MDPI

Article

A COMPLETE CLASSIFICATION AND CLUSTERING MODEL TO ACCOUNT FOR CONTINUOUS AND CATEGORICAL DATA IN PRESENCE OF MISSING VALUES AND OUTLIERS

Guillaume Revillon [1,†,‡] 0000-0001-5963-0531 and Ali Mohammad-Djafari [1,*]

[1] L2S, CentraleSupélec-Univ Paris Saclay, 91192 Gif-sur-Yvette, France; guillaume.revillon@l2s.centralesupelec.fr, Ali.Mohammad-Djafari@l2s.centralesupelec.fr

* Correspondence: guillaume.revillon@l2s.centralesupelec.fr; Tel.: +33 6 19 97 11 17

‡ These authors contributed equally to this work.

Version August 26, 2019 submitted to Journal Not Specified

Abstract: Classification and clustering problems are closely connected with pattern recognition where many general algorithms have been developed and used in various fields. Depending on the complexity of patterns in data, classification and clustering procedures should take into consideration both continuous and categorical data which can be partially missing and erroneous due to mismeasurements and human errors. However, most algorithms cannot 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 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. More precisely, a scale mixture of Normal distributions is updated to handle outliers and missing data issues for any types of data. Then a variational Bayesian inference 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. Eventually, experiments are carried out to exhibit the effectiveness of the proposed model through an application in Electronic Warfare.

Keywords: Classification, Clustering, Mixture Models, Bayesian Framework, Outliers, Missing Data

1. Introduction

Classification and clustering problems are closely connected with pattern recognition [1] where many general algorithms [2–4] have been developed and used in various fields [5,6]. Depending on the complexity of patterns in data, classification and clustering procedures should take into consideration both continuous and categorical data which can be partially missing and erroneous due to mismeasurements and human errors. However, most algorithms cannot handle missing data and imputation methods [7] 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 [8]. Two families of models emerge from finite mixture models fitting mixed-type data :

- The location mixture model [9] that assumes that continuous variables follow a multivariate Gaussian distribution conditionally on both component and categorical variables.

- The underlying variables mixture model [10] that assumes that each discrete variable arises from a latent continuous variable and that all continuous variables follow a Gaussian mixture model.

In this work, the location mixture model approach is retained since it better models relations between continuous and categorical features when data patterns are mostly designed by first choosing patterns of categorical features to achieve a specific goal and then choosing continuous features that meet constraints related to the chosen patterns and the problem environment. Indeed regarding clustering approach, each cluster groups observations that share same combinations of categorical features where continuous features belong to a peculiar subset. Hence, the location mixture model naturally responds to that dependence structure by assuming that continuous variables are normally distributed conditionally to categorical variables. More precisely, a scale mixture of conditional Gaussian distributions [11] is updated to handle outliers and missing data issues for any types of data. Then a variational Bayesian inference [12] 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. An application of the resulting model in Electronic Warfare [13] is proposed to perform Source Emission Identification which is a supreme asset for decision making in military tactical situations. By providing information about the presence of threats, classification and clustering of radar emitters have a significant role ensuring that countermeasures against enemies are well-chosen and enabling detection of unknown radar signals to update databases. As a pulse-to-pulse modulation pattern [14], 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. However, a radar signal is often partially observed due to the presence of many radar emitters in the electromagnetic environment causing mismeasurements and measurement errors. Therefore the proposed model is suitable for radar emitter classification and clustering. The outline of the paper is as follows. Assumptions on mixed-type data are presented in Section 2. Then, the proposed model and inference procedure are introduced in Section 3. Finally, evaluation of the model is proposed through different experiments on radar emitter datasets in Section 4.

2. Mixed-type data

In this section, 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.

2.1. Assumptions on mixed-type data

Data $\boldsymbol{x}$ consist of J observations $(\boldsymbol{x}_j)_{j=1}^J$ gathering continuous features $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j=1}^J$ and categorical features $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j=1}^J$. 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 variables,
- $\boldsymbol{x}_{cj} = \left(x_{cj}^0, \ldots, x_{cj}^{q-1}\right) \in \mathcal{C}_q$ 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 $\forall i \in \{0, \ldots, q-1\}$.

2.2. Distribution of mixed-type data

Considering that the retained approach focuses on conditioning continuous data $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j=1}^J$ according to categorical data $\boldsymbol{x}_c = (\boldsymbol{x}_{cj})_{j=1}^J$, the following joint distribution is introduced

$$\forall j \in \{1, \ldots, J\},\ p(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj}) = \prod_{\boldsymbol{c} \in \mathcal{C}_q} \left(\pi_{\boldsymbol{c}} \mathcal{N}\left(\boldsymbol{x}_{qj} | \boldsymbol{\mu}_{\boldsymbol{c}}, \boldsymbol{\Sigma}\right)\right)^{\delta_{\boldsymbol{x}_{cj}}^{\boldsymbol{c}}} \tag{1}$$

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 a multivariate categorical distribution $\mathcal{MC}(\boldsymbol{x}_{cj}|\boldsymbol{\pi})$ parametrized by weights $\boldsymbol{\pi} = (\pi_{\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}$ and defined by

$$\mathcal{MC}(\boldsymbol{x}_{cj}|\boldsymbol{\pi}) = \prod_{\boldsymbol{c}\in\boldsymbol{C}_q} \pi_{\boldsymbol{c}}^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}} \tag{2}$$

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

$$\sum_{\boldsymbol{c}\in\mathcal{C}_q} \pi_{\boldsymbol{c}} = 1 \text{ and } \delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}} = \begin{cases} 1 \text{ if } x^0_{cj} = c^0, \ldots, x^{q-1}_{cj} = c^{q-1} \\ 0 \text{ otherwise} \end{cases} .$$

This multivariate categorical distribution is proposed to tackle issues related to missing data by modelling a dependence structure for $\boldsymbol{x}_{cj}$ that enables inference on missing categorical features.

2.3. Outlier handling

Outliers are only considered for continuous data $\boldsymbol{x}_q = (\boldsymbol{x}_{qj})_{j=1}^J$ since only reliable categorical variables are assumed to be filled in databases and unreliable ones are processed as missing data. Then, continuous outliers are handled by introducing scale latent variables $\boldsymbol{u} = (u_j)_{j=1}^J$ conditionally to categorical data $\boldsymbol{x}_c$ due to the dependence structure established in (1) such that

$$\forall j \in \{1, \ldots, J\},\ \boldsymbol{x}_{qj}|u_j, \boldsymbol{x}_{cj} \sim \prod_{\boldsymbol{c}\in\mathcal{C}_q} \mathcal{N}\left(\boldsymbol{x}_{qj}|\boldsymbol{\mu}_{\boldsymbol{c}}, u_j^{-1}\boldsymbol{\Sigma}\right)^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}} \text{ and } u_j|\boldsymbol{x}_{cj} \sim \prod_{\boldsymbol{c}\in\mathcal{C}_q} \mathcal{G}\left(u_j|\alpha_{\boldsymbol{c}}, \beta_{\boldsymbol{c}}\right)^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}} ,$$

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

2.4. Missing data handling

Both continuous and categorical data $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j=1}^J$ can be partially observed. Hence $(\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j=1}^J$ are decomposed into observed features $(\boldsymbol{x}^{\text{obs}}_{qj}, \boldsymbol{x}^{\text{obs}}_{cj})_{j=1}^J$ and missing features $(\boldsymbol{x}^{\text{miss}}_{qj}, \boldsymbol{x}^{\text{miss}}_{cj})_{j=1}^J$ such that

$$\forall j \in \{1, \ldots, J\}, \quad \begin{aligned} \boldsymbol{x}_{qj} &= \begin{pmatrix} \boldsymbol{x}^{\text{miss}}_{qj} \\ \boldsymbol{x}^{\text{obs}}_{qj} \end{pmatrix} \text{ with } (\boldsymbol{x}^{\text{miss}}_{qj}, \boldsymbol{x}^{\text{obs}}_{qj}) \in \mathbb{R}^{d^{\text{miss}}_j} \times \mathbb{R}^{d^{\text{obs}}_j} \text{ and } d^{\text{miss}}_j + d^{\text{obs}}_j = d \,, \\ \boldsymbol{x}_{cj} &= \begin{pmatrix} \boldsymbol{x}^{\text{miss}}_{cj} \\ \boldsymbol{x}^{\text{obs}}_{cj} \end{pmatrix} \text{with } (\boldsymbol{x}^{\text{miss}}_{cj}, \boldsymbol{x}^{\text{obs}}_{cj}) \in \mathcal{C}_{q^{\text{miss}}_j} \times \mathcal{C}_{q^{\text{obs}}_j} \text{ and } q^{\text{miss}}_j + q^{\text{obs}}_j = q \,. \end{aligned}$$

where $(\mathbb{R}^{d^{\text{miss}}_j}, \mathcal{C}_{q^{\text{miss}}_j})$ and $(\mathbb{R}^{d^{\text{obs}}_j}, \mathcal{C}_{q^{\text{obs}}_j})$, are disjoint subsets of $(\mathbb{R}^d, \mathcal{C}_q)$ embedding missing features $(\boldsymbol{x}^{\text{miss}}_{qj}, \boldsymbol{x}^{\text{miss}}_{cj})$ and observed features $(\boldsymbol{x}^{\text{obs}}_{qj}, \boldsymbol{x}^{\text{obs}}_{cj})$. Missing continuous data $\boldsymbol{x}^{\text{miss}}_q = (\boldsymbol{x}^{\text{miss}}_{qj})_{j=1}^J$ are handled by taking advantage of properties of the multivariate normal distribution to obtain a distribution for missing values. Due to the dependence structure established in (1), missing continuous data $\boldsymbol{x}^{\text{miss}}_q = (\boldsymbol{x}^{\text{miss}}_{qj})_{j=1}^J$ are distributed conditionally to observed continuous data $\boldsymbol{x}^{\text{obs}}_q = (\boldsymbol{x}^{\text{obs}}_{qj})_{j=1}^J$ and categorical data $\boldsymbol{x}_c$ as follows

$$\forall j \in \{1, \ldots, J\},\ \boldsymbol{x}^{\text{miss}}_{qj}|\boldsymbol{x}^{\text{obs}}_{qj}, \boldsymbol{x}_{cj} \sim \prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{x}^{\text{miss}}_{qj}|\boldsymbol{\mu}^{\boldsymbol{x}^{\text{miss}}_q}_{j\boldsymbol{c}}, \boldsymbol{\Sigma}^{\boldsymbol{x}^{\text{miss}}_q}\right)^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}},\ \boldsymbol{x}^{\text{obs}}_{qj}|\boldsymbol{x}^{\text{obs}}_{qj}, \boldsymbol{x}_{cj} \sim \prod_{\boldsymbol{c}\in\mathcal{C}} \mathcal{N}\left(\boldsymbol{x}^{\text{obs}}_{qj}|\boldsymbol{\mu}^{\boldsymbol{x}^{\text{obs}}_q}_{j\boldsymbol{c}}, \boldsymbol{\Sigma}^{\boldsymbol{x}^{\text{obs}}_q}\right)^{\delta^{\boldsymbol{c}}_{\boldsymbol{x}_{cj}}},$$

where $\forall j \in \{1, \dots, J\},\ \forall \boldsymbol{c} \in \mathcal{C}_q$:

$$\boldsymbol{\mu}_{j\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{miss}}} = \boldsymbol{\mu}_{\boldsymbol{c}}^{\text{miss}} + \boldsymbol{\Sigma}^{\text{cov}}\boldsymbol{\Sigma}^{\text{obs}^{-1}}\left(\boldsymbol{x}_{qj}^{\text{obs}} - \boldsymbol{\mu}_{\boldsymbol{c}}^{\text{obs}}\right),\ \boldsymbol{\mu}_{j\boldsymbol{c}}^{\boldsymbol{x}_q^{\text{obs}}} = \boldsymbol{\mu}_{\boldsymbol{c}}^{\text{obs}},$$

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

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 missing features $\boldsymbol{x}_{cj}^{\text{miss}}$ follow a multivariate categorical distribution conditionally to observed features $\boldsymbol{x}_{cj}^{\text{obs}}$ given by

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

where $\pi_{\boldsymbol{c}^{\text{miss}},\boldsymbol{c}^{\text{obs}}}$ is the joint probability $\pi_{\boldsymbol{c}}$ defined in (2) for $\boldsymbol{c} = (\boldsymbol{c}^{\text{miss}}, \boldsymbol{c}^{\text{obs}}) \in \boldsymbol{C}_{q_j^{\text{miss}}} \times \boldsymbol{C}_{q_j^{\text{obs}}}$.

3. Model and inference

In this section, the proposed model is briefly presented as a hierarchical latent variable model handling missing values and outliers. Then, the inference procedure is developed through a variational Bayesian approximation. At last, classification and clustering algorithms are introduced by using the proposed model.

3.1. Model

According to a dataset $\boldsymbol{x}^{\text{obs}}$ of i.i.d observations, independent latent variables $\boldsymbol{h} = (\boldsymbol{x}^{\text{miss}}, \boldsymbol{u}, \boldsymbol{z})$, parameters $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the K clusters and assumptions on mixed data defined in subsection 2.1, the complete likelihood of the proposed mixture model can be expressed as

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

where

- $\boldsymbol{x}^{\text{obs}} = (\boldsymbol{x}_{qj}^{\text{obs}}, \boldsymbol{x}_{cj}^{\text{obs}})_{j=1}^{J}$ are the observed features,
- $\boldsymbol{x}^{\text{miss}} = (\boldsymbol{x}_{qj}^{\text{miss}}, \boldsymbol{x}_{cj}^{\text{miss}})_{j=1}^{J}$ are the latent variables modelling the missing features,
- $\boldsymbol{z} = (z_j)_{j=1}^{J}$ the independent labels for continuous and categorical observations $\boldsymbol{x} = (\boldsymbol{x}_{qj}, \boldsymbol{x}_{cj})_{j=1}^{J}$
- $\boldsymbol{u} = (u_j)_{j=1}^{J}$ the scale latent variables handling outliers for quantitative data $\boldsymbol{x}_q$ and distributed according to a Gamma distribution with shape and rate parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_{k\boldsymbol{c}}, \beta_{k\boldsymbol{c}})_{(k,\boldsymbol{c})\in\{1,\dots,K\}\times\mathcal{C}_q}$,
- $\boldsymbol{a} = (a_k)_{k=1}^{K}$ are the weights related to component distributions,
- $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = ((\boldsymbol{\mu}_{k\boldsymbol{c}})_{\boldsymbol{c}\in\mathcal{C}_q}, \boldsymbol{\Sigma}_k)_{k=1}^{K}$ the mean and the variance parameters of quantitative data $\boldsymbol{x}_q$ for each cluster,
- $\boldsymbol{\pi} = (\boldsymbol{\pi}_k)_{k=1}^{K}$ the weights of the multivariate Categorical distribution of categorical data $\boldsymbol{x}_c$ for each cluster.

Eventually, the Bayesian framework imposes to specify a prior distribution $p(\boldsymbol{\Theta}|K)$ for $\boldsymbol{\Theta}$ which is chosen as

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

where $\mathcal{D}(\cdot|\cdot)$ and $\mathcal{IW}(\cdot|\cdot)$ denote the Dirichlet and Inverse-Wishart distributions and $p(\cdot,\cdot|p,q,s,r)$ is a particular distribution designed to avoid a non-closed-form posterior distribution for $(\boldsymbol{\alpha},\boldsymbol{\beta})$ such that $\forall(\alpha,\beta)\in\mathbb{R}^{*+}\times\mathbb{R}^{*+},\ p(\alpha,\beta|p,q,s,r)\propto p^{\alpha-1}e^{-q\beta}\beta^{s\alpha}\Gamma(\alpha)^{-r}$.

3.2. Variational Bayesian Inference

The intractable posterior distribution $P = p(\boldsymbol{h},\boldsymbol{\Theta}|\boldsymbol{x}^{\text{obs}},K)$ is approximated by a tractable one $Q = q(\boldsymbol{h},\boldsymbol{\Theta}|K)$ whose parameters are chosen via a variational principle to minimize the Kullback-Leibler (KL) divergence

$$KL\left[Q||P\right] = \int q(\boldsymbol{h},\boldsymbol{\Theta}|K)\log\left(\frac{q(\boldsymbol{h},\boldsymbol{\Theta}|K)}{p(\boldsymbol{h},\boldsymbol{\Theta}|\boldsymbol{x}^{\text{obs}},K)}\right)\partial\boldsymbol{h}\partial\boldsymbol{\Theta} = \log p(\boldsymbol{x}^{\text{obs}}|K) - \mathcal{L}(q|K)$$

with $\mathcal{L}(q|K)$ a lower bound for the log evidence $\log p(\boldsymbol{x}^{\text{obs}}|K)$ given by

$$\mathcal{L}(q|K) = \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h},\boldsymbol{\Theta}|K)\right] - \mathbb{E}_{\boldsymbol{h},\boldsymbol{\Theta}}\left[\log q(\boldsymbol{h},\boldsymbol{\Theta}|K)\right]\ , \tag{3}$$

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

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

Thereafter, the algorithm iteratively updates the variational posteriors by increasing the bound $\mathcal{L}(q|K)$. Even if latent variables $\boldsymbol{h}$ and parameters $\boldsymbol{\Theta}$ are assumed to be independent a posteriori, their conditional structures are preserved as follows

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

Eventually, the following conjugate variational posterior distributions are obtained according to the previous assumptions

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

$$q(\boldsymbol{\Theta}|K) = \mathcal{D}(\boldsymbol{a}|\tilde{\boldsymbol{\kappa}})\prod_{k=1}^{K}\mathcal{D}(\boldsymbol{\pi}|\tilde{\boldsymbol{\pi}}_k)\prod_{\boldsymbol{c}\in\mathcal{C}_q} p(\alpha_{k\boldsymbol{c}},\beta_{k\boldsymbol{c}}|\tilde{p}_k,\tilde{q}_k,\tilde{s}_k,\tilde{r}_k)\mathcal{N}\left(\boldsymbol{\mu}_{k\boldsymbol{c}}|\tilde{\boldsymbol{\mu}}_{k\boldsymbol{c}},\tilde{\eta}_{k\boldsymbol{c}}^{-1}\boldsymbol{\Sigma}_k\right)\mathcal{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k,\tilde{\boldsymbol{\Sigma}}_k)\ .$$

Their respective parameters are estimated during the VBE and VBM-steps by developing expectations $\mathbb{E}_{\boldsymbol{\Theta}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h}|\boldsymbol{\Theta},K)\right]$ and $\mathbb{E}_{\boldsymbol{h}}\left[\log p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h},\boldsymbol{\Theta}|K)\right]$.

3.3. 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

$$k^* = \arg\max_{k=1}^{K} q(z^* = k|K)$$

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

$$k^* = \arg\max_{k=1}^{K} q(z = k|K) \ .$$

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

$$\tilde{K} = \arg\max_{K} \log p(\boldsymbol{x}|K) = \arg\max_{K} \log \int p(\boldsymbol{x}, \boldsymbol{\Theta}|K) d\boldsymbol{\Theta} \ . \tag{4}$$

Unfortunately, $\log p(\boldsymbol{x}|K)$ is intractable and the lower bound in (3) is preferred to penalized likelihood criteria [8,15,16] 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 (4) is chosen as the maximizer of the lower bound $\mathcal{L}(q|K)$:

$$\tilde{K} = \arg\max_{K} \ \mathcal{L}(q|K) \ . \tag{5}$$

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

4. Application

In this section, the proposed method is performed on a radar emitter dataset. For comparison, a standard neural network (NN), the k-nearest neighbours (KNN) algorithm, Random Forests (RdF) the k-means algorithm are also evaluated. Two experiments are carried out to evaluate classification and clustering performance with respect to a range of percentages of missing values.

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 (pulse features) and a quartet of categorical features (pulse modulations) listed among 42 combinations of the categorical features. 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. 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. Therefore, imputation methods [17] are used to handle missing data for comparison algorithms. As for continuous missing data, they are handled through the Mean and k-nearest neighbours imputation methods whereas missing categorical data are handled through

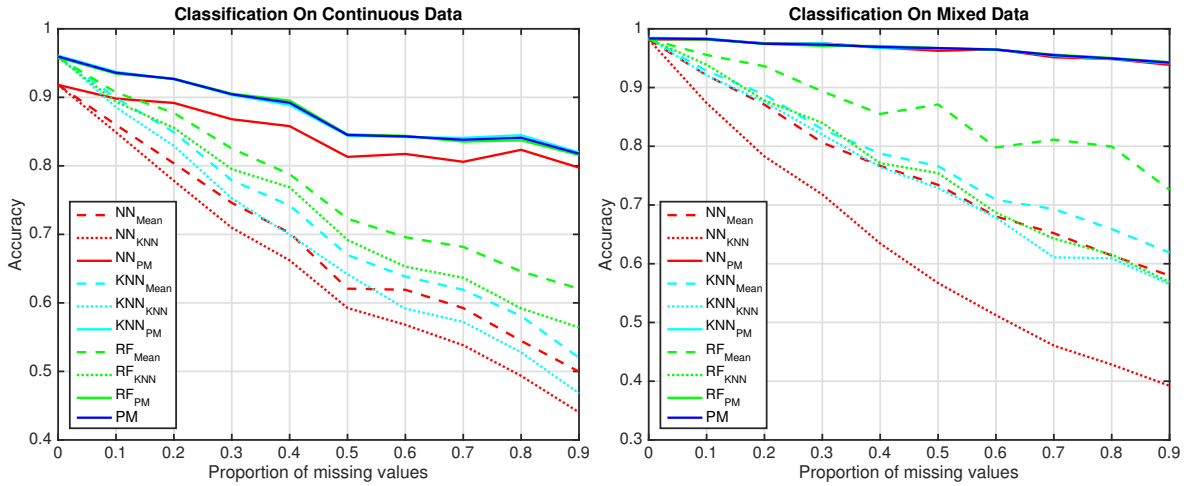

Figure 1. Classification performance are presented for the proposed model (PM) in blue, the NN in red, the RnF in green and the KNN in cyan. For each figure, solid lines represent accuracies with a posteriori reconstructed missing data, doted dashed lines stand for accuracies with mean/mode imputation whereas dashed lines show accuracies with KNN imputation for the comparison algorithms.

the k-nearest neighbours and mode imputation methods. These imputation methods are compared with the proposed approach where missing continuous data are reconstructed through the variational posterior marginal mean of missing continuous data given by $\forall j \in \{1, \ldots, J\}$,

$$\tilde{\boldsymbol{x}}_{qj}^{\text{miss}} = \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{qj}}^{\text{miss}}}\left[\int q(\boldsymbol{x}_{qj}^{\text{miss}}, u_j, \boldsymbol{x}_{cj} z_j) \partial u_j \partial \boldsymbol{x}_{cj} \partial z_j\right] = \sum_{k=1}^{K} \tilde{r}_{jk} \sum_{\boldsymbol{c}^{\text{obs}} \in \mathcal{C}_{q_j^{\text{obs}}}} \delta_{\boldsymbol{x}_{cj}^{\text{obs}}}^{\boldsymbol{c}^{\text{obs}}} \sum_{\boldsymbol{c}^{\text{miss}} \in \mathcal{C}_{q_j^{\text{miss}}}} \tilde{r}_{jk\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_c^{\text{miss}}} \tilde{\boldsymbol{\mu}}_{jk\boldsymbol{c}^{\text{obs}}\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_q^{\text{miss}}} \tag{6}$$

and missing categorical data are reconstructed through the variational posterior marginal mode of missing categorical data given by $\forall j \in \{1, \ldots, J\}$,

$$\tilde{\boldsymbol{x}}_{cj}^{\text{miss}} = \arg \max_{\boldsymbol{c}^{\text{miss}} \in \mathcal{C}_{q_j^{\text{miss}}}} \int q(\boldsymbol{x}_{cj}^{\text{miss}}, z_j) dz_j = \arg \max_{\boldsymbol{c}^{\text{miss}} \in \mathcal{C}_{q_j^{\text{miss}}}} \sum_{k=1}^{K} \tilde{r}_{jk} \tilde{r}_{jk\boldsymbol{c}^{\text{miss}}}^{\boldsymbol{x}_c^{\text{miss}}} . \tag{7}$$

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. 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 (6)-(7). For the classification experiment, results are shown in Figure 1. Without missing data, both algorithms cannot perfectly classify the 55 radar emitters for the 2 datasets. Indeed, both algorithms reach accuracies of 90% for the continuous 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 and some intervals of continuous features. 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% and 95% for 90% of deleted continuous and mixed values whereas accuracies of comparison algorithms are lower than 65% and 75% with missing data imputation from standard methods. 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 1 when comparison algorithms are applied on data reconstructed by

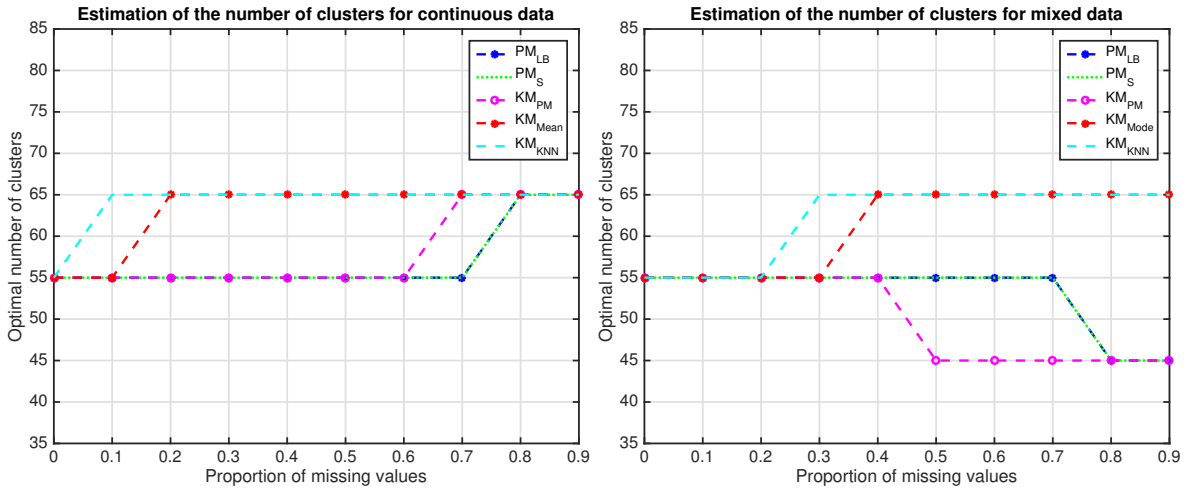

Figure 2. Estimation of the number of clusters using the lower bound (LB) and the silhouette score (S) for the proposed model and only the silhouette score (S) for the k-means algorithm.

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 data even for 90% of deleted 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.

4.3. Clustering experiment

The clustering experiment tests the ability of each algorithm to find the true number of clusters $\tilde{K}$ among $\{35, \ldots, 85\}$. The lower bound (3) and the average Silhouette score [18] are criteria used to select the optimal number of clusters for the proposed model and the k-means algorithm. Results of the clustering experiment are visible on Figure 2 which 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. In presence of missing values, the average Silhouette score mainly selects $K = 65$ when the k-means algorithm is run on the 2 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.

5. Conclusion

In this paper, a mixture model handling both continuous data and categorical data is developed. More precisely, an approach based on the conditional Gaussian mixture model is investigated by establishing conditional relations between continuous and categorical data. Benefiting from a dependence structure designed for mixed-type 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 Bayesian approximation where variational posterior distributions are proposed for continuous and categorical missing data. Experiments point out that the proposed approach can handle mixed-type data even in presence of missing values and can outperform standard algorithms in classification and 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.

References

1. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer-Verlag: Berlin, Heidelberg, 2006.
2. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **1979**, *28*, 100–108.
3. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X.; others. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 1996, Vol. 96, pp. 226–231.
4. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
5. Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. Density-Based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data mining and knowledge discovery* **1998**, *2*, 169–194.
6. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern recognition letters* **2010**, *31*, 651–666.
7. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525.
8. Biernacki, C.; Celeux, G.; Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* **2000**, *22*, 719–725.
9. Lawrence, C.J.; Krzanowski, W.J. Mixture separation for mixed-mode data. *Statistics and Computing* **1996**, *6*, 85–92. doi:10.1007/BF00161577.
10. Everitt, B. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters* **1988**, *6*, 305 – 309. doi:https://doi.org/10.1016/0167-7152(88)90004-1.
11. Andrews, D.F.; Mallows, C.L. Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* **1974**, *36*, 99–102.
12. Waterhouse, S.; MacKay, D.; Robinson, T.; others. Bayesian methods for mixtures of experts. *Advances in neural information processing systems* **1996**, pp. 351–357.
13. Schleher, D.C. Introduction to Electronic Warfare. Technical report, Eaton Corp., AIL Div., Deer Park, NY, 1986.
14. Richards, M.A. *Fundamentals of radar signal processing*; McGraw-Hill Education, 2005.
15. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Springer, 1998; pp. 199–213.
16. Schwarz, G.; others. Estimating the dimension of a model. *The annals of statistics* **1978**, *6*, 461–464.
17. García-Laencina, P.J.; Sancho-Gómez, J.L.; Figueiras-Vidal, A.R. Pattern classification with missing data: a review. *Neural Computing and Applications* **2010**, *19*, 263–282.
18. Kaufman, L.; Rousseeuw, P.J. *Finding groups in data: an introduction to cluster analysis*; Vol. 344, John Wiley & Sons, 2009.