Online EM for mixtures of multiple scale Student distributions

Florence Forbes Hien D. Nguyen

9th November 2021

Abstract

The multiple scale Student distribution is a generalization of the multivariate Student distribution that allows to model multiple tail weights and can more flexibly account for outliers in a multivariate setting. Mixtures of such distributions are useful for clustering tasks where clusters may not be elliptical and have been used in various applications. Such mixtures require more parameters than the standard Student or Gaussian mixtures and the EM algorithm used to estimate the mixture parameters involves more complex numerical optimizations for some of the M-steps. Consequently, when the number of samples to be clustered becomes large, applying EM on the whole data set (Batch EM) may become costly both in terms of time and memory requirements. A natural approach to bypass this issue is to consider an online version of the algorithm, that can incorporate the samples online or in mini batches. Such an online EM algorithm has been proposed and theoretically studied in the literature but most works restrict to distributions in the exponential family. In addition, the practical implementation of the algorithm may not be straightforward. In this paper, we show that the multiple scale Student distribution can be cast into the framework of a tractable online EM and we propose an explicit design of the algorithm for mixtures. The resulting online EM is illustrated on simulated data and a real data experiment involving MRI measurements for the detection of subtle brain anomalies in patients suffering from Parkinson disease.

Keywords: Online EM algorithm, Gaussian scale mixture, Multivariate generalized t-distribution, Outlier detection, Parkinson Disease.

1 Introduction

A popular way to approach clustering tasks is via a parametric finite mixture model. The vast majority of the work on such mixtures has been based on Gaussian mixture models (see *e.g.* Fraley and Raftery [2002]). However, in some applications the tails of Gaussian distributions are shorter than appropriate or parameter estimations are affected by atypical observations (outliers). To address this issue, mixtures of multivariate Student t-distributions have been proposed and used for clustering. In contrast to the Gaussian case, no closed-form solution exists for the t-distribution but tractability is maintained, both in the univariate and multivariate case, via the use of the EM algorithm [McLachlan and Peel, 2000, Bishop and Svensen, 2005, Archambeau and Verleysen, 2007]. The algorithm makes use of the representation of the t-distribution as a so-called *infinite mixture of scaled Gaussians* or *Gaussian scale mixture*, which is of the form:

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\vartheta}) = \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)\; f_W(w;\boldsymbol{\vartheta})\; \mathrm{d}w \tag{1}$$

where $\mathcal{N}_M(\ .\ ;\boldsymbol{\mu},\boldsymbol{\Sigma}/w)$ denotes the M-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/w$ and f_W is the probability distribution of a univariate positive variable W referred to hereafter as the weight variable. When f_W is a Gamma distribution $\mathcal{G}(\nu/2,\nu/2)$ where ν denotes the degrees of freedom (dof), the standard multivariate t-distribution is recovered. We shall also denote the Gamma distribution when the variable is X by $\mathcal{G}(x;\alpha,\beta) = x^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\beta x)\beta^\alpha$ where Γ is the Gamma function. The standard t-distribution belongs to the class of elliptically contoured distributions (see for instance Fang et al. [2002] for a definition of elliptical distributions) and all marginals are t-distributions with the same dof parameter ν and hence the same amount of tailweight. It is not possible to account for very different tail behaviors across dimensions, such as a Gaussian (infinite dof) tail in one dimension and a Cauchy (dof=1) tail in an other dimension [Azzalini and Genton, 2008]. More generally, elliptical distributions are limited by the type of elliptical shapes they allow. Alternative distributions, with a large variety of shapes, exist such as those using copula modelling. Unfortunately copula models become rapidly intractable when more than 2 variables have to be jointly modeled. It is thus important to design models that are both flexible in shapes and tractable in higher dimension. This is the case of the multiple scale t-distribution (MST) introduced by Forbes and Wraith [2014], which goes far beyond the standard t-distribution in terms of possible (not restricted to elliptical) shapes.

The MST distribution [Forbes and Wraith, 2014] comes from a generalization of (1). Denoting respectively by $\boldsymbol{D}$ and $\boldsymbol{A} = diag(A_1, \ldots, A_M)$ the matrix of eigenvectors of $\boldsymbol{\Sigma}$ and the diagonal matrix with the corresponding eigenvalues, we can write $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$. The scaled Gaussian part in (1) is replaced $\mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)$, where $\boldsymbol{\Delta}_{\mathbf{w}} = \text{diag}(w_1^{-1}, \ldots, w_M^{-1})$ is a $M \times M$ diagonal matrix whose diagonal components are the inverse weights $\{w_1^{-1}, \ldots, w_M^{-1}\}$. It follows a generalized multiple scale Gaussian distribution,

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\vartheta}) = \int_0^\infty \int_0^\infty \mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T) f_{W_1}(w_1; \vartheta_1) \ldots f_{W_M}(w_M; \vartheta_M) dw_1 \ldots dw_M . \quad (2)$$

As a particular case, when for $m \in [M]$ ($[M]$ denotes the set of integers from 1 to M), $\vartheta_m = \nu_m$ and $f_{W_m}(w_m; \vartheta_m)$ is set to $\mathcal{G}(w_m; \nu_m/2, \nu_m/2)$, the resulting density denoted by $\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$ with $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_M\}$ is referred to as the multiple scale t-distribution (MST):

$$\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) = \prod_{m=1}^{M} \frac{\Gamma((\nu_m + 1)/2)}{\Gamma(\nu_m/2)(A_m \nu_m \pi)^{1/2}} \left(1 + \frac{[\boldsymbol{D}^T(\mathbf{y} - \boldsymbol{\mu})]_m^2}{A_m \nu_m}\right)^{-(\nu_m+1)/2}, \quad (3)$$

where $[\boldsymbol{D}^T(\mathbf{y} - \boldsymbol{\mu})]_m^2$ denotes the squared mth element of vector $[\boldsymbol{D}^T(\mathbf{y} - \boldsymbol{\mu})]$.

It is clear from (3) that this distribution is not of the elliptical form (see Fang et al. [2002]).

The MST distribution allows in particular different tailweights in each dimension thanks to M possibly different dof parameters. As described in Forbes and Wraith [2014], an EM algorithm can be designed to learn mixtures of MST and has been successfully applied in various practical clustering tasks, *e.g.* Arnaud et al. [2018], Munoz Ramirez et al. [2019], Zheng et al. [2019]. However, the application may be limited to samples of moderate sizes due to the additional time and memory requirements implied by the additional dof and more complex M-steps. To bypass this potential limitation, we propose an online version of the algorithm that allows to incorporate observations one after the other or in subsets of moderate sizes (mini batches).

Online EM algorithms exist and have been studied in previous work [Cappé and Moulines, 2009, Le Corff and Fort, 2013, Nguyen et al., 2020, Nguyen and Forbes, 2021]. In this work we consider the online EM of Cappé and Moulines [2009] for which we have already proposed useful mini-batch variants [Nguyen et al., 2020] and implementations [Nguyen and Forbes, 2021]. Nevertheless, the adaptation of the algorithm to MST mixtures is not straightforward as it requires first to exhibit an exponential family form for

the MST distribution and then to check the tractability of the subsequent optimization and updating steps. ECM theory? to be checked

The paper is outlined as follows. The general description of the online EM algorithm is given is Section 2. The exponential forms of a MST and mixture of MST distributions are given respectively in Section 3 and 4. Section 5 provides illustrations on simulated data and on a data set coming from brain MRI acquisitions for subjects suffering from Parkinson disease.

2 Online EM algorithm

Due to the changing nature of the acquisition and volume of data, online and incremental variants of EM and EM-like algorithms have become increasingly popular. Examples of such algorithms include those described in Cappé and Moulines [2009], Maire et al. [2017], Karimi et al. [2019a,b], Fort et al. [2020b], Kuhn et al. [2020], Nguyen et al. [2020], and Allassonniere and Chevalier [2021], among others. As an archetype of such algorithms, we shall consider the online EM algorithm of Cappé and Moulines [2009].

Suppose that we observe a sequence of n independent and identically distributed (IID) replicates of some random variable $\mathbf{Y} \in \mathbb{Y} \subseteq \mathbb{R}^M$, for $M \in \mathbb{N} = \{1, 2, \ldots\}$ (i.e., $(\mathbf{Y}_i)_{i=1}^n$), where $\mathbf{Y}$ is the visible component of the pair $\mathbf{X}^\top = \left(\mathbf{Y}^\top, \boldsymbol{Z}^\top\right)$, where $\boldsymbol{Z} \in \mathbb{H}$ is a hidden (latent) variable, and $\mathbb{H} \subseteq \mathbb{R}^l$, for $l \in \mathbb{N}$. That is, each $\mathbf{Y}_i$ $(i \in [n] = \{1, \ldots, n\})$ is the visible component of a pair $\mathbf{X}_i^\top = \left(\mathbf{Y}_i^\top, \boldsymbol{Z}_i^\top\right) \in \mathbb{X}$. In the context of online learning, we observe the sequence $(\mathbf{Y}_i)_{i=1}^n$ one observation at a time, in sequential order.

Suppose that $\mathbf{Y}$ arises from some data generating process (DGP) that is characterised by a probability density function (PDF) $f(\mathbf{y}; \boldsymbol{\theta})$, which is parameterised by a parameter vector $\boldsymbol{\theta} \in \mathbb{T} \subseteq \mathbb{R}^p$, for $p \in \mathbb{N}$. Specifically, the sequence of data arises from a DGP that is characterised by an unknown parameter vector $\boldsymbol{\theta}_0 \in \mathbb{T}$. Using the sequence $(\mathbf{Y}_i)_{i=1}^n$, one wishes to sequentially estimate the parameter vector $\boldsymbol{\theta}_0$. The method of Cappé and Moulines [2009] assumes the following restrictions regarding the DGP of $\mathbf{Y}$.

(A1) The complete-data likelihood corresponding to the pair $\mathbf{X}$ is of the exponential family form:

$$f_c(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x}) \exp\left\{ [\boldsymbol{s}(\mathbf{x})]^\top \boldsymbol{\phi}(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}) \right\}, \tag{4}$$

where $h : \mathbb{R}^{M+l} \to [0, \infty)$, $\psi : \mathbb{R}^p \to \mathbb{R}$, $\boldsymbol{s} : \mathbb{R}^{M+l} \to \mathbb{R}^q$, and $\boldsymbol{\phi} : \mathbb{R}^p \to \mathbb{R}^q$, for $q \in \mathbb{N}$.

(A2) The function

$$\bar{\boldsymbol{s}}\left(\mathbf{y};\boldsymbol{\theta}\right)=\mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{s}\left(\mathbf{X}\right)|\mathbf{Y}=\mathbf{y}\right] \tag{5}$$

is well-defined for all $\mathbf{y}\in\mathbb{Y}$ and $\boldsymbol{\theta}\in\mathbb{T}$, where $\mathrm{E}_{\boldsymbol{\theta}}\left[\cdot|\mathbf{Y}=\mathbf{y}\right]$ is the conditional expectation under the assumption that $\mathbf{X}$ arises from the DGP characterised by $\boldsymbol{\theta}$.

(A3) There is a convex subset $\mathbb{S}\subseteq\mathbb{R}^{q}$, which satisfies the properties:

(i) for all $\boldsymbol{s}\in\mathbb{S}$, $\mathbf{y}\in\mathbb{Y}$, and $\boldsymbol{\theta}\in\mathbb{T}$,

$$\left(1-\gamma\right)\boldsymbol{s}+\gamma\bar{\boldsymbol{s}}\left(\mathbf{y};\boldsymbol{\theta}\right)\in\mathbb{S},$$

for any $\gamma\in\left(0,1\right)$, and

(ii) for any $\boldsymbol{s}\in\mathbb{S}$, the function

$$Q\left(\boldsymbol{s};\boldsymbol{\theta}\right)=\boldsymbol{s}^{\top}\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)-\psi\left(\boldsymbol{\theta}\right) \tag{6}$$

has a unique global maximiser on $\mathbb{T}$, which is denote by

$$\bar{\boldsymbol{\theta}}\left(\boldsymbol{s}\right)=\underset{\boldsymbol{\theta}\in\mathbb{T}}{\arg\max}\,Q\left(\boldsymbol{s};\boldsymbol{\theta}\right). \tag{7}$$

Let $\left(\gamma_{i}\right)_{i=1}^{n}$ be a sequence of learning rates in $\left(0,1\right)$ and let $\boldsymbol{\theta}^{(0)}\in\mathbb{T}$ be an initial estimate of $\boldsymbol{\theta}_{0}$. For each $i\in[n]$, the method of Cappé and Moulines [2009] proceeds by computing

$$\boldsymbol{s}^{(i)}=\gamma_{i}\bar{\boldsymbol{s}}\left(\mathbf{Y}_{i};\boldsymbol{\theta}^{(i-1)}\right)+\left(1-\gamma_{i}\right)\boldsymbol{s}^{(i-1)}, \tag{8}$$

and

$$\boldsymbol{\theta}^{(i)}=\bar{\boldsymbol{\theta}}\left(\boldsymbol{s}^{(i)}\right), \tag{9}$$

where $\boldsymbol{s}^{(0)}=\bar{\boldsymbol{s}}\left(\mathbf{Y}_{1};\boldsymbol{\theta}^{(0)}\right)$. As an output, the algorithm produces a sequence of estimators of $\boldsymbol{\theta}_{0}$: $\left(\boldsymbol{\theta}^{(i)}\right)_{i=1}^{n}$.

Under these assumptions and additional ones not detailed here, Cappé and Moulines [2009] proved that the sequences $\left(\boldsymbol{s}^{(i)}\right)_{i=1}^{\infty}$ and $\left(\boldsymbol{\theta}^{(i)}\right)_{i=1}^{\infty}$, computed via the algorithm defined by (8) and (9), permit a convergence result to stationary points of the likelihood (cf. Cappé and Moulines, 2009, Thm. 1 for a more precise statement).

In order to apply the above framework, we first need to identify an exponential family form (4). This case is already very broad. It is not restricted

to exponential family distributions but to any distribution that admits a data augmentation scheme that yields a complete likelihood of the exponential family form. For instance, in Section 4, we show that finite mixtures of exponential family distributions have complete-data likelihoods in the exponential family form. This is also the case for the Student distribution, which is not an exponential family distribution, but admits a data augmented likelihood that permits the use of the online EM algorithm (see Nguyen and Forbes [2021]). For data augmentation schemes that do not yield complete data likelihoods in the exponential family form, the online EM algorithm of Cappé and Moulines [2009] cannot be applied, since it is based on the linearity of the complete-data log-likelihood, with respect to some sufficient statistic that may be latent but for which a conditional expectation exists (with respect to the observed data). Thus, it is not immediately clear whether the online EM and its variants can be extended to handle distributions without complete data exponential family representation, although the recent works of Karimi et al. [2019b] and Fort et al. [2020a] demonstrate how penalization and regularization can be incorporated within the online EM framework. Outside of online EM algorithms, the related online MM (minorisation–maximisation) algorithms of Mairal [2013] and Razaviyayn et al. [2016] can be used to estimate the parameters of generic distributions. However, they can be restrictive in their own ways, such as by requiring certain strong convexity criteria to be met.

The implementation of the algorithm requires then two important quantities the functions $\bar{\boldsymbol{s}}$ in (5) and $\bar{\boldsymbol{\theta}}$ in (7) which are necessary to define the updating equations for sequences $\left(\boldsymbol{s}^{(i)}\right)_{i=1}^{\infty}$ and $\left(\boldsymbol{\theta}^{(i)}\right)_{i=1}^{\infty}$. The next section provides these quantities for a single MST distribution from which the mixture case can be deduced as then explained in Section 4.

3 Exponential family form of a MST distribution

The t-distribution can be seen as the marginal of a product of a scaled Gaussian PDF and a gamma PDF, which can be expressed in an exponential family form. Similarly for the MST distribution, $\boldsymbol{W}^T = (W_1, \ldots, W_M)$ is a natural latent variable so that setting $\mathbf{x}^T = (\mathbf{y}^T, w_1, \ldots, w_M)$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\nu})$ with $\boldsymbol{\nu} = \nu_1, \ldots, \nu_M$, the complete data likelihood writes:

$$f_c(\mathbf{x}; \boldsymbol{\theta}) \;=\; \mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T) \prod_{m=1}^{M} \mathcal{G}\left(w_m; \frac{\nu_m}{2}, \frac{\nu_m}{2}\right) . \qquad (10)$$

The gamma distribution belongs to the exponential family and its PDF can be written as,

$$\mathcal{G}\left(w_m; \frac{\nu_m}{2}, \frac{\nu_m}{2}\right) = h_{W_m}(w_m) \exp\left(s_{W_m}(w_m)\phi_{W_m}(\nu_m) - \psi_{W_m}(\nu_m)\right) ,$$

with $s_{W_m}(w_m) = \log w_m - w_m$, $\phi_{W_m}(\nu_m) = \nu_m/2$, $h_{W_m}(w_m) = 1/w_m$, $\psi_{W_m}(\nu_m) = \log\Gamma(\nu_m/2) - \nu_m/2\log(\nu_m/2)$. It follows that the exponential form of the product $\prod_{m=1}^{M}\mathcal{G}(w_m; \nu_m/2, \nu_m/2)$ is

$$\prod_{m=1}^{M}\mathcal{G}\left(w_m; \frac{\nu_m}{2}, \frac{\nu_m}{2}\right) = h_W(\boldsymbol{w}) \exp\left([\boldsymbol{s}_W(\boldsymbol{w})]^T\boldsymbol{\phi}_W(\boldsymbol{\nu}) - \psi_W(\boldsymbol{\nu})\right),$$

with $[\boldsymbol{s}_W(\boldsymbol{w})]^T = (s_{W_1}(w_1), \ldots, s_{W_M}(w_M))$, $\boldsymbol{\phi}_W(\boldsymbol{\nu})^T = (\nu_1/2, \ldots, \nu_M/2)$ and $\psi_W(\boldsymbol{\nu}) = \sum_{m=1}^{M}(\log\Gamma(\nu_m/2) - \nu_m/2\log(\nu_m/2))$.

For the Gaussian part in (10), it is convenient to use the equivalent expression below, denoting by $A_1, \ldots, A_M$ the diagonal elements of $\boldsymbol{A}$ and $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M$ the columns of matrix $\boldsymbol{D}$,

$$\mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T) = \prod_{m=1}^{M}\mathcal{N}_1([\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m; 0, A_m w_m^{-1}),$$

with

$$\begin{aligned}[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m^2 &= \mathbf{y}^T\boldsymbol{d}_m\boldsymbol{d}_m^T\mathbf{y} + \boldsymbol{\mu}^T\boldsymbol{d}_m\boldsymbol{d}_m^T\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{d}_m\boldsymbol{d}_m^T\mathbf{y} \\ &= vec(\boldsymbol{d}_m\boldsymbol{d}_m^T)^T vec(\mathbf{y}\mathbf{y}^T) + vec(\boldsymbol{d}_m\boldsymbol{d}_m^T)^T vec(\boldsymbol{\mu}\boldsymbol{\mu}^T) - 2\boldsymbol{\mu}^T\boldsymbol{d}_m\boldsymbol{d}_m^T\mathbf{y} .\end{aligned}$$

Notation $vec(\cdot)$ denotes the vectorisation operator, which converts a matrix to a column vector. Some useful properties of thi operator are recalled in Appendix 6.

It follows the exponential form of the complete likelihood (10):

$$f_c(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{y}, \boldsymbol{w}) \exp\left([\boldsymbol{s}(\mathbf{y}, \boldsymbol{w})]^T\boldsymbol{\phi}(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\nu}) - \psi(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\nu})\right)$$

with $\boldsymbol{s}(\cdot,\cdot)$ a function from $\mathbb{R}^{2M}$ to $\mathbb{R}^q$ with $q = (M + M^2 + 2)M$,

$$\boldsymbol{s}(\mathbf{y}, \boldsymbol{w}) = \begin{bmatrix} w_1\mathbf{y} \\ w_1 vec(\mathbf{y}\mathbf{y}^T) \\ w_1 \\ \log w_1 \\ \vdots \\ w_M\mathbf{y} \\ w_M vec(\mathbf{y}\mathbf{y}^T) \\ w_M \\ \log w_M \end{bmatrix} \tag{11}$$

$$\boldsymbol{\phi}(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\nu}) = \begin{bmatrix} \frac{\boldsymbol{d}_1 \boldsymbol{d}_1^T \boldsymbol{\mu}}{A_1} \\ -\frac{vec(\boldsymbol{d}_1 \boldsymbol{d}_1^T)}{2A_1} \\ -\frac{vec(\boldsymbol{d}_1 \boldsymbol{d}_1^T)^T vec(\boldsymbol{\mu}\boldsymbol{\mu}^T)}{2A_1} - \frac{\nu_1}{2} \\ \frac{1+\nu_1}{2} \\ \vdots \\ \frac{\boldsymbol{d}_M \boldsymbol{d}_M^T \boldsymbol{\mu}}{A_M} \\ -\frac{vec(\boldsymbol{d}_M \boldsymbol{d}_M^T)}{2A_M} \\ -\frac{vec(\boldsymbol{d}_M \boldsymbol{d}_M^T)^T vec(\boldsymbol{\mu}\boldsymbol{\mu}^T)}{2A_M} - \frac{\nu_M}{2} \\ \frac{1+\nu_M}{2} \end{bmatrix} \tag{12}$$

$$\psi(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\nu}) = \sum_{m=1}^{M} \left(\frac{\log A_m}{2} + \log \Gamma(\frac{\nu_m}{2}) - \frac{\nu_m}{2} \log(\frac{\nu_m}{2}) \right) .$$

The expression of ψ does not actually depend on $\boldsymbol{D}$ and $\boldsymbol{\mu}$. The exact form of h is not important for the algorithm.

With this exponential family form, we can now define and compute the following quantities. First, $\bar{\boldsymbol{\theta}}(\boldsymbol{s})$ is defined as the unique maximizer of function $Q(\boldsymbol{s}, \boldsymbol{\theta}) = \boldsymbol{s}^T \boldsymbol{\phi}(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta})$ where $\boldsymbol{s}$ is a vector that matches the definition and dimension of $\boldsymbol{\phi}(\boldsymbol{\theta})$ in (12) and can be conveniently written as

$$\boldsymbol{s} = \begin{bmatrix} \boldsymbol{s}_{11} \\ vec(\boldsymbol{S}_{21}) \\ s_{31} \\ s_{41} \\ \vdots \\ \boldsymbol{s}_{1M} \\ vec(\boldsymbol{S}_{2M}) \\ s_{3M} \\ s_{4M} \end{bmatrix} \tag{13}$$

with for each $m \in [M]$, $\boldsymbol{s}_{1m}$ is a M-dimensional vector, $\boldsymbol{S}_{2m}$ is a $M \times M$ matrix, s_{3m} and s_{4m} are scalars.

Parameters are then updated by maximizing Q with respect to $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\nu})$. We can define $\bar{\boldsymbol{\theta}}(\boldsymbol{s})$ as the root of the first-order condition

$$\boldsymbol{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\,\boldsymbol{s} - \frac{\partial \psi}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbf{0}, \tag{14}$$

where $\boldsymbol{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \partial\boldsymbol{\phi}/\partial\boldsymbol{\theta}$ is the Jacobian of $\boldsymbol{\phi}$, with respect to $\boldsymbol{\theta}$, as a function of $\boldsymbol{\theta}$. Computing gradients leads to $\bar{\boldsymbol{\theta}}(\boldsymbol{s}) = (\bar{\boldsymbol{\mu}}(\boldsymbol{s}), \bar{\boldsymbol{A}}(\boldsymbol{s}), \bar{\boldsymbol{D}}(\boldsymbol{s}), \bar{\boldsymbol{\nu}}(\boldsymbol{s}))$ given by the solution in $(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\nu})$ of the following equations. Parameters $(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{D})$ and $\boldsymbol{\nu}$ can be optimized separately.

$$\begin{aligned}
\boldsymbol{\mu} &= \left(\sum_{m=1}^{M} \frac{s_{3m}}{A_m}(\boldsymbol{d}_m\boldsymbol{d}_m^T)\right)^{-1}\left(\sum_{m=1}^{M}\frac{\boldsymbol{d}_m\boldsymbol{d}_m^T}{A_m}\boldsymbol{s}_{1m}\right)\\
&= \left(\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{S}_3\boldsymbol{D}^T\right)^{-1}\left(\sum_{m=1}^{M}\frac{\boldsymbol{d}_m\boldsymbol{d}_m^T}{A_m}\boldsymbol{s}_{1m}\right)\\
&= \boldsymbol{D}\boldsymbol{A}\boldsymbol{S}_3^{-1}\boldsymbol{D}^T\left(\sum_{m=1}^{M}\frac{\boldsymbol{d}_m\boldsymbol{d}_m^T}{A_m}\boldsymbol{s}_{1m}\right)\\
&= \boldsymbol{D}\boldsymbol{S}_3^{-1}\boldsymbol{v}\ ,
\end{aligned} \tag{15}$$

where $\boldsymbol{S}_3 = diag(s_{31}, \dots s_{3M})$ and $\boldsymbol{v}$ is a vector defined as $\boldsymbol{v}^T = (\boldsymbol{d}_1^T\boldsymbol{s}_{11}, \dots, \boldsymbol{d}_M^T\boldsymbol{s}_{1M})$. For matrix $\boldsymbol{A}$, we get for each $m \in [M]$,

$$\begin{aligned}
A_m &= vec(\boldsymbol{S}_{2m})^T vec(\boldsymbol{d}_m\boldsymbol{d}_m^T) + s_{3m} vec(\boldsymbol{d}_m\boldsymbol{d}_m^T)^T vec(\boldsymbol{\mu}\boldsymbol{\mu}^T) - 2\boldsymbol{s}_{1m}^T\boldsymbol{d}_m\boldsymbol{d}_m^T\boldsymbol{\mu}\\
&= \boldsymbol{d}_m^T(\boldsymbol{S}_{2m} + s_{3m}\boldsymbol{\mu}\boldsymbol{\mu}^T - 2\boldsymbol{\mu}\boldsymbol{s}_{1m}^T)\boldsymbol{d}_m\\
&= \boldsymbol{d}_m^T\boldsymbol{S}_{2m}\boldsymbol{d}_m - \frac{(\boldsymbol{d}_m^T\boldsymbol{s}_{1m})^2}{s_{3m}},
\end{aligned} \tag{16}$$

where the last equality is obtained by replacing $\boldsymbol{\mu}$ by its expression in (15). Equations (15) and (16) express the optimal $\boldsymbol{\mu}$ and $\boldsymbol{A}$ as functions of the optimal $\boldsymbol{D}$.

The maximisation in $\boldsymbol{D}$ has to take into account the orthogonality of $\boldsymbol{D}$. Plugin-in the expressions of $\boldsymbol{\mu}$ and $\boldsymbol{A}$ above and omitting parts that depend

on $\boldsymbol{\nu}$ only, it comes,

$$\begin{aligned}
\boldsymbol{D} &= \arg \max_{\boldsymbol{D}, \boldsymbol{D}^T\boldsymbol{D}=Id} \sum_{m=1}^{M} A_m^{-1} \boldsymbol{d}_m^T \left(\boldsymbol{\mu} \boldsymbol{s}_{1m}^T - \frac{1}{2} \boldsymbol{S}_{2m} - \frac{1}{2} s_{3m} \boldsymbol{\mu}\boldsymbol{\mu}^T \right) \boldsymbol{d}_m - \frac{1}{2} \log A_m \\
&= \arg \min_{\boldsymbol{D}, \boldsymbol{D}^T\boldsymbol{D}=Id} M + \sum_{m=1}^{M} \log A_m \\
&= \arg \min_{\boldsymbol{D}, \boldsymbol{D}^T\boldsymbol{D}=Id} \sum_{m=1}^{M} \log \left(\boldsymbol{d}_m^T (\boldsymbol{S}_{2m} - \frac{\boldsymbol{s}_{1m}\boldsymbol{s}_{1m}^T}{s_{3m}}) \boldsymbol{d}_m \right) \qquad (17)
\end{aligned}$$

The second and third equalities result from the plugin of (15) and (16) showing in two steps that this makes the first sum constant.

Each term in the log can be easily minimized individually: $\boldsymbol{d}_m$ is the norm-1 egeinvector associated to the smallest egeinvalue of matrix $\boldsymbol{S}_{2m} - \frac{\boldsymbol{s}_{1m}\boldsymbol{s}_{1m}^T}{s_{3m}}$ provided the matrix is full-rank otherwise the eigenvalue is 0 and there may be a full space of solutions (?). But the problem is that all these $\boldsymbol{d}_m$ vectors are not necessarily orhogonal. So first attempt: use Lagrangian multipliers to express the orthogonality constraints. This seems to work fine and gives a fixed point equation, made of M equations of the form $\boldsymbol{d}_m = F_m(\boldsymbol{D})$, but probably tedious to solve at each iteration (?). More specifically I found for each m, the optimal $\boldsymbol{d}_m$'s should satisfy, $\boldsymbol{d}_m = \boldsymbol{B}_m^{-1} \sum_{l=1}^{M} (\boldsymbol{d}_l^T \boldsymbol{B}_m \boldsymbol{d}_m)\, \boldsymbol{d}_l$ where $\boldsymbol{B}_m = \boldsymbol{S}_{2m} - \frac{\boldsymbol{s}_{1m}\boldsymbol{s}_{1m}^T}{s_{3m}}$ which has to be invertible then.

Other solution? in the end it's a sort of online ECM algorithm?

For the degrees-of-freedom parameters $\boldsymbol{\nu}$, as in the standard t-distribution case, we have to solve the following equation in ν_m for each $m \in [M]$,

$$s_{4m} - s_{3m} - \Psi^{(0)}(\nu_m/2) + \log(\nu_m/2) + 1 \quad = \quad 0\,, \qquad (18)$$

where $\Psi^{(0)}$ is the digamma function defined as $\Psi^{(0)}(\cdot) = \frac{d \log \Gamma(\cdot)}{d\cdot}$.

A second important quantity in the online EM algorithm is $\bar{\boldsymbol{s}}(\mathbf{y}, \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}} [\boldsymbol{s}(\mathbf{X}) | \mathbf{Y} = \mathbf{y}]$. This quantity requires to compute the following expectations for all $m \in [M]$, $\mathrm{E}_{\boldsymbol{\theta}} [W_m | \mathbf{Y} = \mathbf{y}]$ and $\mathrm{E}_{\boldsymbol{\theta}} [\log W_m | \mathbf{Y} = \mathbf{y}]$. More specifically in the update iteration (8) defining the algorithm, these expectations need to be computed for $\mathbf{y} = \mathbf{y}_i$ the observation at iteration i. We therefore denote these expectations respectively by

$$u_{im}^{(i-1)} = \mathrm{E}_{\boldsymbol{\theta}^{(i-1)}} [W_m | \mathbf{Y} = \mathbf{y}_i] = \frac{\alpha_m^{(i-1)}}{\beta_m^{(i-1)}}$$

and

$$\tilde{u}_{im}^{(i-1)} = \mathrm{E}_{\boldsymbol{\theta}^{(i-1)}}\left[\log W_m | \mathbf{Y} = \mathbf{y}_i\right] = \Psi^{(0)}(\alpha_m^{(i-1)}) - \log \beta_m^{(i-1)} \ ,$$

where

$$\alpha_m^{(i-1)} = \frac{\nu_m^{(i-1)} + 1}{2}$$

and

$$\beta_m^{(i-1)} = \frac{\nu_m^{(i-1)}}{2} + \frac{\left(\boldsymbol{d}_m^{(i-1)T}(\mathbf{y}_i - \boldsymbol{\mu}^{(i-1)})\right)^2}{2A_m^{(i-1)}} \ .$$

The update of $\boldsymbol{s}^{(i)}$ in (8) follows then from the update for each $m \in [M]$,

$$\begin{aligned}
\boldsymbol{s}_{1m}^{(i)} &= \gamma_i \ u_{im}^{(i-1)} \ \mathbf{y}_i + (1-\gamma_i) \ \boldsymbol{s}_{1m}^{(i-1)}, \\
\boldsymbol{S}_{2m}^{(i)} &= \gamma_i \ u_{im}^{(i-1)} \ \mathbf{y}_i\mathbf{y}_i^\top + (1-\gamma_i) \ \boldsymbol{S}_{2m}^{(i-1)}, \\
s_{3m}^{(i)} &= \gamma_i \ u_{im}^{(i-1)} + (1-\gamma_i) \ s_{3m}^{(i-1)}, \\
s_{4m}^{(i)} &= \gamma_i \ \widetilde{u}_{im}^{(i-1)} + (1-\gamma_i) \ s_{4m}^{(i-1)},
\end{aligned}$$

starting from

$$\begin{aligned}
\boldsymbol{s}_{1m}^{(1)} &= u_{1m}^{(0)} \ \mathbf{y}_1, \\
\boldsymbol{S}_{2m}^{(1)} &= u_{1m}^{(0)} \ \mathbf{y}_1\mathbf{y}_1^\top, \\
s_{3m}^{(1)} &= u_{1m}^{(0)}, \\
s_{4m}^{(1)} &= \widetilde{u}_{1m}^{(0)}.
\end{aligned}$$

4 Online EM for mixtures of MST distributions

We now consider a mixture of K MST distributions. As described in Nguyen and Forbes [2021], the online EM for mixtures can be easily derived from the online EM for a single component. For a mixture with K components, let $Z \in [K]$ be a categorical latent random variable, such that $\Pr(Z = z) = \pi_z$. Let denote by $\boldsymbol{\theta}_{\mathcal{M}}$ the mixture parameters which contain the pairs $(\pi_z, \boldsymbol{\theta}_z)$ for $z \in [K]$, with π_z the weight of component z and $\boldsymbol{\theta}_z = (\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z, \boldsymbol{\nu}_z)$ the distribution specific parameters for component z. The mixture PDF is,

$$p(\mathbf{y}; \boldsymbol{\theta}_{\mathcal{M}}) = \sum_{z=1}^{K} \pi_z \mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z, \boldsymbol{\nu}_z) \ .$$

Upon defining $\mathbf{X}^\top = \left(\mathbf{Y}^\top, \boldsymbol{W}^T, Z\right)$, we can write the complete-data likelihood in the exponential family form (cf. Nguyen et al., 2020, Prop. 2):

$$f_c\left(\mathbf{x};\boldsymbol{\theta}_{\mathcal{M}}\right) = h\left(\mathbf{y},\boldsymbol{w}\right)\exp\left\{\sum_{\zeta=1}^{K}\mathbf{1}_{\{z=\zeta\}}\left[\log\pi_\zeta + \left[\boldsymbol{s}\left(\mathbf{y},\boldsymbol{w}\right)\right]^\top\boldsymbol{\phi}\left(\boldsymbol{\theta}_\zeta\right) - \psi\left(\boldsymbol{\theta}_\zeta\right)\right]\right\}$$
$$= h_{\mathcal{M}}\left(\mathbf{x}\right)\exp\left\{\left[\boldsymbol{s}_{\mathcal{M}}\left(\mathbf{x}\right)\right]^\top\boldsymbol{\phi}_{\mathcal{M}}\left(\boldsymbol{\theta}_{\mathcal{M}}\right) - \psi_{\mathcal{M}}\left(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}}}\right)\right\},$$

where $h_{\mathcal{M}}\left(\mathbf{x}\right) = h\left(\mathbf{y},\boldsymbol{w}\right)$, $\psi_{\mathcal{M}}\left(\boldsymbol{\theta}\right) = 0$,

$$\boldsymbol{s}_{\mathcal{M}}\left(\mathbf{x}\right) = \begin{bmatrix}\mathbf{1}_{\{z=1\}}\\ \mathbf{1}_{\{z=1\}}\boldsymbol{s}\left(\mathbf{y},\boldsymbol{w}\right)\\ \vdots\\ \mathbf{1}_{\{z=K\}}\\ \mathbf{1}_{\{z=K\}}\boldsymbol{s}\left(\mathbf{y},\boldsymbol{w}\right)\end{bmatrix}, \text{ and } \boldsymbol{\phi}_{\mathcal{M}}\left(\boldsymbol{\theta}\right) = \begin{bmatrix}\log\pi_1 - \psi\left(\boldsymbol{\theta}_1\right)\\ \boldsymbol{\phi}\left(\boldsymbol{\theta}_1\right)\\ \vdots\\ \log\pi_K - \psi\left(\boldsymbol{\theta}_K\right)\\ \boldsymbol{\phi}\left(\boldsymbol{\theta}_K\right)\end{bmatrix}. \quad (19)$$

In the MST case, $\boldsymbol{s}(\mathbf{y},\boldsymbol{w})$ and the $\boldsymbol{\phi}(\boldsymbol{\theta}_z)$'s are respectively given by (11) and (12) and are of dimension $q = (M + M^2 + 2)M$. To match the definition of $\boldsymbol{s}_{\mathcal{M}}\left(\mathbf{x}\right)$ above, we introduce vector $\boldsymbol{s}_{\mathcal{M}}$ with the following notation:

$$\boldsymbol{s}_{\mathcal{M}}^\top = (s_{01}, \boldsymbol{s}_{\mathcal{M}1}^\top, \ldots, s_{0K}, \boldsymbol{s}_{\mathcal{M}K}^\top),$$

and for $z \in [K]$:

$$\boldsymbol{s}_{\mathcal{M}z}^\top = (s_{1z}, \ldots, s_{qz}),$$

where $\boldsymbol{s}_{\mathcal{M}z}$ corresponds to one MST distribution and has the structure given in (13). Vector $\boldsymbol{s}_{\mathcal{M}z} \in \mathbb{S}$, for an appropriate open convex set $\mathbb{S}$, as defined in (A3). Then $\boldsymbol{s}_{\mathcal{M}} \in \mathbb{S}_{\mathcal{M}}$, where $\mathbb{S}_{\mathcal{M}} = \left((0,\infty)\times\mathbb{S}\right)^K$ is an open and convex product space.

As noted by Cappé and Moulines [2009], the finite mixture model demonstrates the importance of the role played by the set $\mathbb{S}$ (and thus $\mathbb{S}_{\mathcal{M}}$) in Assumption (A3). In the sequel, we require that s_{0z} be strictly positive, for each $z \in [K]$. These constraints define $\mathbb{S}_{\mathcal{M}}$, which is open and convex if $\mathbb{S}$ is open and convex. Via (19), the objective function $Q_{\mathcal{M}}$ for the mixture complete-data likelihood, of form (6), can be written as

$$Q_{\mathcal{M}}(\boldsymbol{s}_{\mathcal{M}},\boldsymbol{\theta}_{\mathcal{M}}) = \boldsymbol{s}_{\mathcal{M}}^\top\boldsymbol{\phi}_{\mathcal{M}}(\boldsymbol{\theta}_{\mathcal{M}}) = \sum_{z=1}^{K} s_{0z}(\log\pi_z - \psi(\boldsymbol{\theta}_z)) + \boldsymbol{s}_{\mathcal{M}z}^\top\boldsymbol{\phi}(\boldsymbol{\theta}_z).$$

Whatever the form of the component PDF, the maximisation with respect to π_z yields the mapping

$$\bar{\pi}_z(\boldsymbol{s}_{\mathcal{M}}) \;=\; \frac{s_{0z}}{\sum_{\zeta=1}^{K} s_{0\zeta}}\,. \quad (20)$$

Then, for each $z \in [K]$,

$$\begin{aligned}\frac{\partial Q_{\mathcal{M}}}{\partial \boldsymbol{\theta}_z}\left(\boldsymbol{s}_{\mathcal{M}}, \boldsymbol{\theta}_{\mathcal{M}}\right) &= -s_{0z}\frac{\partial \psi}{\partial \boldsymbol{\theta}_z}\left(\boldsymbol{\theta}_z\right) + \boldsymbol{J}_{\phi}(\boldsymbol{\theta}_z)\boldsymbol{s}_{\mathcal{M}z} \\ &= s_{0z}\left(\boldsymbol{J}_{\phi}(\boldsymbol{\theta}_z)\left[\frac{\boldsymbol{s}_{\mathcal{M}z}}{s_{0z}}\right] - \frac{\partial \psi}{\partial \boldsymbol{\theta}_z}\right) \\ &= s_{0z}\,\frac{\partial Q}{\partial \boldsymbol{\theta}_z}\left(\left[\frac{\boldsymbol{s}_{\mathcal{M}z}}{s_{0z}}\right], \boldsymbol{\theta}_z\right),\end{aligned}$$

where Q is the objective function of form (6) corresponding to a single component PDF. Since $s_{0z} > 0$, for all $z \in [K]$, it follows that the maximisation of $Q_{\mathcal{M}}$ can be conducted by solving

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_z}\left(\left[\frac{\boldsymbol{s}_{\mathcal{M}z}}{s_{0z}}\right], \boldsymbol{\theta}_z\right) = \mathbf{0},$$

with respect to $\boldsymbol{\theta}_z$, for each z. Therefore, it is enough to show that for a single component PDF, there exists a root of the equation above, $\bar{\boldsymbol{\theta}}(\boldsymbol{s})$, with respect to $\boldsymbol{s}$, in order to find a solution for the maximiser of the mixture objective $Q_{\mathcal{M}}$. That is, we can set

$$\bar{\boldsymbol{\theta}}_{\mathcal{M}}(\boldsymbol{s}_{\mathcal{M}}) \;=\; \begin{bmatrix} \bar{\pi}_1(\boldsymbol{s}_{\mathcal{M}}) \\ \bar{\boldsymbol{\theta}}\left(\boldsymbol{s}_{\mathcal{M}1}/s_{01}\right) \\ \vdots \\ \bar{\pi}_K(\boldsymbol{s}_{\mathcal{M}}) \\ \bar{\boldsymbol{\theta}}\left(\boldsymbol{s}_{\mathcal{M}K}/s_{0K}\right) \end{bmatrix}, \tag{21}$$

where $\bar{\pi}_z(\boldsymbol{s}_{\mathcal{M}})$ is given in (20) and $\bar{\boldsymbol{\theta}}$ is the expression found for one single MST component, *i.e.* given by equations (15) to (18).

To complete the online algorithm, we need to specify the quantity $\bar{\boldsymbol{s}}_{\mathcal{M}}(\mathbf{y}; \boldsymbol{\theta}_{\mathcal{M}}) = \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}}\left[\boldsymbol{s}_{\mathcal{M}}\left(\mathbf{Y}, \boldsymbol{W}, Z\right)|\mathbf{Y} = \mathbf{y}\right]$. Using the definition of $\boldsymbol{s}_{\mathcal{M}}(\mathbf{x})$ in (19), we need to compute,

$$r_{ik} = \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}}\left[\mathbf{1}_{\{Z=k\}}|\mathbf{Y} = \mathbf{y}_i\right] = p(Z = k|\mathbf{y}_i) = \frac{\pi_k\, \mathcal{MS}(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\nu}_k)}{p(\mathbf{y}_i; \boldsymbol{\theta}_{\mathcal{M}})},$$

$$\mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}}\left[\mathbf{1}_{\{Z=k\}}\boldsymbol{s}(\mathbf{Y}, \boldsymbol{W})|\mathbf{Y} = \mathbf{y}_i\right] = r_{ik}\,\mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}}\left[\boldsymbol{s}(\mathbf{Y}, \boldsymbol{W})|\mathbf{Y} = \mathbf{y}_i, Z = k\right]\ .$$

The last expectation requires to compute for each $m \in [M]$ and each $k \in [K]$, $u_{imk} = \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}}\left[W_m|\mathbf{Y} = \mathbf{y}_i, Z = k\right]$ and $\tilde{u}_{imk} = \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}}\left[\log W_m|\mathbf{Y} = \mathbf{y}_i, Z = k\right]$ respectively by

$$u_{imk}^{(i-1)} = \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}^{(i-1)}}\left[W_m|\mathbf{Y} = \mathbf{y}_i, Z = k\right] = \frac{\alpha_{mk}^{(i-1)}}{\beta_{mk}^{(i-1)}}$$

and

$$\tilde{u}_{imk}^{(i-1)} = \mathrm{E}_{\boldsymbol{\theta}_{\mathcal{M}}^{(i-1)}}\left[\log W_m | \mathbf{Y} = \mathbf{y}_i, Z = k\right] = \Psi^{(0)}(\alpha_{mk}^{(i-1)}) - \log \beta_{mk}^{(i-1)} \ ,$$

where

$$\alpha_{mk}^{(i-1)} = \frac{\nu_{mk}^{(i-1)} + 1}{2}$$

and

$$\beta_{mk}^{(i-1)} = \frac{\nu_{mk}^{(i-1)}}{2} + \frac{\left(\boldsymbol{d}_{mk}^{(i-1)T}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(i-1)})\right)^2}{2A_{mk}^{(i-1)}} \ .$$

Finally the updated $\boldsymbol{s}_{\mathcal{M}}^{(i)}$ is a vector made of K subvectors, for $k \in [K]$, $(s_{0k}^{(i)}, \boldsymbol{s}_{\mathcal{M}k}^{(i)T})$ where the scalar $s_{0k}^{(i)}$ is updated as

$$s_{0k}^{(i)} = \gamma_i r_{ik}^{(i-1)} + (1 - \gamma_i) s_{0k}^{(i-1)}$$

and each vector $\boldsymbol{s}_{\mathcal{M}k}^{(i)T}$ is made of M subvectors of size $M^2 + M + 2$ updated as, for $m \in [M[$,

$$\begin{aligned}
\boldsymbol{s}_{1mk}^{(i)} &= \gamma_i \ r_{ik}^{(i-1)} \ u_{imk}^{(i-1)} \ \mathbf{y}_i + (1 - \gamma_i) \ \boldsymbol{s}_{1mk}^{(i-1)}, \\
\boldsymbol{S}_{2mk}^{(i)} &= \gamma_i \ r_{ik}^{(i-1)} \ u_{imk}^{(i-1)} \ \mathbf{y}_i \mathbf{y}_i^{\top} + (1 - \gamma_i) \ \boldsymbol{S}_{2mk}^{(i-1)}, \\
s_{3mk}^{(i)} &= \gamma_i \ r_{ik}^{(i-1)} \ u_{imk}^{(i-1)} + (1 - \gamma_i) \ s_{3mk}^{(i-1)}, \\
s_{4mk}^{(i)} &= \gamma_i \ r_{ik}^{(i-1)} \ \widetilde{u}_{imk}^{(i-1)} + (1 - \gamma_i) \ s_{4mk}^{(i-1)},
\end{aligned}$$

5 Application to Parkinson disease

See section 3 of Munoz Ramirez et al. [2019].

6 Conclusion and future work

We have proposed

Appendix

Some useful properties of the *vec* operator **?**

$$trace(\boldsymbol{A}^T \boldsymbol{B}) = vec(\boldsymbol{A})^T vec(\boldsymbol{B})$$

$$\mathbf{x}^T \boldsymbol{A} \mathbf{x} = vec(\boldsymbol{A})^T vec(\mathbf{x}\mathbf{x}^T)$$

$$\frac{dvec(\boldsymbol{\mu}\boldsymbol{\mu}^T)}{d\boldsymbol{\mu}} = 2\boldsymbol{\mu}$$

References

S Allassonniere and J Chevalier. A new class of stochastic EM algorithms. Escaping local maxima and handling inractable sampling. *Computational Statistics and Data Analysis*, 159:107159, 2021.

C. Archambeau and M. Verleysen. Robust Bayesian clustering. *Neural Networks*, 20(1):129–138, 2007.

A. Arnaud, F. Forbes, N. Coquery, N. Collomb, B. Lemasson, and E. Barbier. Fully Automatic Lesion Localization and Characterization: Application to Brain Tumors Using Multiparametric Quantitative MRI Data. *IEEE Transactions on Medical Imaging*, 37(7):1678–1689, 2018.

A. Azzalini and M. G. Genton. Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76(1): 106–129, 2008.

C. M. Bishop and M. Svensen. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005.

O Cappé and E Moulines. On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society B*, 71:593–613, 2009.

H-B. Fang, K-T. Fang, and S. Kotz. The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1):1–16, 2002.

Florence Forbes and Darren Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. *Statistics and Computing*, 24(6):971–984, 2014.

G Fort, E Moulines, and H-T Wai. A stochastic path-integrated differential estimator expectation maximization algorithm. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020a.

G Fort, E Moulines, and H-T Wai. A stochastic path-integrated differential estimator expectation maximization algorithm. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020b.

C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

B Karimi, B Miasojedow, E Moulines, and H-T Wai. Non-asymptotic analysis of biased stochastic approximation scheme. *Proceedings of Machine Learning Research*, 99:1–31, 2019a.

B Karimi, H-T Wai, R Moulines, and M Lavielle. On the global convergence of (fast) incremental expectation maximization methods. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019b.

E Kuhn, C Matias, and T Rebafka. Properties of the stochastic approximation EM algorithm with mini-batch sampling. *Statistics and Computing*, 30:1725–1739, 2020.

S Le Corff and G Fort. Online expectation maximization based algorithms for interence in hidden Markov models. *Electronic Journal of Statistics*, 7:763–792, 2013.

J Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, 2013.

F Maire, E Moulines, and S Lefebvre. Online EM for functional data. *Computational Statistics and Data Analysis*, 111:27–47, 2017.

G.J. McLachlan and D. Peel. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

V. Munoz Ramirez, F. Forbes, J. Arbel, A. Arnaud, and M. Dojat. Quantitative MRI characterization of brain abnormalities in de novo Parkinsonian patients. In *ISBI 2019 - IEEE International Symposium on Biomedical Imaging*, Venice, Italy, 2019.

H. D. Nguyen and F. Forbes. Global implicit function theorems and the online expectation-maximisation algorithm. *Under revision for ANZJ*, 2021.

H D Nguyen, F Forbes, and G J McLachlan. Mini-batch learning of exponential family finite mixture models. *Statistics and Computing*, 30:731–748, 2020.

M Razaviyayn, M Sanjabi, and Z-Q Luo. A stochastic successive minization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *Mathematical Programming*, 157:515–545, 2016.

F Zheng, M. Jalbert, F. Forbes, S. Bonnet, A. Wojtusciszyn, S. Lablanche, and P-Y. Benhamou. Characterization of Daily Glycemic Variability in Subjects with Type 1 Diabetes with mixture of metrics. *Diabetes Technology and Therapeutics*, 2019.