Some results on maximum likelihood from incomplete data: finite sample properties and an improved M-estimator for resampling method

Budhi Surya (budhi.surya@vuw.ac.nz)  
Victoria University of Wellington

Research Article

**Keywords:** Conditional observed information, Incomplete data, Maximum likelihood, M-estimator, Sandwich estimator

**Posted Date:** February 16th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2580049/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

**Additional Declarations:** No competing interests reported.
Some results on maximum likelihood from incomplete data: finite sample properties and an improved M-estimator for resampling method

Budhi Surya

School of Mathematics and Statistics, Victoria University of Wellington, Gate 6, Kelburn PDE, Wellington 6140, New Zealand

Abstract

This paper presents some results on the classical maximum likelihood (ML) parameter estimation from incomplete data. Finite sample properties of incomplete-data conditional observed information matrices are established. They possess positive definiteness and the same Loewner partial ordering as the expected information matrices do. An explicit form of the observed Fisher information (OFI) is derived for the calculation of standard errors of the ML estimates. It simplifies Louis (1982) general formula for the OFI matrix. To prevent from getting an incorrect inverse of the OFI matrix, which may be attributed by the lack of sparsity and large size of the matrix, a monotone convergent recursive equation for the inverse of OFI matrix is developed which in turn generalizes the algorithm of Hero and Fessler (1994) for the Cramér-Rao lower bound. To improve the estimation, in particular when applying repeated sampling to incomplete data, a robust M-estimator is introduced. A closed form sandwich estimator of covariance matrix is proposed to provide the standard errors of the M-estimator. By the resulting loss of information presented in finite-sample incomplete data, the sandwich estimator produces smaller standard errors for the M-estimator than the ML estimates. In the case of complete information or absence of re-sampling, the M-estimator coincides with the ML estimates. Application to parameter estimation of a finite-mixture of Markov jump processes is discussed to verify the results. The simulation study confirms the accuracy and asymptotic properties of both the ML estimates and M-estimator.

Keywords: Conditional observed information, Incomplete data, Maximum likelihood, M-estimator, Sandwich estimator, 2020 MSC: 60J20, 60J27, 62M09, 62H30

1. Introduction

Large sample properties of maximum likelihood estimates (MLE) of statistical model parameters have been well documented in literature. See for e.g. Cramér (1946), Newey and McFadden (1994), and Van der Vaart (2000). The properties state that as the sample size increases, the MLE has asymptotic (multivariate) normal distribution with mean equal to the true parameter value whereas the covariance matrix is given by the inverse of the Fisher information of observed data. These fine properties of MLE were first shown by Fisher (1925) and later established rigorously, among others, by Cramér (1946). For unbiased estimators, the information matrix corresponds to the Cramér-Rao lower bound. See for e.g. p.489 of Cramér (1946) or p.2162 of Newey and McFadden (1994). For finite sample application, the observed Fisher information has been widely used to find the standard errors of the MLE with good accuracy (Efron and Hinkley, 1978). Although the large sample properties have been well developed for complete data, some further studies are needed to understand finite/large sample properties of MLE when applied to incomplete data and using resampling method.

The EM algorithm developed by Dempster et al. (1977) for ML parameter estimation under incomplete data has been widely applied over various fields in literature. It is a robust and powerful tool for statistical analysis with missing data (Little and Rubin, 2020). The algorithm provides an iterative approach to obtain the ML estimates of the model parameters in a way that avoids necessary regularity conditions on the log-likelihood function in terms of the existence of its second derivative and invertibility of the corresponding Hessian matrix. Otherwise, if the regularity conditions are satisfied for each observation, one may employ the Newton-Raphson approach, i.e., the Fisher scoring method to find the MLE also iteratively, see for e.g. Osborne (1992), Hastie et al. (2009), and Takai (2020). There are two steps in the EM estimation. The first step, the $E$-step, involves valuation under current parameter estimate of conditional expectation of the log-likelihood given observed data, whereas the second step, the $M$-step, deals with optimizing the conditional expectation. In each iteration, the algorithm increases the value of log-likelihood. The appealing monotone convergence property of the observed-data log-likelihood gives a higher degree of stability for the convergence of the EM algorithm (Wu, 1983). See McLachlan and Krishnan (2008) for recent
developments, extensions and applications of EM algorithm.

However, the EM algorithm only provides point estimates of parameters (McLachlan and Krishnan, 2008). Unlike the Fisher scoring method, it does not automatically produce the covariance matrix of MLE. Additional steps are required to find the covariance matrix. For finite independent data, the expected information is replaced by the observed information matrix specified by the second derivative of the observed-data log-likelihood function (Efron and Hinkley, 1978) whose inverse provides estimated covariance matrix.

In general, the second derivative is difficult to evaluate directly. One of major contribution on the evaluation of the observed information was given by Louis (1982) in which a general matrix formula was proposed. Notice that the Louis’ formula involves conditional expectation of outer product of the complete-data score function which may be complicated to simplify. Meng and Rubin (1991) derived the covariance matrix using the fundamental identity given in Dempster et al. (1977) relating the log-likelihood of observed data, the EM-update criterion function and the conditional expectation of conditional log-likelihood of complete data given its incomplete observation. In working out the information matrix, they applied in the M-step of the EM-algorithm a first-order Taylor approximation around current parameter estimate to the EM-update function. This approximation was first noted in Meilijson (1989) in an attempt to provide a fast improvement to the EM algorithm. The covariance matrix derived in Meng and Rubin (1991) is given by the difference between the complete information matrix and an incomplete one, although in a slightly more complex form than Louis’ formula. Based on the Taylor approximation discussed in Meilijson (1989), Jamshidian and Jennrich (2000) proposed a numerical differentiation method to evaluate the covariance matrix for the incomplete data. A rather direct calculation of observed information matrix was discussed in Oakes (1999) based on taking derivatives of the fundamental EM identity.

Research contribution

However, inverting the observed Fisher information matrix may be problematic due to lack of sparsity and the (large) size of the information matrix parameter. This might result in getting an incorrect inverse of the matrix. To overcome this problem, a monotone convergent iterative scheme is proposed to calculate the inverse. The key to deriving the iteration is the use of a derived information-loss inequality for observed information matrices, which generalizes the Orchard and Woodbury (1972) inequality for the corresponding expected information matrices. Furthermore, to improve the parameter estimation, in particular when applying repeated sampling such as the Bootstrapping method (Efron and Hastie, 2021), this paper proposes a robust M-estimator with smaller standard errors than that of the MLE. A closed form sandwich estimator of finite-sample covariance matrix is introduced to provide the standard errors of the M-estimator. The sandwich estimator reserves itself as the lower bound for the inverse of the observed Fisher information. It is slightly different from Huber sandwich estimator (Huber, 1967; Freedman, 2006; Little and Rubin, 2020) for model misspecification under incomplete data. In the absence of repeated sampling, the M-estimator coincides with the MLE.

This paper is organized as follows. Section 2 discusses ML parameter estimation from incomplete data and respective problems of the estimation. Section 3 presents the main results and contributions of this paper. An example on finite mixture of Markov jump processes is presented in Section 4. A series of simulation studies based on the example of Section 4 are performed in Section 5 to verify the main results of Section 3 before concluding this paper in Section 6.

2. Maximum likelihood estimation from incomplete data

2.1. The likelihood function of incomplete data

Let X and Y be two random vectors defined on the same probability space $(\Omega, \mathcal{F}, P)$. Denote by $(X, S)$ and $(Y, T)$ the corresponding measurable state spaces of X and Y and by $T : X \to Y$ a many-to-one mapping from X to Y. Suppose that a complete-data vector $x \in X$ is partially observed through an incomplete-data vector $y = T(x)$ in Y. Assume that there exist probability density functions $f_x(x|\theta)$ and $f_{xy}(y|\theta)$ corresponding to the complete data $x \in X$ and its incomplete observation $y \in Y$, respectively. Here $\theta$ represents a vector of parameters on a parameter space $\Theta$, with $d = |\theta|$, characterizing the distribution of X. For formality, we assume that $\Theta$ is a compact set. Define $X(y) = \{x \in X : T(x) = y\} \in S$. Then,

$$f_{xy}(y|\theta) = \int_{X(y)} f_x(x|\theta) \lambda(dx), \quad (1)$$

where $\lambda$ is a $\sigma$-finite measure on S, absolutely continuous with respect to the probability distribution $P \circ X^{-1}$ with the density function $f_x(x|\theta)$ (the Radon-Nikodym derivative). See e.g. Halmos and Savage (1949) for details. Notice that our description of the marginal distribution (1) for incomplete data is slightly more general than the one employed in McLachlan and Krishnan (2008).

Following the identity (1), the conditional probability density function $f(x|y, \theta)$ of the complete-data X given its incomplete observation Y is therefore given by

$$f(x|y, \theta) = \frac{f_x(x|\theta)}{f_{xy}(y|\theta)}. \quad (2)$$

Assumption 1. Without loss of generality, we assume throughout the remaining of this paper that the log-likelihood function $\log f_x(x|\theta)$ is twice continuously differentiable w.r.t $\theta$ and for all $\theta \in \Theta$ and for all $m \in \{0, 1, 2\}$,

$$\int_{X(y)} \left| \frac{\partial^m \log f_x(x|\theta)}{\partial \theta^m} \right| f(x|y, \theta) \lambda(dx) < \infty. \quad (A1)$$


By independence of the observations
Proposition 1.

2.2. M-criterion for maximum likelihood estimation

To prove the claim, consider the almost surely under , where is partially observed, represented by an incomplete-data vector for . And each observation is also independent. Namely, if is independent of for , then is independent of . The log-likelihood contribution of the incomplete observation is represented by an incomplete-data vector .

is used to get an estimator defined as the global maximizer of the log-likelihood function.

This method proves the convergence of to the true value as the sample size increases, where the convergence occurs almost surely under , where .

2.2. M-criterion for maximum likelihood estimation

The following proposition establishes the above claim.

Proposition 1. By independence of the observations ,

implies that as the sample size increases.

Proof. To prove the claim, consider the 

\[ M_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} \log f_{o}(Y^{k|\theta}). \]  

See Van der Vaart (2000). By independence of ,

\[ M_n(\theta) \Rightarrow M(\theta) = \mathbb{E}[\log f_{o}(Y^{k|\theta})] \]

which holds for a generic sample . Furthermore, from the Shannon-Kolmogorov information inequality, it holds for any and generic ,

\[ R(\theta^{0}, \theta) := \mathbb{E}[\log \left( \frac{f_{o}(Y^{k|\theta})}{f_{o}(Y^{k|\theta^{0}})} \right)|\theta] > 0, \]

see p. 113 in Ferguson (1996). Thus, for any ,

\[ \mathbb{E}[\log f_{o}(Y^{k|\theta})] < M(\theta) \iff \theta = \arg \max_{\theta \in \Theta} M(\theta). \]

Since is the global maximizer of and the latter converges with probability one to , it follows that gets closer and closer to the global maximizer of as increases, which by compactness of implies that \[ \hat{\theta}_{n} \Rightarrow \theta_{0}. \]

By (A1) and the first order Euler condition, the MLE is found as the solution of the systems of equation

\[ 0 = S_{n}(\theta) := \frac{1}{n} \sum_{k=1}^{n} \frac{\partial \log f_{o}(Y^{k|\theta})}{\partial \theta}. \]  

Assuming continuous differentiability of the score function , for large , the consistency of allows one to apply the first-order Taylor approximation around to , see e.g. Friedman (2006), to arrive at

\[ \hat{\theta}_{n} = \theta_{0} + J_{f}^{-1}(\theta_{0})S_{n}(\theta_{0}), \]

where \[ J_{f}(\theta) = -\frac{1}{n} \sum_{k=1}^{n} \frac{\partial^2 \log f_{o}(Y^{k|\theta})}{\partial \theta^2}. \]

By independence of , application of the Slutsky’s lemma (see, Lemma 2.8 in Van der Vaart (2000)) and the Central Limit Theorem (CLT), it is known that \[ \sqrt{n}S_{n}(\theta_{0}) \sim N(0, I_{f}(\theta_{0})), \]

as and . Thus, \[ I_{f}(\theta) = \mathbb{E}\left[ -\frac{\partial^2 \log f_{o}(Y^{k|\theta})}{\partial \theta^2} \right| \theta \]

being the expected Fisher information matrix. See, e.g. Cramér (1946), Newey and McFadden (1994) and Van der Vaart (2000). As a result, it follows from (5) that the MLE has the \( \sqrt{n} \)-asymptotic normal distribution

\[ \sqrt{n}(\hat{\theta}_{n} - \theta_{0}) \sim N(0, I_{f}^{-1}(\theta_{0})). \]

2.3. Some estimation difficulties for incomplete data

However, for incomplete data, it is generally difficult to obtain the MLE explicitly. There are some difficulties:

(i) solves the systems of (nonlinear) equations

\[ 0 = S_{n}(\hat{\theta}_{n}) = \frac{1}{n} \sum_{k=1}^{n} \int_{\mathcal{X}(\theta^{0})} \frac{\partial f_{o}(x^{k|\theta})}{\partial \theta} \lambda(dx^{k}), \]

from which it is difficult to pull out \( \hat{\theta}_{n} \) explicitly.

(ii) Since \( \theta_{0} \) is unknown, the estimator (5) is not applicable. Hence, should be derived recursively.

(iii) The observed Fisher information \( I_{f}(\theta) \) and the score function \( S_{n}(\theta) \) are not explicit. Therefore, the recursive Fisher scoring method (see, Hastie et al. (2009)) derived from (5) is difficult to implement.

(iv) Although Louis (1982) formula is available for evaluating the observed Fisher information \( I_{f}(\theta) \), it does not simplify the conditional expectation of outer product of the complete-data score function.

(v) In general the observed information matrix \( I_{f}(\theta) \) lacks of sparsity and the size of \( \theta \) could be large. Hence, inverting \( I_{f}(\theta) \) directly might be difficult and may result in an incorrect inverse \( I_{f}^{-1}(\theta) \).

(vi) Due to the resulting loss of information presented in incomplete data \( Y \), the standard errors of the MLE \( \hat{\theta}_{n} \) is larger than that of using complete data \( X \).

(vii) The EM algorithm may be slow in its convergence and might be difficult to get an explicit iteration form due to possible nonlinearity of the likelihood function in terms of the model parameter.
This paper attempts to solve the above problems and proposes improved estimation of the true value \( \theta_0 \) in terms of smaller standard error of its estimator, in particular, by applying repeated sampling method.

3. Main contributions

To overcome the difficulties (i)-(vi), we consider estimating the true value \( \theta_0 \) based on the incomplete observation \( \{Y^k, k = 1, \ldots, n\} \) using the M-criterion

\[
\mathcal{M}_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} E \left[ \log f_x(X^k|\theta) | Y^k, \theta_0 \right].
\]

where \( E[\bullet|\theta_0] \) refers to the expectation operator associated with the underlying probability measure \( \mathbb{P}[\bullet|\theta_0] \) from which the complete-data \( \{X^k : 1 \leq k \leq n\} \) were generated under \( \theta_0 \). Note that such M-criterion function as \( \mathcal{M}_n(\theta) \) is not discussed in Van der Vaart (2000).

An estimator \( \hat{\theta}_n^0 \) of \( \theta_0 \) is defined as the maximizer of the M-criterion \( \mathcal{M}_n(\theta) \) over \( \theta \in \Theta \), as the solution of

\[
0 = \mathcal{I}_n(\theta) := \frac{1}{n} \sum_{k=1}^{n} E \left[ \frac{\partial \log f_x(X^k|\theta)}{\partial \theta} | Y^k, \theta_0 \right].
\]

Define \( \mathcal{I}_n(\theta) := E[\partial \log f_x(X^k|\theta) \partial \theta] \). It is clear that \( \mathcal{I}_n(\theta_0) = 0 \). Thus, the M-estimator \( \hat{\theta}_n^0 \) can simply be written as

\[
\hat{\theta}_n^0 = \mathcal{I}_n^{-1}(\mathcal{I}_n(\theta_0)).
\]

As the M-estimator, we may apply similar arguments to the proof of Proposition 1 and the following result to establish the consistency of the M-estimator \( \hat{\theta}_n^0 \).

**Lemma 1.** Let \( \mathcal{M}(\theta) = E[\log f_x(X|\theta)|\theta_0] \). Then, for \( \theta_0 \in \Theta \), \( \sup_{\theta \in \Theta} \mathcal{M}(\theta) \leq \mathcal{M}(\theta_0) \). Equivalently, \( \mathcal{M}(\theta) = 0 \) follows from \( \mathcal{M}(\theta_0) = 0 \).

**Proof.** The proof follows from concavity of log-function and application of Jensen’s inequality, i.e.,

\[
\mathcal{M}(\theta) - \mathcal{M}(\theta_0) \leq \log E \left[ \frac{f_x(X|\theta)}{f_x(X|\theta_0)} | \theta_0 \right] = 0,
\]

where the last equality is due to the fact that the likelihood ratio \( \frac{f_x(X|\theta)}{f_x(X|\theta_0)} \) corresponds to the Radon-Nikodym derivative of changing underlying probability measure from \( \mathbb{P}[\bullet|\theta_0] \) to \( \mathbb{P}[\bullet|\theta] \) which results in \( E \left[ \log f_x(X|\theta_0) \right] = 1 \). \( \mathcal{M}(\theta_0) = 0 \) follows from (A1) and taking account that \( E[\partial \log f_x(X|\theta) \partial \theta] = 0 \).

Alternatively, by independence of \( \{Y^k\} \), the law of large number shows that \( \mathcal{M}_n(\theta) \to \mathcal{I}(\theta) \) for any \( \theta \in \Theta \). As the result, \( \hat{\theta}_n^0 \to \theta_0 \) and we have asymptotically that \( \theta \approx \mathcal{I}_n^{-1}(\mathcal{I}(\theta)). \) By consistency of \( \hat{\theta}_n^0 \) and the regularity condition (A1), we may apply the first order Taylor approximation around \( \theta_0 \) to \( \mathcal{I}_n(\hat{\theta}_n^0) \), see Freedman (2006), to obtain

\[
\hat{\theta}_n^0 = \theta_0 + J_n^{-1}(\theta_0) \mathcal{I}_n(\theta_0),
\]

where \( J_n(\theta) \) defines the conditional observed information

\[
J_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} E \left[ - \frac{\partial^2 \log f_x(X^k|\theta)}{\partial \theta^2} | Y^k, \theta \right],
\]

which by independence of \( \{Y^k\} \) converges to the complete-data expected information \( I_n(\theta) = E[ - \frac{\partial^2 \log f_x(X|\theta)}{\partial \theta^2} | \theta] \).

In the section below, finite sample properties of the information matrices \( J_n(\theta) \) and \( J_n(\theta) \) are studied. They will be used to derive recursive estimations for the MLE \( \hat{\theta}_n \) and for the Cr\^{a}mer-Rao lower bound \( J_n^{-1}(\hat{\theta}_n) \) for the finite-sample covariance matrix of the MLE. In particular, to obtain the limiting normal distribution of \( \hat{\theta}_n \).

3.1. Finite sample properties of information matrices

The results below will be used to derive an explicit form of the observed information \( J_n(\theta) \), to establish the inequality concerned with resulting loss of information in a finite-sample incomplete data, to obtain the \( \sqrt{n} \)-limiting normal distribution of the M-estimator \( \hat{\theta}_n \), and its asymptotic efficiency compared to the MLE \( \hat{\theta}_n \). They have remained largely unexamined in literature on statistical analysis of incomplete data, see e.g., McLachlan and Krishnan (2008) and Little and Rubin (2020) for details.

**Proposition 2.** For any \( \theta \in \Theta \) and \( Y \in \mathcal{Y} \),

\[
E \left[ \frac{\partial \log f_x(X|\theta)}{\partial \theta} | Y, \theta \right] = 0.
\]

**Proof.** Using the conditional density function (2),

\[
E \left[ \frac{\partial \log f_x(X|\theta)}{\partial \theta} | Y = y, \theta \right] = \int_{X^{(y)}} \frac{\partial \log f_x(x|y, \theta)}{\partial \theta} f_x(x|y, \theta) dx = \int_{X^{(y)}} \frac{\partial f_x(x|y, \theta)}{\partial \theta} dx,
\]

from which the claim follows on account of the regularity condition (A1) along with the identities (1) and (2).

Applying (11) to (2) after taking the logarithm,

\[
\frac{\partial \log f_x(X|\theta)}{\partial \theta} = \frac{\partial \log f_x(Y|\theta)}{\partial \theta} + \frac{\partial \log f_x(X|Y, \theta)}{\partial \theta},
\]

leads to the following identity, which corresponds to eqn. (3.44) in Section 3.7 of McLachlan and Krishnan (2008).

**Corollary 1.** For any \( \theta \in \Theta \) and \( Y \in \mathcal{Y} \),

\[
E \left[ \frac{\partial \log f_x(X|\theta)}{\partial \theta} | Y, \theta \right] = \frac{\partial \log f_x(Y|\theta)}{\partial \theta}.
\]
The identity (13) identifies \( \mathcal{D}_k(\theta_0) \) as \( S_k(\theta) \) (4). The result of Proposition 2 leads to the following fact.

**Corollary 2.** For any \((\theta, \theta_j) \in \Theta \) and \( Y \in \mathcal{Y} \),

\[
\text{Cov}\left( \frac{\partial \log f(X|Y)}{\partial \theta_i}, \frac{\partial \log f(X|Y)}{\partial \theta_j} \right| Y, \theta) = 0.
\]

The result above shows that \( \frac{\partial \log f(X|Y)}{\partial \theta_i} \) and \( \frac{\partial \log f(X|Y)}{\partial \theta_j} \) are conditionally independent given incomplete data \( Y \).

To show the Loewner partial ordering of \( J_i(\theta) \) and \( J_{ij}(\theta) \) and to derive an explicit form of the observed information matrix \( J_{ij}(\theta) \), the results below are required.

**Theorem 1.** The conditional information matrix

\[
J_{ij}(\theta, \theta_j) := \mathbb{E}\left[ \frac{\partial^2 \log f(X|Y)}{\partial \theta_i \partial \theta_j} \right| Y, \theta]
\]

is positive definite for any \( \theta \in \Theta \) satisfying the equation

\[
J_{ij}(\theta, \theta_j) = \mathbb{E}\left[ \frac{\partial \log f(X|Y)}{\partial \theta_i} \frac{\partial \log f(X|Y)}{\partial \theta_j} \right| Y, \theta].
\]

**Proof.** To prove the identity, by the chain rule the identity below holds for any \((\theta, \theta_j) \in \Theta \) and \( y \in \mathcal{Y} \),

\[
\begin{aligned}
\int_{\mathcal{Y}} \left[ \frac{\partial^2 \log f(x|y, \theta)}{\partial \theta_i \partial \theta_j} \right] f(x|y, \theta) \lambda(dx) &= \int_{\mathcal{Y}} \left[ \frac{\partial \log f(x|y, \theta)}{\partial \theta_i} \right] \frac{\partial \log f(x|y, \theta)}{\partial \theta_j} f(x|y, \theta) \lambda(dx) \\
&+ \int_{\mathcal{Y}} \left[ \frac{\partial \log f(x|y, \theta)}{\partial \theta_i} \right] \frac{\partial \log f(x|y, \theta)}{\partial \theta_j} f(x|y, \theta) \lambda(dx)
\end{aligned}
\]

Therefore, following the regularity condition (A1),

\[
\mathbb{E}\left[ \frac{\partial^2 \log f(X|Y)}{\partial \theta_i \partial \theta_j} \right| Y = y, \theta]
\]

\[
= -\mathbb{E}\left[ \left( \frac{\partial \log f(X|Y)}{\partial \theta_i} \right) \left( \frac{\partial \log f(X|Y)}{\partial \theta_j} \right) \right| Y = y, \theta]
\]

\[
= \mathbb{E}\left[ \left( \frac{\partial \log f(X|Y)}{\partial \theta_i} \right) \left( \frac{\partial \log f(X|Y)}{\partial \theta_j} \right) \right| Y = y, \theta]
\]

which leads to the identity on account of (11). To show that \( J_{ij}(\theta) \) is positive definite, let \( S_{ij}(\theta) = \frac{\partial \log f(X|Y)}{\partial \theta_i} \frac{\partial \log f(X|Y)}{\partial \theta_j} \frac{\partial \log f(X|Y)}{\partial \theta_i} \). Thus, \( S_{ij}(\theta) = S_{ij}(\theta)S_{ij}(\theta)^T \). Hence, for any \( 0 \neq z \in \mathbb{R}^d \), \( z^T S_{ij}(\theta) z > 0 \) showing the information matrix \( J_{ij}(\theta) \) is positive definite.

Taking expectation w.r.t. \( \mathbb{P}[\bullet|\theta] \) on both sides of the last identity leads to the fact that \( -\frac{\partial^2 \log f(X|Y)}{\partial \theta_i \partial \theta_j} \) is the observed Fisher information, see Schervish (1995).

**Theorem 2** (Resulting loss of information presented in incomplete data). For any \( \theta \in \Theta \) and \( Y \in \mathcal{Y} \),

\[
J_{ij}(\theta) > J_{ij}(\theta).
\]

**Proof.** Let \( J_i(\theta, \theta_j) \) be the \((i, j)\)-element of \( J_i(\theta) \), similarly defined for \( J_j(\theta, \theta_i) \). Taking derivative of (12) yields

\[
-\frac{\partial^2 \log f(X|Y)}{\partial \theta_i \partial \theta_j} = -\frac{\partial^2 \log f(X|Y)}{\partial \theta_i \partial \theta_j} + \frac{\partial^2 \log f(X|Y)}{\partial \theta_i \partial \theta_j}.
\]

By taking conditional expectation \( \mathbb{E}[\bullet|Y, \theta] \), we obtain,

\[
J_{ij}(\theta, \theta_j) = J_i(\theta, \theta_j) - J_j(\theta, \theta_i).
\]

The inequality (14) follows from the fact that the information matrix \( J_{ij}(\theta) \) is positive definite, by Theorem 1.

By Theorem 2 we deduce the following inequality corresponding to the resulting information loss presented in incomplete-data. The inequality was discussed in Orchard and Woodbury (1972), Blahut (1987), Theorem 2.86 of Schervish (1995), and McLachlan and Krishnan (2008).

\[
I_i(\theta) > I_j(\theta) > 0,
\]

with \( I_i(\theta) = \mathbb{E}[J_i(\theta|\theta)] \) and \( I_j(\theta) = \mathbb{E}[J_j(\theta|\theta)] \), the expected information matrices. Note that positive definiteness of \( I_i(\theta) \) is due to the fact that it is the expected Fisher information matrix satisfying \( I_i(\theta) = \mathbb{E}[\left( \frac{\partial \log f(X|Y)}{\partial \theta} \right) \left( \frac{\partial \log f(X|Y)}{\partial \theta} \right)^T] > 0 \).

By applying the conditional probability density (2) and the above identities, we derive the following result.

**Proposition 3.** For any \((\theta, \theta_j) \in \Theta \) and \( Y \in \mathcal{Y} \),

\[
J_{ij}(\theta, \theta_j) = \mathbb{E}\left[ \left( \frac{\partial \log f(X|Y)}{\partial \theta_i} \right) \left( \frac{\partial \log f(X|Y)}{\partial \theta_j} \right) \right| Y, \theta]
\]

\[
= \mathbb{E}\left[ \left( \frac{\partial \log f(X|Y)}{\partial \theta_i} \right) \left( \frac{\partial \log f(X|Y)}{\partial \theta_j} \right) \right| Y, \theta]
\]

\[
\mathbb{E}\left[ \left( \frac{\partial \log f(X|Y)}{\partial \theta_i} \right) \left( \frac{\partial \log f(X|Y)}{\partial \theta_j} \right) \right| Y, \theta]
\]

which leads to the identity on account of (11). To show that \( J_{ij}(\theta) \) is positive definite, let \( S_{ij}(\theta) = \frac{\partial \log f(X|Y)}{\partial \theta_i} \frac{\partial \log f(X|Y)}{\partial \theta_j} \frac{\partial \log f(X|Y)}{\partial \theta_i} \). Thus, \( J_{ij}(\theta) = S_{ij}(\theta)S_{ij}(\theta)^T \). Hence, for any \( 0 \neq z \in \mathbb{R}^d \), \( z^T S_{ij}(\theta) z > 0 \) showing the information matrix \( J_{ij}(\theta) \) is positive definite.

The result below presents explicit form of the \((i, j)\)-element \( J_i(\theta, \theta_j) \) = \(-\frac{\partial^2 \log f(X|Y)}{\partial \theta_i \partial \theta_j} \), with \((\theta, \theta_j) \in \Theta \) of the observed Fisher information matrix \( J_i(\theta) \).
Theorem 3. [(i)] The $(i, j)$-element of the observed information $I_{ij}(\theta)$ for incomplete data $Y = \bigcup_{k=1}^{n} Y^k$ is

\[
I_{ij}(\theta, \theta) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ \frac{\partial^2 \log f_k(Y^k|\theta)}{\partial \theta_i \partial \theta_j} \right] - \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ \left( \frac{\partial \log f_k(Y^k|\theta)}{\partial \theta_i} \right) \left( \frac{\partial \log f_k(Y^k|\theta)}{\partial \theta_j} \right) \right] + \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ \frac{\partial \log f_k(Y^k|\theta)}{\partial \theta_i} \right] \mathbb{E} \left[ \frac{\partial \log f_k(Y^k|\theta)}{\partial \theta_j} \right].
\]

[(ii)] Both $J_i(\theta_\hat{n})$ and $J_i(\theta_n)$ are positive definite with

\[
J_i(\theta_\hat{n}) > J_i(\theta_n) > 0. \tag{16}
\]

Proof. Let $J_i^2(\theta_i, \theta_j) = \frac{\partial^2 \log f_i(Y^i|\theta)}{\partial \theta_i \partial \theta_j}$. Similarly for $J_i^2(\theta_i, \theta_j)$. From (12),

\[
\frac{\partial^2 \log f_i(Y^i|\theta)}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 \log f_i(Y^i|\theta)}{\partial \theta_i \partial \theta_j} - \frac{\partial \log f_i(Y^i|\theta)}{\partial \theta_i} \frac{\partial \log f_i(Y^i|\theta)}{\partial \theta_j}.
\]

Taking conditional expectation $\mathbb{E} \left[ \frac{\partial^2 \log f_i(Y^i|\theta)}{\partial \theta_i \partial \theta_j} \right]$ we obtain $J_i^2(\theta_i, \theta_j) = J_i^2(\theta_i, \theta_j)$. The proof of (i) is complete by the result of Proposition 3 on account that $J_i(\theta_i, \theta_j) = \frac{1}{n} \sum_{k=1}^{n} J_i(\theta_i, \theta_j) - \frac{1}{n} \sum_{k=1}^{n} J_i(\theta_i, \theta_j)$. The claim (ii) is due to (14) on account that $\theta_\hat{n}$ maximizes (3). □

Remark 1. Notice that the information matrix $I_{ij}(\theta)$ takes a slightly different form than Louis (1982) matrix formula. The main differences with the latter is that it simplifies the calculation of conditional expectation of the outer product of the complete-data score function appeared in the Louis’ matrix formula. And most notably, it directly verifies the asymptotic consistency of $I_{ij}(\theta)$ to the incomplete-data Fisher information matrix $I_{ij}(\theta)$ as the sample size $n$ increases. Also, the derivation is much simplified compared to the approach of Louis (1982) and McCullagh and Krishnan (2008).

The expression of $(i, j)$-element of information matrix $I_{ij}(\theta)$ agrees with that of given by Frydman and Surya (2023) for the finite mixture of Markov jump processes.

Corollary 3. For any incomplete data $Y \in \mathcal{Y}$,

\[
J_i(\theta_\hat{n}) = \mathbb{E} \left[ \frac{\partial^2 \log f_i(X|\theta_\hat{n})}{\partial \theta_i \partial \theta_j} \right] > M_i(\theta_\hat{n}) = \mathbb{E} \left[ \left( \frac{\partial \log f_i(X|\theta_\hat{n})}{\partial \theta_i} \right) \left( \frac{\partial \log f_i(X|\theta_\hat{n})}{\partial \theta_j} \right) \right]. \tag{17}
\]

Corollary 3 extends the result on equivalence between unconditional variance of complete-data score function and the expected Fisher information, see Schervish (1995).

3.2. Recursive algorithms for $J_i^{-1}(\theta_\hat{n})$ and MLE $\hat{\theta}_n$

This section discusses recursive estimations of the MLE $\hat{\theta}_n$ and the inverse of observed information matrix $J_i^{-1}(\theta_\hat{n})$ used for deriving the covariance matrix of estimation error $\hat{\theta}_n - \theta_0$.

3.2.1. Recursive calculation of $J_i^{-1}(\theta_\hat{n})$

Hero and Fessler (1994) proposed a recursive equation for the valuation of the Cramér-Rao lower bound $I^{-1}(\theta)$, the inverse of expected Fisher information $I(\theta)$. The method avoids taking direct inverse of $I(\theta)$, which in general may be difficult to invert than the information matrix $I(\theta)$.

However, their result is not immediately applicable for incomplete data in general since the expected information matrices $I(\theta)$ and $I(\theta)$ may not be available in closed form. We generalize their results for the inverse of observed information matrix $J_i^{-1}(\theta_\hat{n})$ using conditional observed information matrices $J_i(\theta_\hat{n})$ and $J_i(\theta_n)$. The key to deriving the recursive equation for the inverse $J_i^{-1}(\theta)$ is the inequality (16).

Theorem 4. Let $\{\Psi_\ell\}_{\ell \geq 0}$ be a sequence of $(d \times d)$-matrices, with $d = |\Theta|$, and $\Psi_0 = 0$ satisfying

\[
\Psi_{\ell+1} = A(\theta_\hat{n})\Psi_\ell + B(\theta_\hat{n}), \tag{18}
\]

for $A(\theta_\hat{n})$, $B(\theta_\hat{n}) \in \mathbb{R}^{d \times d}$. Then, $\{\Psi_\ell\}_{\ell \geq 0}$ converges with root of convergence $r(A(\theta_\hat{n}))$ to $\Psi = J_i^{-1}(\theta_\hat{n})$ with

\[
A(\theta_\hat{n}) = \left[ I - J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n}) \right] \quad \text{and} \quad B(\theta_\hat{n}) = J_i^{-1}(\theta_\hat{n}).
\]

Furthermore, the convergence is monotone in the sense

\[
\Psi_\ell < \Psi_{\ell+1} \leq \Psi \quad \text{for } \ell = 0, 1, \ldots
\]

Proof. Since by Theorem 3(iii) $J_i(\theta_\hat{n})$ and $J_i(\theta_n)$ are positive definite, it follows from (14) and Theorem 7.2.1 on p. 438 of Horn and Johnson (2013) that $0 < J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n}) < I$. Therefore, all eigenvalues of $I - J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n})$ are positive and strictly less than one. See Corollary 1.3.4 in Horn and Johnson (2013). By Corollary 5.6.16 of Horn and Johnson (2013), it leads to

\[
\Psi = (\sum_{\ell=0}^{\infty} [I - J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n})])^{-1} J_i^{-1}(\theta_\hat{n}) - \Psi
\]

Since all eigenvalues of $I - J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n})$ are positive and strictly less than one, we then obtain

\[
\Psi_{\ell+1} - \Psi = [I - J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n})] \Psi \ell + J_i^{-1}(\theta_\hat{n}) - \Psi
\]

as $\ell \to \infty$, with root of convergence factor given by the maximum absolute eigenvalues of $I - J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n})$. Thus, $\{\Psi_\ell\}_{\ell \geq 0}$ converges to $\Psi$ as $\ell \to \infty$. To show the convergence is monotone, recall that

\[
\Psi_{\ell+1} - \Psi = [I - J_i^{-1}(\theta_\hat{n})J_i(\theta_\hat{n})] \Psi \ell - \Psi_{\ell+1}
\]

which is positive definite by Corollary 7.7.4(a) of Horn and Johnson (2013) and by the inequality (16). □
3.2.2. Fisher scoring method for incomplete data

One of the widely used recursive scheme for parameter estimation is the Fisher scoring method, see for e.g. Osborne (1992), Hastie et al. (2009) and Takai (2020). The scheme follows from (5) by replacing the true value $\theta_0$ by an estimate $\theta_0$ obtained after $\ell$ iterations and use $\theta_{\ell+1}$ in place of $\theta_{\ell}$.

The main difficulty in dealing with incomplete data is that the information matrix $J_1(\theta)$ is difficult to evaluate. To overcome this difficulty, Takai (2020) proposed the use of lower-bound of Bohnig and Lindsay (1988). However, although it eases the difficulty, the lower-bound does not actually correspond to the observed information matrix $J_1(\theta)$.

For this reason, we use the explicit form of information matrix $J_1(\theta)$ stated in Theorem 3. The scheme reads

$$\theta^{\ell+1} = \theta^\ell + J_1^{-1}(\theta^\ell)S_n(\theta^\ell),$$

provided the matrix $J_1(\theta)$ is invertible for any $\theta \in \Theta$. Taking account of the inequality (14) and the identity (13), the above Fisher scoring method can be improved using (10) by replacing $\theta_0$ by an estimate $\hat{\theta}$ obtained after $\ell$–iteration and use $\theta^{\ell+1}$ in place of $\hat{\theta}$. The recursive scheme of the estimation is discussed in the section below.

3.2.3. EM-Gradient algorithm

The recursive equation corresponds to the EM-Gradient algorithm proposed by Lange (1995). The recursion provides the fastest Newton-Raphson algorithm for solving the M-step iteratively which has quadratic convergence compared to the linear convergence in the EM algorithm. See Wu (1983) for the convergence properties of the EM algorithm. The EM-Gradient algorithm (19) yields an estimate $\hat{\theta}$ for $\theta_0$ which serves as the lower bound to that of by the incomplete-data Fisher scoring method, see for e.g. Osborne (1992), McLachlan and Krishnan (2008) and Takai (2020), whose estimated standard error is given by $J_1^{-1}(\hat{\theta})$.

The EM-Gradient algorithm

(i) **Initial step** Set an initial value $\hat{\theta}_0$ of $\theta_0$.

(ii) **E-step**, after $\ell$–th iteration, evaluate using the current estimate $\hat{\theta}$ the conditional expectations

$$S_n(\hat{\theta}) := \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ \frac{\partial \log f_k(\chi_k; \hat{\theta})}{\partial \theta} \right] | Y^k, \hat{\theta}].$$

$$J_1(\hat{\theta}) := \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ - \frac{\partial^2 \log f_k(\chi_k; \hat{\theta})}{\partial \theta^2} \right] | Y^k, \hat{\theta}].$$

(iii) **M-step**, get an update $\hat{\theta}^{\ell+1}$ using the identity

$$\hat{\theta}^{\ell+1} = \hat{\theta}^\ell + J_1^{-1}(\hat{\theta}^\ell)S_n(\hat{\theta}^\ell).$$

(iv) **Stop** if $||\hat{\theta}^{\ell+1} - \hat{\theta}|| < \varepsilon$, with $\varepsilon > 0$. Otherwise, go back to (ii) and replace $\hat{\theta}$ by $\hat{\theta}^{\ell+1}$.

Remark 2. Thus, if the convergence criterion $||\hat{\theta}^{\ell+1} - \hat{\theta}|| < \varepsilon$ is reached, it follows from (19) and (20) that at its convergence the recursive estimator $\hat{\theta}^{\infty}$ corresponds to the MLE $\theta_0$ since $\hat{\theta}^{\infty}$ solves the score equation $S_n(\hat{\theta}^{\infty}) = 0.$

3.2.4. Revised BHHH algorithm

To deal with invertibility of the information matrix $J_1(\theta)$ in iteration (20), we may replace $J_1(\theta)$ by $M_1(\theta)$ (17) to get

$$\hat{\theta}^{\ell+1} = \hat{\theta}^\ell + M_1^{-1}(\hat{\theta}^\ell)S_n(\hat{\theta}^\ell),$$

taking advantage that $M_1(\theta)$ is positive definite, hence is invertible. At the convergence of iteration (21), $\hat{\theta}^{\infty}$ corresponds to the MLE $\hat{\theta}_n$ at which $S_n(\hat{\theta}^{\infty}) = 0$. Iteration (21) extends the algorithm of Berndt et al. (1974) for incomplete data.

Definition 1 (Finite-sample sandwich estimator).

$$\Sigma_n(\theta_0) = J_1^{-1}(\theta_0)J_1(\hat{\theta}_n)J_1^{-1}(\theta_0),$$

defines a finite-sample consistent efficient sandwich estimator of covariance matrix for the MLE $\hat{\theta}_n$.

Note that (22) is slightly different from Huber sandwich estimator $V_n(\hat{\theta}) = J_1^{-1}(\hat{\theta})K_n(\hat{\theta})J_1^{-1}(\hat{\theta})$, with the matrix $K_n(\theta)$ defined by $\frac{1}{n} \sum_{k=1}^{n} \left( \frac{\partial \log f_k(\chi_k; \theta)}{\partial \theta} \right)^\prime \left( \frac{\partial \log f_k(\chi_k; \theta)}{\partial \theta} \right)$, for model misspecification under incomplete data. See e.g. Huber (1967), Freedman (2006), and Little and Rubin (2020). In contrast to the Huber sandwich estimator, (22) does not involve the inverse of $J_1(\theta)$ which in general may be more difficult to evaluate than $J_1(\theta)$.

Theorem 5 (\(\sqrt{n}\)-consistent limiting normal distribution of the M-estimator $\hat{\theta}_n$). By Assumption (A1),

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\rightarrow} N(0, \Sigma(\theta_0)),$$

with the asymptotic covariance matrix $\Sigma(\theta_0)$ defined by

$$\Sigma(\theta_0) = I_1^{-1}(\theta_0)J_1(\theta_0)I_1^{-1}(\theta_0).$$

Proof. As a consistent estimator of $\theta_0$, the asymptotic normality of the M-estimator $\hat{\theta}_n$ follows from (9)-(10) and equality between $J_n(\hat{\theta}_n)$ (8) and $S_n(\theta_0)$ (4). Since $\sqrt{n}S_n(\theta_0) \sim N(0, I(\theta_0))$, application of Slutsky’s lemma and CLT gives the limiting distribution of $\hat{\theta}_n$.

By Assumption (A1) and consistency of the MLE $\hat{\theta}_n$, the asymptotic covariance $\Sigma(\theta_0)$ can be consistently estimated
by the sandwich estimator $\Sigma_{n}(\hat{\theta}_{n})$ (22). Furthermore, positive definiteness of $J_{x}(\hat{\theta}_{n})$ and $J_{s}(\hat{\theta}_{n})$, hence is invertible (see Theorem 7.2.1 on p. 438 of Horn and Johnson (2013)), implies following (14) that $\Sigma_{n}(\hat{\theta}_{n})$ produces smaller standard errors of the M-estimator $\hat{\theta}_{n}$ than that of the MLE $\hat{\theta}_{n}$ given by the inverse observed Fisher information $J_{x}^{-1}(\hat{\theta}_{n})$, which corresponds to the Cramér-Rao lower bound for the covariance matrix of the MLE.

**Proposition 4.** The sandwich estimator $\Sigma_{n}(\hat{\theta}_{n})$ satisfies the Loewner partial matrix ordering

\[ J_{x}^{-1}(\hat{\theta}_{n}) > J_{s}^{-1}(\hat{\theta}_{n}) > \Sigma_{n}(\hat{\theta}_{n}) > 0. \]  

(23)

### 3.4. Improved estimation by repeated sampling

The results of Theorem 5 and Proposition 4 shows that the M-estimator $\hat{\theta}_{n}$ (10) has a posterior distribution of the true value $\theta_{0}$ as it attains standard error smaller than that of the MLE $\hat{\theta}_{n}$ (5) given by the inverse $J_{x}^{-1}(\hat{\theta}_{n})$ of the observed Fisher information $J_{x}(\hat{\theta}_{n})$; even smaller than $J_{s}^{-1}(\hat{\theta}_{n})$. However, the M-estimator $\hat{\theta}_{n}$ involves $\theta_{0}$.

To construct $\hat{\theta}_{n}$, suppose that it is possible to draw (independently) $K$ samples $Y_{k} = \{Y_{k, \ell} : \ell = 1, \ldots, n\}$ of size $n$. For this purpose, one may use the Bootstrapping method. From each $k$-th subsample $Y_{k}$ an MLE is derived by applying iterative Fisher scoring method (19), the EM-Gradient algorithm (20) or the EM algorithm, which results in $K$ independent sets of MLEs $\{\hat{\theta}_{n}^{(k)} : \ell = 1, \ldots, K\}$. Define $\hat{\theta} := \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_{n}^{(k)}$. Since $\sqrt{n}(\hat{\theta}_{n}^{(k)} - \theta_{0}) \sim N(0, I_{x}^{-1}(\theta_{0}))$, it follows that $\hat{\theta} \Rightarrow \theta_{0}$ as $K$ increases where the convergence occurs with probability one. Then, apply (10) to each sample $\{Y_{k}\}$ to obtain

\[ \hat{\theta}_{n} = \hat{\theta} + J_{s}^{-1}(\hat{\theta})S_{n}(\hat{\theta}), \]  

(24)

where $S_{n}(\hat{\theta})$ and $J_{s}(\hat{\theta})$ are defined respectively by

\[ S_{n}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_{x}(Y_{k}^{(i)}|\theta), \]

\[ J_{s}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \frac{\partial^{2}}{\partial \theta^{2}} \log f_{x}(X_{k}^{(i)}|\theta) \right]. \]

whilst $X_{k}^{(i)}$ is the complete observation of $Y_{k}$. It is worth mentioning that although $S_{n}(\hat{\theta}_{0}) = 0$, $\ell \in \{1, \ldots, K\}$, $S_{n}(\hat{\theta}) \neq 0$. Thus, by applying Lemma 2.8 and Theorem 2.7 (iii) in Van der Vaart (2000), it follows from (24) that

\[ \sqrt{n}(\hat{\theta}_{n} - \theta_{0}) \sim N(0, \Sigma(\theta_{0})). \]  

(25)

For an estimate of $\Sigma(\theta_{0})$, we use the sandwich estimator:

\[ \Sigma_{n}(\hat{\theta}) = \hat{\mathcal{J}}_{s}^{-1}(\hat{\theta}) \hat{\mathcal{J}}_{s}(\hat{\theta}) \hat{\mathcal{J}}_{s}^{-1}(\hat{\theta}), \]

where $\hat{\mathcal{J}}_{s}(\hat{\theta})$ and $\hat{\mathcal{J}}_{s}(\hat{\theta})$ are defined respectively by

\[ \hat{\mathcal{J}}_{s}(\hat{\theta}) = \frac{1}{K} \sum_{k=1}^{K} J_{s,k}(\hat{\theta}) \quad \text{and} \quad \hat{\mathcal{J}}_{s}(\hat{\theta}) = \frac{1}{K} \sum_{k=1}^{K} J_{s,k}(\hat{\theta}), \]

with $J_{s,k}(\hat{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_{x}(Y_{k}^{(i)}|\theta)$ calculated using the result of Theorem 3. Implementation of this estimation method is discussed in more details in Section 5.

**Remark 3.** Notice that the arguments used above are not applicable in order to improve the standard errors of the MLE $\hat{\theta}_{n}$ when applied to (5) to derive the estimator

\[ \hat{\theta}_{n} = \hat{\theta} + J_{s}^{-1}(\hat{\theta})S_{n}(\hat{\theta}), \]

where $J_{s,k}(\hat{\theta})$ and $\hat{\mathcal{J}}_{s}(\hat{\theta})$ coincide, with estimated covariance matrix $\Sigma_{n}(\hat{\theta})$.

### 4. Finite-mixture of Markov jump processes

To exemplify the results of Section 3, we consider maximum likelihood estimation of the distribution parameters of a finite mixture of Markov jump process $X = \{X_{t} : t \geq 0\}$ discussed in Frydman and Surya (2023). See Surya (2018, 2022b) for distributional properties of the finite mixture of Markov jump processes. The mixture process is a complex stochastic model which can be used to describe a sequence of events where the occurrence of an event depends not only on the current state, but also on the current time and past observations of the process. It may be considered as a nontrivial generalization of the Markov jump process (see, e.g., Norris, 2009) and has distributional equivalence with a conditional Markov jump process (Surya, 2022a). The later model allows the process to switch the transition rates from a finite number of transition matrices $Q_{m} = (q_{xy,m} : x, y \in S)$, $m = 1, \ldots, M$ when it moves from any phase $x$ of the state space $S = \{1, \ldots, p\}$, to another state $y \in S$ with switching probability depending on the current state, time, and its past information/sample paths. The later summarizes observable quantities of $X$ concerning the number of transitions $N_{xy}$ between states $(x, y) \in S$, occupation time $T_{xy}$ in each state $x \in S$, and initial state indicator $B_{0}$, having value one if $X_{0} = x$, or zero otherwise. Figure 1 depicts the transition diagram of a generalized Markov mixture model.

Beside the transition matrices $\{Q_{m} : m = 1, \ldots, M\}$, the distribution of $X$ is characterized by an initial probability $\alpha$ with $\alpha_{x} = \mathbb{P}(X_{0} = x)$ satisfying $\sum_{x=1}^{p} \alpha_{x} = 1$, and regime-selecting probability $\phi_{xy,m} = \mathbb{P}(X_{0} = x, X_{0}^{[m]}|X_{0} = y)$ which is the probability of making an initial transition w.r.t
a Markov process \( X^{(m)} = \{ X_t^{(m)} : t \geq 0 \} \), with transition matrix \( Q_m \), starting from a state \( x \in \mathcal{S} \). As the underlying Markov processes \( (X^{(m)}, Q_m) \), \( m = 1, \ldots, M \) are defined on the same state space \( \mathcal{S} \), there is a hidden information \( \Phi \) regarding which underlying Markov process that drives the movement of \( X \) when it makes a jump. The random variable \( \Phi \) has a categorical distribution with \( \mathbb{P}(\Phi = m) = p_m \) and \( \sum_m p_m = 1 \). The pair \( (X^k, \Phi^k) \) accounts for complete observation of \( k \)-th paths. Define \( \{ \Phi, \theta \} \) on \( \mathcal{S} \times \mathbb{R}^m \), with transition \( \Phi = 1 \) for each sample path \( X^k \). Consider \( n \) independent paths \( \{ X^k : k = 1, \ldots, n \} \) of \( X \) (generated data or real dataset). Following Frydman and Surya (2023), the complete-data log-likelihood function of \( (X^k, \Phi^k) \) is given by

\[
\log f_c(X^k, \Phi^k \mid \theta) = \sum_{m=1}^M \sum_{x=1}^p \Phi_{k,m} B^k_x \log \phi_{x,m} + \sum_{m=1}^M \sum_{x=1}^p \sum_{y,x=1}^p \Phi_{k,m} \left[ X^k_x \log q_{x,y} - q_{x,y} T^k_x \right], \tag{26}
\]

where \( \{ \Phi_{k,m} : x \in \mathcal{S}, 1 \leq m \leq M \} \) are subject to the constraint \( \sum_m \Phi_{k,m} = 1 \) for each \( x \in \mathcal{S} \). Note that \( X^k_x, T^k_x \) and \( B^k_x \) are observable quantities of the \( k \)-th sample path \( X^k \). It follows from the log-likelihood function (26) that the proportion \( \alpha_x \) can be estimated separately by \( \hat{\alpha}_x = \overline{B}_x / n \), with \( \overline{B}_x := \sum_k B^k_x. \) Since \( \hat{\alpha}_x \) does not involve \( \Phi \), it is therefore excluded from the estimation of the other parameters \( \theta \). Let \( \theta_0 = (\Phi_{0,m}, \Phi_{0,m}'_m : x, y \in \mathcal{S}, 1 \leq m \leq M) \) of the mixture.

### 4.1. The observed information matrix \( J_c(\theta) \)

To derive the elements of information matrices \( J_c(\theta) \) and \( J_c(\theta) \), define \( \hat{\Phi}_{k,m}(\theta) = \sum_{x=1}^p \Phi_{k,m}(\theta) B^k_x \), where

\[
\hat{\Phi}_{k,m}(\theta) = \frac{f_c(X^k, \Phi^k = m \mid \theta)}{\sum_{m=1}^M f_c(X^k, \Phi^k = m \mid \theta)}. \tag{27}
\]

Similarly defined for \( \hat{\Phi}_{k,m}(\theta) \) and \( \hat{\Phi}_{k,m}(\theta) \). Following the complete-data log-likelihood (26), we obtain

\[
J_c(\theta) = \frac{1}{n} \sum_{k=1}^n \left[ \frac{\partial^2 \log f_c(X^k, \Phi^k \mid \theta)}{\partial \phi_{x,m} \partial \phi_{y,l}} \right] = \left[ \frac{\overline{B}_{x,l}(\theta)}{n \phi_{x,m}^2} \hat{\delta}_{m}(l) + \frac{\overline{B}_{y,l}(\theta)}{n \phi_{y,l}^2} \hat{\delta}_{y,l}(m) \right] \delta_{y,l},
\]

with \( \delta_{y,l} = 1 \) if \( y = l \) or zero otherwise. Furthermore,

\[
J_c(q_{x,y}, q_{r,v}) = \frac{\delta_{r,v}(y) \delta_{x,y}(l)}{n q_{x,y} q_{r,v}} \hat{\Phi}_{x,y}(\theta),
\]

and

\[
J_c(q_{x,y}, q_{r,v}) = 0.
\]

It follows from the matrix structure that \( J_c(\theta) \) is of block diagonal and positive definite, hence is invertible. To be more precise, notice that the submatrix \( J_c(\theta_{x,m}, \theta_{y,l}) \) is of block diagonal with \( J_c(\theta_{x,m}, \theta_{x,m}) = 0 \) for \( y \neq x \) and

\[
J_c(\theta_{x,l}, \theta_{x,m}) = \begin{cases} \frac{\overline{B}_{x,l}(\theta)}{n \phi_{x,m}^2}, & \ell = m, \\ \frac{\overline{B}_{x,l}(\theta)}{n \phi_{y,l}^2}, & \ell \neq m, \end{cases}
\]

for \( \ell, m = 1, \ldots, M - 1 \). Thus, the \( (M - 1) \times (M - 1) \)-matrix \( [J_c(\theta_{x,l})]_{x,m} = J_c(\theta_{x,l}, \theta_{x,m}) \) reads as

\[
J_c(\theta_{x,l}) = D_n(\theta_{x,l}) + \beta_{x,M} \mathbf{1} \mathbf{1}'
\]

where \( D_n(\theta_{x,l}) \) is a diagonal matrix with \( D_n(\theta_{x,l})_{l,l} = \theta_{x,l} \), \( \beta_{x,M} = \overline{B}_{x,M}(\theta)/(n \phi_{x,M}^2) \) and \( \beta_{x,M} = \overline{B}_{x,M}(\theta)/(n \phi_{y,M}^2) \), whilst \( \mathbf{1} \) is a \((M - 1)\)-vector of one. Using the Sherman and Morrison (1950) formula, the inverse of \( J_c(\theta_{x,l}) \) is given by

\[
J_c^{-1}(\theta_{x,l}) = D_n^{-1}(\theta_{x,l}) + \frac{\beta_{x,M} D_n^{-1}(\theta_{x,l}) \mathbf{1} \mathbf{1}' D_n^{-1}(\theta_{x,l})}{1 + \beta_{x,M} \mathbf{1} \mathbf{1}' D_n^{-1}(\theta_{x,l})}.
\]
Furthermore, given that \( D_n^{-1}(\phi) \) is a diagonal matrix, the \((\ell, m)\)-element of the inverse \( J_n^{-1}(\phi) \) simplifies into

\[
|J_n^{-1}(\phi)|_{\ell,m} = \begin{cases} \frac{1}{d_{\phi,\ell}^2(a_\ell + b_{\phi,\ell} x_{m-1}^2 + c_{\phi,\ell})}, & \ell = m, \\ -\frac{b_{\phi,\ell}}{d_{\phi,\ell}^2(a_\ell + b_{\phi,\ell} x_{m-1}^2 + c_{\phi,\ell})}, & \ell \neq m. \end{cases}
\]

By invariance of \( J_n(\phi) \), for \( \theta \in \Theta \), one can use the EM-Gradient algorithm for faster convergent estimation of \( \theta_0 \).

### 4.2. The observed information matrix \( J_n(\phi) \)

The elements of the observed information matrix \( J_n(\phi) \) is given in Proposition 5 of Frydman and Surya (2023):

\[
J_n(\phi_{x,m}, \phi_{y,n}) = \delta(x) \sum_{k=1}^{K} \frac{\hat{\psi}_{x,m}(\theta) \hat{\psi}_{y,n}(\theta)}{\hat{\phi}_{x,m}} B_{x,m},
\]

where \( \hat{\psi}_{x,m}(\theta) = \hat{\phi}_{x,m} \theta - \theta_{m,n} \hat{\phi}_{x,m}(\theta) \). Moreover,

\[
J_n(q_{x,m}, q_{y,n}) = \hat{\delta}(r) \delta(v) \sum_{k=1}^{K} \frac{\hat{F}_{x,m}(\theta)}{\hat{q}_{x,m}} \frac{\hat{A}_{x,m}(\theta)}{\hat{q}_{x,m}} \frac{\hat{B}_{x,m}(\theta)}{\hat{q}_{x,m}},
\]

with \( \hat{\psi}_{x,m} = \hat{\phi}_{x,m} \theta - \theta_{m,n} \hat{\phi}_{x,m}(\theta) \), and

\[
J_n(q_{x,m}, q_{y,n}) = -\sum_{k=1}^{K} \hat{F}_{x,m}(\theta) \left( \delta_{x}(r) - \hat{\phi}_{x,m}(\theta) \right) \frac{\hat{A}_{x,m}(\theta)}{\hat{q}_{x,m}} \frac{\hat{B}_{x,m}(\theta)}{\hat{q}_{x,m}} + \delta_{x}(r) \sum_{k=1}^{K} \hat{F}_{x,m}(\theta) \frac{\hat{A}_{x,m}(\theta)}{\hat{q}_{x,m}} \frac{\hat{B}_{x,m}(\theta)}{\hat{q}_{x,m}}.
\]

Observe that unlike \( J_n(\phi) \), the observed information \( J_n(\theta) \) is in general not a sparse matrix, hence is more difficult to invert than \( J_n(\phi) \). In such case, we will use the iterative scheme (18) to get the inverse matrix \( J_n^{-1}(\phi) \).

### 4.3. The M-estimator \( \hat{\theta}_n \) and the EM algorithm

On recalling that \((X^k, \Phi^k)\) constitutes a complete dataset for the \( k \)-th observation, the M-criterion (7) reads

\[
\mathcal{M}_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ \log f(x^k, \Phi^k | \theta) | X^k, \theta_0 \right].
\]

As \( \hat{\theta}_n \) maximizes the M-criterion \( \mathcal{M}_n(\theta) \), we obtain

\[
\frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ \frac{\partial \log f(x^k, \Phi^k | \theta)}{\partial \phi_{x,m}} | X^k, \theta_0 \right] = \frac{\hat{B}_{x,m}(\theta)}{n \phi_{x,m}} - \frac{\hat{\phi}_{x,m}(\theta)}{n \phi_{x,m}},
\]

for \( 1 \leq m \leq M - 1 \) and

\[
\frac{1}{n} \sum_{k=1}^{n} \mathbb{E} \left[ \frac{\partial \log f(x^k, \Phi^k | \theta)}{\partial q_{x,m}} | X^k, \theta_0 \right] = \frac{\hat{N}_{x,m}(\theta)}{n q_{x,m}} - \frac{T_{x,m}(\theta)}{n}.
\]

### 5. Simulation study

This section verifies the main results of Section 3 using the example of conditional Markov jump process discussed in Section 4. For this purpose, the model parameters \( \theta_0 = (\phi^0_{x,m}, q^0_{x,m} : (x, r, \ell, m) \in S, \ell, m \in \{1, \ldots, M\} \) are set to have the following values used in Surya (2022a).

#### 5.1. Parameter value of \( \theta_0 \)

For simulation study, let \( S = \{1, 2, 3\} \) and \( M = 3 \). The value of initial probabilities \( a^1_i \) and \( \phi^0_{x,m} \) are presented in Table 1. The intensity matrices \( Q^0_1, Q^0_2 \) and \( Q^0_3 \) for the regime
5.2. Simulation and estimation results

We see following intensity matrices

\[ Q^1 = \begin{pmatrix} -2.0 & 1.2 & 0.8 \\ 0.2 & -0.4 & 0.2 \\ 1.2 & 1.8 & -3.0 \end{pmatrix}, \quad Q^2 = \begin{pmatrix} -3.0 & 2.4 & 0.6 \\ 0.2 & -0.4 & 0.2 \\ 0.4 & 1.6 & -2.0 \end{pmatrix}, \]

and

\[ Q^3 = \begin{pmatrix} -4.0 & 1.6 & 2.4 \\ 0.2 & -0.4 & 0.2 \\ 3.0 & 2.0 & -5.0 \end{pmatrix}, \]

respectively. We see following intensity matrices \( Q^m, m = 1, 2, 3 \), that each regime \( X^{(m)} \) has different expected state occupation time and the probability of making an immediate jump from one state to another, except for the transition from state 2. In the latter case, it is difficult to identify which underlying Markov jump process that drives the dynamics of \( X \) when it moves out of state 2.

Table 1: Parameter values for \( a^*_m \) and \( \phi^*_{x,n}, m = 1, 2, 3 \).

<table>
<thead>
<tr>
<th>State (x)</th>
<th>( a^*_x )</th>
<th>( \phi^*_{x,1} )</th>
<th>( \phi^*_{x,2} )</th>
<th>( \phi^*_{x,3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/3</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>1/3</td>
<td>0.25</td>
<td>0.55</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>1/3</td>
<td>0.6</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

of size \( n = 4000 \) each are generated. See Frydman and Surya (2023) for algorithm of generating sample paths. To each set \( X_k \) of paths, the MLE \( \hat{\theta}_k \) is found using the EM algorithm (28) and the EM-Gradient algorithm (20). In carrying out the statistical computation, the R language (2022) was used.

Figure 2 compares in each iteration the value of incomplete-data log-likelihood \( f_0(X_k|\hat{\theta}_k) \) of a randomly chosen set \( X_k \) of sample paths as a function of the current estimate \( \hat{\theta}_k \) obtained using the EM algorithm (28) against that of derived using the EM-Gradient algorithm (20). We observe that the EM-Gradient algorithm reaches its convergence faster than the EM algorithm as it requires less iterations to converge. Figure 3 shows estimation error \( ||\hat{\theta}_k - \bar{\theta}_k|| \) of the two algorithms. As we can see from the figure that the EM-Gradient algorithm is able to correct estimation error to reach its convergence faster than the EM algorithm. The explanation for this observation is due to the fact that the EM-algorithm is equivalent to the Fisher scoring iteration (19) whose successive estimation error \( ||\hat{\theta}_k - \bar{\theta}_k|| \) is determined by the inverse \( J_k^{-1}(\hat{\theta}_k) \), whereas for the EM-Gradient by \( J_k^{-1}(\hat{\theta}_k) \), which is smaller than \( J_k^{-1}(\hat{\theta}_k) \) in the sense of Loewner partial matrix ordering (23), attributed by the resulting loss of information presented in incomplete data.

To verify the asymptotic properties of the MLEs \( \hat{\theta}_k \), the mean squared error MSE \( = \frac{1}{K} \sum_{k=1}^{K}(\hat{\theta}_k - \theta_\ell)^2 \) is calculated. Subsequently, the information matrices \( J_k := J_k(\hat{\theta}_k) \) and \( J_\ell := J_\ell(\hat{\theta}_\ell) \) are evaluated. Then each set of matrices \( J_k : k = 1, \ldots, K \) and \( J_\ell : k = 1, \ldots, K \) are averaged to obtain \( \bar{J}_k = \frac{1}{K} \sum_{k=1}^{K} J_k \) and \( \bar{J}_\ell = \frac{1}{K} \sum_{k=1}^{K} J_\ell \). The inverse \( \bar{J}_\ell^{-1} \) of the averaged Fisher information \( \bar{J}_\ell \) is used to get estimated standard error of the MLE \( \hat{\theta}_\ell \), the Kolmogorov-Smirnov (KS) goodness-of-fit for the asymptotic distribution of the MLEs \( \hat{\theta}_k \). The table shows the convergence of \( \hat{\theta}_\ell \) to the true value \( \theta_\ell \) of the iterative estimator \( \Psi_\ell (18) \) of \( J_\ell^{-1}(\hat{\theta}_\ell) \) is used for \( \ell = 50 \) iteration steps to calculate the standard error of the MLE \( \hat{\theta}_\ell \). They are both compared to the sandwich estimator \( \hat{\Sigma}_n = \bar{J}_\ell^{-1} \bar{J}_\ell \). Note that in the computation of \( \Psi_\ell \), the matrices \( \bar{J}_k \) and \( \bar{J}_\ell \) are used in place of \( J_k(\hat{\theta}_\ell) \) and \( J_\ell(\hat{\theta}_\ell) \), respectively. Furthermore, since we consider the standard error of \( \hat{\theta}_\ell \), all covariance matrices \( \bar{J}_\ell^{-1}, \Psi_\ell \) and \( \Sigma_n \) are standardized by the sample size \( n \).

Table 2 presents the estimation results for the MLE \( \hat{\theta}_\ell \), the standard errors and the Kolmogorov-Smirnov (KS) goodness-of-fit for the asymptotic distribution of the MLEs \( \hat{\theta}_k \).
<table>
<thead>
<tr>
<th>$\theta$</th>
<th>True Value $\hat{\theta}_n$</th>
<th>Est. $\hat{\theta}_n$</th>
<th>RMSE</th>
<th>$\sqrt{\frac{\hat{\theta}_n}{n}}$</th>
<th>$\sqrt{\frac{\theta}{n}}$</th>
<th>$\sqrt{\frac{\Sigma}{n}}$</th>
<th>KS test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_{1,1}$</td>
<td>0.5000</td>
<td>0.5001</td>
<td>1.9653</td>
<td>1.9331</td>
<td>1.9331</td>
<td>1.0767</td>
<td>0.8670</td>
</tr>
<tr>
<td>$\phi_{1,2}$</td>
<td>0.3000</td>
<td>0.3011</td>
<td>1.7579</td>
<td>1.7687</td>
<td>1.7687</td>
<td>1.0084</td>
<td>0.4458</td>
</tr>
<tr>
<td>$\phi_{1,3}$</td>
<td>0.2500</td>
<td>0.2497</td>
<td>1.7939</td>
<td>1.8184</td>
<td>1.8184</td>
<td>0.8739</td>
<td>0.0557</td>
</tr>
<tr>
<td>$\phi_{2,1}$</td>
<td>0.5500</td>
<td>0.5502</td>
<td>1.8725</td>
<td>1.8627</td>
<td>1.8627</td>
<td>1.1420</td>
<td>0.8420</td>
</tr>
<tr>
<td>$\phi_{2,2}$</td>
<td>0.6000</td>
<td>0.6008</td>
<td>1.7988</td>
<td>1.7581</td>
<td>1.7581</td>
<td>1.1113</td>
<td>0.7539</td>
</tr>
<tr>
<td>$\phi_{3,1}$</td>
<td>0.1000</td>
<td>0.0995</td>
<td>1.2848</td>
<td>1.2371</td>
<td>1.2371</td>
<td>0.5909</td>
<td>0.7419</td>
</tr>
<tr>
<td>$\phi_{3,2}$</td>
<td>1.2000</td>
<td>1.2006</td>
<td>1.3919</td>
<td>1.6015</td>
<td>1.6015</td>
<td>1.1406</td>
<td>0.6706</td>
</tr>
<tr>
<td>$\phi_{3,3}$</td>
<td>0.8000</td>
<td>0.7998</td>
<td>1.1935</td>
<td>1.2252</td>
<td>1.2252</td>
<td>0.9338</td>
<td>0.9835</td>
</tr>
<tr>
<td>$\phi_{3,4}$</td>
<td>0.2000</td>
<td>0.1999</td>
<td>0.2497</td>
<td>0.2511</td>
<td>0.2511</td>
<td>0.1964</td>
<td>0.8575</td>
</tr>
<tr>
<td>$\phi_{3,5}$</td>
<td>0.2000</td>
<td>0.1996</td>
<td>0.2433</td>
<td>0.2506</td>
<td>0.2506</td>
<td>0.1965</td>
<td>0.2385</td>
</tr>
<tr>
<td>$\phi_{3,6}$</td>
<td>1.2000</td>
<td>1.1984</td>
<td>1.9141</td>
<td>2.0510</td>
<td>2.0510</td>
<td>1.3234</td>
<td>0.3910</td>
</tr>
<tr>
<td>$\phi_{3,7}$</td>
<td>1.8000</td>
<td>1.7968</td>
<td>2.0537</td>
<td>2.1452</td>
<td>2.1452</td>
<td>1.7263</td>
<td>0.0755</td>
</tr>
<tr>
<td>$\phi_{3,8}$</td>
<td>2.4000</td>
<td>2.3989</td>
<td>4.2487</td>
<td>4.0711</td>
<td>4.0711</td>
<td>2.4568</td>
<td>0.6955</td>
</tr>
<tr>
<td>$\phi_{3,9}$</td>
<td>0.6000</td>
<td>0.5990</td>
<td>1.8485</td>
<td>1.7905</td>
<td>1.7905</td>
<td>1.2667</td>
<td>0.7836</td>
</tr>
<tr>
<td>$\phi_{3,10}$</td>
<td>0.2000</td>
<td>0.1998</td>
<td>0.3052</td>
<td>0.2882</td>
<td>0.2882</td>
<td>0.2268</td>
<td>0.1171</td>
</tr>
<tr>
<td>$\phi_{3,11}$</td>
<td>0.2000</td>
<td>0.2000</td>
<td>0.2838</td>
<td>0.2884</td>
<td>0.2884</td>
<td>0.2268</td>
<td>0.7685</td>
</tr>
<tr>
<td>$\phi_{3,12}$</td>
<td>0.4000</td>
<td>0.3999</td>
<td>1.3428</td>
<td>1.3536</td>
<td>1.3536</td>
<td>0.8441</td>
<td>0.6424</td>
</tr>
<tr>
<td>$\phi_{3,13}$</td>
<td>1.6000</td>
<td>1.5994</td>
<td>2.2547</td>
<td>2.2802</td>
<td>2.2802</td>
<td>1.8279</td>
<td>0.5511</td>
</tr>
<tr>
<td>$\phi_{3,14}$</td>
<td>1.6000</td>
<td>1.5987</td>
<td>2.5430</td>
<td>2.4654</td>
<td>2.4654</td>
<td>2.1908</td>
<td>0.3697</td>
</tr>
<tr>
<td>$\phi_{3,15}$</td>
<td>2.4000</td>
<td>2.4019</td>
<td>3.3403</td>
<td>3.2976</td>
<td>3.2976</td>
<td>2.5562</td>
<td>0.9697</td>
</tr>
<tr>
<td>$\phi_{3,16}$</td>
<td>0.2000</td>
<td>0.2001</td>
<td>0.3471</td>
<td>0.3200</td>
<td>0.3200</td>
<td>0.2788</td>
<td>0.8565</td>
</tr>
<tr>
<td>$\phi_{3,17}$</td>
<td>0.2000</td>
<td>0.2000</td>
<td>0.2993</td>
<td>0.3198</td>
<td>0.3198</td>
<td>0.2787</td>
<td>0.6572</td>
</tr>
<tr>
<td>$\phi_{3,18}$</td>
<td>3.0000</td>
<td>3.0025</td>
<td>3.6650</td>
<td>4.0338</td>
<td>4.0338</td>
<td>3.2132</td>
<td>0.1953</td>
</tr>
<tr>
<td>$\phi_{3,19}$</td>
<td>2.0000</td>
<td>1.9986</td>
<td>3.1396</td>
<td>3.0636</td>
<td>3.0636</td>
<td>2.7344</td>
<td>0.8978</td>
</tr>
</tbody>
</table>

Table 2: Comparison between true value $\theta$, the MLE $\hat{\theta}_n$, and its standard errors using inverse of observed Fisher information $\frac{\hat{\Sigma}^{-1}(\hat{\theta}_n)}{n}$, iterative estimator $\frac{\Psi_{\ell}}{\hat{\theta}_n}$ (18), and the sandwich estimator $\frac{\hat{\Sigma}^{-1}(\hat{\theta}_n)}{n}$ for $K = 200$ independent repeated sampling of sample paths of size $n = 4000$. The last column provides the p-values of the Kolmogorov-Smirnov statistics.

5.2. Improved estimation by the M-estimator $\hat{\theta}_n$

To improve the estimation and reduce variability in the standard errors, we consider the M-estimator $\hat{\theta}_n$ defined as the maximizer of the M-criterion $\hat{\theta}_n(\theta) := \frac{1}{n} \sum_{k=1}^{n} \log f(X_k, \hat{\theta}_n | \theta) | X_k, \hat{\theta}_n |$ as discussed in Section 3.4. The components $\hat{\theta}_{k,m}$ and $\hat{\theta}_{k,m}$ of $\hat{\theta}_n$ are obtained by applying each randomly generated set $X_k = \{X_k^j : \ell = 1, \ldots, n\}, k \in \{k_1, \ldots, K\}$, of sample paths to the estimator (27) by replacing the true value $\theta_0$ by the MLE $\hat{\theta}_n$. The estimator $\hat{\theta}_n = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_n$ is reserved as the estimate of $\theta_0$. The mean squared error $\frac{1}{K} \sum_{k=1}^{K} (\hat{\theta}_n - \theta_0)^2$ and the observed information matrices $J_{\hat{\theta}} := J_{\hat{\theta}}(\hat{\theta}_n)$ and $J_{\hat{\theta}} := J_{\hat{\theta}}(\hat{\theta}_n)$ are evaluated based on $\hat{\theta}_n$. The standard errors from MLE are compared to those provided by the inverse $\hat{\Sigma}^{-1}$ of the averaged observed information matrix $\hat{\Sigma}$, the iterative estimator $\hat{\Sigma}$ (18) (for $\ell = 50$ iterations) and the sandwich estimator $\hat{\Sigma} = \hat{\Sigma} \hat{\Sigma}_n$ with $\hat{\Sigma}_n = \frac{1}{K} \sum_{k=1}^{K} J_{\hat{\theta}}$. As before, in the computation of $\Psi_{\ell}$, the matrices $\hat{\Sigma}$ and $\hat{\Sigma}_n$ are used in place of $J_{\hat{\theta}}(\hat{\theta}_n)$ and $J_{\hat{\theta}}(\hat{\theta}_n)$, respectively.

The results are presented in Table 3. In contrast to the results presented in Table 2, the M-estimator $\hat{\theta}_n$ shows some improvements on the estimation in terms of overall being closer to the true value $\theta_0$ than the MLE $\hat{\theta}_n$ to $\theta_0$ with smaller standard errors (RMSEs) than RMSEs of $\hat{\theta}_n$, given by $\hat{\Sigma}^{-1}(\hat{\theta}_n)$. The table shows the convergence of $\Psi_{\ell}$ to $\hat{\theta}_n$, confirming the result of Theorem 4. The convergence of $\hat{\theta}_n$ to $\theta_0$ makes the RMSEs close to the sample variance of $\hat{\theta}_n$. We observe that RMSEs are much closer to the standard errors produced by the sandwich estimator $\Sigma_n$ than those given by $\hat{\Sigma}^{-1}(\hat{\theta}_n)$. The last column corresponds to the p-values of the KS statistics. All p-values are larger than $\alpha = 5\%$ confirming the asymptotic normal distribution of the M-estimator $\hat{\theta}_n$ with finite-sample covariance matrix $\Sigma_n$ as stated in Theorem 5, in particular (25) for the repeated sampling method.
### Table 3: Comparison between true value $\theta_0$, the M-estimator $\hat{\theta}_n$ and its standard errors using inverse of observed Fisher information $\sqrt{\frac{\hat{\theta}_n}{\hat{J}_{\theta\theta}}}$, iterative estimator $\frac{\hat{\theta}_n}{\hat{J}_{\theta\theta}}$ (18), and the sandwich estimator $\frac{\hat{\theta}_n}{\hat{J}_{\theta\theta}}$ for $K = 200$ independent repeated sampling of sample paths of size $n = 4000$. The last column provides the p-values of the Kolmogorov-Smirnov statistics.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>True Value</th>
<th>Est. $\frac{\hat{\theta}<em>n}{\hat{J}</em>{\theta\theta}}$</th>
<th>RMSE</th>
<th>$\sqrt{\frac{\hat{\theta}<em>n}{\hat{J}</em>{\theta\theta}}}$</th>
<th>$\sqrt{\frac{\hat{\theta}<em>n}{\hat{J}</em>{\theta\theta}}}$</th>
<th>$\sqrt{\frac{\hat{\theta}<em>n}{\hat{J}</em>{\theta\theta}}}$</th>
<th>KS test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_{1,1}$</td>
<td>0.5000</td>
<td>0.5004</td>
<td>1.1175</td>
<td>1.0779</td>
<td>1.9388</td>
<td>1.9387</td>
<td>0.6145</td>
</tr>
<tr>
<td>$\phi_{1,2}$</td>
<td>0.3000</td>
<td>0.3003</td>
<td>1.0836</td>
<td>1.0098</td>
<td>1.7741</td>
<td>1.7741</td>
<td>0.7321</td>
</tr>
<tr>
<td>$\phi_{2,1}$</td>
<td>0.2500</td>
<td>0.2501</td>
<td>0.8180</td>
<td>0.8773</td>
<td>1.8214</td>
<td>1.8214</td>
<td>0.4728</td>
</tr>
<tr>
<td>$\phi_{2,2}$</td>
<td>0.5500</td>
<td>0.5488</td>
<td>1.1116</td>
<td>1.1545</td>
<td>1.8659</td>
<td>1.8658</td>
<td>0.3376</td>
</tr>
<tr>
<td>$\phi_{3,1}$</td>
<td>0.6000</td>
<td>0.6006</td>
<td>1.1202</td>
<td>1.1110</td>
<td>1.7662</td>
<td>1.7662</td>
<td>0.4826</td>
</tr>
<tr>
<td>$\phi_{3,2}$</td>
<td>0.1000</td>
<td>0.1007</td>
<td>0.6369</td>
<td>0.5987</td>
<td>1.2513</td>
<td>1.2513</td>
<td>0.1117</td>
</tr>
<tr>
<td>$q_{12,1}$</td>
<td>1.2000</td>
<td>1.2001</td>
<td>1.0790</td>
<td>1.1395</td>
<td>1.6035</td>
<td>1.6035</td>
<td>0.6775</td>
</tr>
<tr>
<td>$q_{13,1}$</td>
<td>0.8000</td>
<td>0.8003</td>
<td>0.8610</td>
<td>0.9335</td>
<td>1.2261</td>
<td>1.2261</td>
<td>0.6608</td>
</tr>
<tr>
<td>$q_{21,1}$</td>
<td>0.2000</td>
<td>0.1998</td>
<td>0.1957</td>
<td>0.1962</td>
<td>0.2513</td>
<td>0.2513</td>
<td>0.3156</td>
</tr>
<tr>
<td>$q_{23,1}$</td>
<td>0.2000</td>
<td>0.1999</td>
<td>0.1989</td>
<td>0.1963</td>
<td>0.2508</td>
<td>0.2508</td>
<td>0.6091</td>
</tr>
<tr>
<td>$q_{31,1}$</td>
<td>1.2000</td>
<td>1.1991</td>
<td>1.3643</td>
<td>1.3217</td>
<td>2.0585</td>
<td>2.0585</td>
<td>0.8254</td>
</tr>
<tr>
<td>$q_{32,1}$</td>
<td>1.8000</td>
<td>1.7982</td>
<td>1.6917</td>
<td>1.7252</td>
<td>2.1467</td>
<td>2.1467</td>
<td>0.2406</td>
</tr>
<tr>
<td>$q_{12,2}$</td>
<td>2.4000</td>
<td>2.3979</td>
<td>2.4644</td>
<td>2.4559</td>
<td>4.0896</td>
<td>4.0896</td>
<td>0.7365</td>
</tr>
<tr>
<td>$q_{13,2}$</td>
<td>0.6000</td>
<td>0.5989</td>
<td>1.2403</td>
<td>1.2670</td>
<td>1.7960</td>
<td>1.7960</td>
<td>0.5133</td>
</tr>
<tr>
<td>$q_{21,2}$</td>
<td>0.2000</td>
<td>0.2000</td>
<td>0.2261</td>
<td>0.2269</td>
<td>0.2889</td>
<td>0.2889</td>
<td>0.5564</td>
</tr>
<tr>
<td>$q_{23,2}$</td>
<td>0.2000</td>
<td>0.2002</td>
<td>0.2249</td>
<td>0.2269</td>
<td>0.2892</td>
<td>0.2892</td>
<td>0.0983</td>
</tr>
<tr>
<td>$q_{31,2}$</td>
<td>0.4000</td>
<td>0.3999</td>
<td>0.8562</td>
<td>0.8450</td>
<td>1.3570</td>
<td>1.3570</td>
<td>0.6364</td>
</tr>
<tr>
<td>$q_{32,2}$</td>
<td>1.6000</td>
<td>1.6003</td>
<td>1.7598</td>
<td>1.8286</td>
<td>2.2857</td>
<td>2.2857</td>
<td>0.8173</td>
</tr>
<tr>
<td>$q_{12,3}$</td>
<td>1.6000</td>
<td>1.6005</td>
<td>2.1762</td>
<td>2.1921</td>
<td>2.4698</td>
<td>2.4698</td>
<td>0.8779</td>
</tr>
<tr>
<td>$q_{13,3}$</td>
<td>2.4000</td>
<td>2.4017</td>
<td>2.6142</td>
<td>2.5553</td>
<td>3.3039</td>
<td>3.3039</td>
<td>0.4707</td>
</tr>
<tr>
<td>$q_{21,3}$</td>
<td>0.2000</td>
<td>0.1998</td>
<td>0.3005</td>
<td>0.2788</td>
<td>0.3205</td>
<td>0.3205</td>
<td>0.3840</td>
</tr>
<tr>
<td>$q_{23,3}$</td>
<td>0.2000</td>
<td>0.1999</td>
<td>0.2668</td>
<td>0.2785</td>
<td>0.3199</td>
<td>0.3199</td>
<td>0.8889</td>
</tr>
<tr>
<td>$q_{31,3}$</td>
<td>3.0000</td>
<td>3.0006</td>
<td>3.1549</td>
<td>3.2128</td>
<td>4.0344</td>
<td>4.0344</td>
<td>0.7093</td>
</tr>
<tr>
<td>$q_{32,3}$</td>
<td>2.0000</td>
<td>2.0008</td>
<td>2.7380</td>
<td>2.7344</td>
<td>3.0657</td>
<td>3.0657</td>
<td>0.4681</td>
</tr>
</tbody>
</table>

### 6. Concluding Remarks

This paper developed some new results on maximum likelihood estimation from incomplete data. The novelty of the approach is based on the Halmos and Savage (1949) work on the theory of sufficient statistics. The results have largely remained unexamined as far as the literatures on incomplete data are concerned, see e.g., Schervish (1995), McLachlan and Krishnan (2008), and Little and Rubin (2020). Conditional observed information matrices are introduced and utilized to their greater extent to which their finite-sample properties are established. They possess the same Loewner partial matrix ordering properties as the expected Fisher information matrices do. In particular, we show the resulting loss of information holds for finite sample of incomplete data. In its new form, the observed Fisher information of incomplete data simplifies Louis (1982) formula for the same matrix in terms of simplifying the conditional expectation of the outer product of the complete-data score function. Also, it directly verifies asymptotic consistency of the matrix. Our method of derivation of the matrix is direct and much simplified compared to those given by Louis (1982) and McLachlan and Krishnan (2008). To avoid getting an incorrect inverse of the observed information matrix, which may be attributed by the lack of sparsity and large size of the matrix, a monotone convergence recursive equation for estimator of the inverse matrix is developed. An improved estimation by the M-estimator is proposed using repeated sampling method. A consistent sandwich estimator of covariance matrix of the M-estimator is proposed using the conditional information matrices. It extends the Huber sandwich estimator (Huber, 1967; Freedman, 2006; Little and Rubin, 2020) to model misspecification under incomplete data. The main appealing feature of the proposed sandwich estimator is that unlike its counterpart, it does not involve the inverse of observed Fisher information matrix. The standard errors of the M-estimator produced by the sandwich estimator are smaller than those given by Louis (1982) and McLachlan and Krishnan (2008). To avoid getting an incorrect inverse of the observed information matrix, which may be attributed by the lack of sparsity and large size of the matrix, a monotone convergence recursive equation for estimator of the inverse matrix is developed. An improved estimation by the M-estimator is proposed using repeated sampling method. A consistent sandwich estimator of covariance matrix of the M-estimator is proposed using the conditional information matrices. It extends the Huber sandwich estimator (Huber, 1967; Freedman, 2006; Little and Rubin, 2020) to model misspecification under incomplete data. The main appealing feature of the proposed sandwich estimator is that unlike its counterpart, it does not involve the inverse of observed Fisher information matrix. The standard errors of the M-estimator produced by the sandwich estimator are smaller than those of the MLE given by the inverse of the observed information matrix. The difference is attributed to the resulting information loss presented in the incomplete data. The simulation study on a complex stochastic model of finite mixture of
Markov jump processes confirms the accuracy of the main results presented in Sections 3 and 4. We believe that the results should offer potential for variety of applications for maximum likelihood estimation from incomplete data.

Acknowledgments

Part of this work was carried out while Budh Surya was visiting the Department of Technology, Operations and Statistics of New York University Stern School of Business in September 2019. He thanks Professor Halina Frydman for the invitation and for the hospitality provided during his stay at the NYU Stern.

References


