IET Radar, Sonar & Navigation

Research Article

IET Journals
The Institution of Engineering and Technology

Radar emitters classification and clustering with a scale mixture of normal distributions

ISSN 1751-8784
Received on 25th May 2018
Revised 11th August 2018
Accepted on 29th August 2018
E-First on 23rd October 2018
doi: 10.1049/iet-rsn.2018.5202
www.ietdl.org

Guillaume Revillon[1,2] ✉, Ali Mohammad-Djafari[1], Cyrille Enderli[2]

[1]Laboratoire des Signaux Systèmes (L2S), Université Paris-Sud-CNRS-CentraleSupélec-Université Paris-Saclay, 3 rue Joliot Curie, 91192 Gif-sur-Yvette, France

[2]Thales Defense Mission Systems (DMS), 2 avenue Gay Lussac, 78990 Elancourt, France

✉ *E-mail: guillaume.revillon@l2s.centralesupelec.fr*

Abstract: In this study, a scale mixture of normal distributions model is developed for classification and clustering of radar emitters. A radar signal is characterised by a pulse-to-pulse modulation pattern and is often partially observed. The proposed model can classify and cluster different radar emitters even in the presence of outliers and missing values. The classification method, based on a mixture model, 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. The inference is processed through a variation Bayesian approximation. Some numerical experiments on realistic data show that the proposed method provides more accurate results than state-of-the-art classification algorithms.

1 Introduction

In electronic warfare (EW) [1], radar signals identification is a crucial component of electronic support measures (ESM) systems [2]. ESM functions allow 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 [3]. Furthermore, electronic intelligence 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 signal 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 [4].

REC 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, the authors of [5–8] propose various neural classification approaches based on the PDW structure of observed signals whereas Yang *et al*. [9] introduce a hybrid radar emitter recognition method based on rough *k*-means (KM) and relevance vector machine and Chen [10] develop an efficient classification method using a weighted-xgboost model for complex radar signals in large datasets. As regards the clustering problem, He *et al*. [11] develop a dynamic clustering algorithm that uses designed distances and dynamic cluster centres and does not require fixing the number of classes which depends on the input data, Zhou *et al*. [12] 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.

In contrast, SEI aims to extract distinctive features in the process of signal processing to identify even a single copy of an emission source. For a classification purpose, Chen *et al*. [13] develop a fuzzy classifier, which handles unintentional jitter modulations, Dudczyk and Kawalec [14] use a graphical representation of PRI to discover PRI distortions peculiar to each copy and Shi [15] use the kernel canonical correlation analysis to achieve SEI. Kawalec and Owczarek [16] suggest extracting features from intrapulse data to find non-intentional modulations in the receiving signals and the authors of [17, 18] focus on fractal features of radar signals to identify copies. Finally, Dudczyk [4]

proposes an algorithm based on hierarchical agglomerative clustering for large data sets that provide dendrograms used to analyse clustering results and to select the number of clusters.

However, information required for SEI such as intrapulse data or exhaustive PDW is not always available in databases provided by operational entities. Generally, the biggest difficulties for SEI result from the lack of a precise and detailed description of a source emission model in databases [19]. Therefore, this work only focuses on REC to meet operational constraints.

Most of the time, ESM systems receive mixtures of signals from different radar emitters in the electromagnetic environment. Then a radar signal, described by a pulse-to-pulse modulation pattern, is often partially observed. On the one hand, deinterleaving techniques [20] cannot manage to group all the pulses that belong to the same emitters and impute some of them to other emitters. On the other hand, even if the deinterleaving performs well, EW sensors deficiency and low signal-to-noise ratio values in sensors can also cause either missing measurements or measurement errors [21]. When measurements are known to be wrong, such as approximate PW measurements, considering them as missing measurements can be a more reliable approach than using them or discarding them. Furthermore, military databases are filled by human beings and can also be imperfect by gathering outliers and missing data.

Classification and clustering problems are strictly connected with pattern recognition [22] and more general algorithms such as Random Forests [23], density-based spatial clustering of applications with noise algorithm (DBSCAN) [24] and KM

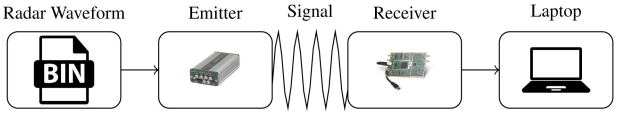

Fig. 1 *Diagram of the acquisition system*

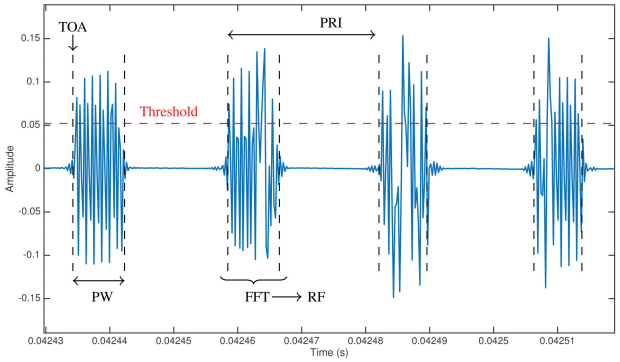

Fig. 2 *Acquired pulses from a radar emitter where the three features (PRI, PW, RF) are shown in the figure*

algorithm [25] are also considered as state-of-the-art since they are used in various fields [26, 27]. However, these practical and general algorithms cannot handle missing data and imputation methods [28] 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. Here, an approach based on mixture models is preferred since mixture models provide a mathematically based, flexible and meaningful framework for a wide variety of classification and clustering requirements [29]. More precisely, a scale mixture of normal distributions [30] 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 [31]. On the other hand, dependencies between features can easily be modelled through a multivariate Gaussian distribution in order to infer on missing values by benefitting from attractive Gaussian properties. Exact inference in that Bayesian approach is unfortunately intractable, therefore a variational Bayesian (VB) inference [32] is used to find the approximate posterior distribution of parameters and to provide a lower bound (LB) on the model log evidence used as a criterion for selecting the number of clusters.

The outline of the paper is as follows. An experimental protocol is developed in Section 2 to acquire unclassified realistic data. Then, the proposed model (PM) is explained in Section 3 and the inference procedure is derived in Section 4. Finally, the evaluation of the model is proposed through different experiments on realistic experimental data in Section 5.

2 Data acquisition

In this section, we will describe the experimental protocol used to acquire unclassified realistic data from different radar emitters. This protocol consists of 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 [33], this setup was chosen because it allows quick development and experimentation tasks on RFs 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 (random variation)/hopping (systematic variation) frequencies and jittered/staggered PRI. The developed system is presented in Fig. 1. To meet hardware constraints, the RF range was mapped to a 4 MHz bandwidth and patterns of Time of Arrival (TOA) and PW were slightly modified but their dynamics was preserved.

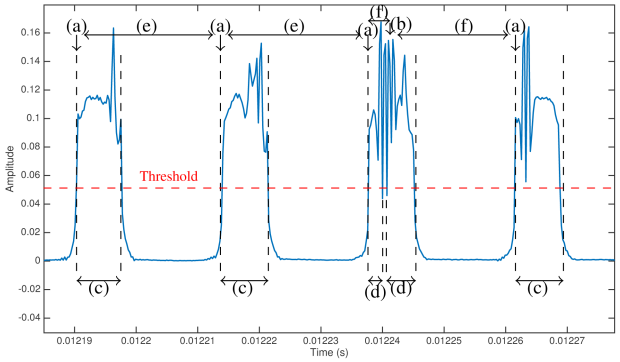

Fig. 3 *Outliers formation during primary parameters measurement on real data*

(a), ***(c)***, ***(e)*** Are respectively exact TOA, PW and PRI, ***(b)***, ***(d)***, ***(f)*** Are outliers for TOA, PW and PRI

Then, a threshold algorithm, provided by Davies and Hollands [34], is used to detect pulses in the recorded signal $s(t)$ of duration T and to extract its PDW. 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 algorithm. Fig. 2 shows the parameters measurement on real data. For a given recorded signal s gathering n_s pulses, the following PDW matrix is obtained

$$\mathrm{PDW} = \begin{pmatrix} \mathrm{RF}_1 & \mathrm{PW}_1 & \mathrm{PRI}_1 \\ \vdots & \vdots & \vdots \\ \mathrm{RF}_m & \mathrm{PW}_m & \mathrm{PRI}_m \\ \vdots & \vdots & \vdots \\ \mathrm{RF}_{n_s} & \mathrm{PW}_{n_s} & \mathrm{PRI}_{n_s} \end{pmatrix}, \tag{1}$$

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

SDR platforms are imperfect [35] and their defects can introduce outliers due to measurement errors. Hardware imperfections are visible in Fig. 3 where the third pulse is cut into two pulses which leads to the formation of PRI and PW outliers. Furthermore, since experiments took place in real outside conditions, other signals and reflections can disturb the acquisition [34].

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 (1) and an observation vector $\boldsymbol{x}_j$ is defined according to (2) such that

$$\boldsymbol{x}_j = (\bar{\mathrm{RF}}_j, \bar{\mathrm{PW}}_j, \bar{\mathrm{PRI}}_j), \tag{2}$$

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 as

$$\bar{\mathrm{RF}}_j = \frac{1}{n_j} \sum_{m=1}^{n_j} \mathrm{RF}_m, \tag{3}$$

$$\bar{\mathrm{PW}}_j = \frac{1}{n_j} \sum_{m=1}^{n_j} \mathrm{PW}_m, \tag{4}$$

$$\bar{\mathrm{PRI}}_j = \frac{1}{n_j} \sum_{m=1}^{n_j} \mathrm{PRI}_m. \tag{5}$$

Once all observation vectors $(\boldsymbol{x}_j)_{j=1}^{J}$ have been constructed, they are normalised to meet constraints of machine learning algorithms. Fig. 4 shows the distribution of 150 normalised observation vectors of a radar emitter, where three outliers are visible.

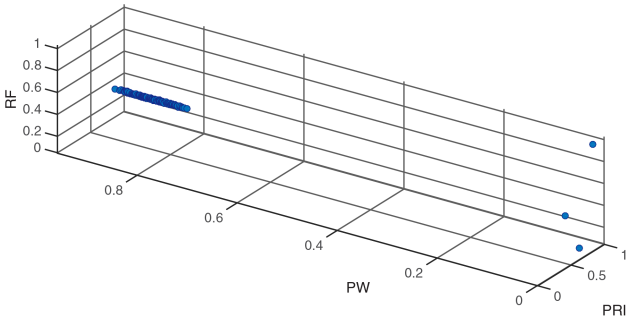

Fig. 4 *Presence of outliers in observations of a radar emitter*

3 Model

In this section, the standard Gaussian mixture model (GMM) is briefly presented as a hierarchical latent variable model before introducing missing values and outliers modelling. Finally, the PM [36] is developed and approaches for classification and clustering with a mixture model are explained.

3.1 Latent variable model

A GMM [37] is a natural framework for classification and clustering. It can be formalised as

$$p(\boldsymbol{x}|\boldsymbol{\Theta},\mathscr{K}) = \sum_{k\in\mathscr{K}} a_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{6}$$

where $\boldsymbol{x}\in\mathbb{R}^d$ is an observation, $\mathscr{K}=\{1,\ldots,K\}$ is a finite and known set of clusters and $\boldsymbol{\Theta}=(\boldsymbol{a},(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)_{k\in\mathscr{K}})$, 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 kth component distribution with a weight a_k where $a_k\geq 0$ and $\sum_{k\in\mathscr{K}} a_k = 1$.

The GMM can be formalised as a latent model since the component label associated with each data point is unobserved. To this end, a categorical variable $z\in\mathscr{K}$ can be considered to describe the index of the component distribution generating the observation variable $\boldsymbol{x}$. Then, the mixture distribution (6) is expressed as

$$p(\boldsymbol{x}|\boldsymbol{\Theta},\mathscr{K}) = \sum_{z\in\mathscr{K}} p(\boldsymbol{x}|z,\boldsymbol{\Theta},\mathscr{K})p(z|\boldsymbol{\Theta},\mathscr{K}), \tag{7}$$

where

$$p(\boldsymbol{x}|z,\boldsymbol{\Theta},\mathscr{K}) = \prod_{k\in\mathscr{K}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{\delta_z^k}, \tag{8}$$

$$p(z|\boldsymbol{\Theta},\mathscr{K}) = \mathcal{C}at(z|\boldsymbol{a}) = \prod_{k\in\mathscr{K}} a_k^{\delta_z^k} \tag{9}$$

and δ_z^k denotes the Kronecker symbol which is 1 if $z=k$ and 0 otherwise.

3.2 Missing values

Missing values can be handled by decomposing the features vector $\boldsymbol{x}\in\mathbb{R}^d$ into observed features $\boldsymbol{x}^{\text{obs}}\in\mathbb{R}^{d_{\text{obs}}}$ and missing features modelled by a latent variable $\boldsymbol{x}^{\text{miss}}\in\mathbb{R}^{d_{\text{miss}}}$ such that $1\leq d_{\text{obs}}\leq d$ and $d_{\text{miss}} = d - d_{\text{obs}}$. Reminding that conditionally to its index cluster the features vector $\boldsymbol{x}$ is Gaussian distributed

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

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

$$p(\boldsymbol{x}^{\text{miss}}|\boldsymbol{x}^{\text{obs}},z,\boldsymbol{\Theta},\mathscr{K})) = \prod_{k\in\mathscr{K}} \mathcal{N}\left(\boldsymbol{x}^{\text{miss}}|\epsilon_k^{\text{miss}},\boldsymbol{\Delta}_k^{\text{miss}}\right)^{\delta_z^k}, \tag{10}$$

where

$$\begin{aligned}\epsilon_k^{\text{miss}} &= \boldsymbol{\mu}_k^{\text{miss}} + \boldsymbol{\Sigma}_k^{\text{cov}}\boldsymbol{\Sigma}_k^{\text{obs}^{-1}}(\boldsymbol{x}^{\text{obs}} - \boldsymbol{\mu}_k^{\text{obs}}),\\ \boldsymbol{\Delta}_k^{\text{miss}} &= \boldsymbol{\Sigma}_k^{\text{miss}} - \boldsymbol{\Sigma}_k^{\text{cov}}\boldsymbol{\Sigma}_k^{\text{obs}^{-1}}\boldsymbol{\Sigma}_k^{\text{cov}'}.\end{aligned} \tag{11}$$

Then, the joint distribution of $(\boldsymbol{x}^{\text{miss}},\boldsymbol{x}^{\text{obs}})$ is derived from (10) such that

$$p(\boldsymbol{x}^{\text{miss}},\boldsymbol{x}^{\text{obs}}|z,\mathscr{K}) = \prod_{k\in\mathscr{K}} \left[\mathcal{N}\left(\boldsymbol{x}^{\text{miss}}|\epsilon_k^{\text{miss}},\boldsymbol{\Delta}_k^{\text{miss}}\right)p_k\left(\boldsymbol{x}^{\text{obs}}\right)\right]^{\delta_z^k}$$

with

$$p_k\left(\boldsymbol{x}^{\text{obs}}\right) = \mathcal{N}\left(\boldsymbol{x}^{\text{obs}}|\boldsymbol{\mu}_k^{\text{obs}},\boldsymbol{\Delta}_k^{\text{obs}}\right),$$

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

3.3 Outliers

Outliers in a GMM can be handled by introducing a latent variable u to scale each mixture component covariance matrix $\boldsymbol{\Sigma}_k$. That family of mixture models is known as scale mixtures of normal distributions [30]. Introducing the latent positive variable u into (8), the following scale component distribution is obtained

$$p(\boldsymbol{x}|u,z,\boldsymbol{\Theta},\mathscr{K}) = \prod_{k\in\mathscr{K}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,u^{-1}\boldsymbol{\Sigma}_k)^{\delta_z^k}, \tag{12}$$

and the joint distribution of $(\boldsymbol{x},u)$ is derived from (12) such that

$$p(\boldsymbol{x},u|z,\boldsymbol{\Theta},\mathscr{K}) = \prod_{k\in\mathscr{K}} \left[\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,u^{-1}\boldsymbol{\Sigma}_k)p_k(u)\right]^{\delta_z^k}, \tag{13}$$

where $p_k(u)$ is the prior distribution of u conditionally to $z=k$.

For the sake of keeping conjugacy between prior and posterior distributions of u, a Gamma distribution $\mathcal{G}(u|\alpha_k,\beta_k)$ with shape and rate parameters (α_k,β_k) is chosen for $p_k(u) = p(u|z=k) = \mathcal{G}(u|\alpha_k,\beta_k)$. Integrating (13) out u, the resulting marginal $p(\boldsymbol{x}|z,\boldsymbol{\Theta},\mathscr{K})$ is a heavy-tailed distribution known as the Student-t distribution [38].

3.4 Proposed model

Combining (8), (10) and (12), the following joint latent representation is obtained

$$\begin{aligned}p(\boldsymbol{x}^{\text{obs}},\boldsymbol{h}|\boldsymbol{\Theta},\mathscr{K})) = \prod_{k\in\mathscr{K}} \Big[&a_k\mathcal{N}\left(\boldsymbol{x}^{\text{miss}}|\epsilon_k^{\text{miss}},u^{-1}\boldsymbol{\Delta}_k^{\text{miss}}\right)\\ &\mathcal{N}\left(\boldsymbol{x}^{\text{obs}}|\boldsymbol{\mu}_k^{\text{obs}},u^{-1}\boldsymbol{\Delta}_k^{\text{obs}}\right)\mathcal{G}(u|\alpha_k,\beta_k)\Big]^{\delta_z^k},\end{aligned}$$

where $\boldsymbol{h} = (\boldsymbol{x}^{\text{miss}},u,z)$ are the latent variables.

Finally, assuming a dataset $\boldsymbol{X}\in\mathbb{R}^{d\times J}$ of i.i.d observations $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_J)$ and independent latent data $\boldsymbol{H} = (\boldsymbol{X}^{\text{miss}},\boldsymbol{u},\boldsymbol{z})$, the complete likelihood function can be expressed as

$$\begin{aligned}p(\boldsymbol{X}^{\text{obs}},\mathbf{H}|\boldsymbol{\Theta},\mathscr{K})) = \prod_{j\in\mathscr{J}}\prod_{k\in\mathscr{K}} \Big[&a_k\mathcal{N}\left(\boldsymbol{x}_j^{\text{miss}}|\epsilon_k^{\text{miss}},u_j^{-1}\boldsymbol{\Delta}_k^{\text{miss}}\right)\\ &\mathcal{N}\left(\boldsymbol{x}_j^{\text{obs}}|\boldsymbol{\mu}_k^{\text{obs}},u_j^{-1}\boldsymbol{\Delta}_k^{\text{obs}}\right)\mathcal{G}(u_j|\alpha_k,\beta_k)\Big]^{\delta_{z_j}^k},\end{aligned}$$

where $\mathscr{J}=\{1,\ldots,J\}$, $\boldsymbol{X}^{\text{obs}} = \left\{\boldsymbol{x}_1^{\text{obs}},\ldots,\boldsymbol{x}_J^{\text{obs}}\right\}$, $\boldsymbol{X}^{\text{miss}} = \left\{\boldsymbol{x}_1^{\text{miss}},\ldots,\boldsymbol{x}_J^{\text{miss}}\right\}$, $\boldsymbol{z} = \{z_j\}_{j\in\mathscr{J}}$ is the discrete variable

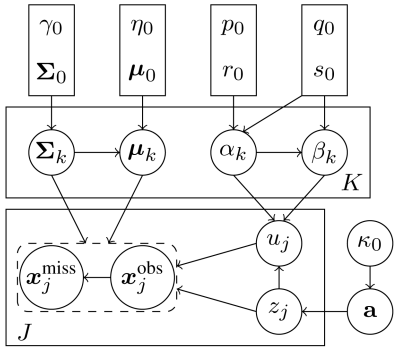

Fig. 5 *Graphical representation of the PM. The arrows represent conditional dependencies between the random variables. The K-plate represents the K mixture components and the J-plate the independent identically distributed observations $\boldsymbol{x}_j$, the scale variables u_j and the indicator variables z_j*

introduced to indicate which cluster the data $\boldsymbol{x}_j$ belongs to and $\boldsymbol{u} = \{u_j\}_{j \in \mathcal{J}}$ is the scale variable associated with $\boldsymbol{x}_j$.

Finally, the Bayesian framework imposes to specify a prior distribution $p(\boldsymbol{\Theta}|\mathcal{K})$ for the parameters $\boldsymbol{\Theta} = (\boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The resulting conjugate prior is

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

with

$$\begin{cases} p(\boldsymbol{a}|\mathcal{K}) = \mathcal{D}(\boldsymbol{a}|\boldsymbol{k}_0), \\ p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{K}) = \prod_{k \in \mathcal{K}} p(\alpha_k, \beta_k|p_0, q_0, s_0, r_0), \\ p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{K}) = \prod_{k \in \mathcal{K}} \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \eta_0^{-1}\Sigma_k)\mathcal{IW}(\boldsymbol{\Sigma}_k|\gamma_0, \boldsymbol{\Sigma}_0), \end{cases}$$

where $\boldsymbol{a}$ follows a Dirichlet distribution, $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ a normal-inverse-Wishart distribution and $p(\alpha_k, \beta_k|p_0, q_0, s_0, r_0)$ is defined below. To avoid a non-closed-form posterior distribution for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the following conditional prior is introduced:

$$p(\alpha, \beta|p_0, q_0, s_0, r_0) = p(\beta|\alpha, s_0, q_0)p(\alpha|p_0, q_0, s_0, r_0), \tag{14}$$

where $p_0, q_0, s_0, r_0 > 0$ and

$$p(\beta|\alpha, s_0, q_0) = \mathcal{G}(\beta|s_0\alpha + 1, q_0),$$

$$p(\alpha|p_0, q_0, s_0, r_0) = \frac{1}{M_0}\frac{p_0^{\alpha-1}\Gamma(s_0\alpha+1)}{q_0^{s_0\alpha+1}\Gamma(\alpha)^{r_0}}\mathbb{I}_{\{\alpha>0\}},$$

where

$$M_0 = \int \frac{p_0^{\alpha-1}\Gamma(s_0\alpha+1)}{q_0^{s_0\alpha+1}\Gamma(\alpha)^{r_0}}\mathbb{I}_{\{\alpha>0\}}\,\mathrm{d}\alpha\,. \tag{15}$$

The normalisation constant M_0 is intractable and a Laplace approximation method [39] is derived to estimate it. The directed acyclic graph of the PM is shown in Fig. 5.

3.5 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 of 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 previously 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 maximised likelihood criterion given by

$$\tilde{K} = \arg\max_{K} \log p(\boldsymbol{X}|K), \tag{16}$$

where

$$p(\boldsymbol{X}|K) = \int p(\boldsymbol{X}, \boldsymbol{\Theta}|K)\mathrm{d}\boldsymbol{\Theta}\,. \tag{17}$$

Unfortunately, (17) is intractable and many penalised likelihood criteria such as Akaike Information Criterion (AIC) [40], Bayesian Information Criterion (BIC) [41] and Integrated Completed Likelihood (ICL) [29] have been proposed. In this study, an LB for (17) is found in Section 4 and is preferred to other criteria since it does not depend on asymptotical assumptions and does not require maximum-likelihood estimates.

Then, according to $\{K_{\min}, \ldots, K_{\max}\}$, a priori range of numbers of clusters, 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. Then, $\tilde{K}$ is chosen as the maximiser of the LB introduced in the next section. After determining $\tilde{K}$, only $\boldsymbol{z}^{\tilde{K}}$ and $\boldsymbol{\Theta}^{\tilde{K}}$ are kept as estimated labels and parameters.

4 Inference

In this section, a brief introduction to variational Bayes is proposed before detailing variational posterior distributions related to latent variables $\boldsymbol{H}$ and parameters $\boldsymbol{\Theta}$ and elements of the variational LB.

4.1 Introduction to variational Bayes

VB can be viewed as a Bayesian generalisation of the expectation–maximisation algorithm [42] combined with a mean field approach [43]. It consists of approximating the intractable posterior distribution $P = p(\boldsymbol{H}, \boldsymbol{\Theta}|\boldsymbol{X}, \mathcal{K})$ by a tractable one $Q = q(\boldsymbol{H}, \boldsymbol{\Theta})$ whose parameters are chosen through a variational principle to minimise the Kullback–Leibler (KL) divergence

$$\mathrm{KL}[Q||P] = \int q(\boldsymbol{H}, \boldsymbol{\Theta})\log\left(\frac{q(\boldsymbol{H}, \boldsymbol{\Theta})}{p(\boldsymbol{H}, \boldsymbol{\Theta}|\boldsymbol{X}, \mathcal{K})}\right)\mathrm{d}\boldsymbol{H}\,\mathrm{d}\boldsymbol{\Theta}\,.$$

Noting that $p(\boldsymbol{H}, \boldsymbol{\Theta}|\boldsymbol{X}, \mathcal{K}) = (p(\boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}|\mathcal{K})/p(\boldsymbol{X}|\mathcal{K}))$, the KL divergence can be written as

$$\mathrm{KL}[Q||P] = \log p(\boldsymbol{X}|\mathcal{K}) - \mathcal{L}(Q|\mathcal{K})$$

with

$$\mathcal{L}(Q|\mathcal{K}) = \mathbb{E}_q[\log p(\boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}|\mathcal{K})] - \mathbb{E}_q[\log q(\boldsymbol{H}, \boldsymbol{\Theta})], \tag{18}$$

where $\mathbb{E}_q[\,\cdot\,]$ denotes the expectation with respect to q. As an LB for the log evidence, $\mathcal{L}(Q|\mathcal{K})$ [31] can be used as a validation criterion to choose the number of classes $\tilde{K}$ in (16) such that

$$\tilde{K} = \arg\max_{K} \mathcal{L}(Q|\mathcal{K})\,. \tag{19}$$

Then, minimising the KL divergence is equivalent to maximising $\mathscr{L}(Q|\mathscr{K})$. Assuming that $q(\boldsymbol{H},\boldsymbol{\Theta})$ can be factorised over the latent variables $\boldsymbol{H}$ and the parameters ($\boldsymbol{\Theta}$), a free-form maximisation with respect to $q(\boldsymbol{H})$ and $q(\boldsymbol{\Theta})$ leads to the following update rules:

$$\text{VBE}-\text{step}: q(\boldsymbol{H}) \propto \exp(\mathbb{E}_{\boldsymbol{\Theta}}[\log p(\boldsymbol{X},\boldsymbol{H}|\boldsymbol{\Theta},\mathscr{K})]),$$

$$\text{VBM}-\text{step}: q(\boldsymbol{\Theta}) \propto \exp(\mathbb{E}_{\boldsymbol{H}}[\log p(\boldsymbol{\Theta},\boldsymbol{X},\boldsymbol{H}|\mathscr{K})])\,.$$

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 $\mathscr{L}(Q|\mathscr{K})$. Running the algorithm steps, each posterior distribution is obtained in the following subsection.

4.2 Variational posterior distributions

Posterior distributions are similarly obtained from classical posteriors related in [31, 44, 45]. However, in this study, missing values are incorporated as latent variables in posterior calculations and a posterior distribution for missing data is proposed. Then, noting that

$$\begin{aligned}p(\boldsymbol{X},\mathbf{H}|\boldsymbol{\Theta},\mathscr{K}) &= p(\boldsymbol{X}|\boldsymbol{u},z,\boldsymbol{\Theta},\mathscr{K})p(\boldsymbol{u}|z,\boldsymbol{\Theta},\mathscr{K})p(z|\boldsymbol{\Theta},\mathscr{K}),\\ p(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{H},\mathscr{K}) &= p(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{H},\mathscr{K})p(\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{X},\boldsymbol{H},\mathscr{K})\\ &\quad\; p(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{X},\mathbf{H},\mathscr{K}),\end{aligned}$$

the following factorised forms are similarly chosen for

$$q(\boldsymbol{H}) = q(\boldsymbol{X}^{\text{miss}}|\boldsymbol{u},z)q(\boldsymbol{u}|z)q(z),$$

$$q(\boldsymbol{\Theta}) = q(\boldsymbol{a})q(\boldsymbol{\mu},\boldsymbol{\Sigma})q(\boldsymbol{\alpha},\boldsymbol{\beta}),$$

where due to conjugacy properties

$$\begin{cases} q(\boldsymbol{X}^{\text{miss}}|\boldsymbol{u},z) = \prod\limits_{j\in\mathscr{J}}\prod\limits_{k\in\mathscr{K}}\mathscr{N}\Big(\boldsymbol{x}_j^{\text{miss}}|\tilde{\epsilon}_{jk}^{\text{miss}},u_j^{-1}\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}\Big)^{\delta_{z_j}^k},\\ q(\boldsymbol{u}|z) = \prod\limits_{j\in\mathscr{J}}\prod\limits_{k\in\mathscr{K}}\mathscr{G}(\tilde{\alpha}_{jk},\tilde{\beta}_{jk})^{\delta_{z_j}^k},\\ q(z) = \prod\limits_{j\in\mathscr{J}}\mathscr{C}at(z_j|\tilde{\boldsymbol{r}}_j),\\ q(\boldsymbol{a}) = \mathscr{D}(\boldsymbol{a}|\tilde{\boldsymbol{k}}),\\ q(\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod\limits_{k\in\mathscr{K}}p(\alpha_k,\beta_k|\tilde{p}_k,\tilde{q}_k,\tilde{s}_k,\tilde{r}_k),\\ q(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod\limits_{k\in\mathscr{K}}\mathscr{N}(\boldsymbol{\mu}_k|\tilde{\boldsymbol{\mu}}_k,\tilde{\eta}_k^{-1}\boldsymbol{\Sigma}_k)\mathscr{IW}(\boldsymbol{\Sigma}_k|\tilde{\gamma}_k,\tilde{\boldsymbol{\Sigma}}_k)\,. \end{cases} \tag{20}$$

Update rules for hyper-parameters defined in (20) are detailed below.

4.3 Variational posterior distributions for latent variables

The Variational Bayes Expectation (VBE) step can be computed by developing the expectation

$$\mathbb{E}_{\boldsymbol{\Theta}}[\log p(\boldsymbol{X},\boldsymbol{H}|\boldsymbol{\Theta},K)] = \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{K}}\delta_{z_j}^k f_k(\boldsymbol{x}_j,u_j), \tag{21}$$

where

$$\begin{aligned}f_k(\boldsymbol{x}_j,u_j) = {}& \frac{-\mathbb{E}_\Theta[\log|\boldsymbol{\Sigma}_k|]}{2} - \frac{d}{2}(\log 2\pi - \log u_j)\\ &+\mathbb{E}_{\boldsymbol{\Theta}}[\log a_k] - \frac{u_j}{2}\mathbb{E}_{\boldsymbol{\Theta}}\big[\mathscr{D}(\boldsymbol{x}_j,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\big]\\ &+\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k]\mathbb{E}_{\boldsymbol{\Theta}}[\log\beta_k] - \mathbb{E}_{\boldsymbol{\Theta}}[\log\Gamma(\alpha_k)]\\ &+(\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k]-1)\log u_j - \mathbb{E}_{\boldsymbol{\Theta}}[\beta_k]u_j\,.\end{aligned} \tag{22}$$

Conditionally to $z_j = k$ and a given u_j, $\boldsymbol{x}_j$ follows a Gaussian distribution with mean $\tilde{\boldsymbol{\mu}}_k$ and covariance matrix $\tilde{\gamma}_k^{-1}u_j^{-1}\tilde{\boldsymbol{\Sigma}}_k$ since

$$\begin{aligned}\mathbb{E}_\Theta[\mathscr{D}(\boldsymbol{x}_j,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)] &= \mathbb{E}_\Theta[(\boldsymbol{x}_j-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_k)]\\ &= \tilde{\gamma}_k(\boldsymbol{x}_j-\tilde{\boldsymbol{\mu}}_k)^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{x}_j-\tilde{\boldsymbol{\mu}}_k) + \frac{d}{\tilde{\eta}_k}\,.\end{aligned}$$

Therefore, the distribution for $\boldsymbol{x}_j^{\text{miss}}$ in (20) is deduced from (10) and (11)

$$q(\boldsymbol{x}_j^{\text{miss}}|u_j,z=k) = \mathscr{N}\Big(\tilde{\epsilon}_{jk}^{\text{miss}},u_j^{-1}\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}\Big),$$

where

$$\tilde{\epsilon}_{jk}^{\text{miss}} = \tilde{\boldsymbol{\mu}}_k^{\text{miss}} + \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}}\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}}(\boldsymbol{x}_j^{\text{obs}} - \tilde{\boldsymbol{\mu}}_k^{\text{obs}}), \tag{23}$$

$$\tilde{\boldsymbol{\Delta}}_k^{\text{miss}} = \frac{\tilde{\boldsymbol{\Sigma}}_k^{\text{miss}} - \tilde{\boldsymbol{\Sigma}}_k^{\text{cov}}\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}^{-1}}\tilde{\boldsymbol{\Sigma}}_k^{\text{cov}'}}{\tilde{\gamma}_k}\,. \tag{24}$$

Auxiliary variables $\tilde{\boldsymbol{x}}_j \in \mathbb{R}^d$ and $\boldsymbol{\Delta}_k^{x_j} \in \mathbb{R}^{d\times d}$ are introduced for the Variational Bayes Maximisation (VBM) step such that

$$\tilde{\boldsymbol{x}}_j = \mathbb{E}_{\mathbf{H}}[\boldsymbol{x}_j] = \begin{pmatrix}\tilde{\epsilon}_{jk}^{\text{miss}}\\ \boldsymbol{x}_j^{\text{obs}}\end{pmatrix},$$

$$\boldsymbol{\Delta}_k^{x_j} = \mathbb{E}_{\mathbf{H}}[\boldsymbol{x}_j\boldsymbol{x}_j^{\mathrm{T}}] - \tilde{\boldsymbol{x}}_j\tilde{\boldsymbol{x}}_j^{\mathrm{T}} = \begin{pmatrix}\tilde{\boldsymbol{\Delta}}_k^{\text{miss}} & \mathbf{0}^{d_{\text{miss}}^j\times d_{\text{obs}}^j}\\ \mathbf{0}^{d_{\text{obs}}^j\times d_{\text{miss}}^j} & \mathbf{0}^{d_{\text{obs}}^j\times d_{\text{obs}}^j}\end{pmatrix}.$$

Marginalising over $\boldsymbol{x}_j^{\text{miss}}$, (22) becomes

$$\begin{aligned}\log\int\exp\{f_k(\boldsymbol{x}_j,u_j)\}\mathrm{d}\boldsymbol{x}_j^{\text{miss}} = {}& -\frac{\mathbb{E}_\Theta[\log|\boldsymbol{\Sigma}_k|]}{2}\\ &-\frac{d_{\text{obs}}^j}{2}(\log 2\pi - \log u_j)\\ &+\mathbb{E}_{\boldsymbol{\Theta}}[\log a_k] + \frac{\log|\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}|}{2}\\ &-\frac{u_j}{2}\Big(\mathscr{D}\Big(\boldsymbol{x}_j^{\text{obs}},\tilde{\mu}_k^{\text{obs}},\tilde{\boldsymbol{\Delta}}_k^{\text{obs}}\Big) + \frac{d}{\tilde{\eta}_k}\Big)\\ &+\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k]\mathbb{E}_{\boldsymbol{\Theta}}[\log\beta_k]\\ &-\mathbb{E}_{\boldsymbol{\Theta}}[\log\Gamma(\alpha_k)]\\ &+(\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k]-1)\log u_j - \mathbb{E}_{\boldsymbol{\Theta}}[\beta_k]u_j,\end{aligned} \tag{25}$$

where d_{obs}^j is the dimension of $\boldsymbol{x}_j^{\text{obs}}$

$$\tilde{\boldsymbol{\Delta}}_k^{\text{obs}} = \frac{\Big(\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}-1} + 2\times\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}-1}\tilde{\boldsymbol{\Sigma}}_k^{\text{cov}'}\Big(\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}\Big)^{-1}\tilde{\boldsymbol{\Sigma}}_k^{\text{cov}}\tilde{\boldsymbol{\Sigma}}_k^{\text{obs}-1}\Big)^{-1}}{\tilde{\gamma}_k}$$

and

$$\mathscr{D}(\boldsymbol{x},\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = (\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\,.$$

Then the conditional posterior for u_j in (20) is deduced from (25) such that

$$q(u_j|z_j=k) \sim \mathscr{G}(\tilde{\alpha}_{jk},\tilde{\beta}_{jk}),$$

where

$$\tilde{\alpha}_{jk} = \mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] + \frac{d_{\text{obs}}^j}{2}, \tag{26}$$

$$\tilde{\beta}_{jk} = \frac{1}{2}\left(\mathscr{D}\left(\boldsymbol{x}_j^{\text{obs}}, \tilde{\boldsymbol{\mu}}_k^{\text{obs}}, \tilde{\boldsymbol{\Delta}}_k^{\text{obs}}\right) + \frac{d}{\tilde{\eta}_k}\right) + \mathbb{E}_{\boldsymbol{\Theta}}[\beta_k]\,. \tag{27}$$

Expectations of $u_j|z_j$ are derived from the Gamma distribution properties such that

$$\mathbb{E}_{\boldsymbol{H}}[u_j] = \frac{\tilde{\alpha}_{jk}}{\tilde{\beta}_{jk}}, \quad \mathbb{E}_{\boldsymbol{H}}[\log u_j] = \psi(\tilde{\alpha}_{jk}) - \log\tilde{\beta}_{jk},$$

where $\psi(\cdot)$ is the digamma function. Noting that

$$\log q(\boldsymbol{z}) = \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{K}} \delta_{z_j}^k \log\tilde{r}_{jk},$$

$\tilde{r}_{jk} = q(z_j = k)$, called responsibility is obtained as follows:

$$\begin{aligned}\tilde{r}_{jk} &\propto \int \exp f_k(\boldsymbol{x}_j, u_j)\,\mathrm{d}\boldsymbol{x}_j^{\text{miss}}\,\mathrm{d}u_j \\ &\propto \frac{\exp(\mathbb{E}_{\boldsymbol{\Theta}}[\log a_k] + \mathbb{E}_{\boldsymbol{\Theta}}[\log\beta_k]\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k])\Gamma\left(\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] + (d_{\text{obs}}^j/2)\right)}{\exp\left(\left(\left(\mathbb{E}_{\boldsymbol{\Theta}}[\log|\boldsymbol{\Sigma}_k|] - \log|\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}|\right)/2\right) + \mathbb{E}_{\Theta}[\log\Gamma(\alpha_k)]\right)} \\ &\quad\times\left[\mathbb{E}_{\boldsymbol{\Theta}}[\beta_k] + \frac{\mathscr{D}\left(\boldsymbol{x}_j^{\text{obs}}, \tilde{\boldsymbol{\mu}}_k^{\text{obs}}, \tilde{\boldsymbol{\Delta}}_k^{\text{obs}}\right) + (d/\tilde{\eta}_k)}{2}\right]^{-\left(\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] + (d_{\text{obs}}^j/2)\right)}.\end{aligned} \tag{28}$$

The expectation of $\delta_{z_j}^k$ is deduced from the categorical distribution property such that

$$\mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k] = \tilde{r}_{jk}\,.$$

4.4 Variational posterior distributions for parameters

The VBM step can be computed by developing the expectation

$$\begin{aligned}\mathbb{E}_{\boldsymbol{H}}[\log p(\boldsymbol{\Theta}, \boldsymbol{X}, \boldsymbol{H}|K)] = \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{K}} \mathbb{E}_{\boldsymbol{H}}\big[\delta_{z_j}^k(&\log a_k \\ &+\log\mathscr{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &+\log\mathscr{G}(u_j|\alpha_k, \beta_k))\big] + \log p(\boldsymbol{\Theta}|\mathscr{K})\end{aligned}$$

and gathering common terms to obtain hyper-parameters defined in (20) such that

$$\begin{aligned}&\tilde{k}_k = k_0 + J\bar{\pi}_k, \quad \tilde{\eta}_k = \eta_0 + J\bar{\omega}_k, \\ &\tilde{p}_k = p_0\exp\left(J\bar{\delta}_k\right), \quad \tilde{q}_k = q_0 + J\bar{\omega}_k, \\ &\tilde{r}_k = r_0 + J\bar{\pi}_k, \quad \tilde{s}_k = s_0 + J\bar{\pi}_k, \\ &\tilde{\boldsymbol{\mu}}_k = \frac{\eta_0\boldsymbol{\mu}_0 + J\bar{\omega}_k\boldsymbol{\mu}_k^x}{\tilde{\eta}_k}, \\ &\tilde{\boldsymbol{\Sigma}}_k = \boldsymbol{\Sigma}_0 + \frac{J\bar{\omega}_k\eta_0}{\tilde{\eta}_k}(\boldsymbol{\mu}_k^x - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k^x - \boldsymbol{\mu}_0)^{\mathrm{T}} + J\bar{\omega}_k\boldsymbol{\Sigma}_k^x + \boldsymbol{\Sigma}_k^m \\ &\tilde{\gamma}_k = \gamma_0 + J\bar{\pi}_k,\end{aligned} \tag{29}$$

where auxiliary variables are obtained as follows:

$$\bar{\pi}_k = \frac{1}{J}\sum_{j\in\mathscr{J}} \mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k],$$

$$\bar{\omega}_k = \frac{1}{J}\sum_{j\in\mathscr{J}} \mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k]\mathbb{E}_{\boldsymbol{H}}[u_j],$$

$$\bar{\delta}_k = \frac{1}{J}\sum_{j\in\mathscr{J}} \mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k]\mathbb{E}_{\boldsymbol{H}}[\log u_{jk}]\,.$$

$$\boldsymbol{\mu}_k^x = \frac{1}{J\bar{\omega}_k}\sum_{j\in\mathscr{J}} \mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k]\mathbb{E}_{\boldsymbol{H}}[u_j]\tilde{\boldsymbol{x}}_j,$$

$$\boldsymbol{\Sigma}_k^x = \frac{1}{J\bar{\omega}_k}\sum_{j\in\mathscr{J}} \mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k]\mathbb{E}_{\boldsymbol{H}}[u_j](\tilde{\boldsymbol{x}}_j - \boldsymbol{\mu}_k^x)(\tilde{\boldsymbol{x}}_j - \boldsymbol{\mu}_k^x)^{\mathrm{T}},$$

$$\boldsymbol{\Sigma}_k^m = \sum_{j\in\mathscr{J}} \mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k]\boldsymbol{\Delta}_k^{x_j}\,.$$

Using the properties of the Dirichlet and the Inverse Wishart distribution, the following expectations are defined

$$\mathbb{E}_{\boldsymbol{\Theta}}[\log a_k] = \psi(\tilde{\kappa}_k) - \psi\left(\sum_{k'=1}^{K}\tilde{\kappa}_{k'}\right),$$

$$\mathbb{E}_{\boldsymbol{\Theta}}[\boldsymbol{\Sigma}_k^{-1}] = \tilde{\gamma}_k\tilde{\boldsymbol{\Sigma}}_k^{-1},$$

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

4.5 LB elements and expectations

The LB (18) is proven to increase at each VB iteration and its difference between two iterations can be used as a stop criterion. The introduction of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ slightly modifies the LB since the prior distribution (14) as well as the posterior distributions (20) have to be taken into account. LB elements are presented below.

$$\begin{aligned}\mathbb{E}_q[\log p(\boldsymbol{X}, \boldsymbol{H}, \Theta|K)] = &\sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{K}} \mathbb{E}_{\boldsymbol{H}}[\delta_{z_j}^k]\{\mathbb{E}_{\Theta}[\log a_k] \\ &-\frac{\mathbb{E}_{\boldsymbol{\Theta}}[\log|\boldsymbol{\Sigma}_k|]}{2} - \frac{d}{2}(\log 2\pi - \mathbb{E}_{\boldsymbol{H}}[\log u_{jk}]) \\ &-\mathbb{E}_{\boldsymbol{\Theta}}[\log\Gamma(\alpha_k)] \\ &+\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k]\mathbb{E}_{\boldsymbol{\Theta}}[\log\beta_k] - \frac{\mathbb{E}_{\boldsymbol{H}}[u_{jk}]}{2}\mathbb{E}_{\boldsymbol{\Theta},\boldsymbol{H}}\left[\mathscr{D}(\boldsymbol{x}_j, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right] \\ &+(\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] - 1)\mathbb{E}_{\boldsymbol{u}}[\log u_j] - \mathbb{E}_{\boldsymbol{\Theta}}[\beta_k]\mathbb{E}_{\boldsymbol{u}}[u_j]\} \\ &+\sum_{k\in\mathscr{K}}(\kappa_0 - 1)\mathbb{E}_{\boldsymbol{\Theta}}[\log a_k] - \frac{d}{2}(\log 2\pi - \log\eta_0) \\ &-\frac{\mathbb{E}_{\boldsymbol{\Theta}}[\log|\boldsymbol{\Sigma}_k|]}{2} - \frac{\eta_0}{2}\mathbb{E}_{\boldsymbol{\Theta}}[\mathscr{D}(\boldsymbol{\mu}_k, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_k)] \\ &-\frac{\mathrm{tr}\{\boldsymbol{\Sigma}_0\mathbb{E}_{\boldsymbol{\Theta}}[\boldsymbol{\Sigma}_k^{-1}]\}}{2} \\ &-\frac{\gamma_0 + d + 1}{2}\mathbb{E}_{\boldsymbol{\Theta}}[\log|\boldsymbol{\Sigma}_k|] + (\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] - 1)\log p_0 \\ &-r_0\mathbb{E}_{\boldsymbol{\Theta}}[\log\Gamma(\alpha_k)] + s_0\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k]\mathbb{E}_{\boldsymbol{\Theta}}[\log\beta_k] - q_0\mathbb{E}_{\boldsymbol{\Theta}}[\beta_k] \\ &+K(\log c_{\mathscr{D}}(\kappa_0) + \log c_{\mathscr{IW}}(\gamma_0, \boldsymbol{\Sigma}_0) - \log M_0),\end{aligned}$$

$$\begin{aligned}\mathbb{E}_q[\log q(\boldsymbol{H}, \Theta|K)] = &\sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{K}} \mathbb{E}_{\mathbf{H}}[\delta_{z_j}^k]\{\log\tilde{r}_{jk} + \tilde{\alpha}_k\log\tilde{\beta}_k \\ &-\log\Gamma(\tilde{\alpha}_k) + (\tilde{\alpha}_k - 1)\mathbb{E}_{\boldsymbol{u}}[\log u_{jk}] - \tilde{\beta}_k\mathbb{E}_{\boldsymbol{u}}[u_j]\} \\ &+\sum_{k\in\mathscr{K}}(\tilde{\kappa}_k - 1)\mathbb{E}_{\Theta}[\log a_k] - \frac{d}{2}(\log 2\pi - \log\tilde{\eta}_k) - \log M_k \\ &-\frac{\mathrm{tr}\{\tilde{\Sigma}_k\mathbb{E}_{\Theta}[\Sigma_k^{-1}]\}}{2} - \frac{\tilde{\gamma}_k + d + 1}{2}\mathbb{E}_{\Theta}[\log|\Sigma_k|] - \tilde{q}_k\mathbb{E}_{\Theta}[\beta_k] \\ &-\frac{\mathbb{E}_{\Theta}[\log|\Sigma_k|]}{2} + (\mathbb{E}_{\Theta}[\alpha_k] - 1)\log\tilde{p}_k - \tilde{r}_k\mathbb{E}_{\Theta}[\log\Gamma(\alpha_k)] \\ &+\tilde{s}_k\mathbb{E}_{\Theta}[\alpha_k]\mathbb{E}_{\Theta}[\log\beta_k] + \log c_{\mathscr{D}}(\tilde{\kappa}) + \log c_{\mathscr{IW}}(\tilde{\gamma}_k, \tilde{\Sigma}_k),\end{aligned}$$

where $c_{\mathscr{D}}(\cdot)$ and $c_{\mathscr{IW}}(\cdot,\cdot)$ are the normalisation constants of the Dirichlet and Inverse Wishart distributions.

Posterior expectations of β_k are derived from the posterior Gamma distribution (20) properties and can easily be computed conditionally to α_k

Table 1 Initialisation of hyper-parameter values

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

$$\mathbb{E}_{\boldsymbol{\Theta}}[\beta_k] = \frac{\tilde{s}_k \mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] + 1}{\tilde{q}_k},$$

$$\mathbb{E}_{\boldsymbol{\Theta}}[\log \beta_k] = \mathbb{E}_{\boldsymbol{\Theta}}[\psi(\tilde{s}_k \alpha_k + 1)] - \log \tilde{q}_k .$$

However, expectations depending on α_k are intractable

$$\mathbb{E}_{\boldsymbol{\Theta}}[\psi(\tilde{s}_k \alpha_k + 1)] = \int \psi(\tilde{s}_k \alpha_k + 1) p(\alpha_k | \tilde{p}_k, \tilde{r}_k) \mathrm{d}\alpha_k, \tag{30}$$

$$\mathbb{E}_{\boldsymbol{\Theta}}[\alpha_k] = \int \alpha_k p(\alpha_k | \tilde{p}_k, \tilde{r}_k) \mathrm{d}\alpha_k, \tag{31}$$

$$\mathbb{E}_{\boldsymbol{\Theta}}[\log \Gamma(\alpha_k)] = \int \log \Gamma(\alpha_k) p(\alpha_k | \tilde{p}_k, \tilde{r}_k) \mathrm{d}\alpha_k . \tag{32}$$

Since LB calculation is required as a stop criterion, expectations (30), (31) and (32) have to be approximated. As for the normalising constant $M0$ defined in (15), a Laplace approximation is performed for each expectation.

5 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 (KNN) algorithm, random forests (RdF) the KM algorithm and the DBSCAN are also evaluated. Two experiments are carried out to evaluate classification and clustering performances 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 performances are shown to exhibit the effectiveness of the PM. For each experiment, hyper-parameters are initialised as in Table 1 and 100 simulations are performed to take into account randomness of data deletion.

5.1 Data

Real data are acquired from the system detailed in Section 2. 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. The dataset is shown in Fig. 6. However, extra missing values are added to evaluate the limits of the proposed approach. Missing information is 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 [46]. In this study, two classical imputation methods, based on statistical analysis and machine learning, are performed. First, the mean imputation consists of filling a missing component of an observation by the average of observed values of that component. This method has the obvious disadvantage that it underrepresents the variability and also ignores correlations between observations [47]. Then, imputation can be processed through a KNN method [48] 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 [28]. These imputation methods are compared with

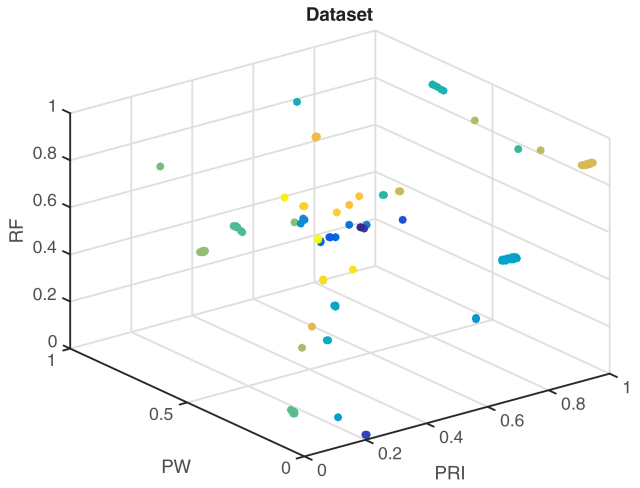

Fig. 6 *Dataset gathering 6300 observations from 42 radar emitters. Some clusters are completely separable whereas some others share features and cannot be linearly separated*

the proposed approach in terms of classification, clustering and reconstruction performances. For the comparison of reconstruction performances, mean-squared errors between original data and previous imputation methods are compared with the mean-squared error between original data and the variational posterior marginal mean of missing data given by

$$\begin{aligned} \forall j \in \mathcal{J}, \quad \mathbb{E}_q[\boldsymbol{x}_j^{\text{miss}}] &= \mathbb{E}_q\left(\int \int q(\boldsymbol{x}_j^{\text{miss}}, u_j, z_j) \, \mathrm{d}u_j \, \mathrm{d}z_j \right) \\ &= \sum_{k \in \mathcal{K}} \tilde{r}_{jk} \tilde{\epsilon}_{jk}^{\text{miss}} . \end{aligned} \tag{33}$$

5.2 Classification experiment

The classification experiment evaluates the ability of each algorithm to assign unlabelled data to one of the K classes trained by a set of labelled data. As developed in Section 3.5, the classification task is decomposed into a training step and a prediction step defined in procedures 1 and 2 (see Figs. 7 and 8). The training step consists of estimating variational parameters of $q(\boldsymbol{\Theta})$ defined in (20) given a set of training data with known labels. As for the prediction step, it results in associating new data to the class that maximises 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 PM 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 accurate metric is chosen for the classification experiment and observations belonging to the reject class are considered as misclassification errors.

For the classification experiment, results are shown in Fig. 9. Without missing data, both algorithms perfectly classify the 42 radar emitters. When the proportion of missing values increases, the PM outperforms comparisons, algorithms and achieves an accuracy of 85% for 40% of deleted values whereas the accuracy

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 $\tilde{\alpha}_{jk}$, $\tilde{\beta}_{jk}$, $\tilde{\boldsymbol{\epsilon}}_{jk}^{\text{miss}}$, $\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}$ (23-24-26-27)
Update $\tilde{\kappa}_k$, $\tilde{\eta}_k$, $\tilde{\gamma}_k$, $\tilde{p}_k$, $\tilde{r}_r$, $\tilde{s}_k$, $\tilde{q}_k$, $\tilde{\boldsymbol{\mu}}_k$, $\tilde{\boldsymbol{\Sigma}}_k$ (29)
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

Fig. 7 *Procedure 1: classification: training 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 $\tilde{\alpha}_{jk}$, $\tilde{\beta}_{jk}$, $\tilde{\boldsymbol{\epsilon}}_{jk}^{\text{miss}}$, $\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}$, $\tilde{r}_{jk}$ (23-24-26-27-28)
return $\tilde{\boldsymbol{z}}^{\text{pred}}$ such that each $\tilde{z}_j^{\text{pred}} = \arg\max_{k\in\mathcal{K}} \tilde{r}_{jk}$

Fig. 8 *Procedure 2: classification: prediction step*

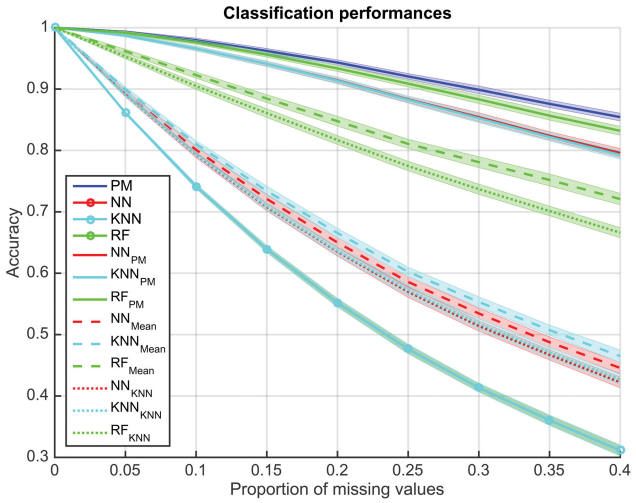

Fig. 9 *Classification performances are presented for the PM in blue, the NN in red, the RnF in green and the KNN in cyan. The solid lines represent the average accuracies with discarded observations for the NN, the RnF and the KNN, the dashed lines stand for the average accuracies with mean imputation for the NN, the RnF and the KNN whereas the dotted lines show average accuracies with KNN imputation. Shaded error regions represent standard deviations of accuracies*

of NN and KNN is <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 PM reveals that the proposed method embeds a more efficient inference method than other imputation methods. That result is confirmed in Fig. 9 where comparison algorithms are applied to the data reconstructed by the PM. Indeed when the proposed inference is chosen, performances of NN and KNN increase up to 80% for 40% of deleted values and the RnF has almost the same performances than the PM. Fig. 9 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 in Fig. 10 where the PM exhibits a lower mean-squared error for missing data imputation than the mean and KNN imputation methods. The effectiveness of the PM can be explained by the fact that missing data imputation methods can create outliers that deteriorate performances of classification algorithms whereas the inference on missing data and labels prediction are jointly estimated in the PM. Indeed, embedding the inference procedure into the model framework allows properties of the model, such as outliers handling, to counterbalance drawbacks of imputation

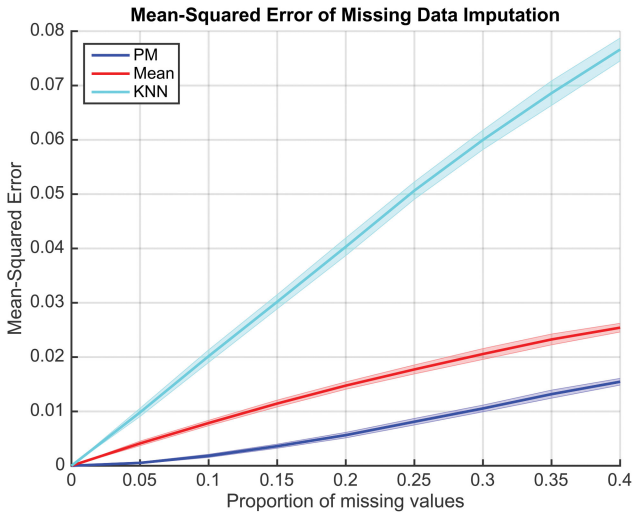

Fig. 10 *Mean-squared errors of missing data imputation methods are presented in blue for the PM, in red for the NN and in cyan for the KNN. Solid lines are average mean-squared errors and shaded error regions represent standard deviations of mean-squared errors*

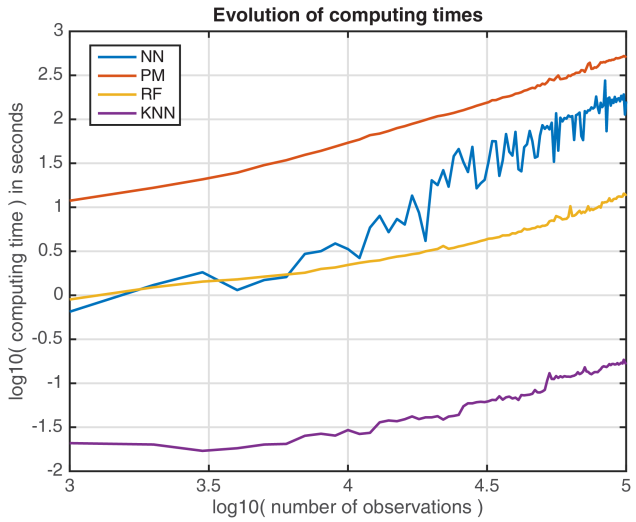

Fig. 11 *Evolution of computing times taken by model learning for RnF, K-NNalgorithm, NN and the PM*

methods such as outliers creation. Concerning the computational burden of the proposed approach, Fig. 11 shows the evolution of computing times taken by model learning of the PM and comparison algorithms according to different numbers of observations. Considering that the learning of the PM is done offline and that its code can be drastically optimised since it is only developed under Matlab, the computational burden of the proposed approach is acceptable. Indeed the PM 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.

5.3 Clustering experiments

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 Section 3.5, 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 (20) and labels of data in a mixture of K components. Then, the value of K that maximises the LB (19) is retained as the posterior number of clusters as well as its associated parameters (see Figs. 12 and 13).

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 $\tilde{\alpha}_{jk}$, $\tilde{\beta}_{jk}$, $\tilde{\boldsymbol{\epsilon}}_{jk}^{\text{miss}}$, $\tilde{\boldsymbol{\Delta}}_k^{\text{miss}}$, $\tilde{r}_{jk}$ (23-24-26-27-28)
Update $\tilde{\kappa}_k$, $\tilde{\eta}_k$, $\tilde{\gamma}_k$, $\tilde{p}_k$, $\tilde{r}_r$, $\tilde{s}_k$, $\tilde{q}_k$, $\tilde{\boldsymbol{\mu}}_k$, $\tilde{\boldsymbol{\Sigma}}_k$ (29)
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

Fig. 12 *Procedure 3: semi-supervised classification*

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 $\tilde{L}$
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 $\tilde{\boldsymbol{\Theta}}^{\tilde{K}} = \left(\tilde{\kappa}_k, \tilde{\eta}_k, \tilde{\gamma}_k, \tilde{p}_k, \tilde{r}_r, \tilde{s}_k, \tilde{q}_k, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k\right)_{k=1}^{\tilde{K}}$ such that $\tilde{K} = \arg\max_K \mathcal{L}^K$

Fig. 13 *Procedure 4: clustering procedure*

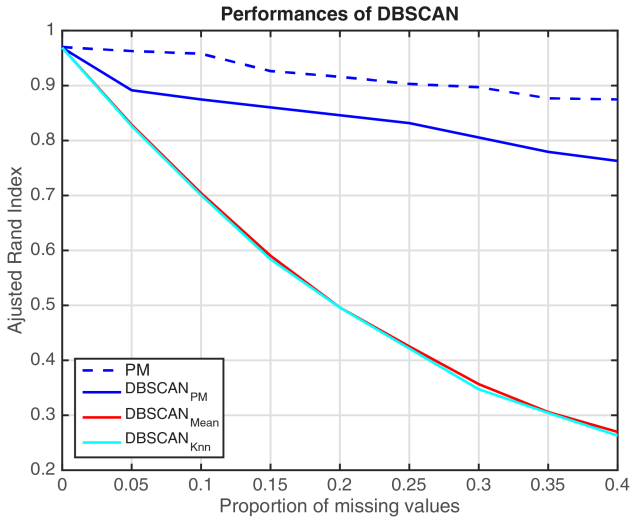

Fig. 14 *Performances of the PM compared with DBSCAN for K = 42 according to different proportions of missing values and imputation methods on realistic data*

According to the dataset visualised in Fig. 6, $K_{\min}$ 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 = 8×10^{-3} by using a heuristic proposed in the original paper [24]. A supervised initialisation is retained for the PM due to its sensitivity to initialisation. It consists of initialising prior component means $\boldsymbol{\mu}_0$ from results of a KM 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.

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 $K \in \{K_{\min}, \ldots, K_{\max}\}$. Performances are evaluated through the adjusted rand index (ARI) [49] that compare estimated partitions of data with the ground-truth. Results of the

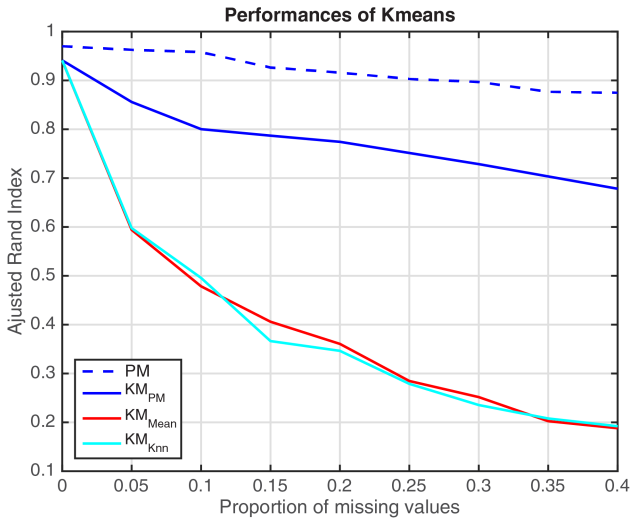

Fig. 15 *Performances of the PM compared with a KM algorithm for K = 42 according to different proportions of missing values and imputation methods on realistic data*

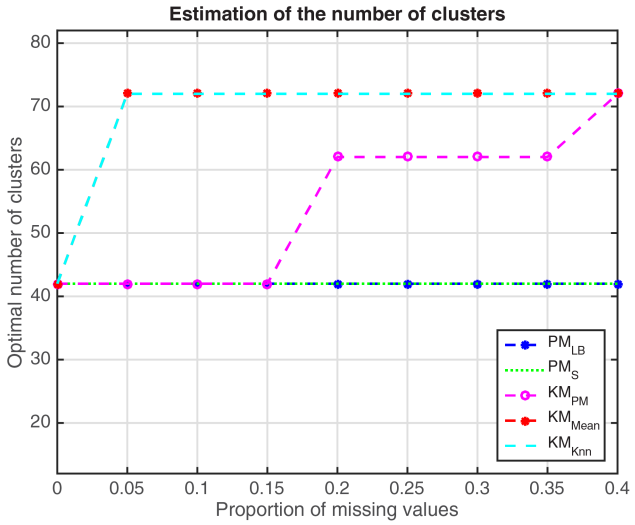

Fig. 16 *Estimation of the number of clusters using the LB and the silhouette score (S) for the PM and only the silhouette score (S) for the KM algorithm*

first experiment on realistic data are shown in Figs. 14 and 15. Without the presence of missing values, performances of DBSCAN and the PM are similar to an ARI of 97% (Fig. 14) whereas the KM algorithm ARI reaches 95% (Fig. 15). When the proportion of missing values increases, the PM outperforms both DBSCAN and KM and achieves an ARI of 87% for 40% of deleted values whereas the ARI of comparison algorithms is <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 Figs. 14 and 15 where DBSCAN and KM are applied on data reconstructed by the PM. Indeed, performances 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 LB (19) and the average Silhouette score [50] are criteria used to select the optimal number of clusters for the PM and the KM 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 realistic data are visible in Figs. 16 and 17. Fig. 17 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

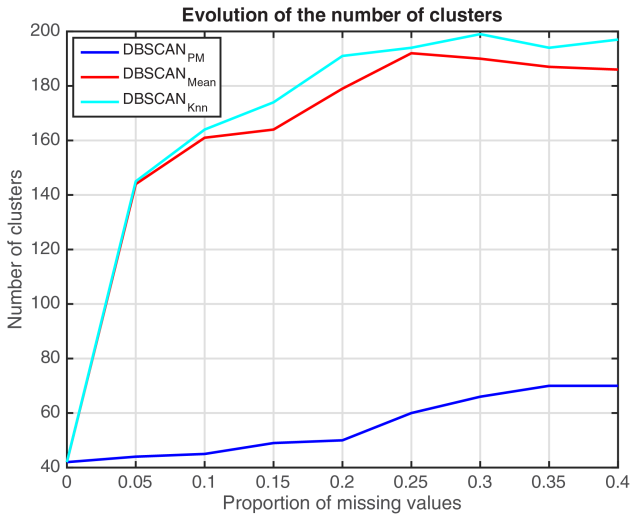

Fig. 17 *Estimation of the number of clusters by DBSCAN according to mean imputation, k-NN imputation and posterior reconstruction of the PM*

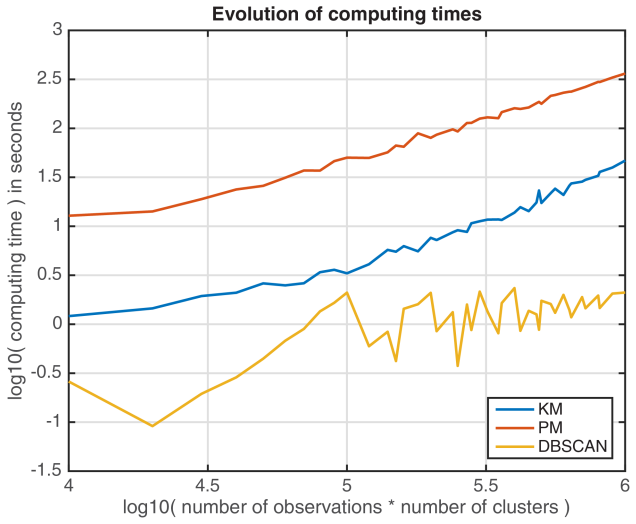

Fig. 18 *Evolution of computing times for DBSCAN, KM algorithm and the PM*

manages outliers by creating new clusters, results in Fig. 17 can be used to evaluate the performances of imputations methods. For mean and k-NN imputation methods, DBSCAN estimates the number of clusters >140 as a proportion of missing values is ≥5%. When DBSCAN is run on the posterior reconstruction (33), the estimated number of clusters stays under 50 until 20% of missing values and reaches 70% for 40% of missing values. These performances indicate that the proposed approach creates fewer outliers than other imputation methods by providing a more robust inference on missing data since DBSCAN localises fewer outliers in the posterior reconstruction (33) than in standard imputation methods. Fig. 16 presents numbers of clusters selected by the LB and average Silhouette scores for the PM and KM algorithm according to different proportions of missing values and imputation methods. Without missing data, the correct number of clusters ($K = 42$) is selected by the two criteria for the KM algorithm and the PM. In the presence of missing values, the average Silhouette score always selects $K = 72$ when the KM algorithm is run on data completed by standard imputation methods. When the KM algorithm performs clustering on the posterior reconstruction (33), 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 >20%. Eventually, when the PM 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 PM. As previously, the PM provides a more robust inference on missing data since the average Silhouette score chooses the more representative number of clusters when the KM algorithm is run on the posterior reconstruction (33) than on data completed by standard imputation methods. Furthermore, since the LB 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.

Fig. 18 shows the evolution of computing times taken by model learning of the PM and comparison algorithms according to different numbers of clusters and observations. As the model learning in Section 5.2, 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 KM algorithm, the computational burden of the proposed approach is still acceptable and meets operational requirements.

6 Conclusion

In this study, we propose a mixture model to classify and cluster radar emitters from different types. Radar signals are often partially observed due to imperfect conditions of acquisition and deficient hardware. 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. Benefiting from Gaussian properties and the introduction of latent variables, the PM 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. 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 showed that the proposed approach handles both outliers and missing values and can outperform standard algorithms in 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.

7 References

[1] Schleher, D.C.: 'Introduction to electronic warfare'. Tech. Rep., Eaton Corp., AIL Div., Deer Park, NY, 1986

[2] Rogers, J.: 'ESM processor system for high pulse density radar environments', *IEE Proc. F, Commun. Radar Signal Process.*, 1985, **132**, (7), pp. 621–625

[3] Wiley, R.G.: '*Electronic intelligence: the analysis of radar signals*' (Artech House, Inc., Dedham, MA, 1982), 250p

[4] Dudczyk, J.: 'Radar emission sources identification based on hierarchical agglomerative clustering for large data sets', *J. Sens.*, 2016, **2016**, pp. 1–9

[5] Shieh, C.-S., Lin, C.-T.: 'A vector neural network for emitter identification', *IEEE Trans. Antennas Propag.*, 2002, **50**, (8), pp. 1120–1127

[6] Petrov, N., Jordanov, I., Roe, J.: 'Radar emitter signals recognition and classification with feedforward networks', *Procedia Comput. Sci.*, 2013, **22**, pp. 1192–1200

[7] Li, H., Jin, W.D., Liu, H.D., *et al.*: 'Work mode identification of airborne phased array radar based on the combination of multi-level modeling and deep learning'. 2016 35th Chinese Control Conf. (CCC), Chengdu, China, July 2016, pp. 7005–7010

[8] Sun, J.: 'Radar emitter classification based on unidimensional convolutional neural network', *IET Radar Sonar Navig.*, 2018, **12**, pp. 862–867, Available at http://digital-library.theiet.org/content/journals/10.1049/iet-rsn.2017.0547

[9] Yang, Z., Wu, Z., Yin, Z., *et al.*: 'Hybrid radar emitter recognition based on rough k-means classifier and relevance vector machine', *Sensors*, 2013, **13**, (1), pp. 848–864. Available at http://www.mdpi.com/1424-8220/13/1/848

[10] Chen, W.: 'Radar emitter classification for large data set based on weighted-xgboost', *IET Radar Sonar Navig.*, 2017, **11**, pp. 1203–1207(4). Available at http://digitallibrary.theiet.org/content/journals/10.1049/iet-rsn.2016.0632

[11] He, A.-L., Zeng, D.-G., Wang, J., *et al.*: 'Multi-parameter signal sorting algorithm based on dynamic distance clustering', *J. Electron. Sci. Technol.*, 2009, **7**, (3), pp. 249–253

[12] Zhou, D., Wang, X., Cheng, S., *et al.*: 'An online multisensor data fusion framework for radar emitter classification', *Int. J. Aerosp. Eng.*, 2016, **2016**, pp. 1–16

[13] Chen, Y.M., Lin, C.-M., Hsueh, C.-S.: 'Emitter identification of electronic intelligence system using type-2 fuzzy classifier', *Syst. Sci. Control Eng., Open Access J.*, 2014, **2**, (1), pp. 389–397

[14] Dudczyk, J., Kawalec, A.: 'Specific emitter identification based on graphical representation of the distribution of radar signal parameters', *Bull. Pol. Acad. Sci. Tech. Sci.*, 2015, **63**, (2), pp. 391–396
[15] Shi, Y.: 'Kernel canonical correlation analysis for specific radar emitter identification', *Electron. Lett.*, 2014, **50**, pp. 1318–1320(2). Available at http://digital-library.theiet.org/content/journals/10.1049/el.2014.1458
[16] Kawalec, A., Owczarek, R.: 'Radar emitter recognition using intrapulse data'. 15th Int. Conf. on Microwaves, Radar and Wireless Communications, 2004. MIKON-2004, Warsaw, Poland, 2004, vol. 2, pp. 435–438
[17] Germain, M., Bénié, G.B., Boucher, J.-M., *et al.*: 'Contribution of the fractal dimension to multiscale adaptive filtering of SAR imagery', *IEEE Trans. Geosci. Remote Sens.*, 2003, **41**, (8), pp. 1765–1772
[18] Dudczyk, J., Kawalec, A.: 'Identification of emitter sources in the aspect of their fractal features', *Bull. Pol. Acad. Sci. Tech. Sci.*, 2013, **61**, (3), pp. 623–628
[19] Dudczyk, J., Kawalec, A.: 'Fast-decision identification algorithm of emission source pattern in database', *Bull. Pol. Acad. Sci. Tech. Sci.*, 2015, **63**, (2), pp. 385–389
[20] Milojevic, D.J., Popovic, B.M.: 'Improved algorithm for the deinterleaving of radar pulses', *IEE Proc. F, Radar Signal Process.*, 1992, **139**, (1), pp. 98–104
[21] Keshavarzi, M., Pezeshk, A.M.: 'A simple geometrical approach for deinterleaving radar pulse trains'. 2016 UKSim-AMSS 18th Int. Conf. on Computer Modelling and Simulation (UKSim), Cambridge, UK, 2016, pp. 172–177
[22] Bishop, C.M.: '*Pattern recognition and machine learning (information science and statistics)*' (Springer-Verlag, Berlin, Heidelberg, 2006)
[23] Breiman, L.: 'Random forests', *Mach. Learn.*, 2001, **45**, (1), pp. 5–32
[24] Ester, M., Kriegel, H.-P., Sander, J., *et al.*: 'A density-based algorithm for discovering clusters in large spatial databases with noise', *KDD*, 1996, **96**, (34), pp. 226–231
[25] Hartigan, J.A., Wong, M.A.: 'Algorithm AS 136: a k-means clustering algorithm', *J. R. Stat. Soc. C, Appl. Stat.*, 1979, **28**, (1), pp. 100–108
[26] Sander, J., Ester, M., Kriegel, H.-P., *et al.*: 'Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications', *Data Min. Knowl. Discov.*, 1998, **2**, (2), pp. 169–194
[27] Jain, A.K.: 'Data clustering: 50 years beyond K-means', *Pattern Recognit. Lett.*, 2010, **31**, (8), pp. 651–666
[28] Troyanskaya, O., Cantor, M., Sherlock, G., *et al.*: 'Missing value estimation methods for DNA microarrays', *Bioinformatics*, 2001, **17**, (6), pp. 520–525
[29] Biernacki, C., Celeux, G., Govaert, G.: 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (7), pp. 719–725
[30] Andrews, D.F., Mallows, C.L.: 'Scale mixtures of normal distributions', *J. R. Stat. Soc. B, Methodol.*, 1974, **36**, (1), pp. 99–102. Available at http://www.jstor.org/stable/2984774
[31] Archambeau, C., Verleysen, M.: 'Robust Bayesian clustering', *Neural Netw.*, 2007, **20**, (1), pp. 129–138
[32] Waterhouse, S., MacKay, D., Robinson, T., *et al.*: 'Bayesian methods for mixtures of experts', *Adv. Neural. Inf. Process. Syst.*, 1996, **8**, pp. 351–357
[33] Seute, H., Enderli, C., Grandin, J.-F., *et al.*: 'Experimental analysis of time deviation on a passive localization system'. 2016 Sensor Signal Processing for Defence (SSPD), Edinburgh, UK, 2016, pp. 1–5
[34] Davies, C., Hollands, P.: 'Automatic processing for ESM', *IEE Proc. F, Commun. Radar Signal Process.*, 1982, **129**, (3), pp. 164–171
[35] Fettweis, G., Löhning, M., Petrovic, D., *et al.*: 'Dirty RF: a new paradigm', *Int. J. Wirel. Inf. Netw.*, 2007, **14**, (2), pp. 133–148
[36] Revillon, G., Mohammad-Djafari, A., Enderli, C.: 'Radar emitters classification and clustering with a scale mixture of normal distributions'. 2018 IEEE Radar Conf., Oklahoma City, OK, USA, May 2018, no. 4455
[37] Jordan, M.I., Jacobs, R.A.: 'Hierarchical mixtures of experts and the EM algorithm', *Neural Comput.*, 1994, **6**, (2), pp. 181–214
[38] Dumitru, M., Li, W., Gac, N., *et al.*: 'Performance comparison of Bayesian iterative algorithms for three classes of sparsity enforcing priors with application in computed tomography'. 2017 IEEE Int. Conf. on Image Processing, Beijing, China, 2017
[39] Tierney, L., Kadane, J.B.: 'Accurate approximations for posterior moments and marginal densities', *J. Am. Stat. Assoc.*, 1986, **81**, (393), pp. 82–86
[40] Akaike, H.: 'Information theory and an extension of the maximum likelihood principle', in '*Selected papers of Hirotugu Akaike*' (Springer, New York, NY, 1998), pp. 199–213
[41] Schwarz, G.: 'Estimating the dimension of a model', *Ann. Stat.*, 1978, **6**, (2), pp. 461–464
[42] Dempster, A.P., Laird, N.M., Rubin, D.B.: 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Stat. Soc. B, Methodol.*, 1977, **39**, pp. 1–38
[43] Opper, M., Saad, D.: '*Advanced mean field methods: theory and practice*' (MIT press, Cambridge, MA, 2001)
[44] Svensén, M., Bishop, C.M.: 'Robust Bayesian mixture modelling', *Neurocomputing*, 2005, **64**, pp. 235–252
[45] McLachlan, G., Peel, D.: '*Finite mixture models*' (John Wiley & Sons, New York, NY, 2004)
[46] García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R.: 'Pattern classification with missing data: a review', *Neural Comput. Appl.*, 2010, **19**, (2), pp. 263–282
[47] Schafer, J.L.: '*Analysis of incomplete multivariate data*' (CRC press, London, UK, 1997)
[48] Hastie, T., Tibshirani, R., Sherlock, G., *et al.*: 'Imputing missing data for gene expression arrays', December 2001, vol. **1**
[49] Hubert, L., Arabie, P.: 'Comparing partitions', *J. Classif.*, 1985, **2**, (1), pp. 193–218
[50] Kaufman, L., Rousseeuw, P.J.: '*Finding groups in data: an introduction to cluster analysis*', vol. **344** (John Wiley & Sons, New York, NY, 2009)