where

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{h}|\boldsymbol{\Theta},\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^{k}\right]\Bigg(&\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \sum_{\boldsymbol{c}\in\mathcal{C}_q}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right]\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log \pi_{k\boldsymbol{c}}\right]\\ &- \frac{1}{2}\Bigg(d(\log 2\pi - \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right]) + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right]\\ &+ \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[u_j(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})\right]\Bigg) + \mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right]\\ &+ (\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]-1)\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] - \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_{k\boldsymbol{c}})\right] - \mathbb{E}_{\boldsymbol{h}}\left[u_j\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_{k\boldsymbol{c}}\right]\Bigg)\Bigg)\end{aligned}$$

with

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[u_j(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{qj}-\boldsymbol{\mu}_{k\boldsymbol{c}})\right] = \mathbb{E}_{\boldsymbol{h}}\left[u_j\right]\Bigg(&(\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right]-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})^T\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}(\mathbb{E}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right]-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})\\ &\frac{d}{\tilde{\eta}_{k\boldsymbol{c}}}\Bigg) + \mathrm{Trace}\left(\mathbb{V}_{\boldsymbol{h}}\left[\boldsymbol{x}_{qj}\right]\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}\right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right] = &\sum_{k\in\mathcal{K}}(\kappa_{0_k}-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log c_{\mathcal{D}}(\boldsymbol{\kappa}_0) - \frac{1}{2}\Bigg((\gamma_0+d+1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right]\\ &+ \mathrm{Trace}\left(\boldsymbol{\Sigma}_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Sigma}_k^{-1}\right]\right)\Bigg) + c_{\mathcal{IW}}(\gamma_0,\boldsymbol{\Sigma}_0) + \sum_{\boldsymbol{c}\in\mathcal{C}_q}\frac{1}{2}\Bigg(d(\log\eta_{0_{k\boldsymbol{c}}}-\log 2\pi)\\ &- \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] - \eta_{0_{k\boldsymbol{c}}}\mathbb{E}_{\boldsymbol{\Theta}}\left[\left(\boldsymbol{\mu}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right)\right]\Bigg) - \log M_0\\ &+ (\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]-1)\log p_0 - r_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_{k\boldsymbol{c}})\right] + s_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right]\\ &- q_0\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_{k\boldsymbol{c}}\right] + (\pi_{0_{k\boldsymbol{c}}}-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\pi_{k\boldsymbol{c}}\right] + \log c_{\mathcal{D}}(\boldsymbol{\pi}_0)\ .\end{aligned}$$

with

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\left(\boldsymbol{\mu}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right)\right] = \left(\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right)^T\tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1}\left(\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}-\boldsymbol{\mu}_{0_{k\boldsymbol{c}}}\right) + \frac{d}{\tilde{\eta}_{k\boldsymbol{c}}}\ .$$

As for the entropy term, the following decompositon is obtained

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_q^{\mathrm{miss}},\boldsymbol{u},\boldsymbol{x}_c^{\mathrm{miss}},\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\\ &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_q^{\mathrm{miss}}|\boldsymbol{u},\boldsymbol{x}_c^{\mathrm{miss}},\boldsymbol{z},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{u}|\boldsymbol{x}_c^{\mathrm{miss}},\boldsymbol{z},\mathcal{K})\right]\\ &+ \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_c^{\mathrm{miss}}|\boldsymbol{z},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\end{aligned}$$

where

$$\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{\boldsymbol{c}\in\mathcal{C}_q}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right]\frac{1}{2}\Bigg(d_j^{\text{miss}}(\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] - \log 2\pi - 1) - \log|\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_{qj}}|\Bigg),$$

$$\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{\boldsymbol{c}\in\mathcal{C}_q}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\right]\Big(\tilde{\alpha}_{jk\boldsymbol{c}}\log\tilde{\beta}_{jk\boldsymbol{c}} - \log\Gamma(\tilde{\alpha}_{jk\boldsymbol{c}}) + (\tilde{\alpha}_{jk\boldsymbol{c}}-1)\mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] - \tilde{\beta}_{jk\boldsymbol{c}}\mathbb{E}_{\boldsymbol{h}}\left[u_j\right]\Big),$$

$$\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z},\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}}\right]\log\tilde{r}_{jk\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_c^{\text{miss}}}$$

$$\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\tilde{r}_{jk}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right] = {}& \sum_{k\in\mathcal{K}}(\tilde{\kappa}_k-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log c_{\mathcal{D}}(\tilde{\boldsymbol{\kappa}}_k) - \frac{1}{2}\Bigg((\tilde{\gamma}_k+d+1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] \\ & + \text{Trace}\left(\tilde{\boldsymbol{\Sigma}}_k\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Sigma}_k^{-1}\right]\right)\Bigg) + c_{\mathcal{IW}}(\tilde{\gamma}_k,\tilde{\boldsymbol{\Sigma}}_k) + \sum_{\boldsymbol{c}\in\mathcal{C}_q}\frac{1}{2}\Bigg(d(\log\tilde{\eta}_{k\boldsymbol{c}} - \log 2\pi) \\ & - \mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] - \tilde{\eta}_{k\boldsymbol{c}}\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{\mu}_{k\boldsymbol{c}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_{k\boldsymbol{c}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})\right]\Bigg) - \log M_k \\ & + (\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]-1)\log\tilde{p}_{k\boldsymbol{c}} - \tilde{r}_{k\boldsymbol{c}}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\Gamma(\alpha_{k\boldsymbol{c}})\right] + \tilde{s}_{k\boldsymbol{c}}\mathbb{E}_{\boldsymbol{\Theta}}\left[\alpha_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right] \\ & - \tilde{q}_{k\boldsymbol{c}}\mathbb{E}_{\boldsymbol{\Theta}}\left[\beta_{k\boldsymbol{c}}\right] + (\tilde{\pi}_{k\boldsymbol{c}}-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\pi_{k\boldsymbol{c}}\right] + \log c_{\mathcal{D}}(\tilde{\boldsymbol{\pi}}_k),\end{aligned}$$

with

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[(\boldsymbol{\mu}_{k\boldsymbol{c}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_{k\boldsymbol{c}}-\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}})\right] = \frac{d}{\tilde{\eta}_{k\boldsymbol{c}}}.$$

3.3.5 Expectations from variational distributions

Expectations developed in variational calculations are derived from properties of variational posterior distributions and are obtained as follows. Categorical distribution properties lead to

$$\forall j\in\mathcal{J},\ \forall k\in\mathcal{K},\ \forall\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}:$$

$$\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] = \tilde{r}_{jk},$$

$$\mathbb{E}_{\boldsymbol{h}}\left[\delta_{\boldsymbol{x}_{cj}^{\text{miss}}}^{\boldsymbol{c}^{\text{miss}}}\right] = \tilde{r}_{jk\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_c^{\text{miss}}}.$$

Dirichlet distribution properties lead to

$$\forall k\in\mathcal{K},\ \forall\boldsymbol{c}\in\mathcal{C}_q:$$

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k\in\mathcal{K}}\tilde{\kappa}_k\right),$$

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\pi_{k\boldsymbol{c}}\right] = \psi(\tilde{\pi}_{k\boldsymbol{c}}) - \psi\left(\sum_{\boldsymbol{c}\in\mathcal{C}_q}\tilde{\pi}_{k\boldsymbol{c}}\right),$$

where $\psi(\cdot)$ is the digamma function. Gamma distribution properties lead to

$$\begin{aligned}
&\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \forall \boldsymbol{c} \in \mathcal{C}_q : \\
&\qquad \mathbb{E}_{\boldsymbol{h}}\left[u_j\right] = \frac{\tilde{\alpha}_{jk\boldsymbol{c}}}{\tilde{\beta}_{jk\boldsymbol{c}}} , \\
&\qquad \mathbb{E}_{\boldsymbol{h}}\left[\log u_j\right] = \psi(\tilde{\alpha}_{jk\boldsymbol{c}}) - \log \tilde{\beta}_{jk\boldsymbol{c}} .
\end{aligned}$$

Normal distribution properties lead to

$$\begin{aligned}
&\forall k \in \mathcal{K},\ \forall \boldsymbol{c} \in \mathcal{C}_q : \\
&\qquad \mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_{k\boldsymbol{c}}\right] = \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}} , \\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_{k\boldsymbol{c}}\boldsymbol{\mu}_{k\boldsymbol{c}}^T\right] = \mathbb{V}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_{k\boldsymbol{c}}\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_{k\boldsymbol{c}}\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\mu}_{k\boldsymbol{c}}\right]^T \\
&\qquad\qquad = \tilde{\eta}_{k\boldsymbol{c}}^{-1}\boldsymbol{\Sigma}_k + \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}^T ,
\end{aligned}$$

Inverse Wishart distribution properties lead to

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Sigma}_k^{-1}\right] = \tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1} , \\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\log|\boldsymbol{\Sigma}_k|\right] = \log|\tilde{\boldsymbol{\Sigma}}_k| - \sum_{i=1}^{d}\psi\left(\frac{\tilde{\gamma}_k + 1 - i}{2}\right) - d\log 2 .
\end{aligned}$$

Posterior expectations of $\boldsymbol{\beta}$ can easily be computed conditionally to $\boldsymbol{\alpha}$ such that $\forall(k, \boldsymbol{c}) \in \mathcal{K} \times \mathcal{C}_q$:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}}[\beta_{k\boldsymbol{c}}] = \frac{\tilde{s}_{k\boldsymbol{c}}\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_{k\boldsymbol{c}}] + 1}{\tilde{q}_{k\boldsymbol{c}}} , \\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\beta_{k\boldsymbol{c}}\right] = \mathbb{E}_{\boldsymbol{\Theta}}[\psi\left(\tilde{s}_{k\boldsymbol{c}}\alpha_{k\boldsymbol{c}} + 1\right)] - \log\tilde{q}_{k\boldsymbol{c}} .
\end{aligned}$$

However, expectations depending on $\alpha_{k\boldsymbol{c}}$ are intractable

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}}[\psi\left(\tilde{s}_{k\boldsymbol{c}}\alpha_{k\boldsymbol{c}} + 1\right)] = \int \psi\left(\tilde{s}_{k\boldsymbol{c}}\alpha_{k\boldsymbol{c}} + 1\right) p(\alpha_{k\boldsymbol{c}}|\tilde{p}_{k\boldsymbol{c}}, \tilde{r}_{k\boldsymbol{c}})d\alpha_{k\boldsymbol{c}} , \\
&\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_{k\boldsymbol{c}}] = \int \alpha_{k\boldsymbol{c}} p(\alpha_{k\boldsymbol{c}}|\tilde{p}_{k\boldsymbol{c}}, \tilde{r}_{k\boldsymbol{c}})d\alpha_{k\boldsymbol{c}} , \\
&\mathbb{E}_{\boldsymbol{\Theta}}[\log\Gamma(\alpha_{k\boldsymbol{c}})] = \int \log\Gamma(\alpha_{k\boldsymbol{c}}) p(\alpha_{k\boldsymbol{c}}|\tilde{p}_{k\boldsymbol{c}}, \tilde{r}_{k\boldsymbol{c}})d\alpha_{k\boldsymbol{c}} .
\end{aligned}$$

As in previous chapter, the deterministic method introduced by [TK86] is applied to estimate these expectations.

3.4 Experiments

In this section, the proposed method is performed on 3 sets of realistic simulated data which are composed of continuous, categorical and mixed data. For comparison, a standard neural network (NN), the k-nearest neighbours (KNN) algorithm, Random Forests (RdF) the k-means algorithm and the DBSCAN are also evaluated. Two experiments are carried out to evaluate classification and clustering performance with respect to a range of percentages of missing values. First, characteristics for realistic data acquisition and imputation methods for missing data are detailed. Then, both experiments are described with their error measure and their performance are shown to exhibit the effectiveness of the proposed model.

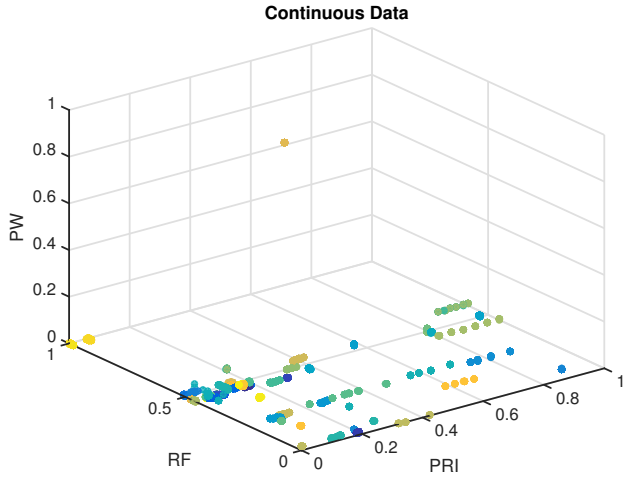


Figure 3.7: Dataset gathering 5500 continuous observations from 55 radar emitters. Some clusters are completely separable whereas some others share features and cannot be linearly separated.

Table 3.2: Categorical features observed in categorical data where 42 different combinations of these features are shared by the 55 emitters.

Intrapulse	Pulse-to-Pulse PRI	Pulse-to-pulse RF	Scan
Barker 13	Dwell	Agility	Circular
Barker 7	Triangular Dwell	Agility Burst	Sector
Chirp	Complex	Diversity	None
Diversity	HFR	FMICW	
Double Chirp	Increasing Wobble	None	
FMCW	Jitter		
Trapeze Chirp	Sinus Jitter		
Phase Code	Stagger		
S Law	Wobble		
Trapeze	None		
None			

3.4.1 Data

Realistic data are generated from an operational database gathering 55 radar emitters presenting various patterns. Each pattern consists of a sequence of pulses which are defined by a triplet of continuous features $\boldsymbol{x}_q$ = (RF,PW,PRI) and a fourtet of categorical features $\boldsymbol{x}_c$ referring to pulse-to-pulse modulations of RF and PRI, intrapulse modulations of RF and scanning types. 42 combinations of the categorical features are observed among the 55 emitters and they are composed of modulations listed in Table 3.2. For each radar emitter, 100 observations $(\boldsymbol{x}_j)_{j=1}^{100}$ are simulated from its pattern of pulses such that an observation $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$ is made up of continuous features $\boldsymbol{x}_{qj}$ and categorical features $\boldsymbol{x}_{cj}$ related to one of the pulses. Then, continuous observations are noised by applying a multivariate Gaussian noise with a diagonal covariance matrix whose diagonal elements are $[\sigma^2_{RF}, \sigma^2_{PW}, \sigma^2_{PRI}] = [1\text{MHz}, 50\text{ns}, 1\mu\text{s}]$. Negative features issued from the generated noise are thresholded to zero. Hence, outliers are embedded in observations due to the thresholding step. The dataset is shown in Figure 3.7. Moreover, extra missing values are added to evaluate limits of the proposed approach by randomly deleting coordinates of $(\boldsymbol{x}_{qj})_{j=1}^{100}$ and $(\boldsymbol{x}_{cj})_{j=1}^{100}$ for each of the 55 radar emitters. Percentages of deletion range from 5%

to 90%. At last the continuous dataset, categorical dataset and the mixed dataset are composed of $(\boldsymbol{x}_{qj})_{j=1}^{5500}$, $(\boldsymbol{x}_{cj})_{j=1}^{5500}$ and $(\boldsymbol{x}_j)_{j=1}^{5500} = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j=1}^{5500}$. As in the previous chapter, imputation methods [GLSGFV10] are used to handle missing data for comparison algorithms. Mean and k-nearest neighbours imputation methods are still implemented for continuous data. Regarding missing categorical data, they are handled through the k-nearest neighbours and mode imputation methods. The mode imputation consists in filling a missing component of an observation by the mode of observed values of that component. This method has the obvious disadvantage that it under represents the variability and also ignores correlations between observations [Sch97]. These imputation methods are compared with the proposed approach in terms of classification, clustering and reconstruction performance. For the comparison of reconstruction performance on continuous data, mean-squared errors between original continuous data and previous imputation methods are compared with the mean-squared error between original continuous data and the variational posterior marginal mean of missing continuous data given by

$$\begin{aligned} \forall j \in \mathcal{J},\ \tilde{\boldsymbol{x}}_{qj}^{\text{miss}} &= \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{qj}}^{\text{miss}}}\left[\int q(\boldsymbol{x}_{qj}^{\text{miss}}, u_j, \boldsymbol{x}_{cj} z_j)\partial u_j \partial \boldsymbol{x}_{cj} \partial z_j\right] \\ &= \sum_{k\in\mathcal{K}} \tilde{r}_{jk} \sum_{\boldsymbol{c}^{\text{obs}}\in\mathcal{C}_{q_j^{\text{obs}}}} \delta_{\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}} \sum_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \tilde{r}_{jk\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_c^{\text{miss}}} \tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}^{\text{obs}}\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_q^{\text{miss}}} . \end{aligned} \tag{3.23}$$

As for categorical data, reconstruction performance are evaluated through the comparison of Jaccard distances between original categorical data and imputation methods against Jaccard distances [Jac01] between original categorical data and the variational posterior marginal mode of missing categorical data given by

$$\begin{aligned} \forall j \in \mathcal{J},\ \tilde{\boldsymbol{x}}_{cj}^{\text{miss}} &= \arg\max_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \int q(\boldsymbol{x}_{cj}^{\text{miss}}, z_j) dz_j \\ &= \arg\max_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}} \sum_{k\in\mathcal{K}} \tilde{r}_{jk} \tilde{r}_{jk\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_c^{\text{miss}}} . \end{aligned} \tag{3.24}$$

3.4.2 Classification experiment

The classification experiment evaluates the ability of each algorithm to assign unlabeled data to one of the K classes trained by a set of labeled data. As in the previous chapter, the classification task is decomposed into a training step and a prediction step defined in procedures 3.1 and 3.2. The training step consists in estimating variational parameters of $q(\boldsymbol{\Theta})$ given a set of training data with known labels. As for the prediction step, it results in associating new data to the class that maximizes their posterior probabilities. Since comparison algorithms do not handle datasets including missing values, a complete dataset is used to enable their training. During the prediction step, incomplete observations are completed thanks to the mean and KNN imputation methods and the posterior reconstructions defined in (3.23)-(3.24). Standard configurations provided by Matlab are chosen for the RnF, the NN and the KNN algorithm. The proposed model and comparisons algorithms are trained on 70% of the initial database without extra missing values and tested on the remaining 30% of the database whose elements are randomly deleted according to different proportions of missing values. The RnF gathers 50 trees. The NN is composed of one hidden layer of 70 neurons and a softmax output layer and is trained with a cross-entropy loss. An accuracy metric is chosen for the classification experiment and observations belonging to the reject class are considered as misclassification errors. At last, hyper-parameters are initialised as in Table 3.3.

Table 3.3: Initialisation of hyper-parameters values for classification on mixed data

κ_0	π_0	η_0	γ_0	p_0	r_0	q_0	s_0	$\boldsymbol{\mu}_{0\boldsymbol{c}}$	$\boldsymbol{\Sigma}_0$
0.5	0.5	10^{-4}	4	0.9	1	1	1	$[0,0,0]$	I_3

Procedure 3.1 Classification procedure on mixed data : Training step

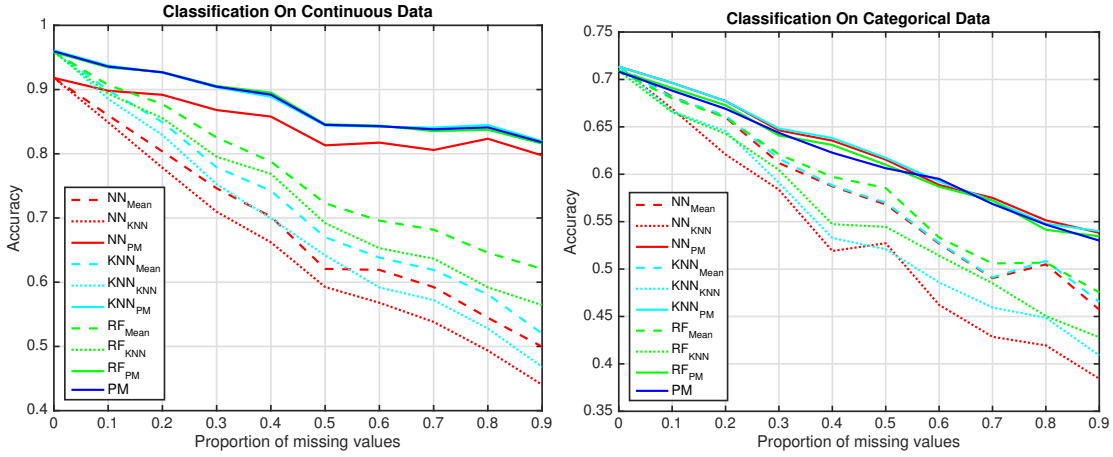
Input: Training set $\boldsymbol{x}^{\text{train}}$ and associated labels $\boldsymbol{z}^{\text{train}}$
Output: Learned parameters $\tilde{\boldsymbol{\Theta}}_{\text{train}}$
  Initialise $\boldsymbol{\kappa_0}$, $\boldsymbol{\pi}_0$, γ_0, η_0 , $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, p_0, r_0, s_0 and q_0
  for iter $= 1$ **to** itermax **do**
    Update $\tilde{\alpha}_{jk\boldsymbol{c}}$, $\tilde{\beta}_{jk\boldsymbol{c}}$, $\tilde{r}_{jk\boldsymbol{c}}^{\boldsymbol{x}_{cj}^{\text{miss}}}$, $\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_{q}^{\text{miss}}}$, $\tilde{\boldsymbol{\Sigma}}_{k}^{\boldsymbol{x}_{qj}^{\text{miss}}}$
    Update $\tilde{\kappa}_k$, $\tilde{\eta}_{k\boldsymbol{c}}$, $\tilde{\gamma}_k$, $\tilde{p}_{k\boldsymbol{c}}$, $\tilde{r}_{k\boldsymbol{c}}$, $\tilde{s}_{k\boldsymbol{c}}$, $\tilde{q}_{k\boldsymbol{c}}$, $\tilde{\pi}_{k\boldsymbol{c}}$, $\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}$, $\tilde{\boldsymbol{\Sigma}}_k$
    Calculate the lower bound $\mathcal{L}$
    if $\mathcal{L}_{\text{iter}} - \mathcal{L}_{\text{iter}-1} \leq \text{tol} \times \mathcal{L}_{\text{iter}-1}$ **then**
      return $\tilde{\boldsymbol{\Theta}}_{\text{train}} = \left(\tilde{\kappa}_k, \tilde{\eta}_{k\boldsymbol{c}}, \tilde{\gamma}_k, \tilde{p}_{k\boldsymbol{c}}, \tilde{r}_{k\boldsymbol{c}}, \tilde{s}_{k\boldsymbol{c}}, \tilde{q}_{k\boldsymbol{c}}, \tilde{\pi}_{k\boldsymbol{c}}, \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}, \tilde{\boldsymbol{\Sigma}}_k\right)_{(k,\boldsymbol{c})\in\mathcal{K}\times\mathcal{C}_q}$
    end if
  end for

For the classification experiment, results are shown in Figure 3.8 where classification performance are exhibited for the 3 datasets. Without missing data, both algorithms cannot perfectly classify the 55 radar emitters for the 3 datasets. Indeed, both algorithms reach accuracies of 90% for the continuous dataset, 75% for the categorical dataset and 98% for the mixed dataset. These performance can be explained by the non total separability of continuous and categorical datasets since the 55 emitters share 42 combinations of categorical features (Table 3.2) and (RF,PRI,PW) intervals as shown in Figure 3.7. Nonetheless when mixed data are taken into consideration, the dataset becomes more separable leading to higher performance of both algorithms. When the proportion of missing values increases, the proposed model outperforms comparisons algorithms for each dataset. It achieves accuracies of 80%, 55% and 95% for 90% of deleted continuous, categorical and mixed values whereas accuracies of comparison algorithms are lower than 65%, 50% and 75% with missing data imputation from standard methods. As in the previous chapter, these higher performance of the proposed model reveal that the proposed method embeds a more efficient inference method than other imputation methods. That result is confirmed on Figure 3.8 when comparison algorithms are applied on data reconstructed by the proposed model. Indeed when the proposed inference is chosen, comparison algorithms share the same performance than the proposed model and manage to handle missing even for 90% of deleted values. Finally, this efficiency is shown on Figure 3.9 where data reconstructed by the proposed model exhibit lower mean-squared errors and Jaccard distances for missing data imputation than the standard imputation methods. Indeed, the lowest mean-squared errors and Jaccard distances are obtained by the proposed model reconstruction on the mixed dataset, which demonstrate

Procedure 3.2 Classification procedure on mixed data: Prediction step

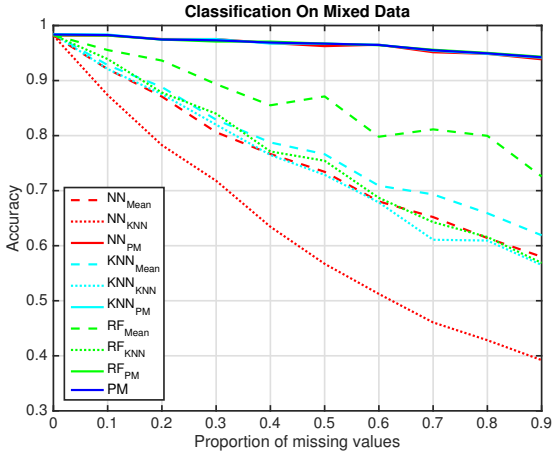
Input: Unlabelled dataset $\boldsymbol{x}^{\text{pred}}$ and learned parameters $\tilde{\boldsymbol{\Theta}}^{\text{train}}$
Output: Predicted labels $\tilde{\boldsymbol{z}}^{\text{pred}}$
  Update $\tilde{\alpha}_{jk\boldsymbol{c}}$, $\tilde{\beta}_{jk\boldsymbol{c}}$, $\tilde{r}_{jk\boldsymbol{c}}^{\boldsymbol{x}_{cj}^{\text{miss}}}$, $\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_{q}^{\text{miss}}}$, $\tilde{\boldsymbol{\Sigma}}_{k}^{\boldsymbol{x}_{qj}^{\text{miss}}}$, $\tilde{r}_{jk}$
  return $\tilde{\boldsymbol{z}}^{\text{pred}}$ such that each $\tilde{z}_j^{\text{pred}} = \arg\max_{k\in\mathcal{K}} \tilde{r}_{jk}$

that missing data imputation is even more efficient when both continuous and categorical are jointly modeled. Furthermore, a correlation between higher performance of the proposed model and the quality of its reconstructions can be noticed for any percentage of missing values. Then, effectiveness of the proposed model can be explained by the fact that missing data imputation methods can create outliers that deteriorate performance of classification algorithms whereas the inference on missing data and labels prediction are jointly estimated in the proposed model. Indeed, embedding the inference procedure into the model framework allows properties of the model, such as outliers handling, to counterbalance drawbacks of imputation methods such as outlier creation.



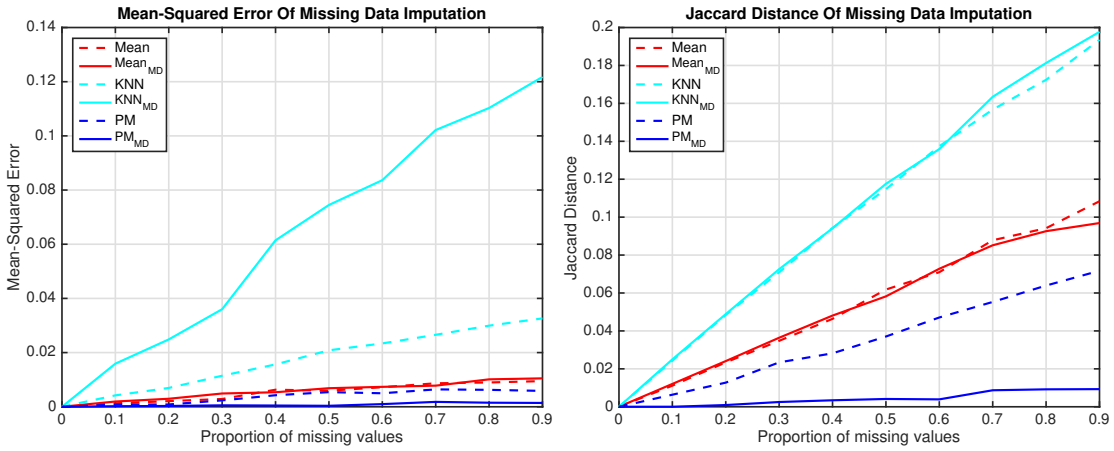
(a) Classification on continuous data.

(b) Classification on categorical data.



(c) Classification on mixed data.

Figure 3.8: Classification performance are presented for the proposed model (PM) in blue, the NN in red, the RnF in green and the KNN in cyan. Figure (a) exhibits classification performance when only continuous data are taken into consideration. Figure (b) exhibits classification performance when only categorical data are taken into consideration. Figure (c) exhibits classification performance when both continuous and categorical data are taken into consideration. For each figure, the solid lines represent accuracies with a posteriori reconstructed missing data for the NN, the RnF and the KNN, the doted dashed lines stands for accuracies with mean/mode imputation for the NN, the RnF and the KNN whereas the dashed lines shows accuracies with KNN imputation.



(a) Mean-squared errors of imputation methods and the posterior reconstruction (3.23).

(b) Jaccard distances of imputation methods and the posterior reconstruction (3.24).

Figure 3.9: Evaluation of imputation methods and posterior reconstructions (3.23-3.24) while considering continuous, categorical and mixed data. Performance of reconstructions are presented in red for the Mean and Mode imputations, in cyan for the KNN imputation and in blue for the proposed model. Figure (a) exhibits mean-squared errors related to imputation methods and the posterior reconstruction (3.23) when continuous data (dashed lines) and mixed data (solid lines) are considered. Figure (b) exhibits Jaccard distances related to imputation methods and the posterior reconstruction (3.24) when categorical data (dashed lines) and mixed data (solid lines) are considered.

3.4.3 Clustering experiment

The clustering experiment is composed of two experiments that aim to exhibit the clustering ability of each algorithm according to an a priori number of clusters $K \in \{K_{\min}, \ldots, K_{\max}\}$. As developed in previous chapters, the clustering algorithm is decomposed into two parts. First, a semi-supervised classification is performed for each K ranges from $K_{\min}$ to $K_{\max}$ to estimate variational parameters of $q(\boldsymbol{\Theta}, \boldsymbol{H})$ and labels in a mixture of K components. Then, the value of K that maximizes the lower bound is retained as the posterior number of clusters as well as its associated parameters.

Procedure 3.3 Semi-supervised classification procedure on mixed data

Input: Unlabelled dataset $\boldsymbol{x}$ and number of classes K
Output: Labels $\tilde{\boldsymbol{z}}$ and parameters $\tilde{\boldsymbol{\Theta}}$

Initialise $\boldsymbol{\kappa}_0$, $\boldsymbol{\pi}_0$, γ_0, η_0 , $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, p_0, r_0, s_0 and q_0
for iter $= 1$ **to** itermax **do**
  Update $\tilde{\alpha}_{jk\boldsymbol{c}}$, $\tilde{\beta}_{jk\boldsymbol{c}}$, $\tilde{r}_{jk\boldsymbol{c}}^{\boldsymbol{x}_{cj}^{\text{miss}}}$, $\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_{q}^{\text{miss}}}$, $\tilde{\boldsymbol{\Sigma}}_{k}^{\boldsymbol{x}_{qj}^{\text{miss}}}$, $\tilde{r}_{jk}$
  Update $\tilde{\kappa}_k$, $\tilde{\eta}_{k\boldsymbol{c}}$, $\tilde{\gamma}_k$, $\tilde{p}_{k\boldsymbol{c}}$, $\tilde{r}_{k\boldsymbol{c}}$, $\tilde{s}_{k\boldsymbol{c}}$, $\tilde{q}_{k\boldsymbol{c}}$, $\tilde{\pi}_{k\boldsymbol{c}}$, $\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}$, $\tilde{\boldsymbol{\Sigma}}_k$
  Calculate the lower bound $\mathcal{L}$
  if $\mathcal{L}_{\text{iter}} - \mathcal{L}_{\text{iter}-1} \leq \text{tol} \times \mathcal{L}_{\text{iter}-1}$ **then**
    return $\tilde{\boldsymbol{\Theta}} = \left(\tilde{\kappa}_k, \tilde{\eta}_{k\boldsymbol{c}}, \tilde{\gamma}_k, \tilde{p}_{k\boldsymbol{c}}, \tilde{r}_{k\boldsymbol{c}}, \tilde{s}_{k\boldsymbol{c}}, \tilde{q}_{k\boldsymbol{c}}, \tilde{\pi}_{k\boldsymbol{c}}, \tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}, \tilde{\boldsymbol{\Sigma}}_k\right)_{(k,\boldsymbol{c})\in\mathcal{K}\times\mathcal{C}_q}$ and $\tilde{\boldsymbol{z}}$ such that each $\tilde{z}_j = \arg\max_{k\in\mathcal{K}} \tilde{r}_{jk}$
  end if
end for

According to the dataset visualised in Figure 3.7, $K_{\min}$ and $K_{\max}$ are set to 35 and 85 in

Procedure 3.4 Clustering procedure on mixed data

Input: Unlabelled dataset $\boldsymbol{x}$ and a priori range of numbers of clusters $K \in \{K_{\min}, \ldots, K_{\max}\}$
Output: Labels $\tilde{\boldsymbol{z}}$, parameters $\tilde{\boldsymbol{\Theta}}$ and optimal number of clusters $\tilde{K}$
  for $K = K_{\min}$ **to** $K_{\max}$ **do**
    Perform semi-supervised classification with K classes
    Stock labels $\tilde{\boldsymbol{z}}^K$, parameters $\tilde{\boldsymbol{\Theta}}^K$ and $\mathcal{L}^K$
  end for
  return $\tilde{\boldsymbol{\Theta}}^{\tilde{K}}$ and $\tilde{\boldsymbol{z}}^{\tilde{K}}$ such that $\tilde{K} = \arg\max_K \mathcal{L}^K$

Table 3.4: Initialisation of hyper-parameters values for clustering on mixed data

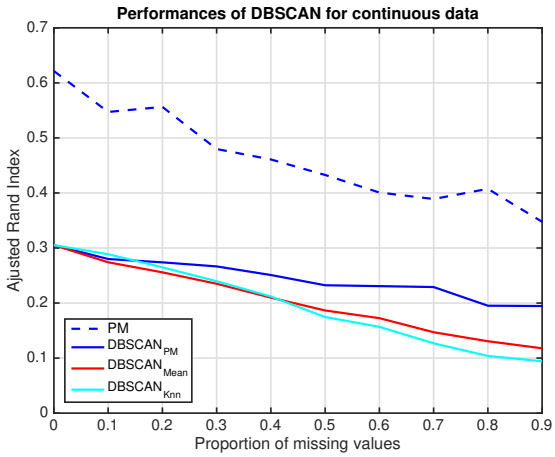
κ_0	π_0	η_0	γ_0	p_0	r_0	q_0	s_0
0.5	0.5	100	4	0.9	1	1	1

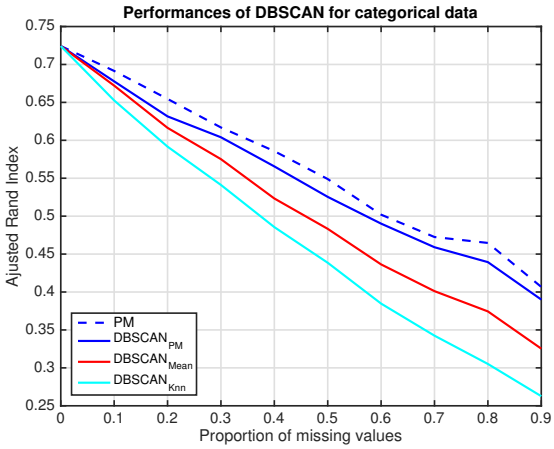
order to evaluate the impact of the a priori number of clusters on data clustering. Parameters of DBSCAN are set to Minpts = 5 and eps = 0.01 by using an heuristic proposed in the original paper [EKS+96]. A supervised initialisation is retained for the proposed model due to its sensitivity to initialisation. It consists in initialising prior component means $\boldsymbol{\mu}_{0\boldsymbol{c}}$ from results of a k-means algorithm and prior component covariance matrices $\boldsymbol{\Sigma}_0$ from diagonal matrices whose diagonal elements are variances of observed features. Other hyper-parameters are initialised as in Table 3.4. Since comparison algorithms do not handle observations with missing values and do not provide a clustering result for them, missing data are reconstructed through the mean, the KNN and the proposed model imputation methods before running these algorithms.

The first clustering experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters $K \in \{K_{\min}, \ldots, K_{\max}\}$. Performance are evaluated through the Adjusted Rand Index (ARI) [HA85] that compares estimated partitions of data with the ground-truth. Results of the first experiment on the 3 datasets are shown in Figures 3.10 and 3.11. Without the presence of missing values, performance of DBSCAN, k-means and the proposed model are similar with ARIs of 62%, 72% and 88% for the continuous, categorical and mixed datasets. As in the classification experiment, these performance are explained by the non total separability of continuous and categorical datasets since the 55 emitters share 42 combinations of categorical features (Table 3.2) and (RF,PRI,PW) intervals as shown in Figure 3.7. Once again when mixed data are taken into consideration, the dataset becomes more separable leading to higher performance of both algorithms (ARI = 88%). When the proportion of missing values increases, the proposed model outperforms both DBSCAN and k-means and achieves ARIs of 35% ,40% and 58% on continuous, categorical and mixed datasets for 90% of deleted values whereas the ARIs of comparison algorithms with standard missing data imputation are lower than 35% on each dataset. As in the classification experiment, these higher performance reveal that the proposed method embeds a more efficient inference method than other imputation methods. That result is confirmed on both Figure 3.10 and Figure 3.11 where DBSCAN and k-means are applied on data reconstructed by the proposed model. Indeed, performance of both algorithms increase up to performance of the proposed model for any percentages of deleted values when the proposed inference is chosen.

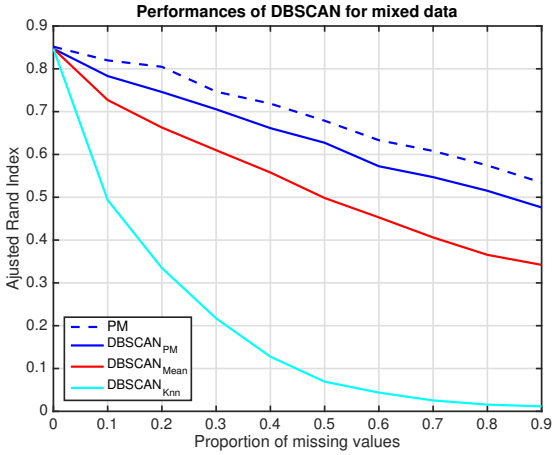
The second experiment tests the ability of each algorithm to find the true number of clusters $\tilde{K}$ among $\{K_{\min}, \ldots, K_{\max}\}$. The lower bound (3.22) and the average Silhouette score [KR09] are

criteria used to select the optimal number of clusters for the proposed model and the k-means algorithm. Indeed, the ARI cannot be used since it requires the ground-truth and DBSCAN automatically selects a number of clusters for a given dataset. Results of the second experiment on the 3 datasets are visible on Figures Figure 3.12 and 3.13. Figure 3.12 presents numbers of clusters selected by the lower bound and average Silhouette scores for the proposed model and k-means algorithm according to different proportions of missing values and imputation methods. Without missing data, the correct number of clusters (K=55) is selected by the two criteria for the k-means algorithm and the proposed model when continuous and mixed data are clustered. As for categorical data, both criteria select 45 as the optimal number of clusters since the 42 combinations of categorical features (Table 3.2) shared by the 55 emitters constitute 42 distinct clusters. In presence of missing values, the average Silhouette score mainly selects $K = 65$ when the k-means algorithm is run on the 3 datasets completed by standard imputation methods. When, the k-means algorithm performs clustering on the posterior reconstructions, the average Silhouette score correctly selects $K = 55$ until 60% of missing values for continuous data and 40% of missing values for mixed data. Eventually when the proposed model does clustering, the two criteria select the correct number of clusters $K = 55$ until 70% of missing values for continuous and mixed data. These results show two main advantages of the proposed model. As previously, the proposed model provides a more robust inference on missing data since the average Silhouette score chooses more representative number of clusters when the k-means algorithm is run on the posterior reconstructions than on data completed by standard imputation methods. Furthermore, since the lower bound criterion also selects the correct number of clusters as the average Silhouette score, it can be used as a valid criterion for selecting the optimal number of clusters and does not require extra computational costs as the Silhouette score since it is computed during the model parameter estimation. Finally, the proposed approach provides a more robust inference on missing data and a criterion for selecting the optimal number of clusters without extra computations. As for the Figure 3.13, it shows the evolution of the number of clusters estimated by DBSCAN according to different proportions of missing values and imputation methods. Since DBSCAN automatically estimates the number of clusters and manages outliers by creating new clusters, results on Figure 3.13 can be used to evaluate performance of imputations methods. For mean, mode and k-NN imputation methods, DBSCAN estimates a number of clusters greater than the number estimated for the proposed model according to any proportion of missing values. These performance indicate that the proposed approach creates less outliers than other imputation methods by providing a more robust inference on missing data since DBSCAN localizes less outliers in the posterior reconstructions (3.23-3.24) than in standard imputation methods.



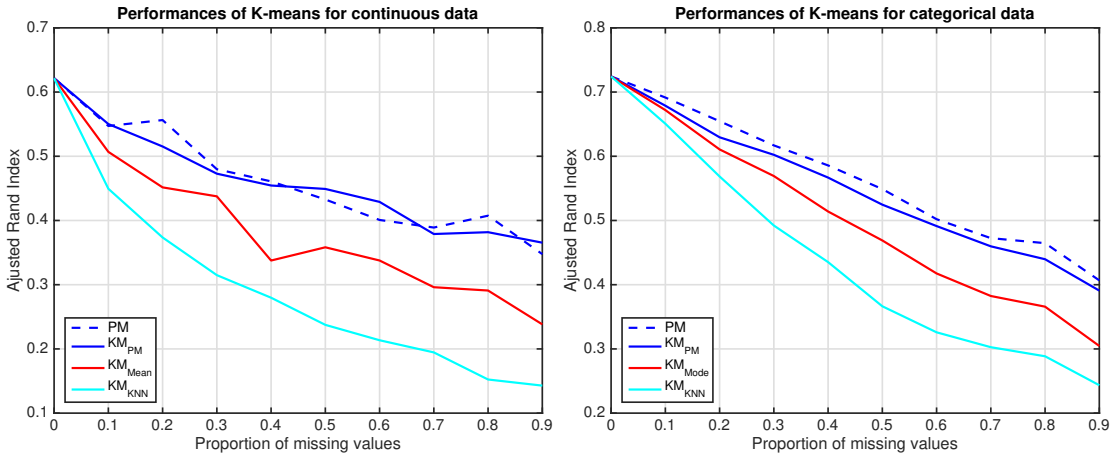


(a) Clustering performance for continuous data. (b) Clustering performance for categorical data.

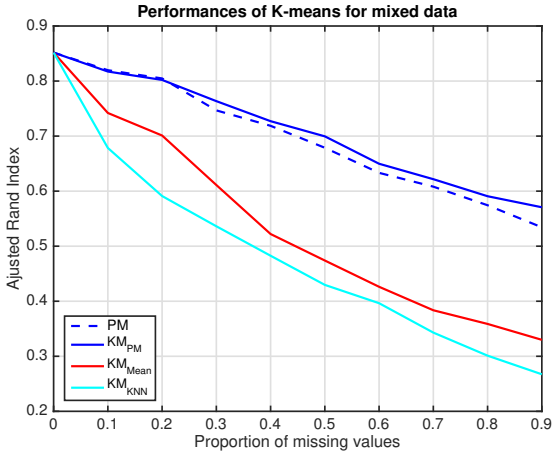


(c) Clustering performance for mixed data.

Figure 3.10: Performance of the proposed model compared with DBSCAN according to different proportions of missing values and imputation methods. The number of clusters K is fixed at 45 for categorical data and 55 for continuous and mixed data.

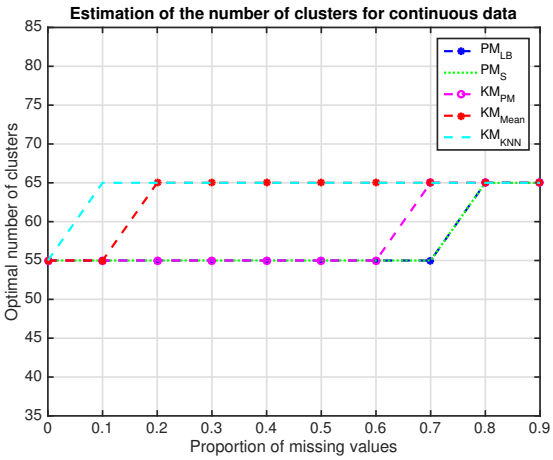


(a) Clustering performance for continuous data. (b) Clustering performance for categorical data.

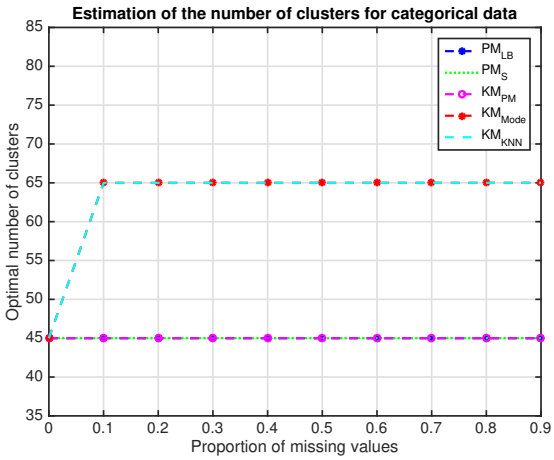


(c) Clustering performance for mixed data.

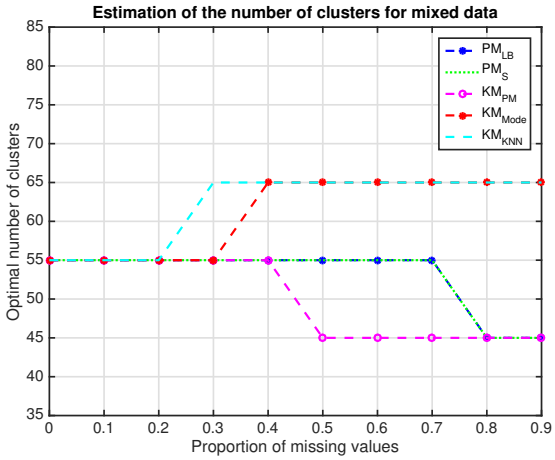
Figure 3.11: Performance of the proposed model compared with k-means algorithm according to different proportions of missing values and imputation methods. The number of clusters K is fixed at 45 for categorical data and 55 for continuous and mixed data.



(a) Estimation of the number of clusters for continuous data.

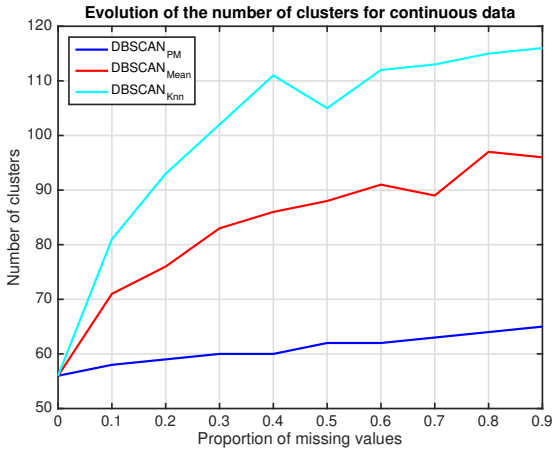


(b) Estimation of the number of clusters for categorical data.

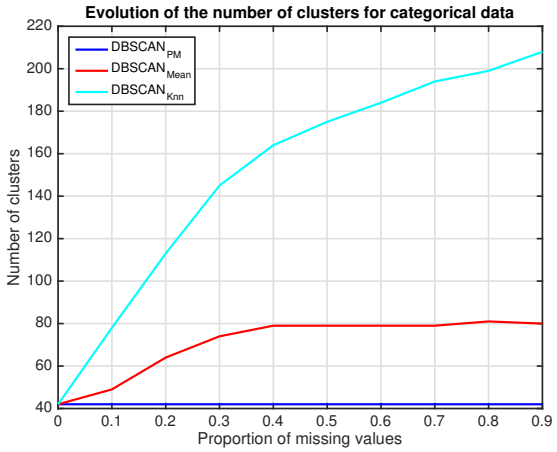


(c) Estimation of the number of clusters for mixed data.

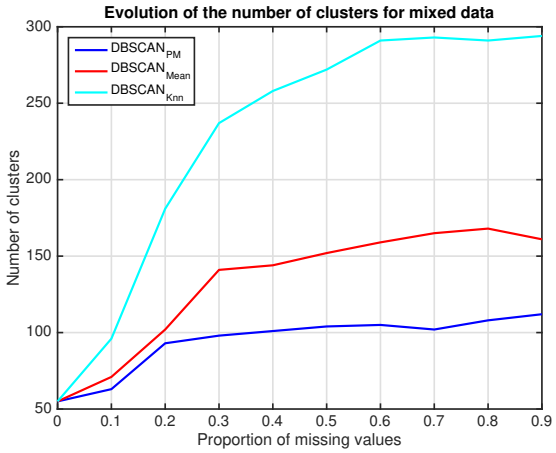
Figure 3.12: Estimation of the number of clusters using the lower bound (LB) and the silhouette score (S) for the proposed model and only the silhouette score (S) for the k-means algorithm.



(a) Estimation of the number of clusters for continuous data.



(b) Estimation of the number of clusters for categorical data.



(c) Estimation of the number of clusters for mixed data.

Figure 3.13: Estimation of the number of clusters by DBSCAN according to mean imputation, k-NN imputation and posterior reconstruction of the proposed model.

3.5 Conclusion

In this chapter, various modulations of radar emitter patterns have been presented. These modulations can be used as categorical data in the classification and clustering tasks. Hence, a mixture model handling both continuous data and categorical data has been developed. An approach based on the Location Mixture Model has been investigated by establishing conditional relations between continuous and categorical data. Benefiting from a dependence structure designed for mixed data, the proposed model shows its efficiency for inferring on missing data, performing classification and clustering tasks and selecting the correct number of clusters. Since the posterior distribution is intractable, model learning is processed through a variational Bayes inference where variational posterior distributions are proposed for continuous and categorical missing values. Experiments show that the proposed approach handles mixed data even in presence of missing values and can outperform standard algorithms in clustering tasks. Indeed the main advantage of our approach is that it enables the counterbalance of imputation methods drawbacks by embedding the inference procedure into the model framework.

Chapter 4

Temporal Evolution Data

Continuous data describing radar emitters waveforms such as the Carrier Frequency, the Pulse Width and the Pulse Repetition Interval have been previously taken into account in order to cluster radar emitters. Nonetheless, the Pulse Description World gathers other features such the Amplitude whose relation with the Time of Arrival reflects the scanning behaviour of a radar emitter. Therefore, this temporal relation can be exploited to cluster radar emitters. Depending on the scanning type, this relation can be represented by either a parabola or a piecewise parabola. These two relations have to be included into the mixture distribution to take advantage of the temporal behaviour of each radar emitter. This chapter contains three sections which focus on three different cases. The first section deals with data where only emitters having a parabolic scanning behaviour are observed. The second section introduces the case where temporal evolution data are only distributed according to piecewise parabolic relations. As for the last section, it is about the case where any type of scanning behaviours can be observed in data. Each section presents the model integrating radar temporal evolution data and its inference procedure before proposing a more complete model taking into consideration temporal evolution data and mixed data. Eventually, experiments are carried out to exhibit performance of the proposed approach. In this chapter, radar temporal evolution data consist of J pulses gathering J amplitudes $\boldsymbol{x}_t = (x_{tj})_{j\in\mathcal{J}}$ and J times of arrival $\boldsymbol{t} = (t_j)_{j\in\mathcal{J}}$ from K distinct emitters.

Contents

4.1 Parabolic data

In this section, K emitters presenting parabolic scanning behaviours are considered. Therefore, the main objective is to develop a mixture model which can build K distinct clusters formed by K parabolas and to define an inference procedure for parameter estimation. Then, the proposed model is enhanced with the mixture model designed for mixed data in Chapter 3 in order to improve clustering performance. Finally, experiments on synthetic and real data are carried out to exhibit performance of the proposed approach.

4.1.1 Model

Before introducing a mixture model that handles parabolic data, the parabolic relation between amplitudes and times of arrival is defined. Then, the mixture model is developed into a Bayesian framework.

Parabola Equation

The parabolic relation between the amplitude x_{tj} of the j^{th} pulse and its time of arrival t_j, visible on Figure 4.1, can be described by the following parabolic equation

$$x_{tj} = at_j^2 + bt_j + c + \epsilon \tag{4.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a measurement noise introduced to model defects of materials. Since the measurement noise ϵ is only embedded in materials, the variance parameter σ^2 is independent from $\boldsymbol{x}_t$ and $\boldsymbol{t}$. Equation (4.1) can be reformulated as a linear regression problem such that

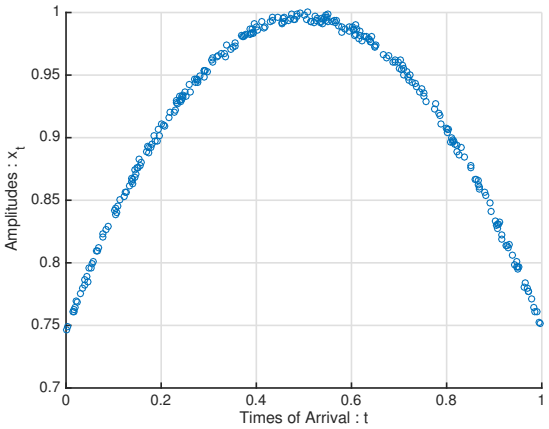


Figure 4.1: Simulated data where amplitudes $\boldsymbol{x}_t$ and times of arrival $\boldsymbol{t}$ are distributed according to a parabolic relation.

$$x_{tj} = \boldsymbol{\Phi}(t_j)^T \boldsymbol{\omega} + \epsilon \tag{4.2}$$

with $\boldsymbol{\Phi}(t_j) = (t_j^2, t_j, 1)^T$ the vector containing polynomial transformations of t_j and $\boldsymbol{\omega} = (a, b, c)^T$ the vector of regression parameters. Since the measurement error is assumed to be Gaussian, the amplitude x_{tj} is distributed according to a normal distribution centered in $\boldsymbol{\Phi}(t_j)^T \boldsymbol{\omega}$ with variance σ^2

$$x_{tj} \sim \mathcal{N}\left(x_{tj} | \boldsymbol{\Phi}(t_j)^T \boldsymbol{\omega}, \sigma^2\right) . \tag{4.3}$$

Mixture model

Since each radar emitter has its own scanning behaviour, K unique parabolas exist in data and they are configured with K regression parameters $\boldsymbol{\omega} = (\boldsymbol{\omega}_k)_{k\in\mathcal{K}}$. Then, each amplitude x_{tj} belongs to one of these parabolas which is related to a specific emitter. In other words, conditionally to its label z_j and its time of arrival t_j, the amplitude x_{tj} is distributed according to (4.3) such that the component distribution is defined by

$$x_{tj}|t_j, z_j = k \sim \mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) \tag{4.4}$$

Recalling that $p(z_j = k) = a_k$ where $\boldsymbol{a} = (a_k)_{k\in\mathcal{K}}$ are the weights related to component distributions, the mixture model is obtained from (4.4) such that

$$\forall j \in \mathcal{J},\ p(x_{tj}|t_j, \boldsymbol{\Theta}) = \sum_{k\in\mathcal{K}} a_k \mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) \tag{4.5}$$

where $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{\omega}, \sigma^2)$ is the set of parameters.

Bayesian framework

As in chapters 2 and 3, a Bayesian framework is used to estimate parameters $\boldsymbol{\Theta}$. Assuming datasets $(\boldsymbol{x}_t, \boldsymbol{t})$ of i.i.d observations $(x_{tj}, t_j)_{j\in\mathcal{J}}$ and independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$, the complete likelihood associated to (4.5) is defined by

$$p(\boldsymbol{x}_t, \boldsymbol{z}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \left(a_k \mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right)\right)^{\delta_{z_j}^k}$$

Eventually, the prior distribution required for $\boldsymbol{\Theta}$ is chosen as

$$p(\boldsymbol{\Theta}|\mathcal{K}) = p(\boldsymbol{a}|\mathcal{K})p(\boldsymbol{\omega}|\sigma^2, \mathcal{K})p(\sigma^2)$$

where $\boldsymbol{a}$ follows a Dirichlet distribution, each $\boldsymbol{\omega}_k$ follows a Normal distribution and σ^2 follows an Inverse Gamma distribution such that

$$\begin{cases} p(\mathbf{a}|\mathcal{K}) = \mathcal{D}(\mathbf{a}|\kappa_0)\ , \\ p(\boldsymbol{\omega}|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\boldsymbol{\omega}_0, \sigma^2\boldsymbol{\Lambda}_0\right)\ , \\ p(\sigma^2) = \mathcal{IG}(\sigma^2|\xi_1^0, \xi_2^0)\ . \end{cases}$$

The resulting mixture model is shown on Figure 4.2.

4.1.2 Inference

The Variational Bayes (VB) procedure is derived to estimate parameters of the mixture model defined in (4.5). Variational posterior distributions are obtained from the VB Expectation (VBE) and VB Maximization (VBM) steps and a lower bound on the log evidence is defined to master the convergence of the VB procedure.

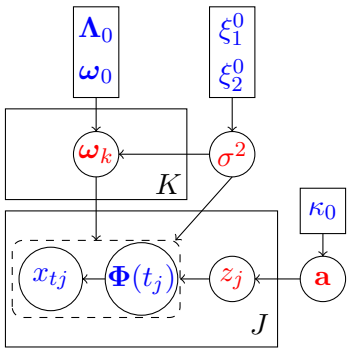


Figure 4.2: Graphical representation of the proposed mixture model handling parabolic data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations (x_{tj}, t_j) decomposed into the amplitude x_{tj} and the polynomial transformation $\boldsymbol{\Phi}(t_j)$ and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red.

Variational posterior distributions

As previously, a factorized posterior distribution $q(\boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K}) = q(\boldsymbol{z}|\mathcal{K})q(\boldsymbol{\Theta}|\mathcal{K})$ is chosen as an approximation of the intractable posterior joint distribution $p(\boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{x}_t, \boldsymbol{t}, \mathcal{K})$ such that latent variables $\boldsymbol{z}$ and parameters $\boldsymbol{\Theta}$ are a posteriori independent and $q(\boldsymbol{\Theta}|\mathcal{K}) = q(\boldsymbol{a}|\mathcal{K})q(\boldsymbol{\omega}|\sigma^2\mathcal{K})q(\sigma^2)$. According to VB assumptions, the following conjugate variational posterior distributions are obtained from the VB procedure

$$\begin{cases} q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j) \ , \\ q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}}) \ , \\ q(\boldsymbol{\omega}|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\tilde{\boldsymbol{\omega}}_k, \sigma^2\tilde{\boldsymbol{\Lambda}}_k\right) \ , \\ q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1, \tilde{\xi}_2) \ . \end{cases} \tag{4.6}$$

Their respective parameters are estimated during the VBE and VBM steps.

VBE-step

The VBE-step consists in deriving the following expectation

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - \frac{1}{2}\Bigg(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\right]\Bigg)\Bigg) \\ &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \log \rho_{jk} \end{aligned} \tag{4.7}$$

where

$$\log \rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - \frac{1}{2}\left(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\right]\right) . \tag{4.8}$$

Hence, a categorical distribution for labels $\boldsymbol{z}$ is deduced from (4.7) such that

$$q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j)$$

and their parameters $(\tilde{\boldsymbol{r}}_j)_{j\in\mathcal{J}}$ are obtained from (4.8) as follows

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \tilde{r}_{jk} = \frac{\rho_{jk}}{\sum\limits_{k\in\mathcal{K}} \rho_{jk}} .$$

VBM-step

The VBM-step consists in deriving the following expectation

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right] =& \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] + \log p(\boldsymbol{\Theta}|\mathcal{K}) \\
&= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\Bigg(\log a_k - \frac{1}{2}\Bigg(\log 2\pi + \log\sigma^2 \\
&+ \frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\Bigg)\Bigg) + \sum_{k\in\mathcal{K}}(\kappa_k^0 - 1)\log a_k \\
&- \frac{1}{2}\left(3(\log 2\pi + \log\sigma^2) + \log|\boldsymbol{\Lambda}_0| + (\boldsymbol{\omega}_k - \boldsymbol{\omega}_0)^T\frac{\boldsymbol{\Lambda}_0^{-1}}{\sigma^2}(\boldsymbol{\omega}_k - \boldsymbol{\omega}_0)\right) \\
&- (\xi_1^0 + 1)\log\sigma^2 - \frac{\xi_2^0}{\sigma^2} + \log c_{\mathcal{D}}(\boldsymbol{\kappa}^0) + \log c_{\mathcal{IG}}(\xi_1^0, \xi_2^0) .
\end{aligned} \tag{4.9}$$

By factorizing terms related to $\boldsymbol{a}$ in (4.9), the following Dirichlet distribution is obtained

$$q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}})$$

where

$$\forall k \in \mathcal{K},\ \tilde{\kappa}_k = \kappa_k^0 + \sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{z}}\left[\delta_{z_j}^k\right] .$$

By aggregating terms related to each $\boldsymbol{\omega}_k$ in (4.9), a Normal distribution is obtained for each $\boldsymbol{\omega}_k$ such that

$$q(\boldsymbol{\omega}|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\tilde{\boldsymbol{\omega}}_k, \sigma^2\tilde{\boldsymbol{\Lambda}}_k\right)$$

where $\forall k \in \mathcal{K}$,

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_k &= \left(\sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{z}}\left[\delta_{z_j}^k\right]\boldsymbol{\Phi}(t_j)\boldsymbol{\Phi}(t_j)^T + \boldsymbol{\Lambda}_0^{-1}\right)^{-1} , \\
\tilde{\boldsymbol{\omega}}_k &= \tilde{\boldsymbol{\Lambda}}_k\left(\sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{z}}\left[\delta_{z_j}^k\right] x_{tj}\boldsymbol{\Phi}(t_j) + \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\omega}_0\right) .
\end{aligned}$$

Eventually, an Inverse Gamma distribution is deduced from (4.9) such that

$$q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1, \tilde{\xi}_2)$$

where

$$\begin{aligned}
\tilde{\xi}_1 &= \xi_1^0 + \frac{J}{2} , \\
\tilde{\xi}_2 &= \xi_2^0 + \frac{1}{2}\sum_{k\in\mathcal{K}}\left(\sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{z}}\left[\delta_{z_j}^k\right] x_{tj}^2 + \boldsymbol{\omega}_0^T\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\omega}_0 - \tilde{\boldsymbol{\omega}}_k^T\tilde{\boldsymbol{\Lambda}}_k^{-1}\tilde{\boldsymbol{\omega}}_k\right) .
\end{aligned}$$

Lower bound

Recalling that the lower bound on the log evidence is given by

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right] - \mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right]$$

where $\mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right]$ is the free energy and $\mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right]$ is the entropy of the approximate posterior $q(\boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})$. The free energy can be developed as

$$\mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right] = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right]$$

where

$$\mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{z}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{z}}\left[\delta_{z_j}^k\right] \log \rho_{jk}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right] = \sum_{k\in\mathcal{K}} (\kappa_k^0 - 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - \frac{1}{2}\Bigg(3(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right]) + \log|\boldsymbol{\Lambda}_0| \\ + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right](\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right] - \boldsymbol{\omega}_0)^T\boldsymbol{\Lambda}_0^{-1}(\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right] - \boldsymbol{\omega}_0) + \text{Trace}\left(\tilde{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_0^{-1}\right)\Bigg) \\ - (\xi_1^0 + 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right] - \xi_2^0\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] + \log c_{\mathcal{D}}(\boldsymbol{\kappa}^0) + \log c_{\mathcal{IG}}(\xi_1^0, \xi_2^0)\ .\end{aligned}$$

As for the entropy term, the following decompositon is obtained

$$\mathbb{E}_{\boldsymbol{z},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right] = \mathbb{E}_{\boldsymbol{z}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]$$

where

$$\mathbb{E}_{\boldsymbol{z}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{z}}\left[\delta_{z_j}^k\right] \log \tilde{r}_{jk}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right] = \sum_{k\in\mathcal{K}} (\tilde{\kappa}_k - 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - \frac{1}{2}\left(3(1 + \log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right]) + \log|\tilde{\boldsymbol{\Lambda}}_k|\right) \\ - (\tilde{\xi}_1 + 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right] - \tilde{\xi}_2\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] + \log c_{\mathcal{D}}(\tilde{\boldsymbol{\kappa}}) + \log c_{\mathcal{IG}}(\tilde{\xi}_1, \tilde{\xi}_2)\ .\end{aligned}$$

Expectations

Expectations developed in variational calculations are derived from properties of variational posterior distributions and are obtained as follows. Categorical distribution properties lead to

$$\begin{aligned}&\forall j \in \mathcal{J},\ \forall k \in \mathcal{K}: \\ &\qquad\qquad \mathbb{E}_{\boldsymbol{z}}\left[\delta_{z_j}^k\right] = \tilde{r}_{jk}\ .\end{aligned}$$

Dirichlet distribution properties lead to

$$\begin{aligned}&\forall k \in \mathcal{K}: \\ &\qquad \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k\in\mathcal{K}} \tilde{\kappa}_k\right)\ ,\end{aligned}$$

where $\psi(\cdot)$ is the digamma function. Normal distribution properties lead to

$$\begin{aligned}
&\forall k \in \mathcal{K} : \\
&\quad \mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right] = \tilde{\boldsymbol{\omega}}_k \ , \\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k \boldsymbol{\omega}_k^T\right] = \mathbb{V}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right]^T \\
&\qquad\qquad = \sigma^2 \tilde{\boldsymbol{\Lambda}}_k + \tilde{\boldsymbol{\omega}}_k \tilde{\boldsymbol{\omega}}_k^T \ ,
\end{aligned}$$

Inverse Gamma distribution properties lead to

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] &= \frac{\tilde{\xi}_1}{\tilde{\xi}_2} \ , \\
\mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right] &= \log \tilde{\xi}_2 - \psi(\tilde{\xi}_1) \ .
\end{aligned}$$

Using all these properties, the following expectation can be calculated as

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K} :$$

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T \boldsymbol{\omega}_k)^2}{\sigma^2}\right] = \frac{\tilde{\xi}_1 \left(x_{tj} - \boldsymbol{\Phi}(t_j)^T \tilde{\boldsymbol{\omega}}_k\right)^2}{\tilde{\xi}_2} + \text{Trace}\left(\boldsymbol{\Phi}(t_j)^T \tilde{\boldsymbol{\Lambda}}_k \boldsymbol{\Phi}(t_j)\right)$$

4.1.3 Complete model

A model integrating parabolic data and mixed data is now presented. By taking into consideration any types of available data, the resulting model can fit data better and can estimate more accurate clusters. First, data formalism and assumptions are detailed. Then, the resulting mixture model and its inference procedure are developed.

Data and assumptions

In this part, data consist of J pulses gathering J amplitudes $\boldsymbol{x}_t = (x_{tj})_{j\in\mathcal{J}}$ associated to J times of arrival $\boldsymbol{t} = (t_j)_{j\in\mathcal{J}}$, J continuous features $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ and J categorical features $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ from K distinct emitters. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj})$ the j^{th} observation vector of mixed variables where

- $\boldsymbol{x}_{qj} \in \mathbb{R}^d$ is a vector of d continuous radar features such as the Radio Frequency, the Pulse Width, the Azimuth or the Pulse Repetition Interval,
- $\boldsymbol{x}_{cj} = (x_{cj_0}, \ldots, x_{cj_{q-1}}) \in C_q$ is a vector of q categorical radar modulations such as intra-pulse modulations or pulse-to-pulse modulations,
- $x_{tj} \in \mathbb{R}$ is a continuous variable modeling the Amplitude.

For each pulse j, the temporal evolution variable x_{tj} and mixed variables $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$ are assumed to be independent conditionally to each cluster $k \in \mathcal{K}$

$$\forall j \in \mathcal{J},\ (\boldsymbol{x}_q, \boldsymbol{x}_c)|z_j = k \perp\!\!\!\perp x_t|z_j = k \ . \tag{4.10}$$

with z_j the latent variable modeling the label of the j^{th} observation vector $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj})$. Moreover, the temporal data $(x_{tj}, t_j)_{j\in\mathcal{J}}$ are distributed according to a parabolic relation and the quantitative data $(\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ are normally distributed conditionally to categorical data $(\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$. Both quantitative and categorical data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ can be partially observed. Hence $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$

are decomposed into observed features $(\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj}^{\text{obs}})_{j\in\mathcal{J}}$ and missing features $(\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{miss}})_{j\in\mathcal{J}}$ such that

$$\forall j \in \mathcal{J}, \quad \begin{aligned} \boldsymbol{x}_{qj} &= \begin{pmatrix} \boldsymbol{x}_{qj}^{\text{miss}} \\ \boldsymbol{x}_{qj}^{\text{obs}} \end{pmatrix} \text{ with } (\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{qj}^{\text{obs}}) \in \mathbb{R}^{d_j^{\text{miss}}} \times \mathbb{R}^{d_j^{\text{obs}}} \text{ and } d_j^{\text{miss}} + d_j^{\text{obs}} = d \,, \\ \boldsymbol{x}_{cj} &= \begin{pmatrix} \boldsymbol{x}_{cj}^{\text{miss}} \\ \boldsymbol{x}_{cj}^{\text{obs}} \end{pmatrix} \text{with } (\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) \in \mathcal{C}_{q_j^{\text{miss}}} \times \mathcal{C}_{q_j^{\text{obs}}} \text{ and } q_j^{\text{miss}} + q_j^{\text{obs}} = q \,. \end{aligned}$$

Mixture model

According to the independence assumption (4.10), the distribution of mixed data (Chapter 3) and the parabolic relation between temporal evolution data, the component distribution results in

$$\forall j \in \mathcal{J},\ p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}|z_j = k) = p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}|z_j = k)p(x_{tj}|z_j = k)$$

where

$$\forall j \in \mathcal{J}, \quad \begin{aligned} p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}|z_j = k) &= \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \,, \\ p(x_{tj}|t_j, z_j = k) &= \mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) \end{aligned} \tag{4.11}$$

with

- $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$ the scale latent variables handling outliers for quantitative data $\boldsymbol{x}_q$ and distributed according to a Gamma distribution with shape and rate parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})_{(k,\boldsymbol{c})\in\mathcal{K}\times\mathcal{C}_q}$ conditionally to categorical data $\boldsymbol{x}_c$ and labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$,
- $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = ((\boldsymbol{\mu}_{k\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}, \boldsymbol{\Sigma}_k)_{k\in\mathcal{K}}$ the mean and the variance parameters of quantitative data $\boldsymbol{x}_q$ for each cluster,
- $\boldsymbol{\pi} = (\boldsymbol{\pi}_k)_{k\in\mathcal{K}}$ the weights of the multivariate Categorical distribution of categorical data $\boldsymbol{x}_c$ for each cluster,
- $\boldsymbol{\omega} = (\boldsymbol{\omega}_k)_{k\in\mathcal{K}}$ the regression parameters for temporal evolution data $\boldsymbol{x}_t$ for each cluster,
- σ^2 the variance of the measurement noise related to temporal evolution data $\boldsymbol{x}_t$.

Recalling that $p(z_j = k) = a_k$ where $\boldsymbol{a} = (a_k)_{k\in\mathcal{K}}$ are the weights related to component distributions, the mixture model is obtained from (4.11) such that $\forall j \in \mathcal{J}$,

$$p(\boldsymbol{x}_j, u_j|t_j, \boldsymbol{\Theta}) = \sum_{k\in\mathcal{K}} a_k \mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right) \mathcal{G}(u_j|\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \tag{4.12}$$

where $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{\omega}, \sigma^2, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the set of parameters.

Bayesian Framework

As in chapters 2 and 3, a Bayesian framework is used to estimate parameters $\boldsymbol{\Theta}$. Assuming datasets $(\boldsymbol{x} = (\boldsymbol{x}_q, \boldsymbol{x}_c, \boldsymbol{x}_t), \boldsymbol{t})$ of i.i.d observations $(\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}), t_j)_{j\in\mathcal{J}}$, independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$ and scale latent variables $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$, the complete likelihood associated to (4.12) is defined by

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \left(a_k \mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right) \mathcal{G}(u_j|\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \right)^{\delta_{z_j}^k}$$

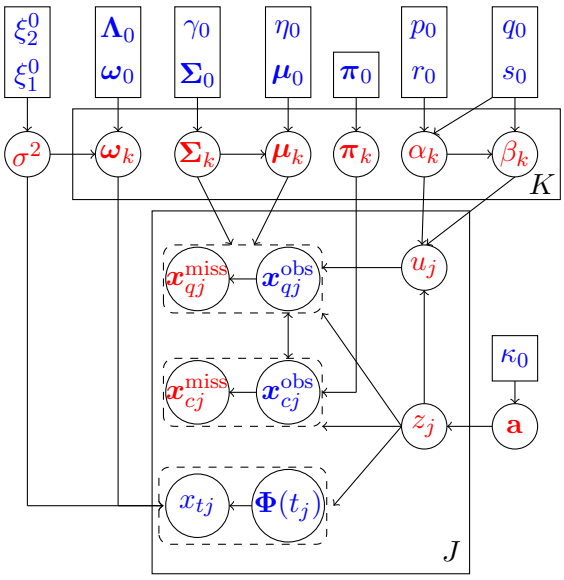


Figure 4.3: Graphical representation of the proposed model integrating temporal evolution data and mixed-type data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}, t_j)$ decomposed into temporal evolution data (x_{tj}, t_j) and mixed-type data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$, the scale variables u_j and the indicator variables z_j. Known quantities, respectively unknown quantities, are in blue, respectively in red.

Eventually, the prior distribution required for $\boldsymbol{\Theta}$ is chosen as

$$p(\boldsymbol{\Theta}|\mathcal{K}) = p(\boldsymbol{a}|\mathcal{K})p(\boldsymbol{\omega}|\sigma^2,\mathcal{K})p(\sigma^2)p(\boldsymbol{\pi}|\mathcal{K})p(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{K})p(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})$$

where

$$\begin{cases} p(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\boldsymbol{\kappa}_0) \ , \\ p(\boldsymbol{\pi}|\mathcal{K}) = \displaystyle\prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{\pi}_k|\boldsymbol{\pi}_0) \ , \\ p(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K}) = \displaystyle\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}_q} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\boldsymbol{\mu}_0,\eta_0^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{IW}(\boldsymbol{\Sigma}_k|\gamma_0,\boldsymbol{\Sigma}_0) \ , \\ p(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{K}) = \displaystyle\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}_q} p(\alpha_{k\boldsymbol{c}},\beta_{k\boldsymbol{c}}|p_0,q_0,s_0,r_0) \ , \\ p(\boldsymbol{\omega}|\sigma^2,\mathcal{K}) = \displaystyle\prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\boldsymbol{\omega}_0,\sigma^2\boldsymbol{\Lambda}_0\right) \ , \\ p(\sigma^2) = \mathcal{IG}(\sigma^2|\xi_1^0,\xi_2^0) \ . \end{cases}$$

Graphical representation of the proposed model is shown in Figure 4.3.

Inference

As previously, a factorized posterior distribution $q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K}) = q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}|\mathcal{K})q(\boldsymbol{\Theta}|\mathcal{K})$ is chosen as an approximation of the intractable posterior joint distribution $p(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{x}, \boldsymbol{t}, \mathcal{K})$ such that latent variables $\boldsymbol{h} = (\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z})$ and parameters $\boldsymbol{\Theta}$ are a posteriori independent and

$$\begin{aligned} q(\boldsymbol{h}|\mathcal{K}) &= q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K})q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z},\mathcal{K})q(\boldsymbol{z}|\mathcal{K}) \ , \\ q(\boldsymbol{\Theta}|\mathcal{K}) &= q(\boldsymbol{a}|\mathcal{K})q(\boldsymbol{\omega}|\sigma^2,\mathcal{K})q(\sigma^2)q(\boldsymbol{\pi}|\mathcal{K})q(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{K})q(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K}) \ . \end{aligned}$$

According to VB assumptions, the following conjugate variational posterior distributions are obtained from the VB procedure

$$
\begin{cases}
q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{miss}}|\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}}, u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k} , \\
q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}},\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{G}\left(u_j|\tilde{\alpha}_{jk\boldsymbol{c}},\tilde{\beta}_{jk\boldsymbol{c}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k} , \\
q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{MC}(\boldsymbol{x}_{cj}^{\text{miss}}|\tilde{\boldsymbol{r}}_{jk}^{\boldsymbol{x}_c^{\text{miss}}})^{\delta_{z_j}^k} \\
q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j) , \\
q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}}) , \\
q(\boldsymbol{\pi}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{\pi}|\tilde{\boldsymbol{\pi}}_k) , \\
q(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}},\tilde{\eta}_{k\boldsymbol{c}}^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k,\tilde{\boldsymbol{\Sigma}}_k) , \\
q(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{K}) = \prod_{k\in\mathcal{K}} p(\alpha_{k\boldsymbol{c}},\beta_{k\boldsymbol{c}}|\tilde{p}_k,\tilde{q}_k,\tilde{s}_k,\tilde{r}_k) , \\
q(\boldsymbol{\omega}|\sigma^2,\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\tilde{\boldsymbol{\omega}}_k,\sigma^2\tilde{\boldsymbol{\Lambda}}_k\right) , \\
q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1,\tilde{\xi}_2) .
\end{cases}
$$

Their respective parameters are estimated during the VBE and VBM steps by developing expectations $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right]$ and $\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right]$. Noting that

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] =& \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t|\boldsymbol{t},\boldsymbol{z},\boldsymbol{\omega},\sigma^2,\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_q,\boldsymbol{u},\boldsymbol{x}_c|\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma},\mathcal{K})\right] \\
&+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{z}|\boldsymbol{a},\mathcal{K})\right] ,
\end{aligned}
\tag{4.13}
$$

and

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right] =& \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t,\boldsymbol{\omega},\sigma^2|\boldsymbol{t},\boldsymbol{z},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_q,\boldsymbol{u},\boldsymbol{x}_c,\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{z},\mathcal{K})\right] \\
&+ \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{z},\boldsymbol{a}|\mathcal{K})\right] ,
\end{aligned}
\tag{4.14}
$$

the VBE (4.13) and VBM (4.14) steps can be independently derived for latent variables and parameters related to temporal evolution data $\boldsymbol{x}_t$ and mixed data $(\boldsymbol{x}_q,\boldsymbol{x}_c)$. Therefore, variational posterior distributions of latent variables $(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}})$ and parameters $(\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})$ related to mixed data $(\boldsymbol{x}_q,\boldsymbol{x}_c)$ are obtained as in Chapter 3 by deriving green expectations in (4.13) and (4.14). As for $(\boldsymbol{\omega},\sigma^2)$, their variational posterior distribution are obtained as in subsection 4.1.2 by developing blue expectations in (4.13) and (4.14). As in subsection 4.1.2 or in Chapter 3, the Dirichlet posterior distribution of $\boldsymbol{a}$ is deduced from the red expectation in (4.14). Eventually, the variational distribution of labels $\boldsymbol{z}$ is obtained by marginalising over latent variables in the green expectation and developing both blue and red expectations in (4.13) such that

$$
\int \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right]\partial\boldsymbol{x}_q^{\text{miss}}\partial\boldsymbol{u}\partial\boldsymbol{x}_c^{\text{miss}} = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\delta_{z_j}^k\log\rho_{jk}
$$

where $\log \rho_{jk}$ is deduced from red, blue and green expectations in (4.13) as follows

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \log \rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log \rho^t_{jk} + \log \rho^{qc}_{jk}\ . \tag{4.15}$$

with

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] &= \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k\in\mathcal{K}} \tilde{\kappa}_k\right)\ , \\
\log \rho^t_{jk} &= -\frac{1}{2}\left(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\right]\right)\ , \\
\log \rho^{qc}_{jk} &= \log \sum_{\boldsymbol{c}^{\text{miss}} \in \mathcal{C}_{q_j^{\text{miss}}}} \rho_{k\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_{cj}^{\text{miss}}}\ .
\end{aligned}$$

The red term $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right]$ is deduced from properties of the Dirichlet distribution, the blue term $\log \rho^t_{jk}$ is deduced from (4.7) and (4.8) in subsection 4.1.2 and the green term $\log \rho^{qc}_{jk}$ has been detailed in Chapter 3. Hence, $\boldsymbol{z}$ is distributed a posteriori according to a product of Categorical distributions parametrized by $\tilde{\boldsymbol{r}} = (\tilde{r}_{jk})_{(j,k)\in\mathcal{J}\times\mathcal{K}}$ given by

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \tilde{r}_{jk} = \frac{\rho_{jk}}{\sum\limits_{k\in\mathcal{K}} \rho_{jk}} \tag{4.16}$$

The lower bound on the log evidence is still required to master the VB inference and can be also decomposed into terms related to temporal evolution data (blue terms), mixed data (green terms) and labels $\boldsymbol{z}$ (red terms). This decomposition is obtained as follows

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{u},\boldsymbol{z},\boldsymbol{\Theta}|\mathcal{K})\right]$$

where the free energy can be developed as

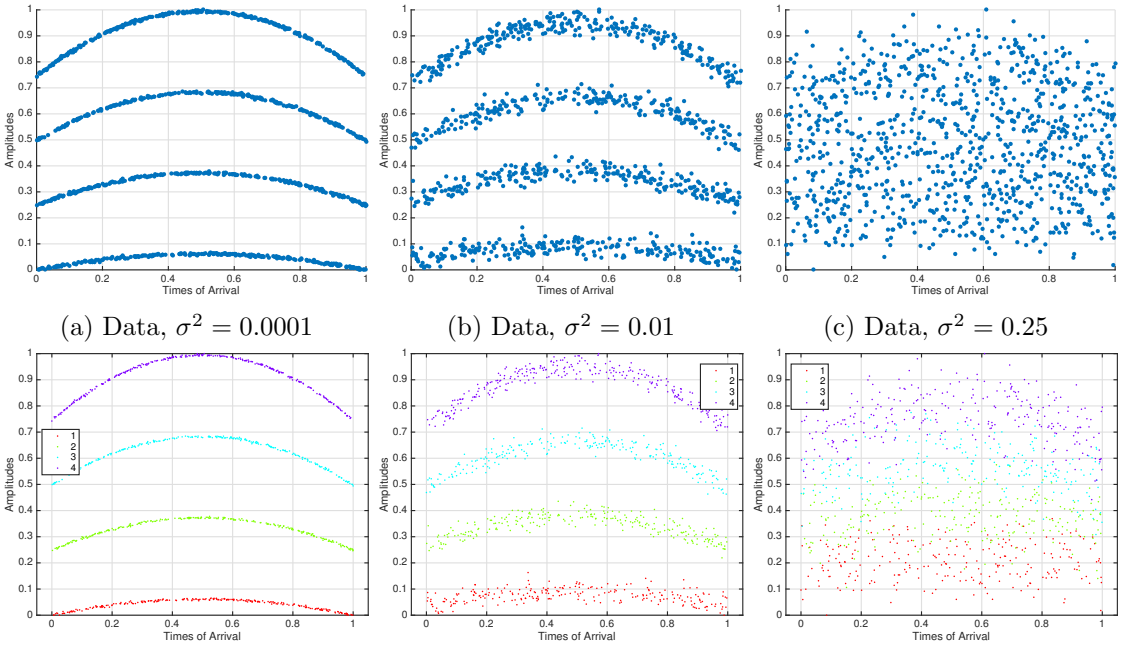
$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\omega},\sigma^2|\mathcal{K})\right] \\
+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{a},\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})\right]
\end{aligned}$$

and the entropy as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{u},\boldsymbol{z},\boldsymbol{\Theta}|\mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\omega},\sigma^2|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{u}|\boldsymbol{z},\mathcal{K})\right] \\
+ \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{a},\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})\right]\ .
\end{aligned}$$

Blue terms, respectively green terms, have been previously detailed in subsection 4.1.2 , respectively in Chapter 3. As for red terms, they are detailed below :

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta^k_{z_j}\right] \log \rho_{jk}\ , \\
\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta^k_{z_j}\right] \log \tilde{r}_{jk}\ .
\end{aligned}$$



(a) Data, $\sigma^2 = 0.0001$ (b) Data, $\sigma^2 = 0.01$ (c) Data, $\sigma^2 = 0.25$

(d) Ground-Truth, $\sigma^2 = 0.0001$ (e) Ground-Truth, $\sigma^2 = 0.01$ (f) Ground-Truth, $\sigma^2 = 0.25$

Figure 4.4: Synthetic parabolic data generated from different values of the variance parameter σ^2. Figures (a), (b) and (c) present unlabeled data where 4 parabolas are generated. Ground-truth are visible on Figures (d), (e) and (f).

4.1.4 Experiments

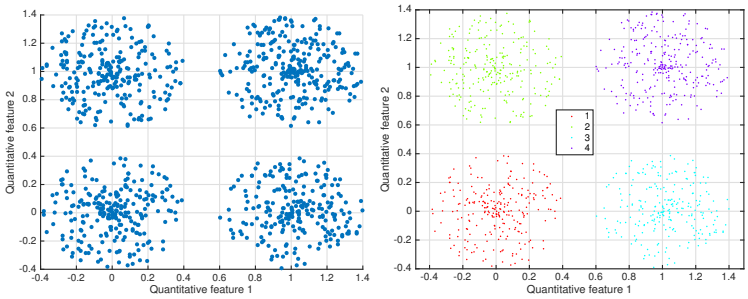
Two experiments are carried out to evaluate clustering performance with respect to a set of synthetic data and a set of real data. In the first experiment, only temporal evolution data are taken into consideration in the clustering procedure. Then, both temporal evolution data and quantitative data are considered in the second one. For comparison, the spectral clustering [VL07] and the k-means algorithm from [HW79] are also evaluated. First, characteristics of data, comparison algorithms and evaluation metrics are detailed. Then, both experiments are described and performance are shown to exhibit the effectiveness of the proposed model.

Data, algorithms and metrics

Both synthetic and real data are composed of temporal evolution data related to amplitudes which are distributed according to a parabolic relation and quantitative data related to continuous radar features which are jointly distributed according to a multivariate normal distribution. In synthetic data, temporal evolution data are generated by sampling a set of data from four parabolas directed by

$$\boldsymbol{\omega} = \begin{pmatrix} -1 & -2 & -3 & -4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

and quantitative data are generated by sampling a set of data from four well-separated bivariate clusters with centers $[0,0]^T$, $[1,0]^T$,$[0,1]^T$ and $[1,1]^T$ and identity covariance matrices. Three synthetic datasets are generated with respect to a range of values of σ^2. These datasets are shown in Figures 4.4 and 4.5 where each radar emitter is represented by a parabola (Figure 4.4) and a Gaussian cluster (Figure 4.5). Real data are extracted from operational recordings which



(a) Quantitative Data (b) Ground-Truth

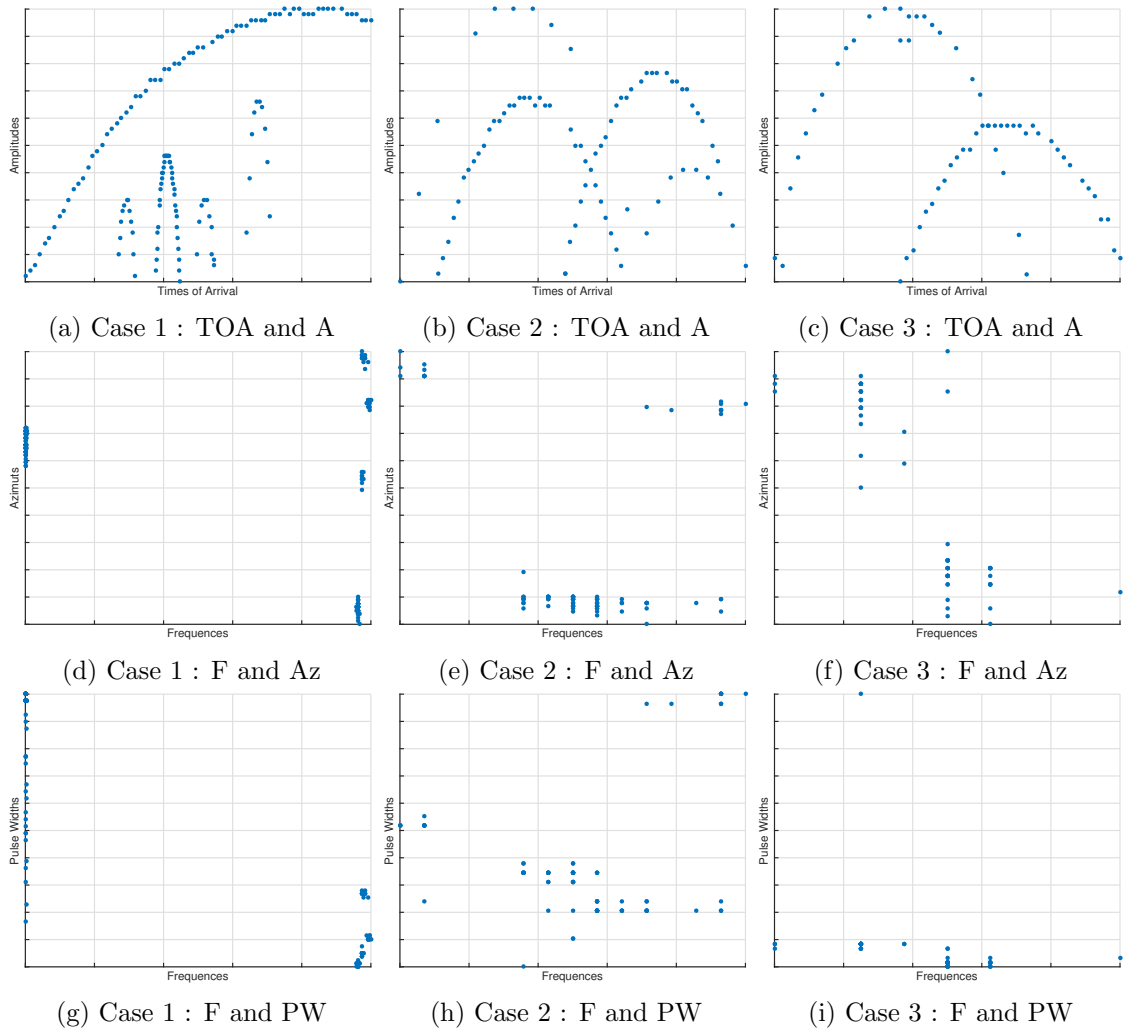
Figure 4.5: Synthetic quantitative data generated from 4 multivariate normal distributions. Figure (a) shows unlabeled data and Figure (b) exhibits the ground-truth.

include unlabeled pulses that are mainly described by their Pulse Description World (PDW) composed of

- Time Of Arrival (TOA) ,
- Amplitude (A) ,
- Radio Frequency (RF) ,
- Pulse Width (PW) ,
- Azimuth (Az) .

Therefore, temporal evolution data are pairs (TOA,A) and quantitative data are triplets (RF,PW,Az). Unfortunately, these real data are classified and values of PDW cannot be released. Hence, axes of figures related to real data are not displayed. Three cases are obtained from real recordings and they are visible on Figure 4.6. Different numbers of parabolas and (RF,PW,Az) clusters can be observed according to a chosen real case. These differences result in different numbers of emitters for the 3 cases such that 5, 4 and 2 emitters are identified in the cases 1, 2 and 3. Eventually, synthetic and real data are linearly transformed by a min-max normalization to meet algorithms requirements.

Except for the k-means algorithm, an initialisation is required for clustering algorithms that are involved in these experiments. The similarity graph required for the spectral clustering is obtained from a k-nearest neighbor graph as suggested in [VL07] where the number of neighbors k is chosen as the product of the log number of observations and the number of clusters. As for the proposed model, a supervised initialisation is retained due to its sensitivity to initialisation. First, prior hyperparameters ξ_1^0 and ξ_2^0 are initialised such that the prior mean $\mathbb{E}[\frac{1}{\sigma^2}] = \frac{\xi_1^0}{\xi_2^0}$ of the variance parameter σ^2 is equal to the inverse of the determinant of the covariance matrix of temporal evolution data points. This choice is motivated by the fact that the determinant of the covariance matrix can be interpreted as the generalized variance that reflects the overall spread of the data. Setting $\xi_2^0 = 1$, ξ_1^0 is initialised as the inverse of the generalized variance of the sample of temporal evolution data. Then, prior component means $\boldsymbol{\mu}_0$, respectively covariance matrices $\boldsymbol{\Sigma}_0$, are initialised from results of a k-means algorithm on quantitative data, respectively from diagonal matrices whose diagonal elements are variances of quantitative data. Other hyper-parameters are initialised as in Table 4.1.



(a) Case 1 : TOA and A (b) Case 2 : TOA and A (c) Case 3 : TOA and A

(d) Case 1 : F and Az (e) Case 2 : F and Az (f) Case 3 : F and Az

(g) Case 1 : F and PW (h) Case 2 : F and PW (i) Case 3 : F and PW

Figure 4.6: Real data obtained from 3 operational cases. Figures (a), (b) and (c) exhibit distributions of Times of Arrival (TOA) and Amplitudes (A). Figures (d), (e) and (f) exhibit distributions of Frequencies (F) and Azimuths (Az). Figures (g), (h) and (i) exhibit distributions of Frequencies (F) and Pulse Widths (PW).

Performance on synthetic data are evaluated through the Adjusted Rand Index (ARI) [HA85] that compares estimated partitions of data with the ground-truth and the Silhouette Coefficient [KR09] which does not require the ground-truth and provides a higher score when clusters are dense and well separated. Performance on real data are only evaluated through the Silhouette score since the ground-truth is not available for each case.

Experiments and results

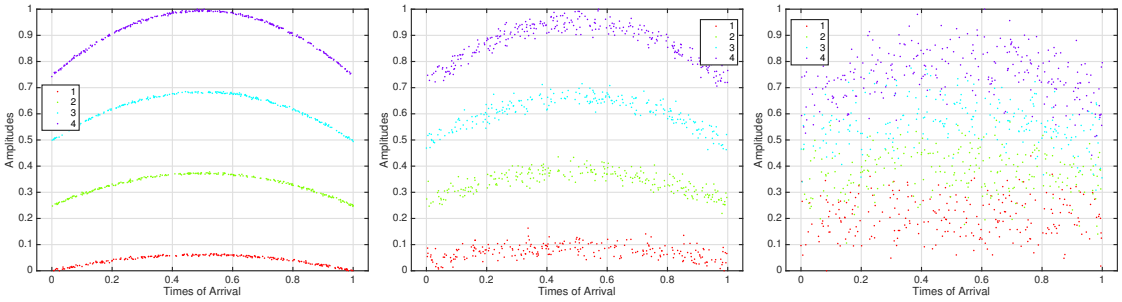
The first experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters K when only temporal evolution data are taken into consideration. According to datasets visualised in Figure 4.4 and Figure 4.6, K is set to 4 for synthetic data and to 5, 4 and 2 for the three real cases. Results of the first experiment on synthetic data are shown in Figure 4.7 and in Table 4.2. The proposed model and the spectral clustering succeed in clustering synthetic data for $\sigma^2 \in \{0.0001, 0.01\}$ since the ground-truth partition is recovered in Figure 4.7 with an ARI equals to 1 visible on Table 4.2. The lower performance of

Table 4.1: Initialisation of hyper-parameters values for clustering on parabolic data

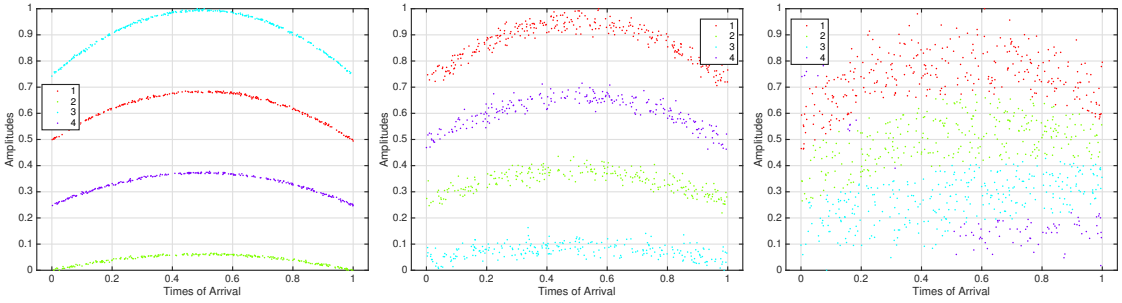
$\boldsymbol{\omega}_0$	$\boldsymbol{\Lambda}_0$	κ_0	η_0	γ_0	p_0	r_0	q_0	s_0
$(0,0,0)^T$	I_3	0.5	100	1	1	1	1	1

the k-means algorithm (ARI = 0.33) can be explained by the fact that the k-means algorithm creates convex and isotropic clusters that cannot handle the parabolic structure of the generated data. This limitation is emphasized by higher Silhouette Coefficients of the k-means algorithm since the Silhouette Coefficient is generally higher for convex clusters. Moreover, the lower Silhouette Coefficients of the ground-truth for $\sigma^2 \in \{0.0001, 0.01\}$ confirm the non-convexity of the data. Even if all algorithms poorly perform when data are embedded in noise ($\sigma^2 = 0.25$), the proposed algorithm estimates clusters with a more parabolic shape than other algorithms which build more isotropic clusters (Subfigures (f), (i) and (l) in Figure 4.2). Indeed the Silhouette Coefficient of the proposed model ($S = 0.14$) is closer to the Silhouette Coefficient of the ground-truth ($S = 0.10$) than Silhouette Coefficients of spectral clustering ($S = 0.53$) and k-means ($S = 0.56$). Results of the first experiment on real data are shown in Figure 4.8 and in Table 4.3. Interpretation of algorithm performance through Silhouette coefficients and visual representations is complex since the Silhouette coefficient enhances algorithms that create convex clusters whereas the visual representations of estimated clusters tend to choose algorithms that create clusters with a parabolic shape. As in the example of real case 1, the proposed model succeeds in finding the radar emitter whose scanning behaviour is described by the red parabola (Subfigure (d) in Figure 4.8) whereas spectral clustering and k-means find that this parabola belongs to many emitters (Subfigures (g) and (j) in Figure 4.8). Nonetheless, spectral clustering and k-means provide higher Silhouette Coefficient ($S = 0.46$ and $S = 0.57$) than the proposed model ($S = 0.05$).

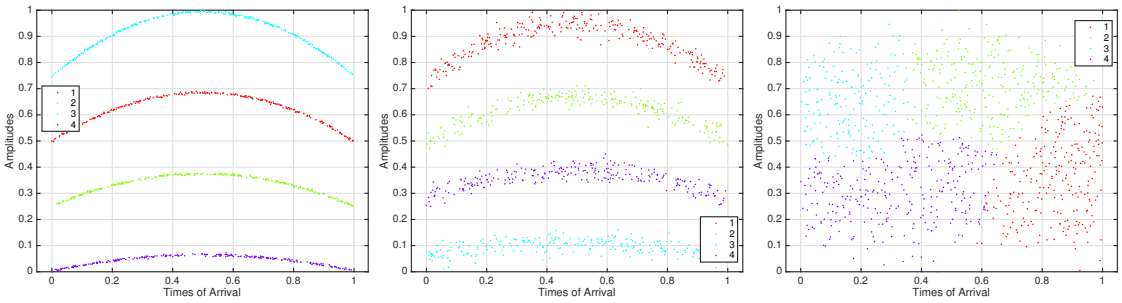
The second experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters K when all types of data are taken into consideration. The number of clusters K is still set to 4 for synthetic data and to 5, 4 and 2 for the three real cases. Results of the second experiment on synthetic data are shown in Figure 4.9 and in Table 4.4. All algorithms succeed in clustering synthetic data for $\sigma^2 \in \{0.0001, 0.01, 0.25\}$ since the ground-truth partition is recovered in Figure 4.9 with an ARI equals to 1 visible on Table 4.4. Adding quantitative information enables algorithms to recover the ground-truth for any value of σ^2. Results of the second experiment on real data are shown in Figure 4.10 and in Table 4.5. The proposed model perfectly estimates clusters for the three cases whereas spectral clustering and k-means cannot manage to recover the correct clusters in real case 1 (Subfigures (d), (g) and (j)). Indeed, the proposed model finds the five different emitters by exploiting quantitative features while spectral clustering and k-means cannot identify the emitter whose temporal evolution features are distributed according to the parabola which is slowly increasing (red parabola in Subfigure (d)). Nonetheless, k-means and spectral clustering have higher Silhouette coefficients than the proposed model since they provide more convex clusters than the proposed method.



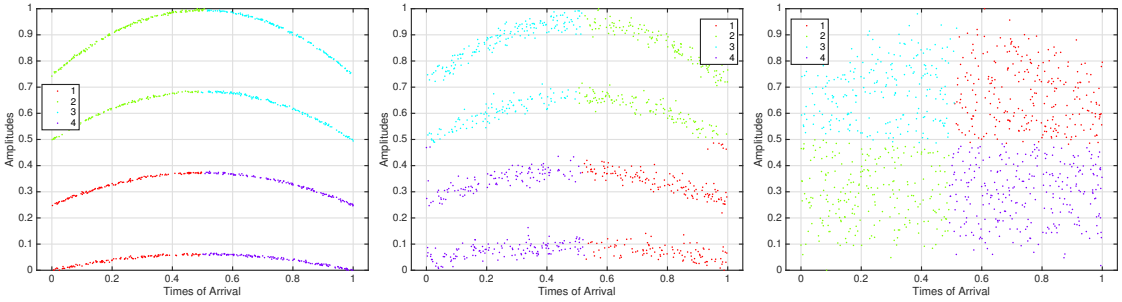
(a) Ground-Truth, $\sigma^2 = 0.0001$ (b) Ground-Truth, $\sigma^2 = 0.01$ (c) Ground-Truth, $\sigma^2 = 0.25$



(d) PM, $\sigma^2 = 0.0001$ (e) PM, $\sigma^2 = 0.01$ (f) PM, $\sigma^2 = 0.25$



(g) SC, $\sigma^2 = 0.0001$ (h) SC, $\sigma^2 = 0.01$ (i) SC, $\sigma^2 = 0.25$



(j) KM, $\sigma^2 = 0.0001$ (k) KM, $\sigma^2 = 0.01$ (l) KM, $\sigma^2 = 0.25$

Figure 4.7: Results on synthetic data during the first experiment when only temporal evolution data are considered. Figures (a), (b) and (c) show synthetic data generated with different values of the variance σ^2. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).

Table 4.2: Adjusted Rand Index (ARI) and Silhouette coefficient (S) values for the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the first experiment on synthetic data when only temporal evolution data are considered.

	ARI			S			
	PM	SC	KM	Data	PM	SC	KM
$\sigma^2 = 0.0001$	1	1	0.34	0.32	0.32	0.32	0.63
$\sigma^2 = 0.01$	1	1	0.33	0.27	0.27	0.27	0.63
$\sigma^2 = 0.25$	0.42	0.23	0.27	0.10	0.14	0.53	0.56

Table 4.3: Silhouette coefficients of the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the first experiment on real data when only temporal evolution data are considered.

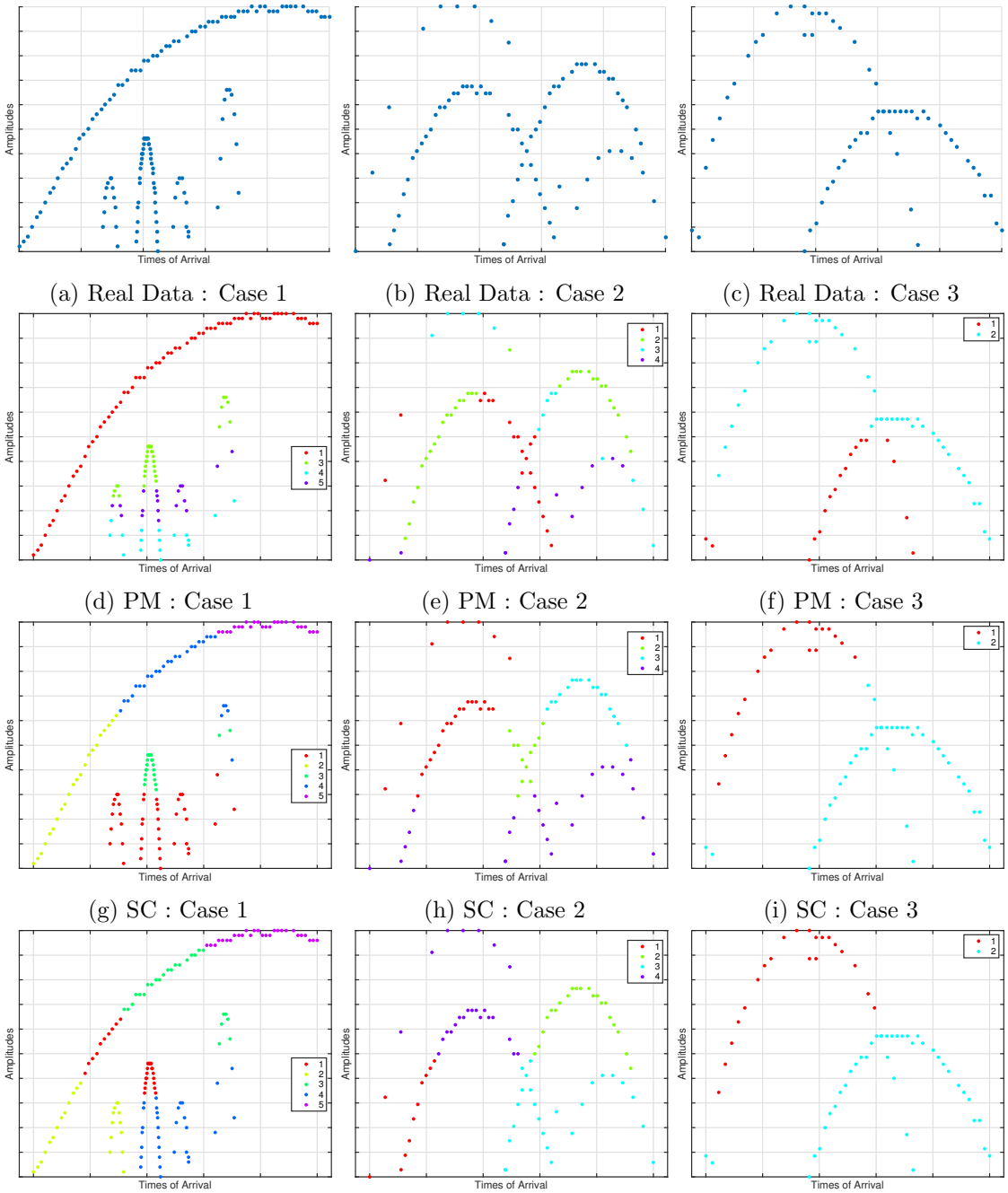
	Silhouette Coefficient		
	PM	SC	KM
Case 1	0.05	0.46	0.57
Case 2	0.12	0.28	0.57
Case 3	0.26	0.61	0.61

Table 4.4: Adjusted Rand Index (ARI) and Silhouette coefficient (S) values for the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the experiment on synthetic data when all types of data are considered.

	ARI			S			
	PM	SC	KM	Data	PM	SC	KM
$\sigma^2 = 0.0001$	1	1	1	0.76	0.76	0.76	0.76
$\sigma^2 = 0.01$	1	1	1	0.75	0.75	0.75	0.75
$\sigma^2 = 0.25$	1	1	1	0.73	0.73	0.73	0.73

Table 4.5: Silhouette coefficients of the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the second experiment on real data when all types of data are considered.

	Silhouette Coefficient		
	PM	SC	KM
Case 1	0.77	0.75	0.81
Case 2	0.57	0.32	0.61
Case 3	0.71	0.73	0.74



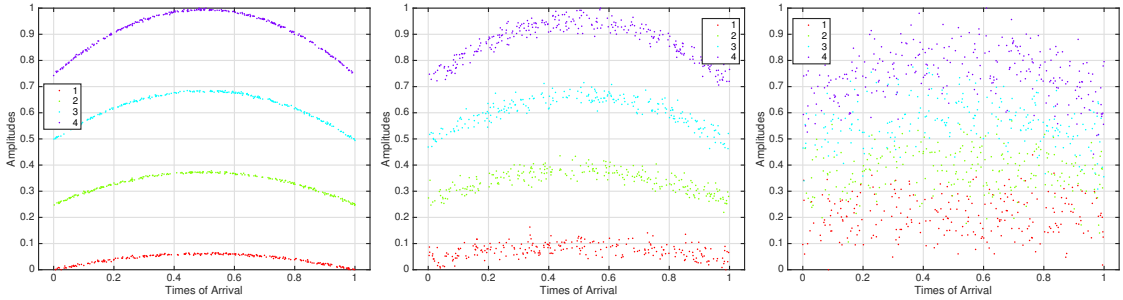
(a) Real Data : Case 1 (b) Real Data : Case 2 (c) Real Data : Case 3

(d) PM : Case 1 (e) PM : Case 2 (f) PM : Case 3

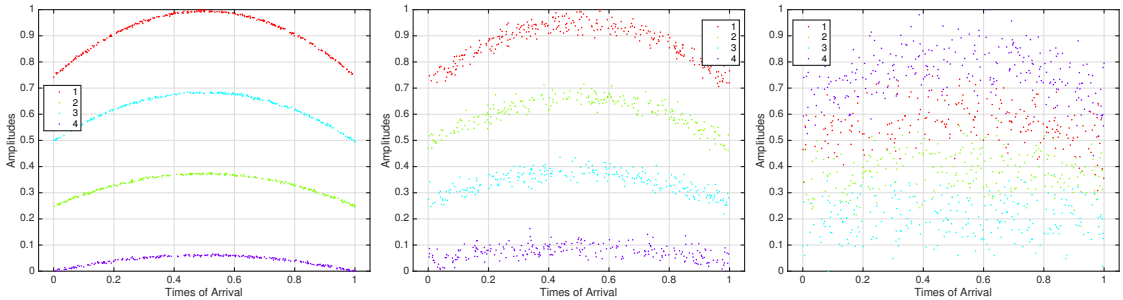
(g) SC : Case 1 (h) SC : Case 2 (i) SC : Case 3

(j) KM : Case 1 (k) KM : Case 2 (l) KM : Case 3

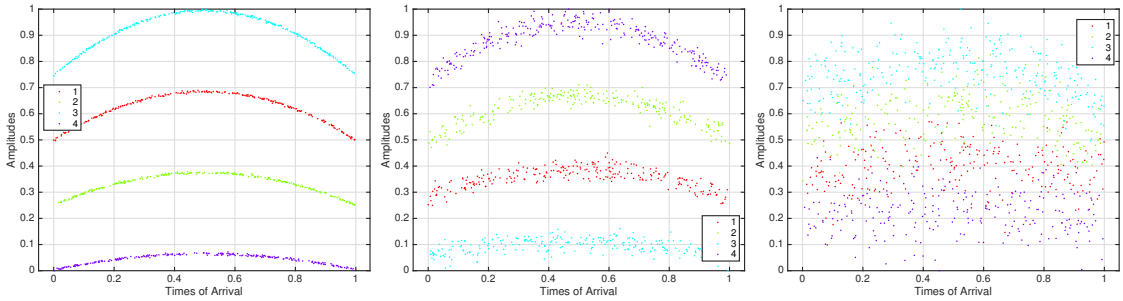
Figure 4.8: Results on real parabolic data during the first experiment when only temporal evolution data are considered. Figures (a), (b) and (c) show real data in different cases. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).



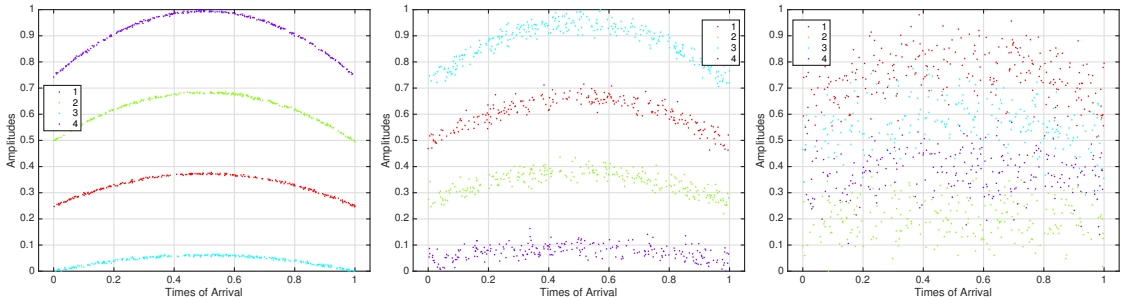
(a) Ground-Truth, $\sigma^2 = 0.0001$ (b) Ground-Truth, $\sigma^2 = 0.01$ (c) Ground-Truth, $\sigma^2 = 0.25$



(d) PM, $\sigma^2 = 0.0001$ (e) PM, $\sigma^2 = 0.01$ (f) PM, $\sigma^2 = 0.25$

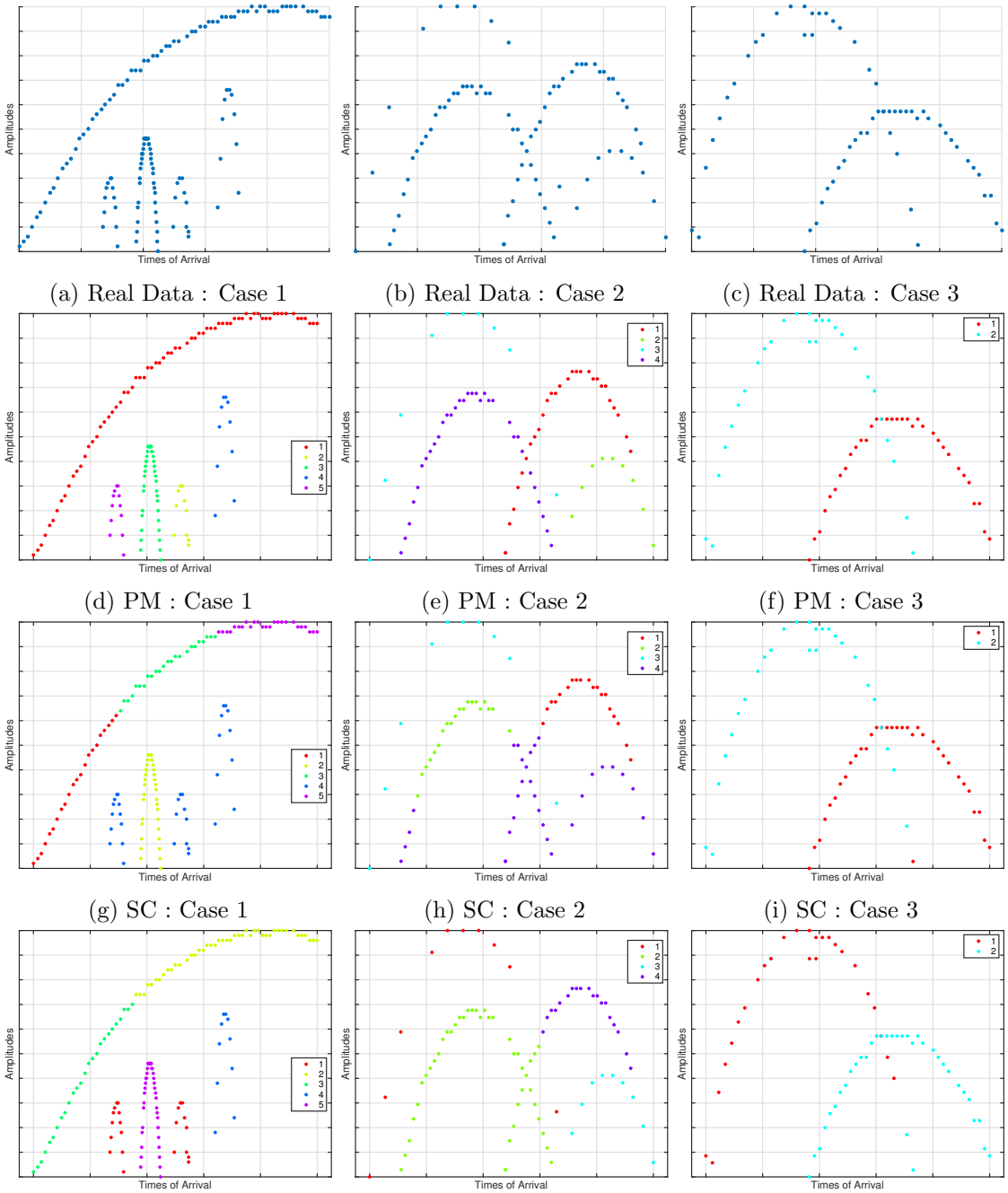


(g) SC, $\sigma^2 = 0.0001$ (h) SC, $\sigma^2 = 0.01$ (i) SC, $\sigma^2 = 0.25$



(j) KM, $\sigma^2 = 0.0001$ (k) KM, $\sigma^2 = 0.01$ (l) KM, $\sigma^2 = 0.25$

Figure 4.9: Results on synthetic parabolic data when all types of data are considered. Figures (a), (b) and (c) show synthetic data generated with different values of the variance σ^2. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).



(a) Real Data : Case 1 (b) Real Data : Case 2 (c) Real Data : Case 3

(d) PM : Case 1 (e) PM : Case 2 (f) PM : Case 3

(g) SC : Case 1 (h) SC : Case 2 (i) SC : Case 3

(j) KM : Case 1 (k) KM : Case 2 (l) KM : Case 3

Figure 4.10: Results on real parabolic data when any types of data are considered. Figures (a), (b) and (c) show real data in different cases. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).

4.2 Piecewise parabolic data

In this section, K emitters presenting piecewise parabolic scanning behaviours are considered. Therefore, the main objective is to develop a mixture model which can build K distinct clusters formed by K piecewise parabolas with P piecewises and to define an inference procedure for parameter estimation. Then, the proposed model is enhanced with the mixture model designed for mixed data in Chapter 3 in order to improve clustering performance. Finally, experiments on synthetic data are carried out to exhibit performance of the proposed approach.

4.2.1 Model

Before introducing a mixture model that handles piecewise parabolic data, the piecewise parabolic relation between amplitudes and times of arrival is defined. Then, the mixture model is developed into a Bayesian framework.

Piecewise parabola equation

The piecewise parabolic relation between amplitudes $(x_{tj})_{j\in\mathcal{J}}$ and times of arrival $(t_j)_{j\in\mathcal{J}}$ gives form to a set of P piecewises of constant amplitudes $(\mu_p^t)_{p\in\mathcal{P}}$ that are linked by a parabolic relation (Figure 4.11). Each piecewise μ_p^t gathers pulses $(x_{tj}, t_j)_{j\in\mathcal{J}_p}$ whose amplitude x_{tj} is equal to μ_p^t and where $\mathcal{J}_p$ is the set of indexes of pulses that belong to the p^{th} piecewise. These P sets $\mathcal{J}_p$ of pulses are disjoint and constitute a partition of $\mathcal{J}$ such $\bigcup_p \mathcal{J}_p = \mathcal{J}$. Finally, the piecewises $(\mu_p^t)_{p\in\mathcal{P}}$ belong to a parabola parametrized by $\boldsymbol{\omega}$ and $\left(\min_{j\in\mathcal{J}_p} t_j\right)_{p\in\mathcal{P}}$ which are the times of the first pulses belonging to the piecewises. That definition can be translated into the following system

$$\forall j \in \mathcal{J}_p, \begin{cases} x_{tj} = \mu_p^t + \epsilon \\ \mu_p^t = \boldsymbol{\Phi}\left(\min_{j\in\mathcal{J}_p} t_j\right)^T \boldsymbol{\omega} \end{cases} \tag{4.17}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a measurement noise introduced to model defects of materials, $\mathcal{J}_p = \{j \in \mathcal{J}, |x_{tj} - \mu_p^t| \leq \sigma\}$ and $\boldsymbol{\Phi}(\cdot)$, respectively $\boldsymbol{\omega}$, are the polynomial transformation, respectively the regression parameter, defined in (4.2). Since the measurement error is assumed to be Gaussian, amplitude $(x_{tj})_{j\in\mathcal{J}_p}$ are distributed according to a normal distribution centered in μ_p^t with variance σ^2

$$\forall j \in \mathcal{J}_p,\ x_{tj} \sim \mathcal{N}\left(x_{tj}|\mu_p^t, \sigma^2\right)\ . \tag{4.18}$$

If $(\mathcal{J}_p)_{p\in\mathcal{P}}$ and $(\mu_p^t)_{p\in\mathcal{P}}$ are known, the regression parameter $\boldsymbol{\omega}$ is the solution of a linear problem given by

$$\boldsymbol{\mu}^t = \boldsymbol{\Phi}(\mathcal{X})^T\boldsymbol{\omega}$$

where $\boldsymbol{\Phi}(\mathcal{X})$ is a $3 \times P$ matrix whose columns are the P polynomial transformations $\left(\boldsymbol{\Phi}\left(\min_{j\in\mathcal{J}_p} t_j\right)\right)_{p\in\mathcal{P}}$ and $\boldsymbol{\mu}^t = (\mu_1^t, \ldots, \mu_P^t)^T$. Then the regression parameter $\boldsymbol{\omega}$ is obtained such that

$$\boldsymbol{\omega} = \left(\boldsymbol{\Phi}(\mathcal{X})\boldsymbol{\Phi}(\mathcal{X})^T\right)^{-1}\boldsymbol{\Phi}(\mathcal{X})\boldsymbol{\mu}^t\ .$$

Mixture model

Since each radar emitter has its own scanning behaviour, K unique piecewise parabolas exist in data and they are configured with K regression parameters $\boldsymbol{\omega} = (\boldsymbol{\omega}_k)_{k\in\mathcal{K}}$ and K sets of piecewises $\boldsymbol{\mu}^t = (\mu_{kp}^t)_{(p,k)\in\mathcal{P}\times\mathcal{K}}$. Then, each amplitude x_{tj} belongs to one of these sets of piecewises which is

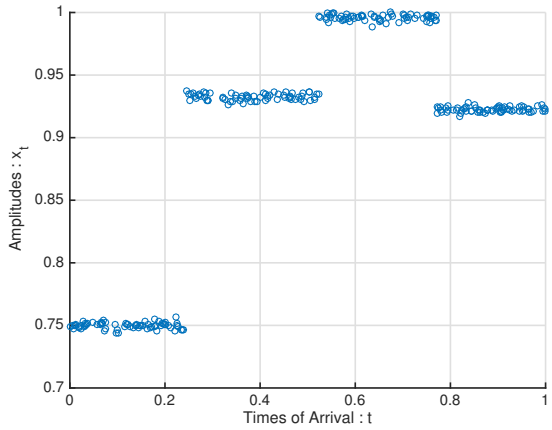


Figure 4.11: Simulated data where amplitudes $\boldsymbol{x}_t$ and times of arrival $\boldsymbol{t}$ are distributed according to a piecewise parabolic relation defined with $P = 4$ piecewises.

related to a specific emitter. In other words, conditionally to its label z_j and its affiliations to one of the piecewises, the amplitude x_{tj} is distributed according to (4.18) such that the component distribution is defined by

$$\forall j \in \mathcal{J}_p,\ x_{tj}|z_j = k \sim \mathcal{N}\left(x_{tj}|\mu_{kp}^t, \sigma^2\right) \tag{4.19}$$

In order to model its affiliations to one of the piecewises, a latent discrete variable y_j is introduced such that (4.19) becomes

$$\forall j \in \mathcal{J},\ x_{tj}|y_j = p, z_j = k \sim \mathcal{N}\left(x_{tj}|\mu_{kp}^t, \sigma^2\right) \tag{4.20}$$

where y_j belongs to $\mathcal{P}$ and follows, conditionally to $z_j = k$, a categorical distribution with weights $\boldsymbol{b}_k = (b_{k1}, \ldots, b_{kP})$. Therefore the initial component distribution (4.19) can be reformulated as a mixture model such that

$$p(x_{tj}|z_j = k, \boldsymbol{\Theta}, \mathcal{K}) = \sum_{p\in\mathcal{P}} b_{kp}\mathcal{N}\left(x_{tj}|\mu_{kp}^t, \sigma^2\right)$$

Recalling that $p(z_j = k) = a_k$ where $\boldsymbol{a} = (a_k)_{k\in\mathcal{K}}$ are the weights related to component distributions, the proposed mixture model is a mixture of mixture models given by

$$\forall j \in \mathcal{J},\ p(x_{tj}|\boldsymbol{\Theta}) = \sum_{k\in\mathcal{K}} a_k \sum_{p\in\mathcal{P}} b_{kp}\mathcal{N}\left(x_{tj}|\mu_{kp}^t, \sigma^2\right) \tag{4.21}$$

where $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\mu}^t, \sigma^2)$ is the set of parameters.

Bayesian framework

As in previous chapters, a Bayesian framework is used to estimate parameters $\boldsymbol{\Theta}$. Assuming datasets $(\boldsymbol{x}_t, \boldsymbol{t})$ of i.i.d observations $(x_{tj}, t_j)_{j\in\mathcal{J}}$ and independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$ and $\boldsymbol{y} = (y_j)_{j\in\mathcal{J}}$ for clusters and piecewises, the complete likelihood associated to (4.21) is defined by

$$p(\boldsymbol{x}_t, \boldsymbol{h}|\boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\left(a_k \prod_{p\in\mathcal{P}}\left(b_{kp}\mathcal{N}\left(x_{tj}|\mu_{kp}^t, \sigma^2\right)\right)^{\delta_{y_j}^p}\right)^{\delta_{z_j}^k}$$

where $\boldsymbol{h} = (\boldsymbol{y}, \boldsymbol{z})$ is the set of latent variables. Eventually, the prior distribution required for $\boldsymbol{\Theta}$ is chosen as

$$p(\boldsymbol{\Theta}|\mathcal{K}) = p(\boldsymbol{a}|\mathcal{K})p(\boldsymbol{b}|\mathcal{K})p(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K})p(\sigma^2)$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ follow a Dirichlet distribution, each μ_{kp}^t follows a Normal distribution and σ^2 follows an Inverse Gamma distribution such that

$$\begin{cases} p(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\mathbf{a}|\kappa_0) \ , \\ p(\boldsymbol{b}|\mathcal{K}) = \prod\limits_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|o_0) \ , \\ p(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K}) = \prod\limits_{k\in\mathcal{K}} \prod\limits_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\mu_0^t, \tau_0^{-1}\sigma^2\right) \ , \\ p(\sigma^2) = \mathcal{IG}(\sigma^2|\xi_1^0, \xi_2^0) \ . \end{cases}$$

The resulting mixture model is shown on Figure 4.12.

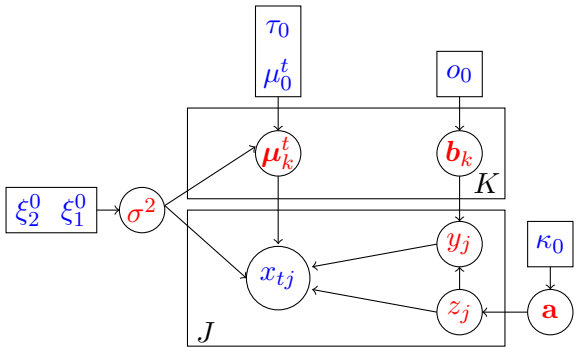


Figure 4.12: Graphical representation of the proposed mixture model handling piecewise parabolic data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed amplitudes x_j and the indicator variables (z_j, y_j). Known quantities, respectively unknown quantities, are in blue, respectively in red.

4.2.2 Inference

The Variational Bayes (VB) procedure is derived to estimate parameters of the mixture model defined in (4.21). Variational posterior distributions are obtained from the VB Expectation (VBE) and VB Maximization (VBM) steps and a lower bound on the log evidence is defined to master the convergence of the VB procedure.

Variational posterior distributions

As previously, a factorized posterior distribution $q(\boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K}) = q(\boldsymbol{h}|\mathcal{K})q(\boldsymbol{\Theta}|\mathcal{K})$ is chosen as an approximation of the intractable posterior joint distribution $p(\boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{x}_t, \mathcal{K})$ such that latent variables $\boldsymbol{h}$ and parameters $\boldsymbol{\Theta}$ are a posteriori independent and

$$\begin{aligned} q(\boldsymbol{h}|\mathcal{K}) &= q(\boldsymbol{y}|\boldsymbol{z}, \mathcal{K})q(\boldsymbol{z}|\mathcal{K}) \ , \\ q(\boldsymbol{\Theta}|\mathcal{K}) &= q(\boldsymbol{a}|\mathcal{K})q(\boldsymbol{b}|\mathcal{K})q(\boldsymbol{\mu}^t|\sigma^2\mathcal{K})q(\sigma^2) \ . \end{aligned}$$

According to VB assumptions, the following conjugate variational posterior distributions are obtained from the VB procedure

$$\begin{cases} q(\boldsymbol{y}|\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{C}at(y_j|\tilde{\boldsymbol{r}}_{jk}^{y})^{\delta_{z_j}^k} \ , \\ q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j) \ , \\ q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}}) \ , \\ q(\boldsymbol{b}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|\tilde{\boldsymbol{o}}_k) \ , \\ q(\boldsymbol{\mu}^t|\sigma^2,\mathcal{K}) = \prod_{k\in\mathcal{K}}\prod_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\tilde{\mu}_{kp}^t, \tilde{\tau}_{kp}^{-1}\sigma^2\right) \ , \\ q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1,\tilde{\xi}_2) \ . \end{cases}$$

Their respective parameters are estimated during the VBE and VBM steps.

VBE-step

The VBE-step consists in deriving the following expectation

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t,\boldsymbol{h}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \sum_{p\in\mathcal{P}} \delta_{y_j}^p \Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\Bigg(\log 2\pi \\ &\quad + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj}-\mu_{kp}^t)^2}{\sigma^2}\right]\Bigg)\Bigg)\Bigg) \\ &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta_{z_j}^k \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \sum_{p\in\mathcal{P}} \delta_{y_j}^p \log\rho_{jkp}^y \right) \end{aligned} \tag{4.22}$$

where

$$\log\rho_{jkp}^y = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\left(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj}-\mu_{kp}^t)^2}{\sigma^2}\right]\right) \ . \tag{4.23}$$

Hence, a categorical distribution for piecewise labels $\boldsymbol{y}$ is deduced from (4.22) conditionally to cluster labels $\boldsymbol{z}$ such that

$$q(\boldsymbol{y}|\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{C}at(y_j|\tilde{\boldsymbol{r}}_{jk}^{y})^{\delta_{z_j}^k}$$

and their parameters $(\tilde{\boldsymbol{r}}_{jk}^{y})_{j\in\mathcal{J}}$ are obtained from (4.23) as follows

$$\forall j\in\mathcal{J},\ \forall k\in\mathcal{K},\ \forall p\in\mathcal{P},\ \tilde{r}_{jkp}^{y} = \frac{\rho_{jkp}^y}{\sum_{p\in\mathcal{P}}\rho_{jkp}^y} \ .$$

Marginalising over $\boldsymbol{y}$ in (4.22), a categorical is obtained for cluster labels $\boldsymbol{z}$ such that

$$q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j)$$

with

$$\forall j\in\mathcal{J},\ \forall k\in\mathcal{K},\ \tilde{r}_{jk} = \frac{\rho_{jk}}{\sum_{k\in\mathcal{K}}\rho_{jk}} \ .$$

where

$$\rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log\sum_{p\in\mathcal{P}}\rho_{jkp}^y \ .$$

VBM-step

The VBM-step consists in deriving the following expectation

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] + \log p(\boldsymbol{\Theta}|\mathcal{K}) \\ &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\Bigg(\log a_k + \sum_{p\in\mathcal{P}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\Bigg(\log 2\pi \\ &+ \log\sigma^2 + \frac{(x_{tj} - \mu_{kp}^t)^2}{\sigma^2}\Bigg)\Bigg)\Bigg) + \sum_{k\in\mathcal{K}}(\kappa_k^0 - 1)\log a_k + \log c_{\mathcal{D}}(\boldsymbol{\kappa}^0) \\ &- (\xi_1^0 + 1)\log\sigma^2 - \frac{\xi_2^0}{\sigma^2} + \log c_{\mathcal{IG}}(\xi_1^0, \xi_2^0) + \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}}(o_{kp}^0 - 1)\log b_{kp} \\ &- \frac{1}{2}\left(\log 2\pi + \log\sigma^2 + \frac{\tau_0}{\sigma^2}(\mu_{kp}^t - \mu_0^t)^2\right) + \sum_{k\in\mathcal{K}}\log c_{\mathcal{D}}(\boldsymbol{o}_k^0)\ .\end{aligned} \tag{4.24}$$

By factorizing terms related to $\boldsymbol{a}$ in (4.24), the following Dirichlet distribution is obtained

$$q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}})$$

where

$$\forall k \in \mathcal{K},\ \tilde{\kappa}_k = \kappa_k^0 + \sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\ .$$

Following the same reasoning, $\boldsymbol{b}$ is distributed according to a product of Dirichlet distributions given by

$$q(\boldsymbol{b}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|\tilde{\boldsymbol{o}}_k)$$

where

$$\forall k \in \mathcal{K},\ \forall p \in \mathcal{P},\ \tilde{o}_{kp} = o_{kp}^0 + \sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]\ .$$

By aggregating terms related to each μ_{kp}^t in (4.24), a Normal distribution is obtained for each μ_{kp}^t such that

$$q(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}}\prod_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\tilde{\mu}_{kp}^t, \tilde{\tau}_{kp}^{-1}\sigma^2\right)$$

where $\forall k \in \mathcal{K}$ and $\forall p \in \mathcal{P}$

$$\begin{aligned}\tilde{\tau}_{kp} &= \sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right] + \tau_0\ , \\ \tilde{\mu}_{kp}^t &= \frac{\sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]x_{tj} + \tau_0\mu_0^t}{\tilde{\tau}_{kp}}\ .\end{aligned}$$

Eventually, an Inverse Gamma distribution is deduced from (4.24) such that

$$q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1, \tilde{\xi}_2)$$

where

$$\begin{aligned}\tilde{\xi}_1 &= \xi_1^0 + \frac{J}{2}\ , \\ \tilde{\xi}_2 &= \xi_2^0 + \frac{1}{2}\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}}\left(\sum_{j\in\mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]x_{tj}^2 + \tau_0{\mu_0^t}^2 - \tilde{\tau}_{kp}(\tilde{\mu}_{kp}^t)^2\right)\ .\end{aligned}$$

Lower bound

Recalling that the lower bound on the log evidence is given by

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})\right]$$

where $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right]$ is the free energy and $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})\right]$ is the entropy of the approximate posterior $q(\boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})$. The free energy can be developed as

$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right]$$

where

$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \sum_{p\in\mathcal{P}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]\log \rho_{jkp}^y\right)$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right] = &\sum_{k\in\mathcal{K}}(\kappa_k^0 - 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - (\xi_1^0 + 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] - \xi_2^0\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] + \log c_{\mathcal{D}}(\boldsymbol{\kappa}^0)\\ &+ \log c_{\mathcal{IG}}(\xi_1^0, \xi_2^0) + \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}}(o_{kp}^0 - 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\Bigg(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right]\\ &+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right]\tau_0(\mathbb{E}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right] - \mu_0^t)^2 + \tau_0\tilde{\tau}_{kp}^{-1}\Bigg) + \sum_{k\in\mathcal{K}}\log c_{\mathcal{D}}(\boldsymbol{o}_k^0)\ .\end{aligned}$$

As for the entropy term, the following decompositon is obtained

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{y}, \boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\\ &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{y}|\boldsymbol{z}, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{y}|\boldsymbol{z}, \mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\sum_{p\in\mathcal{P}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]\log\tilde{r}_{jkp}^y\ ,\\ \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\tilde{r}_{jk}\ .\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right] = &\sum_{k\in\mathcal{K}}(\tilde{\kappa}_k - 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - (\tilde{\xi}_1 + 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] - \tilde{\xi}_2\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] + \log c_{\mathcal{D}}(\tilde{\boldsymbol{\kappa}})\\ &+ \log c_{\mathcal{IG}}(\tilde{\xi}_1, \tilde{\xi}_2) + \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}}(\tilde{o}_{kp} - 1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\left(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] + 1\right)\\ &+ \sum_{k\in\mathcal{K}}\log c_{\mathcal{D}}(\tilde{\boldsymbol{o}}_k)\ .\end{aligned}$$

Expectations

Expectations developed in variational calculations are derived from properties of variational posterior distributions and are obtained as follows. Categorical distribution properties lead to

$$\begin{aligned}\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \forall p \in \mathcal{P}:&\\ \mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right] &= \tilde{r}_{jkp}^y\ ,\\ \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] &= \tilde{r}_{jk}\ .\end{aligned}$$

Dirichlet distribution properties lead to

$$\forall k \in \mathcal{K},\ \forall p \in \mathcal{P} :$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k\in\mathcal{K}} \tilde{\kappa}_k\right) ,$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] = \psi(\tilde{o}_{kp}) - \psi\left(\sum_{p\in\mathcal{P}} \tilde{o}_{kp}\right) .$$

where $\psi(\cdot)$ is the digamma function. Normal distribution properties lead to

$$\begin{aligned}\forall k \in \mathcal{K},\ \forall p \in \mathcal{P} :&\\ \mathbb{E}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right] &= \tilde{\mu}_{kp}^t ,\\ \mathbb{E}_{\boldsymbol{\Theta}}\left[(\mu_{kp}^t)^2\right] &= \mathbb{V}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right]^2\\ &= \sigma^2\tilde{\tau}_{kp}^{-1} + (\tilde{\mu}_{kp}^t)^2 ,\end{aligned}$$

Inverse Gamma distribution properties lead to

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] &= \frac{\tilde{\xi}_1}{\tilde{\xi}_2} ,\\ \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] &= \log\tilde{\xi}_2 - \psi(\tilde{\xi}_1) .\end{aligned}$$

Using all these properties, the following expectation can be calculated as

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \forall p \in \mathcal{P} :$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \mu_{kp}^t)^2}{\sigma^2}\right] = \frac{\tilde{\xi}_1\left(x_{tj} - \tilde{\mu}_{kp}^t\right)^2}{\tilde{\xi}_2} + \tilde{\tau}_{kp}^{-1} .$$

4.2.3 Complete model

A model integrating piecewise parabolic data and mixed data is now presented. By taking into consideration any types of available data, the resulting model can fit data better and can estimate more accurate clusters. First, data formalism and assumptions are detailed. Then, the resulting mixture model and its inference procedure are developed.

Data and assumptions

In this part, data consist of J pulses gathering J amplitudes $\boldsymbol{x}_t = (x_{tj})_{j\in\mathcal{J}}$ associated to J times of arrival $\boldsymbol{t} = (t_j)_{j\in\mathcal{J}}$, J continuous features $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ and J categorical features $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ from K distinct emitters. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj})$ the j^{th} observation vector of mixed variables where

- $\boldsymbol{x}_{qj} \in \mathbb{R}^d$ is a vector of d continuous radar features such as the Radio Frequency, the Pulse Width, the Azimuth or the Pulse Repetition Interval,
- $\boldsymbol{x}_{cj} = (x_{cj_0}, \ldots, x_{cj_{q-1}}) \in C_q$ is a vector of q categorical radar modulations such as intra-pulse modulations or pulse-to-pulse modulations,
- $x_{tj} \in \mathbb{R}$ is a continuous variable modeling the Amplitude.

For each pulse j, the temporal evolution variable x_{tj} and mixed variables $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$ are assumed to be independent conditionally to each cluster $k \in \mathcal{K}$

$$\forall j \in \mathcal{J},\ (\boldsymbol{x}_q, \boldsymbol{x}_c)|z_j = k \perp\!\!\!\perp x_t|z_j = k\ . \tag{4.25}$$

with z_j the latent variable modeling the label of the j^{th} observation vector $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj})$. Moreover, the temporal evolution data $(x_{tj}, t_j)_{j\in\mathcal{J}}$ are distributed according to a piecewise parabolic relation and the quantitative data $(\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ are normally distributed conditionally to categorical data $(\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$. Both quantitative and categorical data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ can be partially observed. Hence $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ are decomposed into observed features $(\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj}^{\text{obs}})_{j\in\mathcal{J}}$ and missing features $(\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{miss}})_{j\in\mathcal{J}}$ such that

$$\forall j \in \mathcal{J},\quad \begin{aligned} \boldsymbol{x}_{qj} &= \begin{pmatrix} \boldsymbol{x}_{qj}^{\text{miss}} \\ \boldsymbol{x}_{qj}^{\text{obs}} \end{pmatrix} \text{ with } (\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{qj}^{\text{obs}}) \in \mathbb{R}^{d_j^{\text{miss}}} \times \mathbb{R}^{d_j^{\text{obs}}} \text{ and } d_j^{\text{miss}} + d_j^{\text{obs}} = d\ , \\ \boldsymbol{x}_{cj} &= \begin{pmatrix} \boldsymbol{x}_{cj}^{\text{miss}} \\ \boldsymbol{x}_{cj}^{\text{obs}} \end{pmatrix} \text{with } (\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) \in \mathcal{C}_{q_j^{\text{miss}}} \times \mathcal{C}_{q_j^{\text{obs}}} \text{ and } q_j^{\text{miss}} + q_j^{\text{obs}} = q\ . \end{aligned}$$

Mixture model

According to the independence assumption (4.25), the distribution of mixed data (Chapter 3) and the piecewise parabolic relation between temporal evolution data, the component distribution results in

$$\forall j \in \mathcal{J},\ p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}|u_j, z_j = k) = p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}|u_j, z_j = k)p(x_{tj}|z_j = k)$$

where

$$\forall j \in \mathcal{J},\quad \begin{aligned} p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}|u_j, z_j = k) &= \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}}\ , \\ p(x_{tj}|z_j = k) &= \prod_{p\in\mathcal{P}} \left(b_{kp}\mathcal{N}\left(x_{tj}|\mu_{kp}^t, \sigma^2\right)\right)^{\delta_{y_j}^p} \end{aligned} \tag{4.26}$$

with

- $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$ the scale latent variables handling outliers for quantitative data $\boldsymbol{x}_q$ and distributed according to a Gamma distribution with shape and rate parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})_{(k,\boldsymbol{c})\in\mathcal{K}\times\mathcal{C}_q}$ conditionally to categorical data $\boldsymbol{x}_c$ and labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$,
- $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = ((\boldsymbol{\mu}_{k\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}, \boldsymbol{\Sigma}_k)_{k\in\mathcal{K}}$ the mean and the variance parameters of quantitative data $\boldsymbol{x}_q$ for each cluster,
- $\boldsymbol{\pi} = (\boldsymbol{\pi}_k)_{k\in\mathcal{K}}$ the weights of the multivariate Categorical distribution of categorical data $\boldsymbol{x}_c$ for each cluster,
- $\boldsymbol{y} = (y_j)_{j\in\mathcal{J}}$ the latent variables indicating the p^{th} piecewise temporal evolution data $\boldsymbol{x}_t$ belong to,
- $\boldsymbol{b} = ((b_{kp})_{p\in\mathcal{P}})_{k\in\mathcal{K}}$ the weights of the Categorical distribution of latent variables $\boldsymbol{y}$,
- $\boldsymbol{\mu}^t = ((\mu_{kp}^t)_{p\in\mathcal{P}})_{k\in\mathcal{K}}$ the set of piecewises for temporal evolution data $\boldsymbol{x}_t$ for each cluster,
- σ^2 the variance of the measurement noise related to temporal evolution data $\boldsymbol{x}_t$.

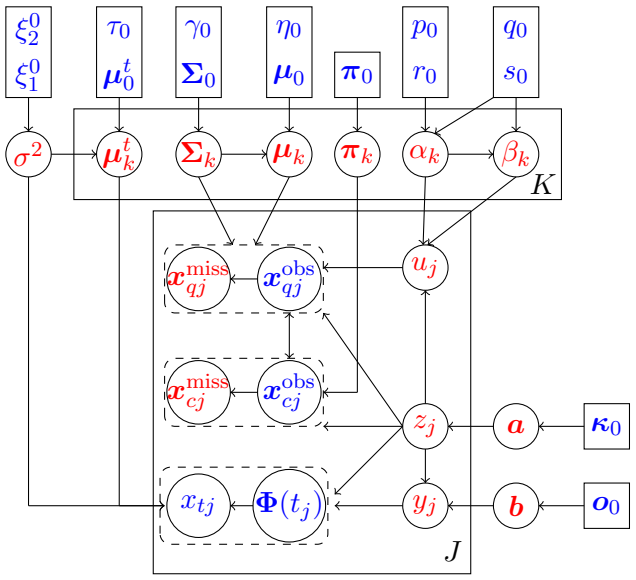


Figure 4.13: Graphical representation of the proposed model integrating temporal evolution data and mixed-type data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}, t_j)$ decomposed into temporal evolution data (x_{tj}, t_j) and mixed-type data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$, the scale variables u_j and the indicator variables (y_j, z_j). Known quantities, respectively unknown quantities, are in blue, respectively in red.

Recalling that $p(z_j = k) = a_k$ where $\boldsymbol{a} = (a_k)_{k \in \mathcal{K}}$ are the weights related to component distributions, the mixture model is obtained from (4.26) such that $\forall j \in \mathcal{J}$,

$$p(\boldsymbol{x}_j, u_j, y_j | \boldsymbol{\Theta}) = \sum_{k \in \mathcal{K}} a_k \prod_{p \in \mathcal{P}} \left(b_{kp} \mathcal{N}\left(x_{tj} | \mu_{kp}^t, \sigma^2\right) \right)^{\delta_{y_j}^p} \prod_{\boldsymbol{c} \in \mathcal{C}} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj} | \boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1} \boldsymbol{\Sigma}_k\right) \mathcal{G}(u_j | \alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}}) \right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \tag{4.27}$$

where $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\mu}^t, \sigma^2, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the set of parameters.

Bayesian framework

As in previous chapters 2 and 3, a Bayesian framework is used to estimate parameters $\boldsymbol{\Theta}$. Assuming datasets $(\boldsymbol{x} = (\boldsymbol{x}_q, \boldsymbol{x}_c, \boldsymbol{x}_t), \boldsymbol{t})$ of i.i.d observations $(\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}), t_j)_{j \in \mathcal{J}}$, independent labels $\boldsymbol{z} = (z_j)_{j \in \mathcal{J}}$, piecewise indicators $\boldsymbol{y} = (y_j)_{j \in \mathcal{J}}$ and scale latent variables $\boldsymbol{u} = (u_j)_{j \in \mathcal{J}}$, the complete likelihood associated to (4.27) is defined by

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{z} | \boldsymbol{\Theta}, \mathcal{K}) = \prod_{j \in \mathcal{J}} \prod_{k \in \mathcal{K}} \Bigg(a_k \prod_{p \in \mathcal{P}} \left(b_{kp} \mathcal{N}\left(x_{tj} | \mu_{kp}^t, \sigma^2\right) \right)^{\delta_{y_j}^p} \\ \times \prod_{\boldsymbol{c} \in \mathcal{C}_q} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj} | \boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1} \boldsymbol{\Sigma}_k\right) \mathcal{G}(u_j | \alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}}) \right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \Bigg)^{\delta_{z_j}^k} \end{aligned}$$

Eventually, the prior distribution required for $\boldsymbol{\Theta}$ is chosen as

$$p(\boldsymbol{\Theta} | \mathcal{K}) = p(\boldsymbol{a} | \mathcal{K}) p(\boldsymbol{b} | \mathcal{K}) p(\boldsymbol{\mu}^t | \sigma^2, \mathcal{K}) p(\sigma^2) p(\boldsymbol{\pi} | \mathcal{K}) p(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathcal{K}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathcal{K})$$

where

$$\begin{cases} p(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\boldsymbol{\kappa}_0) \ , \\ p(\boldsymbol{b}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|\boldsymbol{o}_0) \ , \\ p(\boldsymbol{\pi}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{\pi}_k|\boldsymbol{\pi}_0) \ , \\ p(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{\boldsymbol{c}\in\mathcal{C}_q} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\boldsymbol{\mu}_0, \eta_0^{-1}\boldsymbol{\Sigma}_k\right) \mathcal{IW}(\boldsymbol{\Sigma}_k|\gamma_0,\boldsymbol{\Sigma}_0) \ , \\ p(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{\boldsymbol{c}\in\mathcal{C}_q} p(\alpha_{k\boldsymbol{c}},\beta_{k\boldsymbol{c}}|p_0,q_0,s_0,r_0) \ , \\ p(\boldsymbol{\mu}^t|\sigma^2,\mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\mu_0^t, \tau_0^{-1}\sigma^2\right) \ , \\ p(\sigma^2) = \mathcal{IG}(\sigma^2|\xi_1^0,\xi_2^0) \ . \end{cases}$$

Graphical representation of the proposed model is shown in Figure 4.13.

Inference

As previously, a factorized posterior distribution $q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K}) = q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{z}|\mathcal{K})q(\boldsymbol{\Theta}|\mathcal{K})$ is chosen as an approximation of the intractable posterior joint distribution $p(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{x}^{\text{obs}}, \mathcal{K})$ such that latent variables $\boldsymbol{h} = (\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{z})$ and parameters $\boldsymbol{\Theta}$ are a posteriori independent and

$$
\begin{aligned}
q(\boldsymbol{h}|\mathcal{K}) &= q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K})q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K})q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z}, \mathcal{K})q(\boldsymbol{y}|\boldsymbol{z}, \mathcal{K})q(\boldsymbol{z}|\mathcal{K}) \ , \\
q(\boldsymbol{\Theta}|\mathcal{K}) &= q(\boldsymbol{a}|\mathcal{K})q(\boldsymbol{b}|\mathcal{K})q(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K})q(\sigma^2)q(\boldsymbol{\pi}|\mathcal{K})q(\boldsymbol{\alpha}, \beta|\mathcal{K})q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K}) \ .
\end{aligned}
$$

According to VB assumptions, the following conjugate variational posterior distributions are obtained from the VB procedure

$$
\begin{cases}
q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{miss}}|\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}}, u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k} \ , \\
q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{G}\left(u_j|\tilde{\alpha}_{jk\boldsymbol{c}}, \tilde{\beta}_{jk\boldsymbol{c}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k} \ , \\
q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{MC}(\boldsymbol{x}_{cj}^{\text{miss}}|\tilde{\boldsymbol{r}}_{jk}^{\boldsymbol{x}_c^{\text{miss}}})^{\delta_{z_j}^k} \\
q(\boldsymbol{y}|\boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{C}at(y_j|\tilde{\boldsymbol{r}}_{jk}^{y})^{\delta_{z_j}^k} \ , \\
q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j) \ , \\
q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}}) \ , \\
q(\boldsymbol{\pi}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{\pi}|\tilde{\boldsymbol{\pi}}_k) \ , \\
q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k\in\mathcal{K}}\prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}, \tilde{\eta}_{k\boldsymbol{c}}^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k, \tilde{\boldsymbol{\Sigma}}_k) \ , \\
q(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \prod_{k\in\mathcal{K}} p(\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}}|\tilde{p}_k, \tilde{q}_k, \tilde{s}_k, \tilde{r}_k) \ , \\
q(\boldsymbol{b}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|\tilde{\boldsymbol{o}}_k) \ , \\
q(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}}\prod_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\tilde{\mu}_{kp}^t, \tilde{\tau}_{kp}^{-1}\sigma^2\right) \ , \\
q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1, \tilde{\xi}_2) \ .
\end{cases}
\tag{4.28}
$$

Their respective parameters are estimated during the VBE and VBM steps by developing expectations $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right]$ and $\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right]$. Noting that

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right] =& \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{y}|\boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\mu}^t, \sigma^2, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_q, \boldsymbol{u}, \boldsymbol{x}_c|\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{K})\right] \\
&+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{z}|\boldsymbol{a}, \mathcal{K})\right] \ ,
\end{aligned}
\tag{4.29}
$$

and

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right] =& \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{b}, \boldsymbol{\mu}^t, \sigma^2|\boldsymbol{z}, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_q, \boldsymbol{u}, \boldsymbol{x}_c, \boldsymbol{\pi}, \boldsymbol{\alpha}, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{z}, \mathcal{K})\right] \\
&+ \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{z}, \boldsymbol{a}|\mathcal{K})\right] \ ,
\end{aligned}
\tag{4.30}
$$

the VBE (4.29) and VBM (4.30) steps can be independently derived for latent variables and parameters related to temporal evolution data $\boldsymbol{x}_t$ and mixed data $(\boldsymbol{x}_q, \boldsymbol{x}_c)$. Therefore, variational posterior distributions of latent variables $(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}})$ and parameters $(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ related to mixed data $(\boldsymbol{x}_q, \boldsymbol{x}_c)$ are obtained as in Chapter 3 by deriving green expectations in (4.29) and (4.30). As for $(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{\mu}^t, \sigma^2)$, their variational posterior distribution are obtained as in subsection 4.2.2 by developing blue expectations in (4.29) and (4.30). As in subsection 4.2.2 or in Chapter 3, the Dirichlet posterior distribution of $\boldsymbol{a}$ is deduced from the red expectation in (4.30). Eventually, the variational distribution of labels $\boldsymbol{z}$ is obtained by marginalising over latent variables in the green expectation and developing both blue and red expectations in (4.29) such that

$$\int \mathbb{E}_{\boldsymbol{\Theta}} \left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right] \partial \boldsymbol{x}_q^{\text{miss}} \partial \boldsymbol{y} \partial \boldsymbol{u} \partial \boldsymbol{x}_c^{\text{miss}} = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \delta_{z_j}^k \log \rho_{jk}$$

where $\log \rho_{jk}$ is deduced from red, blue and green expectations in (4.29) as follows

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \log \rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}} \left[\log a_k\right] + \log \rho_{jk}^t + \log \rho_{jk}^{qc} \,. \tag{4.31}$$

with

$$\begin{aligned}
&\forall j \in \mathcal{J},\ \forall k \in \mathcal{K}, \\
&\qquad \mathbb{E}_{\boldsymbol{\Theta}} \left[\log a_k\right] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k \in \mathcal{K}} \tilde{\kappa}_k\right) \,, \\
&\qquad \log \rho_{jk}^t = \log \sum_{p \in \mathcal{P}} \rho_{jkp}^y \,, \\
&\qquad \log \rho_{jk}^{qc} = \log \sum_{\boldsymbol{c}^{\text{miss}} \in \mathcal{C}_{q_j^{\text{miss}}}} \rho_{k\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_{cj}^{\text{miss}}} \,.
\end{aligned}$$

The red term $\mathbb{E}_{\boldsymbol{\Theta}} \left[\log a_k\right]$ is deduced from properties of the Dirichlet distribution, the blue term $\log \rho_{jk}^t$ is deduced from (4.22) and (4.23) in subsection 4.2.2 and the green term $\log \rho_{jk}^{qc}$ has been detailed in Chapter 3. Hence, $\boldsymbol{z}$ is distributed a posteriori according to a product of Categorical distributions parametrized by $\tilde{\boldsymbol{r}} = (\tilde{r}_{jk})_{(j,k) \in \mathcal{J} \times \mathcal{K}}$ given by

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \tilde{r}_{jk} = \frac{\rho_{jk}}{\sum\limits_{k \in \mathcal{K}} \rho_{jk}} \tag{4.32}$$

The lower bound on the log evidence is still required to master the VB inference and can be also decomposed into terms related to temporal evolution data (blue terms), mixed data (green terms) and labels $\boldsymbol{z}$ (red terms). This decomposition is obtained as follows

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right]$$

where the free energy can be developed as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\Theta}, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}} \left[\log p(\boldsymbol{b}, \boldsymbol{\mu}^t, \sigma^2|\mathcal{K})\right] \\
+ \mathbb{E}_{\boldsymbol{\Theta}} \left[\log p(\boldsymbol{a}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K})\right]
\end{aligned}$$

and the entropy as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log q(\boldsymbol{b}, \boldsymbol{\mu}^t, \sigma^2|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log q(\boldsymbol{y}|\boldsymbol{z}, \mathcal{K})\right] \\
+ \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{u}|\boldsymbol{z}, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log q(\boldsymbol{z}|\mathcal{K})\right] \\
+ \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}} \left[\log q(\boldsymbol{a}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K})\right] \,.
\end{aligned}$$

Blue terms, respectively green terms, have been previously detailed in subsection 4.2.2 , respectively in Chapter 3. As for red terms, they are detailed below :

$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\rho_{jk}\ ,$$
$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\tilde{r}_{jk}\ .$$

4.2.4 Experiments

Two experiments are carried out to evaluate clustering performance with respect to a set of synthetic data. In the first experiment, only temporal evolution data are taken into consideration in the clustering procedure. Then, both temporal evolution data and quantitative data are considered in the second one. For comparison, the spectral clustering [VL07] and the k-means algorithm from [HW79] are also evaluated. First, characteristics of data, comparison algorithms and evaluation metrics are detailed. Then, both experiments are described and performance are shown to exhibit the effectiveness of the proposed model.

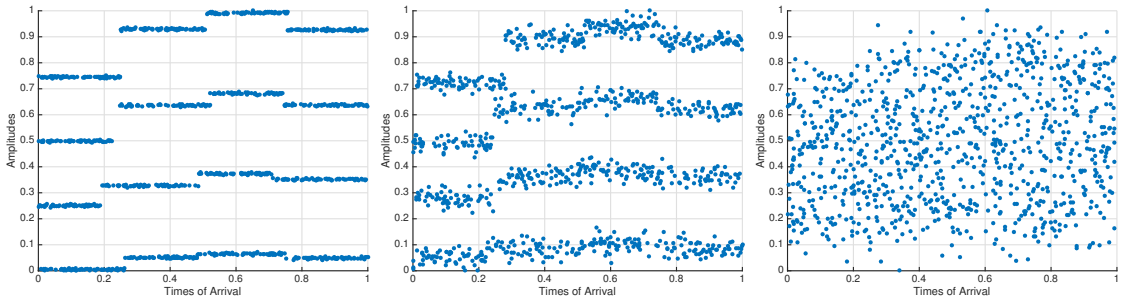
Data, algorithms and metrics

Synthetic data are composed of temporal evolution data related to amplitudes which are distributed according to a piecewise parabolic relation and quantitative data related to continuous radar features which are jointly distributed according to a multivariate normal distribution. Temporal evolution data are generated by sampling a set of data from four piecewise parabolas directed by

$$\boldsymbol{\omega} = \begin{pmatrix} -1 & -2 & -3 & -4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}\ .$$

For each piecewise parabola, $p = 4$ piecewises are obtained by dividing the time interval in p equal subintervals and assigning to each piecewise the value of the parabola at the minimal time of its related time subinterval. Quantitative data are generated by sampling a set of data from four well-separated bivariate clusters with centers $[0,0]^T$, $[1,0]^T$,$[0,1]^T$ and $[1,1]^T$ and identity covariance matrices. Three synthetic datasets are generated with respect to a range of values of σ^2 and are linearly transformed by a min-max normalization to meet algorithms requirements. These datasets are shown in Figures 4.14 and 4.15 where each radar emitter is represented by a piecewise parabola (Figure 4.14) and a Gaussian cluster (Figure 4.15).

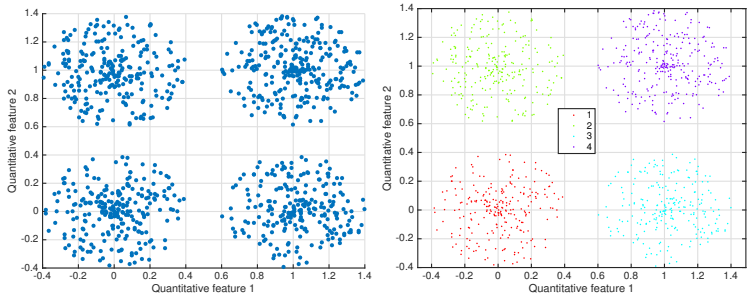
Except for the k-means algorithm, an initialisation is required for clustering algorithms that are involved in these experiments. The similarity graph required for the spectral clustering is obtained from a k-nearest neighbor graph as suggested in [VL07] where the number of neighbors k is chosen as the product of the log number of observations and the number of clusters. As for the proposed model, a supervised initialisation is retained due to its sensitivity to initialisation. First, prior hyperparameters ξ_1^0 and ξ_2^0 are initialised such that the prior mean $\mathbb{E}[\frac{1}{\sigma^2}] = \frac{\xi_1^0}{\xi_2^0}$ of the variance parameter σ^2 is equal to the inverse of the determinant of the covariance matrix of temporal evolution data points. This choice is motivated by the fact that the determinant of the covariance matrix can be interpreted as the generalized variance that reflects the overall spread of the data. Setting $\xi_2^0 = 1$, ξ_1^0 is initialised as the inverse of the generalized variance of the sample of temporal evolution data. In addition, prior piecewise means μ_0^t are initialised from results of a k-means algorithm on temporal evolution data. Then, prior component means



(a) Data, $\sigma^2 = 0.0001$ (b) Data, $\sigma^2 = 0.01$ (c) Data, $\sigma^2 = 0.25$

(d) Ground-Truth, $\sigma^2 = 0.0001$ (e) Ground-Truth, $\sigma^2 = 0.01$ (f) Ground-Truth, $\sigma^2 = 0.25$

Figure 4.14: Synthetic piecewise parabolic data generated from different values of the variance parameter σ^2. Figures (a), (b) and (c) present unlabeled data where 4 piecewise parabolas are generated. Ground-truth are visible on Figures (d), (e) and (f).



(a) Quantitative Data (b) Ground-Truth

Figure 4.15: Synthetic quantitative data generated from 4 multivariate normal distributions. Figure (a) shows unlabeled data and Figure (b) exhibits the ground-truth.

Table 4.6: Initialisation of hyper-parameters values for clustering on piecewise parabolic data

τ_0	κ_0	η_0	γ_0	p_0	r_0	q_0	s_0
1	0.5	100	1	1	1	1	1

Table 4.7: Adjusted Rand Index (ARI) and Silhouette coefficient (S) values for the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the first experiment on synthetic data when only temporal evolution data are considered.

	ARI			S			
	PM	SC	KM	Data	PM	SC	KM
$\sigma^2 = 0.0001$	0.73	0.73	0.35	0.30	0.35	0.35	0.64
$\sigma^2 = 0.01$	0.76	0.84	0.34	0.26	0.31	0.33	0.64
$\sigma^2 = 0.25$	0.54	0.26	0.27	0.10	0.14	0.53	0.56

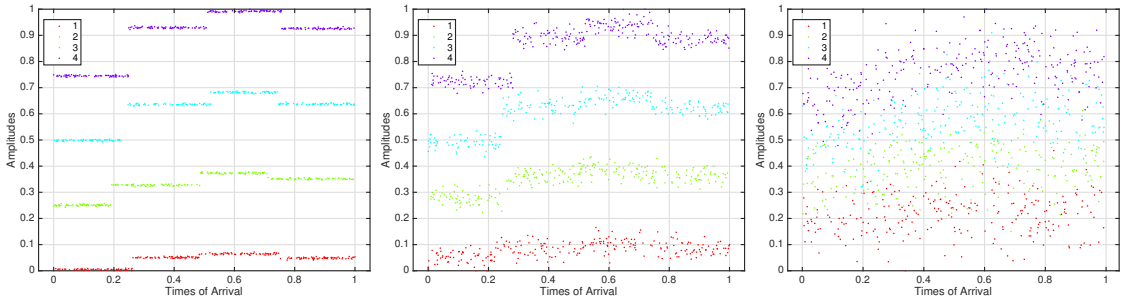
$\boldsymbol{\mu}_0$, respectively covariance matrices $\boldsymbol{\Sigma}_0$, are initialised from results of a k-means algorithm on quantitative data, respectively from diagonal matrices whose diagonal elements are variances of quantitative data. Other hyper-parameters are initialised as in Table 4.6.

Performance on synthetic data are evaluated through the Adjusted Rand Index (ARI) [HA85] that compares estimated partitions of data with the ground-truth and the Silhouette Coefficient [KR09] which does not require the ground-truth and provides a higher score when clusters are dense and well separated.

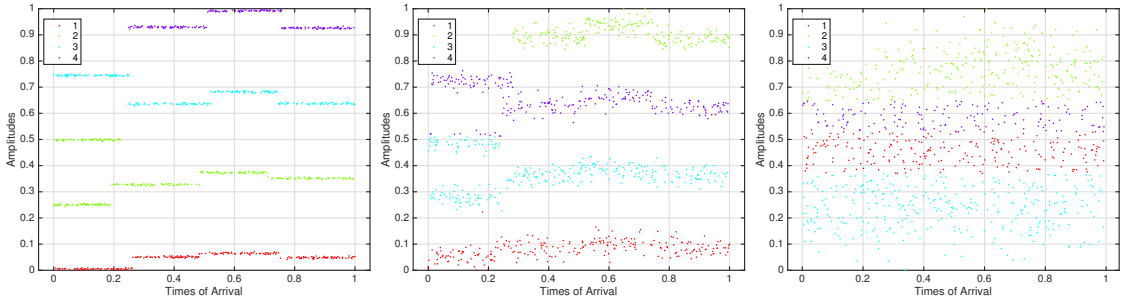
Experiments and results

The first experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters K when only temporal evolution data are taken into consideration. According to datasets visualised in Figure 4.14, K is set to 4 for synthetic data. Results of the first experiment on synthetic data are shown in Figure 4.16 and in Table 4.7. When $\sigma^2 \in \{0.0001, 0.01\}$, the proposed model and the spectral clustering have similar performance in clustering synthetic data with an ARI equals to 0.73 while the k-means algorithm has the lowest performance (ARI = 0.35) in creating convex and isotropic clusters that cannot handle the piecewise parabolic structure of the generated data. This limitation is emphasized by higher Silhouette Coefficients of the k-means algorithm whereas the non-convexity of the data is confirmed by the lower Silhouette Coefficients of the ground-truth. Even if all algorithms poorly perform when data are embedded in noise ($\sigma^2 = 0.25$), the proposed algorithm estimates clusters with a more parabolic shape than other algorithms which build more isotropic clusters (Subfigures (f), (i) and (l) in Figure 4.7). Indeed the Silhouette Coefficient of the proposed model ($S = 0.14$) is closer to the Silhouette Coefficient of the ground-truth ($S = 0.10$) than Silhouette Coefficients of spectral clustering ($S = 0.53$) and k-means ($S = 0.56$).

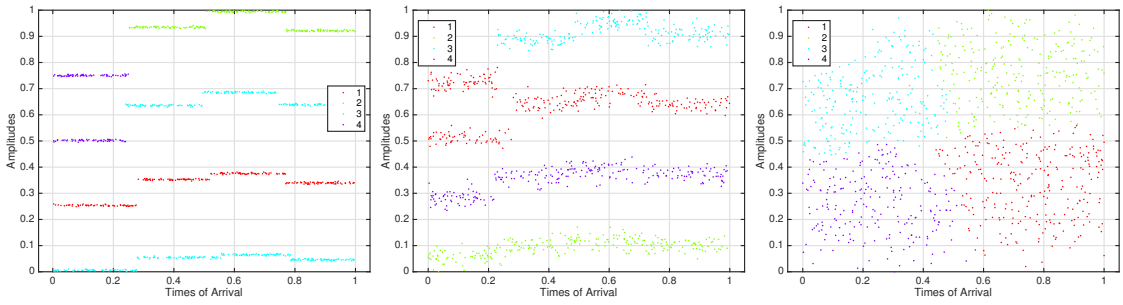
The second experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters K when all types of data are taken into consideration. The number of clusters K is still set to 4 for synthetic data.. Results of the second experiment on synthetic data are shown in Figure 4.17 and in Table 4.8. All algorithms succeed in clustering synthetic data for $\sigma^2 \in \{0.0001, 0.01, 0.25\}$ since the ground-truth partition is recovered in Figure 4.17 with an ARI equals to 1 visible on Table 4.8. Adding quantitative information enables algorithms to recover the ground-truth for any value of σ^2.



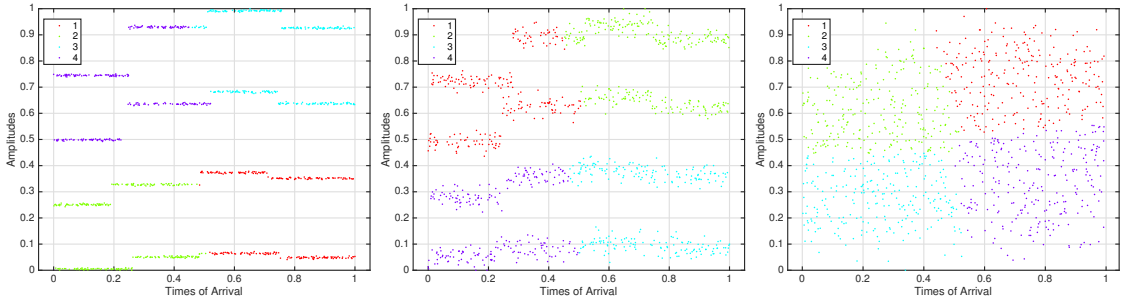
(a) Ground-Truth, $\sigma^2 = 0.0001$ (b) Ground-Truth, $\sigma^2 = 0.01$ (c) Ground-Truth, $\sigma^2 = 0.25$



(d) PM, $\sigma^2 = 0.0001$ (e) PM, $\sigma^2 = 0.01$ (f) PM, $\sigma^2 = 0.25$

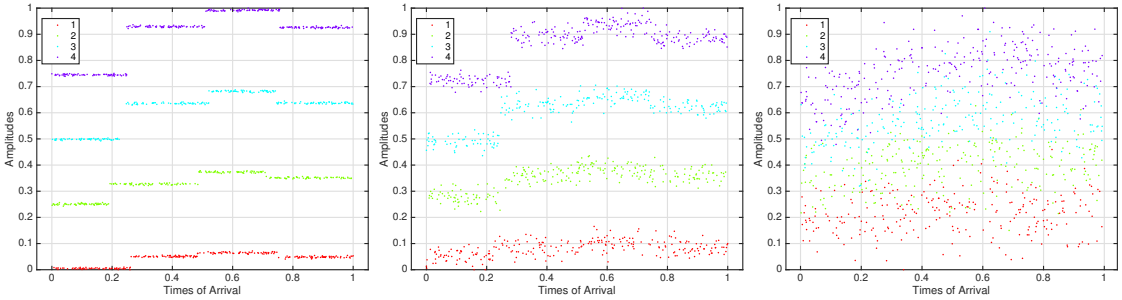


(g) SC, $\sigma^2 = 0.0001$ (h) SC, $\sigma^2 = 0.01$ (i) SC, $\sigma^2 = 0.25$

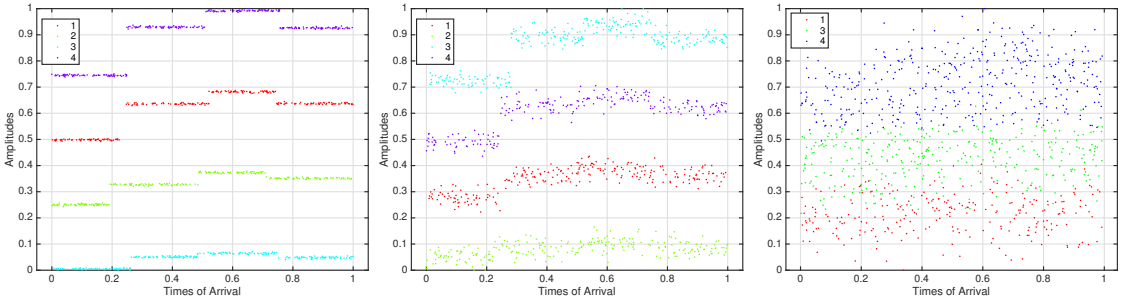


(j) KM, $\sigma^2 = 0.0001$ (k) KM, $\sigma^2 = 0.01$ (l) KM, $\sigma^2 = 0.25$

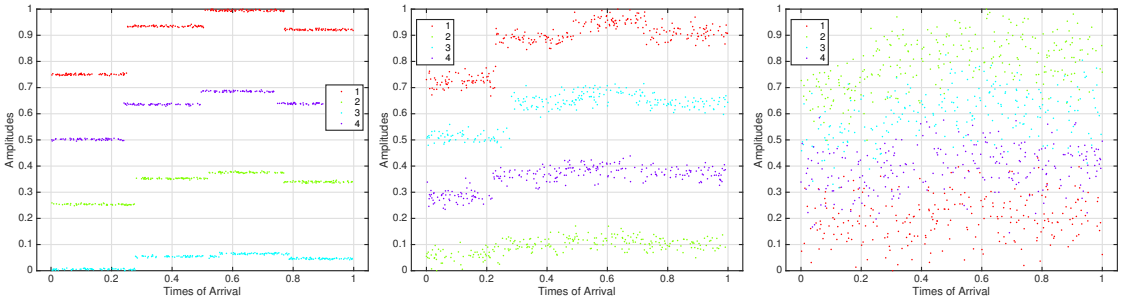
Figure 4.16: Results on synthetic parabolic piecewise data when only temporal evolution data are considered. Figures (a), (b) and (c) show synthetic data generated with different values of the variance σ^2. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).



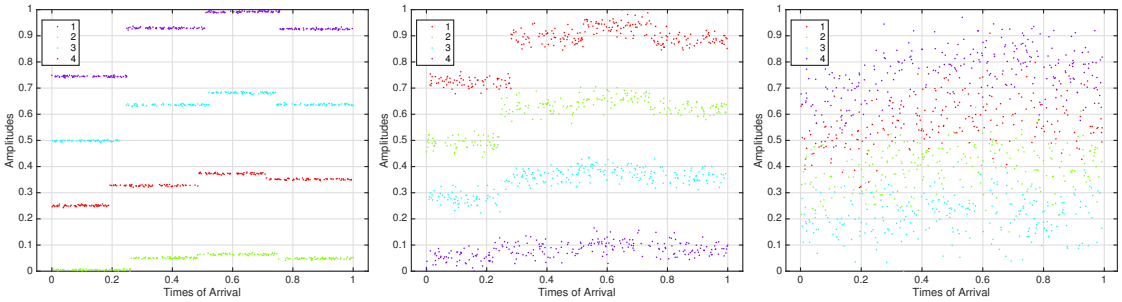
(a) Ground-Truth, $\sigma^2 = 0.0001$ (b) Ground-Truth, $\sigma^2 = 0.01$ (c) Ground-Truth, $\sigma^2 = 0.25$



(d) PM, $\sigma^2 = 0.0001$ (e) PM, $\sigma^2 = 0.01$ (f) PM, $\sigma^2 = 0.25$



(g) SC, $\sigma^2 = 0.0001$ (h) SC, $\sigma^2 = 0.01$ (i) SC, $\sigma^2 = 0.25$



(j) KM, $\sigma^2 = 0.0001$ (k) KM, $\sigma^2 = 0.01$ (l) KM, $\sigma^2 = 0.25$

Figure 4.17: Results on synthetic data when all types of data are considered. Figures (a), (b) and (c) show synthetic data generated with different values of the variance σ^2. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).

Table 4.8: Adjusted Rand Index (ARI) and Silhouette coefficient (S) values for the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the second experiment on synthetic data when all types of data are considered.

	ARI			S			
	PM	SC	KM	Data	PM	SC	KM
$\sigma^2 = 0.0001$	1	1	1	0.76	0.76	0.76	0.76
$\sigma^2 = 0.01$	1	1	1	0.75	0.75	0.75	0.75
$\sigma^2 = 0.25$	0.64	1	1	0.73	0.45	0.73	0.73

4.3 Parabolic and piecewise parabolic data

In this section, we consider that both scanning behaviours are observed among the K emitters. Hence, parabolic and piecewise parabolic relations are observed in data and have to be taken into account in the clustering procedure by developing a mixture model that can build K distinct clusters formed by either parabolas or piecewise parabolas. Then, the proposed model is enhanced with the mixture model designed for mixed data in Chapter 3 in order to improve clustering performance. Finally, experiments on synthetic data are carried out to exhibit performance of the proposed approach.

4.3.1 Model

Definitions of parabolic relation and piecewise parabolic have been previously introduced in Sections 4.1 and 4.2. Now, data can be modeled either by a parabolic relation (4.1) or a piecewise parabolic relation (4.17). Since the measurement noise ϵ is still Gaussian, amplitudes $(x_{tj})_{j\in\mathcal{J}}$ are normally distributed according to (4.3) and (4.18) such that

$$\forall j \in \mathcal{J},\ x_{tj}|t_j \sim \begin{cases} \mathcal{N}\left(x_{tj}|\mathbf{\Phi}(t_j)^T\boldsymbol{\omega}, \sigma^2\right) \text{ if } (x_{tj}, t_j) \text{ have a parabolic relation .} \\ \exists p \in \mathcal{P},\ \mathcal{N}\left(x_{tj}|\mu_p^t, \sigma^2\right) \text{ if } (x_{tj}, t_j) \text{ have a piecewise parabolic relation .} \end{cases} \tag{4.33}$$

Figure 4.18 presents data where amplitudes $\boldsymbol{x}_t$ and times of arrival $\boldsymbol{t}$ are distributed according to a parabolic relation and a piecewise parabolic relation from two distinct emitters.

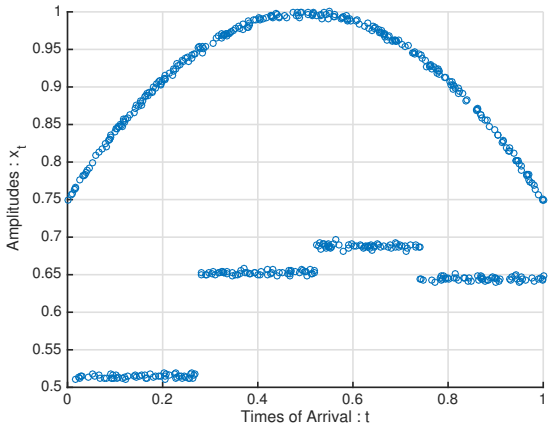


Figure 4.18: Data generated from two distinct emitters presenting a parabolic scanning behaviour and a piecewise parabolic scanning behaviour. Hence, amplitudes $\boldsymbol{x}_t$ and times of arrival $\boldsymbol{t}$ are distributed according to a parabolic relation and a piecewise parabolic relation defined with $P = 4$ piecewises.

Mixture model

Since each radar emitter has its own scanning behaviour, radar emitters can be distinguished into two groups where K_0 unique parabolas exist in the first group and K_1 piecewise parabolas exist in the second one such that $K = K_0 + K_1$. Therefore K_0 from K regression parameters $\boldsymbol{\omega} = (\boldsymbol{\omega}_k)_{k\in\mathcal{K}}$ and K_1 from K sets of piecewises $\boldsymbol{\mu}^t = (\mu_{kp}^t)_{(p,k)\in\mathcal{P}\times\mathcal{K}}$ have to be estimated. Then, each amplitude x_{tj} belongs to one of these sets which is related to a specific emitter. In other words, conditionally to its label z_j, the amplitude x_{tj} is distributed according to (4.33) such that

the component distribution is defined by

$$\forall j \in \mathcal{J},\ x_{tj}|t_j, z_j = k \sim \begin{cases} \mathcal{N}\left(x_{tj}|\mathbf{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) \text{ if } k \in \mathcal{K}_0\ . \\ \exists p \in \mathcal{P},\ \mathcal{N}\left(x_{tj}|\mu^t_{kp}, \sigma^2\right) \text{ if } k \in \mathcal{K}_1\ . \end{cases} \tag{4.34}$$

In order to model its affiliations to one of the groups, a latent discrete variable w_j is introduced such that (4.34) becomes

$$\forall j \in \mathcal{J},\ \begin{array}{l} x_{tj}|t_j, w_j = 0, z_j = k \sim \mathcal{N}\left(x_{tj}|\mathbf{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) \\ x_{tj}|y_j = p, w_j = 1, z_j = k \sim\ \mathcal{N}\left(x_{tj}|\mu^t_{kp}, \sigma^2\right) \end{array} \tag{4.35}$$

where $w_j \in \{0, 1\}$ follows, conditionally to $z_j = k$, a categorical distribution with weights $\boldsymbol{c}_k = (c_{k0}, c_{k1})$ and y_j is the latent variable defined in (4.20) that follows, conditionally to $z_j = k$ and $w_j = 1$, a categorical distribution with weights $\boldsymbol{b}_k = (b_{k1}, \ldots, b_{kP})$. Therefore the initial component distribution (4.34) can be reformulated as

$$p(x_{tj}|t_j, z_j = k, \mathbf{\Theta}, \mathcal{K}) = c_{k0}\mathcal{N}\left(x_{tj}|\mathbf{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) + c_{k1}\sum_{p\in\mathcal{P}} b_{kp}\mathcal{N}\left(x_{tj}|\mu^t_{kp}, \sigma^2\right)$$

Recalling that $p(z_j = k) = a_k$ where $\boldsymbol{a} = (a_k)_{k\in\mathcal{K}}$ are the weights related to component distributions, the proposed mixture model is a mixture of mixture models given by

$$\forall j \in \mathcal{J},\ p(x_{tj}|t_j, \mathbf{\Theta}) = \sum_{k\in\mathcal{K}} a_k \left(c_{k0}\mathcal{N}\left(x_{tj}|\mathbf{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right) + c_{k1}\sum_{p\in\mathcal{P}} b_{kp}\mathcal{N}\left(x_{tj}|\mu^t_{kp}, \sigma^2\right)\right) \tag{4.36}$$

where $\mathbf{\Theta} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\omega}, \boldsymbol{\mu}^t, \sigma^2)$ is the set of parameters.

Bayesian framework

As in previous chapters, a Bayesian framework is used to estimate parameters $\mathbf{\Theta}$. Assuming datasets $(\boldsymbol{x}_t, \boldsymbol{t})$ of i.i.d observations $(x_{tj}, t_j)_{j\in\mathcal{J}}$ and independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$, $\boldsymbol{w} = (w_j)_{j\in\mathcal{J}}$ and $\boldsymbol{y} = (y_j)_{j\in\mathcal{J}}$ for clusters, scanning types and piecewises, the complete likelihood associated to (4.36) is defined by

$$p(\boldsymbol{x}_t, \boldsymbol{h}|\mathbf{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \left(a_k \left(c_{k0}\mathcal{N}\left(x_{tj}|\mathbf{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right)\right)^{\delta^0_{w_j}} \left(c_{k1}\prod_{p\in\mathcal{P}} \left(b_{kp}\mathcal{N}\left(x_{tj}|\mu^t_{kp}, \sigma^2\right)\right)^{\delta^p_{y_j}}\right)^{\delta^1_{w_j}}\right)^{\delta^k_{z_j}}.$$

where $\boldsymbol{h} = (\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z})$ is the set of latent variables. Eventually, the prior distribution required for $\mathbf{\Theta}$ is chosen as

$$p(\mathbf{\Theta}|\mathcal{K}) = p(\boldsymbol{a}|\mathcal{K})p(\boldsymbol{b}|\mathcal{K})p(\boldsymbol{c}|\mathcal{K})p(\boldsymbol{\omega}|\sigma^2, \mathcal{K})p(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K})p(\sigma^2)$$

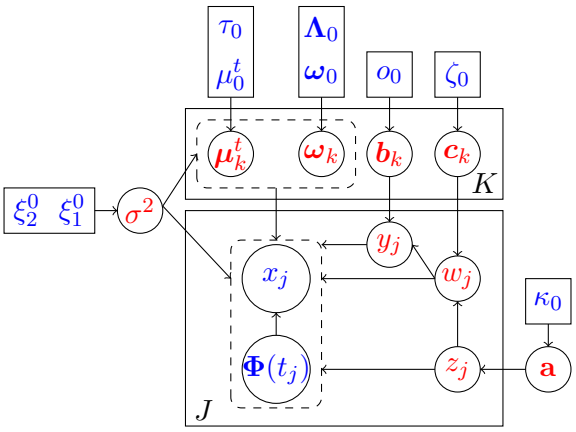


Figure 4.19: Graphical representation of the proposed mixture model handling parabolic and piecewise parabolic data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations (x_{tj}, t_j) decomposed into the amplitude x_j and the polynomial transformation $\boldsymbol{\Phi}(t_j)$ and the indicator variables (y_j, w_j, z_j). Known quantities, respectively unknown quantities, are in blue, respectively in red.

where $\boldsymbol{a}$, $\boldsymbol{b}_k$ and $\boldsymbol{c}_k$ follow a Dirichlet distribution, each $\boldsymbol{\omega}_k$ and μ_{kp}^t follow a Normal distribution and σ^2 follows an Inverse Gamma distribution such that

$$\begin{cases} p(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\mathbf{a}|\kappa_0) \ , \\ p(\boldsymbol{b}|\mathcal{K}) = \prod\limits_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|o_0) \ , \\ p(\boldsymbol{c}|\mathcal{K}) = \prod\limits_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{c}_k|\zeta_0) \ , \\ p(\boldsymbol{\omega}|\sigma^2,\mathcal{K}) = \prod\limits_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\boldsymbol{\omega}_0, \sigma^2\boldsymbol{\Lambda}_0\right) \ , \\ p(\boldsymbol{\mu}^t|\sigma^2,\mathcal{K}) = \prod\limits_{k\in\mathcal{K}}\prod\limits_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\mu_0^t, \tau_0^{-1}\sigma^2\right) \ , \\ p(\sigma^2) = \mathcal{IG}(\sigma^2|\xi_1^0, \xi_2^0) \ . \end{cases}$$

The resulting mixture model is shown on Figure 4.19.

4.3.2 Inference

The Variational Bayes (VB) procedure is derived to estimate parameters of the mixture model defined in (4.36). Variational posterior distributions are obtained from the VB Expectation (VBE) and VB Maximization (VBM) steps and a lower bound on the log evidence is defined to master the convergence of the VB procedure.

Variational posterior distributions

As previously, a factorized posterior distribution $q(\boldsymbol{h}, \boldsymbol{\Theta}|\mathcal{K}) = q(\boldsymbol{h}|\mathcal{K})q(\boldsymbol{\Theta}|\mathcal{K})$ is chosen as an approximation of the intractable posterior joint distribution $p(\boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{x}_t, \mathcal{K})$ such that latent variables $\boldsymbol{h}$ and parameters $\boldsymbol{\Theta}$ are a posteriori independent and

$$\begin{aligned} q(\boldsymbol{h}|\mathcal{K}) &= q(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{z}, \mathcal{K})q(\boldsymbol{w}|\boldsymbol{z}, \mathcal{K})q(\boldsymbol{z}|\mathcal{K}) \ , \\ q(\boldsymbol{\Theta}|\mathcal{K}) &= q(\boldsymbol{a}|\mathcal{K})q(\boldsymbol{b}|\mathcal{K})q(\boldsymbol{c}|\mathcal{K})q(\boldsymbol{\omega}|\sigma^2, \mathcal{K})q(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K})q(\sigma^2) \ . \end{aligned}$$

According to VB assumptions, the following conjugate variational posterior distributions are obtained from the VB procedure

$$\begin{cases} q(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{C}at(y_j|\tilde{\boldsymbol{r}}^y_{jk})^{\delta^1_{w_j}\delta^k_{z_j}} \ , \\ q(\boldsymbol{w}|\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{C}at(y_j|\tilde{\boldsymbol{r}}^w_{jk})^{\delta^k_{z_j}} \ , \\ q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j) \ , \\ q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}}) \ , \\ q(\boldsymbol{b}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|\tilde{\boldsymbol{o}}_k) \ , \\ q(\boldsymbol{c}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{c}_k|\tilde{\boldsymbol{\zeta}}_k) \ , \\ q(\boldsymbol{\omega}|\sigma^2,\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\tilde{\boldsymbol{\omega}}_k,\sigma^2\tilde{\boldsymbol{\Lambda}}_k\right) \ , \\ q(\boldsymbol{\mu}^t|\sigma^2,\mathcal{K}) = \prod_{k\in\mathcal{K}}\prod_{p\in\mathcal{P}} \mathcal{N}\left(\mu^t_{kp}|\tilde{\mu}^t_{kp},\tilde{\tau}^{-1}_{kp}\sigma^2\right) \ , \\ q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1,\tilde{\xi}_2) \ . \end{cases}$$

Their respective parameters are estimated during the VBE and VBM steps.

VBE-step

The VBE-step consists in deriving the following expectation

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t,\boldsymbol{h}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta^k_{z_j}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \delta^0_{w_j}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{k0}\right] - \frac{1}{2}\Bigg(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj}-\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\right]\Bigg)\Bigg) + \delta^1_{w_j}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{k1}\right] + \sum_{p\in\mathcal{P}}\delta^p_{y_j}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] \\ &\quad - \frac{1}{2}\Bigg(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj}-\mu^t_{kp})^2}{\sigma^2}\right]\Bigg)\Bigg)\Bigg)\Bigg) \\ &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \delta^k_{z_j}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \delta^0_{w_j}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{k0}\right] - \frac{1}{2}\Bigg(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj}-\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\right]\Bigg)\Bigg) + \delta^1_{w_j}\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{k1}\right] + \sum_{p\in\mathcal{P}}\delta^p_{y_j}\log\rho^y_{jkp}\Bigg)\Bigg) \end{aligned} \tag{4.37}$$

where

$$\log\rho^y_{jkp} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\left(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj}-\mu^t_{kp})^2}{\sigma^2}\right]\right) \ . \tag{4.38}$$

Hence, a categorical distribution for piecewise labels $\boldsymbol{y}$ is deduced from (4.37) conditionally to indexes $\boldsymbol{z}$ and $\boldsymbol{w}$ such that

$$q(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z},\mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}} \mathcal{C}at(y_j|\tilde{\boldsymbol{r}}^y_{jk})^{\delta^1_{w_j}\delta^k_{z_j}}$$

and their parameters $(\tilde{\boldsymbol{r}}^y_{jk})_{j\in\mathcal{J}}$ are obtained from (4.38) as follows

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \forall p \in \mathcal{P},\ \tilde{r}^y_{jkp} = \frac{\rho^y_{jkp}}{\sum\limits_{p\in\mathcal{P}} \rho^y_{jkp}} \ .$$

Marginalising over $\boldsymbol{y}$ in (4.37), the expectation (4.37) becomes

$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] = \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}} \delta^k_{z_j} \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \delta^0_{w_j} \log \rho^w_{jk0} + \delta^1_{w_j} \log \rho^w_{jk1} \right) \tag{4.39}$$

where

$$\begin{aligned} \log \rho^w_{jk0} &= \mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{k0}\right] - \frac{1}{2}\Bigg(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log \sigma^2\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T \boldsymbol{\omega}_k)^2}{\sigma^2} \right] \Bigg) \ , \\ \log \rho^w_{jk1} &= \mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{k1}\right] + \log \sum_{p\in\mathcal{P}} \rho^y_{jkp} \ . \end{aligned} \tag{4.40}$$

Then, a categorical distribution for scanning type labels $\boldsymbol{w}$ is deduced from (4.39) conditionally to cluster labels $\boldsymbol{z}$ such that

$$q(\boldsymbol{w}|\boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \mathcal{C}at(w_j|\tilde{\boldsymbol{r}}^w_{jk})^{\delta^k_{z_j}}$$

and their parameters $(\tilde{\boldsymbol{r}}^y_{jk})_{j\in\mathcal{J}}$ are obtained from (4.40) as follows

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \forall i \in \{0,1\},\ \tilde{r}^w_{jki} = \frac{\rho^w_{jki}}{\sum\limits_{i\in\{0,1\}} \rho^w_{jki}} \ .$$

Eventually by marginalising over $\boldsymbol{w}$ in (4.39) a categorical distribution is obtained for cluster labels $\boldsymbol{z}$ such that

$$q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j)$$

with

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \tilde{r}_{jk} = \frac{\rho_{jk}}{\sum\limits_{k\in\mathcal{K}} \rho_{jk}} \ .$$

where

$$\rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log \sum_{i\in\{0,1\}} \rho^w_{jki} \ .$$

VBM-step

The VBM-step consists in deriving the following expectation

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta}|\boldsymbol{t}, \mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}|\boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] + \log p(\boldsymbol{\Theta}|\mathcal{K}) \\
&= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\Bigg(\log a_k + \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^0\right]\Bigg(\log c_{k0} - \frac{1}{2}\Bigg(\log 2\pi + \log\sigma^2 \\
&+ \frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\Bigg)\Bigg) + \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^1\right]\Bigg(\log c_{k1} + \sum_{p\in\mathcal{P}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]\Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] \\
&- \frac{1}{2}\Bigg(\log 2\pi + \log\sigma^2 + \frac{(x_{tj} - \mu_{kp}^t)^2}{\sigma^2}\Bigg)\Bigg)\Bigg)\Bigg) + \sum_{k\in\mathcal{K}}(\kappa_k^0 - 1)\log a_k + \log c_{\mathcal{D}}(\boldsymbol{\kappa}^0) \\
&- \frac{1}{2}\left(3(\log 2\pi + \log\sigma^2) + \log|\boldsymbol{\Lambda}_0| + (\boldsymbol{\omega}_k - \boldsymbol{\omega}_0)^T\frac{\boldsymbol{\Lambda}_0^{-1}}{\sigma^2}(\boldsymbol{\omega}_k - \boldsymbol{\omega}_0)\right) \\
&- (\xi_1^0 + 1)\log\sigma^2 - \frac{\xi_2^0}{\sigma^2} + \log c_{\mathcal{IG}}(\xi_1^0, \xi_2^0) + \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}}(o_{kp}^0 - 1)\log b_{kp} \\
&- \frac{1}{2}\left(\log 2\pi + \log\sigma^2 + \frac{\tau_0}{\sigma^2}(\mu_{kp}^t - \mu_0^t)^2\right) \\
&+ \sum_{k\in\mathcal{K}}\sum_{i\in\{0,1\}}(\zeta_{ki}^0 - 1)\log c_{ki} + \sum_{k\in\mathcal{K}}\log c_{\mathcal{D}}(\boldsymbol{o}_k^0) + \log c_{\mathcal{D}}(\boldsymbol{\zeta}_k^0)\ .
\end{aligned} \tag{4.41}$$

By factorizing terms related to $\boldsymbol{a}$ in (4.41), the following Dirichlet distribution is obtained

$$q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}})$$

where

$$\forall k \in \mathcal{K},\ \tilde{\kappa}_k = \kappa_k^0 + \sum_{j\in\mathcal{J}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\ .$$

Following the same reasoning, $\boldsymbol{b}$ and $\boldsymbol{c}$ are distributed according to a product of Dirichlet distributions given by

$$\begin{aligned}
q(\boldsymbol{b}|\mathcal{K}) &= \prod_{k\in\mathcal{K}}\mathcal{D}(\boldsymbol{b}_k|\tilde{\boldsymbol{o}}_k)\ , \\
q(\boldsymbol{c}|\mathcal{K}) &= \prod_{k\in\mathcal{K}}\mathcal{D}(\boldsymbol{c}_k|\tilde{\boldsymbol{\zeta}}_k)\ ,
\end{aligned}$$

where

$$\forall k \in \mathcal{K},\ \forall p \in \mathcal{P},\ \tilde{o}_{kp} = o_{kp}^0 + \sum_{j\in\mathcal{J}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^1\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]$$

and

$$\forall k \in \mathcal{K},\ \forall i \in \{0,1\},\ \tilde{\zeta}_{ki} = \zeta_{ki}^0 + \sum_{j\in\mathcal{J}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^i\right]\ .$$

By aggregating terms related to each μ_{kp}^t and $\boldsymbol{\omega}_k$ in (4.41), a Normal distribution is obtained for each μ_{kp}^t and $\boldsymbol{\omega}_k$ such that

$$\begin{aligned}
q(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K}) &= \prod_{k\in\mathcal{K}}\prod_{p\in\mathcal{P}}\mathcal{N}\left(\mu_{kp}^t|\tilde{\mu}_{kp}^t, \tilde{\tau}_{kp}^{-1}\sigma^2\right)\ , \\
q(\boldsymbol{\omega}|\sigma^2, \mathcal{K}) &= \prod_{k\in\mathcal{K}}\mathcal{N}\left(\boldsymbol{\omega}_k|\tilde{\boldsymbol{\omega}}_k, \sigma^2\tilde{\boldsymbol{\Lambda}}_k\right)\ ,
\end{aligned}$$

where $\forall k \in \mathcal{K}$ and $\forall p \in \mathcal{P}$

$$\begin{aligned}
\tilde{\tau}_{kp} &= \sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^1\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right] + \tau_0 \ , \\
\tilde{\mu}_{kp}^t &= \frac{\sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^1\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right] x_{tj} + \tau_0 \mu_0^t}{\tilde{\tau}_{kp}} \ , \\
\tilde{\boldsymbol{\Lambda}}_k &= \left(\sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^0\right] \boldsymbol{\Phi}(t_j) \boldsymbol{\Phi}(t_j)^T + \boldsymbol{\Lambda}_0^{-1} \right)^{-1} \ , \\
\tilde{\boldsymbol{\omega}}_k &= \tilde{\boldsymbol{\Lambda}}_k \left(\sum_{j \in \mathcal{J}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^0\right] x_{tj} \boldsymbol{\Phi}(t_j) + \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\omega}_0 \right) \ .
\end{aligned}$$

Eventually, an Inverse Gamma distribution is deduced from (4.41) such that

$$q(\sigma^2) = \mathcal{IG}(\sigma^2 | \tilde{\xi}_1, \tilde{\xi}_2)$$

where

$$\begin{aligned}
&\tilde{\xi}_1 = \xi_1^0 + \frac{J}{2} \ , \\
&\tilde{\xi}_2 = \xi_2^0 + \frac{1}{2} \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} \Bigg(\sum_{p \in \mathcal{P}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^1\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right] x_{tj}^2 + \tau_0 (\mu_0^t)^2 - \tilde{\tau}_{kp} (\tilde{\mu}_{kp}^t)^2 + \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^0\right] x_{tj}^2 \\
&\quad + \boldsymbol{\omega}_0^T \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\omega}_0 - \tilde{\boldsymbol{\omega}}_k^T \tilde{\boldsymbol{\Lambda}}_k^{-1} \tilde{\boldsymbol{\omega}}_k \Bigg) \ .
\end{aligned}$$

Lower bound

Recalling that the lower bound on the log evidence is given by

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta} | \boldsymbol{t}, \mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h}, \boldsymbol{\Theta} | \mathcal{K})\right]$$

where $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta} | \boldsymbol{t}, \mathcal{K})\right]$ is the free energy and $\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h}, \boldsymbol{\Theta} | \mathcal{K})\right]$ is the entropy of the approximate posterior $q(\boldsymbol{h}, \boldsymbol{\Theta} | \mathcal{K})$. The free energy can be developed as

$$\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h}, \boldsymbol{\Theta} | \boldsymbol{t}, \mathcal{K})\right] = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h} | \boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta} | \mathcal{K})\right]$$

where

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t, \boldsymbol{h} | \boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K})\right] = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] \Bigg(& \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^0\right] \log \rho_{jk0}^w \\
& + \mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^1\right] \Bigg(\mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{k1}\right] + \sum_{p \in \mathcal{P}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right] \log \rho_{jkp}^y \Bigg) \Bigg)
\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{\Theta}|\mathcal{K})\right] = &\sum_{k\in\mathcal{K}}(\kappa_k^0-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - (\xi_1^0+1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] - \xi_2^0\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] + \log c_{\mathcal{D}}(\boldsymbol{\kappa}^0)\\ &+ \log c_{\mathcal{IG}}(\xi_1^0,\xi_2^0) - \frac{1}{2}\Big(3(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right]) + \log|\boldsymbol{\Lambda}_0| + \text{Trace}\left(\tilde{\boldsymbol{\Lambda}}_k\boldsymbol{\Lambda}_0^{-1}\right)\\ &+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right](\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right]-\boldsymbol{\omega}_0)^T\boldsymbol{\Lambda}_0^{-1}(\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right]-\boldsymbol{\omega}_0)\Big)\\ &+ \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}}(o_{kp}^0-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\Big(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right]\\ &+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right]\tau_0(\mathbb{E}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right]-\mu_0^t)^2 + \tau_0\tilde{\tau}_{kp}^{-1}\Big) + \sum_{k\in\mathcal{K}}\sum_{i\in\{0,1\}}(\zeta_{ki}^0-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{ki}\right]\\ &+ \sum_{k\in\mathcal{K}}\log c_{\mathcal{D}}(\boldsymbol{o}_k^0) + \log c_{\mathcal{D}}(\boldsymbol{\zeta}_k^0)\ .\end{aligned}$$

As for the entropy term, the following decompositon is obtained

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h},\boldsymbol{\Theta}|\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{y},\boldsymbol{w},\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\\ &= \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{w}|\boldsymbol{z},\mathcal{K})\right]\\ &+ \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right]\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^1\right]\sum_{p\in\mathcal{P}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right]\log\tilde{r}_{jkp}^y\ ,\\ \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{w}|\boldsymbol{z},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\sum_{i\in\{0,1\}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^i\right]\log\tilde{r}_{jki}^w\ ,\\ \mathbb{E}_{\boldsymbol{h}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\tilde{r}_{jk}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log q(\boldsymbol{\Theta}|\mathcal{K})\right] = &\sum_{k\in\mathcal{K}}(\tilde{\kappa}_k-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] - (\tilde{\xi}_1+1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] - \tilde{\xi}_2\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] + \log c_{\mathcal{D}}(\tilde{\boldsymbol{\kappa}})\\ &+ \log c_{\mathcal{IG}}(\tilde{\xi}_1,\tilde{\xi}_2) - \frac{1}{2}\left(3(1+\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right]) + \log|\tilde{\boldsymbol{\Lambda}}_k|\right)\\ &+ \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}}(\tilde{o}_{kp}-1)\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] - \frac{1}{2}\left(\log 2\pi + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] + 1\right)\\ &+ \sum_{k\in\mathcal{K}}\log c_{\mathcal{D}}(\tilde{\boldsymbol{o}}_k) + c_{\mathcal{D}}(\tilde{\boldsymbol{\zeta}}_k)\ .\end{aligned}$$

Expectations

Expectations developed in variational calculations are derived from properties of variational posterior distributions and are obtained as follows. Categorical distribution properties lead to

$$\begin{aligned}&\forall j\in\mathcal{J},\ \forall k\in\mathcal{K},\ \forall p\in\mathcal{P},\ \forall i\in\{0,1\}:\\ &\qquad\mathbb{E}_{\boldsymbol{h}}\left[\delta_{y_j}^p\right] = \tilde{r}_{jkp}^y\ ,\\ &\qquad\mathbb{E}_{\boldsymbol{h}}\left[\delta_{w_j}^i\right] = \tilde{r}_{jki}^w\ ,\\ &\qquad\mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right] = \tilde{r}_{jk}\ .\end{aligned}$$

Dirichlet distribution properties lead to

$$\forall k \in \mathcal{K},\ \forall p \in \mathcal{P},\ \forall i \in \{0,1\} :$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k\in\mathcal{K}} \tilde{\kappa}_k\right) ,$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log b_{kp}\right] = \psi(\tilde{o}_{kp}) - \psi\left(\sum_{p\in\mathcal{P}} \tilde{o}_{kp}\right) ,$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\log c_{ki}\right] = \psi(\tilde{\zeta}_{ki}) - \psi\left(\sum_{i\in\{0,1\}} \tilde{\zeta}_{ki}\right) .$$

where $\psi(\cdot)$ is the digamma function. Normal distribution properties lead to

$$\begin{aligned}
&\forall k \in \mathcal{K},\ \forall p \in \mathcal{P} :\\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right] = \tilde{\mu}_{kp}^t ,\\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[(\mu_{kp}^t)^2\right] = \mathbb{V}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\mu_{kp}^t\right]^2\\
&\qquad\qquad = \sigma^2\tilde{\tau}_{kp}^{-1} + (\tilde{\mu}_{kp}^t)^2 ,\\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right] = \tilde{\boldsymbol{\omega}}_k ,\\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\boldsymbol{\omega}_k^T\right] = \mathbb{V}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right]\mathbb{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\omega}_k\right]^T\\
&\qquad\qquad = \sigma^2\tilde{\boldsymbol{\Lambda}}_k + \tilde{\boldsymbol{\omega}}_k\tilde{\boldsymbol{\omega}}_k^T ,
\end{aligned}$$

and Inverse Gamma distribution properties lead to

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{1}{\sigma^2}\right] = \frac{\tilde{\xi}_1}{\tilde{\xi}_2} ,\\
&\mathbb{E}_{\boldsymbol{\Theta}}\left[\log\sigma^2\right] = \log\tilde{\xi}_2 - \psi(\tilde{\xi}_1) .
\end{aligned}$$

Using all these properties, the following expectations can be calculated as

$$\forall j \in \mathcal{J},\ \forall k \in \mathcal{K},\ \forall p \in \mathcal{P} :$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \mu_{kp}^t)^2}{\sigma^2}\right] = \frac{\tilde{\xi}_1\left(x_{tj} - \tilde{\mu}_{kp}^t\right)^2}{\tilde{\xi}_2} + \tilde{\tau}_{kp}^{-1} ,$$
$$\mathbb{E}_{\boldsymbol{\Theta}}\left[\frac{(x_{tj} - \boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k)^2}{\sigma^2}\right] = \frac{\tilde{\xi}_1\left(x_{tj} - \boldsymbol{\Phi}(t_j)^T\tilde{\boldsymbol{\omega}}_k\right)^2}{\tilde{\xi}_2} + \mathrm{Trace}\left(\boldsymbol{\Phi}(t_j)^T\tilde{\boldsymbol{\Lambda}}_k\boldsymbol{\Phi}(t_j)\right) .$$

4.3.3 Complete model

A model integrating parabolic data, piecewise parabolic data and mixed data is now presented. By taking into consideration any types of available data, the resulting model can fit data better and can estimate more accurate clusters. First, data formalism and assumptions are detailed. Then, the resulting mixture model and its inference procedure are developed.

Data and assumptions

In this part, data consist of J pulses gathering J amplitudes $\boldsymbol{x}_t = (x_{tj})_{j\in\mathcal{J}}$ associated to J times of arrival $\boldsymbol{t} = (t_j)_{j\in\mathcal{J}}$, J continuous features $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ and J categorical features $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ from K distinct emitters. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj})$ the j^{th} observation vector of mixed variables where

- $\boldsymbol{x}_{qj} \in \mathbb{R}^d$ is a vector of d continuous radar features such as the Radio Frequency, the Pulse Width, the Azimuth or the Pulse Repetition Interval,
- $\boldsymbol{x}_{cj} = (x_{cj_0}, \ldots, x_{cj_{q-1}}) \in C_q$ is a vector of q categorical radar modulations such as intra-pulse modulations or pulse-to-pulse modulations,
- $x_{tj} \in \mathbb{R}$ is a continuous variable modeling the Amplitude.

For each pulse j, the temporal evolution variable x_{tj} and mixed variables $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$ are assumed to be independent conditionally to each cluster $k \in \mathcal{K}$

$$\forall j \in \mathcal{J},\ (\boldsymbol{x}_q, \boldsymbol{x}_c)|z_j = k \perp\!\!\!\perp x_t|z_j = k\ . \tag{4.42}$$

with z_j the latent variable modeling the label of the j^{th} observation vector $\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj})$. Moreover, the temporal evolution data $(x_{tj}, t_j)_{j\in\mathcal{J}}$ are distributed according to either a parabolic relation or a piecewise parabolic relation and the quantitative data $(\boldsymbol{x}_{qj})_{j\in\mathcal{J}}$ are normally distributed conditionally to categorical data $(\boldsymbol{x}_{cj})_{j\in\mathcal{J}}$. Both quantitative and categorical data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ can be partially observed. Hence $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j\in\mathcal{J}}$ are decomposed into observed features $(\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj}^{\text{obs}})_{j\in\mathcal{J}}$ and missing features $(\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{miss}})_{j\in\mathcal{J}}$ such that

$$\forall j \in \mathcal{J},\quad \begin{aligned} \boldsymbol{x}_{qj} &= \begin{pmatrix} \boldsymbol{x}_{qj}^{\text{miss}} \\ \boldsymbol{x}_{qj}^{\text{obs}} \end{pmatrix} \text{ with } (\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{qj}^{\text{obs}}) \in \mathbb{R}^{d_j^{\text{miss}}} \times \mathbb{R}^{d_j^{\text{obs}}} \text{ and } d_j^{\text{miss}} + d_j^{\text{obs}} = d\ , \\ \boldsymbol{x}_{cj} &= \begin{pmatrix} \boldsymbol{x}_{cj}^{\text{miss}} \\ \boldsymbol{x}_{cj}^{\text{obs}} \end{pmatrix} \text{with } (\boldsymbol{x}_{cj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{obs}}) \in \mathcal{C}_{q_j^{\text{miss}}} \times \mathcal{C}_{q_j^{\text{obs}}} \text{ and } q_j^{\text{miss}} + q_j^{\text{obs}} = q\ . \end{aligned}$$

Mixture model

According to the independence assumption (4.42), the distribution of mixed data (Chapter 3) and the parabolic and piecewise parabolic relations between temporal evolution data, the component distribution results in

$$\forall j \in \mathcal{J},\ p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}|z_j = k) = p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}|z_j = k)p(x_{tj}|z_j = k)$$

where

$$\forall j \in \mathcal{J},\quad \begin{aligned} p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}|z_j = k) &= \prod_{\boldsymbol{c}\in\mathcal{C}_q} \left(\pi_{k\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}}\ , \\ p(x_{tj}|t_j, z_j = k) &= \left(c_{k0}\mathcal{N}\left(\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right)\right)^{\delta_{w_j}^0} \left(c_{k1}\prod_{p\in\mathcal{P}}\left(b_{kp}\mathcal{N}\left(x_{tj}|\mu_{kp}^t, \sigma^2\right)\right)^{\delta_{y_j}^p}\right)^{\delta_{w_j}^1} \end{aligned} \tag{4.43}$$

with

- $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$ the scale latent variables handling outliers for quantitative data $\boldsymbol{x}_q$ and distributed according to a Gamma distribution with shape and rate parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})_{(k,\boldsymbol{c})\in\mathcal{K}\times\mathcal{C}_q}$ conditionally to categorical data $\boldsymbol{x}_c$ and labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$,

- $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = ((\boldsymbol{\mu}_{k\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}, \boldsymbol{\Sigma}_k)_{k\in\mathcal{K}}$ the mean and the variance parameters of quantitative data $\boldsymbol{x}_q$ for each cluster,
- $\boldsymbol{\pi} = (\boldsymbol{\pi}_k)_{k\in\mathcal{K}}$ the weights of the multivariate Categorical distribution of categorical data $\boldsymbol{x}_c$ for each cluster,
- $\boldsymbol{y} = (y_j)_{j\in\mathcal{J}}$ the latent variables indicating the p^{th} piecewise temporal evolution data $\boldsymbol{x}_t$ belong to,
- $\boldsymbol{w} = (w_j)_{j\in\mathcal{J}}$ the latent variables indicating if temporal evolution data $\boldsymbol{x}_t$ are distributing according to a parabolic relation or a piecewise parabolic relation,
- $\boldsymbol{b} = ((b_{kp})_{p\in\mathcal{P}})_{k\in\mathcal{K}}$ the weights of the Categorical distribution of latent variables $\boldsymbol{y}$,
- $\boldsymbol{c} = (c_{k0}, c_{k1})_{k\in\mathcal{K}}$ the weights of the Categorical distribution of latent variables $\boldsymbol{w}$,
- $\boldsymbol{\mu}^t = ((\mu^t_{kp})_{p\in\mathcal{P}})_{k\in\mathcal{K}}$ the set of piecewises for temporal evolution data $\boldsymbol{x}_t$ for each cluster,
- $\boldsymbol{\omega} = (\boldsymbol{\omega}_k)_{k\in\mathcal{K}}$ the regression parameters for temporal evolution data $\boldsymbol{x}_t$ for each cluster,
- σ^2 the variance of the measurement noise related to temporal evolution data $\boldsymbol{x}_t$.

Recalling that $p(z_j = k) = a_k$ where $\boldsymbol{a} = (a_k)_{k\in\mathcal{K}}$ are the weights related to component distributions, the mixture model is obtained from (4.43) such that $\forall j \in \mathcal{J}$,

$$p(\boldsymbol{x}_j, u_j, y_j | t_j, \boldsymbol{\Theta}) = \sum_{k\in\mathcal{K}} a_k \left(c_{k0}\mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right)\right)^{\delta^0_{w_j}} \left(c_{k1}\prod_{p\in\mathcal{P}}\left(b_{kp}\mathcal{N}\left(x_{tj}|\mu^t_{kp}, \sigma^2\right)\right)^{\delta^p_{y_j}}\right)^{\delta^1_{w_j}} \times \prod_{\boldsymbol{c}\in\mathcal{C}_q}\left(\pi_{k\boldsymbol{c}}\mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{G}(u_j|\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})\right)^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}} \tag{4.44}$$

where $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\mu}^t, \boldsymbol{\omega}, \sigma^2, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the set of parameters.

Bayesian framework

As in previous chapters, a Bayesian framework is used to estimate parameters $\boldsymbol{\Theta}$. Assuming datasets $(\boldsymbol{x} = (\boldsymbol{x}_q, \boldsymbol{x}_c, \boldsymbol{x}_t), \boldsymbol{t})$ of i.i.d observations $(\boldsymbol{x}_j = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}), t_j)_{j\in\mathcal{J}}$, independent labels $\boldsymbol{z} = (z_j)_{j\in\mathcal{J}}$, indicators $(\boldsymbol{y}, \boldsymbol{w}) = (y_j, w_j)_{j\in\mathcal{J}}$ and scale latent variables $\boldsymbol{u} = (u_j)_{j\in\mathcal{J}}$, the complete likelihood associated to (4.44) is defined by

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{z} | \boldsymbol{t}, \boldsymbol{\Theta}, \mathcal{K}) = \prod_{j\in\mathcal{J}}\prod_{k\in\mathcal{K}}\left(a_k\left(c_{k0}\mathcal{N}\left(x_{tj}|\boldsymbol{\Phi}(t_j)^T\boldsymbol{\omega}_k, \sigma^2\right)\right)^{\delta^0_{w_j}}\left(c_{k1}\prod_{p\in\mathcal{P}}\left(b_{kp}\mathcal{N}\left(x_{tj}|\mu^t_{kp}, \sigma^2\right)\right)^{\delta^p_{y_j}}\right)^{\delta^1_{w_j}} \times \prod_{\boldsymbol{c}\in\mathcal{C}_q}\left(\pi_{k\boldsymbol{c}}\mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{k\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{G}(u_j|\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})\right)^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}}\right)^{\delta^k_{z_j}}.$$

Eventually, the prior distribution required for $\boldsymbol{\Theta}$ is chosen as

$$p(\boldsymbol{\Theta}|\mathcal{K}) = p(\boldsymbol{a}|\mathcal{K})p(\boldsymbol{b}|\mathcal{K})p(\boldsymbol{c}|\mathcal{K})p(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K})p(\boldsymbol{\omega}|\sigma^2, \mathcal{K})p(\sigma^2)p(\boldsymbol{\pi}|\mathcal{K})p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K})p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K})$$

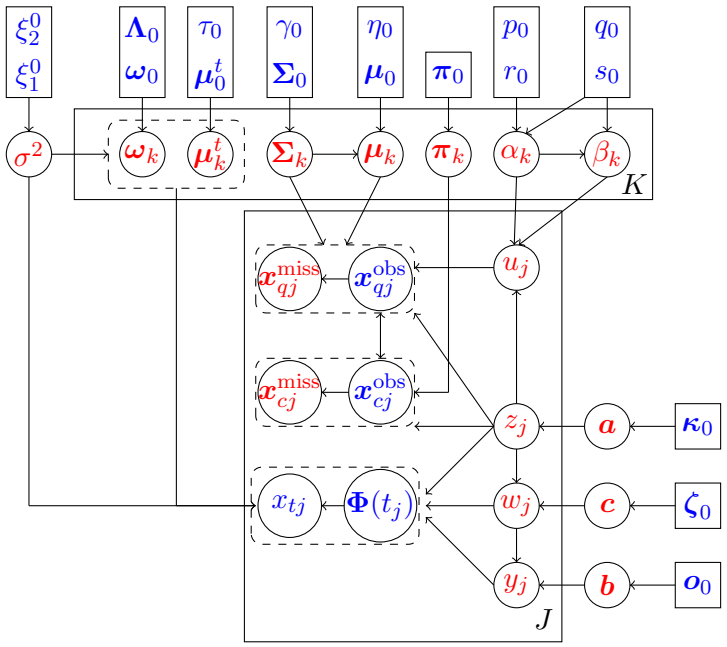


Figure 4.20: Graphical representation of the proposed model integrating temporal evolution data and mixed-type data. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}, x_{tj}, t_j)$ decomposed into temporal evolution data (x_{tj}, t_j) and mixed-type data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})$, the scale variables u_j and the indicator variables (y_j, w_j, z_j). Known quantities, respectively unknown quantities, are in blue, respectively in red.

where

$$
\begin{cases}
p(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\boldsymbol{\kappa}_0) \ , \\
p(\boldsymbol{b}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|\boldsymbol{o}_0) \ , \\
p(\boldsymbol{c}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{c}_k|\boldsymbol{\zeta}_0) \ , \\
p(\boldsymbol{\pi}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{\pi}_k|\boldsymbol{\pi}_0) \ , \\
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{\boldsymbol{c}\in\mathcal{C}_q} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\boldsymbol{\mu}_0, \eta_0^{-1}\boldsymbol{\Sigma}_k\right) \mathcal{IW}(\boldsymbol{\Sigma}_k|\gamma_0, \boldsymbol{\Sigma}_0) \ , \\
p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{\boldsymbol{c}\in\mathcal{C}_q} p(\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}}|p_0, q_0, s_0, r_0) \ , \\
p(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\mu_0^t, \tau_0^{-1}\sigma^2\right) \ , \\
p(\boldsymbol{\omega}|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\boldsymbol{\omega}_0, \sigma^2\boldsymbol{\Lambda}_0\right) \ , \\
p(\sigma^2) = \mathcal{IG}(\sigma^2|\xi_1^0, \xi_2^0) \ .
\end{cases}
$$

Graphical representation of the proposed model is shown in Figure 4.20.

Inference

As previously, a factorized posterior distribution $q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\Theta}|\mathcal{K}) = q(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z}|\mathcal{K})q(\boldsymbol{\Theta}|\mathcal{K})$ is chosen as an approximation of the intractable posterior joint distribution $p(\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\Theta}|\boldsymbol{x}^{\text{obs}}, \mathcal{K})$ such that latent variables $\boldsymbol{h} = (\boldsymbol{x}_q^{\text{miss}}, \boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z})$ and parameters $\boldsymbol{\Theta}$ are a posteriori independent and

$$
\begin{aligned}
q(\boldsymbol{h}|\mathcal{K}) &= q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K})q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K})q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z}, \mathcal{K})q(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{z}, \mathcal{K})q(\boldsymbol{w}|\boldsymbol{z}, \mathcal{K})q(\boldsymbol{z}|\mathcal{K}) \ , \\
q(\boldsymbol{\Theta}|\mathcal{K}) &= q(\boldsymbol{a}|\mathcal{K})q(\boldsymbol{b}|\mathcal{K})q(\boldsymbol{c}|\mathcal{K})q(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K})q(\boldsymbol{\omega}|\sigma^2, \mathcal{K})q(\sigma^2)q(\boldsymbol{\pi}|\mathcal{K})q(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K})q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K}) \ .
\end{aligned}
$$

According to VB assumptions, the following conjugate variational posterior distributions are obtained from the VB procedure

$$
\begin{cases}
q(\boldsymbol{x}_q^{\text{miss}}|\boldsymbol{u}, \boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{x}_{qj}^{\text{miss}}|\tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}}, u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\boldsymbol{x}_q^{\text{miss}}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k} \ , \\
q(\boldsymbol{u}|\boldsymbol{x}_c^{\text{miss}}, \boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{G}\left(u_j|\tilde{\alpha}_{jk\boldsymbol{c}}, \tilde{\beta}_{jk\boldsymbol{c}}\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}\delta_{z_j}^k} \ , \\
q(\boldsymbol{x}_c^{\text{miss}}|\boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \mathcal{MC}(\boldsymbol{x}_{cj}^{\text{miss}}|\tilde{\boldsymbol{r}}_{jk}^{\boldsymbol{x}_c^{\text{miss}}})^{\delta_{z_j}^k} \\
q(\boldsymbol{y}|\boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \mathcal{C}at(y_j|\tilde{\boldsymbol{r}}_{jk}^y)^{\delta_{w_j}^1\delta_{z_j}^k} \ , \\
q(\boldsymbol{w}|\boldsymbol{z}, \mathcal{K}) = \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{K}} \mathcal{C}at(w_j|\tilde{\boldsymbol{r}}_{jk}^w)^{\delta_{z_j}^k} \ , \\
q(\boldsymbol{z}|\mathcal{K}) = \prod_{j\in\mathcal{J}} \mathcal{C}at(z_j|\tilde{\boldsymbol{r}}_j) \ , \\
q(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}}) \ , \\
q(\boldsymbol{\pi}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{\pi}|\tilde{\boldsymbol{\pi}}_k) \ , \\
q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}}, \tilde{\eta}_{k\boldsymbol{c}}^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k, \tilde{\boldsymbol{\Sigma}}_k) \ , \\
q(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \prod_{k\in\mathcal{K}} p(\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}}|\tilde{p}_k, \tilde{q}_k, \tilde{s}_k, \tilde{r}_k) \ , \\
q(\boldsymbol{b}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{b}_k|\tilde{\boldsymbol{o}}_k) \ , \\
q(\boldsymbol{c}|\mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{D}(\boldsymbol{c}_k|\tilde{\boldsymbol{\zeta}}_k) \ , \\
q(\boldsymbol{\mu}^t|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}} \prod_{p\in\mathcal{P}} \mathcal{N}\left(\mu_{kp}^t|\tilde{\mu}_{kp}^t, \tilde{\tau}_{kp}^{-1}\sigma^2\right) \ , \\
q(\boldsymbol{\omega}|\sigma^2, \mathcal{K}) = \prod_{k\in\mathcal{K}} \mathcal{N}\left(\boldsymbol{\omega}_k|\tilde{\boldsymbol{\omega}}_k, \sigma^2\tilde{\boldsymbol{\Lambda}}_k\right) \ , \\
q(\sigma^2) = \mathcal{IG}(\sigma^2|\tilde{\xi}_1, \tilde{\xi}_2) \ .
\end{cases}
\tag{4.45}
$$

Their respective parameters are estimated during the VBE and VBM steps by developing expectations $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right]$ and $\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right]$. Noting that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] =& \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_t,\boldsymbol{y},\boldsymbol{w}|\boldsymbol{z},\boldsymbol{t},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\omega},\boldsymbol{\mu}^t,\sigma^2,\mathcal{K})\right]\\ &+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}_q,\boldsymbol{u},\boldsymbol{x}_c|\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma},\mathcal{K})\right]\\ &+ \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{z}|\boldsymbol{a},\mathcal{K})\right],\end{aligned} \tag{4.46}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right] =& \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_t,\boldsymbol{y},\boldsymbol{w},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\omega},\boldsymbol{\mu}^t,\sigma^2|\boldsymbol{z},\boldsymbol{t},\mathcal{K})\right]\\ &+ \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}_q,\boldsymbol{u},\boldsymbol{x}_c,\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{z},\mathcal{K})\right]\\ &+ \mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{z},\boldsymbol{a}|\mathcal{K})\right],\end{aligned} \tag{4.47}$$

the VBE (4.46) and VBM (4.47) steps can be independently derived for latent variables and parameters related to temporal evolution data $\boldsymbol{x}_t$ and mixed data $(\boldsymbol{x}_q,\boldsymbol{x}_c)$. Therefore, variational posterior distributions of latent variables $(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{u},\boldsymbol{x}_c^{\text{miss}})$ and parameters $(\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})$ related to mixed data $(\boldsymbol{x}_q,\boldsymbol{x}_c)$ are obtained as in Chapter 3 by deriving green expectations in (4.46) and (4.47). As for $(\boldsymbol{y},\boldsymbol{w},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\omega},\boldsymbol{\mu}^t,\sigma^2)$, their variational posterior distribution are obtained as in subsection 4.3.2 by developing blue expectations in (4.46) and (4.47). As in subsection 4.3.2 or in Chapter 3, the Dirichlet posterior distribution of $\boldsymbol{a}$ is deduced from the red expectation in (4.47). Eventually, the variational distribution of labels $\boldsymbol{z}$ is obtained by marginalising over latent variables in the green expectation and developing both blue and red expectations in (4.46) such that

$$\int \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right]\partial\boldsymbol{x}_q^{\text{miss}}\partial\boldsymbol{y}\partial\boldsymbol{w}\partial\boldsymbol{u}\partial\boldsymbol{x}_c^{\text{miss}} = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\delta_{z_j}^k\log\rho_{jk}$$

where $\log\rho_{jk}$ is deduced from red, blue and green expectations in (4.46) as follows

$$\forall j\in\mathcal{J},\ \forall k\in\mathcal{K},\ \log\rho_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] + \log\rho_{jk}^t + \log\rho_{jk}^{qc}\,. \tag{4.48}$$

with

$$\begin{aligned}&\forall j\in\mathcal{J},\ \forall k\in\mathcal{K},\\ &\qquad \mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k\in\mathcal{K}}\tilde{\kappa}_k\right),\\ &\qquad \log\rho_{jk}^t = \log\sum_{p\in\mathcal{P}}\rho_{jkp}^w,\\ &\qquad \log\rho_{jk}^{qc} = \log\sum_{\boldsymbol{c}^{\text{miss}}\in\mathcal{C}_{q_j^{\text{miss}}}}\rho_{k\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_{cj}^{\text{miss}}}.\end{aligned}$$

The red term $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log a_k\right]$ is deduced from properties of the Dirichlet distribution, the blue term $\log\rho_{jk}^t$ is deduced from (4.37) and (4.38) in subsection 4.3.2 and the green term $\log\rho_{jk}^{qc}$ has been detailed in Chapter 3. Hence, $\boldsymbol{z}$ is distributed a posteriori according to a product of Categorical distributions parametrized by $\tilde{\boldsymbol{r}} = (\tilde{r}_{jk})_{(j,k)\in\mathcal{J}\times\mathcal{K}}$ given by

$$\forall j\in\mathcal{J},\ \forall k\in\mathcal{K},\ \tilde{r}_{jk} = \frac{\rho_{jk}}{\sum_{k\in\mathcal{K}}\rho_{jk}} \tag{4.49}$$

The lower bound on the log evidence is still required to master the VB inference and can be also decomposed into terms related to temporal evolution data (blue terms), mixed data (green terms) and labels $\boldsymbol{z}$ (red terms). This decomposition is obtained as follows

$$\mathcal{L}(q|\mathcal{K}) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z},\boldsymbol{\Theta}|\mathcal{K})\right]$$

where the free energy can be developed as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z},\boldsymbol{\Theta}|\boldsymbol{t},\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{b},\boldsymbol{c},\boldsymbol{\omega},\boldsymbol{\mu}^t,\sigma^2|\mathcal{K})\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{a},\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})\right]\end{aligned}$$

and the entropy as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z},\boldsymbol{\Theta}|\mathcal{K})\right] &= \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{b},\boldsymbol{c},\boldsymbol{\omega},\boldsymbol{\mu}^t,\sigma^2|\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{z},\mathcal{K})\right] \\ &\quad + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{x}_q^{\text{miss}},\boldsymbol{x}_c^{\text{miss}},\boldsymbol{u}|\boldsymbol{z},\mathcal{K})\right] + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] \\ &\quad + \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{a},\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{K})\right] .\end{aligned}$$

Blue terms, respectively green terms, have been previously detailed in subsection 4.3.2 , respectively in Chapter 3. As for red terms, they are detailed below :

$$\begin{aligned}\mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{y},\boldsymbol{w},\boldsymbol{z}|\boldsymbol{t},\boldsymbol{\Theta},\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\rho_{jk} , \\ \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{z}|\mathcal{K})\right] &= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{h}}\left[\delta_{z_j}^k\right]\log\tilde{r}_{jk} .\end{aligned}$$

4.3.4 Experiments

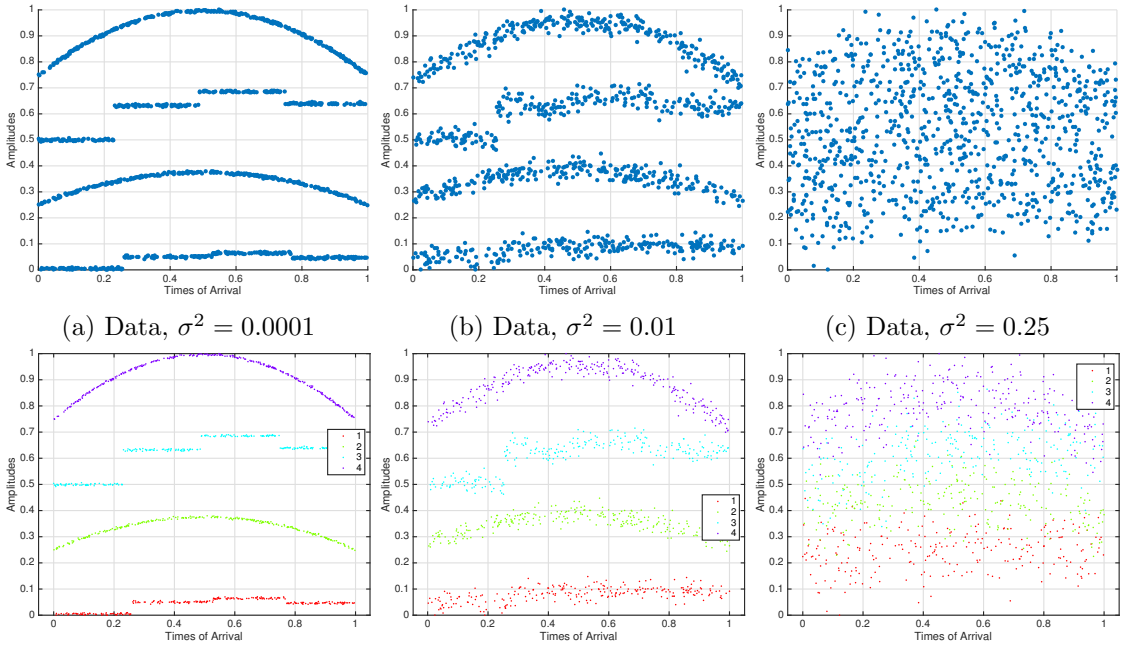
Two experiments are carried out to evaluate clustering performance with respect to a set of synthetic data. In the first experiment, only temporal evolution data are taken into consideration in the clustering procedure. Then, both temporal evolution data and quantitative data are considered in the second one. For comparison, the spectral clustering [VL07] and the k-means algorithm from [HW79] are also evaluated. First, characteristics of data, comparison algorithms and evaluation metrics are detailed. Then, both experiments are described and performance are shown to exhibit the effectiveness of the proposed model.

Data, algorithms and metrics

Synthetic data are composed of temporal evolution data related to amplitudes which are distributed according to either a parabolic relation or a piecewise parabolic relation and quantitative data related to continuous radar features which are jointly distributed according to a multivariate normal distribution. Temporal evolution data are generated by sampling a set of data from 2 parabolas and 2 piecewise parabolas directed by

$$\boldsymbol{\omega} = \begin{pmatrix} -1 & -2 & -3 & -4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} .$$

For each piecewise parabola, $p = 4$ piecewises are obtained by dividing the time interval in p equal subintervals and assigning to each piecewise the value of the parabola at the minimal time of its related time subinterval. Quantitative data are generated by sampling a set of data from



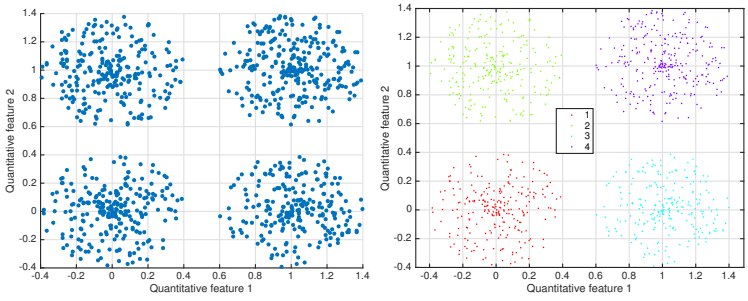
(a) Data, $\sigma^2 = 0.0001$ (b) Data, $\sigma^2 = 0.01$ (c) Data, $\sigma^2 = 0.25$

(d) Ground-Truth, $\sigma^2 = 0.0001$ (e) Ground-Truth, $\sigma^2 = 0.01$ (f) Ground-Truth, $\sigma^2 = 0.25$

Figure 4.21: Synthetic parabolic and piecewise parabolic data generated from different values of the variance parameter σ^2. Figures (a), (b) and (c) present unlabeled data where 4 parabolas are generated. Ground-truth are visible on Figures (d), (e) and (f).

four well-separated bivariate clusters with centers $[0, 0]^T$, $[1, 0]^T$,$[0, 1]^T$ and $[1, 1]^T$ and identity covariance matrices. Three synthetic datasets are generated with respect to a range of values of σ^2 and are linearly transformed by a min-max normalization to meet algorithms requirements. These datasets are shown in Figures 4.21 and 4.22.

Except for the k-means algorithm, an initialisation is required for clustering algorithms that are involved in these experiments. The similarity graph required for the spectral clustering is obtained from a k-nearest neighbor graph as suggested in [VL07] where the number of neighbors k is chosen as the product of the log number of observations and the number of clusters. As for the proposed model, a supervised initialisation is retained due to its sensitivity to initialisation. First, prior hyperparameters ξ_1^0 and ξ_2^0 are initialised such that the prior mean $\mathbb{E}[\frac{1}{\sigma^2}] = \frac{\xi_1^0}{\xi_2^0}$ of



(a) Quantitative Data (b) Ground-Truth

Figure 4.22: Synthetic quantitative data generated from 4 multivariate normal distributions. Figure (a) shows unlabeled data and Figure (b) exhibits the ground-truth.

Table 4.9: Initialisation of hyper-parameters values for clustering on parabolic and piecewise parabolic data

$\boldsymbol{\omega}_0$	$\boldsymbol{\Lambda}_0$	τ_0	κ_0	η_0	γ_0	p_0	r_0	q_0	s_0
$(0,0,0)^T$	I_3	1	0.5	100	1	1	1	1	1

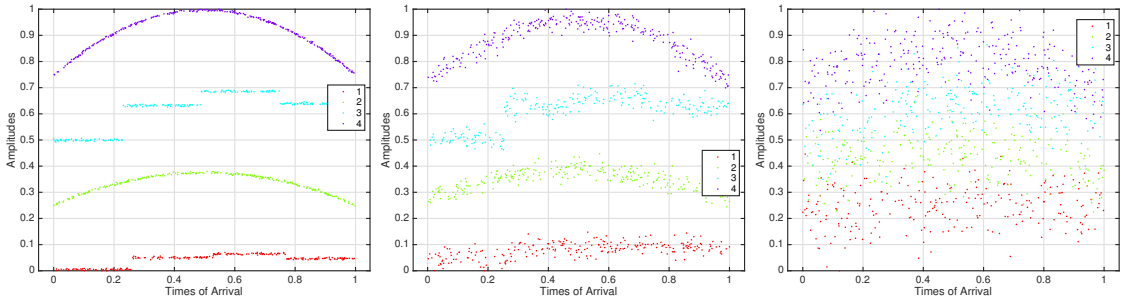
the variance parameter σ^2 is equal to the inverse of the determinant of the covariance matrix of temporal evolution data points. This choice is motivated by the fact that the determinant of the covariance matrix can be interpreted as the generalized variance that reflects the overall spread of the data. Setting $\xi_2^0 = 1$, ξ_1^0 is initialised as the inverse of the generalized variance of the sample of temporal evolution data. In addition, prior piecewise means μ_0^t are initialised from results of a k-means algorithm on temporal evolution data. Then, prior component means $\boldsymbol{\mu}_0$, respectively covariance matrices $\boldsymbol{\Sigma}_0$, are initialised from results of a k-means algorithm on quantitative data, respectively from diagonal matrices whose diagonal elements are variances of quantitative data. Other hyper-parameters are initialised as in Table 4.9.

Performance on synthetic data are evaluated through the Adjusted Rand Index (ARI) [HA85] that compares estimated partitions of data with the ground-truth and the Silhouette Coefficient [KR09] which does not require the ground-truth and provides a higher score when clusters are dense and well separated.

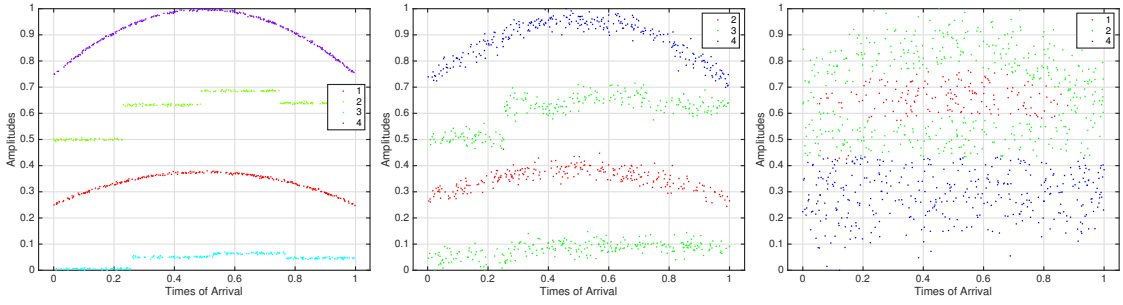
Experiments and results

The first experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters K when only temporal evolution data are taken into consideration. According to datasets visualised in Figure 4.21, K is set to 4 for synthetic data. Results of the first experiment on synthetic data are shown in Figure 4.23 and in Table 4.10. When $\sigma^2 \in \{0.0001, 0.01\}$, the proposed model and the spectral clustering have similar performance in clustering synthetic data with an ARI equals to 0.73 while the k-means algorithm has the lowest performance (ARI = 0.35) in creating convex and isotropic clusters that cannot handle the parabolic and piecewise parabolic structures of the generated data. This limitation is emphasized by higher Silhouette Coefficients of the k-means algorithm whereas the non-convexity of the data is confirmed by the lower Silhouette Coefficients of the ground-truth. Even if all algorithms poorly perform when data are embedded in noise ($\sigma^2 = 0.25$), the proposed algorithm estimates clusters with a more parabolic shape than other algorithms which build more isotropic clusters (Subfigures (f), (i) and (l) in Figure 4.10). Indeed the Silhouette Coefficient of the proposed model ($S = 0.14$) is closer to the Silhouette Coefficient of the ground-truth ($S = 0.10$) than Silhouette Coefficients of spectral clustering ($S = 0.53$) and k-means ($S = 0.56$).

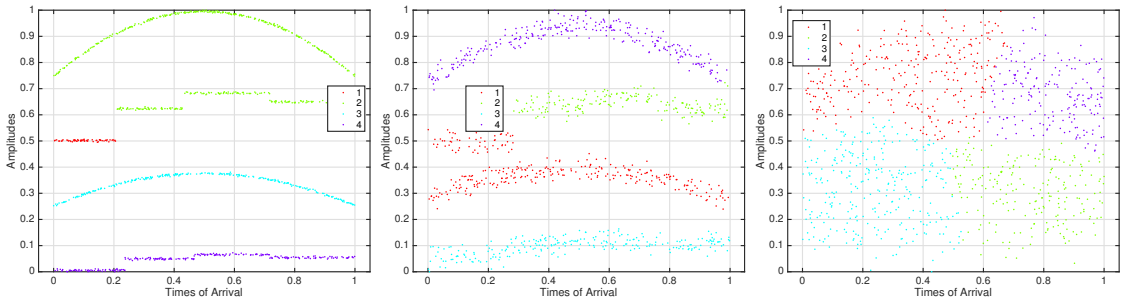
The second experiment aims to determine the ability of each algorithm to restore the true clusters according to an a priori number of clusters K when all types of data are taken into consideration. The number of clusters K is still set to 4 for synthetic data.. Results of the second experiment on synthetic data are shown in Figure 4.24 and in Table 4.11. All algorithms succeed in clustering synthetic data for $\sigma^2 \in \{0.0001, 0.01, 0.25\}$ since the ground-truth partition is recovered in Figure 4.24 with an ARI equals to 1 visible on Table 4.11. Adding quantitative information enables algorithms to recover the ground-truth for any value of σ^2.



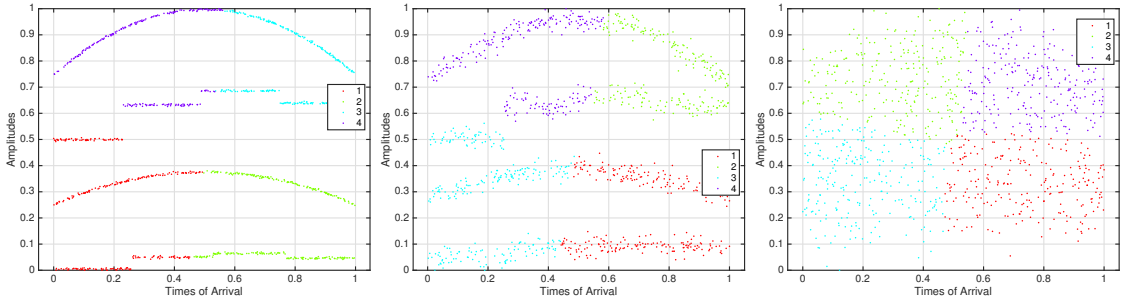
(a) Ground-Truth, $\sigma^2 = 0.0001$ (b) Ground-Truth, $\sigma^2 = 0.01$ (c) Ground-Truth, $\sigma^2 = 0.25$



(d) PM, $\sigma^2 = 0.0001$ (e) PM, $\sigma^2 = 0.01$ (f) PM, $\sigma^2 = 0.25$

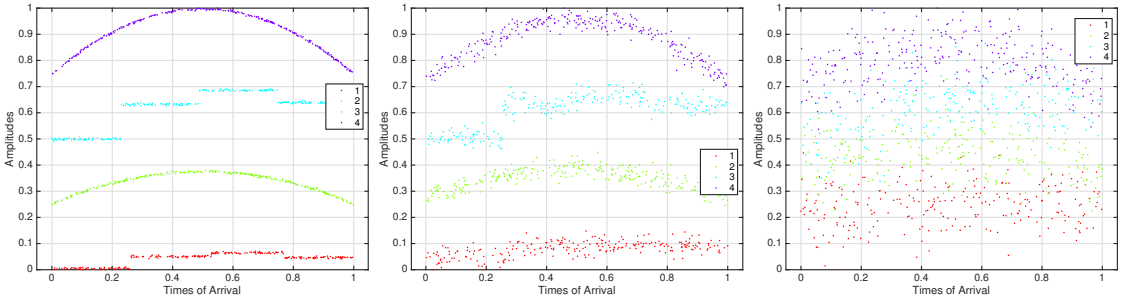


(g) SC, $\sigma^2 = 0.0001$ (h) SC, $\sigma^2 = 0.01$ (i) SC, $\sigma^2 = 0.25$

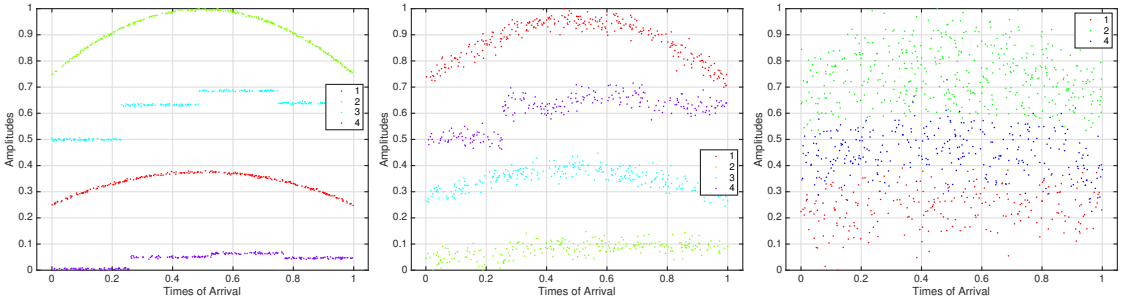


(j) KM, $\sigma^2 = 0.0001$ (k) KM, $\sigma^2 = 0.01$ (l) KM, $\sigma^2 = 0.25$

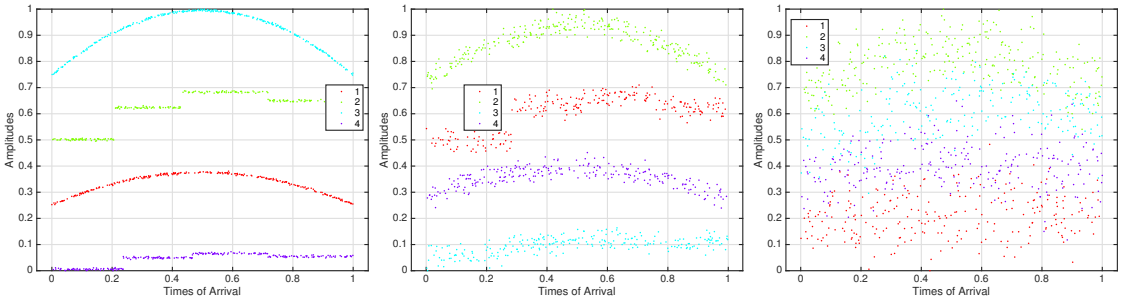
Figure 4.23: Results on synthetic data when only temporal evolution data are considered. Figures (a), (b) and (c) show synthetic data generated with different values of the variance σ^2. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).



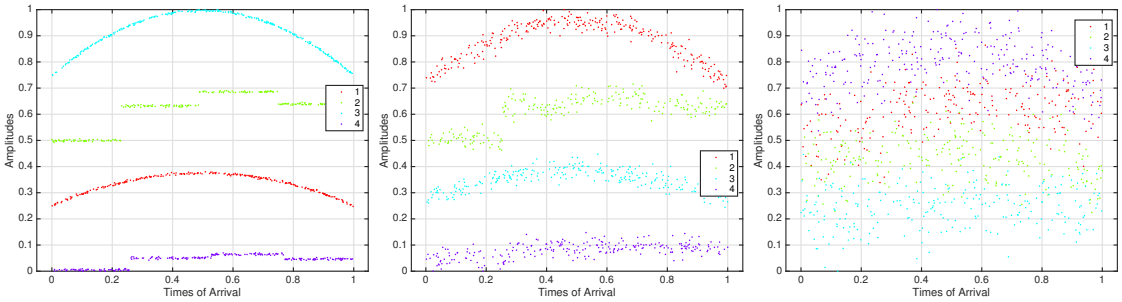
(a) Ground-Truth, $\sigma^2 = 0.0001$ (b) Ground-Truth, $\sigma^2 = 0.01$ (c) Ground-Truth, $\sigma^2 = 0.25$



(d) PM, $\sigma^2 = 0.0001$ (e) PM, $\sigma^2 = 0.01$ (f) PM, $\sigma^2 = 0.25$



(g) SC, $\sigma^2 = 0.0001$ (h) SC, $\sigma^2 = 0.01$ (i) SC, $\sigma^2 = 0.25$



(j) KM, $\sigma^2 = 0.0001$ (k) KM, $\sigma^2 = 0.01$ (l) KM, $\sigma^2 = 0.25$

Figure 4.24: Results on synthetic data when any types of data are considered. Figures (a), (b) and (c) show synthetic data generated with different values of the variance σ^2. Figures (d), (e) and (f) show clustering results of the proposed model (PM). Figures (g), (h) and (i) show clustering results of the spectral clustering (SC). Figures (j), (k) and (l) show clustering results of the k-means algorithm (KM).

Table 4.10: Adjusted Rand Index (ARI) and Silhouette coefficient (S) values for the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the first experiment on synthetic data when only temporal evolution data are considered.

	ARI			S			
	PM	SC	KM	Data	PM	SC	KM
$\sigma^2 = 0.0001$	0.73	0.73	0.35	0.30	0.35	0.35	0.64
$\sigma^2 = 0.01$	0.76	0.84	0.34	0.26	0.31	0.33	0.64
$\sigma^2 = 0.25$	0.54	0.26	0.27	0.10	0.14	0.53	0.56

Table 4.11: Adjusted Rand Index (ARI) and Silhouette coefficient (S) values for the proposed model (PM), the spectral clustering (SC) and the k-means algorithm (KM) during the first experiment on synthetic data when only temporal evolution data are considered.

	ARI			S			
	PM	SC	KM	Data	PM	SC	KM
$\sigma^2 = 0.0001$	0.73	0.73	0.35	0.30	0.35	0.35	0.64
$\sigma^2 = 0.01$	0.76	0.84	0.34	0.26	0.31	0.33	0.64
$\sigma^2 = 0.25$	0.54	0.26	0.27	0.10	0.14	0.53	0.56

4.4 Conclusion

In this chapter, two types of scanning behaviours have been presented. They can be observed in data when amplitudes and times of arrival shared either a parabolic relation or a piecewise parabolic relation. Since scanning behaviours fully characterise radar emitters, they can be taken into consideration to cluster radar emitters. Hence, mixture models handling parabolic data and piecewise parabolic data have been developed. Three approaches have been investigated : the first one assumes that temporal evolution data are only distributed according to parabolic relations, the second one mainly focuses on piecewise parabolic data and the last one handles both types of relations. Moroever in each approach, the proposed mixture model is enhanced with the mixture model designed for mixed data in Chapter 3 in order to improve clustering performance. Then, parameter estimation has been derived from the Variational Bayesian inference and experiments on real and synthetic data have exhibited the effectiveness of these three approaches.

Chapter 5

Conclusion and perspectives

Radar emitter identification is a crucial function of ESM systems since it prevents enemy forces from surprise attacks by detecting enemy radar signals and it improves military databases by analyzing unknown signals. Depending on radar emitter function and geopolitical context, radar emitters can emit complex signals based on pulse-to-pulse modulation patterns. Radar signal patterns can be decomposed into continuous features given by the continuous parameters of radar pattern pulses and categorical features that represent modulations of pulse sequences. Furthermore, radar signals are often partially observed in the electromagnetic environment due to failures of deinterleaving techniques or sensors deficiencies. Therefore, a framework handling any types of data has been developed in this work to perform classification and clustering of radar emitters even in presence of outliers and missing data.

State-of-the-art algorithms have been reviewed in Chapter 1. They perform either classification or clustering by learning boundaries that separate data into heterogeneous clusters or by learning underlying structure of data to constitute clusters. However, neither of these algorithms provides an internal framework that infers on missing data and copes with any degrees of supervision or any types of data. Therefore, an approach based on mixture models has been proposed. Theoretical aspects of mixture models have been introduced by detailing ways of modeling and estimating them for a generic type of data. Indeed, mixture models benefit from a flexible and probabilistic framework to handle outliers and missing data by introducing a latent space where each latent variable focuses on a specific constraint. However, the resulting model is not tractable and model learning is processed through Variational Bayes Approximation. Eventually whatever degree of supervision is required, the number of classes K and parameters can be estimated to perform classification and clustering tasks. Nonetheless, the Variational Bayes Appromixation tends to under-estimate uncertainties related to variational estimation. To this end, the Expectation propagation algorithm [Min01], which minimizes the reversed Kullback-Leibler divergence, can be implemented in future works to improve estimation accuracy.

Chapter 2 has focused on classification and clustering of continuous data with a scale mixture of Normal distributions accounting for missing data and outliers. Benefiting from Gaussian properties and the introduction of latent variables, the proposed model has shown its efficiency for inferring on missing data, performing classification and clustering tasks and selecting the correct number of clusters in a dataset obtained from an experimental protocol generating realistic data. A major contribution in this chapter is the incorporation of latent variables handling missing data and provided with a variational posterior distribution that leads to a more effective inference on missing data. As pointed out in both experiments, the effectiveness of the proposed model results from the fact that standard missing data imputation methods can create outliers that deteriorate

performance of classification and clustering algorithms whereas in the proposed model inference on missing data and labels prediction are jointly estimated. Indeed, embedding the inference procedure into the model framework allows properties of the model, such as outliers handling, to counterbalance drawbacks of imputation methods such as outlier creation. As for outliers handling, a full Bayesian approach has been adopted to avoid using the deterministic variable ν_k parametrizing the distribution of the latent variable u modeling outliers. This full Bayesian treatment enables a less restrictive modeling of data since parameters of the latent variable u are estimated into the Variational Bayes Approximation framework instead of being updated *via* an optimization procedure. Despite these advantages, the proposed model has a higher computational cost than comparison algorithms especially during the model learning step. Hence, a parallelization of the proposed model would be useful in order to reduce its computational burden.

As for Chapter 3, it has presented the general case where both continuous and categorical data are used for classification and clustering tasks. This chapter has precisely focused on modeling dependencies between continuous and categorical data in order to infer on missing data while performing classification and clustering. To this end, an approach based on the Location Mixture Model has been investigated on by establishing conditional relations between continuous and categorical data to tackle issues related to outliers and missing data. The developed approach has exhibited its effectiveness for inferring on missing data, performing classification and clustering tasks and selecting the correct number of clusters even for high proportions of missing values. As pointed out in Chapter 2, the proposed approach enables joint estimation of missing components and labels by embedding the inference procedure into the model framework. Indeed, estimating the missing components conditionally to their labels proves to be more effective than standard imputation methods which do not take into consideration information related to the data partition. Moreover, experiments have pointed out that using continuous and categorical data can really improve classification and clustering performance than considering either only continuous data or only categorical data. Indeed, higher performance on mixed data lie in a more relevant separation of clusters obtained by taking advantage of the more complex structure of mixed-type data. In this work, the Location Mixture Model assumption has been naturally considered since radar pattern designers use to define pulse modulation sequences (categorical radar data) before assigning pulse values (continuous radar data). However in the Electronic Warfare context, continuous radar data are first measured by sensors before being processed to deduce categorical data related to modulation patterns. Therefore, it would be interesting to investigate the Underlying Response Variable approach [Eve88, EM90] to assess assumption on conditioning categorical radar data by continuous radar data.

In Chapter 4, two types of scanning behaviours have been presented. They can be observed in radar data when amplitudes and times of arrival of radar emitters share either a parabolic or a piecewise parabolic relation. Since scanning behaviours fully characterise radar emitters, they have been taken into consideration to cluster radar emitters. To this end, both types of relation have been integrated into the mixture model framework by modeling the parabolic relation as a Bayesian regression and the piecewise parabolic relation as a mixture of normal distributions. As in Chapters 2 and 3, estimation of parameters has been processed through the Variational Bayesian inference and experiments on real and synthetic data have exhibited the effectiveness of the proposed model. Indeed, the resulting model enables creation of non isotropic clusters which better fit parabolic data than standard algorithms as the k-means algorithm. Nonetheless, when radar emitters share similar scanning behaviours, namely when amplitude parabolas of radar emitters intersect during their scanning period, the proposed model cannot perfectly identify radar emitters. Therefore, the proposed mixture model has been enhanced with the

mixture model designed for mixed data in Chapter 3 in order to improve clustering performance. Then, the complete model has managed to separate radar emitters in real operational cases by taking advantage of the whole available information related to radar emitters. In this work, only radar emitters presenting parabolic scanning behaviours have been clustered in real data cases. Hence, piecewise parabolic scanning behaviours remain to be evaluated on real data cases in future practical studies. Moreover, a radar signal pattern could be modeled as a Markovian process through its temporal evolution, continuous and categorical features since it is mainly defined as a pattern of pulses whose features share sequential and conditional relations. To this end, a clustering method, that integrates a Markovian process while handling missing data and outliers, could be developed in future works.

To conclude, the different mixture models developed in this work have focused on performing classification and clustering on various types of real and simulated data while handling outliers and missing values. These models have managed to reach high performance in classification and clustering tasks even in presence of large proportions of missing data and have proposed an effective inference procedure to reconstruct missing features. These models have been assessed on datasets gathering around 50 radar emitters according to the different experiments. Nonetheless, real databases may contain thousands of radar emitters. Therefore, the proposed models have to be tested on larger databases before being integrated into industrial and operational products.

Publications

Journal Paper

- Revillon, Guillaume; Mohammad-Djafari, Ali; Enderli, Cyrille: 'Radar emitters classification and clustering with a scale mixture of normal distributions', IET Radar, Sonar & Navigation, 2019, 13, (1), p. 128-138, DOI: 10.1049/iet-rsn.2018.5202

International Conferences

- G. Revillon, A. Mohammad-Djafari and C. Enderli, "Radar emitters classification and clustering with a scale mixture of normal distributions," 2018 IEEE Radar Conference (RadarConf18), Oklahoma City, OK, 2018, pp. 1371-1376. doi: 10.1109/RADAR.2018.8378764
- G. Revillon, A. Mohammad-Djafari and C. Enderli, "Radar Emitters Clustering With Outliers and Missing Data", International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (Max Ent 2018), London, UK, 2018

Extended abstract in french

Dans le contexte de la Guerre Electronique, l'identification des signaux radar est un atout majeur de la prise de décisions tactiques liées au théâtre d'opérations militaires. En fournissant des informations sur la présence de menaces, la classification et le partitionnement des signaux radar ont alors un rôle crucial assurant un choix adapté des contre-mesures dédiées à ces menaces et permettant la détection de signaux radar inconnus pour la mise à jour des bases de données. Les systèmes de Mesures de Soutien Electronique enregistrent la plupart du temps des mélanges de signaux radar provenant de différents émetteurs présents dans l'environnement électromagnétique. Le signal radar, décrit par un motif de modulations impulsionnelles, est alors souvent partiellement observé du fait de mesures manquantes et aberrantes. Le processus d'identification se fonde sur l'analyse statistique des paramètres mesurables du signal radar qui le caractérisent tant quantitativement que qualitativement. De nombreuses approches mêlant des techniques de fusion de données et d'apprentissage statistique ont été développées. Cependant, ces algorithmes ne peuvent pas à la fois effectuer la classification ainsi que le partionnement des émetteurs radar et gérer les données manquantes. A cette fin, des méthodes de substitution de données sont requises en amont de la classification et du partitionnement mais leur utilisation entraine l'apparition de nouvelles valeurs aberrantes. L'objectif principal de cette thèse est alors de définir un modèle de classification et partitionnement intégrant la gestion des valeurs aberrantes et manquantes présentes dans tout type de données. Une approche fondée sur les modèles de mélange de lois de probabilité est proposée dans cette thèse. Les modèles de mélange fournissent un formalisme mathématique flexible favorisant l'introduction de variables latentes permettant la gestion des données aberrantes et la modélisation des données manquantes dans les problèmes de classification et de partitionnement. L'apprentissage du modèle ainsi que la classification et le partitionnement sont réalisés dans un cadre d'inférence bayésienne où une méthode d'approximation variationnelle est introduite afin d'estimer la loi jointe a posteriori des variables latentes et des paramètres. Des expériences sur diverses données montrent que la méthode proposée fournit de meilleurs résultats que les algorithmes standards.

Le premier chapitre de cette thèse présente les différents algorithmes de l'état de l'art en matière de classification et de partitionnement. Tant par l'apprentissage de frontières au sein des données que par l'apprentissage d'une structure sous-jacente des données, ces algorithmes répondent aux problématiques de classification et partitionnement en créant des groupes hétérogènes d'observations. Cependant, aucun de ces algorithmes ne peut à la fois intégrer des données continues et catégorielles, gérer les contraintes liées aux données manquantes et s'adapter à différents degrés de supervision. L'approche fondée sur les modèles de mélange de lois de probabilité est alors introduite afin de palier ces divers problèmes. En effet, les modèles de mélange fournissent un cadre probabiliste favorisant l'introduction de variables latentes permettant l'intégration de données de tout type, la gestion de données aberrantes ainsi que la modélisation des données manquantes dans les problèmes de classification et de partitionnement. Néanmoins, cette approche requiert l'utilisation d'une méthode d'approximation bayésienne appelée Variational Bayes et

dont les aspects théoriques sont détaillés dans ce chapitre.

Le deuxième chapitre introduit un modèle de mélange de lois gaussiennes afin de prendre en compte les données continues représentant les paramètres physiques des impulsions. Ce modèle de mélange gaussien est ensuite mis à jour *via* l'introduction de variables latentes modélisant les valeurs aberrantes et manquantes afin d'obtenir un modèle robuste à ces deux contraintes. L'inférence est menée au travers de la méthode Variational Bayes permettant d'obtenir une approximation de la distribution jointe a posteriori des paramètres et des variables latentes du modèle. Enfin, ce modèle est testé sur des données acquises à l'aide d'un protocole expérimental fournissant des données réalistes intégrant les contraintes des systèmes d'acquisition opérationnels.

Le troisième chapitre intègre les données categorielles au modèle précédent en conditionnant les variables continues d'une observation par ses variables catégorielles. Le modèle obtenu est alors un mélange de lois gaussiennes conditionnelles intégrant également des variables latentes modélisant les valeurs manquantes propres aux observations catégorielles. L'inference est à nouveau faite par le biais de l'approximation variationelle bayésienne afin d'obtenir la distribution jointe a posterior des paramètres et variables latentes du modèle. Les performances du modèle proposé sont ensuite testées sur des données générées à partir d'une base de données réelles comportant 55 émetteurs radar avec des motifs impulsionnels variés.

Enfin, le dernier chapitre se focalise sur le caractère temporel des données impulsionnelles. En effet, l'évolution temporelle des amplitudes liées aux impulsions d'un émetteur radar présente une forme parabolique qui peut être exploitée afin d'améliorer la classification et le partionnement des émetteurs radar. Dans ce but, cette relation parabolique est modélisée par le biais d'une régression parabolique bayésienne intégrée au modèle de mélange. Les paramètres du modèle sont alors estimés par le biais de la précédente méthode d'approximation variationnelle et le modèle résultant est testé sur des données synthétiques et réelles provenant de différents cas opérationnels.

En conclusion, les différents modèles développés dans cette thèse ont permis la classification et le partionnement d'émetteurs radar caractérisés par des motifs impulsionnels présentant des valeurs manquantes et aberrantes tant continues que catégorielles. Ces modèles ont démontré leur efficacité en réalisant de bonnes performances sur des bases de données synthétiques et réelles même en présence d'une grande proportion de valeurs manquantes. Néanmoins, les bases de données réelles peuvent contenir des milliers d'émetteurs radar et les modèles proposés doivent alors être mis à l'épreuve sur de plus grandes bases de données avant d'être intégrés dans de futurs produits industriels.

Bibliography

[AKA91] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, Jan 1991. 7

[Aka98] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998. 19

[Alt92] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. 8

[AM74] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102, 1974. 2, 26, 29

[Att99] Hagai Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 21–30. Morgan Kaufmann Publishers Inc., 1999. 18

[AV07] Cédric Archambeau and Michel Verleysen. Robust Bayesian clustering. *Neural Networks*, 20(1):129–138, 2007. 26, 29, 30, 33

[AWT18] AWT Global. Radar scan types, 2012-2018. 58

[Bar53] RH Barker. Group synchronization of binary digital systems. *Communication theory*, pages 273–287, 1953. 56

[BCG00] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000. 2, 15, 19, 26

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992. 9

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. 2, 15, 16

[BR93] Jeffrey D. Banfield and Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993. 6

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 2, 9, 26

[Bre17] Leo Breiman. *Classification and regression trees*. Routledge, 2017. 8

[But98] Joseph MacKay Butler. *Tracking and control in multi-function radar*. PhD thesis, University of London, 1998. 1

[BWW15] Marion Byrne, Kruger White, and Jason Williams. Scheduling multifunction radar for search and tracking. In *Information Fusion (Fusion), 2015 18th International Conference on*, pages 945–952. IEEE, 2015. 1

[Che17] Wenbin Chen. Radar emitter classification for large data set based on weighted-xgboost. *IET Radar, Sonar & Navigation*, 11:1203–1207(4), August 2017. 1, 26

[CLH14] Yee Ming Chen, Chih-Min Lin, and Chi-Shun Hsueh. Emitter identification of electronic intelligence system using type-2 fuzzy classifier. *Systems Science & Control Engineering: An Open Access Journal*, 2(1):389–397, 2014. 1

[Cos84] John P Costas. A study of a class of detection waveforms having nearly ideal range. Doppler ambiguity properties. *Proceedings of the IEEE*, 72(8):996–1009, 1984. 55

[Cox58] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958. 7

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 9

[DGK04] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556. ACM, 2004. 11, 12

[DH82] CL Davies and P Hollands. Automatic processing for ESM. In *IEE Proceedings F (Communications, Radar and Signal Processing)*, volume 129, pages 164–171. IET, 1982. 22, 23

[DK13] Janusz Dudczyk and Adam Kawalec. Identification of emitter sources in the aspect of their fractal features. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 61(3):623–628, 2013. 1

[DK15a] J Dudczyk and A Kawalec. Fast-decision identification algorithm of emission source pattern in database. *Bulletin of the Polish Academy of Sciences Technical Sciences*, 63(2):385–389, 2015. 1

[DK15b] Janusz Dudczyk and Adam Kawalec. Specific emitter identification based on graphical representation of the distribution of radar signal parameters. *Bulletin of the Polish Academy of Sciences Technical Sciences*, 63(2):391–396, 2015. 1

[DLGMD17] Mircea Dumitru, Wang Li, Nicolas Gac, and Ali Mohammad-Djafari. Performance comparison of Bayesian iterative algorithms for three classes of sparsity enforcing priors with application in computed tomography. In *2017 IEEE International Conference on Image Processing*, 2017. 31

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. 16, 18

[Dud16] Janusz Dudczyk. Radar emission sources identification based on hierarchical agglomerative clustering for large data sets. *Journal of Sensors*, 2016, 2016. 1

[EKS+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996. 2, 11, 26, 46, 79

[EM90] Brian S Everitt and C Merette. The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics*, 17(3):283–297, 1990. 60, 148

[Eve88] B.S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, 6(5):305 – 309, 1988. 60, 148

[FHT+00] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000. 9

[Fis36] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188, 1936. 6

[FLP+07] Gerhard Fettweis, Michael Löhning, Denis Petrovic, Marcus Windisch, Peter Zillmann, and Wolfgang Rave. Dirty RF: A new paradigm. *International Journal of Wireless Information Networks*, 14(2):133–148, 2007. 23

[Fre95] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995. 9

[FS97] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997. 9

[FZH62] R Frank, S Zadoff, and R Heimiller. Phase shift pulse codes with good periodic correlation properties. *IRE Transactions on Information Theory*, 8(6):381–382, 1962. 56

[GBB+03] Mickaël Germain, Goze B Bénié, J-M Boucher, Samuel Foucher, Ko Fung, and Kalifa Goïta. Contribution of the fractal dimension to multiscale adaptive filtering of SAR imagery. *IEEE transactions on geoscience and remote sensing*, 41(8):1765–1772, 2003. 1

[GLSGFV10] Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010. 41, 75

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985. 46, 79, 100, 122, 142

[HJ99] Lynette Hunt and Murray Jorgensen. Theory & methods: Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171, 1999. 60, 61

[HMDH18] Sibylle Hess, Katharina Morik, Wouter Duivesteijn, and Philipp-Jan Honysz. The SpectACl of nonconvex clustering: a spectral approach to density-based clustering. 10 2018. 13

[Ho98] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, Aug 1998. 9

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. 10

[HTS+01] Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing missing data for gene expression arrays. 1, 12 2001. 41

[Hua98] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, Sep 1998. 10

[HW79] John A Hartigan and Manchek A Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. 2, 10, 26, 98, 120, 140

[HZWT09] Ai-Ling He, De-Guo Zeng, Jun Wang, and Bin Tang. Multi-parameter signal sorting algorithm based on dynamic distance clustering. *Journal of Electronic Science and Technology*, 7(3):249–253, 2009. 1, 26

[J.88] Engel J. Polytomous logistic regression. *Statistica Neerlandica*, 42(4):233–252, 1988. 7

[Jac01] P. Jaccard. Comparative study of the floral distribution in a portion of the alps and jura. *The Company Vaudoise Bulletin of Natural Sciences*, 37(5):547–579, 1901. 75

[Jai10] Anil K Jain. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8):651–666, 2010. 2, 26

[JJ94] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994. 15, 26

[KO04] A Kawalec and R Owczarek. Radar emitter recognition using intrapulse data. In *Microwaves, Radar and Wireless Communications, 2004. MIKON-2004. 15th International Conference on*, volume 2, pages 435–438. IEEE, 2004. 1

[KP16] Mahmoud Keshavarzi and Amir Mansour Pezeshk. A simple geometrical approach for deinterleaving radar pulse trains. In *Computer Modelling and Simulation (UKSim), 2016 UKSim-AMSS 18th International Conference on*, pages 172–177. IEEE, 2016. 2

[KR87] Leonard Kaufmann and Peter Rousseeuw. Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 01 1987. 11

[KR09] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. 46, 79, 100, 122, 142

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 10

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 10

[Lic13] M. Lichman. UCI machine learning repository, 2013. ix, 7

[LJLC16] H. Li, W. D. Jin, H. D. Liu, and T. W. Chen. Work mode identification of airborne phased array radar based on the combination of multi-level modeling and deep learning. In *2016 35th Chinese Control Conference (CCC)*, pages 7005–7010, July 2016. 1, 26

[LK96] C. J. Lawrence and W. J. Krzanowski. Mixture separation for mixed-mode data. *Statistics and Computing*, 6(1):85–92, Mar 1996. 60

[LM04] Nadav Levanon and Eli Mozeson. *Radar signals*. John Wiley & Sons, 2004. 55, 56

[Mar89] H. K. Mardia. New techniques for the deinterleaving of repetitive sequences. *IEE Proceedings F - Radar and Signal Processing*, 136(4):149–154, Aug 1989. 2

[MBV17] Matthieu Marbac, Christophe Biernacki, and Vincent Vandewalle. Model-based clustering of Gaussian copulas for mixed data. *Communications in Statistics - Theory and Methods*, 46(23):11635–11656, 2017. 61

[Min01] Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 147

[MK94] J. B. Moore and V. Krishnamurthy. Deinterleaving pulse trains using discrete-time stochastic dynamic-linear models. *IEEE Transactions on Signal Processing*, 42(11):3092–3103, Nov 1994. 2

[Mor12] Isabella Morlini. A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Advances in Data Analysis and Classification*, 6(1):5–28, 2012. 61

[MP43] WS. McClulloch and W. Pitts. A logical calculus of the ideas immanent in neurons activity. *Bulletin of mathematical biophysics*, 5:115–133, 1943. 10

[MP69] Marvin Minsky and Seymour Papert. Perceptrons. 1969. 10

[MP92] D. J. Milojevic and B. M. Popovic. Improved algorithm for the deinterleaving of radar pulses. *IEE Proceedings F - Radar and Signal Processing*, 139(1):98–104, Feb 1992. 2

[MP04] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004. 33

[NW14] T. M. Nguyen and Q. M. J. Wu. Bounded asymmetrical Student's-t mixture model. *IEEE Transactions on Cybernetics*, 44(6):857–869, June 2014. 29, 30

[OS01] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001. 18

[PJR13] Nedyalko Petrov, Ivan Jordanov, and Jon Roe. Radar emitter signals recognition and classification with feedforward networks. *Procedia Computer Science*, 22:1192–1200, 2013. 1, 26

[PM00] David Peel and Geoffrey J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000. 30

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 14

[QR78] Richard E. Quandt and James B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American statistical Association*, 73(364):730–738, 1978. 26

[RHW85] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985. 10

[RHW88] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988. 10

[Ric05] Mark A Richards. *Fundamentals of radar signal processing*. McGraw-Hill Education, 2005. 1

[RM05] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005. 5, 6

[Rog85] JAV Rogers. ESM processor system for high pulse density radar environments. In *IEE Proceedings F (Communications, Radar and Signal Processing)*, volume 132, pages 621–625. IET, 1985. 1

[Rok10] Lior Rokach. *Pattern classification using ensemble methods*, volume 75. World Scientific, 2010. 8, 9

[Ros58] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. 10, 26

[RR17] Monia Ranalli and Roberto Rocci. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*, 110:87 – 102, 2017. 61

[S+78] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978. 19

[SB05] Markus Svensén and Christopher M. Bishop. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005. 29, 30, 33

[Sch86] D. Curtis Schleher. Introduction to Electronic Warfare. Technical report, Eaton Corp., AIL Div., Deer Park, NY, 1986. 1

[Sch90] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990. 9

[Sch97] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997. 41, 75

[SEG+16] Hugo Seute, Cyrille Enderli, Jean-Francois Grandin, Ali Khenchaf, and Jean-Christophe Cexus. Experimental analysis of time deviation on a passive localization system. In *Sensor Signal Processing for Defence (SSPD), 2016*, pages 1–5. IEEE, 2016. 22

[SEKX98] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data mining and knowledge discovery*, 2(2):169–194, 1998. 2, 26

[Shi14] Ya Shi. Kernel canonical correlation analysis for specific radar emitter identification. *Electronics Letters*, 50:1318–1320(2), August 2014. 1

[SL02] Ching-Sung Shieh and Chin-Teng Lin. A vector neural network for emitter identification. *IEEE Transactions on Antennas and Propagation*, 50(8):1120–1127, 2002. 1, 26

[Sun18] Jun Sun. Radar emitter classification based on unidimensional convolutional neural network. *IET Radar, Sonar & Navigation*, April 2018. 1, 26

[SZKL17] J. Sun, A. Zhou, S. Keates, and S. Liao. Simultaneous Bayesian clustering and feature selection through Student's t mixtures model. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99):1–13, 2017. 29, 30

[TCS+01] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. 2, 26, 41

[TK86] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American statistical association*, 81(393):82–86, 1986. 40, 73

[Tur63] R Turyn. On Barker codes of even length. *Proceedings of the IEEE*, 51(9):1256–1256, 1963. 56

[VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. 12, 98, 99, 120, 140, 141

[VM02] Jeroen K Vermunt and Jay Magidson. Latent class cluster analysis. *Applied latent class analysis*, 11:89–106, 2002. 60

[VP98] MP Veyssieres and Richard E Plant. Identification of vegetation state and transition domains in California's hardwood rangelands. *University of California*, 101, 1998. 5

[WB99] Alan Willse and Robert J. Boik. Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, 9(2):111–121, Apr 1999. 60

[Wil82] Richard G Wiley. Electronic Intelligence: the analysis of radar signals. *Dedham, MA, Artech House, Inc., 1982. 250 p*, 1982. 1

[WMR+96] Steve Waterhouse, David MacKay, Tony Robinson, et al. Bayesian methods for mixtures of experts. *Advances in neural information processing systems*, pages 351–357, 1996. 2, 18

[WS00] Martin J Wainwright and Eero P Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In *Advances in neural information processing systems*, pages 855–861, 2000. 29

[YWY+13] Zhutian Yang, Zhilu Wu, Zhendong Yin, Taifan Quan, and Hongjian Sun. Hybrid radar emitter recognition based on rough k-means classifier and relevance vector machine. *Sensors*, 13(1):848–864, 2013. 1, 26

[Z+63] Solomon Zadoff et al. Phase coded signal receiver, July 2 1963. US Patent 3,096,482. 56

[ZWCZ16] Dongqing Zhou, Xing Wang, Siyi Cheng, and Xi Zhang. An online multisensor data fusion framework for radar emitter classification. *International Journal of Aerospace Engineering*, 2016, 2016. 1, 26

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Gestion des incertitudes en identification des modes radar

Mots clés : émetteurs radar ,classification,partitionnement,valeurs aberrantes,données manquantes,modèles de mélange

Résumé : En Guerre Electronique, l'identification des signaux radar est un atout majeur de la prise de décisions tactiques liées au théâtre d'opérations militaires. En fournissant des informations sur la présence de menaces, la classification et le partitionnement des signaux radar ont alors un rôle crucial assurant un choix adapté des contre-mesures dédiées à ces menaces et permettant la détection de signaux radar inconnus pour la mise à jour des bases de données. Les systèmes de Mesures de Soutien Electronique enregistrent la plupart du temps des mélanges de signaux radar provenant de différents émetteurs présents dans l'environnement électromagnétique. Le signal radar, décrit par un motif de modulations impulsionnelles, est alors souvent partiellement observé du fait de mesures manquantes et aberrantes. Le processus d'identification se fonde sur l'analyse statistique des paramètres mesurables du signal radar qui le caractérisent tant quantitativement que qualitativement. De nombreuses approches mêlant des techniques de fusion de données et d'apprentissage statistique ont été développées. Cependant, ces algorithmes ne peuvent pas gérer les données manquantes et des méthodes de substitution de données sont requises afin d'utiliser ces derniers. L'objectif principal de cette thèse est alors de définir un modèle de classification et partitionnement intégrant la gestion des valeurs aberrantes et manquantes présentes dans tout type de données. Une approche fondée sur les modèles de mélange de lois de probabilité est proposée dans cette thèse. Les modèles de mélange fournissent un formalisme mathématique flexible favorisant l'introduction de variables latentes permettant la gestion des données aberrantes et la modélisation des données manquantes dans les problèmes de classification et de partitionnement. L'apprentissage du modèle ainsi que la classification et le partitionnement sont réalisés dans un cadre d'inférence bayésienne où une méthode d'approximation variationnelle est introduite afin d'estimer la loi jointe a posteriori des variables latentes et des paramètres. Des expériences sur diverses données montrent que la méthode proposée fournit de meilleurs résultats que les algorithmes standards.

Title : Uncertainty in radar emitter classification and clustering

Keywords : radar emitter,classification,clustering,outliers,missing data,mixture models

Abstract : In Electronic Warfare, radar signals identification is a supreme asset for decision making in military tactical situations. By providing information about the presence of threats, classification and clustering of radar signals have a significant role ensuring that countermeasures against enemies are well-chosen and enabling detection of unknown radar signals to update databases. Most of the time, Electronic Support Measures systems receive mixtures of signals from different radar emitters in the electromagnetic environment. Hence a radar signal, described by a pulse-to-pulse modulation pattern, is often partially observed due to missing measurements and measurement errors. The identification process relies on statistical analysis of basic measurable parameters of a radar signal which constitute both quantitative and qualitative data. Many general and practical approaches based on data fusion and machine learning have been developed and traditionally proceed to feature extraction, dimensionality reduction and classification or clustering. However, these algorithms can not handle missing data and imputation methods are required to generate data to use them. Hence, the main objective of this work is to define a classification/clustering framework that handles both outliers and missing values for any types of data. Here, an approach based on mixture models is developed since mixture models provide a mathematically based, flexible and meaningful framework for the wide variety of classification and clustering requirements. The proposed approach focuses on the introduction of latent variables that give us the possibility to handle sensitivity of the model to outliers and to allow a less restrictive modelling of missing data. A Bayesian treatment is adopted for model learning, supervised classification and clustering and inference is processed through a variational Bayesian approximation since the joint posterior distribution of latent variables and parameters is untractable. Some numerical experiments on synthetic and real data show that the proposed method provides more accurate results than standard algorithms.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France