Global implicit function theorems and the online Expectation-Maximisation algorithm

Hien Duy Nguyen, Florence Forbes

Global implicit function theorems and the online expectation–maximisation algorithm

Hien Duy Nguyen[1,2*] and Florence Forbes[3]

University of Queensland and Inria Grenoble Rhône-Alpes

Summary

The expectation–maximisation (EM) algorithm framework is an important tool for statistical computation. Due to the changing nature of data, online and mini-batch variants of EM and EM-like algorithms have become increasingly popular. The consistency of the estimator sequences that are produced by these EM variants often rely on an assumption regarding the continuous differentiability of a parameter update function. In many cases, the parameter update function is not in closed form and may only be defined implicitly, which makes the verification of the continuous differentiability property difficult. We demonstrate how a global implicit function theorem can be used to verify such properties in the cases of finite mixtures of distributions in the exponential family, and more generally, when the component specific distributions admit data augmentation schemes, within the exponential family. We then illustrate the use of such a theorem in the cases of mixtures of beta distributions, gamma distributions, fully-visible Boltzmann machines and Student distributions. Via numerical simulations, we provide empirical evidence towards the consistency of the online EM algorithm parameter estimates in such cases.

Key words: online algorithm; expectation–maximimisation algorithm; global implicit function theorem; Student distribution; mixture models

1. Introduction

Since their introduction by Dempster, Laird & Rubin (1977), the expectation–maximisation (EM) algorithm framework has become an important tool for the conduct of maximum likelihood estimation (MLE) for complex statistical models. Comprehensive accounts of EM algorithms and their variants can be found in the volumes of McLachlan & Krishnan (2008) and Lange (2016).

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 & Moulines (2009), Maire, Moulines

*Author to whom correspondence should be addressed.
[1] School of Mathematics and Physics, University of Queensland, St. Lucia 4067, Queensland Australia
[2] Department of Mathematics and Statistics, La Trobe University, Bundoora 3086, Victoria Australia
Email: h.nguyen7@uq.edu.au
[3] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

& Lefebvre (2017), Karimi et al. (2019a,b), Fort, Moulines & Wai (2020a), Kuhn, Matias & Rebafka (2020), Nguyen, Forbes & McLachlan (2020), and Allassonniere & Chevalier (2021), among others. As an archetype of such algorithms, we shall consider the online EM algorithm of Cappé & Moulines (2009) as a primary example.

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

Suppose that $\boldsymbol{Y}$ arises from some data generating process (DGP) that is characterised by a probability density function (PDF) $f(\boldsymbol{y}; \boldsymbol{\theta})$, 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 $(\boldsymbol{Y}_i)_{i=1}^n$, one wishes to sequentially estimate the parameter vector $\boldsymbol{\theta}_0$. The method of Cappé & Moulines (2009) assumes the following restrictions regarding the DGP of $\boldsymbol{Y}$.

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

$$f_c(\boldsymbol{x}; \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{ [\boldsymbol{s}(\boldsymbol{x})]^\top \boldsymbol{\phi}(\boldsymbol{\theta}) - \boldsymbol{\psi}(\boldsymbol{\theta}) \right\}, \tag{1}$$

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

(A2) The function

$$\bar{\boldsymbol{s}}(\boldsymbol{y}; \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{s}(\boldsymbol{X}) \,|\boldsymbol{Y} = \boldsymbol{y}\right] \tag{2}$$

is well-defined for all $\boldsymbol{y} \in \mathbb{Y}$ and $\boldsymbol{\theta} \in \mathbb{T}$, where $\mathrm{E}_{\boldsymbol{\theta}}[\cdot|\boldsymbol{Y} = \boldsymbol{y}]$ is the conditional expectation under the assumption that $\boldsymbol{X}$ 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 $\boldsymbol{s} \in \mathbb{S}$, $\boldsymbol{y} \in \mathbb{Y}$, and $\boldsymbol{\theta} \in \mathbb{T}$,

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

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

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

$$Q\left(\boldsymbol{s};\boldsymbol{\theta}\right) = \boldsymbol{s}^{\top}\boldsymbol{\phi}\left(\boldsymbol{\theta}\right) - \psi\left(\boldsymbol{\theta}\right) \tag{3}$$

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

$$\bar{\boldsymbol{\theta}}\left(\boldsymbol{s}\right) = \underset{\boldsymbol{\theta}\in\mathbb{T}}{\arg\max}\, Q\left(\boldsymbol{s};\boldsymbol{\theta}\right). \tag{4}$$

Let $(\gamma_i)_{i=1}^n$ be a sequence of learning rates in $(0,1)$ 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 & Moulines (2009) proceeds by computing

$$\boldsymbol{s}^{(i)} = \gamma_i \bar{\boldsymbol{s}}\left(\boldsymbol{Y}_i;\boldsymbol{\theta}^{(i-1)}\right) + \left(1-\gamma_i\right)\boldsymbol{s}^{(i-1)}, \tag{5}$$

and

$$\boldsymbol{\theta}^{(i)} = \bar{\boldsymbol{\theta}}\left(\boldsymbol{s}^{(i)}\right), \tag{6}$$

where $\boldsymbol{s}^{(0)} = \bar{\boldsymbol{s}}\left(\boldsymbol{Y}_1;\boldsymbol{\theta}^{(0)}\right)$. As an output, the algorithm produces a sequence of estimators of $\boldsymbol{\theta}_0$: $\left(\boldsymbol{\theta}^{(i)}\right)_{i=1}^n$.

Suppose that the true DGP of $(\boldsymbol{Y}_i)_{i=1}^n$ is characterised by the probability measure Pr_0, where we write $\mathrm{E}_{\mathrm{Pr}_0}$ to indicate the expectation according to this DGP. We write

$$\boldsymbol{\eta}\left(\boldsymbol{s}\right) = \mathrm{E}_{\mathrm{Pr}_0}\left[\bar{\boldsymbol{s}}\left(\boldsymbol{Y};\bar{\boldsymbol{\theta}}\left(\boldsymbol{s}\right)\right)\right] - \boldsymbol{s},$$

and define the roots of $\boldsymbol{\eta}$ as $\mathbb{O} = \{\boldsymbol{s} \in \mathbb{S} : \boldsymbol{\eta}\left(\boldsymbol{s}\right) = \boldsymbol{0}\}$. Further, let

$$l\left(\boldsymbol{\theta}\right) = \mathrm{E}_{\mathrm{Pr}_0}\left[\log f\left(\boldsymbol{Y};\boldsymbol{\theta}\right)\right]$$

and define the sets

$$\mathbb{U}_{\mathbb{O}} = \left\{l\left(\bar{\boldsymbol{\theta}}\left(\boldsymbol{s}\right)\right) : \boldsymbol{s} \in \mathbb{O}\right\}$$

and

$$\mathbb{M}_{\mathbb{T}} = \left\{\hat{\boldsymbol{\theta}} \in \mathbb{T} : \frac{\partial l}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}\right\}.$$

Denote the distance between the real vector $\boldsymbol{a}$ and the set $\mathbb{B}$ by

$$\mathrm{dist}\left(\boldsymbol{a},\mathbb{B}\right) = \inf_{\boldsymbol{b}\in\mathbb{B}} \left\|\boldsymbol{a}-\boldsymbol{b}\right\|,$$

where $\|\cdot\|$ is the Euclidean norm. Further, denote the complement of set $\mathbb{B}$ by $\mathbb{B}^{\complement}$, and make the following assumptions:

(A4) The set $\mathbb{T}$ is convex and open, and $\boldsymbol{\phi}$ and ψ are both twice continuously differentiable with respect to $\boldsymbol{\theta} \in \mathbb{T}$.

(A5) The function $\bar{\boldsymbol{\theta}}$ is continuously differentiable, with respect to $\boldsymbol{s} \in \mathbb{S}$.

(A6) For some $r > 2$ and compact subset $\mathbb{K} \subset \mathbb{S}$,

$$\sup_{\boldsymbol{s} \in \mathbb{K}} \mathrm{E}_{\mathrm{Pr}_0} \left[\left| \bar{\boldsymbol{s}} \left(\boldsymbol{Y}; \bar{\boldsymbol{\theta}} \left(\boldsymbol{s} \right) \right) \right|^r \right] < \infty.$$

(A7) The sequence $(\gamma_i)_{i=1}^{\infty}$ satisfies the condition that $\gamma_i \in (0, 1)$, for each $i \in \mathbb{N}$,

$$\sum_{i=1}^{\infty} \gamma_i = \infty, \text{ and } \sum_{i=1}^{\infty} \gamma_i^2 < \infty.$$

(A8) The value $\boldsymbol{s}^{(0)}$ is in $\mathbb{S}$, and, with probability 1,

$$\limsup_{i \to \infty} \left\| \boldsymbol{s}^{(i)} \right\| < \infty, \text{ and } \liminf_{i \to \infty} \operatorname{dist} \left(\boldsymbol{s}^{(i)}, \mathbb{S}^{\complement} \right) = 0.$$

(A9) The set $\mathbb{U}_{\mathbb{O}}$ is nowhere dense.

Under Assumptions (A1)–(A9), Cappé & 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 (5) and (6), permit the conclusion that

$$\lim_{i \to \infty} \operatorname{dist} \left(\boldsymbol{s}^{(i)}, \mathbb{O} \right) = 0, \text{ and } \lim_{i \to \infty} \operatorname{dist} \left(\boldsymbol{\theta}^{(i)}, \mathbb{M}_{\mathbb{T}} \right) = 0, \tag{7}$$

with probability 1, when computed using an IID sequence $(\boldsymbol{Y}_i)_{i=1}^{\infty}$, with DGP characterised by measure Pr_0 (cf. Cappé & Moulines 2009, Thm. 1).

The result can be interpreted as a type of consistency for the estimator $\boldsymbol{\theta}^{(n)}$, as $n \to \infty$. Indeed if Pr_0 can be characterised by the PDF $f\left(\boldsymbol{y}; \boldsymbol{\theta}_0\right)$ in the family of PDFs $f\left(\boldsymbol{y}; \boldsymbol{\theta}\right)$, where the family is identifiable in the sense that $f\left(\boldsymbol{y}; \boldsymbol{\theta}\right) = f\left(\boldsymbol{y}; \boldsymbol{\theta}_0\right)$ for all $\boldsymbol{y} \in \mathbb{Y}$, if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then $\boldsymbol{\theta}_0 \in \mathbb{M}_{\mathbb{T}}$ and $\boldsymbol{\theta}_0$ is the only value minimising $l(\cdot)$. If there is no other stationary point, then the result guarantees that $\boldsymbol{\theta}^{(n)} \to \boldsymbol{\theta}_0$, as $n \to \infty$. If the family is not identifiable, in addition to other stationary points, $\mathbb{M}_{\mathbb{T}}$ could contain several minimisers of $l(\cdot)$, in addition to $\boldsymbol{\theta}_0$. This situation is illustrated in Section 4. In any case, a lack of identifiability does not affect the nature of $\mathbb{O}$, due to Proposition 1 in Cappé & Moulines (2009), which states that any two different parameter values $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$ in $\mathbb{M}_{\mathbb{T}}$, with $f\left(\cdot; \boldsymbol{\theta}'\right) = f\left(\cdot; \boldsymbol{\theta}''\right)$, correspond to the same $\boldsymbol{s} \in \mathbb{O}$.

It is evident that when satisfied, Assumptions (A1)–(A9) provide a strong guarantee of correctness for the online EM algorithm and thus it is desirable to validate them in any particular application. In this work, we are particularly interested in the validation of (A5), since it is a key assumption in the algorithm of Cappé & Moulines (2009) and variants of it are also assumed in order to provided theoretical guarantees for many online and mini-batch EM-like algorithms, including those that appear in the works that have been cited above.

In typical applications, the validation of (A5) is conducted by demonstrating that $Q(\boldsymbol{\theta}; \boldsymbol{s})$ can be maximised in closed form, and then showing that the closed form maximiser $\bar{\boldsymbol{\theta}}(\boldsymbol{s})$ is a continuously differentiable function and hence satisfies (A5). This can be seen, for example, in the Poisson finite mixture model and normal finite mixture regression model examples of Cappé & Moulines (2009) and the exponential finite mixture model and multivariate normal finite mixture model examples of Nguyen, Forbes & McLachlan (2020).

However, in some important scenarios, no closed form solution for $\bar{\boldsymbol{\theta}}(\boldsymbol{s})$ exists, such as when $\boldsymbol{Y}$ arises from beta or gamma distributions, when $\boldsymbol{Y}$ has a Boltzmann law (cf. Sundberg 2019, Ch. 6), such as when $\boldsymbol{Y}$ arises from a fully-visible Boltzmann machine (cf. Hyvarinen 2006, and Bagnall et al. 2020), or when data arise from variance mixtures of normal distributions. In such cases, by (4), we can define $\bar{\boldsymbol{\theta}}(\boldsymbol{s})$ as the root of the first-order condition

$$\boldsymbol{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\, \boldsymbol{s} - \frac{\partial \psi}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbf{0}, \tag{8}$$

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 $\boldsymbol{\theta}$.

To verify (A5), we are required to show that there exists a continuously differentiable function $\bar{\boldsymbol{\theta}}(\boldsymbol{s})$ that satisfies (8), in the sense that

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

for all $\boldsymbol{s} \in \mathbb{S}$. Such a result can be established via the use of a global implicit function theorem.

Recently, global implicit function theorems have been used in the theory of indirect inference to establish limit theorems for implicitly defined estimators (see, e.g., Phillips 2012, and Frazier, Oka & Zhu 2019). In this work, we demonstrate how the global implicit function theorem of Arutyunov & Zhukovskiy (2019) can be used to validate (A5) when applying the online EM algorithm of Cappé & Moulines (2009) to compute the MLE when data arise from the beta, gamma, and Student distributions, or from a fully-visible Boltzmann machine. Simulation results are presented to provide empirical evidence towards the exhibition of theoretical guarantee (7). Discussions are also provided regarding the implementation of online EM algorithms to mean, variance, and mean and variance mixtures of normal distributions (see, e.g., Lee & McLachlan 2021 for details regarding such distributions). More

generally, we show that it is straightforward to consider mixtures of the aforementioned distributions. We show that the problem of checking Assumption (A5) for such mixtures reduces to checking (A5) for their component distributions.

The remainder of the paper proceeds as follows. In Section 2, we provide a discussion regarding global implicit function theorems and present the main tool that we will use for the verification of (A5). In Section 3, we consider finite mixtures of distributions with complete likelihoods in the exponential family form. Here, we also illustrate the applicability of the global implicit theorem to the validation of (A5) in the context of the online EM algorithm for the computation of the MLE in the gamma and Student distribution contexts. Additional illustrations for the beta distribution and the fully-visible Boltzmann machine model are provided in Appendices A.2 and A.3. Numerical simulations are presented in Section 4. Conclusions are finally drawn in Section 5. Additional technical results and illustrations are provided in the Appendix.

2. Global implicit function theorems

Implicit function theorems are among the most important analytical results from the perspective of applied mathematics; see, for example, the extensive exposition of Krantz & Parks (2003). The following result from Zhang & Ge (2006) is a typical (local) implicit function theorem for real-valued functions.

Theorem 1. Local implicit function theorem. *Let $\boldsymbol{g} : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}^p$ be a function and $\mathbb{V} \times \mathbb{W} \subset \mathbb{R}^q \times \mathbb{R}^p$ be a neighbourhood of $(\boldsymbol{v}_0, \boldsymbol{w}_0) \in \mathbb{R}^q \times \mathbb{R}^p$, for $p, q \in \mathbb{N}$. Further, let $\boldsymbol{g}$ be continuous on $\mathbb{V} \times \mathbb{W}$ and continuously differentiable with respect to $\boldsymbol{w} \in \mathbb{W}$, for each $\boldsymbol{v} \in \mathbb{V}$. If*

$$\boldsymbol{g}(\boldsymbol{v}_0, \boldsymbol{w}_0) = \boldsymbol{0} \text{ and } \det\left[\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{w}}(\boldsymbol{v}_0, \boldsymbol{w}_0)\right] \neq \boldsymbol{0},$$

then there exists a neighbourhood $\mathbb{V}_0 \subset \mathbb{V}$ of $\boldsymbol{v}_0$ and a unique continuous mapping $\boldsymbol{\chi} : \mathbb{V}_0 \to \mathbb{R}^p$, such that $\boldsymbol{g}(\boldsymbol{v}, \boldsymbol{\chi}(\boldsymbol{v})) = \boldsymbol{0}$ and $\boldsymbol{\chi}(\boldsymbol{v}_0) = \boldsymbol{w}_0$. Moreover, if $\boldsymbol{g}$ is also continuously differentiable, jointly with respect to $(\boldsymbol{v}, \boldsymbol{w}) \in \mathbb{V} \times \mathbb{W}$, then $\boldsymbol{\chi}$ is also continuously differentiable.

We note that Theorem 1 is local in the sense that the existence of the continuously differentiable mapping $\boldsymbol{\chi}$ is only guaranteed within an unknown neighbourhood $\mathbb{V}_0$ of the root $\boldsymbol{v}_0$. This is insufficient for the validation of (A5), since (in context) the existence of a continuously differentiable mapping is required to be guaranteed for all $\mathbb{V}$, regardless of the location of the root $\boldsymbol{v}_0$.

Since the initial works of Sandberg (1981) and Ichiraku (1985), the study of conditions under which global versions of Theorem 1 can be established has become popular in the

mathematics literature. Some state-of-the-art variants of global implicit function theorems for real-valued functions can be found in the works of Zhang & Ge (2006), Galewski & Koniorczyk (2016), Cristea (2017), and Arutyunov & Zhukovskiy (2019), among many others. In this work, we make use of the following version of Arutyunov & Zhukovskiy (2019, Thm. 6), and note that other circumstances may call for different global implicit function theorems.

Theorem 2. Global implicit function theorem. *Let* $\boldsymbol{g} : \mathbb{V} \times \mathbb{R}^p \to \mathbb{R}^r$*, where* $\mathbb{V} \subseteq \mathbb{R}^q$ *and* $p, q, r \in \mathbb{N}$ *and make the following assumptions:*

(B1) *The mapping* $\boldsymbol{g}$ *is continuous.*

(B2) *The mapping* $\boldsymbol{g}(\boldsymbol{v}, \cdot)$ *is twice continuously differentiable with respect to* $\boldsymbol{w} \in \mathbb{R}^p$*, for each* $\boldsymbol{v} \in \mathbb{V}$*.*

(B3) *The mappings* $\partial \boldsymbol{g}/\partial \boldsymbol{w}$ *and* $\partial^2 \boldsymbol{g}/\partial \boldsymbol{w}^2$ *are continuous, jointly with respect to* $(\boldsymbol{v}, \boldsymbol{w}) \in \mathbb{V} \times \mathbb{R}^p$*.*

(B4) *There exists a root* $(\boldsymbol{v}_0, \boldsymbol{w}_0) \in \mathbb{V} \times \mathbb{R}^p$ *of the mapping* $\boldsymbol{g}$*, in the sense that* $\boldsymbol{g}(\boldsymbol{v}_0, \boldsymbol{w}_0) = \mathbf{0}$*.*

(B5) *For all pairs* $(\boldsymbol{v}', \boldsymbol{w}') \in \mathbb{V} \times \mathbb{R}^p$*, the linear operator defined by the Jacobian evaluated at* $(\boldsymbol{v}', \boldsymbol{w}')$*:* $\partial \boldsymbol{g}/\partial \boldsymbol{w}(\boldsymbol{v}', \boldsymbol{w}')$*, is surjective.*

Under Assumptions (B1)–(B5), there exists a continuous mapping $\boldsymbol{\chi} : \mathbb{V} \to \mathbb{R}^p$*, such that* $\boldsymbol{\chi}(\boldsymbol{v}_0) = \boldsymbol{w}_0$ *and* $\boldsymbol{g}(\boldsymbol{v}, \boldsymbol{\chi}(\boldsymbol{v})) = \mathbf{0}$*, for any* $\boldsymbol{v} \in \mathbb{V}$*. Furthermore, if* $\mathbb{V}$ *is an open subset of* $\mathbb{R}^d$ *and the mapping* $\boldsymbol{g}$ *is twice continuously differentiable, jointly with respect to* $(\boldsymbol{v}, \boldsymbol{w}) \in \mathbb{V} \times \mathbb{R}^p$*, then* $\boldsymbol{\chi}$ *can be chosen to be continuously differentiable.*

We note that the stronger conclusions of Theorem 2 requires stronger hypotheses on the function $\boldsymbol{g}$, when compared to Theorem 1. Namely, it is requires $\boldsymbol{g}$ to have continuous second-order derivatives in all arguments in Theorem 2, whereas only the first derivatives are required in Theorem 1. Assumption (B5) may be abstract in nature, but can be replaced by the practical condition that

$$\det \left[\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{w}} (\boldsymbol{v}', \boldsymbol{w}') \right] \neq 0, \tag{9}$$

for all $(\boldsymbol{v}', \boldsymbol{w}') \in \mathbb{V} \times \mathbb{R}^p$, when $p = r$, since a square matrix operator is bijective if and only if it is invertible. When $p > r$, Assumption (B5) can be validated by checking that

$$\operatorname{rank} \left[\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{w}} (\boldsymbol{v}', \boldsymbol{w}') \right] = r,$$

for all $(\boldsymbol{v}', \boldsymbol{w}') \in \mathbb{V} \times \mathbb{R}^p$ (cf. Yang 2015, Def. 2.1). We thus observe that the assumptions of Theorem 2, although strong, can often be relatively simple to check.

3. Applications of the global implicit function theorem

We now proceed to demonstrate how Theorem 2 can be used to validate Assumption (A5) for the application of the online EM algorithm in various finite mixture scenarios of interest.

We recall the notation from Section 1. Suppose that $\boldsymbol{Y}$ is a random variable that has a DGP characterised by a $K \in \mathbb{N}$ component finite mixture model (cf. McLachlan & Peel 2000), where each mixture component has a PDF of the form $f(\boldsymbol{y}; \boldsymbol{\vartheta}_z)$, for $z \in [K]$, and $f(\boldsymbol{y}; \boldsymbol{\vartheta}_z)$ has exponential family form, as defined in (A1). That is, $\boldsymbol{Y}$ has PDF

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{z=1}^{K} \pi_z f(\boldsymbol{y}; \boldsymbol{\vartheta}_z) = \sum_{z=1}^{K} \pi_z h(\boldsymbol{y}) \exp\left\{ [\boldsymbol{s}(\boldsymbol{y})]^\top \boldsymbol{\phi}(\boldsymbol{\vartheta}_z) - \psi(\boldsymbol{\vartheta}_z) \right\}, \tag{10}$$

where $\pi_z > 0$ and $\sum_{z=1}^{K} \pi_z = 1$, and $\boldsymbol{\theta}$ contains the concatenation of elements $(\pi_z, \boldsymbol{\vartheta}_z)$, for $z \in [K]$.

Remark 1. *We note that the component density $f(\boldsymbol{y}; \boldsymbol{\vartheta}_z)$ in (10) can be replaced by a complete-data likelihood $f_c(\boldsymbol{x}'; \boldsymbol{\vartheta}_z)$ of exponential family form, where $\boldsymbol{X}' = (\boldsymbol{Y}, \boldsymbol{U})^\top$ is a further latent variable representation via the augmented random variable $\boldsymbol{U}$, and where $\boldsymbol{Y}$ is the observed random variable, as previously denoted. This is the case when $\boldsymbol{Y}$ arises from a finite mixture of Student distributions. Although the Student distribution is not within the exponential family, its complete-data likelihood, when considered as a Gaussian scale mixture, can be written as a product of a scaled Gaussian PDF and a gamma PDF, which can be expressed in an exponential family form. We illustrate this scenario in Section 3.2.1.*

Remark 2. *Another type of missing (latent) variable occurs when we have to face missing observations. We can consider the case of IID vectors of observations, where some of the elements are missing. Checking assumption (A5) is the same as in the fully observed case but the computation of $\bar{\boldsymbol{s}}$ is different as it requires an account of the missing data imputation. An illustration in the multivariate Gaussian case is given in Appendix A.7.*

Let $Z \in [K]$ be a categorical latent random variable, such that $\Pr(Z = z) = \pi_z$. Then, upon defining $\boldsymbol{X}^\top = \left(\boldsymbol{Y}^\top, Z\right)$, we can write the complete-data likelihood in the exponential

family form (cf. Nguyen, Forbes & McLachlan 2020, Prop. 2):

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

where the subscript m stands for 'mixture', and where $h_m\left(\boldsymbol{x}\right) = h\left(\boldsymbol{y}\right)$, $\psi_m\left(\boldsymbol{\theta}\right) = 0$,

$$\boldsymbol{s}_m\left(\boldsymbol{x}\right) = \begin{bmatrix} \mathbf{1}_{\{z=1\}} \\ \mathbf{1}_{\{z=1\}}\boldsymbol{s}\left(\boldsymbol{y}\right) \\ \vdots \\ \mathbf{1}_{\{z=K\}} \\ \mathbf{1}_{\{z=K\}}\boldsymbol{s}\left(\boldsymbol{y}\right) \end{bmatrix}, \text{ and } \boldsymbol{\phi}_m\left(\boldsymbol{\theta}\right) = \begin{bmatrix} \log\pi_1 - \psi\left(\boldsymbol{\vartheta}_1\right) \\ \boldsymbol{\phi}\left(\boldsymbol{\vartheta}_1\right) \\ \vdots \\ \log\pi_K - \psi\left(\boldsymbol{\vartheta}_K\right) \\ \boldsymbol{\phi}\left(\boldsymbol{\vartheta}_K\right) \end{bmatrix}. \tag{11}$$

Recall that $\boldsymbol{\theta}$ contains the pairs $(\pi_z, \boldsymbol{\vartheta}_z)$ $(z \in [K])$ and $q \in \mathbb{N}$ is the dimension of the component specific sufficient statistics $\boldsymbol{s}(\boldsymbol{y})$. We introduce the following notation, for $z \in [K]$:

$$\boldsymbol{s}_z^\top = (s_{1z}, \ldots, s_{qz}),$$
$$\text{and } \boldsymbol{s}_m^\top = (s_{01}, \boldsymbol{s}_1^\top, \ldots, s_{0K}, \boldsymbol{s}_K^\top),$$

where $\boldsymbol{s}_z \in \mathbb{S}$, for an appropriate open convex set $\mathbb{S}$, as defined in (A3). Then $\boldsymbol{s}_m \in \mathbb{S}_m$, where $\mathbb{S}_m = \left((0,\infty)\times\mathbb{S}\right)^K$ is an open and convex product space.

As noted by Cappé & Moulines (2009), the finite mixture model demonstrates the importance of the role played by the set $\mathbb{S}$ (and thus $\mathbb{S}_m$) in Assumption (A3). In the sequel, we require that s_{0z} be strictly positive, for each $z \in [K]$. These constraints define $\mathbb{S}_m$, which is open and convex if $\mathbb{S}$ is open and convex. Via (11), the objective function Q_m for the mixture complete-data likelihood, of form (3), can be written as

$$Q_m(\boldsymbol{s}_m, \boldsymbol{\theta}) = \boldsymbol{s}_m^\top\boldsymbol{\phi}_m(\boldsymbol{\theta}) = \sum_{z=1}^{K} s_{0z}(\log\pi_z - \psi(\boldsymbol{\vartheta}_z)) + \boldsymbol{s}_z^\top\boldsymbol{\phi}(\boldsymbol{\vartheta}_z).$$

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

$$\bar{\pi}_z(\boldsymbol{s}_m) = \frac{s_{0z}}{\sum_{\zeta=1}^{K} s_{0\zeta}}.$$

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

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

where Q is the objective function of form (3) corresponding to the component PDFs. Since $s_{0z} > 0$, for all $z \in [K]$, it follows that the maximisation of Q_m can be conducted by solving

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

with respect to $\boldsymbol{\vartheta}_z$, for each z. Therefore, it is enough to show that for the component PDFs, there exists a continuously differentiable root of the equation above, $\bar{\boldsymbol{\vartheta}}(\boldsymbol{s})$, with respect to $\boldsymbol{s}$, in order to verify (A5) for the maximiser of the mixture objective Q_m. That is, we can set

$$\bar{\boldsymbol{\theta}}_m(\boldsymbol{s}_m) = \begin{bmatrix} \bar{\pi}_1(\boldsymbol{s}_m) \\ \bar{\boldsymbol{\vartheta}}(\boldsymbol{s}_1/s_{01}) \\ \vdots \\ \bar{\pi}_K(\boldsymbol{s}_m) \\ \bar{\boldsymbol{\vartheta}}(\boldsymbol{s}_K/s_{0K}) \end{bmatrix},$$

which is continuously differentiable if $\bar{\boldsymbol{\vartheta}}$ is continuously differentiable. In the sequel, we illustrate how Theorem 2 can be applied, with $\mathbb{V} = \mathbb{S}$, to establish the existence of continuous and differentiable functions $\bar{\boldsymbol{\vartheta}}$ in various scenarios.

3.1. The gamma distribution

We firstly suppose that $Y \in (0, \infty)$ is characterised by the PDF

$$f(y; \boldsymbol{\theta}) = \varsigma(y; k, \theta) = \frac{1}{\Gamma(k)\theta^k} y^{k-1} \exp\{-y/\theta\},$$

where $\boldsymbol{\theta}^\top = (\theta, k) \in (0, \infty)^2$, which has an exponential family form, with $h(y) = 1$, $\psi(\boldsymbol{\theta}) = \log\Gamma(k) + k\log\theta$, $\boldsymbol{s}(y) = (\log y, y)^\top$, and $\boldsymbol{\phi}(\boldsymbol{\theta}) = (k-1, -1/\theta)^\top$. Here, $\Gamma(\cdot)$ denotes the gamma function. The objective function Q in (A3) can be written as

$$Q(\boldsymbol{s}; \boldsymbol{\theta}) = s_1(k-1) - \frac{s_2}{\theta} - \log\Gamma(k) - k\log\theta,$$

where $\boldsymbol{s}^\top = (s_1, s_2) \in \mathbb{R} \times (0, \infty)$.

Using the first-order condition (8), we can define $\bar{\boldsymbol{\theta}}$ as a solution of the system of equations:

$$\frac{\partial Q}{\partial k} = s_1 - \Psi^{(0)}(k) - \log\theta = 0, \tag{12}$$

$$\frac{\partial Q}{\partial \theta} = \frac{s_2}{\theta^2} - \frac{k}{\theta} = 0, \tag{13}$$

where $\Psi^{(r)}(k) = \mathrm{d}^{r+1} \log\Gamma(k) / \mathrm{d}k^{r+1}$, is the rth-order polygamma function (see, e.g., Olver et al. 2010, Sec. 5.15).

The existence and uniqueness of $\bar{\boldsymbol{\theta}}$ can be proved using Proposition 2 (from Appendix A.1). Firstly note that $\boldsymbol{\phi}(\boldsymbol{\theta}) = (k - 1, -1/\theta)^\top \in \mathbb{P} = (-1, \infty) \times (-\infty, 0)$, which is an open set and hence we have regularity. Then, setting $\boldsymbol{\Phi}^\top = (\Phi_1, \Phi_2)$, we obtain

$$\boldsymbol{\delta}(\boldsymbol{\Phi}) = \begin{bmatrix} \Psi^{(0)}(1 + \Phi_1) + \log(-1/\Phi_2) \\ -(1 + \Phi_1)/\Phi_2 \end{bmatrix}.$$

For any $\boldsymbol{s}^\top = (s_1, s_2)$, we can solve $\boldsymbol{\delta}(\boldsymbol{\Phi}) = \boldsymbol{s}$ with respect to $\boldsymbol{\Phi}$, which yields: $\Phi_2 = -(1 + \Phi_1)/s_2$ and requires the root of $\Psi^{(0)}(1 + \Phi_1) + \log s_2 - \log(1 + \Phi_1) = s_1$, which is solvable for any $\boldsymbol{s}$ satisfying $s_1 - \log s_2 < 0$, since both $\Psi^{(0)}(\cdot)$ and $\log(\cdot)$ are continuous, and $\Psi^{(0)}(a) - \log(a)$ is increasing in $a \in (0, \infty)$ and has limits of $-\infty$ and 0, as $a \to 0$ and $a \to \infty$, respectively, by Guo et al. (2015, Eqns. 1.5 and 1.6). Thus, $\bar{\boldsymbol{\theta}}$ exists and is unique when

$$\boldsymbol{s} \in \mathbb{S} = \{\boldsymbol{s} = (s_1, s_2) \in \mathbb{R} \times (0, \infty) : s_2 > 0, s_1 - \log s_2 < 0\}. \tag{14}$$

Assuming $\boldsymbol{s} \in \mathbb{S}$, we can proceed to solve (13) with respect to θ, to obtain

$$\theta = \frac{s_2}{k}, \tag{15}$$

which substitutes into (12) to yield:

$$s_1 - \Psi^{(0)}(k) - \log s_2 + \log k = 0. \tag{16}$$

Notice that θ, as defined by (15), is continuously differentiable with respect to k, and thus if k is a continuously differentiable function of $\boldsymbol{s}$, then θ is also a continuous differentiable function of $\boldsymbol{s}$. Hence, we are required to show that there exists a continuously differentiable root of (16), with respect to k, as a function of $\boldsymbol{s}$.

We wish to apply Theorem 2 to show that there exists a continuously differentiable solution of (16). Let

$$g\left(\boldsymbol{s}, w\right) = s_1 - \Psi^{(0)}\left(e^w\right) - \log s_2 + w, \tag{17}$$

where $k = e^w$. We reparameterise with respect to w, since Theorem 2 requires the parameter to be defined over the entire domain $\mathbb{R}$. Notice that (B1)–(B3) are easily satisfied by considering existence and continuity of $\Psi^{(r)}$ over $(0, \infty)$, for all $r \geq 0$. Assumption (B4) is satisfied when $\boldsymbol{s} \in \mathbb{S}$, since it is satisfied if $\bar{\boldsymbol{\theta}}$ exists. Next, to assess (B5), we require the derivative:

$$\frac{\partial g}{\partial w} = 1 - e^w \Psi^{(1)}\left(e^w\right) = 1 - k\Psi^{(1)}\left(k\right). \tag{18}$$

By the main result of Ronning (1986), we have the fact that $-k\Psi^{(1)}\left(k\right)$ is negative and strictly increasing for all $k > 0$. Using an asymptotic expansion, it can be shown that $-k\Psi^{(1)}\left(k\right) \to -1$, as $k \to \infty$ (see the proof of Batir 2005, Lem. 1.2). Thus, (18) is negative for all w, implying that (B5) is validated.

Finally, we establish the existence of a continuously differentiable function $\chi\left(\boldsymbol{s}\right)$, such that $g\left(\boldsymbol{s}, \chi\left(\boldsymbol{s}\right)\right) = 0$ by noting that g is twice continuously differentiable jointly in $(\boldsymbol{s}, w)$. We thus validate (A5) in this scenario by setting

$$\bar{\boldsymbol{\theta}}\left(\boldsymbol{s}\right) = \begin{bmatrix} s_2 / \exp\left\{\chi\left(\boldsymbol{s}\right)\right\} \\ \exp\left\{\chi\left(\boldsymbol{s}\right)\right\} \end{bmatrix},$$

where $\chi\left(\boldsymbol{s}\right)$ is a continuously differentiable root of (17), as guaranteed by Theorem 2.

3.2. Variance mixtures of normal distributions

Variance, or scale mixtures of normal distributions refer to the family of distributions with PDFs that are generated by scaling the covariance matrix of a Gaussian PDF by a positive scalar random variable U. A recent review of such distributions can be found in Lee & McLachlan (2021). Although such distributions are not necessarily in the exponential family, we show that they can be handled within the online EM setting presented in this paper.

Indeed, if U admits an exponential family form, a variance mixture of normal distributions admits a hierarchical representation whose joint distribution, after data augmentation, belongs to the exponential family. We present the general form in this section and illustrate its use by deriving an online EM algorithm for the Student distribution.

Let $f_u(u;\boldsymbol{\theta}_u)$ denote the PDF of U, depending on some parameters $\boldsymbol{\theta}_u$, and admitting an exponential family representation

$$f_u(u;\theta_u) = \; h_u\left(u\right)\exp\left\{\left[\boldsymbol{s}_u\left(u\right)\right]^\top \boldsymbol{\phi}_u\left(\boldsymbol{\theta}_u\right) - \psi_u\left(\boldsymbol{\theta}_u\right)\right\}.$$

If $\boldsymbol{Y}$ is a characterized by a variance mixture of a normal distributions, then with $\boldsymbol{x}^\top = (\boldsymbol{y}^\top, u)$ and $\boldsymbol{\theta}^\top = (\boldsymbol{\mu}^\top, \text{vec}(\boldsymbol{\Sigma})^\top, \boldsymbol{\theta}_u^\top)$, we can write $f_c(\boldsymbol{x};\boldsymbol{\theta})$ as the product of a scaled Gaussian PDF and f_u:

$$f_c(\boldsymbol{x};\boldsymbol{\theta}) = \varphi(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/u)\; f_u(u;\boldsymbol{\theta}_u),$$

where $\varphi(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/u)$ is the PDF of a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}/u$. Here, vec$(\cdot)$ denotes the vectorisation operator, which converts matrices to column vectors.

Using the exponential family forms of both PDFs (see Nguyen, Forbes & McLachlan 2020 for the Gaussian representation), it follows that

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

where $h\left(\boldsymbol{x}\right) = (2\pi/u)^{-d/2} h_u(u)$, $\psi\left(\boldsymbol{\theta}\right) = \log\det\left[\boldsymbol{\Sigma}\right]/2 + \psi_u(\boldsymbol{\theta}_u)$,

$$\boldsymbol{s}\left(\boldsymbol{x}\right) = \begin{bmatrix} u\boldsymbol{y} \\ u\text{vec}(\boldsymbol{y}\boldsymbol{y}^\top) \\ u \\ \boldsymbol{s}_u(u) \end{bmatrix} \text{ and } \boldsymbol{\phi}\left(\boldsymbol{\theta}\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}) \\ -\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\phi}_u(\boldsymbol{\theta}_u) \end{bmatrix}. \tag{19}$$

Depending on the statistics defining $\boldsymbol{s}_u(u)$, the representation above can be made more compact; see, for example, the Student distribution case, below.

Consider the objective function $Q(\boldsymbol{s};\boldsymbol{\theta})$, as per (A3), with $\boldsymbol{s}^\top = (\boldsymbol{s}_1^\top, \text{vec}(\boldsymbol{S}_2)^\top, s_3, \boldsymbol{s}_4^\top)$, where $\boldsymbol{s}_1$ and $\boldsymbol{s}_4$ are real vectors, $\boldsymbol{S}_2$ is a matrix (all of appropriate dimensions) and s_3 is a strictly positive scalar. An interesting property is that whatever the mixing PDF f_u, when maximising Q, closed-form expressions are available for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\bar{\boldsymbol{\mu}}(\boldsymbol{s}) = \frac{\boldsymbol{s}_1}{s_3}, \tag{20}$$

$$\bar{\boldsymbol{\Sigma}}(\boldsymbol{s}) = \boldsymbol{S}_2 - \frac{\boldsymbol{s}_1\boldsymbol{s}_1^\top}{s_3}\,. \tag{21}$$

The rest of the expression of $\bar{\boldsymbol{\theta}}_u(\boldsymbol{s})$ depends on the specific choice of f_u, as illustrated in the sequel.

Remark 3. *Similarly, others families of distributions can be generated by considering mean mixtures, and mean and variance mixtures of normal distributions. If the mixing distribution belongs to the exponential family, the corresponding complete-data likelihood also belongs to the exponential family and can be handled in a similar manner as above. These exponential family forms are provided in Appendices A.5 and A.6. Examples of such distributions are listed in Lee & McLachlan (2021) but are not discussed further in this work.*

3.2.1. The Student distribution

In contrast to the gamma distribution example, the case of the Student distribution requires the introduction of an additional positive scalar latent variable U. The Student distribution is a variance mixture of normal distributions, where U follows a gamma distribution with parameters, in the previous notation of Section 3.1, $k = \nu/2$ and $\theta = 2/\nu$, where ν is commonly referred to as the degree-of-freedom parameter (dof).

Remark 4. *When the two parameters of the gamma distribution are not linked via the joint parameter ν we obtain a slightly more general form of the Student distribution, which is often referred to as the Pearson type VII or generalised Student distribution. Although this later case may appear more general, the Pearson type VII distribution suffers from an identifiability issue that requires a constraint be placed upon the parameters values, which effectively makes it equivalent in practice to the usual Student distribution. See Fang, Kotz & Ng (1990, Sec. 3.3) for a detailed account regarding the Pearson type VII distribution.*

Maximum likelihood estimation of a Student distribution is usually performed via an EM algorithm. As noted in the previous section, the Student distribution does not belong to the exponential family, but the complete-data likelihood after data augmentation by U, does have exponential family form. Indeed $f_c(\boldsymbol{x};\boldsymbol{\theta})$ is the product of a scaled Gaussian and a gamma PDF, which both belong to the exponential family. More specifically, with $\boldsymbol{x}^\top = (\boldsymbol{y}^\top, u)$ and $\boldsymbol{\theta}^\top = (\boldsymbol{\mu}^\top, \text{vec}(\boldsymbol{\Sigma})^\top, \nu)$:

$$f_c(\boldsymbol{x};\boldsymbol{\theta}) = \varphi(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/u)\, \varsigma\left(u; \frac{\nu}{2}, \frac{2}{\nu}\right).$$

It follows from the more general case (19), that

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

where $h(\boldsymbol{x}) = (2\pi/u)^{-d/2}$, $\psi(\boldsymbol{\theta}) = \log\det[\boldsymbol{\Sigma}]/2 + \log\Gamma(\nu/2) - (\nu/2)\log(\nu/2)$,

$$\boldsymbol{s}(\boldsymbol{x}) = \begin{bmatrix} u\boldsymbol{y} \\ u\text{vec}(\boldsymbol{y}\boldsymbol{y}^\top) \\ u \\ \log u \end{bmatrix}, \text{ and } \boldsymbol{\phi}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}) \\ -\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{\nu}{2} \\ \frac{\nu}{2} - 1 \end{bmatrix}. \tag{22}$$

The closed-form expressions for $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$ are given in (20) and (21), while for the dof parameter, we obtain similar equations as in Section 3.1, which leads to defining $\bar{\nu}(\boldsymbol{s})$ as the solution, with respect to ν, of

$$s_4 - \Psi^{(0)}\left(\frac{\nu}{2}\right) - s_3 + 1 + \log\frac{\nu}{2} = 0.$$

With the necessary restrictions on $\mathbb{S}$ (i.e., that $s_3 > 0$ and that $\boldsymbol{S}_2$ be symmetric positive definite), the same arguments as in Section 3.1 apply and provide verification of (A5). A complete derivation of the online EM algorithm is detailed in Appendix A.4 and a numerical illustration is provided in the next section.

4. Numerical Simulations

We now present empirical evidence towards the exhibition of the consistency conclusion (7) of Cappé & Moulines (2009, Thm. 1). The online EM algorithm is illustrated on the scenarios described in Section 3; that is, for gamma and Student mixtures. For the Student distribution example, the estimation of a single such distribution already requires a latent variable representation, due to the scale mixture form, and the online EM algorithm for this case is provided in Appendix A.4. Both the classes of gamma and Student mixtures have been shown to be identifiable; see for instance Teicher (1963, Prop. 2) regarding finite gamma mixtures and Holzmann, Munk & Gneiting (2006, Example 1) regarding finite Student mixtures. If observations are generated from one of these mixtures, checking for evidence that (7) is true is equivalent to checking that the algorithm generates a sequence of parameters that converges to the parameter values used for the simulation. In contrast, beta mixtures are not identifiable, as proved in Ahmad & Al-Hussaini (1982). Boltzmann machine mixtures are also not identifiable. In these cases, the algorithm may generate parameter sequences that converge to values different from the one used for simulation. However, assumptions implying (7) may still be satisfied and the convergence of the online EM algorithm in these cases can be demonstrated (see Appendix A.8).

The numerical simulations are all conducted in the R programming environment (R Core Team 2020) and simulation scripts are made available at `https://github.com/`

hiendn/onlineEM. In each of the cases, we follow Nguyen, Forbes & McLachlan (2020) in using the learning rates $(\gamma_i)_{i=1}^{\infty}$, defined by $\gamma_i = \left(1 - 10^{-10}\right) \times i^{-6/10}$, which satisfies (A7). This choice of γ_i satisfies the recommendation by Cappé & Moulines (2009) to set $\gamma_i = \gamma_0 i^{-(0.5+\epsilon)}$, with $\epsilon \in (-0.5, 0.5)$ and $\gamma_0 \in (0, 1)$. Our choice agrees with the previous reports of Cappe (2009) and Kuhn, Matias & Rebafka (2020), who demonstrated good performance using the same choice of $\epsilon = 0.1$. Furthermore, Le Corff & Fort (2013) and Allassonniere & Chevalier (2021) showed good good numerical results using $\epsilon = 0.03$ and $\epsilon = 0.15$, respectively. The online EM algorithm is then run for $n = 100000$ to $n = 500000$ iterations (i.e., using a sequence of observations $(\boldsymbol{Y}_i)_{i=1}^{n}$, where $n = 100000$ or $n = 500000$. Where required, the mappings $\bar{\boldsymbol{\theta}}$ are numerically computed using the optim function, which adequately solves the respective optimisation problems, as defined by (4). Random gamma observations are generated using the rgamma function. Random Student observations are generated hierarchically using the rnorm and rgamma functions.

For the gamma mixture distribution scenario, we generate data from a mixture of $K = 3$ gamma distributions using the values $\theta_{0z} = 0.5, 0.05, 0.1,\ k_{0z} = 9, 20, 1,$ and $\pi_{0z} = 0.3, 0.2, 0.5$, for each of the 3 components $z = 1, 2, 3$, respectively. The algorithm is initialised with $\theta_z^{(0)} = 1, 1, 1,\ k_z^{(0)} = 5, 9, 2,$ and $\pi_z^{(0)} = 1/3, 1/3, 1/3,$ for each $z = 1, 2, 3$. For the Student mixture case, we restrict our attention to the $d = 1$ case where $\boldsymbol{\Sigma}$ is a scalar, which we denote by $\sigma^2 > 0$. We generate data from a mixture of $K = 3$ Student distributions using $\mu_{0z} = 0, 3, 6,\ \sigma_{0z}^2 = 1, 1, 1,$ and $\nu_{0z} = 3, 2, 1,$ for $z = 1, 2, 3$. The corresponding mixture weights are taken as $\pi_{0z} = 0.3, 0.5, 0.2,$ for each respective component. The algorithm is initialised with values set to $\mu_z^{(0)} = 1, 4, 7,\ \sigma_z^{2(0)} = 2, 2, 2,$ $\nu_z^{(0)} = 4, 4, 4,\ \pi_z^{(0)} = 1/3, 1/3, 1/3,$ for each component $z = 1, 2, 3$.

Example sequences of online EM parameter estimates $\left(\boldsymbol{\theta}^{(i)}\right)_{i=1}^{n}$ for the gamma and Student simulations are presented in Figures 1 and 2, respectively. As suggested by Cappé & Moulines (2009), the parameter vector is not updated for the first 500 iterations. That is, for $i \leq 500$, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(0)}$, and for $i > 500$, $\boldsymbol{\theta}^{(i)} = \bar{\boldsymbol{\theta}}\left(\boldsymbol{s}^{(i)}\right)$. This is to ensure that the initial elements of the sufficient statistic sequence is stable. Other than ensuring that $\boldsymbol{s}^{(0)} \in \mathbb{S}$ in each case, we did not find it necessary to mitigate against any tendencies towards violations of Assumption (A8). We note that if such violations are problematic, then one can employ a truncation of the sequence $\left(\boldsymbol{s}^{(i)}\right)_{i=0}^{n}$, as suggested in Cappé & Moulines (2009) and considered in Nguyen, Forbes & McLachlan (2020).

From the two figures, we notice that the sequences each approach and fluctuate around the respective generative parameter values $\boldsymbol{\theta}_0$. This provides empirical evidence towards the correctness of conclusion (7) of Cappé & Moulines (2009, Thm. 1), in the cases considered in Section 3. In each of the figures, we also observe the decrease in volatility as the iterations increase. This may be explained by the asymptotic normality of the sequences (cf. Cappé

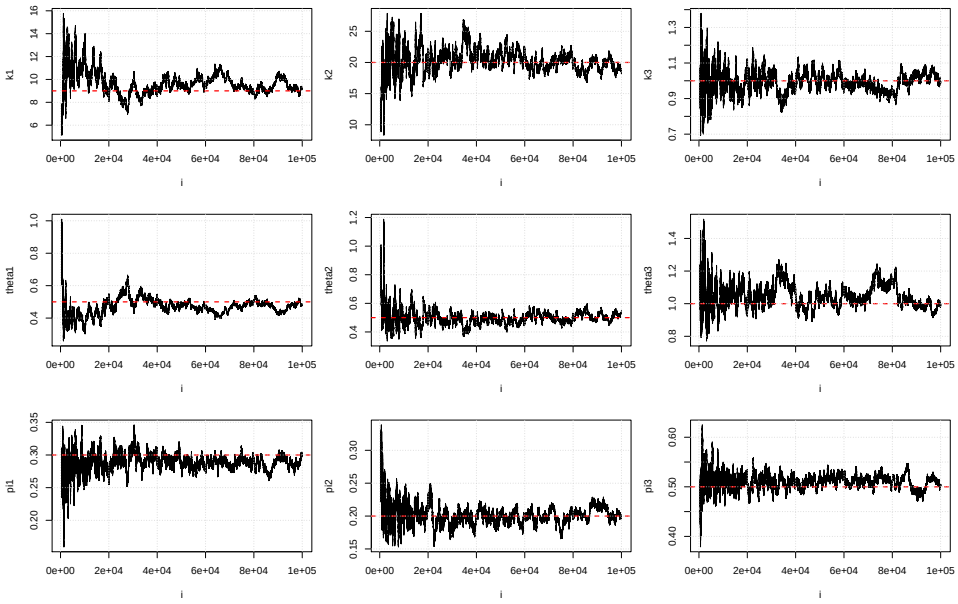

Figure 1. Online EM algorithm estimator sequence $\boldsymbol{\theta}_z^{(i)} = \left(\theta_z^{(i)}, k_z^{(i)}, \pi_z^{(i)}\right)^\top$ ($z \in [3]$), for a mixture of $K = 3$ gamma distributions. The dashed lines indicates the generative parameter values of the DGP. Components are grouped in columns.

& Moulines 2009, Thm. 2), which is generally true under the assumptions of Cappé & Moulines (2009, Thm. 1). For the Student mixture, we consider $n = 500000$ and observed that convergence for the dof parameters may be slower, especially when the dof is larger. This may be due to a flatter likelihood surface for larger values of dof.

5. Conclusion

Assumptions regarding the continuous differentiability of mappings are common for the establishment of consistency results for online and mini-batch EM and EM-like algorithms. As an archetype of such algorithms, we studied the online EM algorithm of Cappé & Moulines (2009), which requires the verification of Assumption (A5) in order for consistency to be establish. We demonstrated that (A5) can be verified in the interesting scenarios when data arises from mixtures of beta distributions, gamma distributions, fully-visible Boltzmann machines and Student distributions, using a global implicit function theorem. Via numerical simulations, we also provide empirical evidence of the convergence of the online EM algorithm in the aforementioned scenarios.

Furthermore, our technique can be used to verify (A5) for other exponential family distributions of interest, that do not have closed form estimators, such as the inverse gamma

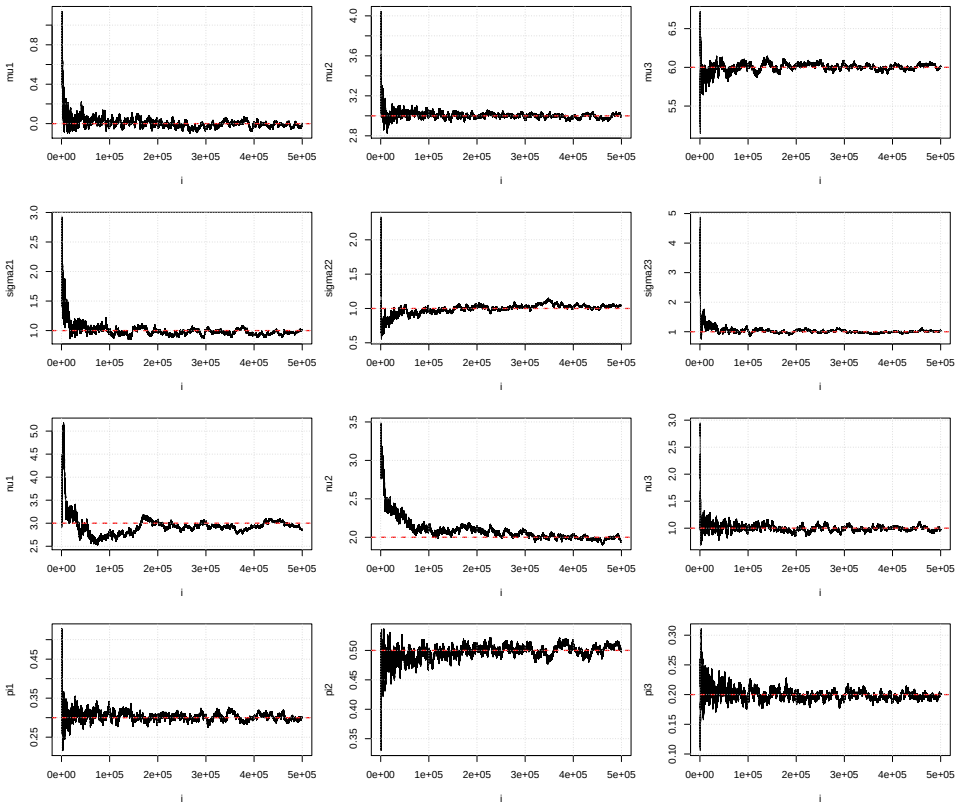

Figure 2. Online EM algorithm sequence of estimator $\boldsymbol{\theta}_z^{(i)} = \left(\mu_z^{(i)}, \sigma_z^{2(i)}, \nu_z^{(i)}, \pi_z^{(i)}\right)^\top$ $(z \in [3])$, for a mixture of $K = 3$ Student distributions. The dashed lines indicates the generative parameter values of the DGP. Components are grouped in columns.

and Wishart distributions, which are widely used in practice. Other models for which our method is applicable include the wide variety of variance, and mean and variance mixtures of normal distributions. We have exclusively studied the verification of assumptions of an online EM algorithm in the IID setting. An interesting question arises as to whether our results apply to online EM algorithms for hidden Markov models (HMMs). Online parameter estimation of HMMs is a challenging task due to the non-trivial dependence between the observations. Recent results in this direction appear in Le Corff & Fort (2013). In this paper, Assumption (A1)(c) is equivalent to our Assumption (A5), where (A1)(c) assumes that a parameter map $\bar{\boldsymbol{\theta}}$ is continuous for convergence of the algorithm. Additionally, to study the rate of convergence of their algorithm, Assumption (A8)(a) is made, which assumes that $\bar{\boldsymbol{\theta}}$ is twice continuously differentiable. Our Theorem 2 can be directly applied to check (A1)(c) but cannot be used to show (A8)(a).

We also note that when the complete-data likelihood or objective function cannot be represented in exponential family form, other online algorithms may be required. The recent works of Karimi et al. (2019b) and Fort, Moulines & Wai (2020b) 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, Sanjabi & Luo (2016) can be used to estimate the parameters of generic distributions. However, these MM algorithms require their own restrictive assumptions, such as the strong convexity of the objective function and related expressions. We defer the exploration of applications of the global implicit function theorem in these settings to future work.

A. Appendix

A.1. Properties of the objective function and maximiser from Assumption (A3)

An expression of form (1) is said to be regular if $\boldsymbol{\phi} : \mathbb{T} \to \mathbb{P}$, where $\mathbb{P}$ is an open subset of $\mathbb{R}^q$. Let $\boldsymbol{\phi}^{-1} : \mathbb{P} \to \mathbb{T}$ denote the inverse function of $\boldsymbol{\phi}$. Call $\mathbb{D} \subseteq \mathbb{R}^q$ the closed convex support of the exponential family form (1) and define it as the smallest closed and convex set such that

$$\inf_{\boldsymbol{\theta} \in \mathbb{T}} \Pr_{\boldsymbol{\theta}} \left(\{ \boldsymbol{x} \in \mathbb{X} : \boldsymbol{s}(\boldsymbol{x}) \in \mathbb{D} \} \right) = 1,$$

where $\Pr_{\boldsymbol{\theta}}$ is the probability measure of $\mathbb{X}$, under the assumption that $\mathbb{X}$ arises from the DGP characterised by $\boldsymbol{\theta}$. Further, let $\operatorname{int} \mathbb{D}$ be the interior of $\mathbb{D}$. The following pair of results are taken from from (Sundberg 2019, Prop. 3.10) and combine (Sundberg 2019, Props. 3.11 and 3.12), respectively (cf. Johansen 1979, Ch. 3, and Barndorff-Neilsen 2014, Ch. 9). The first result provides conditions under which the objective Q is strictly concave, and the second provides conditions under which the maximiser (4) exists and is unique.

Proposition 1. *If* (1) *is regular and* $\boldsymbol{\phi}$ *is bijective, then*

$$Q(\boldsymbol{s}; \boldsymbol{\Phi}) = \boldsymbol{s}^\top \boldsymbol{\Phi} - \psi\left(\boldsymbol{\phi}^{-1}(\boldsymbol{\Phi})\right)$$

is a smooth and strictly concave function of $\boldsymbol{\Phi} \in \mathbb{P}$.

Proposition 2. *If* (1) *is regular then*

$$\boldsymbol{\delta}(\boldsymbol{\Phi}) = \frac{\partial}{\partial \boldsymbol{\Phi}} \psi\left(\boldsymbol{\phi}^{-1}(\boldsymbol{\Phi})\right)$$

is a one-to-one function of $\boldsymbol{\Phi} \in \mathbb{P}$*, where* $\boldsymbol{\delta}(\mathbb{P}) = \{\boldsymbol{\delta}(\boldsymbol{\Phi}) \in \mathbb{R}^q : \boldsymbol{\Phi} \in \mathbb{P}\}$ *is open, and if* $\boldsymbol{\phi}$ *is bijective, then* (4) *exists and is unique if and only if* $\boldsymbol{s} \in \boldsymbol{\delta}(\mathbb{P})$*. Furthermore, we can write* (4) *as the unique root of* $\boldsymbol{s} = \boldsymbol{\delta}(\boldsymbol{\phi}(\boldsymbol{\theta}))$*, and* $\boldsymbol{s} \in \boldsymbol{\delta}(\mathbb{P}) = \operatorname{int} \mathbb{D}$*.*

A.2. The beta distribution

We now consider a beta distributed random variable $Y \in (0,1)$, characterised by the PDF

$$f(y;\boldsymbol{\theta}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

where $\boldsymbol{\theta}^\top = (\alpha,\beta) \in (0,\infty)^2$, which has an exponential family form with $h(y) = y^{-1}(1-y)^{-1}$, $\psi(\boldsymbol{\theta}) = \log\Gamma(\alpha) + \log\Gamma(\beta) - \log\Gamma(\alpha+\beta)$, $\boldsymbol{s}(y) = (\log y, \log(1-y))^\top$, and $\boldsymbol{\phi}(\boldsymbol{\theta}) = (\alpha,\beta)^\top$. The objective function Q in (A3) can be written as

$$Q(\boldsymbol{s};\boldsymbol{\theta}) = s_1\alpha + s_2\beta - \log\Gamma(\alpha) - \log\Gamma(\beta) + \log\Gamma(\alpha+\beta),$$

where $\boldsymbol{s} \in \mathbb{R}^2$.

As in Section 3.1, we can specify conditions for the existence of $\bar{\boldsymbol{\theta}}$ using Proposition 2. Here, there are no problems with regularity, and we can write

$$\boldsymbol{\delta}(\boldsymbol{\phi}(\boldsymbol{\theta})) = \boldsymbol{\delta}(\boldsymbol{\theta}) = \begin{bmatrix} \Psi^{(0)}(\alpha) - \Psi^{(0)}(\alpha+\beta) \\ \Psi^{(0)}(\beta) - \Psi^{(0)}(\alpha+\beta) \end{bmatrix}.$$

Proposition 2 then states that $\bar{\boldsymbol{\theta}}$ exists and is unique when $\boldsymbol{s} \in \mathbb{S} = \boldsymbol{\delta}(\mathbb{P})$, where $\mathbb{P} = (0,\infty)^2$. We may then use the fact that $\boldsymbol{\delta}(\mathbb{P}) = \operatorname{int}\mathbb{D}$ to write

$$\mathbb{S} = \operatorname{int}\mathbb{D} = \left\{\boldsymbol{s} = (s_1,s_2) \in \mathbb{R}^2 : s_1 < 0, s_2 < \log(1-\exp s_1)\right\},$$

since $s_1 = \log y < 0$ and $s_2 = \log(1-y) = \log(1-\exp s_1)$ is a concave function of s_1 and hence has convex hypograph. This is exactly the result of Barndorff-Neilsen (2014, Example 9.2).

Next, we can define $\bar{\boldsymbol{\theta}}$ as the solution of the first-order condition (8):

$$\frac{\partial Q}{\partial \alpha} = s_1 - \Psi^{(0)}(\alpha) + \Psi^{(0)}(\alpha+\beta) = 0,$$

$$\frac{\partial Q}{\partial \beta} = s_2 - \Psi^{(0)}(\beta) + \Psi^{(0)}(\alpha+\beta) = 0.$$

401 To apply Theorem 2, we write

$$\boldsymbol{g}\left(\boldsymbol{s},\boldsymbol{w}\right)=\left[\begin{array}{c} g_1\left(\boldsymbol{s},\boldsymbol{w}\right) \\ g_2\left(\boldsymbol{s},\boldsymbol{w}\right) \end{array}\right]=\left[\begin{array}{c} s_1-\Psi^{(0)}\left(e^a\right)+\Psi^{(0)}\left(e^a+e^b\right) \\ s_2-\Psi^{(0)}\left(e^b\right)+\Psi^{(0)}\left(e^a+e^b\right) \end{array}\right], \tag{23}$$

where $\boldsymbol{w}^\top=(a,b)\in\mathbb{R}^2$ and $(\alpha,\beta)=\left(e^a,e^b\right)$. As in Section 3.1, (B1)–(B3) are validated by the existence and continuity of $\Psi^{(r)}$, for all $r\geq 0$. Assumption (B4) is verified via the existence of $\boldsymbol{\theta}$; that is, when $\boldsymbol{s}\in\mathbb{S}$. To assess (B5), we require the Jacobian

$$\begin{aligned}\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{w}}&=\left[\begin{array}{cc} \frac{\partial g_1}{\partial a} & \frac{\partial g_1}{\partial b} \\ \frac{\partial g_2}{\partial a} & \frac{\partial g_2}{\partial b} \end{array}\right] \\ &=\left[\begin{array}{cc} -\alpha\Psi^{(1)}\left(\alpha\right)+\alpha\Psi^{(1)}\left(\alpha+\beta\right) & \beta\Psi^{(1)}\left(\alpha+\beta\right) \\ \alpha\Psi^{(1)}\left(\alpha+\beta\right) & -\beta\Psi^{(1)}\left(\beta\right)+\beta\Psi^{(1)}\left(\alpha+\beta\right) \end{array}\right],\end{aligned}$$

402 which has determinant

$$\det\left[\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{w}}\right]=\alpha\beta\left\{\Psi^{(1)}\left(\alpha\right)\Psi^{(1)}\left(\beta\right)-\left[\Psi^{(1)}\left(\alpha\right)+\Psi^{(1)}\left(\beta\right)\right]\Psi^{(1)}\left(\alpha+\beta\right)\right\}. \tag{24}$$

403 Here, we know that

$$\Psi^{(1)}\left(\alpha\right)\Psi^{(1)}\left(\beta\right)-\left[\Psi^{(1)}\left(\alpha\right)+\Psi^{(1)}\left(\beta\right)\right]\Psi^{(1)}\left(\alpha+\beta\right)\neq 0, \tag{25}$$

404 since Q is strictly concave with respect to $\boldsymbol{\theta}$, by Proposition 1, and the left-hand side of (25) is
405 the determinant of its Hessian, and thus (24) is non-zero since $\alpha,\beta>0$, thus verifying (B5),
406 using condition (9).

407 We confirm that there exists a continuously differentiable mapping $\boldsymbol{\chi}\left(\boldsymbol{s}\right)$, such that
408 $\boldsymbol{g}\left(\boldsymbol{s},\boldsymbol{\chi}\left(\boldsymbol{s}\right)\right)=\mathbf{0}$, by noting that $\boldsymbol{g}$ is twice differentially continuous in $(\boldsymbol{s},\boldsymbol{w})$ and thus (A5)
409 is validated, by setting

$$\bar{\boldsymbol{\theta}}\left(\boldsymbol{s}\right)=\left[\begin{array}{c} \exp\left\{\chi_1\left(\boldsymbol{s}\right)\right\} \\ \exp\left\{\chi_2\left(\boldsymbol{s}\right)\right\} \end{array}\right],$$

410 where $\boldsymbol{\chi}\left(\boldsymbol{s}\right)=\left(\chi_1\left(\boldsymbol{s}\right),\chi_2\left(\boldsymbol{s}\right)\right)^\top$ is a continuously differentiable root of (23), as guaranteed
411 by Theorem 2.

A.3. The fully-visible Boltzmann machine

We next consider a multivariate example, where $\boldsymbol{Y}^\top = (Y_1, \ldots, Y_d) \in \{-1, 1\}^d$, characterised by the Boltzmann law PDF

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{\exp\left(\sum_{j=1}^d a_j y_j + \sum_{j=2}^d \sum_{k=1}^{j-1} b_{jk} y_j y_k\right)}{\kappa(\boldsymbol{\theta})}, \tag{26}$$

where

$$\kappa(\boldsymbol{\theta}) = \sum_{\boldsymbol{\zeta} \in \{-1,1\}^d} \exp\left(\sum_{j=1}^d a_j \zeta_j + \sum_{j=2}^d \sum_{k=1}^{j-1} b_{jk} \zeta_j \zeta_k\right),$$

$\boldsymbol{\theta}^\top = (a_1, \ldots, a_d, b_{12}, b_{13}, \ldots, b_{d-1,d}) \in \mathbb{R}^{d(d+1)/2}$, and $\boldsymbol{\zeta}^\top = (\zeta_1, \ldots, \zeta_d)$, which has an exponential family form with $h(\boldsymbol{y}) = 1$, $\psi(\boldsymbol{\theta}) = \log \kappa(\boldsymbol{\theta})$, $\boldsymbol{s}(\boldsymbol{y}) = (y_1, \ldots, y_d, y_1 y_2, y_1 y_3, \ldots, y_{d-1} y_d)^\top$, and $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. Models of form (26) are often referred to as fully-visible Boltzmann machines in the machine learning literature (see, e.g., Bagnall et al. 2020).

The objective function Q can be written as:

$$Q(\boldsymbol{s}; \boldsymbol{\theta}) = \sum_{j=1}^{d(d+1)/2} \theta_j s_j - \log \kappa(\boldsymbol{\theta}),$$

where $\boldsymbol{s}^\top = \left(s_1, \ldots, s_{d(d+1)/2}\right)$. Since $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, we have

$$\boldsymbol{\delta}(\boldsymbol{\phi}(\boldsymbol{\theta})) = \boldsymbol{\delta}(\boldsymbol{\theta}) = \frac{\sum_{\boldsymbol{\zeta} \in \{-1,1\}^d} \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{s}(\boldsymbol{\zeta})\right\} \boldsymbol{s}(\boldsymbol{\zeta})}{\sum_{\boldsymbol{\zeta} \in \{-1,1\}^d} \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{s}(\boldsymbol{\zeta})\right\}}.$$

By Proposition 2, $\bar{\boldsymbol{\theta}}$ exists and is unique when $\boldsymbol{s} \in \mathbb{S} = \boldsymbol{\delta}(\mathbb{P})$, where $\mathbb{P} = \mathbb{R}^{d(d+1)/2}$. Again, we can use the fact that $\boldsymbol{\delta}(\mathbb{P}) = \operatorname{int} \mathbb{D}$ to write $\mathbb{S}$ as the interior of the convex hull of the set $\left\{\boldsymbol{s}(\boldsymbol{y}) : \boldsymbol{y} \in \{-1, 1\}^d\right\}$.

To apply Theorem (2), we simply set

$$g(\boldsymbol{s}, \boldsymbol{w}) = \frac{\partial Q}{\partial \boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{w}). \tag{27}$$

Using $\boldsymbol{w} = \boldsymbol{\theta}$, and noting that $\boldsymbol{\theta} \in \mathbb{R}^{d(d+1)/2}$, we conclude that no change of variables is necessary. Since f is composed of the exponential function, with elementary compositions, (B1)–(B3) can be validated. Assumption (B4) is validated by the existence of $\bar{\boldsymbol{\theta}}$, under the assumption that $\boldsymbol{s} \in \mathbb{S}$. Finally, (B5) is validated since the Jacobian of $\boldsymbol{g}$ is the Hessian of Q, which has non-zero determinant since Q is strictly concave by Proposition 1.

Thus, there exists a continuously differentiable mapping $\boldsymbol{\chi}(\boldsymbol{s})$, such that $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{\chi}(\boldsymbol{s})) = \mathbf{0}$, since $\boldsymbol{g}$ is twice differentially continuous in $(\boldsymbol{s}, \boldsymbol{w})$. Therefore (A5) is validated by setting

$$\bar{\boldsymbol{\theta}}(\boldsymbol{s}) = \boldsymbol{\chi}(\boldsymbol{s}),$$

where $\boldsymbol{\chi}(\boldsymbol{s})$ is a continuously differentiable root of (27), as guaranteed by Theorem 2.

A.4. Online EM algorithm for the Student distribution

We provide details regarding the updating equations of $\boldsymbol{s}^{(i)}$ and $\boldsymbol{\theta}^{(i)}$, as defined in (5) and (6). Let $(\boldsymbol{y}_i)_{i=1}^n$ be n realisations of $\boldsymbol{Y}$, introduced sequentially in the algorithm, starting from $\boldsymbol{y}_1$. At iteration i, for previous iteration of the parameter values $\boldsymbol{\theta}^{(i-1)} = (\boldsymbol{\mu}^{(i-1)\top}, \text{vec}\left(\boldsymbol{\Sigma}^{(i-1)}\right)^\top, \nu^{(i-1)})^\top$, we first need to compute

$$\bar{\boldsymbol{s}}\left(\boldsymbol{y}_i; \boldsymbol{\theta}^{(i-1)}\right) = \begin{bmatrix} u_i^{(i-1)} \boldsymbol{y}_i \\ u^{(i-1)} \text{vec}(\boldsymbol{y}_i \boldsymbol{y}_i^\top) \\ u^{(i-1)} \\ \widetilde{u}_i^{(i-1)} \end{bmatrix},$$

where $u_i^{(i-1)} = \mathrm{E}_{\boldsymbol{\theta}^{(i-1)}}[U|\boldsymbol{Y} = \boldsymbol{y}_i]$ and $\widetilde{u}_i^{(i-1)} = \mathrm{E}_{\boldsymbol{\theta}^{(i-1)}}[\log U|\boldsymbol{Y} = \boldsymbol{y}_i]$. Both these quantities have closed-form expressions (see, e.g., Forbes & Wraith 2014):

$$u_i^{(i-1)} = \frac{\nu^{(i-1)} + 1}{\nu^{(i-1)} + (\boldsymbol{y}_i - \boldsymbol{\mu}^{(i-1)})^\top \boldsymbol{\Sigma}^{(i-1)-1} (\boldsymbol{y}_i - \boldsymbol{\mu}^{(i-1)})},$$
$$\widetilde{u}_i^{(i-1)} = \Psi^{(0)}\left(\frac{\nu^{(i-1)}}{2} + \frac{1}{2}\right) - \log\left(\frac{\nu^{(i-1)}}{2} + \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}^{(i-1)})^\top \boldsymbol{\Sigma}^{(i-1)-1} (\boldsymbol{y}_i - \boldsymbol{\mu}^{(i-1)})\right).$$

It follows that

$$\begin{aligned}
\boldsymbol{s}_1^{(i)} &= \gamma_i \, u_i^{(i-1)} \, \boldsymbol{y}_i + (1 - \gamma_i) \, \boldsymbol{s}_1^{(i-1)}, \\
\boldsymbol{S}_2^{(i)} &= \gamma_i \, u_i^{(i-1)} \, \boldsymbol{y}_i \boldsymbol{y}_i^\top + (1 - \gamma_i) \, \boldsymbol{S}_2^{(i-1)}, \\
s_3^{(i)} &= \gamma_i \, u_i^{(i-1)} + (1 - \gamma_i) \, s_3^{(i-1)}, \\
s_4^{(i)} &= \gamma_i \, \widetilde{u}_i^{(i-1)} + (1 - \gamma_i) \, s_4^{(i-1)}.
\end{aligned}$$

Starting from

$$\begin{aligned}
\boldsymbol{s}_1^{(1)} &= u_1^{(0)}\ \boldsymbol{y}_1, \\
\boldsymbol{S}_2^{(1)} &= u_1^{(0)}\ \boldsymbol{y}_1\boldsymbol{y}_1^\top, \\
s_3^{(1)} &= u_1^{(0)}, \\
s_4^{(1)} &= \widetilde{u}_1^{(0)},
\end{aligned}$$

it follows, with $\widetilde{\gamma}_j = \gamma_j \prod_{j<\ell\leq i} (1-\gamma_\ell)$, that

$$\begin{aligned}
\boldsymbol{s}_1^{(i)} &= \sum_{j=1}^{i} \widetilde{\gamma}_j\ u_j^{(j-1)}\ \boldsymbol{y}_j, \\
\boldsymbol{S}_2^{(i)} &= \sum_{j=1}^{i} \widetilde{\gamma}_j\ u_j^{(j-1)}\ \boldsymbol{y}_j\boldsymbol{y}_j^\top, \\
s_3^{(i)} &= \sum_{j=1}^{i} \widetilde{\gamma}_j\ u_j^{(j-1)}, \\
s_4^{(i)} &= \sum_{j=1}^{i} \widetilde{\gamma}_j\ \widetilde{u}_j^{(j-1)}.
\end{aligned}$$

Using the formulas found in Section 3.2.1, we get parameter updates similar to those for the standard EM algorith (see, e.g., McLachlan & Peel 2000):

$$\begin{aligned}
\boldsymbol{\mu}^{(i)} &= \frac{\boldsymbol{s}_1^{(i)}}{s_3^{(i)}} = \frac{\sum_{j=1}^{i} \widetilde{\gamma}_j\ u_j^{(j-1)}\ \boldsymbol{y}_j}{\sum_{j=1}^{i} \widetilde{\gamma}_j\ u_j^{(j-1)}}, \\
\boldsymbol{\Sigma}^{(i)} &= \sum_{j=1}^{i} \widetilde{\gamma}_j\ u_j^{(j-1)}\ \boldsymbol{y}_j\boldsymbol{y}_j^\top\ -\ s_3^{(i)}\ \boldsymbol{\mu}^{(i)}\boldsymbol{\mu}^{(i)\top},
\end{aligned}$$

which are made of typical weighted sums of the observations, where the weights are inversely proportional to the Mahalanobis distance of the observation to the current center of the distribution. The dof parameter update $\nu^{(i)}$ is then defined as the solution, with respect to ν, of

$$s_4^{(i)} - \Psi^{(0)}\left(\frac{\nu}{2}\right) - s_3^{(i)} + 1 + \log\frac{\nu}{2} = 0\ .$$

A.5. Mean mixtures of normal distributions

In this section we provide the exponential family form of the complete-data likelihoods for mean mixtures of normal distributions and the first steps towards the implementation

of an online EM algorithm for the MLE of these distributions. Like the variance mixtures, mean mixtures involve an additional mixing variable U. The full description of the algorithm requires the specification of the mixing distribution and is not provided here.

If $\boldsymbol{Y}$ follows a mean mixture of normal distributions, then with $\boldsymbol{x}^\top = (\boldsymbol{y}^\top, u)$ and $\boldsymbol{\theta}^\top = (\boldsymbol{\mu}^\top, \text{vec}(\boldsymbol{\Sigma})^\top, \boldsymbol{\delta}^\top, \boldsymbol{\theta}_u^\top)$, where $\boldsymbol{\delta}$ is an additional real vector parameter, $f_c(\boldsymbol{x}; \boldsymbol{\theta})$ can be written as the following product of PDFs:

$$f_c(\boldsymbol{x}; \boldsymbol{\theta}) = \varphi(\boldsymbol{y}; \boldsymbol{\mu} + u\boldsymbol{\delta}, \boldsymbol{\Sigma})\ f_u(u; \boldsymbol{\theta}_u).$$

Using the exponential family forms of both distributions, it follows that

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

where $h\left(\boldsymbol{x}\right) = (2\pi)^{-d/2}\ h_u(u)$, $\psi\left(\boldsymbol{\theta}\right) = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}/2 + \log\det\left[\boldsymbol{\Sigma}\right]/2 + \psi_u(\boldsymbol{\theta}_u)$,

$$\boldsymbol{s}\left(\boldsymbol{x}\right) = \begin{bmatrix} \boldsymbol{y} \\ \text{vec}(\boldsymbol{y}\boldsymbol{y}^\top) \\ u\boldsymbol{y} \\ u^2 \\ u \\ \boldsymbol{s}_u(u) \end{bmatrix}, \text{ and } \boldsymbol{\phi}\left(\boldsymbol{\theta}\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}) \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \\ -\frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \\ -\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \\ \boldsymbol{\phi}_u(\boldsymbol{\theta}_u) \end{bmatrix}. \tag{28}$$

Depending on the statistics defining $\boldsymbol{s}_u(u)$, the representation above can be made more compact.

Considering the objective function $Q(\boldsymbol{s}; \boldsymbol{\theta})$, as per (A3), with $\boldsymbol{s}$ denoted by $\boldsymbol{s}^\top = (\boldsymbol{s}_1^\top, \text{vec}(\boldsymbol{S}_2)^\top, \boldsymbol{s}_3^\top, s_4, s_5, \boldsymbol{s}_6^\top)$, where $\boldsymbol{s}_1, \boldsymbol{s}_3, \boldsymbol{s}_6$ are vectors, $\boldsymbol{S}_2$ is a matrix (all of appropriate dimensions), and s_4, s_5 are scalar values. Whatever the mixing distribution f_u, when maximising Q, closed-form expressions are available for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\delta}$:

$$\begin{aligned} \bar{\boldsymbol{\delta}} &= \frac{s_5\boldsymbol{s}_1 - \boldsymbol{s}_3}{s_5^2 - s_4}, \\ \bar{\boldsymbol{\mu}} &= \boldsymbol{s}_1 - s_5\bar{\boldsymbol{\delta}}, \\ \bar{\boldsymbol{\Sigma}} &= \boldsymbol{S}_2 - \bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^\top - s_4\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}^\top - 2s_5\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\delta}}^\top. \end{aligned}$$

The rest of the expression of $\bar{\boldsymbol{\theta}}_u(\boldsymbol{s})$ then depends on f_u.

From the expressions above, it is possible to derive an online EM algorithm, depending on the tractability of the computation of $\bar{\boldsymbol{s}}(\boldsymbol{y}; \boldsymbol{\theta})$. This quantity requires the computation of conditional moments (e.g., $\text{E}[U|\boldsymbol{Y} = \boldsymbol{y}]$ and $\text{E}[U^2|\boldsymbol{Y} = \boldsymbol{y}]$, which may not always be straightforward. As an illustration, this computation is closed-form for a normal mean mixture

considered by Abdi et al. (2021), obtained when f_u is set to an exponential distribution with fixed known parameter (e.g., a standard exponential distribution, with unit rate).

A.6. Mean and variance mixtures of normal distributions

Mean and variance mixtures of normal distributions combine both the mean and variance mixture cases. This family include in particular a variety of skewed and heavy tailed distributions. Examples and related references are given by Lee & McLachlan (2021).

For a mean and variance mixture of normal variable $\boldsymbol{Y}$, with $\boldsymbol{x}^\top = (\boldsymbol{y}^\top, u)$ and $\boldsymbol{\theta}^\top = (\boldsymbol{\mu}^\top, \text{vec}(\boldsymbol{\Sigma})^\top, \boldsymbol{\delta}^\top, \boldsymbol{\theta}_u^\top)$, the complete-data likelihood $f_c(\boldsymbol{x}; \boldsymbol{\theta})$ can be written as the following product of PDFs (note that in the variance part, u is now appearing as a factor):

$$f_c(\boldsymbol{x}; \boldsymbol{\theta}) = \varphi(\boldsymbol{y}; \boldsymbol{\mu} + u\boldsymbol{\delta}, u\boldsymbol{\Sigma})\ f_u(u; \boldsymbol{\theta}_u).$$

Using expressions (28), replacing $\boldsymbol{\Sigma}$ by $u\boldsymbol{\Sigma}$, it follows that

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

where $h\left(\boldsymbol{x}\right) = (u2\pi)^{-d/2}\ h_u(u)$, $\psi\left(\boldsymbol{\theta}\right) = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta} + \log\det\left[\boldsymbol{\Sigma}\right]/2 + \psi_u(\boldsymbol{\theta}_u)$,

$$\boldsymbol{s}\left(\boldsymbol{x}\right) = \begin{bmatrix} u^{-1}\boldsymbol{y} \\ u^{-1}\text{vec}(\boldsymbol{y}\boldsymbol{y}^\top) \\ \boldsymbol{y} \\ u \\ u^{-1} \\ \boldsymbol{s}_u(u) \end{bmatrix}, \text{ and } \boldsymbol{\phi}\left(\boldsymbol{\theta}\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}) \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \\ -\frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \\ -\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\phi}_u(\boldsymbol{\theta}_u) \end{bmatrix}. \tag{29}$$

Depending on the statistics defining $\boldsymbol{s}_u(u)$, the representation above can be made more compact.

Similar derivations as in the previous section can then be carried out, leading to closed-form expressions for updating $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\delta}$, whatever the mixing distribution f_u:

$$\begin{aligned} \bar{\boldsymbol{\delta}} &= \frac{\boldsymbol{s}_1 - s_5\boldsymbol{s}_3}{1 - s_5 s_4}, \\ \bar{\boldsymbol{\mu}} &= \boldsymbol{s}_3 - s_4\bar{\boldsymbol{\delta}}, \\ \bar{\boldsymbol{\Sigma}} &= \boldsymbol{S}_2 - \boldsymbol{s}_1\bar{\boldsymbol{\mu}}^\top - \boldsymbol{s}_3\bar{\boldsymbol{\delta}}^\top. \end{aligned}$$

The remainder of the expression of $\bar{\boldsymbol{\theta}}_u(\boldsymbol{s})$ depends on f_u.

In particular, the mean variance mixtures include the case of generalised hyperbolic and normal inverse Gaussian (NIG) distributions, which correspond to f_u being the PDF of a generalised inverse Gaussian and inverse Gaussian distributions, respectively. In the NIG case, the required conditional moments to implement an online EM algorithm, $\mathrm{E}[U|\boldsymbol{Y}=\boldsymbol{y}]$ and $\mathrm{E}[U^{-1}|\boldsymbol{Y}=\boldsymbol{y}]$, are given in the Appendix of Karlis & Santourian (2009). If f_u is assumed to be an inverse Gaussian distribution, with parameters α and β, then the updates $\bar{\alpha}=(s_4 s_5)^{-1}$ and $\bar{\beta}=s_5^{-1}$ are also closed-form.

A.7. Online EM algorithm with missing observations

We consider the case of IID vectors $(\boldsymbol{Y}_i)_{i=1}^n$ in dimension $d\in\mathbb{N}$, where some of the dimensions may be missing. For a given $\boldsymbol{Y}_i$, let $\boldsymbol{M}_i\in\{0,1\}^d$ be a binary random variable that is bijective to the power set of $[d]$, where each position of $\boldsymbol{M}_i$ indexes whether the corresponding position of observation $\boldsymbol{Y}_i$ is missing. We let $\boldsymbol{m}_i$ denote a realisation of $\boldsymbol{M}_i$ and we abuse set and vector notation to write $\boldsymbol{m}_i\subset[d]$. We assume that we observe $\boldsymbol{M}_i$. We also write $\bar{\boldsymbol{M}}_i=\mathbf{1}-\boldsymbol{M}_i$ and let $\bar{\boldsymbol{m}}_i$ be its realisation. Here, $\bar{\boldsymbol{m}}_i\subset[d]$ indexes the non-missing dimensions of the realisation $\boldsymbol{y}_i$. We then write $\boldsymbol{Y}_{\boldsymbol{M}_i}$ and $\boldsymbol{Y}_{\bar{\boldsymbol{M}}_i}$ to denote the missing and observed sub-vectors of $\boldsymbol{Y}_i$; that is $\boldsymbol{Y}_{\boldsymbol{M}_i}=(Y_{ik})_{k\in\boldsymbol{M}_i}$. The complete data $\boldsymbol{X}_i$ can then be written as $\boldsymbol{X}_i^\top=\left(\boldsymbol{M}_i^\top,\boldsymbol{Y}_{\bar{\boldsymbol{M}}_i}^\top,\boldsymbol{Y}_{\boldsymbol{M}_i}^\top\right)$, where $\boldsymbol{M}_i$ and $\boldsymbol{Y}_{\bar{\boldsymbol{M}}_i}$ are observed. We will also write $\boldsymbol{X}_i^\top=\left(\boldsymbol{M}_i^\top,\boldsymbol{Y}_i^\top\right)$, for brevity.

Let us assume that the missingness mechanism controlling the $\boldsymbol{M}_i$s depends on some parameter $\boldsymbol{\rho}$, which is known or need not to be estimated. The rest of the parameters to be estimated are gathered in $\boldsymbol{\theta}$. We can then write the complete likelihood,

$$f_c(\boldsymbol{x}_i,\boldsymbol{\theta})=f_{\text{miss}}(\boldsymbol{m}_i|\boldsymbol{y}_i;\boldsymbol{\rho})\; f(\boldsymbol{y}_i;\boldsymbol{\theta}),$$

where $f_{\text{miss}}\left(\cdot;\boldsymbol{\rho}\right)$ characterises the missingness distribution. If f is assumed to be in the exponential family, it follows that

$$f_c(\boldsymbol{x}_i,\boldsymbol{\theta})=h_{\boldsymbol{\rho}}(\boldsymbol{m}_i,\boldsymbol{y}_i)\;\exp\{\boldsymbol{s}(\boldsymbol{y}_i)^\top\boldsymbol{\phi}(\boldsymbol{\theta})-\psi(\boldsymbol{\theta})\}\;,$$

which is also in the exponential family. Making explicit the observed and missing parts, we need to compute

$$\bar{\boldsymbol{s}}(\boldsymbol{m}_i,\boldsymbol{y}_{\bar{\boldsymbol{m}}_i};\boldsymbol{\theta})=\mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{s}(\boldsymbol{Y}_{\boldsymbol{m}_i},\boldsymbol{y}_{\bar{\boldsymbol{m}}_i})\;|\boldsymbol{M}_i=\boldsymbol{m}_i,\boldsymbol{Y}_{\bar{\boldsymbol{M}}_i}=\boldsymbol{y}_{\bar{\boldsymbol{m}}_i}\right]\;.$$

The difficulty may be that the conditional distribution of $(\boldsymbol{Y}_{\boldsymbol{m}_i}\;|\boldsymbol{M}_i=\boldsymbol{m}_i,\boldsymbol{Y}_{\bar{\boldsymbol{M}}_i}=\boldsymbol{y}_{\bar{\boldsymbol{m}}_i})$ may not be known or that the expectation of $\boldsymbol{s}$ with respect to this conditional distribution may

not be easy to compute. In the latter case, we can often resort to approximate computation using Monte-Carlo methods.

In any case, it appears that the allowance of missing observations does not change the definitions of $\boldsymbol{s}$, Q, or $\bar{\boldsymbol{\theta}}$, but impacts upon the computation of $\bar{\boldsymbol{s}}$, with the computation now requiring and account of the imputation of the missing observations.

As an illustration, we detail the multivariate Gaussian case, where $\bar{\boldsymbol{s}}$ can be computed explicitly. In this case, omitting the i index in the notation, $\boldsymbol{s}(\boldsymbol{y})$ is a vector of dimension $d + d^2$ made of the concatenation of vector $\boldsymbol{y}$ and vector $\text{vec}(\boldsymbol{y}\boldsymbol{y}^\top)$. It follows that $\bar{\boldsymbol{s}}$ is also a vector of dimension $d + d^2$, $\bar{\boldsymbol{s}} = (\bar{\boldsymbol{s}}_1^\top, \text{vec}^\top(\bar{\boldsymbol{S}}_2))^\top$, where $\bar{\boldsymbol{s}}_1 = (\bar{\boldsymbol{s}}_{1,k})_{k\in[d]}$ is a vector of dimension d and $\bar{\boldsymbol{S}}_2 = (\bar{\boldsymbol{s}}_{2,k,k'})_{k,k'\in[d]}$ is a $d \times d$ matrix.

We then get that

$$\bar{\boldsymbol{s}}_{1,k} = \begin{cases} y_k & \text{if } k \in \bar{\boldsymbol{m}}, \\ \mathrm{E}_{\boldsymbol{\theta}}\left[Y_k | \boldsymbol{y}_{\bar{\boldsymbol{m}}}\right] & \text{if } k \in \boldsymbol{m}. \end{cases}$$

Similarly,

$$\bar{\boldsymbol{s}}_{2,k,k'} = \begin{cases} y_k y_{k'} & \text{if } k, k' \in \bar{\boldsymbol{m}}, \\ \mathrm{E}_{\boldsymbol{\theta}}\left[Y_k Y_{k'} | \boldsymbol{y}_{\bar{\boldsymbol{m}}}\right] & \text{if } k, k' \in \boldsymbol{m}, \\ \mathrm{E}_{\boldsymbol{\theta}}\left[Y_k | \boldsymbol{y}_{\bar{\boldsymbol{m}}}\right] y_{k'} & \text{if } k \in \boldsymbol{m},\, k' \in \bar{\boldsymbol{m}}, \\ \mathrm{E}_{\boldsymbol{\theta}}\left[Y_{k'} | \boldsymbol{y}_{\bar{\boldsymbol{m}}}\right] y_k & \text{if } k \in \bar{\boldsymbol{m}},\, k' \in \boldsymbol{m}. \end{cases}$$

The conditional distributions involved in the computation of $\bar{\boldsymbol{s}}$ are all Gaussian distributions and the expectations required all have explicit expressions.

Similar computations can be made for the multivariate Student distribution where conditional distributions are Student distributions (cf. Ding 2016), but the additional latent variable U leads to more complicated expectations. However these expectations can easily be approximated by Monte Carlo methods. Other kinds of elliptical distributions could be handled in this manner using results giving expressions of the conditional distributions; see, for example, Cambanis, Huang & Simons (1981) and Frahm (2004).

A.8. Additional illustration

A setting similar to that of Section 3 is used for beta distribution and Boltzmann machine mixtures. We illustrate the case of these two non-identifiable mixtures, where it is possible to have convergence of the algorithm, without consistency.

Random beta observations are generated using the `rbeta` function while observations from the fully-visible Boltzmann machine are generated using the `rfvbm` from the package BoltzMM (Jones, Bagnall & Nguyen 2019).

For the beta distribution scenario, we generate data from a mixture of $K = 3$ beta distributions using parameter values $\alpha_{0z} = 3, 2, 9$, respectively for each of the 3 components $z = 1, 2, 3$. Respectively, we set $\beta_{0z} = 1, 2, 1$ and use the mixture weights 0.5, 0.25, and 0.25. The algorithm is initialised with $\alpha_z^{(0)} = 2, 2, 10$, $\beta_z^{(0)} = 1, 1, 2$ and weights all set to 1/3, for each component $z = 1, 2, 3$, respectively. This setting has been chosen so as to illustrate the non-identifiability issue. The sequences plotted in Figure 3 all converge, but not to the parameter values used to generate the observations. This experiment is also an empirical illustration that the satisfaction of assumptions leading to (7) is independent on the identifiablity of the model. Note that starting from different initial values, it is also possible to recover the parameter values used for simulations. We check numerically that the solution exhibited in Figure 3 is indeed equivalent to the generative beta mixture characterized by $\boldsymbol{\theta}_0$. Under the assumption that the sequences in 3 have become mean stationary for large n, we use the means of last 50 observations of each sequence as parameter estimates. We then obtain estimates $\hat{\alpha}_z = 1.99, 1.99, 10.40$, $\hat{\beta}_z = 0.93, 0.93, 1.12$ and $\hat{\pi}_z = 0.4, 0.4, 0.2$, for $z = 1, 2, 3$. The log-likelihood value corresponding to these estimates is then evaluated to be 100521. This value is very close to the log-likelihood value obtained for the simulated data evaluated at the generative parameters $\boldsymbol{\theta}_0$, which is 100526. In addition Figure 4 shows that the two beta mixture pdfs are extremely close.

To illustrate the online EM algorithm for the fully-visible Boltzmann machine, we consider the $d = 2$ case and generate data using parameter values $a_{01z} = 2, 1, 1$ $a_{02z} = 1, 2, 1$, $b_{012z} = -1, 0, 1$ and $\pi_{0z} = 1/3, 1/3, 1/3$, for $z = 1, 2, 3$, respectively. The algorithm is initialised with $a_{1z}^{(0)} = 1, 1, 1$, $a_{2z}^{(0)} = 1, 1, 1$, $b_{12z}^{(0)} = -2, 1, 2$ and $\pi_z^0 = 1/3, 1/3, 1/3$, for each component $z = 1, 2, 3$, respectively. Although some of the initial values are set to the ones used for simulation, the sequences in Figure 5 illustrate an identifiability issue. Similar to our previous approach, we average the last 100 values in the sequences to obtain parameter estimates. The probability mass function of the estimated mixture is then compared to that of the generative mixture. When $d = 2$, this reduces to compare probabilities for the 4 couples $(1, 1)$, $(1, -1)$, $(-1, 1)$, and $(-1, -1)$. For both mixtures, the probabilities are approximately 0.76, 0.17, 0.07, and 0.01, for each respective couple.

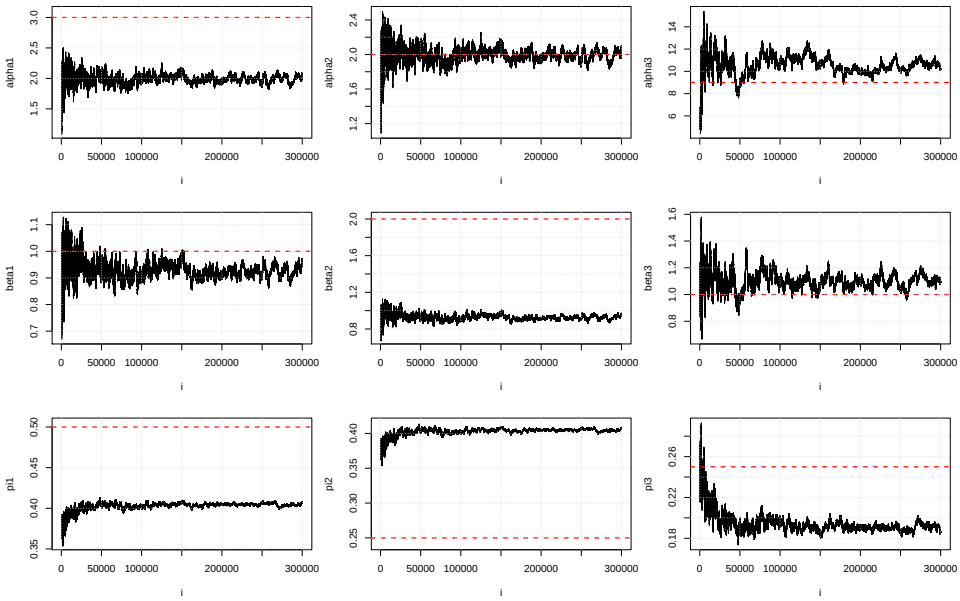

Figure 3. Online EM algorithm sequence of estimator $\boldsymbol{\theta}_z^{(i)} = \left(\alpha_z^{(i)}, \beta_z^{(i)}, \pi_z^{(i)}\right)^\top$ for $z \in [3]$ for a mixture of $K = 3$ beta distributions. The dashed lines indicates the generative parameter values of the DGP. Components are grouped in columns.

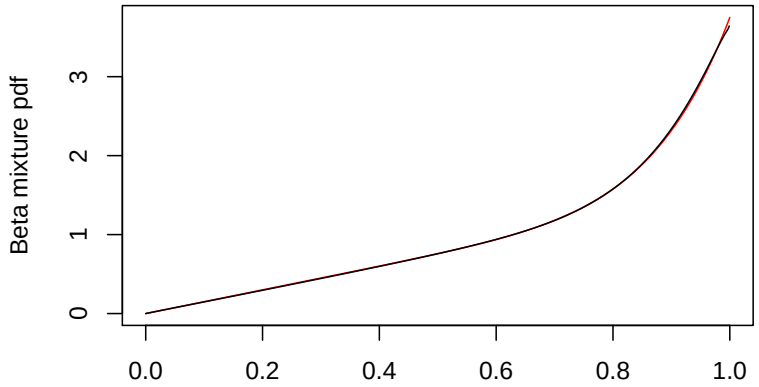

Figure 4. PDFs of beta distributions defined by the DGP parameter vector $\boldsymbol{\theta}_0$ in red and the online EM estimated parameters in black. The proximity of the two pdfs illustrates non-identifiability of beta mixtures.

References

ABDI, M., MADADI, M., BALAKRISHNAN, N. & JAMALIZADEH, A. (2021). Family of mean-mixtures of multivariate normal distributions: Properties, inference and assessment of multivariate skewness. *J. Multivar. Anal.* **181**, 104679.

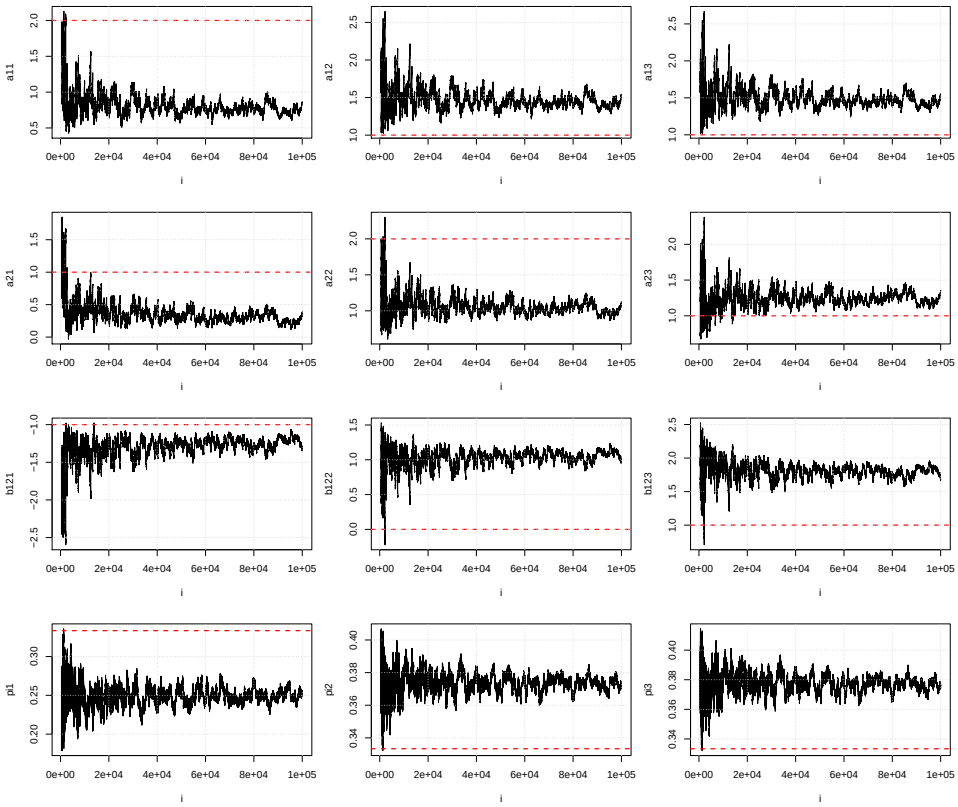

Figure 5. Online EM algorithm sequence of estimator $\boldsymbol{\theta}_z^{(i)} = \left(a_{1z}^{(i)}, a_{2z}^{2(i)}, b_{12z}^{(i)}, \pi_z^{(i)}\right)^\top$ $(z \in [3])$, for a mixture of $K = 3$ Boltzmann machines. The dashed lines indicates the parameter values of the DGP. Components are in column.

AHMAD, K.E.D. & AL-HUSSAINI, E.K. (1982). Remarks on the non-identifiability of mixtures of distributions. *Annals of the Institute of Mathematical Statistics* **34**, 543–544.

ALLASSONNIERE, S. & CHEVALIER, J. (2021). A new class of stochastic EM algorithms. Escaping local maxima and handling inractable sampling. *Computational Statistics and Data Analysis* **159**, 107159.

ARUTYUNOV, A.V. & ZHUKOVSKIY, S.E. (2019). Application of methods of ordinary differential equations to global inverse function theorems. *Differential Equations* **55**, 437–448.

BAGNALL, J.J., JONES, A.T., KARAVARSAMIS, N. & NGUYEN, H.D. (2020). The fully visible Boltzmann machine and the Senate of the 45th Australian Parliament in 2016. *Journal of Computational Social Science* **3**, 55–81.

BARNDORFF-NEILSEN, O. (2014). *Information and Exponential Families in Statistical Theory*. Chichester: Wiley.

BATIR, N. (2005). Some new inequalities for gamma and polygamma functions. *Journal of Inequalities in Pure and Applied Mathematics* **6**, 1–9.

CAMBANIS, S., HUANG, S. & SIMONS, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis* **11**, 368–385.

CAPPE, O. (2009). Online sequential Monte Carlo EM algorithm. In *IEEE/SP 15th Workshop on Statistical Signal Processing*.

CAPPÉ, O. & MOULINES, E. (2009). On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society B* **71**, 593–613.

CRISTEA, M. (2017). On global implicit function theorem. *Journal of Mathematical Analysis and Applications* **456**, 1290–1302.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39**, 1–38.

DING, P. (2016). On the conditional distribution of the multivariate t distribution. `1604.00561`.

FANG, K.T., KOTZ, S. & NG, K.W. (1990). *Symmetric Multivariate And Related Distributions*. London: Chapman and Hall.

FORBES, F. & WRAITH, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Stat. Comput.* **24**, 971–984.

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

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

FRAHM, G. (2004). Generalized elliptical distributions: Theory and applications. Ph.D. thesis, Universität zu Köln.

FRAZIER, D.T., OKA, T. & ZHU, D. (2019). Indirect inference with a non-smooth criterion function. *Journal of Econometrics* **212**, 623–645.

GALEWSKI, M. & KONIORCZYK, M. (2016). On a global implicit function theorem and some applications to integro-differential initial value problems. *Acta Mathematica Hungarica* **148**, 257–278.

GUO, B.N., QI, F., ZHAO, J.L. & LUO, Q.M. (2015). Sharp inequalities for polygamma functions. *Mathematica Slavaca* , 103–120.

HOLZMANN, H., MUNK, A. & GNEITING, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics* **33**, 753–763.

HYVARINEN, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation* **18**, 2283–2292.

ICHIRAKU, S. (1985). A note on global implicit function theorems. *IEEE Transactions on Circuits and Systems* **32**, 503–505.

JOHANSEN, S. (1979). *Introduction to the Theory of Regular Exponential Families*. Copenhagen: Institute of Mathematical Statistics, University of Copenhagen.

JONES, A.T., BAGNALL, J.J. & NGUYEN, H.D. (2019). BoltzMM: an R package for maximum pseudolikelihood estimation of fully-visible Boltzmann machines. *Journal of Open Source Software* **4**, 1193.

KARIMI, B., MIASOJEDOW, B., MOULINES, E. & WAI, H.T. (2019a). Non-asymptotic analysis of biased stochastic approximation scheme. *Proceedings of Machine Learning Research* **99**, 1–31.

KARIMI, B., WAI, H.T., MOULINES, R. & LAVIELLE, M. (2019b). On the global convergence of (fast) incremental expectation maximization methods. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*.

KARLIS, D. & SANTOURIAN, A. (2009). Model-based clustering with non-elliptically contoured distributions. *Stat. Comput.* **19**, 73–83.

KRANTZ, S.G. & PARKS, H.R. (2003). *The Implicit Function Theorem: History, Theory, and Applications*. New York: Birkhauser.

KUHN, E., MATIAS, C. & REBAFKA, T. (2020). Properties of the stochastic approximation EM alpgorithm with mini-batch sampling. *Statistics and Computing* **30**, 1725–1739.

LANGE, K. (2016). *MM Optimization Algorithms.* Philadelphia: SIAM.

LE CORFF, S. & FORT, G. (2013). Online expectation maximization based algorithms for interence in hidden Markov models. *Electronic Journal of Statistics* **7**, 763–792.

LEE, S.X. & MCLACHLAN, G.J. (2021). On mean and/or variance mixtures of normal distributions. *Studies in Classification, Data Analysis, and Knowledge Organization, S. Balzano, G.C. Porzio, R. Salvatore, D. Vistocco, and M. Vichi (Eds.)* .

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

MAIRE, F., MOULINES, E. & LEFEBVRE, S. (2017). Online EM for functional data. *Computational Statistics and Data Analysis* **111**, 27–47.

MCLACHLAN, G.J. & KRISHNAN, T. (2008). *The EM Algorithm And Extensions.* New York: Wiley.

MCLACHLAN, G.J. & PEEL, D. (2000). *Finite Mixture Models.* New York: Wiley.

NGUYEN, H.D., FORBES, F. & MCLACHLAN, G.J. (2020). Mini-batch learning of exponential family finite mixture models. *Statistics and Computing* **30**, 731–748.

OLVER, F.W.J., LOZIER, D.W., BOISVERT, R.F. & CLARK, C.W. (eds.) (2010). *NIST Handbook of Mathematical Functions.* Cambridge: Cambridge University Press.

PHILLIPS, P.C.B. (2012). Folklore theorems, implicit maps, and indirect inference. *Econometrica* **80**, 425–454.

R CORE TEAM (2020). *R: a language and environment for statistical computing.* R Foundation for Statistical Computing.

RAZAVIYAYN, M., SANJABI, M. & LUO, Z.Q. (2016). A stochastic successive minization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *Mathematical Programming* **157**, 515–545.

RONNING, G. (1986). On the curvature of the trigamma function. *Journal of Computational and Applied Mathematics* **15**, 397–399.

SANDBERG, I.W. (1981). Global implicit function theorems. *IEEE Transactions on Circuits and Systems* **28**, 145–149.

SUNDBERG, R. (2019). *Statistical Modelling by Exponential Families.* Cambridge: Cambridge University Press.

TEICHER, H. (1963). Identifiability of finite mixtures. *Annals of Mathematical Statistics* **34**, 1265–1269.

YANG, Y. (2015). *A Concise Text on Advanced Linear Algebra.* Cambridge: Cambridge University Press.

ZHANG, W. & GE, S.S. (2006). A global implicit function theorem without initial point and is applications to control of non-affine systems of high dimensions. *Journal of Mathematical Analysis and Applications* **313**, 251–261.