Method for Computation of the Fisher Information Matrix in the Expectation–Maximization Algorithm

Lingyao Meng
Department of Applied Mathematics & Statistics
The Johns Hopkins University
Baltimore, Maryland 21218

Abstract

The expectation–maximization (EM) algorithm is an iterative computational method to calculate the maximum likelihood estimators (MLEs) from the sample data. It converts a complicated one-time calculation for the MLE of the incomplete data to a series of relatively simple calculations for the MLEs of the complete data. When the MLE is available, we naturally want the Fisher information matrix (FIM) of unknown parameters. The FIM is, in fact, a good measure of the amount of information a sample of data provides and can be used to determine the lower bound of the variance and the asymptotic variance of the estimators. However, one of the limitations of the EM is that the FIM is not an automatic by-product of the algorithm. In this paper, we review some basic ideas of the EM and the FIM. Then we construct a simple Monte Carlo-based method requiring only the gradient values of the function we obtain from the E step and basic operations. Finally, we conduct theoretical analysis and numerical examples to show the efficiency of our method. The key part of our method is to utilize the simultaneous perturbation stochastic approximation method to approximate the Hessian matrix from the gradient of the conditional expectation of the complete-data log-likelihood function.

Key words: Fisher information matrix, EM algorithm, Monte Carlo, Simultaneous perturbation stochastic approximation
1. Introduction

The expectation–maximization (EM) algorithm introduced by Dempster, Laird and Rubin in 1977 is a well-known method to compute the MLE iteratively from the observed data. It is applied to problems in which the observed data is incomplete or the log-likelihood function of the observed data is so complicated that obtaining the MLE from it directly is too difficult. There are some books on EM algorithm such as McLachlan and Krishnan (1997), Watanabe and Yamaguchi (2004), and Gupta and Chen (2011). The superiority of the EM algorithm comes from its simplicity. It obtains a sequence of results from the simple complete data log-likelihood function, thereby avoiding calculations from the complicated incomplete data log-likelihood function. Dempster, Laird, and Rubin have proved that this sequence has the monotonicity property of increase with each iteration. However, because the incomplete (i.e., observed) data log-likelihood function is not directly used in the EM algorithm, we cannot directly obtain the Fisher information matrix (FIM) for the observed data. This paper presents a general method for computing the FIM in the EM setting.

The FIM plays a key role in uncertainty calculation and other aspects of estimation for a wide range of statistical methods and applications, including parameter estimation (e.g., Abt and Welch, 1998; Spall, 2014), system identification (Ljung, 1999, pp. 212–221), and experimental design (Pronzato, 2008; Spall, 2010), among other areas. Hence, it is critical to obtain the FIM in a practical manner. Both the expected FIM (the expectation of the square of the gradient of the incomplete data log-likelihood functions) at the MLE and the observed FIM (the negative Hessian matrix of the incomplete data log-likelihood functions evaluated at the observed data) at the MLE can be used to calculate the Cramer-Rao lower bound and asymptotic distribution of
the MLE. Some methods have been introduced to estimate the FIM. In Spall (2005), for example, the author constructed an efficient method to approximate the expected FIM based on Monte Carlo simulation. Extensions are given in Spall (2008) and Das et al. (2010). Moreover, Berisha and Hero (2015) introduced a method to estimate the FIM directly from the sample data without the density function.

Given the importance of the FIM and the popularity of the EM, some research has been done to calculate the FIM from the EM. Louis (1982) provides a closed form of the observed FIM (the negative Hessian matrix of incomplete data) that requires the negative conditional expectation of the Hessian matrix of the complete data with respect to the incomplete data and the conditional expectation of the square of gradient of complete data log-likelihood function with respect to the incomplete data. In Meng and Rubin (1991), the authors introduced the supplemented EM (SEM) algorithm to calculate the observed FIM.

However, all these methods have limitations. One common disadvantage is that we can only obtain the observed information matrix from these methods, yet the expected Fisher information matrix at MLE is a maximum likelihood estimator of the expected Fisher information matrix at the true value. Furthermore, Cao and Spall (2012) and Cao (2013) demonstrated that:

*Under reasonable conditions similar to standard MLE conditions, the inverse expected FIM outperforms the inverse observed FIM under a mean squared error criterion.*

In addition, the formulas from Louis (1982) work well for some relatively simple problems. Nevertheless, when the log-likelihood function of the complete data is more complicated, the calculating of the negative conditional expectation of the Hessian matrix of the
complete data log-likelihood function with respect to the incomplete data is difficult to accomplish. Additionally, the conditional expectation of the square of the complete data score function is likewise hard to calculate. In the SEM algorithm, similarly, the calculation of the negative conditional expectation of the Hessian matrix is sometimes infeasible. The SEM algorithm, furthermore, requires running the EM algorithm many times, and the number of iterations depends on the number of iterations in the EM algorithm. However, the cost of running the EM algorithm is high, and its convergence rate is slow.

Using simultaneous perturbation stochastic approximation (SPSA) (Spall, 1992, 1998; Bhatnagar et al., 2013), our method takes the numerical differentiation to the gradient of the conditional expectation of the complete-data log-likelihood function. If the gradient is unavailable, we can approximate it with the value of the conditional expectation of the complete-data log-likelihood function, which can be obtained by running the E step. Then through the Monte Carlo simulation, we can approximate the expected FIM by calculating the negative average of all the approximate Hessian matrices.

Section 2 in this article covers the basics of the FIM and the EM algorithm. Section 3 presents our Monte Carlo-based method. The numerical examples and theoretical analysis to show the efficiency of our method are in Sections 4 and 5, respectively.
2. Review of EM Algorithm

In order to have a good understanding for further discussion, we are going to introduce basic information on the EM algorithm, (the expected and the observed) FIM and some existing methods to calculate the FIM in the EM algorithm in this section.

2.1 The EM algorithm

Let \(X\) and \(Y\) be the missing data and the observed (incomplete) data, respectively; so \(Z = (X, Y)\) is the complete data. We also suppose that the density function of \(Z\) is:

\[
p_Z(z|\theta) = p_{X,Y}(x,y|\theta),
\]
and the density function of \(Y\) is \(p_Y(y|\theta)\). Let the log-likelihood function of the complete data be:

\[
L(\theta|Z) = L(\theta|X, Y) = \log\{p_Z(z|\theta)\}.
\]

Let the log-likelihood function of the observed data be:

\[
L_O(\theta|Y) = \log\{p_Y(y|\theta)\}.
\]

The purpose of the EM algorithm is to calculate the MLE of \(\theta\), say \(\theta^*\), from the observed data \(Y\).

The EM algorithm operates as follows. The Expectation (E) step: Calculate the expectation of the log-likelihood function \(L(\theta|X,Y)\) with respect to the conditional distribution of \(X\) given \(Y\) under the current estimate of the parameters \(\theta^{(t)}\),

\[
Q(\theta|\theta^{(t)}) = E[L(\theta|X,Y)|Y,\theta^{(t)}].
\]
The Maximization (M) step: Find the parameters \( \theta \) that maximize \( Q(\theta|\theta^{(t)}) \) for fixed \( \theta^{(t)} \),

\[
\theta^{(t+1)} = \arg \max Q(\theta|\theta^{(t)}).
\]

In each iteration of the EM algorithm, we need to run the E step and the M step only once. This process is repeated until the difference between \( L_O(\theta^{(t+1)}|Y) \) and \( L_O(\theta^{(t)}|Y) \) is less than a particular positive number \( \delta \).

It was shown in Dempster, Laird and Rubin (1977) that the sequence \( L_O(\theta^{(t)}|Y) \) is a nondecreasing sequence. Further, it was shown in Wu (1983) and Vaida (2005) that the log-likelihood converges to the value at \( \theta^* \) and \( \theta^{(t)} \) converges to \( \theta^* \) when \( t \) approaches infinity.

Define the mapping \( M(\theta^{(t)}) = \theta^{(t+1)} \) and \( DM \) is the Jacobian matrix of \( M \) at \( \theta^* \).

### 2.2 The Fisher Information Matrix

The FIM is a good measure of the amount of information the sample data can provide about parameters. Suppose \( p(\theta; x) \) is the density function of the object model and \( L(\theta; x) = \log(p(\theta; x)) \) is the log-likelihood function. We can define the expected FIM as:

\[
E \left[ \frac{\partial L(\theta|X)}{\partial \theta} \frac{\partial L(\theta|X)}{\partial \theta^T} \right].
\]

Under the condition in which we can interchange the order of integration and differentiation,

\[
E \left[ \frac{\partial L(\theta|X)}{\partial \theta} \right] = \int \frac{\partial L(\theta|X)}{\partial \theta} p(\theta; x) dx = \int \frac{\partial p(\theta|x)}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int p(\theta|x) dx = 0.
\]

Therefore the expected FIM also equals:
Additionally, from (2.2.1):

\[
E \left[ \frac{\partial L(\theta|X)}{\partial \theta} \frac{\partial L(\theta|X)}{\partial \theta^T} \right] = -E \left[ \frac{\partial^2 L(\theta|X)}{\partial \theta \partial \theta^T} \right] = -E[H(\theta|X)].
\] (2.2.2)

We then define the observed FIM as:

\[
\frac{\partial^2 L(\theta|X)}{\partial \theta \partial \theta^T} = - \left( \frac{\partial^2 L(\theta|X)}{\partial \theta \partial \theta^T} \right) = -H(\theta|X).
\]

2.3 Some existing methods

Louis (1982) presented the following formula:

\[
I_{\text{Louis}}(\theta^*) = - \left\{ E \left[ \frac{\partial^2 L(\theta|X, Y)}{\partial \theta \partial \theta^T} \right] | Y, \theta^* \right\}_{\theta = \theta^*}
\]

\[
- \left\{ E \left[ \frac{\partial L(\theta|X, Y)}{\partial \theta} \frac{\partial L(\theta|X, Y)}{\partial \theta^T} \right] | Y, \theta^* \right\}_{\theta = \theta^*}
\]

\[
+ \left\{ E \left[ \frac{\partial L(\theta|X, Y)}{\partial \theta} | Y, \theta^* \right] E \left[ \frac{\partial L(\theta|X, Y)}{\partial \theta^T} | Y, \theta^* \right] \right\}_{\theta = \theta^*}.
\]

The formula in Oakes (1999) is:

\[
I_{\text{Oakes}}(\theta^*) = \left\{ \frac{\partial^2 Q(\theta|\theta^*)}{\partial \theta^2} + \frac{\partial^2 Q(\theta|\theta^*)}{\partial \theta \partial \theta^*} \right\}_{\theta = \theta^*},
\]

whereas in the SEM algorithm,

\[
I_{\text{SEM}}(\theta^*) = (I - DM) \left\{ E \left[ - \left( \frac{\partial^2 L(\theta|X, Y)}{\partial \theta \partial \theta^T} \right) | Y, \theta^* \right] \right\}_{\theta = \theta^*}.
\]
where $I$ is the identity matrix.

3. Computation of the Expected FIM in the EM Algorithm

In this section, we continue to use the same notations as in Section 2.1. According to the laws of conditional probability,

$$
p_{X,Y}(x, y|\theta) = p_{X|Y}(x|y, \theta)p_Y(y|\theta).
$$

Then

$$
\log\{p_{X,Y}(x, y|\theta)\} = \log\{p_{X|Y}(x|y, \theta)\} + \log\{p_Y(y|\theta)\}.
$$

Take expectations on both sides of the above equation with respect to the conditional distribution of $X|Y, \theta^{(t)}$, and we have:

$$
Q(\theta|\theta^{(t)}) = E[\log\{p_{X|Y}(x|y, \theta)\} | Y, \theta^{(t)}] + L_O(\theta|Y). \tag{3.1}
$$

If we take the derivative of $\theta$ and let $\theta = \theta^{(t)}$, under the conditions where we can interchange the order of expectation with respect to $X|Y$ and the differentiation of $\theta$, the equation (3.1) implies:

$$
\left[ \frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta} \right]_{\theta = \theta^{(t)}} = E \left[ \frac{\partial p_{X|Y}(x|y, \theta)}{\partial \theta} \frac{1}{p_{X|Y}(x|y, \theta)} \right]_{\theta = \theta^{(t)}} Y, \theta = \theta^{(t)} + \left[ \frac{\partial L_O(\theta|Y)}{\partial \theta} \right]_{\theta = \theta^{(t)}} \tag{3.2}
$$

The first part of the right hand side of the equation (3.2) equals 0. Therefore,

$$
\left[ \frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta} \right]_{\theta = \theta^{(t)}} = \left[ \frac{\partial L_O(\theta|Y)}{\partial \theta} \right]_{\theta = \theta^{(t)}}. \tag{3.3}
$$
While we know nothing about the value of $L_O(\theta|Y)$ or, subsequently, the gradient of $L_O(\theta|Y)$, we can obtain the value of $Q(\theta|\theta^{(t)})$ by running the E step and then the gradient of $Q(\theta|\theta^{(t)})$. Note that both sides of the equation (3.3) can be regarded as functions of $\theta^{(t)}$. Therefore, we define:

$$S(\theta^{(t)}|Y) = \left[ \frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta} \right]_{\theta=\theta^{(t)}}$$ (3.4)

Equation (3.3) implies that we can use $S(\theta|Y)$ to take place of the gradient of $L_O(\theta|Y)$. Then the Hessian matrix can be obtained by numerical differentiation on $S(\theta|Y)$.

There are many numerical differentiation methods, but, due to its efficiency in multivariate problems, we prefer the SPSA method introduced by Spall (1992), Spall (2000), and Spall (2005). Let $\hat{H}^{(k)}(\theta|Y_k)$ be the estimator of the Hessian matrix for the data set $Y_k$ and $G(\theta|Y_k)$ be the gradient of the log-likelihood function for the data set $Y_k$. The formula to approximate the Hessian matrix is:

$$\hat{H}^{(k)}(\theta|Y_k) = \frac{1}{2} \left\{ \frac{\delta G_k}{2} \left[ \Delta^{-1}_{k1}, \Delta^{-1}_{k2}, \ldots, \Delta^{-1}_{kd} \right] + \left( \frac{\delta G_k}{2} \left[ \Delta^{-1}_{k1}, \Delta^{-1}_{k2}, \ldots, \Delta^{-1}_{kd} \right] \right)^T \right\},$$ (3.5)

where $\delta G_k = G(\theta + \Delta_k|Y_k) - G(\theta - \Delta_k|Y_k)$ and $\Delta_k = [\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kd}]^T$ is a random perturbation vector. According to Spall (2005), for any $k$ and any $1 \leq j \leq d$, $\Delta_{kj}$ should satisfy:

- $\Delta_{kj}$ is a mean-zero and a symmetrically distributed random variable;
- every $\Delta_{kj}$ is uniformly bounded and $E[1/|\Delta_{kj}|] < \infty$; and
- The $\Delta_{kj}$ are independent and identically distributed across $k$ and $j$. 


A good choice that satisfies the above conditions is the Bernoulli distribution with half the probability being $c$ and half the probability being $-c$. Note in the right hand side of (3.5), we add the transpose term in order to make the estimator symmetric. Moreover, by combining (3.4) and (3.5), the formula to estimate the Hessian matrix of the incomplete data log-likelihood function at the MLE $\theta^*$ for the data set $Y_k$ is:

$$
\hat{H}^{(k)}(\theta^* | Y_k) = \frac{1}{2} \left\{ \frac{\delta S_k}{2} \left[ \Delta_{k1}^{-1}, \Delta_{k2}^{-1} \ldots \Delta_{kd}^{-1} \right] + \left( \frac{\delta S_k}{2} \left[ \Delta_{k1}^{-1}, \Delta_{k2}^{-1} \ldots \Delta_{kd}^{-1} \right] \right)^T \right\}, \quad (3.6)
$$

where $\delta S_k = S(\theta^* + \Delta_k | Y_k) - S(\theta^* - \Delta_k | Y_k)$. In the case where the value of $S(\theta^* \pm \Delta_k | Y_k)$ is unavailable, we can approximate it by the numerical differentiation on $Q(\theta | \theta^* \pm \Delta_k)$, utilizing the following formula and then substituting the estimator of $S(\theta^* \pm \Delta_k | Y_k)$ into (3.5).

$$
\tilde{S}(\theta^* \pm \Delta_k | Y_k) = \frac{Q(\theta^* \pm \Delta_k + \tilde{\Delta}_k | \theta^* \pm \Delta_k) - Q(\theta^* \pm \Delta_k - \tilde{\Delta}_k | \theta^* \pm \Delta_k)}{2} \begin{pmatrix} \Delta_{k1}^{-1} \\ \Delta_{k2}^{-1} \\ \vdots \\ \Delta_{kd}^{-1} \end{pmatrix}, \quad (3.7)
$$

In equation (3.7), $\tilde{\Delta}_k = [\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2} \ldots, \tilde{\Delta}_{kd}]^T$ is a perturbation vector that satisfies the same conditions as $\Delta_k$. The superiority of the SPSA lies in the fact that for each Hessian matrix estimation, the SPSA requires two values of $S$ when $S$ is available with respect to $2d$ values of $S$ using the traditional finite difference method. When $S$ is unavailable, compared to $2d(d + 1)$ values of $Q$ used in the finite difference method, SPSA only needs four $Q$ values for any dimension $d$.

From (2.2.2), we can approximate the expected FIM by calculating the negative average of all estimations of the Hessian matrix. The algorithm to compute the expected FIM operates as follows:
1. Run the EM algorithm until the difference between $L_O(\theta^{(t^*+1)}|Y)$ and $L_O(\theta^{(t^*)}|Y)$ is less than a particular $\delta$, and then obtain the MLE $\theta^*$, at which we would like to determine the Hessian matrix. Set the small number $c$ in Bernoulli $\pm c$ distribution to generate the perturbation vectors $\Delta_k$ and $\hat{\Delta}_k$ (if needed). Set the number $N$, which is the number of data sets we would like to generate.

2. Generate $N$ data sets by Monte Carlo simulation and denote them by $Y_1, Y_2, ...$ and $Y_N$ respectively.

3. Do one of the following two actions:
   - In the case where the values of $S$ for different $\theta$ and data sets are calculable, generate $N$ perturbation vectors following Bernoulli $\pm c$ distribution and denote them by $\Delta_1, \Delta_2$ and $\Delta_N$ respectively. Set $k = 1$. Jump to step 4.
   - If the values of $S$ are unavailable, generate $2N$ perturbation vectors following Bernoulli $\pm c$ distribution and denote them by $\Delta_1, \Delta_2, ... \Delta_N, \hat{\Delta}_1, \hat{\Delta}_2, ...$ and $\hat{\Delta}_N$ respectively. Set $k = 1$. Jump to step 5.

4. Apply formula (3.6) to calculate $\hat{H}^{(k)}(\theta^*|Y_k)$ using $\Delta_k$ and $Y_k$. Set $k = k + 1$. Jump to step 6.

5. Run the E step of the EM algorithm four times to get $Q(\theta^* + \Delta_k|\theta^* + \Delta_k)$ and $Q(\theta^* + \Delta_k - \hat{\Delta}_k|\theta^* + \Delta_k)$. Apply formula (3.7) to calculate $\hat{S}(\theta^* + \Delta_k|Y_k)$. Substitute $\hat{S}(\theta^* + \Delta_k|Y_k)$ into (3.6) to obtain $\hat{H}^{(k)}(\theta^*|Y_k)$. Set $k = k + 1$. Jump to step 6.

6. Repeat step 4 or 5 until $k = N$. Finally, calculate the negative average of all $\hat{H}^{(k)}(\theta^*|Y_k)$s, which is the approximation of the expected FIM.
We suppose the $\theta^*$ is available after running the EM algorithm. To compute the expected information matrix, the cost of the above algorithm is $N$ data sets, $N$ perturbation vectors, $2N$ $S$ values, and $O(N)$ basic operations if the $S$ values are available. If not, our algorithm requires $N$ data sets, $2N$ perturbation vectors, $4N$ times E step, and $O(N)$ basic operations.

This algorithm also can apply to the observed information matrix. Specifically, we approximate the Hessian matrix for one data set at $\theta^*$ plus/minus different perturbation vectors, and then calculate the negative average of them.

4. Theoretical Analysis

There are two parts in this session. One is the comparison between our algorithm and the SEM algorithm in Meng and Rubin (1991). The other part compares our algorithm with the SP-based method in Spall (2005) to approximate the FIM in general settings.

4.1 Comparison with the SEM algorithm

One important question is: Is our method better than previous methods to calculate the FIM in the EM algorithm? However, it seems impossible to compare our method to all the existing methods; therefore, we consider the relative performance between our method and the famous SEM method.

Recall the SEM method. To approximate $DM$, Meng and Rubin have:
\[
    r^{(t)}_{ij} = \frac{M_j(\theta^*_1, \theta^*_2, ..., \theta^*_i(t), \theta^*_{i+1} ... \theta^*_d) - M_j(\theta^*)}{\theta^*_i(t) - \theta^*_i} 
\]

where \( \theta^* = (\theta^*_1, \theta^*_2, ..., \theta^*_i ... \theta^*_d)^T \) is the MLE acquired from the EM algorithm, and \( \theta^{(t)} = (\theta^{(t)}_1, \theta^{(t)}_2 ... , \theta^{(t)}_i ... \theta^{(t)}_d)^T \) is one point in the parameter space not equal to the \( \theta^* \) in any element.

Then Meng and Rubin obtained \( r_{ij} \), the \( ij \) element of \( DM \) when the sequence \( r^{(t^*)}_{ij}, r^{(t^*+1)}_{ij} \) ... is stable for some \( t^* \).

Intuitively, the SEM method may have three limitations:

- Although we can calculate the complete-data information matrix in some simple examples, such as the Gaussian mixture model and the multinomial model, sometimes the complete-data information matrix is not feasible in practice.

- From Meng and Rubin (1991), SEM requires roughly \((d + 1)/2\) times as much computational time as the EM itself. However, the convergence rate of EM is slow, and the cost is high if it is in high dimensional parameter spaces.

- We can only obtain the observed information matrix from SEM. Although shown by Efron and Hinkley (1978) that the observed information matrix is better than the expected one, it only holds for one-parameter translation families and with conditioning on an ancillary statistic. Moreover in Cao and Spall (2012) and Cao (2013), it was shown that under reasonable conditions similar to standard MLE conditions, the inverse expected FIM outperforms the inverse observed FIM under a mean squared error criterion.
In the case where computation of the second derivative of \( Q(\theta | \theta^{(t)}) \) is not feasible, we cannot utilize the SEM algorithm. As a result, here we would like to compare the SEM algorithm with our SP-based algorithm when the complete data information matrix is available. In that the SEM is used to calculate the observed FIM, we are going to compare the two algorithms in the observed FIM calculation. That is, we would like to compare two algorithms; one approximates \( \mathbf{DM} \) by the SEM, and the other one applies the SPSA to approximate \( \mathbf{DM} \) with different perturbation vectors and calculates the average of them. Note that in the following analysis, we will ignore the cost of basic operations, as the iterations of the EM algorithm are usually the dominant expense.

The formula to calculate the estimation of the \( ij \)th element of \( \mathbf{DM} \) with \( k \)th perturbation vector from SPSA is:

\[
\frac{M_j(\theta^* + \Delta_k) - M_j(\theta^* - \Delta_k)}{2\Delta_{kl}}.
\] (4.1.1)

Using Taylor Expansion and taking expectation to (4.1) we obtain:

\[
E\left[ \sum_{s=1}^{d} \frac{(\mathbf{DM})_{sj}\Delta_{ks}}{\Delta_{kl}} \right] + O(c^2).
\] (4.1.2)

Suppose \( \Delta_{ks} \) follows Bernoulli \( \pm c \) distribution, and because the \( \Delta_{ks} \) are independent, \( E[\Delta_{ks}/\Delta_{kl}] = 1 \) when \( s = i \) and \( E[\Delta_{ks}/\Delta_{kl}] = 0 \) when \( s \neq i \). Therefore, (4.2) equals:

\[
(\mathbf{DM})_{ij} + O(c^2).
\]
In as much as the $c$ here can be arbitrarily small, we can ignore this term. Due to the independence of every element in $\Delta_k$, for $s \neq t$, $\text{cov}(\Delta_{ks}/\Delta_{kt}, \Delta_{kt}/\Delta_{ki}) = 0$. The variance of $\Delta_{ks}/\Delta_{ki}$ is 1 for $s \neq j$ and 0 for $s = j$. Therefore, the variance of (4.1) is $\sum_{s \neq i}[\mathbf{DM}]_{is}^2$.

By the Central Limit Theorem, when we calculate the average of the SPSA term, we can obtain:

$$\frac{\sum_{k=1}^{n} \left( \sum_{s=1}^{d} \frac{\mathbf{DM}_{is} \Delta_{ks}}{\Delta_{ki}} \right)}{n} \sim N \left( (\mathbf{DM})_{ij}, \frac{\sum_{s \neq i}[\mathbf{DM}]_{is}^2}{n} \right).$$

As $(\sum_{s \neq i}[\mathbf{DM}]_{is}(\theta^*)^2) \leq (d - 1)(M_1^1)^2$, where $M_1^1$ is $\max(|DM|)_{ij}$, for $1 \leq j \leq d$ and $j \neq i$. If we let $n = d - 1$, by the 68–95–99.7 rule, the probability of the event, which is the average we obtain that lies in $(\mathbf{DM})_{ij} \pm 2M_1^1$, is at least 95%. When $n = d - 1$, we need $2(d - 1)$ values of $M$ function, which requires $2(d - 1)$ iterations of the EM algorithm.

Next we are going to consider the accuracy of the SEM algorithm when we run $2d$ times EM algorithm. Unfortunately, it is almost impossible to measure it accurately because the cost and accuracy of the SEM depends on the number of iterations of the EM, the starting value of the EM, the starting point of the SEM, and the fraction of missing information. However, we can value it under some reasonable assumptions. In Meng and Rubin (1991), the choice of the starting point of the SEM is the tradeoff between accuracy and computational cost. Meng and Rubin recommended using the second iterate of EM algorithm. Here, we follow their suggestions and make the following assumptions:

- The $M_j$ is second differentiable, supposing that the $ii$th element of second derivative of $M_j$ at $\theta^*$ is $M_i'$. 


• The starting value of the EM algorithm makes $|\theta^{(3)}_i - \theta^*_i| \geq \frac{4M^1_i}{|M'_i|}$ for every $1 \leq i \leq d$ hold, where $\theta^{(3)}_i$ is the $i$th element of $\theta^{(3)}$. Under these assumptions, within the requirements of cost ($2d$ times EM algorithm), applying SEM algorithm, we can obtain:

$$r_{ij}^{(2)} = r_{ij} = \frac{M_j \left( \theta^* \pm \frac{4M^1_i}{|M'_i|} \delta_i \right) - M_j(\theta^*)}{\frac{4M^1_i}{|M'_i|}} \approx (DM)_{ij} \pm 2M^1_i.$$  

We have shown that the probability of the event, which is the approximation we obtain from the SPSA using $2(d - 1)$ times EM algorithm that lies in $(DM)_{ij}(\theta^*) \pm 2M^1_i$, is at least 95%. That is to say, under approximately equal cost ($2(d - 1)$ for the SPSA versus $2d$ for the SEM), we are 95% sure that the $DM$ provided by the SPSA is more accurate.

### 4.2 Comparison with SP-based method in Spall (2005)

The SP-based method in Spall (2005) is also a feasible algorithm to approximate the FIM in EM. In order to approximate the expected (or observed) FIM, while we take the gradient of $Q(\theta|\theta^*)$ in our algorithm using SPSA, the method in Spall (2005) approximate the gradient of the incomplete data log-likelihood function. (Note that enhanced versions of the basic method in Spall, 2005, are considered in Spall, 2008, and Das et al., 2010; we do not consider these enhancements here.)

In this part, we assume the function value ($Q$ in our algorithm versus $L_O$ in Spall (2005)) is the only cost in these algorithms. Furthermore, we suppose one $Q$ value and one $L_O$ share the same cost although the $L_O$ is likely more complicated than $Q$. Otherwise, it may not necessary to apply EM algorithm.
From (3.1), we obtain:

$$L_O(\theta|Y) = Q(\theta|\theta^*) - E[\log\{p_{X|Y}(x|y, \theta)\}|Y, \theta^*].$$

(4.2.1)

Let us define:

$$C(\theta|\theta^*) = E[\log\{p_{X|Y}(x|y, \theta)\}|Y, \theta^*].$$

According to the property of EM algorithm, we have:

$$[\frac{\partial Q(\theta|\theta^*)}{\partial \theta}]_{\theta=\theta^*} = [\frac{\partial L_O(\theta|Y)}{\partial \theta}]_{\theta=\theta^*} = [\frac{\partial C(\theta|\theta^*)}{\partial \theta}]_{\theta=\theta^*} = 0.$$  

(4.2.2)

Based on (4.2.1):

$$L_O(\theta^* + \Delta_k|Y) - L_O(\theta^* - \Delta_k|Y) = \frac{1}{2} \begin{pmatrix} \Delta_{k1}^{-1} \\ \Delta_{k2}^{-1} \\ \vdots \\ \Delta_{kd}^{-1} \end{pmatrix}$$

$$= \frac{Q(\theta^* + \Delta_k|\theta^*) - Q(\theta^* - \Delta_k|\theta^*)}{2} \begin{pmatrix} \Delta_{k1}^{-1} \\ \Delta_{k2}^{-1} \\ \vdots \\ \Delta_{kd}^{-1} \end{pmatrix}$$

$$- \frac{C(\theta^* + \Delta_k|\theta^*) - C(\theta^* - \Delta_k|\theta^*)}{2} \begin{pmatrix} \Delta_{k1}^{-1} \\ \Delta_{k2}^{-1} \\ \vdots \\ \Delta_{kd}^{-1} \end{pmatrix}. \quad (4.2.3)$$

With the assumption that $L_O, Q$, and $C$ are several times differentiable at $\theta^*$, take expectation to (4.2.3) and combine it with (4.2.2), the bias of the expectation of $m$th component of the gradient approximation using the general SP-based method is:
\[ b_{km} = \frac{1}{12} E \left[ \frac{L'''_o(\theta^*(+)) - L'''_o(\theta^*(-))}{\Delta_{km}} \Delta_k \otimes \Delta_k \otimes \Delta_k \right] \]

\[
= \frac{1}{12} E \left[ \left( \frac{\partial^3 Q(\theta^*(+)\mid \theta^* \pm \Delta_k)}{\partial \theta^3} - \frac{\partial^3 C(\theta^*(+)\mid \theta^*)}{\partial \theta^3} \right) + \left( \frac{\partial^3 Q(\theta^*(-)\mid \theta^* \pm \Delta_k)}{\partial \theta^3} - \frac{\partial^3 C(\theta^*(-)\mid \theta^*)}{\partial \theta^3} \right) \right] \Delta_{km} \Delta_k \otimes \Delta_k \otimes \Delta_k \right],
\]

(4.2.4)

where \( \bar{\theta}^{*(\pm)} \) denotes pointes between \( \theta^* \) and \( \theta^* \pm \Delta_k \) and \( \otimes \) is the Kronecker product.

The bias from our algorithm is:

\[
b_{km}' = \frac{1}{12} E \left[ \frac{\partial^3 Q(\bar{\theta}^*(+)\mid \theta^*)}{\partial \theta^3} + \frac{\partial^3 Q(\bar{\theta}^*(-)\mid \theta^*)}{\partial \theta^3} \right] \Delta_{km} \Delta_k \otimes \Delta_k \otimes \Delta_k \right].
\]

(4.2.5)

Given that the c can be small, (4.2.4) and (4.2.5) imply that both methods provide almost unbiased estimator of the gradient, which is 0 in this setting, and the difference between their biases are also ignorable.

5. Numerical Examples

In this section, we introduce two numerical examples, one for the mixture of two Gaussian distributions from Louis (1982) and the other for parameters in a state-space model. In the first example, to calculate the expected FIM and the observed FIM, we use our algorithm; in addition, we also utilize Louis’ formula on the observed FIM. In the state-space model, we apply our algorithm to obtain the expected FIM.
5.1 Mixture of Gaussian Distributions

Let \( Y_1, Y_2 \ldots, Y_n \) be the i.i.d. observed data from the Gaussian mixture model:

\[
p_Y(y|\pi, \mu_1, \mu_2) = (1 - \pi)\phi(y - \mu_1) + \pi\phi(y - \mu_2)
\]

with unknown parameter \( \theta = \{\pi, \mu_1, \mu_2\} \) satisfying:

\[0 \leq \pi \leq 1, \quad -\infty < \mu_1, \mu_2 < +\infty.\]

\( \phi(y) \) is the density function of standard normal distribution. We would like to calculate the expected and observed FIM at \( \theta^* \).

Let \( Z_i = (X_i, Y_i) \) \( i = 1, \ldots, n \) be the complete data where \( X_1, \ldots, X_n \) represents the missing data satisfying:

\[
X_i = \begin{cases} 
0 & \text{when } Y_i \text{ is from } N(\mu_1, 1) \\
1 & \text{when } Y_i \text{ is from } N(\mu_2, 1) 
\end{cases}
\]

Therefore, the density function of the complete data is:

\[
p_{X,Y}(x,y|\pi, \mu_1, \mu_2) = (1 - x)(1 - \pi)\phi(y - \mu_1) + x\pi\phi(y - \mu_2).
\]

It is reasonable to apply the EM algorithm to this problem with the observed data \( Y_i \), the missing data \( X_i \), and the complete data \( Z_i \). The log-likelihood function of the observed data, \( L_O(\theta|Y) \), is:

\[
\log p_Y(y_i|\pi, \mu_1, \mu_2) = \log((1 - \pi)\phi(y_i - \mu_1) + \pi\phi(y_i - \mu_2)).
\]

The log-likelihood function of the complete data, \( L(\theta|Z) = L(\theta|X,Y) \), is:

\[
(1 - x_i)\log(1 - \pi) + x_i\log\pi + (1 - x_i)\log\phi(y_i - \mu_1) + x_i\log\phi(y_i - \mu_2).
\]
Let the conditional expectation of $X_i$ given $Y_i$ and $\theta^{(t)} = (\pi^{(t)}, \mu_1^{(t)}, \mu_2^{(t)})^T$ be $\alpha_i^{(t)}$:

$$
\alpha_i^{(t)} = E[X_i|Y_i, \theta^{(t)}] = \frac{\pi^{(t)} \phi(Y_i - \mu_2^{(t)})}{(1 - \pi^{(t)}) \phi(Y_i - \mu_1^{(t)}) + \pi^{(t)} \phi(Y_i - \mu_2^{(t)})}.
$$

(5.1.1)

From Louis (1982), the EM operator $M(\theta^{(t)})$ is:

$$
M(\theta^{(t)}) = \theta^{(t+1)} = \begin{bmatrix}
\frac{\sum_{i=1}^n \alpha_i^{(t)}}{n} \\
\frac{\sum_{i=1}^n (1 - \alpha_i^{(t)}) Y_i}{(n - \sum_{i=1}^n \alpha_i^{(t)})} \\
\sum_{i=1}^n \alpha_i^{(t)} Y_i / \sum_{i=1}^n \alpha_i^{(t)}
\end{bmatrix}
$$

Set the true $\theta = (\pi, \mu_1, \mu_2)^T = (2/3, 3, 0)^T$ and generate 750 data from the following density function:

$$
p_Y(y|\frac{2}{3}, 0, 3) = \left(1 - \frac{2}{3}\right) \phi(y - 3) + \frac{2}{3} \phi(y).
$$

Set the starting value $\theta^{(0)} = (0.1, 1.0, 1.0)^T$. We obtain $\theta^* = (0.6625, 3.0133, -0.0148)^T$ using EM algorithm after 54 iterations.

To apply our algorithm to compute the expected FIM, firstly, we should know the form of $S(\theta|Y_i)$ for different $\theta' = (\theta_1', \theta_2', \theta_3')^T$:

$$
S(\theta'|Y_i) = \left[ \frac{\partial Q(\theta|\theta')}{\partial \theta} \right]_{\theta=\theta'} = \begin{bmatrix}
\frac{E[X_i|Y_i, \theta']}{\theta_1'} - (1 - E[X_i|Y_i, \theta']) / (1 - \theta_1') \\
(1 - E[X_i|Y_i, \theta']) (Y_i - \theta_2') / E[X_i|Y_i, \theta'] (Y_i - \theta_2') \\
E[X_i|Y_i, \theta'] (Y_i - \theta_3') / \sum_{i=1}^n \alpha_i^{(t)} Y_i / \sum_{i=1}^n \alpha_i^{(t)}
\end{bmatrix}
$$

(5.1.2)

We can calculate $S(\theta'|Y_i)$ for different $\theta'$ and $Y_i$ by substituting (5.1.1) into (5.1.2). Then we generate pseudodata $Y_i$ from the mixture Gaussian distribution based on $\theta^*$. 

20
The expected FIM obtained from our algorithm is:

\[
\begin{pmatrix}
2680.7 & -180.9 & -266.8 \\
-190.9 & 171.4 & -65.5 \\
-266.8 & -65.6 & 396.0
\end{pmatrix}
\]

To compute the observed FIM by Louis’ method, it is required to calculate the complete-data information matrix, which is:

\[
\begin{pmatrix}
X_i + \frac{1-X_i}{(1-\pi)^2} & 0 & 0 \\
0 & 1-X_i & 0 \\
0 & 0 & X_i
\end{pmatrix} = \begin{pmatrix} 3354.1 & 0 & 0 \\
0 & 253.2 & 0 \\
0 & 0 & 496.8 \end{pmatrix},
\]

the conditional expectation of the square of gradient of complete data log-likelihood function and (5.1.2) at \( \theta^* \). After calculation, the observed FIM from Louis’ method, which is the same as the true observed FIM is:

\[
I_{Louis}(\theta^*) = \begin{pmatrix} 2708.3 & -200.1 & -237.1 \\
-200.1 & 176.6 & -59.0 \\
-237.1 & -59.0 & 395.2 \end{pmatrix}.
\]

Based on our algorithm, the estimator of the observed FIM at \( \theta^* \) is:

\[
I_{SPSA}(\theta^*) = \begin{pmatrix} 2709.5 & -205.0 & -235.8 \\
-205.0 & 178.7 & -61.9 \\
-235.8 & -61.9 & 396.5 \end{pmatrix}.
\]

We can measure the accuracy of our method by calculating \( \frac{||I_{Louis}(\theta^*) - I_{SPSA}(\theta^*)||}{||I_{Louis}(\theta^*)||} \), where \( || \cdot || \) is the spectral norm.

\[
\frac{||I_{Louis}(\theta^*) - I_{SPSA}(\theta^*)||}{||I_{Louis}(\theta^*)||} = 0.0029.
\]
\( I_{\text{Louis}}(\theta^*) \) is the true value of the observed FIM evaluated at \( \theta^* \). The result shows that the observed FIM obtained from our SPSA-based method is a good approximation to the true observed FIM evaluated at \( \theta^* \).

5.2 State-space model

A state-space model is a model that uses state variables to represent a physical system. It is defined by the two equations, state equation (5.2.1) and measurement equation (5.2.2).

\[
x_t = A x_{t-1} + w_t, \tag{5.2.1}
\]
\[
y_t = D x_t + v_t, \tag{5.2.2}
\]

for \( t = 1, \ldots, n \). In (5.2.1) and (5.2.2), while \( x_t \) is an unobserved \( p \)-dimensional state process, we can observe \( y_t \), which is a \( q \)-dimensional process. \( A \) is a \( p \times p \) transition matrix and \( D \) is a \( q \times q \) measurement matrix. Noises \( w_t \) and \( v_t \) are typically assumed independent, Gaussian distributed, with means \( 0 \) and covariance matrices \( \text{cov}(w_t) = Q \) and \( \text{cov}(v_t) = R \). We also assume that the mean of \( x_0 \), \( \mu \), and the covariance matrix of \( x_0 \), \( \Sigma \), are known and \( x_0, w_t \) and \( v_t \) are mutually independent. Note that the analytical form of the FIM for arbitrary parameters in a state-space model is difficult to obtain, but the analytical form is available in special cases, as illustrated in Spall and Garner (1990) and Cao (2013, Sect. 3.3).

In this model, \( x_0, x_1, \ldots, x_n, y_1, \ldots, y_n \) is the complete data. However, we only have the observed data, which is \( x_0, y_1, \ldots, y_n \).

From Shumway and Stoffer (1982), neglecting the constant term, the log-likelihood function of the complete data in the state-space model above is:
\[ L(\mu, \Sigma, A, Q, R|x_0, x_1 \ldots, x_n, y_1, \ldots y_n) \]
\[ = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_0 - \mu)^T \Sigma^{-1}(x_0 - \mu) \]
\[ - \frac{n}{2}\log|Q| - \frac{1}{2}\sum_{t=1}^{n} (x_t - Ax_{t-1})^T Q^{-1}(x_t - Ax_{t-1}) - \frac{n}{2}\log|R| \]
\[ - \frac{1}{2}\sum_{t=1}^{n} (y_t - Dx_t)^T R^{-1}(y_t - Dx_t), \]

where \(| * |\) means the determinant of “*”.

In this section, we would like to consider the same setting in Cao (2013). Let \( p = 3 \) and \( q = 1 \). Suppose further that \( \mu, \Sigma, A, D, \) and \( R \) are known and

\[ \mu = (0 \ 0 \ 0)^T, \]
\[ \Sigma = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \]
\[ A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.8 & 0.8 & -0.8 \end{pmatrix}, \]
\[ D = (1 \ 0 \ 0), \]

and \( R = 1 \).

Let \( Q \) be a 3 \times 3 diagonal matrix with diagonal elements \( Q_1, Q_2, \) and \( Q_3 \). We are going to estimate \( \theta = (Q_1, Q_2, Q_3)^T \) with true \( \theta = (1, 1, 1)^T \). Applying the procedure in Shumway and
Stoffer (1982) and setting \( n = 100 \), we obtain \( \theta^* = (0.9372,0.9863,1.0536)^T \) after 75 EM iterations.

After we know the \( \theta^* \), we can use our algorithm to approximate the expected Fisher information matrix at \( \theta^* \). In this study, given that the conditional expectation of complete data log-likelihood function is complicated, we apply (3.7) to obtain the gradient of the conditional expectation and then (3.6) to calculate the Hessian matrix.

When \( N = 10000 \),

\[
I(\theta^*) = \begin{pmatrix} 5.0430 & 2.3190 & 3.1588 \\ 2.3190 & 3.0455 & 1.5283 \\ 3.1588 & 1.5283 & 3.9970 \end{pmatrix},
\]

and \( I_n(\theta^*)^{-1} = \left( \frac{I(\theta^*)}{n} \right)^{-1} = \begin{pmatrix} 48.9167 & -22.0781 & -30.3024 \\ -22.0781 & 50.3417 & -1.9311 \\ -30.3024 & -1.9311 & 49.6131 \end{pmatrix} \).

When \( N = 20000 \),

\[
I(\theta^*) = \begin{pmatrix} 6.5415 & 3.3503 & 4.1475 \\ 3.3503 & 3.8149 & 2.2182 \\ 4.1475 & 2.2182 & 4.5081 \end{pmatrix},
\]

and \( I_n(\theta^*)^{-1} = \left( \frac{I(\theta^*)}{n} \right)^{-1} = \begin{pmatrix} 47.7088 & -22.9402 & -32.6053 \\ -22.9402 & 47.7483 & -2.3891 \\ -32.6053 & -2.3891 & 53.3555 \end{pmatrix} \).

Because there is no closed form for \( I_n(\theta^*)^{-1} \), we apply the approximation in Cao (2013) as the true value for comparison. The result in Cao (2013) is:

\[
I_n(\theta^*)^{-1}_{\text{true}} = \begin{pmatrix} 51.8934 & -24.4471 & -33.7404 \\ -24.4471 & 59.4544 & -3.36401 \\ -33.7404 & -3.36401 & 63.0565 \end{pmatrix}.
\]
When $N = 10000$,

$$\frac{\left| I_n(\theta^*)^{-1} - I_n(\theta^*)^{-1}_{\text{true}} \right|}{\left| I_n(\theta^*)^{-1}_{\text{true}} \right|} = 0.1499.$$ 

When $N = 20000$,

$$\frac{\left| I_n(\theta^*)^{-1} - I_n(\theta^*)^{-1}_{\text{true}} \right|}{\left| I_n(\theta^*)^{-1}_{\text{true}} \right|} = 0.1259.$$ 

The result shows that the accuracy increases with $N$ increases.

6. Conclusions

Calculating the expected/observed Fisher information matrix (FIM) is often not easy in practice. It is even more difficult in the EM algorithm in that the EM algorithm obtained the result from a simple complete data log-likelihood function, not the incomplete data log-likelihood function associated with observed data. This article introduces a Monte Carlo-based method to approximate the expected FIM in the EM algorithm. We generate the pseudodata at $\theta^*$ by the bootstrap method and then apply the numerical differentiation of the gradient of the function obtained from the E step for different pseudodata, acquiring the negative average. Compared to existing methods, for example the approach in Louis (1982) and the SEM in Meng and Rubin (1991), our method does not require the complete-data information matrix, which is sometimes complicated. We only require the values of $S(\theta|Y)$ for different $\theta$ or the values of
$Q(\theta | \theta')$ for different $\theta$. Theoretical analysis as well as numerical analyses are conducted to show the superiority of our method.

References

Abt, M. and Welch, W. J. (1998), “Fisher Information and Maximum-Likelihood Estimation of Covariance Parameters in Gaussian Stochastic Processes,” Canadian Journal of Statistics, 26(1), pp. 127–137.

Berisha, V. and Hero, A. O. (2015), "Empirical Non-Parametric Estimation of the Fisher Information," IEEE Signal Processing Letters, 22(7), 988–992.

Bhatnagar, S., Prasad, H. L., and Prashanth, L. A. (2013), Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods, Springer.

Cao, X. "Relative Performance of Expected and Observed Fisher Information in Covariance Estimation for Maximum Likelihood Estimates," Johns Hopkins University Ph.D. thesis, May 2013.

Cao, X. and Spall, J. C. (2012), “Relative Performance of Expected and Observed Fisher Information in Covariance Estimation for Maximum Likelihood Estimates,” Proceedings of the American Control Conference, Montréal, Canada, 27–29 June 2012, pp. 1871–1876.

Das, S., Spall, J. C., and Ghanem, R. (2010), “Efficient Monte Carlo Computation of Fisher Information Matrix Using Prior Information,” Computational Statistics and Data Analysis, 54(2), pp. 272–289.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society. Series B (Methodological), 39, 1–38.

Gupta, M. R. and Chen, Y. (2011), “Theory and Use of the EM Algorithm,” Foundations and Trends in Signal Processing, 4(3), pp 223–296.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," Journal of the Royal Statistical Society. Series B (Methodological), 44, 226–233.

McLachlan G. J. and Krishnan T. (1997), “The EM Algorithm and Extensions,” John Wiley & Sons.

Meng, X. L. and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," Journal of the American Statistical Association, 86, 899-909.

Oakes, D. (1999), "Direct Calculation of the Information Matrix via the EM," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(2), 479–482.

Pronzato, L. (2008), “Optimal Experimental Design and Some Related Control Problems,” Automatica, vol. 44(2), pp. 303–325.

Shumway, R. H. and Stoffer D. S. (1982), “An Approach to Time Series Smoothing And Forecasting Using the EM Algorithm,” Journal of Time Series Analysis, 3, 253–264.
Spall, J. C. (1992), "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Transactions on Automatic Control*, 37(3), 332–341.

Spall, J. C. (1998), “An Overview of the Simultaneous Perturbation Method for Efficient Optimization,” *Johns Hopkins APL Technical Digest*, 19(4), pp. 482–492.

Spall, J. C. (2000), "Adaptive Stochastic Approximation by the Simultaneous Perturbation Method," *IEEE Transactions on Automatic Control*, 45(10), 1839–1853.

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley, Hoboken, NJ

Spall, J. C. (2005), "Monte Carlo Computation of the Fisher Information Matrix in Nonstandard Settings," *Journal of Computational and Graphical Statistics*, 14, 889–909.

Spall, J. C. (2008), “Improved Methods for Monte Carlo Estimation of the Fisher Information Matrix,” *Proceedings of the American Control Conference*, Seattle, WA, 11–13 June 2008, pp. 2395–2400.

Spall, J. C. (2010), “Factorial Design for Choosing Input Values in Experimentation: Generating Informative Data for System Identification,” *IEEE Control Systems Magazine*, vol. 30(5), pp. 38–53.

Spall, J. C. (2014), “Identification for Systems with Binary Subsystems,” *IEEE Transactions on Automatic Control*, 59(1), pp. 3–17.

Spall, J. C. and Garner, J. P. (1990), “Parameter Identification for State-Space Models with Nuisance Parameters,” *IEEE Transactions on Aerospace and Electronic Systems*, 26(6), pp. 992–998.

Vaida, F. (2005), “Parameter Convergence for EM and MM Algorithms,” *Statistica Sinica*, vol. 15, pp. 831–840.

Wu, C. F. J. (1983), “On the Convergence Properties of the EM Algorithm,” *Annals of Statistics*, vol. 11, pp. 95–103.

Watanabe, M. and Yamaguchi, K. (eds.) (2004), *The EM Algorithm and Related Statistical Models*, CRC Press.