



HAL
open science

Global implicit function theorems and the online Expectation-Maximisation algorithm

Hien Duy Nguyen, Florence Forbes

► **To cite this version:**

Hien Duy Nguyen, Florence Forbes. Global implicit function theorems and the online Expectation-Maximisation algorithm. Australian and New Zealand Journal of Statistics, 2022, 64 (2), pp.255-281. 10.1111/anzs.12356 . hal-03110213v2

HAL Id: hal-03110213

<https://hal.science/hal-03110213v2>

Submitted on 12 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Global implicit function theorems and the online expectation–maximisation algorithm

Hien Duy Nguyen^{1,2*} and Florence Forbes³

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–maximisation 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.

¹School of Mathematics and Physics, University of Queensland, St. Lucia 4067, Queensland Australia

²Department of Mathematics and Statistics, La Trobe University, Bundoora 3086, Victoria Australia
Email: h.nguyen7@uq.edu.au

³Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

16 & Lefebvre (2017), Karimi et al. (2019a,b), Fort, Moulines & Wai (2020a), Kuhn, Matias
 17 & Rebafka (2020), Nguyen, Forbes & McLachlan (2020), and Allasonniere & Chevalier
 18 (2021), among others. As an archetype of such algorithms, we shall consider the online EM
 19 algorithm of Cappé & Moulines (2009) as a primary example.

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

26 Suppose that \mathbf{Y} arises from some data generating process (DGP) that is characterised
 27 by a probability density function (PDF) $f(\mathbf{y}; \boldsymbol{\theta})$, which is parameterised by a parameter
 28 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
 29 is characterised by an unknown parameter vector $\boldsymbol{\theta}_0 \in \mathbb{T}$. Using the sequence $(\mathbf{Y}_i)_{i=1}^n$, one
 30 wishes to sequentially estimate the parameter vector $\boldsymbol{\theta}_0$. The method of Cappé & Moulines
 31 (2009) assumes the following restrictions regarding the DGP of \mathbf{Y} .

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

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

34 where $h : \mathbb{R}^{d+l} \rightarrow [0, \infty)$, $\psi : \mathbb{R}^p \rightarrow \mathbb{R}$, $\mathbf{s} : \mathbb{R}^{d+l} \rightarrow \mathbb{R}^q$, and $\boldsymbol{\phi} : \mathbb{R}^p \rightarrow \mathbb{R}^q$, for
 35 $q \in \mathbb{N}$.

(A2) The function

$$\bar{\mathbf{s}}(\mathbf{y}; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} [\mathbf{s}(\mathbf{X}) | \mathbf{Y} = \mathbf{y}] \quad (2)$$

36 is well-defined for all $\mathbf{y} \in \mathbb{Y}$ and $\boldsymbol{\theta} \in \mathbb{T}$, where $\mathbb{E}_{\boldsymbol{\theta}} [\cdot | \mathbf{Y} = \mathbf{y}]$ is the conditional
 37 expectation under the assumption that \mathbf{X} arises from the DGP characterised by
 38 $\boldsymbol{\theta}$.

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

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

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

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

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

$$Q(\mathbf{s}; \boldsymbol{\theta}) = \mathbf{s}^\top \boldsymbol{\phi}(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}) \quad (3)$$

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

$$\bar{\boldsymbol{\theta}}(\mathbf{s}) = \arg \max_{\boldsymbol{\theta} \in \mathbb{T}} Q(\mathbf{s}; \boldsymbol{\theta}). \quad (4)$$

44 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
 45 $\boldsymbol{\theta}_0$. For each $i \in [n]$, the method of Cappé & Moulines (2009) proceeds by computing

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

46 and

$$\boldsymbol{\theta}^{(i)} = \bar{\boldsymbol{\theta}}(\mathbf{s}^{(i)}), \quad (6)$$

47 where $\mathbf{s}^{(0)} = \bar{\mathbf{s}}(\mathbf{Y}_1; \boldsymbol{\theta}^{(0)})$. As an output, the algorithm produces a sequence of estimators of
 48 $\boldsymbol{\theta}_0$: $(\boldsymbol{\theta}^{(i)})_{i=1}^n$.

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

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

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

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

52 and define the sets

$$\mathbb{U}_0 = \{l(\bar{\boldsymbol{\theta}}(\mathbf{s})) : \mathbf{s} \in \mathbb{O}\}$$

53 and

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

54 Denote the distance between the real vector \mathbf{a} and the set \mathbb{B} by

$$\text{dist}(\mathbf{a}, \mathbb{B}) = \inf_{\mathbf{b} \in \mathbb{B}} \|\mathbf{a} - \mathbf{b}\|,$$

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

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

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

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

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

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

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

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

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

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

64 Under Assumptions (A1)–(A9), Cappé & Moulines (2009) proved that the sequences
65 $(s^{(i)})_{i=1}^{\infty}$ and $(\theta^{(i)})_{i=1}^{\infty}$, computed via the algorithm defined by (5) and (6), permit the
66 conclusion that

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

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

69 The result can be interpreted as a type of consistency for the estimator $\theta^{(n)}$, as $n \rightarrow \infty$.
70 Indeed if Pr_0 can be characterised by the PDF $f(\mathbf{y}; \theta_0)$ in the family of PDFs $f(\mathbf{y}; \theta)$, where
71 the family is identifiable in the sense that $f(\mathbf{y}; \theta) = f(\mathbf{y}; \theta_0)$ for all $\mathbf{y} \in \mathbb{Y}$, if and only if
72 $\theta = \theta_0$, then $\theta_0 \in \mathbb{M}_{\mathbb{T}}$ and θ_0 is the only value minimising $l(\cdot)$. If there is no other stationary
73 point, then the result guarantees that $\theta^{(n)} \rightarrow \theta_0$, as $n \rightarrow \infty$. If the family is not identifiable,
74 in addition to other stationary points, $\mathbb{M}_{\mathbb{T}}$ could contain several minimisers of $l(\cdot)$, in addition
75 to θ_0 . This situation is illustrated in Section 4. In any case, a lack of identifiability does not
76 affect the nature of \mathbb{O} , due to Proposition 1 in Cappé & Moulines (2009), which states that
77 any two different parameter values θ' and θ'' in $\mathbb{M}_{\mathbb{T}}$, with $f(\cdot; \theta') = f(\cdot; \theta'')$, correspond to
78 the same $s \in \mathbb{O}$.

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

85 In typical applications, the validation of (A5) is conducted by demonstrating that
 86 $Q(\boldsymbol{\theta}; \mathbf{s})$ can be maximised in closed form, and then showing that the closed form maximiser
 87 $\bar{\boldsymbol{\theta}}(\mathbf{s})$ is a continuously differentiable function and hence satisfies (A5). This can be seen,
 88 for example, in the Poisson finite mixture model and normal finite mixture regression
 89 model examples of Cappé & Moulines (2009) and the exponential finite mixture model and
 90 multivariate normal finite mixture model examples of Nguyen, Forbes & McLachlan (2020).

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

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

97 where $\mathbf{J}_\phi(\boldsymbol{\theta}) = \partial \phi / \partial \boldsymbol{\theta}$ is the Jacobian of ϕ , with respect to $\boldsymbol{\theta}$, as a function of $\boldsymbol{\theta}$.

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

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

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

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

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

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

124 2. Global implicit function theorems

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

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

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

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

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

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

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

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

152 (B1) *The mapping \mathbf{g} is continuous.*

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

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

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

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

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

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

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

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

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

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

174 3. Applications of the global implicit function theorem

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

178 We recall the notation from Section 1. Suppose that \mathbf{Y} is a random variable that has
 179 a DGP characterised by a $K \in \mathbb{N}$ component finite mixture model (cf. McLachlan & Peel
 180 2000), where each mixture component has a PDF of the form $f(\mathbf{y}; \boldsymbol{\vartheta}_z)$, for $z \in [K]$, and
 181 $f(\mathbf{y}; \boldsymbol{\vartheta}_z)$ has exponential family form, as defined in (A1). That is, \mathbf{Y} has PDF

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

182 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
 183 $z \in [K]$.

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

192 **Remark 2.** Another type of missing (latent) variable occurs when we have to face missing
 193 observations. We can consider the case of IID vectors of observations, where some of the
 194 elements are missing. Checking assumption (A5) is the same as in the fully observed case but
 195 the computation of $\bar{\mathbf{s}}$ is different as it requires an account of the missing data imputation. An
 196 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 $\mathbf{X}^\top = (\mathbf{Y}^\top, Z)$, we can write the complete-data likelihood in the exponential

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

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

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

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

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 $\mathbf{s}(\mathbf{y})$. We introduce the following notation, for $z \in [K]$:

$$\mathbf{s}_z^{\top} = (s_{1z}, \dots, s_{qz}),$$

$$\text{and } \mathbf{s}_m^{\top} = (s_{01}, \mathbf{s}_1^{\top}, \dots, s_{0K}, \mathbf{s}_K^{\top}),$$

197 where $\mathbf{s}_z \in \mathbb{S}$, for an appropriate open convex set \mathbb{S} , as defined in (A3). Then $\mathbf{s}_m \in \mathbb{S}_m$,

198 where $\mathbb{S}_m = ((0, \infty) \times \mathbb{S})^K$ 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

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

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

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

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

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

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

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

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, $\bar{\boldsymbol{\vartheta}}(\mathbf{s})$, with respect to \mathbf{s} , in order to verify (A5) for the maximiser of the mixture objective Q_m . That is, we can set

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

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

202 3.1. The gamma distribution

203 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 $h(y) = 1$, $\psi(\boldsymbol{\theta}) = \log \Gamma(k) + k \log \theta$, $\mathbf{s}(y) = (\log y, y)^\top$, and $\phi(\boldsymbol{\theta}) = (k-1, -1/\theta)^\top$. Here, $\Gamma(\cdot)$ denotes the gamma function. The objective function Q in (A3) can be written as

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

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

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

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

207

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

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

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

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

213 For any $\mathbf{s}^\top = (s_1, s_2)$, we can solve $\delta(\Phi) = \mathbf{s}$ with respect to Φ , which yields: $\Phi_2 =$
214 $-(1 + \Phi_1)/s_2$ and requires the root of $\Psi^{(0)}(1 + \Phi_1) + \log s_2 - \log(1 + \Phi_1) = s_1$, which
215 is solvable for any \mathbf{s} satisfying $s_1 - \log s_2 < 0$, since both $\Psi^{(0)}(\cdot)$ and $\log(\cdot)$ are continuous,
216 and $\Psi^{(0)}(a) - \log(a)$ is increasing in $a \in (0, \infty)$ and has limits of $-\infty$ and 0, as $a \rightarrow 0$ and
217 $a \rightarrow \infty$, respectively, by Guo et al. (2015, Eqns. 1.5 and 1.6). Thus, $\bar{\theta}$ exists and is unique
218 when

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

219

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

$$\theta = \frac{s_2}{k}, \quad (15)$$

221 which substitutes into (12) to yield:

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

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

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

$$g(\mathbf{s}, w) = s_1 - \Psi^{(0)}(e^w) - \log s_2 + w, \quad (17)$$

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 $\mathbf{s} \in \mathbb{S}$, since it is satisfied if $\bar{\boldsymbol{\theta}}$ exists. Next, to assess (B5), we require the derivative:

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

228 By the main result of Ronning (1986), we have the fact that $-k \Psi^{(1)}(k)$ is negative
229 and strictly increasing for all $k > 0$. Using an asymptotic expansion, it can be shown that
230 $-k \Psi^{(1)}(k) \rightarrow -1$, as $k \rightarrow \infty$ (see the proof of Batir 2005, Lem. 1.2). Thus, (18) is negative
231 for all w , implying that (B5) is validated.

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

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

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

236 3.2. Variance mixtures of normal distributions

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

242 Indeed, if U admits an exponential family form, a variance mixture of normal
243 distributions admits a hierarchical representation whose joint distribution, after data
244 augmentation, belongs to the exponential family. We present the general form in this section
245 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

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

If \mathbf{Y} is a characterized by a variance mixture of a normal distributions, then with $\boldsymbol{x}^\top = (\mathbf{y}^\top, u)$ 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 :

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

246 where $\varphi(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/u)$ is the PDF of a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix
 247 $\boldsymbol{\Sigma}/u$. Here, $\text{vec}(\cdot)$ denotes the vectorisation operator, which converts matrices to column
 248 vectors.

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

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

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

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

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

Consider the objective function $Q(\mathbf{s}; \boldsymbol{\theta})$, as per (A3), with $\mathbf{s}^\top = (\mathbf{s}_1^\top, \text{vec}(\mathbf{S}_2)^\top, s_3, \mathbf{s}_4^\top)$, where \mathbf{s}_1 and \mathbf{s}_4 are real vectors, \mathbf{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 $\boldsymbol{\Sigma}$:

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

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

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

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

259 3.2.1. The Student distribution

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

265 **Remark 4.** *When the two parameters of the gamma distribution are not linked via the joint*
 266 *parameter ν we obtain a slightly more general form of the Student distribution, which is often*
 267 *referred to as the Pearson type VII or generalised Student distribution. Although this later*
 268 *case may appear more general, the Pearson type VII distribution suffers from an identifiability*
 269 *issue that requires a constraint be placed upon the parameters values, which effectively makes*
 270 *it equivalent in practice to the usual Student distribution. See Fang, Kotz & Ng (1990, Sec.*
 271 *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 $f_c(\mathbf{x}; \boldsymbol{\theta})$ is the product of a scaled Gaussian and a gamma PDF, which both belong to the exponential family. More specifically, with $\mathbf{x}^\top = (\mathbf{y}^\top, u)$ and $\boldsymbol{\theta}^\top = (\boldsymbol{\mu}^\top, \text{vec}(\boldsymbol{\Sigma})^\top, \nu)$:

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

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

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

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

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

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 $\bar{\nu}(\mathbf{s})$ as the solution, with respect to ν , of

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

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

276

4. Numerical Simulations

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

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

295 hiendn/onlineEM. In each of the cases, we follow Nguyen, Forbes & McLachlan (2020)
 296 in using the learning rates $(\gamma_i)_{i=1}^\infty$, defined by $\gamma_i = (1 - 10^{-10}) \times i^{-6/10}$, which satisfies
 297 (A7). This choice of γ_i satisfies the recommendation by Cappé & Moulines (2009) to set
 298 $\gamma_i = \gamma_0 i^{-(0.5+\epsilon)}$, with $\epsilon \in (-0.5, 0.5)$ and $\gamma_0 \in (0, 1)$. Our choice agrees with the previous
 299 reports of Cappe (2009) and Kuhn, Matias & Rebafka (2020), who demonstrated good
 300 performance using the same choice of $\epsilon = 0.1$. Furthermore, Le Corff & Fort (2013) and
 301 Allasonniere & Chevalier (2021) showed good good numerical results using $\epsilon = 0.03$ and
 302 $\epsilon = 0.15$, respectively. The online EM algorithm is then run for $n = 100000$ to $n = 500000$
 303 iterations (i.e., using a sequence of observations $(\mathbf{Y}_i)_{i=1}^n$, where $n = 100000$ or $n = 500000$.
 304 Where required, the mappings $\bar{\theta}$ are numerically computed using the `optim` function, which
 305 adequately solves the respective optimisation problems, as defined by (4). Random gamma
 306 observations are generated using the `rgamma` function. Random Student observations are
 307 generated hierarchically using the `rnorm` and `rgamma` functions.

308 For the gamma mixture distribution scenario, we generate data from a mixture
 309 of $K = 3$ gamma distributions using the values $\theta_{0z} = 0.5, 0.05, 0.1$, $k_{0z} = 9, 20, 1$, and
 310 $\pi_{0z} = 0.3, 0.2, 0.5$, for each of the 3 components $z = 1, 2, 3$, respectively. The algorithm
 311 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 $z =$
 312 $1, 2, 3$. For the Student mixture case, we restrict our attention to the $d = 1$ case where
 313 Σ is a scalar, which we denote by $\sigma^2 > 0$. We generate data from a mixture of $K = 3$
 314 Student distributions using $\mu_{0z} = 0, 3, 6$, $\sigma_{0z}^2 = 1, 1, 1$, and $\nu_{0z} = 3, 2, 1$, for $z = 1, 2, 3$.
 315 The corresponding mixture weights are taken as $\pi_{0z} = 0.3, 0.5, 0.2$, for each respective
 316 component. The algorithm is initialised with values set to $\mu_z^{(0)} = 1, 4, 7$, $\sigma_z^{2(0)} = 2, 2, 2$,
 317 $\nu_z^{(0)} = 4, 4, 4$, $\pi_z^{(0)} = 1/3, 1/3, 1/3$, for each component $z = 1, 2, 3$.

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

327 From the two figures, we notice that the sequences each approach and fluctuate around
 328 the respective generative parameter values θ_0 . This provides empirical evidence towards the
 329 correctness of conclusion (7) of Cappé & Moulines (2009, Thm. 1), in the cases considered
 330 in Section 3. In each of the figures, we also observe the decrease in volatility as the iterations
 331 increase. This may be explained by the asymptotic normality of the sequences (cf. Cappé

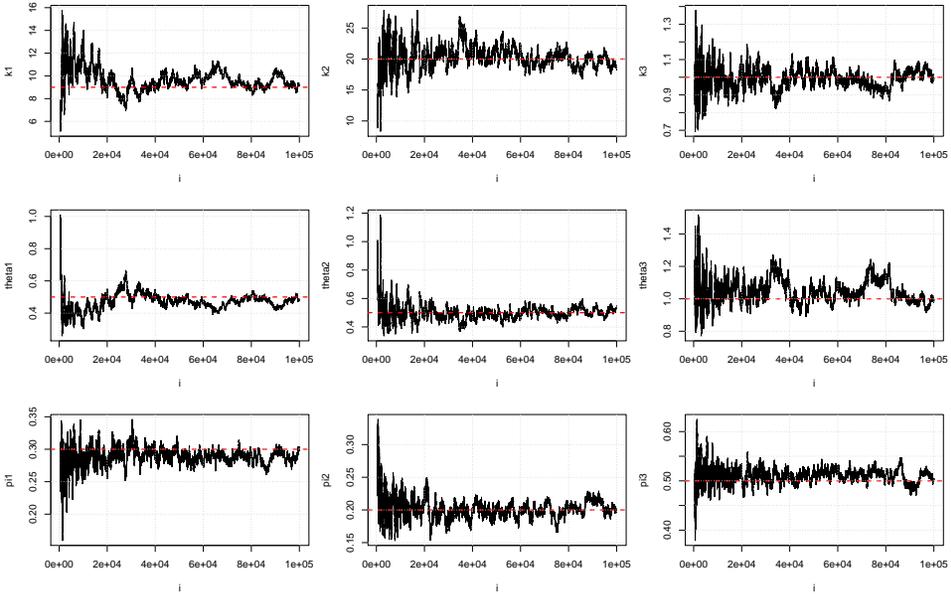


Figure 1. Online EM algorithm estimator sequence $\theta_z^{(i)} = (\theta_z^{(i)}, k_z^{(i)}, \pi_z^{(i)})^\top$ ($z \in [3]$), for a mixture of $K = 3$ gamma distributions. The dashed lines indicates the generative parameter values of the DGP. Components are grouped in columns.

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

336

5. Conclusion

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

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

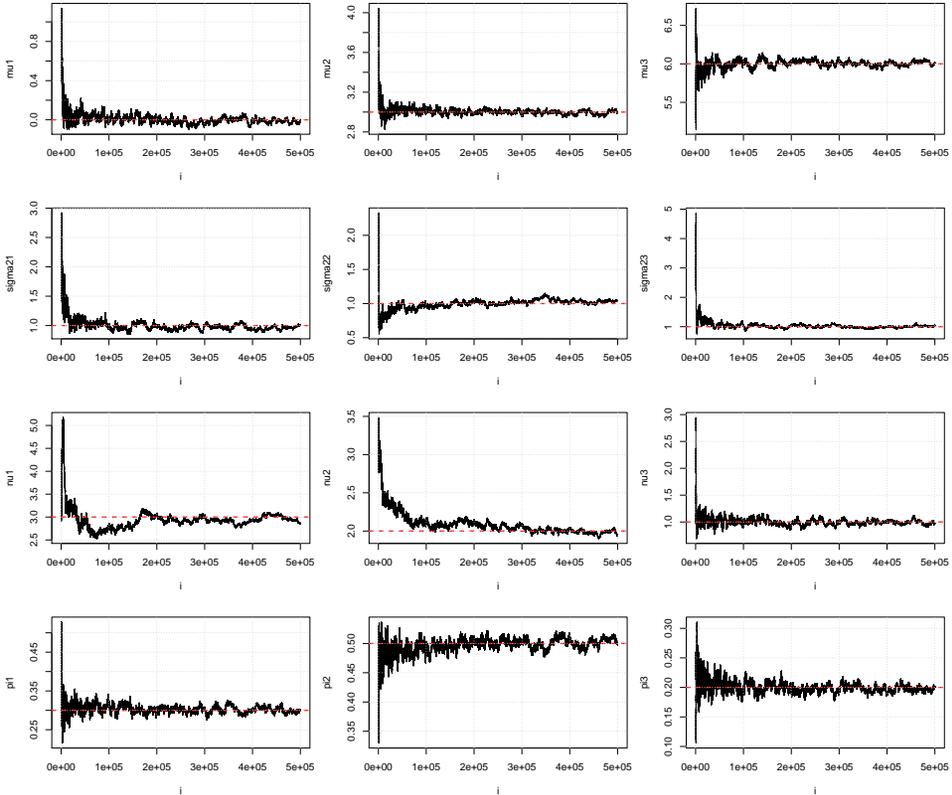


Figure 2. Online EM algorithm sequence of estimator $\theta_z^{(i)} = \left(\mu_z^{(i)}, \sigma_z^{2(i)}, \nu_z^{(i)}, \pi_z^{(i)} \right)^\top$ ($z \in [3]$), for a mixture of $K = 3$ Student distributions. The dashed lines indicates the generative parameter values of the DGP. Components are grouped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

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

360 We also note that when the complete-data likelihood or objective function cannot
 361 be represented in exponential family form, other online algorithms may be required. The
 362 recent works of Karimi et al. (2019b) and Fort, Moulines & Wai (2020b) demonstrate
 363 how penalization and regularization can be incorporated within the online EM framework.
 364 Outside of online EM algorithms, the related online MM (minorisation–maximisation)
 365 algorithms of Mairal (2013) and Razaviyayn, Sanjabi & Luo (2016) can be used to estimate
 366 the parameters of generic distributions. However, these MM algorithms require their own
 367 restrictive assumptions, such as the strong convexity of the objective function and related
 368 expressions. We defer the exploration of applications of the global implicit function theorem
 369 in these settings to future work.

370 A. Appendix

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

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

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

376 where \Pr_{θ} is the probability measure of \mathbb{X} , under the assumption that \mathbb{X} arises from the DGP
 377 characterised by θ . Further, let $\text{int } \mathbb{D}$ be the interior of \mathbb{D} . The following pair of results are
 378 taken from from (Sundberg 2019, Prop. 3.10) and combine (Sundberg 2019, Props. 3.11 and
 379 3.12), respectively (cf. Johansen 1979, Ch. 3, and Barndorff-Neilsen 2014, Ch. 9). The first
 380 result provides conditions under which the objective Q is strictly concave, and the second
 381 provides conditions under which the maximiser (4) exists and is unique.

382 **Proposition 1.** *If (1) is regular and ϕ is bijective, then*

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

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

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

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

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

388 A.2. The beta distribution

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

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

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

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

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

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

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

394 Proposition 2 then states that $\bar{\theta}$ exists and is unique when $\mathbf{s} \in \mathbb{S} = \delta(\mathbb{P})$, where $\mathbb{P} = (0, \infty)^2$.

395 We may then use the fact that $\delta(\mathbb{P}) = \text{int } \mathbb{D}$ to write

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

396 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
 397 hence has convex hypograph. This is exactly the result of Barndorff-Neilsen (2014, Example
 398 9.2).

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

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

400

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

401 To apply Theorem 2, we write

$$\mathbf{g}(\mathbf{s}, \mathbf{w}) = \begin{bmatrix} g_1(\mathbf{s}, \mathbf{w}) \\ g_2(\mathbf{s}, \mathbf{w}) \end{bmatrix} = \begin{bmatrix} s_1 - \Psi^{(0)}(e^a) + \Psi^{(0)}(e^a + e^b) \\ s_2 - \Psi^{(0)}(e^b) + \Psi^{(0)}(e^a + e^b) \end{bmatrix}, \quad (23)$$

where $\mathbf{w}^\top = (a, b) \in \mathbb{R}^2$ and $(\alpha, \beta) = (e^a, e^b)$. 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 $\mathbf{s} \in \mathbb{S}$. To assess (B5), we require the Jacobian

$$\begin{aligned} \frac{\partial \mathbf{g}}{\partial \mathbf{w}} &= \begin{bmatrix} \frac{\partial g_1}{\partial a} & \frac{\partial g_1}{\partial b} \\ \frac{\partial g_2}{\partial a} & \frac{\partial g_2}{\partial b} \end{bmatrix} \\ &= \begin{bmatrix} -\alpha \Psi^{(1)}(\alpha) + \alpha \Psi^{(1)}(\alpha + \beta) & \beta \Psi^{(1)}(\alpha + \beta) \\ \alpha \Psi^{(1)}(\alpha + \beta) & -\beta \Psi^{(1)}(\beta) + \beta \Psi^{(1)}(\alpha + \beta) \end{bmatrix}, \end{aligned}$$

402 which has determinant

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

403 Here, we know that

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

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 $\boldsymbol{\chi}(\mathbf{s})$, such that
408 $\mathbf{g}(\mathbf{s}, \boldsymbol{\chi}(\mathbf{s})) = \mathbf{0}$, by noting that \mathbf{g} is twice differentially continuous in (\mathbf{s}, \mathbf{w}) and thus (A5)
409 is validated, by setting

$$\bar{\boldsymbol{\theta}}(\mathbf{s}) = \begin{bmatrix} \exp \{ \chi_1(\mathbf{s}) \} \\ \exp \{ \chi_2(\mathbf{s}) \} \end{bmatrix},$$

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

412 A.3. The fully-visible Boltzmann machine

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

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

415 where

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

416 $\boldsymbol{\theta}^\top = (a_1, \dots, a_d, b_{12}, b_{13}, \dots, b_{d-1,d}) \in \mathbb{R}^{d(d+1)/2}$, and $\boldsymbol{\zeta}^\top = (\zeta_1, \dots, \zeta_d)$, which
 417 has an exponential family form with $h(\mathbf{y}) = 1$, $\psi(\boldsymbol{\theta}) = \log \kappa(\boldsymbol{\theta})$, $\mathbf{s}(\mathbf{y}) =$
 418 $(y_1, \dots, y_d, y_1 y_2, y_1 y_3, \dots, y_{d-1} y_d)^\top$, and $\phi(\boldsymbol{\theta}) = \boldsymbol{\theta}$. Models of form (26) are often
 419 referred to as fully-visible Boltzmann machines in the machine learning literature (see, e.g.,
 420 Bagnall et al. 2020).

421 The objective function Q can be written as:

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

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

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

423 By Proposition 2, $\bar{\boldsymbol{\theta}}$ exists and is unique when $\mathbf{s} \in \mathbb{S} = \delta(\mathbb{P})$, where $\mathbb{P} = \mathbb{R}^{d(d+1)/2}$. Again,
 424 we can use the fact that $\delta(\mathbb{P}) = \text{int } \mathbb{D}$ to write \mathbb{S} as the interior of the convex hull of the set
 425 $\{\mathbf{s}(\mathbf{y}) : \mathbf{y} \in \{-1, 1\}^d\}$.

426 To apply Theorem (2), we simply set

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

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

432 Thus, there exists a continuously differentiable mapping $\chi(\mathbf{s})$, such that $\mathbf{g}(\mathbf{s}, \chi(\mathbf{s})) =$
 433 $\mathbf{0}$, since \mathbf{g} is twice differentially continuous in (\mathbf{s}, \mathbf{w}) . Therefore (A5) is validated by setting

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

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

435 A.4. Online EM algorithm for the Student distribution

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

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

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

$$u_i^{(i-1)} = \frac{\nu^{(i-1)} + 1}{\nu^{(i-1)} + (\mathbf{y}_i - \boldsymbol{\mu}^{(i-1)})^\top \boldsymbol{\Sigma}^{(i-1)-1} (\mathbf{y}_i - \boldsymbol{\mu}^{(i-1)})},$$

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

It follows that

$$\begin{aligned} \mathbf{s}_1^{(i)} &= \gamma_i u_i^{(i-1)} \mathbf{y}_i + (1 - \gamma_i) \mathbf{s}_1^{(i-1)}, \\ \mathbf{S}_2^{(i)} &= \gamma_i u_i^{(i-1)} \mathbf{y}_i \mathbf{y}_i^\top + (1 - \gamma_i) \mathbf{S}_2^{(i-1)}, \\ \mathbf{s}_3^{(i)} &= \gamma_i u_i^{(i-1)} + (1 - \gamma_i) \mathbf{s}_3^{(i-1)}, \\ \mathbf{s}_4^{(i)} &= \gamma_i \tilde{u}_i^{(i-1)} + (1 - \gamma_i) \mathbf{s}_4^{(i-1)}. \end{aligned}$$

Starting from

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

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

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

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

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

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

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

436 **A.5. Mean mixtures of normal distributions**

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

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

If \mathbf{Y} follows a mean mixture of normal distributions, then with $\mathbf{x}^\top = (\mathbf{y}^\top, u)$ 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, $f_c(\mathbf{x}; \boldsymbol{\theta})$ can be written as the following product of PDFs:

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

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

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

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

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

442 Depending on the statistics defining $s_u(u)$, the representation above can be made more
 443 compact.

Considering the objective function $Q(\mathbf{s}; \boldsymbol{\theta})$, as per (A3), with \mathbf{s} denoted by $\mathbf{s}^\top = (\mathbf{s}_1^\top, \text{vec}(\mathbf{S}_2)^\top, \mathbf{s}_3^\top, s_4, s_5, \mathbf{s}_6^\top)$, where $\mathbf{s}_1, \mathbf{s}_3, \mathbf{s}_6$ are vectors, \mathbf{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 $\boldsymbol{\delta}$:

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

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

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

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

451 **A.6. Mean and variance mixtures of normal distributions**

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

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

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

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

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

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

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

455 Depending on the statistics defining $s_u(u)$, the representation above can be made more
 456 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 :

$$\begin{aligned} \bar{\boldsymbol{\delta}} &= \frac{\mathbf{s}_1 - s_5 \mathbf{s}_3}{1 - s_5 s_4}, \\ \bar{\boldsymbol{\mu}} &= \mathbf{s}_3 - s_4 \bar{\boldsymbol{\delta}}, \\ \bar{\boldsymbol{\Sigma}} &= \mathbf{S}_2 - s_1 \bar{\boldsymbol{\mu}}^\top - s_3 \bar{\boldsymbol{\delta}}^\top. \end{aligned}$$

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

458 In particular, the mean variance mixtures include the case of generalised hyperbolic
 459 and normal inverse Gaussian (NIG) distributions, which correspond to f_u being the PDF of
 460 a generalised inverse Gaussian and inverse Gaussian distributions, respectively. In the NIG
 461 case, the required conditional moments to implement an online EM algorithm, $E[U|\mathbf{Y} = \mathbf{y}]$
 462 and $E[U^{-1}|\mathbf{Y} = \mathbf{y}]$, are given in the Appendix of Karlis & Santourian (2009). If f_u is
 463 assumed to be an inverse Gaussian distribution, with parameters α and β , then the updates
 464 $\bar{\alpha} = (s_4 s_5)^{-1}$ and $\bar{\beta} = s_5^{-1}$ are also closed-form.

465 A.7. Online EM algorithm with missing observations

466 We consider the case of IID vectors $(\mathbf{Y}_i)_{i=1}^n$ in dimension $d \in \mathbb{N}$, where some of the
 467 dimensions may be missing. For a given \mathbf{Y}_i , let $\mathbf{M}_i \in \{0, 1\}^d$ be a binary random variable
 468 that is bijective to the power set of $[d]$, where each position of \mathbf{M}_i indexes whether the
 469 corresponding position of observation \mathbf{Y}_i is missing. We let \mathbf{m}_i denote a realisation of \mathbf{M}_i
 470 and we abuse set and vector notation to write $\mathbf{m}_i \subset [d]$. We assume that we observe \mathbf{M}_i .
 471 We also write $\bar{\mathbf{M}}_i = \mathbf{1} - \mathbf{M}_i$ and let $\bar{\mathbf{m}}_i$ be its realisation. Here, $\bar{\mathbf{m}}_i \subset [d]$ indexes the non-
 472 missing dimensions of the realisation \mathbf{y}_i . We then write $\mathbf{Y}_{\mathbf{M}_i}$ and $\mathbf{Y}_{\bar{\mathbf{M}}_i}$ to denote the missing
 473 and observed sub-vectors of \mathbf{Y}_i ; that is $\mathbf{Y}_{\mathbf{M}_i} = (Y_{ik})_{k \in \mathbf{M}_i}$. The complete data \mathbf{X}_i can then be
 474 written as $\mathbf{X}_i^\top = (\mathbf{M}_i^\top, \mathbf{Y}_{\bar{\mathbf{M}}_i}^\top, \mathbf{Y}_{\mathbf{M}_i}^\top)$, where \mathbf{M}_i and $\mathbf{Y}_{\bar{\mathbf{M}}_i}$ are observed. We will also write
 475 $\mathbf{X}_i^\top = (\mathbf{M}_i^\top, \mathbf{Y}_i^\top)$, for brevity.

Let us assume that the missingness mechanism controlling the \mathbf{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,

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

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

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

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

$$\bar{\mathbf{s}}(\mathbf{m}_i, \mathbf{y}_{\bar{\mathbf{m}}_i}; \boldsymbol{\theta}) = E_\theta [\mathbf{s}(\mathbf{Y}_{\mathbf{m}_i}, \mathbf{y}_{\bar{\mathbf{m}}_i}) | \mathbf{M}_i = \mathbf{m}_i, \mathbf{Y}_{\bar{\mathbf{M}}_i} = \mathbf{y}_{\bar{\mathbf{m}}_i}].$$

476 The difficulty may be that the conditional distribution of $(\mathbf{Y}_{\mathbf{m}_i} | \mathbf{M}_i = \mathbf{m}_i, \mathbf{Y}_{\bar{\mathbf{M}}_i} = \mathbf{y}_{\bar{\mathbf{m}}_i})$
 477 may not be known or that the expectation of \mathbf{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 s , Q , or $\bar{\theta}$, but impacts upon the computation of \bar{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{s} can be computed explicitly. In this case, omitting the i index in the notation, $s(\mathbf{y})$ is a vector of dimension $d + d^2$ made of the concatenation of vector \mathbf{y} and vector $\text{vec}(\mathbf{y}\mathbf{y}^\top)$. It follows that \bar{s} is also a vector of dimension $d + d^2$, $\bar{s} = (\bar{s}_1^\top, \text{vec}^\top(\bar{S}_2))^\top$, where $\bar{s}_1 = (\bar{s}_{1,k})_{k \in [d]}$ is a vector of dimension d and $\bar{S}_2 = (\bar{s}_{2,k,k'})_{k,k' \in [d]}$ is a $d \times d$ matrix.

We then get that

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

Similarly,

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

The conditional distributions involved in the computation of \bar{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.

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

505 For the beta distribution scenario, we generate data from a mixture of $K = 3$ beta
 506 distributions using parameter values $\alpha_{0z} = 3, 2, 9$, respectively for each of the 3 components
 507 $z = 1, 2, 3$. Respectively, we set $\beta_{0z} = 1, 2, 1$ and use the mixture weights 0.5, 0.25, and
 508 0.25. The algorithm is initialised with $\alpha_z^{(0)} = 2, 2, 10$, $\beta_z^{(0)} = 1, 1, 2$ and weights all set to
 509 $1/3$, for each component $z = 1, 2, 3$, respectively. This setting has been chosen so as to
 510 illustrate the non-identifiability issue. The sequences plotted in Figure 3 all converge, but
 511 not to the parameter values used to generate the observations. This experiment is also an
 512 empirical illustration that the satisfaction of assumptions leading to (7) is independent on the
 513 identifiability of the model. Note that starting from different initial values, it is also possible
 514 to recover the parameter values used for simulations. We check numerically that the solution
 515 exhibited in Figure 3 is indeed equivalent to the generative beta mixture characterized by
 516 θ_0 . Under the assumption that the sequences in 3 have become mean stationary for large
 517 n , we use the means of last 50 observations of each sequence as parameter estimates. We
 518 then obtain estimates $\hat{\alpha}_z = 1.99, 1.99, 10.40$, $\hat{\beta}_z = 0.93, 0.93, 1.12$ and $\hat{\pi}_z = 0.4, 0.4, 0.2$,
 519 for $z = 1, 2, 3$. The log-likelihood value corresponding to these estimates is then evaluated
 520 to be 100521. This value is very close to the log-likelihood value obtained for the simulated
 521 data evaluated at the generative parameters θ_0 , which is 100526. In addition Figure 4 shows
 522 that the two beta mixture pdfs are extremely close.

523 To illustrate the online EM algorithm for the fully-visible Boltzmann machine, we
 524 consider the $d = 2$ case and generate data using parameter values $a_{01z} = 2, 1, 1$ $a_{02z} =$
 525 $1, 2, 1$, $b_{012z} = -1, 0, 1$ and $\pi_{0z} = 1/3, 1/3, 1/3$, for $z = 1, 2, 3$, respectively. The algorithm
 526 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
 527 each component $z = 1, 2, 3$, respectively. Although some of the initial values are set to the
 528 ones used for simulation, the sequences in Figure 5 illustrate an identifiability issue. Similar
 529 to our previous approach, we average the last 100 values in the sequences to obtain parameter
 530 estimates. The probability mass function of the estimated mixture is then compared to that of
 531 the generative mixture. When $d = 2$, this reduces to compare probabilities for the 4 couples
 532 $(1, 1)$, $(1, -1)$, $(-1, 1)$, and $(-1, -1)$. For both mixtures, the probabilities are approximately
 533 0.76, 0.17, 0.07, and 0.01, for each respective couple.

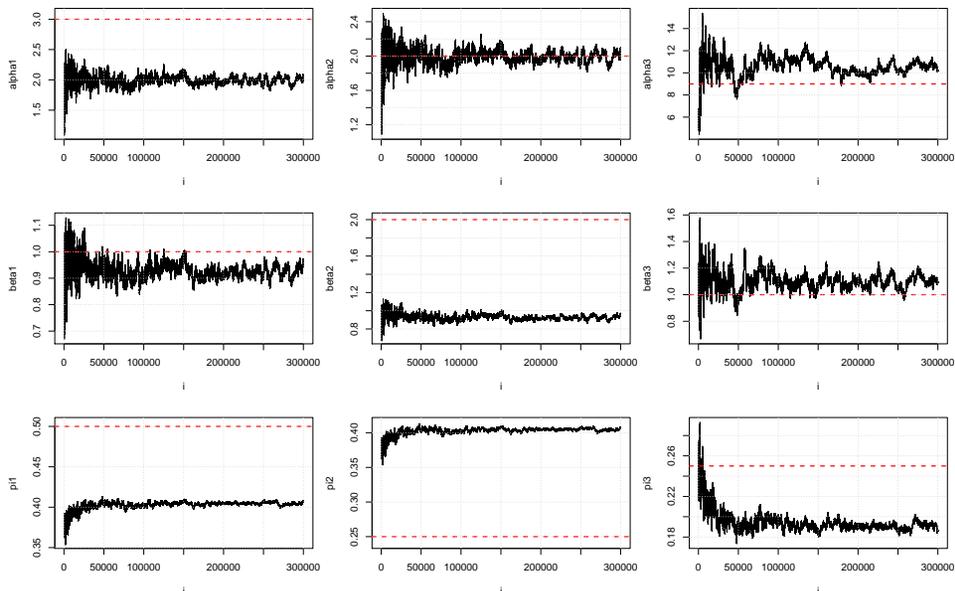


Figure 3. Online EM algorithm sequence of estimator $\theta_z^{(i)} = (\alpha_z^{(i)}, \beta_z^{(i)}, \pi_z^{(i)})^\top$ for $z \in [3]$ for a mixture of $K = 3$ beta distributions. The dashed lines indicates the generative parameter values of the DGP. Components are grouped in columns.

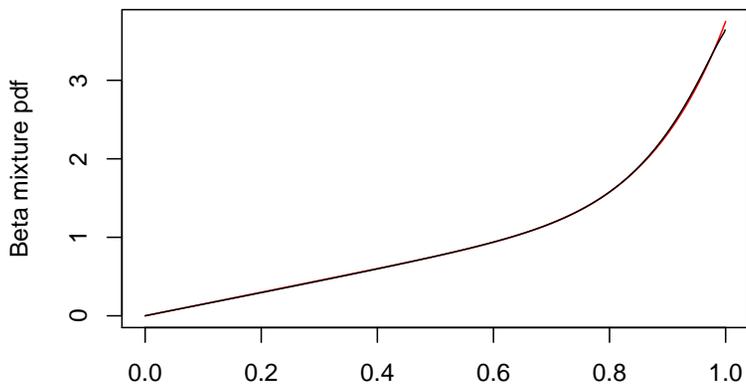


Figure 4. PDFs of beta distributions defined by the DGP parameter vector θ_0 in red and the online EM estimated parameters in black. The proximity of the two pdfs illustrates non-identifiability of beta mixtures.

534

References

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

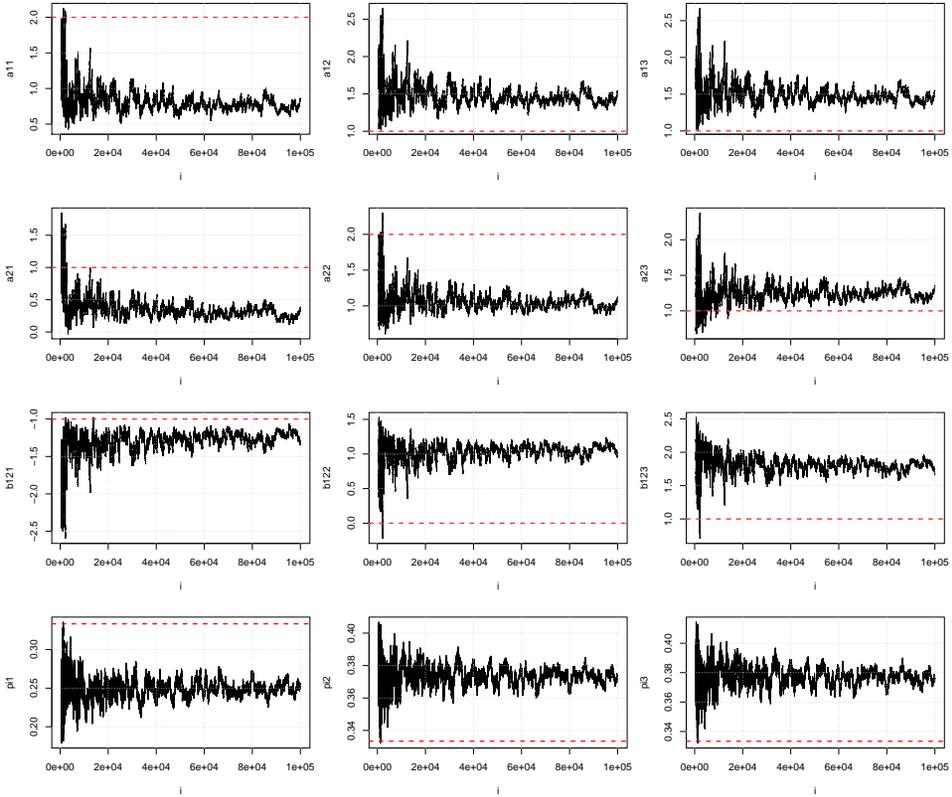


Figure 5. Online EM algorithm sequence of estimator $\theta_z^{(i)} = \left(a_{1z}^{(i)}, a_{2z}^{(i)}, b_{12z}^{(i)}, \pi_z^{(i)} \right)^\top$ ($z \in [3]$), for a mixture of $K = 3$ Boltzmann machines. The dashed lines indicates the parameter values of the DGP. Components are in column.

- 538 AHMAD, K.E.D. & AL-HUSSAINI, E.K. (1982). Remarks on the non-identifiability of mixtures of
539 distributions. *Annals of the Institute of Mathematical Statistics* **34**, 543–544.
- 540 ALLASSONNIERE, S. & CHEVALIER, J. (2021). A new class of stochastic EM algorithms. Escaping local
541 maxima and handling intractable sampling. *Computational Statistics and Data Analysis* **159**, 107159.
- 542 ARUTYUNOV, A.V. & ZHUKOVSKIY, S.E. (2019). Application of methods of ordinary differential equations
543 to global inverse function theorems. *Differential Equations* **55**, 437–448.
- 544 BAGNALL, J.J., JONES, A.T., KARAVARSAMIS, N. & NGUYEN, H.D. (2020). The fully visible Boltzmann
545 machine and the Senate of the 45th Australian Parliament in 2016. *Journal of Computational Social
546 Science* **3**, 55–81.
- 547 BARNDORFF-NEILSEN, O. (2014). *Information and Exponential Families in Statistical Theory*. Chichester:
548 Wiley.
- 549 BATIR, N. (2005). Some new inequalities for gamma and polygamma functions. *Journal of Inequalities in
550 Pure and Applied Mathematics* **6**, 1–9.
- 551 CAMBANIS, S., HUANG, S. & SIMONS, G. (1981). On the theory of elliptically contoured distributions.
552 *Journal of Multivariate Analysis* **11**, 368–385.

- 553 CAPPE, O. (2009). Online sequential Monte Carlo EM algorithm. In *IEEE/SP 15th Workshop on Statistical*
554 *Signal Processing*.
- 555 CAPPÉ, O. & MOULINES, E. (2009). On-line expectation-maximization algorithm for latent data models.
556 *Journal of the Royal Statistical Society B* **71**, 593–613.
- 557 CRISTEA, M. (2017). On global implicit function theorem. *Journal of Mathematical Analysis and*
558 *Applications* **456**, 1290–1302.
- 559 DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the
560 EM algorithm. *Journal of the Royal Statistical Society Series B* **39**, 1–38.
- 561 DING, P. (2016). On the conditional distribution of the multivariate t distribution. 1604.00561.
- 562 FANG, K.T., KOTZ, S. & NG, K.W. (1990). *Symmetric Multivariate And Related Distributions*. London:
563 Chapman and Hall.
- 564 FORBES, F. & WRAITH, D. (2014). A new family of multivariate heavy-tailed distributions with variable
565 marginal amounts of tailweight: application to robust clustering. *Stat. Comput.* **24**, 971–984.
- 566 FORT, G., MOULINES, E. & WAI, H.T. (2020a). A stochastic path-integrated differential estimator
567 expectation maximization algorithm. In *Proceedings of the 34th Conference on Neural Information*
568 *Processing Systems (NeurIPS)*.
- 569 FORT, G., MOULINES, E. & WAI, H.T. (2020b). A stochastic path-integrated differential estimator
570 expectation maximization algorithm. In *Proceedings of the 34th Conference on Neural Information*
571 *Processing Systems (NeurIPS)*.
- 572 FRAHM, G. (2004). Generalized elliptical distributions: Theory and applications. Ph.D. thesis, Universität
573 zu Köln.
- 574 FRAZIER, D.T., OKA, T. & ZHU, D. (2019). Indirect inference with a non-smooth criterion function. *Journal*
575 *of Econometrics* **212**, 623–645.
- 576 GALEWSKI, M. & KONIORCZYK, M. (2016). On a global implicit function theorem and some applications
577 to integro-differential initial value problems. *Acta Mathematica Hungarica* **148**, 257–278.
- 578 GUO, B.N., QI, F., ZHAO, J.L. & LUO, Q.M. (2015). Sharp inequalities for polygamma functions.
579 *Mathematica Slavaca*, 103–120.
- 580 HOLZMANN, H., MUNK, A. & GNEITING, T. (2006). Identifiability of finite mixtures of elliptical
581 distributions. *Scandinavian Journal of Statistics* **33**, 753–763.
- 582 HYVARINEN, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines.
583 *Neural Computation* **18**, 2283–2292.
- 584 ICHIRAKU, S. (1985). A note on global implicit function theorems. *IEEE Transactions on Circuits and*
585 *Systems* **32**, 503–505.
- 586 JOHANSEN, S. (1979). *Introduction to the Theory of Regular Exponential Families*. Copenhagen: Institute
587 of Mathematical Statistics, University of Copenhagen.
- 588 JONES, A.T., BAGNALL, J.J. & NGUYEN, H.D. (2019). BoltzMM: an R package for maximum
589 pseudolikelihood estimation of fully-visible Boltzmann machines. *Journal of Open Source Software*
590 **4**, 1193.
- 591 KARIMI, B., MIASOJEDOW, B., MOULINES, E. & WAI, H.T. (2019a). Non-asymptotic analysis of biased
592 stochastic approximation scheme. *Proceedings of Machine Learning Research* **99**, 1–31.
- 593 KARIMI, B., WAI, H.T., MOULINES, R. & LAVIELLE, M. (2019b). On the global convergence of (fast)
594 incremental expectation maximization methods. In *Proceedings of the 33rd Conference on Neural*
595 *Information Processing Systems (NeurIPS)*.
- 596 KARLIS, D. & SANTOURIAN, A. (2009). Model-based clustering with non-elliptically contoured
597 distributions. *Stat. Comput.* **19**, 73–83.
- 598 KRANTZ, S.G. & PARKS, H.R. (2003). *The Implicit Function Theorem: History, Theory, and Applications*.
599 New York: Birkhauser.
- 600 KUHN, E., MATIAS, C. & REBAFKA, T. (2020). Properties of the stochastic approximation EM algorithm
601 with mini-batch sampling. *Statistics and Computing* **30**, 1725–1739.

- 602 LANGE, K. (2016). *MM Optimization Algorithms*. Philadelphia: SIAM.
- 603 LE CORFF, S. & FORT, G. (2013). Online expectation maximization based algorithms for inference in hidden
604 Markov models. *Electronic Journal of Statistics* **7**, 763–792.
- 605 LEE, S.X. & MCLACHLAN, G.J. (2021). On mean and/or variance mixtures of normal distributions. *Studies*
606 *in Classification, Data Analysis, and Knowledge Organization*, S. Balzano, G.C. Porzio, R. Salvatore,
607 D. Vistocco, and M. Vichi (Eds.) .
- 608 MAIRAL, J. (2013). Stochastic majorization-minimization algorithms for large-scale optimization. In
609 *Advances in Neural Information Processing Systems*.
- 610 MAIRE, F., MOULINES, E. & LEFEBVRE, S. (2017). Online EM for functional data. *Computational*
611 *Statistics and Data Analysis* **111**, 27–47.
- 612 MCLACHLAN, G.J. & KRISHNAN, T. (2008). *The EM Algorithm And Extensions*. New York: Wiley.
- 613 MCLACHLAN, G.J. & PEEL, D. (2000). *Finite Mixture Models*. New York: Wiley.
- 614 NGUYEN, H.D., FORBES, F. & MCLACHLAN, G.J. (2020). Mini-batch learning of exponential family finite
615 mixture models. *Statistics and Computing* **30**, 731–748.
- 616 OLVER, F.W.J., LOZIER, D.W., BOISVERT, R.F. & CLARK, C.W. (eds.) (2010). *NIST Handbook of*
617 *Mathematical Functions*. Cambridge: Cambridge University Press.
- 618 PHILLIPS, P.C.B. (2012). Folklore theorems, implicit maps, and indirect inference. *Econometrica* **80**, 425–
619 454.
- 620 R CORE TEAM (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical
621 Computing.
- 622 RAZAVIYAYN, M., SANJABI, M. & LUO, Z.Q. (2016). A stochastic successive minimization method for
623 nonsmooth nonconvex optimization with applications to transceiver design in wireless communication
624 networks. *Mathematical Programming* **157**, 515–545.
- 625 RONNING, G. (1986). On the curvature of the trigamma function. *Journal of Computational and Applied*
626 *Mathematics* **15**, 397–399.
- 627 SANDBERG, I.W. (1981). Global implicit function theorems. *IEEE Transactions on Circuits and Systems*
628 **28**, 145–149.
- 629 SUNDBERG, R. (2019). *Statistical Modelling by Exponential Families*. Cambridge: Cambridge University
630 Press.
- 631 TEICHER, H. (1963). Identifiability of finite mixtures. *Annals of Mathematical Statistics* **34**, 1265–1269.
- 632 YANG, Y. (2015). *A Concise Text on Advanced Linear Algebra*. Cambridge: Cambridge University Press.
- 633 ZHANG, W. & GE, S.S. (2006). A global implicit function theorem without initial point and its applications to
634 control of non-affine systems of high dimensions. *Journal of Mathematical Analysis and Applications*
635 **313**, 251–261.