Incomplete-data Fisher scoring method with steplength adjustment

Keiji Takai¹

Received: 8 October 2018 / Accepted: 20 January 2020 / Published online: 5 February 2020
© The Author(s) 2020

Abstract

An incomplete-data Fisher scoring method is proposed for parameter estimation in models where data are missing and in latent-variable models that can be formulated as a missing data problem. The convergence properties of the proposed method and an accelerated variant of this method are provided. The main features of this method are its ability to accelerate the rate of convergence by adjusting the steplength, to provide a second derivative of the observed-data log-likelihood function using only the functions used in the proposed method, and the ability to avoid having to explicitly solve the first derivative of the object function. Four examples are presented to demonstrate how the proposed method converges compared with the EM algorithm and its variants. The computing time is also compared.

Keywords Incomplete data · Acceleration · Convergence analysis · EM algorithm · Standard error · Fisher scoring

1 Introduction

As a method for estimating the parameters of probability models with incomplete data, Dempster et al. (1977) formulated the expectation-maximization (EM) algorithm. The EM algorithm has become the standard method of estimating such parameters in the presence of missing data. The EM algorithm has also been extended to estimate parameters for the estimating functions other than the probability models (Elashoff and Ryan 2004) and it can also be used in the context of latent variable models that can be formulated as a missing data problem (e.g., Rubin and Thayer 1982).

Two major limitations of the EM algorithm, which are often emphasized, are its slow convergence and the unavailability of standard errors as a byproduct of the algorithm (McLachlan and Krishnan 2008). Slow convergence means that the increment at one iteration is very small and thus the algorithm might be stopped erroneously (Bentler and Tanaka 1983). To accelerate convergence, various authors have proposed the use of the Newton–Raphson type iteration for the transformed EM algorithm (e.g., Meilijson 1989; Rai

and Matthews 1993; Jamishidian and Jennrich 1993; Lange 1995a, b; Varadhan and Roland 2008) and other methods such as Aitken’s acceleration (e.g., Louis 1982; McLachlan 1995). The unavailability of standard errors occurs because the second derivative of the log-likelihood function cannot be used. As a result, judging whether a convergent point is a local maximum or a saddle point is impossible. Meng and Rubin (1991) proposed the supplemented EM algorithm to provide the second derivative through numerical differentiation.

In this paper, another method for parameter estimation in probability models with incomplete data is proposed, namely, the “incomplete-data Fisher scoring method” (hereafter the IFS method). The IFS method not only overcomes the limitations of the EM algorithm, but it also offers other merits. The IFS method can be accelerated by simple steplength adjustments and the second derivative can be produced through numerical differentiation. The IFS uses the iteration equation,

$$\alpha^{(t+1)} = \alpha^{(t)} + q^{(t)} J_{\text{com}}(\alpha^{(t)})^{-1} \frac{1}{n} \nabla \ell_{\text{obs}}(\alpha^{(t)}) \quad (1)$$

with steplength adjustment, where $\alpha^{(t)}$ is the value of parameter α at the t th iteration, $q^{(t)}$ is the t th steplength, $J_{\text{com}}(\alpha^{(t)})$ is the information matrix for a single complete datum evaluated at $\alpha = \alpha^{(t)}$, and $\nabla \ell_{\text{obs}}(\alpha^{(t)})$ is the first derivative of the observed-data log-likelihood function with missing data of sample size n evaluated at $\alpha = \alpha^{(t)}$. Function $\nabla \ell_{\text{obs}}(\alpha^{(t)})$ can be replaced with $\nabla^{10} Q(\alpha^{(t)} | \alpha^{(t)})$ when necessary, where

This work was supported by JSPS KAKENHI Grant Number JP18K11205.

✉ Keiji Takai
takai@kansai-u.ac.jp

¹ Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka, Japan

$Q()$ is the Q -function of the EM algorithm, which is often easier to calculate than $\nabla \ell_{\text{obs}}(\alpha^{(t)})$, and $\nabla^{10} Q(\alpha^{(t)}|\alpha^{(t)})$ is the first derivative of the Q -function with respect to the first argument and is evaluated at $\alpha = \alpha^{(t)}$. This method is evidently a Fisher scoring method (Osborne 1992) modified for missing data, and also a scaled version of the steepest ascent (Bertsekas 2004). The first advantage of this form is that no exact solution to the first derivative of the observed-data log-likelihood function is required. In cases where an exact solution to the first derivative is not available, the EM algorithm would need to resort to numerical methods such as the steepest ascent (e.g., Lange 1995a). The second advantage of this form is that to develop the IFS method we were able to make use of the vast amount of knowledge that has already been generated with respect to the Newton–Raphson type algorithm. See Bertsekas (2004) on this development.

The IFS method is closely related to the EM algorithm, although the former is derived without starting from the EM map. For further details, this close relationship is exemplified by studies focused on different aspects of the EM algorithm (Titterton 1984; Lange 1995b; Xu and Jordan 1995; Salakhutdinov et al. 2002). Titterton (1984) proposed the same form of the equation as put forward here in terms of the IFS method, as an approximation to the EM algorithm. Lange (1995a) proposed the EM gradient method which converges exactly to Eq. (1) when the distribution belongs to an exponential family of distributions with a natural parameter α . However, application of Eq. (1) is not limited to an exponential family distribution. Xu and Jordan (1995) and Salakhutdinov et al. (2002), among others, expressed the EM algorithm in the form of a scaled version of the steepest ascent method for latent variable models. The IFS method becomes identical to the EM algorithm for certain steplengths as illustrated in the examples which follow. As such, the IFS method can be interpreted as another derivation of the EM algorithm, akin to Minka (1998).

In the next section, two examples are provided to demonstrate the close relationship between the proposed method and the EM algorithm and to emphasize the importance of steplength adjustment. In Sect. 3, a formal definition of the proposed IFS method, and the accelerated IFS, and their derivations are provided. In Sect. 4, properties of the method are described with proofs relegated to the Appendix. In Sect. 5, four numerical examples are shown with the computing time comparison. In Sect. 6, conclusions are offered along with suggestions for future research.

2 Examples

In the following two examples, I show how Eq. (1) of the IFS method is related to the iteration of the conventional EM algorithm (Dempster et al. 1977).

2.1 Normal distribution

Assume that missing data for variable $y = (y_1, y_2)'$ from the two-variate normal distribution $N(\mu, \Sigma)$ are as follows:

$$y_{11}, \dots, y_{1m},$$

$$y_{21}, \dots, y_{2m}, y_{2,m+1}, \dots, y_{2n},$$

where $\mu = (\mu_1, \mu_2)'$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$. The data on variable y_1 are missing from the $(m+1)$ th to the n th cases. For these missing data, we construct a log-likelihood function which is called the “observed-data log-likelihood function” as follows:

$$\begin{aligned} \ell_{\text{obs}}(\alpha) = & -\frac{m}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^m (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \\ & - \frac{n-m}{2} \log \sigma_{22} - \frac{1}{2} \sum_{i=m+1}^n \frac{(y_{2i} - \mu_2)^2}{\sigma_{22}}. \end{aligned} \quad (2)$$

and maximize this function to obtain the maximum likelihood estimates (MLE for short) of the parameters $\alpha = (\mu', \sigma_{11}, \sigma_{12}, \sigma_{22})'$. Note that $\sigma_{12} = \sigma_{21}$. The problem in the maximization is that the function is not easy to handle, e.g., it is not easy to take the derivative

The difficulty in handling the observed-data log-likelihood function can be mitigated by using the complete data and the complete-data log-likelihood function. If the complete data (y_{1i}, y_{2i}) ($i = 1, \dots, n$) were available, the complete-data log-likelihood function,

$$\ell_{\text{com}}(\alpha) = -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^n (y_i - \mu)' \Sigma^{-1} (y_i - \mu),$$

would produce the MLE. However, this is not the case, since the data contain missing values. The basic idea of the EM algorithm is to replace the missing values in $\ell_{\text{com}}(\alpha)$ with their conditional expectation on the observed data. Then, the EM algorithm produces Q -function, $Q(\alpha|\alpha^{(t)}) = E[\ell_{\text{com}}(\alpha)|Y_{\text{obs}}, \alpha^{(t)}]$, where Y_{obs} is the observed data (Dempster et al. 1977).

Deriving the iterations of the EM algorithm by taking the derivative of $Q(\alpha|\alpha^{(t)})$, we have iterations,

$$\begin{aligned} \mu^{(t+1)} = & \frac{1}{n} \left\{ \sum_{i=1}^m y_i + \sum_{i=m+1}^n \left(E \left[\begin{matrix} y_{1i} \\ y_{2i} \end{matrix} \middle| y_{2i}, \alpha^{(t)} \right] \right) \right\}, \\ \text{vec}(\Sigma^{(t+1)}) = & \frac{1}{n} \sum_{i=1}^m \text{vec} \left((y_i - \mu^{(t)})(y_i - \mu^{(t)})' \right) \\ & + \frac{1}{n} \sum_{i=m+1}^n \text{vec} \left(E \left[(y_i - \mu^{(t)})(y_i - \mu^{(t)})' \middle| y_{2i}, \alpha^{(t)} \right] \right), \end{aligned} \quad (3)$$

where $vec(A)$ is a column vector stacking columns in matrix A . Since y_{1i} ($i = m + 1, \dots, n$) are missing values, they are imputed by the conditionally expected value. The iterations can be written as

$$\begin{aligned}\mu^{(t+1)} &= \mu^{(t)} + \frac{1}{n} \left\{ \sum_{i=1}^m (y_i - \mu^{(t)}) \right. \\ &\quad \left. + \sum_{i=m+1}^n \left(E \left[\begin{matrix} y_{1i} - \mu_1^{(t)} \\ y_{2i} - \mu_2^{(t)} \end{matrix} \middle| y_{2i}, \alpha^{(t)} \right] \right) \right\} \\ &= \mu^{(t)} + \frac{1}{n} \Sigma^{(t)} \left\{ \sum_{i=1}^m \left(\Sigma^{(t)} \right)^{-1} (y_i - \mu^{(t)}) \right. \\ &\quad \left. + \sum_{i=m+1}^n \frac{1}{\sigma_{22}^{(t)}} \begin{pmatrix} 0 \\ y_{2i} - \mu_2^{(t)} \end{pmatrix} \right\}.\end{aligned}$$

The function in the curly brackets is the first derivative of $\ell_{\text{obs}}(\alpha)$ with respect to μ . Thus, this iteration can be written as

$$\mu^{(t+1)} = \mu^{(t)} + \frac{1}{n} \Sigma^{(t)} \frac{\partial \ell_{\text{obs}}(\alpha^{(t)})}{\partial \mu}.$$

In a similar way, the EM iteration for Σ becomes

$$vec(\Sigma^{(t+1)}) = vec(\Sigma^{(t)}) + \frac{2}{n} (\Sigma^{(t)} \otimes \Sigma^{(t)}) \frac{\partial \ell_{\text{obs}}(\alpha^{(t)})}{\partial vec(\Sigma)}.$$

(4)

See the Appendix on the equivalence between Eqs. (3) and (4).

In a concise way, by rewriting $\alpha = (\mu', vec(\Sigma)')'$, the iterations of the EM algorithm for a normal distribution are given in the form of Eq. (1) as

$$\alpha^{(t+1)} = \alpha^{(t)} + (n J_{\text{com}}(\alpha^{(t)}))^{-1} \nabla \ell_{\text{obs}}(\alpha^{(t)}),$$

(5)

where $J_{\text{com}}(\alpha) = \begin{pmatrix} \Sigma^{-1} & O \\ O & .5(\Sigma^{-1} \otimes \Sigma^{-1}) \end{pmatrix}$, which is the information matrix of a normal distribution for the complete data of size one. Apparently this is the iteration of Eq. (1) with the unit steplength. The resulting Eq. (5) implies that the expression of this equation might be generally related to the EM algorithm, and that appropriate adjustment of the steplength can accelerate the faster convergence rate than the EM algorithm.

2.2 A Mixture of Poisson distributions

In this example, we consider a mixture of Poisson distributions. This model involves no missing data, but a latent variable. However, the mixture of Poisson distributions can

be formulated as a missing data problem by augmenting the data.

Let $y_1, \dots, y_n$ be data taken independently from a mixture of two Poisson distributions: $Po(y_i | \alpha_1)$ with probability $\pi (> 0)$ and $Po(y_i | \alpha_2)$ with probability $1 - \pi$ ($i = 1, \dots, n$). The parameters to be estimated are $\alpha = (\pi, \alpha_1, \alpha_2)'$. The observed-data log-likelihood function is given as

$$\ell_{\text{obs}}(\alpha) = \sum_{i=1}^n \log \{ \pi Po(y_i | \alpha_1) + (1 - \pi) Po(y_i | \alpha_2) \}.$$

Because taking the derivative is overly cumbersome, an avoidance strategy is needed. The strategy augments the original data by introducing a random binary variable z_i ($i = 1, \dots, n$) which takes one if y_i belongs to $Po(y_i | \alpha_1)$ and zero if y_i belongs to $Po(y_i | \alpha_2)$, and constructs augmented complete data, (y_i, z_i) ($i = 1, \dots, n$). The complete-data log-likelihood function is

$$\begin{aligned}\ell_{\text{com}}(\alpha) &= - \sum_{i=1}^n z_i \alpha_1 + \sum_{i=1}^n z_i y_i \log \alpha_1 - \sum_{i=1}^n (1 - z_i) \alpha_2 \\ &\quad + \sum_{i=1}^n (1 - z_i) y_i \log \alpha_2 \\ &\quad + \sum_{i=1}^n z_i \log \pi + \sum_{i=1}^n (1 - z_i) \log(1 - \pi).\end{aligned}$$

Using the complete-data log-likelihood function, we are able to construct two seemingly different but actually closely related iterations.

Applying the EM algorithm, we obtain the Q -function of the EM algorithm, $Q(\alpha | \alpha^{(t)}) = E[\ell_{\text{com}}(\alpha) | y; \alpha^{(t)}]$. With the standard notation

$$z_i^{(t)} = E[z_i | y_i; \alpha^{(t)}] = \frac{\pi^{(t)} Po(y_i | \alpha_1^{(t)})}{\pi^{(t)} Po(y_i | \alpha_1^{(t)}) + (1 - \pi^{(t)}) Po(y_i | \alpha_2^{(t)})}$$

(See McLachlan and Krishnan 2008), the iterations of the EM algorithm become

$$\begin{aligned}\pi^{(t+1)} &= \frac{\sum_{i=1}^n z_i^{(t)}}{n}, \quad \alpha_1^{(t+1)} = \sum_{i=1}^n \frac{z_i^{(t)} y_i}{\sum_{j=1}^n z_j^{(t)}}, \\ \alpha_2^{(t+1)} &= \sum_{i=1}^n \frac{(1 - z_i^{(t)}) y_i}{\sum_{j=1}^n (1 - z_j^{(t)})}.\end{aligned}$$

Applying the IFS method in Eq. (1), we need to compute $\nabla \ell_{\text{obs}}(\alpha)$. However, in this case, $\nabla \ell_{\text{obs}}(\alpha)$ is hard to compute. Instead, we use the useful relationship that $\nabla \ell_{\text{obs}}(\alpha) = \nabla^{10} Q(\alpha | \alpha^{(t)})$ which makes it less difficult to compute the derivative. Then, we have iterations,

$$\begin{aligned}\pi^{(t+1)} &= \pi^{(t)} + q_1^{(t)} \left(\frac{\sum_{i=1}^n z_i^{(t)}}{n} - \pi^{(t)} \right), \\ \alpha_1^{(t+1)} &= \alpha_1^{(t)} + q_2^{(t)} \frac{1}{n\pi^{(t)}} \sum_{i=1}^n z_i^{(t)} (y_i - \alpha_1^{(t)}), \\ \alpha_2^{(t+1)} &= \alpha_2^{(t)} + q_3^{(t)} \frac{1}{n(1-\pi^{(t)})} \sum_{i=1}^n (1 - z_i^{(t)}) (y_i - \alpha_1^{(t)}),\end{aligned}$$

where we have used different steplengths for each element of the parameters, $q_1^{(t)}$, $q_2^{(t)}$ and $q_3^{(t)}$, and $J_{\text{com}}(\alpha) = \text{diag}(1, \pi, 1 - \pi)$.

The iterations of the EM algorithm again become identical to those of the IFS method by adjusting steplengths. By setting $q_1^{(t)} = 1$, the first iteration equation becomes that of the EM algorithm. The second and third iteration equations become those of the EM algorithm by setting $q_2^{(t)} = n\pi^{(t)} / \sum_{i=1}^n z_i^{(t)}$ and $q_3^{(t)} = n(1 - \pi^{(t)}) / \sum_{i=1}^n (1 - z_i^{(t)})$, respectively. Let the direction provided by the IFS method be $\mathbf{d} = (d_1, d_2, d_3)'$. The direction by the EM algorithm is

$$1 \begin{pmatrix} d_1 \\ 0 \\ 0 \end{pmatrix} + \frac{n\pi^{(t)}}{\sum_{i=1}^n z_i^{(t)}} \begin{pmatrix} 0 \\ d_2 \\ 0 \end{pmatrix} + \frac{n(1 - \pi^{(t)})}{\sum_{i=1}^n (1 - z_i^{(t)})} \begin{pmatrix} 0 \\ 0 \\ d_3 \end{pmatrix}.$$

This is not the steepest ascent direction in terms of the IFS method because of the different steplengths for each element. Steplengths $q_2^{(t)}$ and $q_3^{(t)}$ converge to one as $|\pi^{(t)} - \pi^{(t+1)}| \rightarrow 0$. This means that in the latter stages of convergence, the sequence produced by the IFS method approaches that of the EM algorithm when the IFS method and the EM algorithm start from the common initial values around a convergent point.

Steplengths for the EM algorithm are obtained as follows. The optimum steplength of $q_1^{(t)}$ is a solution to

$$\frac{d}{dq_1^{(t)}} \ell_{\text{obs}}(\pi + q_1^{(t)} d_1, \alpha_1, \alpha_2) = 0,$$

where $d_1 \neq 0$. The left-side term can be expressed as

$$\sum_{i=1}^n \left\{ \frac{z_i^{(t)}}{\pi + q_1^{(t)} d_1} - \frac{1 - z_i^{(t)}}{1 - \pi - q_1^{(t)} d_1} \right\} d_1 = 0,$$

where $z_i^{(t)}$ is fixed as the current estimate, although it is actually a function of $\alpha^{(t)}$. The solution is $q_1^{(t)} = 1$. In a similar way, for $q_2^{(t)}$,

$$\sum_{i=1}^n \frac{z_i^{(t)}}{\alpha_1} (y_i - \alpha_1 - q_2^{(t)} d_2) d_2 = 0$$

leads to $q_1^{(t)} = n\pi^{(t)} / \sum_{i=1}^n z_i^{(t)}$. The steplength $q_3^{(t)}$ becomes identical to that given above in a similar way. Thus, from observation, we can state that the EM algorithm, or the IFS method with these steplengths, proceeds in the same direction as the IFS method in equation (1). The EM algorithm takes (inexact) different steplengths for each element of direction, while the IFS method proceeds on a common steplength. To proceed in the same direction, the common steplength is better than different steplengths for different elements in the same direction. How to determine an appropriate steplength will be discussed in Sect. 3.

3 Definition and derivation

3.1 Definition

As per the foregoing, assume that we have the observed-data log-likelihood function, denoted by $\ell_{\text{obs}}(\alpha)$, for incomplete data of sample size n and $r \times 1$ parameter vector α . The complete-data log-likelihood function for this sample size, denoted by $\ell_{\text{com}}(\alpha)$, is not available for estimation since data contain missing values or the model contains a latent variable. However, the complete-data log-likelihood function can be used as a function of random variables. Thus, its information matrix for a single datum can be computed as $-E[\nabla^2 \ell_{\text{com}}(\alpha^{(t)})] = nJ_{\text{com}}(\alpha^{(t)})$, where E is the expectation with respect to the joint distribution of the complete-data variables (Kenward and Molenberghs 1998). In the following, a theory is developed for independent and identically distributed data, which can also be extended for dependent data in a straightforward manner.

We define the IFS method as follows:

IFS method

Step 0. Determine $\epsilon (> 0)$. Set $t = 0$ and an initial value for $\alpha^{(0)}$.

Step 1. If $\|\nabla \ell_{\text{obs}}(\alpha^{(t)})\| < \epsilon$, do not proceed to the subsequent steps.

Step 2. Compute $\mathbf{d}^{(t)} = n^{-1} J_{\text{com}}(\alpha^{(t)})^{-1} \nabla \ell_{\text{obs}}(\alpha^{(t)})$.

Step 3. Provide the steplength $q^{(t)}$ according to some method and adjust it using the Armijo Rule with $\tilde{\mathbf{d}}^{(t)} = q^{(t)} \mathbf{d}^{(t)}$.

(Armijo Rule) For a given $\beta \in (0, 1)$ and a given $\sigma \in (0, .5)$, find the maximum $s^{(t)} \in \{\beta^\ell | \ell = 0, 1, \dots\}$ such that

$$\ell_{\text{obs}}(\alpha^{(t)} + s^{(t)} \tilde{\mathbf{d}}^{(t)}) - \ell_{\text{obs}}(\alpha^{(t)}) > s^{(t)} \sigma \nabla \ell_{\text{obs}}(\alpha^{(t)})' \tilde{\mathbf{d}}^{(t)}.$$

Step 4. Improve the current estimate $\alpha^{(t)}$ to $\alpha^{(t+1)}$ by $\alpha^{(t+1)} = \alpha^{(t)} + s^{(t)} \tilde{\mathbf{d}}^{(t)}$; set $t \leftarrow t + 1$ and go to Step 1.

In Step 1, the algorithm determines whether or not the convergence criterion is satisfied. In some practical cases, the distance between $\alpha^{(t)}$ and $\alpha^{(t+1)}$ is also used to determine the convergence. This convergence criterion is natural because if $d^{(t)} = 0$, $\nabla \ell_{\text{obs}}(\alpha^{(t)})$ is reduced to zero under the assumption that $J_{\text{com}}(\alpha^{(t)})$ is nonsingular. In Step 2, if $J_{\text{com}}(\alpha^{(t)})$ is positive definite, then $\nabla \ell_{\text{obs}}(\alpha^{(t)})' d^{(t)} > 0$, which means that $d^{(t)}$ is an ascent direction. Positive definiteness is a natural assumption that is usually satisfied. In Step 3, the steplength is determined to ensure that the likelihood increases. An alternative method such as the Wolfe condition (Nocedal and Wright 2006) is also available with the same direction vector $\tilde{d}^{(t)}$. There are various ways to determine $q^{(t)}$ including a random guess. With the results obtained from the previous steps, the current estimate $\alpha^{(t)}$ is improved to $\alpha^{(t+1)}$ in Step 4.

There is no definitive way to determine steplength. An initial steplength might need adjustment to increase the likelihood function, often by shortening the initial steplength. The problem is that the adjusted steplength may render convergence slower than the EM algorithm. Thus, we need a way to identify a good initial steplength that is likely to increase the likelihood function with as few adjustments as possible. The optimal steplength is $q(\geq 0)$, which minimizes $\ell_{\text{obs}}(\alpha^{(t)} + qd^{(t)})$ for a certain direction $d^{(t)}$. However, this steplength cannot be derived exactly because the exact minimization is practically difficult. Instead, an inexact steplength is provided by the following Eq. (6). See Sect. 3.3 for its derivation. The steplength in Eq. (6) accelerates the convergence of the IFS when it is used as the initial steplength. We denote the IFS with Eq. (6) “AIFS” (which stands for “accelerated IFS”). The AIFS might require a steplength adjustment at the initial stage of iterations, while, around a convergent point, the AIFS accepts the initial steplength of Eq. (6) without any adjustment, which is proved in Lemma 1 (ii) in the Appendix. It is also shown that the AIFS converges faster than the EM algorithm (Theorem 2 with the subsequent comments).

To increase the convergence rate of the IFS method, the accelerated IFS (AIFS) method is thus proposed herein.

AIFS method

- Step 0. Determine $\epsilon (> 0)$. Set $t = 0$ and an initial value for $\alpha^{(0)}$.
- Step 1. If $\|\nabla \ell_{\text{obs}}(\alpha^{(t)})\| < \epsilon$, then do not proceed to subsequent steps.
- Step 2. Compute $d^{(t)} = n^{-1} J_{\text{com}}(\alpha^{(t)})^{-1} \nabla \ell_{\text{obs}}(\alpha^{(t)})$.
- Step 3. Provide the steplength $q^{(t)}$ and adjust it using the Armijo Rule with $\tilde{d}^{(t)} = q^{(t)} d^{(t)}$, where

$$q^{(t)} = \frac{nd^{(t)'} J_{\text{com}}(\alpha^{(t)}) d^{(t)}}{d^{(t)'} D(\alpha^{(t)}, \alpha^{(t)} + d^{(t)})}, \quad (6)$$

and $D(\alpha_1, \alpha_2) = \nabla \ell_{\text{obs}}(\alpha_1) - \nabla \ell_{\text{obs}}(\alpha_2)$.

- Step 4. Improve the current estimate $\alpha^{(t)}$ to $\alpha^{(t+1)}$ by $\alpha^{(t+1)} = \alpha^{(t)} + s^{(t)} \tilde{d}^{(t)}$; set $t \leftarrow t + 1$ and go to Step 1.

The only difference between the IFS and AIFS is with respect to $q^{(t)}$. In the former, $q^{(t)}$ is arbitrary, while in the latter $q^{(t)}$ is systematically estimated. The latter brings greater advantage than the former. Not only does the latter accelerate the convergence, but it also provides the standard errors of the estimates. The following IFS-Variance method explains how to do the estimation.

On the termination of either of these methods, a numerical differential produces the second derivative of $\ell_{\text{obs}}(\alpha)$ at the convergent point. Let the convergent point be α^* . Let e_b ($b = 1, \dots, r$) be a $r \times 1$ vector with one only at the b th element and zeros in all other elements, and let $G = [g_{ab}]_{a,b=1,\dots,r}$ be the second derivative of $\ell_{\text{obs}}(\alpha)$ at α^* by numerical differentiation.

IFS-Variance method

- Step 0. Determine $k (> 0)$.
- Step 1. For $a, b = 1, \dots, r$ ($a \leq b$), compute

$$g_{ab} = \frac{D_a(\alpha^* + ke_b, \alpha^* - ke_b)}{2k},$$

where $D_a(\alpha_1, \alpha_2)$ is the a th element of $D(\alpha_1, \alpha_2)$.

- Step 2. $G \leftarrow G + G' - [g_{aa}]_{a=1,\dots,r}$, where $[g_{aa}]$ is a diagonal matrix of elements which are diagonal elements of G .

The standard error matrix of α^* is computed by inverting $-G/n$.

3.2 Derivation of direction vector

Assume that we improve the current estimate $\alpha^{(t)}$ to $\alpha^{(t+1)}$ by using the relation,

$$\alpha^{(t+1)} = \alpha^{(t)} + qd,$$

where $q \geq 0$ is the unknown steplength and d is the unknown direction vector at the t th iteration. They must be chosen so that $\ell_{\text{obs}}(\alpha^{(t+1)}) > \ell_{\text{obs}}(\alpha^{(t)})$.

The IFS method can be derived as a modification of the Fisher scoring method (Osborne 1992). The Fisher scoring method for $\ell_{\text{obs}}(\alpha)$ finds a direction vector d using a quadratic model function:

$$m_F^{(t)}(d) = \ell_{\text{obs}}(\alpha^{(t)}) + \nabla \ell_{\text{obs}}(\alpha^{(t)})' d + \frac{1}{2} d' E[\nabla^2 \ell_{\text{obs}}(\alpha^{(t)})] d,$$

where E denotes the expectation with respect to random variables of data and the corresponding missing-data indicator variables (Kenward and Molenberghs 1998; Seaman et al. 2013). The last term contains $E[\nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})]$ which is often difficult to compute. Thus, to avoid this computation, we invoke the idea of a lower-bound algorithm (Böhning and Lindsay 1988). The algorithm optimizes a quadratic model function which is lower than the original quadratic function. The lower bound of $E[\nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})]$ is $-nJ_{\text{com}}(\boldsymbol{\alpha}^{(t)})$, that is, $E[\nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})] \geq -nJ_{\text{com}}(\boldsymbol{\alpha}^{(t)})$, where this inequality denotes that $\mathbf{x}'(E[\nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})] + nJ_{\text{com}}(\boldsymbol{\alpha}^{(t)}))\mathbf{x} \geq 0$ for an arbitrary vector $\mathbf{x}$ of an appropriate size. See the Appendix for the derivation. Based on this relationship, we have another model which minorizes $m_F^{(t)}(\mathbf{d})$,

$$m_P^{(t)}(\mathbf{d}) = \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) + \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})' \mathbf{d} - \frac{n}{2} \mathbf{d}' J_{\text{com}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}.$$

By taking the first derivative with respect to $\mathbf{d}$, we have

$$\mathbf{d}^{(t)} = J_{\text{com}}(\boldsymbol{\alpha}^{(t)})^{-1} \frac{1}{n} \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}). \quad (7)$$

Using $\nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) = \nabla^{10} Q(\boldsymbol{\alpha}^{(t)} | \boldsymbol{\alpha}^{(t)})$ given in Fisher (1925) and McLachlan and Krishnan (2008), $\mathbf{d}^{(t)}$ can also be written as

$$\mathbf{d}^{(t)} = J_{\text{com}}(\boldsymbol{\alpha}^{(t)})^{-1} \frac{1}{n} \nabla^{10} Q(\boldsymbol{\alpha}^{(t)} | \boldsymbol{\alpha}^{(t)}),$$

where ∇^{10} indicates the first derivative with respect to the first parameter. As observed in the Poisson mixture example, the Q -function of the EM algorithm could be useful to compute the first derivative of the observed-data log-likelihood function.

Another reason why $J_{\text{com}}(\boldsymbol{\alpha})$ appears as a scaled matrix is explored. If the complete data were available, the Fisher scoring method would be given as

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \frac{1}{n} J_{\text{com}}(\boldsymbol{\alpha}^{(t)})^{-1} \nabla \ell_{\text{com}}(\boldsymbol{\alpha}^{(t)}).$$

The Fisher scoring method has several merits (Small and Wang 2003). However, we cannot use it here since we only have the observed data. An available primal estimating function is $\ell_{\text{obs}}(\boldsymbol{\alpha})$. This function is known to produce a consistent estimator that converges to the same point as $\ell_{\text{com}}(\boldsymbol{\alpha})$, that is, to the true value of $\boldsymbol{\alpha}$, under the assumption that data are missing at random (Little and Rubin 2002). When an estimating function is close to $n^{-1} J_{\text{com}}(\boldsymbol{\alpha})^{-1} \nabla \ell_{\text{com}}(\boldsymbol{\alpha})$, we can expect that the estimate sequence will have properties as good as those of $n^{-1} J_{\text{com}}(\boldsymbol{\alpha})^{-1} \nabla \ell_{\text{com}}(\boldsymbol{\alpha})$. Now, it is proposed that a scaled version of $\nabla \ell_{\text{obs}}(\boldsymbol{\alpha})$ should be configured, denoted by $B \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})$, as close to $n^{-1} J_{\text{com}}(\boldsymbol{\alpha})^{-1} \nabla \ell_{\text{com}}(\boldsymbol{\alpha})$ as possible, where B is an appropriate-sized matrix-valued function

of $\boldsymbol{\alpha}$. Let the distance between vectors $\mathbf{a}$ and $\mathbf{b}$ be measured by $\text{dist}(\mathbf{a}, \mathbf{b}) = E[\|\mathbf{a} - \mathbf{b}\|^2]$. The problem to be solved is

$$\min_B \text{dist}(B \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}), n^{-1} J_{\text{com}}(\boldsymbol{\alpha})^{-1} \nabla \ell_{\text{com}}(\boldsymbol{\alpha})).$$

The solution is $B = n^{-1} J_{\text{com}}(\boldsymbol{\alpha})^{-1}$. Thus, $n^{-1} J_{\text{com}}(\boldsymbol{\alpha})^{-1} \nabla \ell_{\text{obs}}(\boldsymbol{\alpha})$ is the best approximation to the direction of the Fisher scoring method for complete data in the sense of expectation. Further, because of its derivation, the IFS direction is closer to $n^{-1} J_{\text{com}}(\boldsymbol{\alpha})^{-1} \nabla \ell_{\text{com}}(\boldsymbol{\alpha})$ than $E[\nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha})]^{-1} \nabla \ell_{\text{obs}}(\boldsymbol{\alpha})$ which is the direction provided by direct application of the Fisher scoring method to $\ell_{\text{obs}}(\boldsymbol{\alpha})$. The direction by the IFS method for incomplete data aims to approach the direction by the Fisher scoring method for complete data.

These two derivations are not based on the EM algorithm. Finally, the EM-based derivation is given to show its close relationship with the EM algorithm. The EM gradient algorithm (Lange 1995a) provides an iteration,

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \left\{ \nabla^{20} Q(\boldsymbol{\alpha}^{(t)} | \boldsymbol{\alpha}^{(t)}) \right\}^{-1} \nabla^{10} Q(\boldsymbol{\alpha}^{(t)} | \boldsymbol{\alpha}^{(t)})$$

where ∇^{20} indicates the second derivative of the Q -function with respect to the former part of $\boldsymbol{\alpha}$. In the same spirit as the Fisher scoring method, we take the expectation of the second derivative of the Q -function with respect to $\mathbf{y}$ and its indicator (Kenward and Molenberghs 1998; Seaman et al. 2013). As a result, the iteration of the same form as Eq. (1) with $q^{(t)} = 1$ is obtained.

Whether the IFS method is derived from the EM algorithm or not, the IFS method uses complete data and a complete-data log-likelihood function just as the EM algorithm does. The EM algorithm uses a complete-data log-likelihood function to derive a surrogate function $Q(\boldsymbol{\alpha} | \boldsymbol{\alpha}^{(t)}) = E[\ell_{\text{com}}(\boldsymbol{\alpha}) | Y_{\text{obs}}, \boldsymbol{\alpha}^{(t)}]$ to $\ell_{\text{obs}}(\boldsymbol{\alpha})$. The Q -function is expected to be easier to handle than $\ell_{\text{obs}}(\boldsymbol{\alpha})$. On the other hand, the IFS method uses the complete-data log-likelihood function to compute the information matrix. It is intriguing that they produce similar iterations as in the previous examples.

3.3 Derivation of steplength

To ensure an increase in the likelihood function, the steplength is determined with an exact or inexact line search. The exact line search finds a steplength maximizing $\ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)} + q\mathbf{d}^{(t)})$ ($q \geq 0$). This type of line search is sometimes impractical because $d\ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)} + q\mathbf{d}^{(t)})/dq = 0$ is difficult to solve. Instead, here it is proposed that an inexact line search should be used. This solves a quadratic approximation of $\ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)} + q\mathbf{d}^{(t)})$,

$$\ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) + q \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})' \mathbf{d}^{(t)} + \frac{q^2}{2} \mathbf{d}^{(t)'} \nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}^{(t)}.$$

The solution is given as

$$q_0 = -\frac{\nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})' \mathbf{d}^{(t)}}{\mathbf{d}^{(t)'} \nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}^{(t)}} = \frac{\mathbf{d}^{(t)'} J_{\text{com}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}^{(t)}}{\mathbf{d}^{(t)'} I_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}^{(t)}},$$

where $I_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) = -n^{-1} \nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})$, and when transforming q_0 to the last equation, we have used $J_{\text{com}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}^{(t)} = n^{-1} \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})$ which results from Eq. (7). It may be practically impossible to use matrix $I_{\text{obs}}(\boldsymbol{\alpha}^{(t)})$, because the form of $\ell_{\text{obs}}(\boldsymbol{\alpha})$ is complicated and taking its derivative is difficult. Instead, we suggest here that the following should be used:

$$q^{(t)} = \frac{n \mathbf{d}^{(t)'} J_{\text{com}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}^{(t)}}{\mathbf{d}^{(t)'} \mathbf{D}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\alpha}^{(t)} + \mathbf{d}^{(t)})}.$$

This estimate is derived by use of this relationship:

$$\nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)} + \mathbf{d}^{(t)}) = \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) + \nabla^2 \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) \mathbf{d}^{(t)} + o(\|\mathbf{d}^{(t)}\|),$$

where $o(\|\mathbf{d}^{(t)}\|)$ indicates a small quantity compared with $\|\mathbf{d}^{(t)}\|$, that is, $o(\|\mathbf{d}^{(t)}\|)/\|\mathbf{d}^{(t)}\| \rightarrow 0$ as $\|\mathbf{d}^{(t)}\| \rightarrow 0$. This relationship indicates that $q^{(t)}$ gets close enough to q_0 for sufficiently small $\|\mathbf{d}^{(t)}\|$'s.

It happens in practice that when, for example, $\|\mathbf{d}^{(t)}\|$ is not sufficiently small, $q^{(t)}$ may not qualify as the steplength. The steplength may be too long, or even negative. In cases where $q^{(t)}$ is too long, the steplength is reduced until the increase in the likelihood value is ensured. Such a steplength is always found provided that the first derivative of $\ell_{\text{obs}}(\boldsymbol{\alpha})$ is not zero, as will be shown in Theorem 1. When $q^{(t)}$ is negative, after changing its sign or replacing it with an arbitrary positive number, appropriate reduction of the steplength results in a steplength that increases the likelihood value. To reduce the initial steplength, the Armijo Rule is used.

4 Properties

We investigate the properties of the (A)IFS method. In the following, the parameter space, denoted by Ω , is a subset of the p -dimensional real space, R^p . A positive integer set is denoted by Z .

To derive the convergence property, we make the following assumptions.

- (A1) There exist positive constants c_1 and c_2 such that $c_1 \|\mathbf{v}\|^2 \leq \mathbf{v}' J_{\text{com}}(\boldsymbol{\alpha}^{(t)}) \mathbf{v} \leq c_2 \|\mathbf{v}\|^2$ for the sequence $\{\boldsymbol{\alpha}^{(t)}\}$ and all $\mathbf{v} \in R^p$.
- (A2) The level set $L_0 = \{\boldsymbol{\alpha} \in \Omega | \ell_{\text{obs}}(\boldsymbol{\alpha}) \geq \ell_{\text{obs}}(\boldsymbol{\alpha}^{(0)}) > -\infty\}$ is compact.

(A3) $\ell_{\text{obs}}(\boldsymbol{\alpha})$ is twice differentiable in the interior of Ω .

Assumption (A1) is not restrictive. Indeed, in many practical cases, it is naturally met. As shown in the previous section, the EM algorithm for some models also requires this assumption, proceeding on the basis that it implicitly holds. Assumption (A2) guarantees that there exists an accumulation point of $\{\boldsymbol{\alpha}^{(t)}\}$ in L_0 . Assumption (A3) is required to conduct the convergence analysis, and is also used for this purpose with respect to the EM algorithm. Twice differentiability is not unnatural because the information matrix based on the twice differential log-likelihood usually exists on the parameter space.

Theorem 1 *The IFS method for $\ell_{\text{obs}}(\boldsymbol{\alpha})$ produces a sequence of $\boldsymbol{\alpha}^{(t)}$. The sequence satisfies the following properties.*

- (i) Assume that $\nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) \neq \mathbf{0}$. The steplength produced by the Armijo Rule is determined within a finite number of iterations and the likelihood function monotonically increases with the steplength.
- (ii) Every limit point of $\{\boldsymbol{\alpha}^{(t)}\}$ is a stationary point.

Proof See Proposition 1.2.1 in Bertsekas (2004). $\square$

Note that Theorem 1 does not imply that the IFS method converges to a global maximum from any initial value. Theorem 1 just indicates that $\boldsymbol{\alpha}^*$ may be a local maximum or a saddle point.

Corollary 1 *If the log-likelihood function is concave with respect to $\boldsymbol{\alpha}$, the convergent point $\boldsymbol{\alpha}^*$ satisfies $\ell_{\text{obs}}(\boldsymbol{\alpha}) \leq \ell_{\text{obs}}(\boldsymbol{\alpha}^*)$ for $\forall \boldsymbol{\alpha} \in \Omega$. Further, if “concavity” can be replaced with “strict concavity,” then $\ell_{\text{obs}}(\boldsymbol{\alpha}) < \ell_{\text{obs}}(\boldsymbol{\alpha}^*)$ for all $\boldsymbol{\alpha} \in \Omega \setminus \{\boldsymbol{\alpha}^*\}$.*

Proof See Appendix. $\square$

The rate of convergence is of interest here. The convergence rate of the proposed method is investigated in terms of the likelihood value $\{\ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})\}_t$. The following notation $o(1)$ indicates a quantity that goes to zero as $\boldsymbol{\alpha}^{(t)}$ goes to $\boldsymbol{\alpha}^*$.

Theorem 2 *Assume that the maximum and minimum eigenvalues of $J_{\text{com}}(\boldsymbol{\alpha}^*)^{-1} I_{\text{obs}}(\boldsymbol{\alpha}^*)$ are λ_1 and λ_r ($1 > \lambda_1 > \lambda_r > 0$). Let $\Delta(\boldsymbol{\alpha}^{(t)}) = |\ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) - \ell_{\text{obs}}(\boldsymbol{\alpha}^*)|$.*

- (i) For IFS, there exists a sufficiently large t_1 such that

$$\Delta(\boldsymbol{\alpha}^{(t+1)}) < (1 - s^{(t)} q^{(t)} \lambda_r)^2 \Delta(\boldsymbol{\alpha}^{(t)}) + o(1), \text{ for } \forall t \geq t_1,$$

where $s^{(t)} q^{(t)} \in (0, 1/\lambda_r)$.

- (ii) For AIFS, there exists a sufficiently large t_2 such that

$$\Delta(\boldsymbol{\alpha}^{(t+1)}) < \left(\frac{\lambda_1 - \lambda_r}{\lambda_1 + \lambda_r} \right)^2 \Delta(\boldsymbol{\alpha}^{(t)}) + o(1), \text{ for } \forall t \geq t_2.$$

Proof See Appendix. $\square$

This theorem states that the AIFS method converges faster than, or at the same rate as, the IFS when

$$s^{(t)}q^{(t)} \leq \frac{2}{\lambda_1 + \lambda_r}$$

around the convergent point α^* . In other words, where steplengths do not satisfy this inequality, the IFS converges faster than the AIFS method. However, before executing the method, it is impossible to know these eigenvalues. In practice, the AIFS should be used from the outset. The EM algorithm (at least the variants appearing in this paper) is regarded as the IFS method with $s^{(t)} = 1$ and $q^{(t)} \rightarrow 1$. Thus, it can be asserted that the IFS with $s^{(t)}q^{(t)} > 1$ converges at a faster rate than the EM algorithm. Note that this inequality does not express the whole sequence of the method/algorithm.

Corollary 2 Let $\|\cdot\|_{I_o}$ be the norm with metric matrix $I_{\text{obs}}(\alpha^*)$. Under the same setting as Theorem 2,

(i) For IFS, there exists a sufficiently large t_1 such that

$$\|\alpha^{(t+1)} - \alpha^*\|_{I_o}^2 < (1 - s^{(t)}q^{(t)}\lambda_p)^2 \|\alpha^{(t)} - \alpha^*\|_{I_o}^2 + o(1)$$

for $\forall t \geq t_1$, where $s^{(t)}q^{(t)} \in (0, 1/\lambda_r)$.

(ii) For AIFS, there exists a sufficiently large t_2 such that

$$\|\alpha^{(t+1)} - \alpha^*\|_{I_o}^2 < \left(\frac{\lambda_1 - \lambda_r}{\lambda_1 + \lambda_r}\right)^2 \|\alpha^{(t)} - \alpha^*\|_{I_o}^2 + o(1)$$

for $\forall t \geq t_2$.

Proof See “Appendix”. $\square$

5 Examples

5.1 Normal distribution

To illustrate how the steplength adjustment works in an empirical context, the AIFS method is compared with the IFS method with $q^{(t)} = 1.5$, and also with the EM algorithm. Note that $q^{(t)} = 1.5$ is an arbitrary number, and the IFS method with $q_1^{(t)} = q_2^{(t)} = 1$ is identical to the EM algorithm. Hereafter, these are referred to as *AIFS*, *IFS*, and *EM*, respectively. Apple crop data from Table 7.4 of Little and Rubin (2002) are used for expository purposes. The data take the same form as in Sect. 2.1 when $n = 18$ and $m = 12$. The three methods to be compared share the initial values, $\mu^{(0)} = (30, 30)'$ and $\Sigma^{(0)} = \text{diag}(100, 100)$, and the convergence criterion $\|\nabla \ell_{\text{obs}}(\alpha^{(t)})\| < 10^{-4}$, where $\text{diag}(x)$ denotes a diagonal matrix with x in its diagonal elements and zeros otherwise.

Table 1 gives $\Delta(\alpha^{(t)}) = |\ell_{\text{obs}}(\alpha^{(t)}) - \ell_{\text{obs}}(\alpha^*)|$, $s^{(t)}q^{(t)}$ of *AIFS*, *IFS*, and $q^{(t)}$ of *EM*. All values are rounded to four decimal places. Thus, note that 0.0000 is not equal to zero, but a rounded zero. Where “–” appears, this denotes an empty cell because the steplength is not used or the computation is terminated. In the initial stages, *EM* comes closest to $\ell_{\text{obs}}(\alpha^*)$. In later stages, when $\Delta(\alpha^{(t)})$ comes closer to zero, *EM* is further from the convergent point than the other two methods, while *AIFS* and *IFS* become closer around α^* . As per Theorem 2, it is confirmed that *AIFS* converges faster than *IFS* and *EM* around the convergent point, and as per this example, it is also confirmed that *AIFS* converges faster than the others even for the whole sequence.

The standard error of the estimate obtained by the (A)IFS method is generated by the IFS-Variance method. In this example, the second derivative of $\ell_{\text{obs}}(\alpha)$ can be explicitly obtained (e.g., Savalei 2010), and thus the standard error matrix is also obtained. The distance between the former and the latter, measured in terms of the sum of squares, was less than 10^{-8} . This indicates that the IFS-Variance method works well here.

5.2 Poisson mixture distribution

Three parameter estimation methods are compared using empirical data: the AIFS, the IFS with the randomly guessed initial value $q^{(t)} = 2$, and the EM algorithm. These are denoted by *AIFS*, *IFS*, and *EM*, respectively. Mortality data from the 1910–1912 *London Times* newspaper (Hasselblad 1969; Titterton 1984) are used for this purpose. The data record the number of deaths per day and the number of days with y_i deaths ($i = 1, \dots, 1096$). The data are a better fit to a two-component Poisson mixture model than a single component model (Lange 1995b). The initial value is the moment estimates $(\pi^{(0)}, \alpha_1^{(0)}, \alpha_2^{(0)}) = (0.2870, 2.582, 1.101)$ and the convergence criterion is $\|\nabla \ell_{\text{obs}}(\alpha^{(t)})\| < 10^{-4}$ for all models.

Figure 1 shows the convergence process for *AIFS*, *IFS*, and *EM*. The horizontal axis indicates the iteration number to the 300th iteration and the vertical axis is $\Delta(\alpha^{(t)})$. Notice that until the 300th iteration *AIFS* converges but the remaining two methods are still in the process of convergence. The log-likelihood values are all -1989.946 for *AIFS*, *IFS*, and *EM* when these methods satisfy the convergence criterion. The value is reported as the maximum log-likelihood function value (Lange 1995b). The number of iterations until convergence is 196, 1474, and 2208 for *AIFS*, *IFS*, and *EM*, respectively. The figure shows that *AIFS* converges much faster than the other two methods and the decrease is not so smooth as the other two; *IFS* and *EM* behave smoothly. Around the convergent point of *IFS* and *EM*, which is not shown in the figure, *IFS* satisfies the convergence criterion faster than *EM*. The limitation of *EM* here is its slow behavior around the con-

Table 1 Convergence comparison for *AIFS*, *IFS*, and *EM* using crop data.

Iteration	<i>AIFS</i>		<i>IFS</i>		<i>EM</i>	
	$\Delta(\alpha^{(t)})$	$s^{(t)}q^{(t)}$	$\Delta(\alpha^{(t)})$	$s^{(t)}q^{(t)}$	$\Delta(\alpha^{(t)})$	$q^{(t)}$
1	43.5351	–	43.5351	–	43.5351	–
2	14.3239	0.9778	17.6348	1.500	14.4074	1.000
3	14.3004	1.2705 ^a	17.1446	1.343 ^a	6.1867	1.000
4	1.8835	1.0718	6.1626	1.500	3.2703	1.000
5	1.6772	2.1539 ^a	2.9889	1.500	1.3871	1.000
6	0.0493	1.1411	0.5008	1.500	0.4884	1.000
7	0.0011	1.4118	0.1192	1.500	0.1586	1.000
8	0.0000	1.8821	0.0219	1.500	0.0524	1.000
9	0.0000	1.2487	0.0045	1.500	0.0182	1.000
10	0.0000	1.6794	0.0009	1.500	0.0066	1.000
11	0.0000	1.2724	0.0002	1.500	0.0024	1.000
12	0.0000	1.6732	0.0000	1.500	0.0009	1.000
13	–	–	0.0000	1.500	0.0003	1.000
14	–	–	0.0000	1.500	0.0001	1.000
15	–	–	0.0000	1.500	0.0000	1.000
16	–	–	0.0000	1.500	0.0000	1.000
17	–	–	0.0000	1.500	0.0000	1.000
18	–	–	0.0000	1.500	0.0000	1.000
19	–	–	–	–	0.0000	1.000
20	–	–	–	–	0.0000	1.000
21	–	–	–	–	0.0000	1.000
22	–	–	–	–	0.0000	1.000
23	–	–	–	–	0.0000	1.000
24	–	–	–	–	0.0000	1.000

^aIndicates that the steplength was adjusted from a starting value

vergent point. It is confirmed that *AIFS* converges faster than *IFS*, whilst *IFS* is faster than *EM* for the whole sequence.

The standard error can be obtained through the IFS-Variance method. In this example, the standard error of the estimates for $(\pi, \alpha_1, \alpha_2)$ is obtained in an explicit form by taking the second derivative of $\ell_{\text{obs}}(\alpha)$. The sum of squares of the difference of all elements between the former and the latter is less than 10^{-6} . Thus, the IFS-Variance method again works well.

5.3 Multivariate *t*-distribution

Multivariate *t*-distributions are often used for data analysis, especially in the context of robust statistics (McLachlan and Peel 2000). Let the observed data, $\mathbf{y}_1, \dots, \mathbf{y}_n$, follow *p*-variate *t*-distribution $t_p(\mu, \Sigma, v)$, where μ is a location vector, Σ is a scale matrix, and *v* is known degrees of freedom. The log-likelihood function of this *t*-distribution with $\mathbf{y}_i$ ($i = 1, \dots, n$) is denoted by $\ell_{\text{obs}}(\alpha)$. In the following, the parameters, μ and the nonredundant elements in Σ , to be estimated are written as α . To estimate α , it is assumed that the complete data are $(\mathbf{y}_i, z_i)$ ($i = 1, \dots, n$) and that

$$\mathbf{y}_i | z_i \sim N_p(\mu, z_i^{-1} \Sigma), \quad z_i \sim v^{-1} \chi_v^2,$$

where $N_p(\mu, z_i^{-1} \Sigma)$ is a normal distribution with mean μ and variance $v^{-1} \Sigma$, and χ_v^2 is a chi-squared distribution with *v* degrees of freedom. The complete-data log-likelihood function is

$$\ell_{\text{com}}(\alpha) \propto -\frac{n}{2} \log |v^{-1} \Sigma| - \sum_{i=1}^n \frac{1}{2} (\mathbf{y}_i - \mu)' (z_i^{-1} \Sigma)^{-1} (\mathbf{y}_i - \mu).$$

The (A)IFS method is derived as follows. The first derivative of $\ell_{\text{obs}}(\alpha^{(t)})$ is

$$\begin{aligned} \nabla \ell_{\text{obs}}(\alpha^{(t)}) &= \sum_{i=1}^n \frac{v+p}{v+h_i^{(t)}} \left(\frac{\Sigma^{(t)-1} (\mathbf{y}_i - \mu^{(t)})}{0.5 \text{vec}((\Sigma^{(t)})^{-1} (\mathbf{y}_i - \mu^{(t)}) (\mathbf{y}_i - \mu^{(t)})' (\Sigma^{(t)})^{-1})} \right) \\ &\quad - \sum_{i=1}^n \left(\frac{\mathbf{0}}{0.5 \text{vec}((\Sigma^{(t)})^{-1})} \right), \end{aligned}$$

where $h_i^{(t)} = (\mathbf{y}_i - \mu^{(t)})' (\Sigma^{(t)})^{-1} (\mathbf{y}_i - \mu^{(t)})$. To derive $\nabla \ell_{\text{obs}}(\alpha^{(t)})$, $\nabla^{10} Q(\alpha^{(t)} | \alpha^{(t)})$ is used, since it is easier to calculate. The inverse matrix of the expected second derivative is

Fig. 1 Convergence comparison for *AIFS*, *IFS*, and *EM* using mortality data

$$J_{\text{com}}(\alpha)^{-1} = \begin{pmatrix} \Sigma & O \\ O & 2\Sigma \otimes \Sigma \end{pmatrix}.$$

Thus, the iteration equation of the IFS method is given as

$$\begin{aligned} \mu^{(t+1)} &= \mu^{(t)} + q_1^{(t)} \frac{1}{n} \sum_{i=1}^n \frac{v+p}{v+h_i^{(t)}} (y_i - \mu^{(t)}), \\ \Sigma^{(t+1)} &= \Sigma^{(t)} + q_2^{(t)} \frac{1}{n} \sum_{i=1}^n \\ &\quad \times \left\{ \frac{v+p}{v+h_i^{(t)}} (y_i - \mu^{(t)})(y_i - \mu^{(t)})' - \Sigma^{(t)} \right\}, \end{aligned}$$

where $q_1^{(t)}$ and $q_2^{(t)}$ indicate the different steplengths.

The IFS method with

$$q_1^{(t)} = n \left(\sum_{i=1}^n \frac{v+p}{v+h_i^{(t)}} \right)^{-1} \quad \text{and} \quad q_2^{(t)} = 1 \quad (8)$$

becomes the iteration of the EM algorithm. These steplengths can be obtained in the same method as in the mixture of Poisson distributions. Let d_1 be the direction for μ . The equation that produces the exact optimum steplength is

$$\frac{d}{dq_1^{(t)}} \ell_{\text{obs}}(\mu^{(t)} + q_1^{(t)} d_1^{(t)}, \Sigma^{(t)}) = 0,$$

This is difficult to solve, thus the following is solved instead:

$$\sum_{i=1}^n \frac{v+p}{v+h_i^{(t)}} d_1^{(t)'} \left(\Sigma^{(t)} \right)^{-1} (y_i - \mu^{(t)} - q_1^{(t)} d_1^{(t)}) = 0,$$

where $h_i^{(t)}$ is fixed at the current estimate, although it is actually a function of $\alpha^{(t)}$. This leads to $q_1^{(t)}$ of Eq. (8). In a similar way for $q_2^{(t)}$, note $q_2^{(t)} = 1$. The fixation is justified when the direction vector according to the IFS method is sufficiently close to zero, that is, when $\alpha^{(t)}$ is near the convergent point. Toward the end, the directions by the EM algorithm and the IFS method are almost identical.

The IFS method with

$$q_1^{(t)} = q_2^{(t)} = n \left(\sum_{i=1}^n \frac{v+p}{v+h_i^{(t)}} \right)^{-1}$$

becomes the iteration of the parameter-expansion EM algorithm (PX-EM algorithm; Liu et al. 1998). This was developed to accelerate the convergence of the EM algorithm. The IFS method with the common steplength (i.e., the PX-EM algorithm) is expected to be faster than that with different steplengths (i.e., the EM algorithm). The problem is how this steplength can be improved.

Here, these four methods are compared: the AIFS method, the IFS method with the initial value $q^{(t)} = 1.2$, the PX-EM algorithm, and the EM algorithm. These are referred to as *AIFS*, *IFS*, *PXEM*, and *EM*, respectively. To compare these methods, data of size $n = 100$ are randomly generated from the multivariate t -distribution with the location parameter $\mu = (1, 2, 3)'$ and the covariance matrix

$$\Sigma = \begin{pmatrix} 1.0 & 0.9 & 0.5 \\ 0.9 & 1.0 & 0.3 \\ 0.5 & 0.3 & 1.0 \end{pmatrix}$$

with six degrees of freedom. For all four computation methods, the initial values are the moment estimates, and the convergence criterion is $\|\nabla \ell_{\text{obs}}(\alpha^{(t)})\| < 10^{-4}$.

Table 2 shows the process of convergence across the four methods in terms of $\Delta(\alpha^{(t)})$ and the steplength chosen to improve the current value at each iteration. Empty cells, “–”, occur for the same reason as per the foregoing normal distribution example. All methods attain the same maximum value of the log-likelihood function. All values are rounded to four decimal places; thus, again, 0.0000 for $\Delta(\alpha^{(t)})$ denotes a rounded zero. The number of iterations until convergence differ across the methods as follows: 10, 15, 11, and 19 for *AIFS*, *IFS*, *PXEM*, and *EM*, respectively. Neither *AIFS* or *IFS* required any steplength adjustments. *AIFS* and *PXEM* require almost the same number of iterations until convergence. From the first to the fourth iteration, *AIFS* is closer to the convergent point than *IFS* and *EM*, but farther than *PXEM*. After the fifth iteration, *AIFS* begins to get closer to the convergent point than *PXEM*. This indicates that *AIFS* improves itself more slowly than *PXEM* in the initial iterations, but more quickly in subsequent iterations. In this

Table 2 Convergence comparison for *AIFS*, *PXEM*, *IFS*, and *EM* using randomly generated data

Iteration	<i>AIFS</i>		<i>IFS</i>		<i>PXEM</i>		<i>EM</i>	
	$\Delta(\alpha^{(t)})$	$s^{(t)}q^{(t)}$	$\Delta(\alpha^{(t)})$	$s^{(t)}q^{(t)}$	$\Delta(\alpha^{(t)})$	$q_1^{(t)}$	$\Delta(\alpha^{(t)})$	$q_1^{(t)}$
1	6.0092	–	6.0092	–	6.0092	–	6.0092	–
2	0.8785	2.0814	0.5194	1.2000	0.1895	0.9240	1.0814	0.9240
3	0.0231	1.6278	0.0562	1.2000	0.0071	0.9858	0.2096	0.9655
4	0.0010	1.5104	0.0064	1.2000	0.0003	0.9974	0.0418	0.9844
5	0.0000	1.3876	0.0007	1.2000	0.0000	0.9995	0.0084	0.9930
6	0.0000	1.4576	0.0001	1.2000	0.0000	0.9999	0.0017	0.9968
7	0.0000	1.3661	0.0000	1.2000	0.0000	1.0000	0.0003	0.9986
8	0.0000	1.4421	0.0000	1.2000	0.0000	1.0000	0.0001	0.9994
9	0.0000	1.3583	0.0000	1.2000	0.0000	1.0000	0.0000	0.9997
10	0.0000	1.4373	0.0000	1.2000	0.0000	1.0000	0.0000	0.9999
11	–	–	0.0000	1.2000	0.0000	1.0000	0.0000	0.9999
12	–	–	0.0000	1.2000	–	–	0.0000	1.0000

example, it is confirmed that *AIFS* converges almost equivalently to *PXEM*, that they are faster than *IFS*, and that *IFS* is faster than *EM*, even for the whole sequence. Although in the application of *PXEM* an extended model is needed that contains the original model as a special case, *AIFS* can be applied to cases where such an extended model cannot be found. In those cases, *AIFS* can substitute for *PXEM*.

5.4 Dirichlet distribution

The (A)IFS method is applied for the parameter estimation of a Dirichlet distribution. The Dirichlet distribution has no explicit form of the iteration for the EM algorithm since it involves the gamma function and, in practical cases, its derivative. The EM algorithm must resort to the Newton–Raphson type iterations, losing its simplicity. On the other hand, the (A)IFS method could compute the parameter without resorting to such a computational method.

Let $\mathbf{y} = (y_1, \dots, y_p)'$ be a data vector taken independently from the Dirichlet distribution with density function,

$$\frac{\Gamma(\sum_{j=1}^p \alpha_j)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \dots \Gamma(\alpha_p)} y_1^{\alpha_1-1} y_2^{\alpha_2-1} \dots y_p^{\alpha_p-1}, \quad (9)$$

where $(\alpha_1, \dots, \alpha_p)$ is the parameter. The Dirichlet random vector is derived by transforming the independent random variables from gamma distributions. Assume that $x_1, \dots, x_p$ are independently positive random variables and each x_j is distributed as a gamma distribution with parameter α_j ($j = 1, \dots, p$). By setting

$$\frac{x_j}{\sum_{j=1}^p x_j} = y_j, \quad j = 1, \dots, p,$$

vector $(y_1, \dots, y_p)$ is distributed as Eq. (9) (Mosimann 1962). $(y_1, \dots, y_p)$ can be regarded as incomplete data and

$(x_1, \dots, x_p)$ as complete data for the estimation of parameters $(\alpha_1, \dots, \alpha_p)$.

Assume that we have data of sample size n taken independently from the Dirichlet distribution, $\mathbf{y}_1, \dots, \mathbf{y}_n$ with $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$. The observed-data log-likelihood function is

$$\begin{aligned} \ell_{\text{obs}}(\alpha) = & n \log \Gamma \left(\sum_{j=1}^p \alpha_j \right) - n \sum_{j=1}^p \log \Gamma(\alpha_j) \\ & + \sum_{j=1}^p (\alpha_j - 1) \sum_{i=1}^n \log y_{ij}, \end{aligned}$$

where $\alpha = (\alpha_1, \dots, \alpha_p)'$. The complete-data log-likelihood function is, ignoring an irrelevant term,

$$\ell_{\text{com}}(\alpha) = -n \sum_{j=1}^p \log \Gamma(\alpha_j) + \sum_{j=1}^p (\alpha_j - 1) \sum_{i=1}^n \log x_{ij}.$$

As mentioned earlier, applying the EM algorithm would produce an inexplicit form of the iteration. In fact, the EM algorithm leads to

$$\begin{aligned} Q(\alpha|\alpha^{(t)}) = & -n \sum_{j=1}^p \log \Gamma(\alpha_j) + \sum_{j=1}^p (\alpha_j - 1) \\ & \sum_{i=1}^n E[\log x_{ij} | \mathbf{y}_i; \alpha^{(t)}]. \end{aligned}$$

This equation cannot be explicitly solved with respect to α since the derivation of $\Gamma(\alpha_j)$ is involved. To derive the M-step, we need to resort to the Newton–Raphson type iterations. In this case, use of the EM algorithm would not bring benefits such as simplicity.

Fig. 2 Convergence comparison for *AIFS*, and *IFS* using duckling diet data. EMG stands for the EM gradient algorithm, which is equivalent to *IFS*

On the other hand, the (A)IFS method does not solve the first derivative of the Q -function. To apply the (A)IFS method, we need a direction vector, $\mathbf{d} = n^{-1} \mathbf{J}_{\text{com}}(\boldsymbol{\alpha})^{-1} \nabla \ell_{\text{obs}}(\boldsymbol{\alpha})$. First, we compute the Fisher information matrix for a single datum.

$$\mathbf{J}_{\text{com}}(\boldsymbol{\alpha}) = -\frac{1}{n} E \left[\frac{\partial \ell_{\text{com}}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \right] = \text{diag}(\Psi_1(\alpha_1), \dots, \Psi_1(\alpha_p)),$$

where $\Psi_1(\alpha_j)$ is the first derivative of a digamma function $\Psi(\alpha_j)$ with respect to α_j . Next, we compute the first derivative of the observed-data log-likelihood function.

$$\begin{aligned} \nabla \ell_{\text{obs}}(\boldsymbol{\alpha}) &= n \text{diag} \left(\Psi_1 \left(\sum_{j=1}^p \alpha_j \right) \right) \\ &\quad - n \text{diag} (\Psi_1(\alpha_1), \dots, \Psi_1(\alpha_p)) + \sum_{i=1}^n \log y_{ij}. \end{aligned}$$

As a result, the direction vector has its j th element as

$$\frac{\Psi_1 \left(\sum_{k=1}^p \alpha_k \right)}{\Psi_1(\alpha_j)} - 1 + \frac{\sum_{k=1}^n y_{kj}/n}{\Psi_1(\alpha_j)}.$$

In this case, the iteration becomes identical to that of the EM gradient algorithm (Lange 1995a) except for steplength.

The AIFS method is compared with the IFS method for the data from Mosimann (1962). The data represent the proportions of the diet ingredients for ducklings. The data have its size $n = 23$, and $p = 3$ variables. The same data are used in the paper about the EM gradient algorithm by Lange

(1995a). The EM gradient algorithm has the same form of the iteration as that of the IFS method shown in Eq. (1). Both the EM gradient algorithm and the IFS method have an arbitrary steplength. I use the steplength of two, because Lange (1995a) suggests the steplength of two approximately halves the number of iterations until convergence for the EM gradient algorithm. Hereafter, the AIFS method, and the IFS method with the steplength of two are referred to as *AIFS*, and *IFS*, respectively.

Starting from the initial value (1.0., 1.0, 1.0), *AIFS* and *IFS* converge to the same value of the observed-data log-likelihood function 73.1250 with the same estimates (3.22, 20.38, 21.69), which is set to be $\boldsymbol{\alpha}^*$ for later use. The convergence criterion is $\|\nabla \ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)})\| < 10^{-4}$ as before. As a result, the numbers of iterations until convergence are 193 when *AIFS* is applied, and 735 when *IFS* is applied. The process of convergence to the 50th iteration is shown in Fig. 2. The horizontal axis indicates the iteration number to the 50th iteration and the vertical axis is $\Delta(\boldsymbol{\alpha}^{(t)}) = |\ell_{\text{obs}}(\boldsymbol{\alpha}^{(t)}) - \ell_{\text{obs}}(\boldsymbol{\alpha}^*)|$. Notice that both methods do not converge yet in the graph. During initial iterations, *IFS* is closer to the convergent point than *AIFS*, and after those initial iterations, *AIFS* becomes closer. This example illustrates that *AIFS* and *IFS* work for models for which the EM algorithm could not be used without resorting to numerical methods, and, on the other hand, that *AIFS* and *IFS* treat such models without counting on the other computational methods. It also shows, with the data of diet ingredients for ducklings, that *IFS*, which is equivalent to the EM gradient algorithm, can be drastically improved to *AIFS* which appropriately adjusts the steplength at each iteration.

5.5 Computing time

In the previous four subsections, from the perspective of the number of iterations until convergence, the (A)IFS was compared with the EM algorithm and its variants. The theorems given above also mainly concern the number of iterations. Another interesting perspective is the computing time. In this subsection, we make the comparison among the AIFS method, the IFS method, the PX-EM algorithm, and the EM algorithm in terms of the computing time using a simulation technique.

We conduct four simulations corresponding to each example above. In each simulation, the same set of data as each example is used with the same initial values to make the same comparison. We make the parameter estimation 1000 times and measure the computing time with the unit of millisecond, obtaining the mean and standard deviation of the computing times. Note that although the multivariate t -distribution uses a randomly generated set of data, the simulation uses a fixed set of data as in Sect. 5.3.

Table 3 Computing time (millisecond) and its standard deviation in parentheses. Note that “dist.” stands for “distribution,” and all the values are rounded

Example	<i>AIFS</i>	<i>IFS</i>	<i>PXEM</i>	<i>EM</i>
Normal dist.	0.4168	0.1115	–	0.0697
In Sect. 5.1	(0.0114)	(0.0030)	–	(0.0020)
Poisson dist.	8.8863	64.4989	–	31.4013
In Sect. 5.2	(0.2183)	(1.4783)	–	(0.7543)
t-dist.	0.2397	0.2464	0.2389	0.3620
In Sect. 5.3	(0.0109)	(0.0061)	(0.0064)	(0.0085)
Dirichlet dist.	0.3292	1.0391	–	–
in Sect. 5.4	(0.0086)	(0.0184)	–	

At each iteration, the (A)IFS method takes more computation than the EM algorithm does. This is because the (A)IFS uses steplength adjustment, that is, the (A)IFS gives an initial steplength and adjusts it if necessary. Thus, if the numbers of iterations until convergence do not vary between the (A)IFS method and the EM algorithm for the same data, the (A)IFS method should take a longer computing time than the EM algorithm. However, this is not the case: the number of iterations until convergence varies between the (A)IFS method and the EM algorithm. As all the examples show, the (A)IFS method takes a smaller number of iterations until convergence. Although the computational amount due to use of the (A)IFS method increases at each iteration, the total computing time should be shorter under the condition that the number of iterations until convergence becomes small enough. Thus, the moderate decrease in the number of iterations due to use of the (A)IFS method would offset the increase in the computational amount due to use of the (A)IFS method, resulting in no change in the total computing time. The large decrease in the number of iterations would shorten the total computing time.

Table 3 shows the result of the simulation. In the table, *AIFS*, and *IFS* denote the AIFS method, the IFS with the initial steplength of 1.5 for the normal distribution example and 2 for the Poisson mixture distribution example and the Dirichlet distribution example, and 1.2 for the multivariate *t*-distribution example. *PXEM*, and *EM* denote the parameter-expansion EM algorithm, and the EM algorithm, respectively. The mean of computing times is given with its standard deviation in parentheses. All values are rounded to four decimal places.

In the normal distribution example, the number of iterations decrease from 24 of *EM* to 17 of *IFS*, and to 12 of *AIFS*. The decrease in the number of iterations is not large enough to offset the increase in the computation amount due to use of *IFS* and *AIFS*. The computing time order may almost reflect the computational amount order. *AIFS* needs more computational amount than *IFS* to compute the initial steplength and

adjust it, and *IFS* needs more than *EM* to adjust the initial steplength of 1.5.

In the rest of the simulation result, *AIFS* always takes the shortest computing time. This reflects that the number of iterations due to use of *AIFS* is small enough, and the computational amount due to *AIFS* increases a little. *AIFS* uses a steplength of a theoretically good initial value, and the selected steplength needs almost no adjustment. *IFS* outperforms *EM* in the multivariate *t*-distribution example, and not in the normal distribution example and the Poisson mixture distribution example. The reason might be that *IFS* uses a steplength of a guessed initial value, and the steplength needs its appropriate adjustment if any.

The result as a whole indicates that the AIFS method works better in terms of computing time as well as the number of iterations. Taking into consideration that the AIFS method can compute the standard error of the estimates without additional programs, the AIFS method is all the more favorable than the IFS method, the EM algorithm, and its variants.

A practical problem in deciding whether to use the AIFS method is that it is unknown before executing the actual computation which of the AIFS method, the IFS method, or the EM algorithm takes the shortest computing time. However, it is beyond the scope of this paper and it will be dealt with in a future paper.

6 Conclusion

In this paper, the IFS method and its accelerated variant, the AIFS method, are proposed. These can be seen as Fisher scoring methods modified for incomplete data, and also as scaled versions of the steepest ascent method (Bertsekas 2004). The noteworthy feature of the IFS method is that it becomes nearly or exactly identical to the EM algorithm, but, at the same time, can be accelerated through steplength adjustment. With the first derivative of $\ell_{\text{obs}}(\alpha)$ for the (A)IFS method, the second derivative is obtained by numerical differentiation, which produces the standard error of the MLE.

It is possible to further modify this proposed method. Observing the process of convergence in the examples provided herein, focusing on the normal distribution and the multivariate *t*-distribution, and switching from the EM algorithm at initial stages to the IFS or AIFS methods at later stages is expected to result in less iterations being required before convergence. This is a kind of “hybrid EM algorithm” (McLachlan and Krishnan 2008). The EM algorithm, at initial stages, can be replaced with the IFS with a fixed $s^{(t)}q^{(t)} = 1$ because it is exactly or nearly identical to the EM algorithm. The advantage of this hybrid method is its smooth transition, achieved by steplength changes. Using the IFS iterations, this transition can be applied to the Dirichlet distribution example as well. As shown in Fig. 2, switching

from the IFS with a certain steplength to the AIFS is expected to result in faster convergence. However, observing the convergence process in the Poisson mixture example in Fig. 1, switching might not be effective for reducing the number of iterations in some cases.

Another modification is the memory-saved IFS method, which is expected to be generally effective if properly employed. Computing the inverse matrix of $J_{\text{com}}(\alpha)$ at every iteration could be computationally onerous, and might cause a longer computing time. To avoid the onerous and long computation, after some iterations, $J_{\text{com}}(\alpha)$ is fixed at its newest value. This is called the “Whittaker’s method” for the Newton–Raphson algorithm (Small and Wang 2003). However, it is unknown beforehand whether to use this method or not and, if the researcher does opt to use this method, when to switch to the AIFS method and when to fix $J_{\text{com}}(\alpha)^{-1}$ is unclear. These are problems to be explored in the future.

This paper is just a starting point for the development of the (A)IFS. This paper focuses on the theoretical aspects of the (A)IFS, namely the convergence theorem (Theorem 1) and the convergence rate (Theorem 2). Beyond the theoretical results, users are likely to be more interested in knowing how much faster the (A)IFS is compared to the EM algorithm in empirical contexts. These practical aspects also need further exploration in the future.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

Equivalence between equations (3) and (4) Define $S = (0, 1)'$. Equation (2) can be written as

$$\begin{aligned}\ell_{\text{obs}}(\alpha) = & -\frac{m}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^m (y_i - \mu) \Sigma^{-1} (y_i - \mu) \\ & - \frac{n-m}{2} \log |S' \Sigma S| \\ & - \frac{1}{2} \sum_{i=m+1}^n (y_{2i} - \mu_2) (S' \Sigma S)^{-1} (y_{2i} - \mu_2)\end{aligned}$$

Taking the derivative with respect to $\text{vec}(\Sigma)$ leads to

$$\begin{aligned}\frac{\partial \ell_{\text{obs}}(\alpha)}{\partial \text{vec}(\Sigma)} = & \frac{1}{2} \sum_{i=1}^m \text{vec} \left(\Sigma^{-1} (y_i - \mu) (y_i - \mu)' \Sigma^{-1} - \Sigma^{-1} \right) \\ & + \frac{1}{2} \sum_{i=m+1}^n \text{vec} \left(\frac{(y_{2i} - \mu_2)^2}{(S' \Sigma S)^2} S S' - S (S' \Sigma S)^{-1} S' \right).\end{aligned}$$

Using the relationship that $\text{vec}(ABC) = (C' \otimes A) \text{vec}(B)$ for matrices A , B , and C of appropriate sizes, we have

$$\begin{aligned}\text{vec}(\Sigma) + \frac{2}{n} (\Sigma \otimes \Sigma) \frac{\partial \ell_{\text{obs}}(\alpha)}{\partial \text{vec}(\Sigma)} = & \frac{1}{n} \sum_{i=1}^m \text{vec}((y_i - \mu)(y_i - \mu)') \\ & + \frac{1}{n} \sum_{i=m+1}^n (y_{2i} - \mu_2)^2 \text{vec} \begin{pmatrix} (\sigma_{12}/\sigma_{22})^2 \sigma_{12}/\sigma_{22} & \\ \sigma_{21}/\sigma_{22} & 1 \end{pmatrix} \\ & + \frac{1}{n} \sum_{i=m+1}^n \begin{pmatrix} \sigma_{11} - (\sigma_{12}^2/\sigma_{22}) & 0 \\ 0 & 0 \end{pmatrix}.\end{aligned}$$

Apparently, this is equivalent to the right-hand side in Eq. (3).

Proof of $E[\nabla^2 \ell_{\text{obs}}(\alpha^{(t)})] \geq -n J_{\text{com}}(\alpha^{(t)})$ We suppress the iteration number of $\alpha^{(t)}$ for convenience. The complete-data log-likelihood function is decomposed into

$$\ell_{\text{com}}(\alpha) = \ell_{\text{obs}}(\alpha) + \ell_{\text{miss}|\text{obs}}(\alpha),$$

where $\ell_{\text{miss}|\text{obs}}(\alpha)$ denotes the distribution of missing data conditional on the observed data. Taking the second derivative and the expectation of both sides gives,

$$E[\nabla^2 \ell_{\text{com}}(\alpha)] = E[\nabla^2 \ell_{\text{obs}}(\alpha)] + E[\nabla^2 \ell_{\text{miss}|\text{obs}}(\alpha)], \quad (10)$$

where E is the expectation with respect to the joint distribution of random variables for the data and the corresponding missing-data indicators (Kenward and Molenberghs 1998; Seaman et al. 2013). The second term on the right side is

$$E[\nabla^2 \ell_{\text{miss}|\text{obs}}(\alpha)] = E_{\text{obs}} \left[E_{\text{miss}|\text{obs}} \left[\nabla^2 \ell_{\text{miss}|\text{obs}}(\alpha) \right] \right],$$

where $E_{\text{miss}|\text{obs}}$ indicates the expectation of the missing-data variables conditional on the observed-data variables and E_{obs} is the expectation of the observed-data variables. Since $-E_{\text{miss}|\text{obs}}[\nabla^2 \ell_{\text{miss}|\text{obs}}(\alpha)]$ is the Fisher information matrix, $E[\nabla^2 \ell_{\text{miss}|\text{obs}}(\alpha)] \leq O$ in the sense that $x' E[\nabla^2 \ell_{\text{miss}|\text{obs}}(\alpha)] x \leq 0$ for any value of vector x . Therefore, with Eq. (10), $E[\nabla^2 \ell_{\text{com}}(\alpha)] \leq E[\nabla^2 \ell_{\text{obs}}(\alpha)]$, that is, $-n J_{\text{com}}(\alpha^{(t)}) \leq E[\nabla^2 \ell_{\text{obs}}(\alpha^{(t)})]$.

Proof of Corollary 1 By the concavity, $\ell_{\text{obs}}(\alpha) - \ell_{\text{obs}}(\alpha^*) \leq \nabla \ell_{\text{obs}}(\alpha^*)'(\alpha - \alpha^*)$ (See Proposition B.3 of Bertsekas 2004). Since $\nabla \ell_{\text{obs}}(\alpha^*) = \mathbf{0}$, the statement holds. The relation $\ell_{\text{obs}}(\alpha) < \ell_{\text{obs}}(\alpha^*)$ directly follows from strict concavity. $\square$

Lemma 1 Under the same setting as Theorem 2,

- (i) For IFS, there exists a sufficiently large t_1 such that $s^{(t)}q^{(t)} \geq 1$ ($t \geq t_1$).
- (ii) For AIFS, there exists a sufficiently large t_2 such that $s^{(t)} = 1$ ($t \geq t_2$).

Proof (i) To simplify notation, iteration numbers are again suppressed in the proof. To prove the lemma, it suffices to show that $s^{(t)}q^{(t)} = 1$ is accepted by the Armijo Rule.

$$\begin{aligned} \ell_{\text{obs}}(\alpha + d) - \ell_{\text{obs}}(\alpha) &= \nabla \ell_{\text{obs}}(\alpha)'d + \frac{1}{2}d'\nabla^2 \ell_{\text{obs}}(\alpha)d + o(\|d\|^2) \\ &= -\frac{n}{2}d'(I_{\text{obs}}(\alpha) - J_{\text{com}}(\alpha))d + \frac{1}{2}d'\nabla \ell_{\text{obs}}(\alpha) + o(\|d\|^2). \end{aligned}$$

Define $y = I_{\text{obs}}(\alpha)^{\frac{1}{2}}d$. Since $J_{\text{com}}(\alpha)$ and $I_{\text{obs}}(\alpha)$ are continuous with respect to α ,

$$\begin{aligned} d'(I_{\text{obs}}(\alpha) - J_{\text{com}}(\alpha))d &= y'(I - I_o^{-\frac{1}{2}}J_cI_o^{-\frac{1}{2}})y + o(\|d\|^2) \\ &< \left(1 - \frac{1}{\lambda_1}\right)\|y\|^2 + o(\|d\|^2). \end{aligned}$$

Thus, there exists a sufficiently large t_1 such that

$$\begin{aligned} \ell_{\text{obs}}(\alpha + d) - \ell_{\text{obs}}(\alpha) &> \frac{n}{2}d'J_{\text{com}}(\alpha)d + o(\|d\|^2) \\ &> \sigma d'\nabla \ell_{\text{obs}}(\alpha) + o(\|d\|^2), \end{aligned}$$

for $\forall t \geq t_1$.

(ii) As $\|d\| \rightarrow 0$, $q^{(t)} \rightarrow q_0$. Thus, it suffices to show $\ell_{\text{obs}}(\alpha + q_0d) - \ell_{\text{obs}}(\alpha) > \sigma q_0d'\nabla \ell_{\text{obs}}(\alpha)$ when t is sufficiently large. In the following, iteration numbers are again suppressed. In the same way as (i),

$$\begin{aligned} \ell_{\text{obs}}(\alpha + q_0d) - \ell_{\text{obs}}(\alpha) &= q_0\nabla \ell_{\text{obs}}(\alpha)'d + \frac{q_0^2}{2}d'\nabla^2 \ell_{\text{obs}}(\alpha)d + o(\|d\|^2) \\ &= \frac{q_0}{2}d'\nabla \ell_{\text{obs}}(\alpha) + o(\|d\|^2) \\ &> \sigma q_0d'\nabla \ell_{\text{obs}}(\alpha) + o(\|d\|^2), \end{aligned}$$

since $\sigma \in (0, .5)$. Because $\|d\| \rightarrow 0$ as $t \rightarrow \infty$, the statement follows.

Proof of Theorem 2 Define $\ell_t = \ell_{\text{obs}}(\alpha^{(t)})$ and $I_{\text{obs}}(\alpha^*) = I_o$ for convenience. Applying the Taylor expansion produces

$$\nabla \ell_t = -nI_o(\alpha^{(t)} - \alpha^*) + o(\|\alpha^{(t)} - \alpha^*\|).$$

Since the direction produced by the IFS method is given as $d^{(t)} = n^{-1}J_{\text{com}}(\alpha^{(t)})^{-1}\nabla \ell_t$,

$$d^{(t)} = -J_{\text{com}}(\alpha^{(t)})^{-1}I_o(\alpha^{(t)} - \alpha^*) + o(\|\alpha^{(t)} - \alpha^*\|).$$

Dividing both sides by $\|\alpha^{(t)} - \alpha^*\|$, for sufficiently large t ,

$$g^{(t)} = -J_c^{-1}I_o e^{(t)} + o(1),$$

where $g^{(t)} = d^{(t)}/\|\alpha^{(t)} - \alpha^*\|$, $J_c = J_{\text{com}}(\alpha^*)$, and $e^{(t)} = (\alpha^{(t)} - \alpha^*)/\|\alpha^{(t)} - \alpha^*\|$. Since $e^{(t)}$ is a unit vector and $\|J_c^{-1}I_o\|$ is nonzero and lower- and upper-bounded, $g^{(t)}$ on the left-side term is also nonzero and bounded.

Since $\nabla \ell_{\text{obs}}(\alpha^*) = 0$, we have

$$\begin{aligned} \frac{\Delta(\alpha^{(t)})}{\|\alpha^{(t)} - \alpha^*\|^2} &= \frac{n}{2}e^{(t)'}I_o e^{(t)} + o(1) \\ &= \frac{n}{2}g^{(t)'}J_cI_o^{-1}J_cg^{(t)} + o(1). \end{aligned}$$

In a similar way, we have

$$\begin{aligned} \frac{\Delta(\alpha^{(t+1)}) - \Delta(\alpha^{(t)})}{\|\alpha^{(t)} - \alpha^*\|^2} &= \frac{-\ell_{t+1} + \ell_t}{\|\alpha^{(t)} - \alpha^*\|^2} \\ &= -\frac{(s^{(t)}q^{(t)})\nabla \ell_t'g^{(t)}}{\|\alpha^{(t)} - \alpha^*\|} \\ &\quad + \frac{(s^{(t)}q^{(t)})^2n}{2}g^{(t)'}I_{\text{obs}}(\alpha^{(t)})g^{(t)} + o(1) \\ &= -(s^{(t)}q^{(t)})ng^{(t)'}J_cg^{(t)} \\ &\quad + \frac{(s^{(t)}q^{(t)})^2n}{2}g^{(t)'}I_o g^{(t)} + o(1). \end{aligned}$$

(i) By Lemma 1, there exists a sufficiently large t_1 such that $(s^{(t)}q^{(t)}) \geq 1$ for $\forall t \geq t_1$. Thus, again maintaining simple notation, let $k = s^{(t)}q^{(t)}$,

$$\begin{aligned} \frac{\Delta(\alpha^{(t+1)}) - \Delta(\alpha^{(t)})}{\Delta(\alpha^{(t)})} &= \frac{-ky'I_o^{\frac{1}{2}}J_c^{-1}J_o^{\frac{1}{2}}y + k^22^{-1}y'\left(J_o^{\frac{1}{2}}J_c^{-1}J_o^{\frac{1}{2}}\right)^2y}{2^{-1}\|y\|^2} + o(1). \end{aligned} \quad (11)$$

There exists an orthogonal matrix P satisfying $P'(I_o^{\frac{1}{2}}J_c^{-1}I_o^{\frac{1}{2}})P = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$, where λ_j is the

eigenvalues of $J_c^{-1}I_o$ ($j = 1, \dots, r$). Letting $\mathbf{x} = P\mathbf{y} = (x_1, \dots, x_p)'$, we have

$$\frac{\Delta(\boldsymbol{\alpha}^{(t+1)}) - \Delta(\boldsymbol{\alpha}^{(t)})}{\Delta(\boldsymbol{\alpha}^{(t)})} = \frac{1}{\|\mathbf{x}\|^2} \sum_{j=1}^p (k^2\lambda_j^2 - 2k\lambda_j)x_j^2 + o(1) \\ < -1 + (1 - k\lambda_r)^2 + o(1),$$

where we have used the inequality, $k^2\lambda_j^2 - 2k\lambda_j \leq k^2\lambda_r^2 - 2k\lambda_r$ for $\forall j \in \{1, \dots, r\}$. We have the desired result.

(ii) By Lemma 1, there exists a sufficiently large t_2 such that $s^{(t)} = 1$. Thus,

$$s^{(t)}q^{(t)} = q^{(t)} = \frac{\mathbf{g}^{(t)'}J_c^{-1}\mathbf{g}^{(t)}}{\mathbf{g}^{(t)'}I_o\mathbf{g}^{(t)}} + o(1)$$

for $\forall t \geq t_2$. Setting $J_c^{\frac{1}{2}}\mathbf{g}^{(t)} = \mathbf{z}$, Eq. (11) becomes

$$-\frac{\|\mathbf{z}\|^4}{(\mathbf{z}'J_c^{\frac{1}{2}}I_o^{-1}J_c^{\frac{1}{2}}\mathbf{z})(\mathbf{z}'J_c^{-\frac{1}{2}}I_oJ_c^{-\frac{1}{2}}\mathbf{z})} \leq -\frac{4\lambda_1\lambda_r}{(\lambda_1 + \lambda_r)^2},$$

by Kantorovich's inequality (Lemma 3.1 in Bertsekas 2004). Therefore, we have the desired result. $\square$

Proof of Corollary 2 (i) Setting $\Delta(\boldsymbol{\alpha}^{(t)}) = .5(\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}^*)'I_o(\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}^*) = .5\|\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}^*\|_{I_o}^2$ in Theorem 2, we have the desired result.

(ii) Proved similarly to (i). $\square$

References

- Bentler, P.M., Tanaka, J.S.: Problems with EM algorithms for ML factor analysis. *Psychometrika* **48**(2), 247–251 (1983)
- Bertsekas, D.P.: *Nonlinear Programming*, 2nd edn. Athena Scientific, MA (2004)
- Böhning, D., Lindsay, B.G.: Monotonicity of quadratic-approximation algorithms. *Ann. I. Stat. Math.* **40**(4), 641–663 (1988). <https://doi.org/10.1007/BF00049423>
- Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **39**(1), 1–38 (1977)
- Elashoff, M.E., Ryan, L.: An EM algorithm for estimating equations. *J. Comput. Graph. Stat.* **13**(1), 48–65 (2004)
- Fisher, R.A.: Theory of statistical estimation. *Proc. Camb. Philos. Soc.* **22**, 700–725 (1925)
- Hasselblad, V.: Estimation of finite mixtures of distributions from the exponential family. *J. Am. Stat. Assoc.* **64**(238), 1459–1471 (1969)
- Jamshidian, M., Jennrich, R.I.: Conjugate gradient acceleration of the EM algorithm. *J. Am. Stat. Assoc.* **88**(421), 221–228 (1993)
- Kenward, M.G., Molenberghs, G.: Likelihood based frequentist inference when data are missing at random. *Stat. Sci.* **13**(3), 236–247 (1998)
- Lange, K.: A gradient algorithm locally equivalent to the EM algorithm. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **57**(2), 425–437 (1995a)
- Lange, K.: A quasi-Newton acceleration of the EM algorithm. *Stat. Sin.* **5**(1), 1–18 (1995b)
- Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, 2nd edn. Wiley, Hoboken (2002)
- Liu, C., Rubin, D.R., Wu, Y.N.: Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **85**(4), 755–770 (1998)
- Louis, A.: Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **44**(2), 226–233 (1982)
- McLachlan, G.: On Aitken's method and other approaches for accelerating convergence of the EM algorithm. In: *Proceedings of the A.C. Aitken Centenary Conference*. University of Otago Press, pp. 201–209 (1995)
- McLachlan, G., Krishnan, T.: *The EM Algorithm and Extensions*, 2nd edn. Wiley, Hoboken (2008)
- McLachlan, G., Peel, D.: *Finite Mixture Models*. Wiley, Hoboken (2000)
- Meilijson, I.: A fast improvement to the EM algorithm on its own terms. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **51**(1), 127–138 (1989)
- Meng, X.L., Rubin, D.B.: Using EM to obtain asymptotic matrices: the SEM algorithm. *J. Am. Stat. Assoc.* **86**(416), 899–909 (1991)
- Minka, T.P.: Expectation-maximization as lower bound maximization. <http://www.cse.psu.edu/~rtc12/CSE598G/papers/minka98expectationmaximization.pdf> (1998). Accessed 12 Jan 2020
- Mosimann, J.E.: On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. *Biometrika* **49**(1 and 2), 65–82 (1962)
- Nocedal, J., Wright, S.J.: *Numerical Optimization*. Springer, Berlin (2006)
- Osborne, R.: Fisher's method of Scoring. *Int. Stat. Rev.* **60**(1), 99–117 (1992)
- Rai, S.N., Matthews, D.E.: Improving the EM algorithm. *Biometrics* **49**(2), 587–591 (1993)
- Rubin, D.B., Thayer, D.T.: EM algorithm for ML factor analysis. *Psychometrika* **47**(1), 69–76 (1982)
- Salakhutdinov, R., Roweis, S., Ghahramani, Z.: Relationship between gradient and EM steps in latent variable models. <http://www.utstat.toronto.edu/~rsalakhu/papers/report.pdf> (2002). Accessed 12 Jan 2020
- Savalei, V.: Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychol. Methods* **15**(4), 352–367 (2010)
- Seaman, S., Galati, J., Jackson, D., Carlin, J.: What is meant by “Missing at Random”? *Stat. Sci.* **28**(2), 257–268 (2013)
- Small, C.G., Wang, J.: *Numerical Methods for Nonlinear Estimating Equations*, Oxford Statistical Science Series, vol. 29. Oxford University Press, Oxford (2003)
- Titterton, D.M.: Recursive parameter estimation using incomplete data. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **46**(2), 257–267 (1984)
- Varadhan, R., Roland, C.: Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.* **35**, 335–353 (2008)
- Xu, L., Jordan, M.I.: On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **8**(1), 129–151 (1995)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.