

# The Fisher Information Matrix: Performance Measure and Monte Carlo-Based Computation<sup>1</sup>

James C. Spall  
The Johns Hopkins University  
Applied Physics Laboratory  
11100 Johns Hopkins Road  
Laurel, Maryland 20723-6099  
Ph.: 240-228-4960  
Fax: 240-228-6661  
E-mail: james.spall@jhuapl.edu

## ABSTRACT

The Fisher information matrix summarizes the amount of information in the data relative to the quantities of interest. There are many applications of the information matrix in modeling, systems analysis, and estimation, including confidence region calculation, input design, prediction bounds, and “noninformative” priors for Bayesian analysis. This paper reviews some basic principles associated with the information matrix, presents a resampling-based method for computing the information matrix together with some new theory related to efficient implementation, and presents some numerical results. The resampling-based method relies on an efficient technique for estimating the Hessian matrix, introduced as part of the adaptive (“second-order”) form of the simultaneous perturbation stochastic approximation (SPSA) optimization algorithm.

**KEY WORDS:** *Monte Carlo simulation; Cramér-Rao bound; simultaneous perturbation; antithetic random numbers.*

## 1. INTRODUCTION

The Fisher information matrix plays a central role in the practice and theory of identification and estimation. This matrix provides a summary of the amount of information in the data relative to the quantities of interest. Some of the specific applications of the information matrix include confidence region calculation for parameter estimates, the determination of inputs in experimental design, providing a bound on the best possible performance in an adaptive system based on unbiased parameter estimates (such as a control system), producing uncertainty bounds on predictions (such as with a neural network), and determining noninformative prior distributions (Jeffreys’ prior) for Bayesian analysis. Unfortunately, the analytical calculation of the information matrix is often difficult or impossible. This is especially the

case with nonlinear models such as neural networks. This paper describes a Monte Carlo resampling-based method for computing the information matrix. This method applies in problems of arbitrary difficulty and is relatively easy to implement.

Section 2 provides some formal background on the information matrix and summarizes two key properties that closely connect the information matrix to the covariance matrix of general parameter estimates. This connection provides the prime rationale for applications of the information matrix in the areas of uncertainty regions for parameter estimation, experimental design, and predictive inference. Section 3 describes the Monte Carlo resampling-based approach. Section 4 presents some theory in support of the method, including a result that provides the basis for an optimal implementation of the Monte Carlo method. Section 5 discusses an implementation based on antithetic random numbers, which can sometimes result in variance reduction. Section 6 describes some numerical results and Section 7 gives some concluding remarks.

## 2. FISHER INFORMATION MATRIX: DEFINITION AND KEY PROPERTIES

Consider a collection of  $n$  random vectors  $\mathbf{z}^{(n)} \equiv [z_1, z_2, \dots, z_n]^T$ . Let us assume that the *general form* for the joint probability density or probability mass (or hybrid density/mass) function for the random data matrix  $\mathbf{z}^{(n)}$  is known, but that this function depends on an unknown vector  $\boldsymbol{\theta}$ . Let the probability density/mass function for  $\mathbf{z}^{(n)}$  be  $p_{\mathbf{z}}(\boldsymbol{\zeta}|\boldsymbol{\theta})$

---

<sup>1</sup>**Acknowledgments:** This work was partially supported by DARPA contract MDA972-96-D-0002 in support of the Advanced Simulation Technology Thrust Area, U.S. Navy Contract N00024-98-D-8124, and the JHU/APL IRAD Program. A more complete version of this paper is available upon request.

where  $\zeta$  (“zeta”) is a dummy matrix representing the possible outcomes for the elements in  $\mathbf{Z}^{(n)}$  (in  $p_{\mathbf{Z}}(\zeta|\boldsymbol{\theta})$ , the index  $n$  on  $\mathbf{Z}^{(n)}$  is being suppressed for notational convenience). The corresponding likelihood function, say  $\ell(\boldsymbol{\theta}|\zeta)$ , satisfies

$$\ell(\boldsymbol{\theta}|\zeta) = p_{\mathbf{Z}}(\zeta|\boldsymbol{\theta}). \quad (2.1)$$

With the definition of the likelihood function in (2.1), we are now in a position to present the Fisher information matrix. The expectations below are with respect to the data set  $\mathbf{Z}^{(n)}$ .

The  $p \times p$  information matrix  $\mathbf{F}_n(\boldsymbol{\theta})$  for a differentiable log-likelihood function is given by

$$\mathbf{F}_n(\boldsymbol{\theta}) \equiv E \left( \frac{\partial \log \ell}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \log \ell}{\partial \boldsymbol{\theta}^T} \middle| \boldsymbol{\theta} \right). \quad (2.2)$$

In the case where the underlying data  $\{z_1, z_2, \dots, z_n\}$  are independent (and even in many cases where the data may be dependent), the magnitude of  $\mathbf{F}_n(\boldsymbol{\theta})$  will grow at a rate proportional to  $n$  since  $\log \ell$  will represent a sum of  $n$  random terms. The bounded quantity  $\mathbf{F}(\boldsymbol{\theta})/n$  is employed as an average information matrix over all measurements. Note also that when the data depend on some inputs  $\mathbf{x}_i$ , then  $\mathbf{F}_n(\boldsymbol{\theta})$  also depends on these inputs, i.e.,  $\mathbf{F}_n(\boldsymbol{\theta}) = \mathbf{F}_n(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ . For notational convenience—and since many applications depend on cases (such as i.i.d. data) where there are no varying inputs—we suppress this dependence and write  $\mathbf{F}_n(\boldsymbol{\theta})$  for the information matrix. In optimal input design, however, this dependence on the  $\mathbf{x}_i$  is critical (e.g., Atkinson and Donev, 1992, Chap. 10; Ljung, 1999, Chap. 13; Spall, 2003, Chap. 17).

Except for relatively simple problems, however, the form in (2.2) is generally not useful in the practical calculation of the information matrix. Computing the expectation of a product of multivariate nonlinear functions is usually a hopeless task. A well-known equivalent form follows by assuming that  $\log \ell$  is twice differentiable in  $\boldsymbol{\theta}$ . That is, the Hessian matrix

$$\mathbf{H}(\boldsymbol{\theta}|\zeta) \equiv \frac{\partial^2 \log^2 \ell(\boldsymbol{\theta}|\zeta)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

is assumed to exist. Further, assume that the likelihood function is “regular” in the sense that standard conditions such as in Wilks (1962, pp. 408–411; pp. 418–419) or Bickel and Doksum (1977, pp. 126–127) hold. One of these conditions is that the set  $\{\zeta: \ell(\boldsymbol{\theta}|\zeta) > 0\}$  does not depend on  $\boldsymbol{\theta}$ . A fundamental implication of the regularity for the likelihood is that the necessary interchanges of differentiation and

integration are valid. Then, the information matrix is related to the Hessian matrix of  $\log \ell$  through:

$$\mathbf{F}_n(\boldsymbol{\theta}) = -E \left[ \mathbf{H}(\boldsymbol{\theta}|\mathbf{Z}^{(n)}) \middle| \boldsymbol{\theta} \right]. \quad (2.3)$$

The form in (2.3) is usually more amenable to calculation than the product-based form in (2.2).

Note that in some applications, the *observed* information matrix at a particular data set  $\mathbf{Z}^{(n)}$  (i.e.,  $-\mathbf{H}(\boldsymbol{\theta}|\mathbf{Z}^{(n)})$ ) may be easier to compute and/or preferred from an inference point of view relative to the actual information matrix  $\mathbf{F}_n(\boldsymbol{\theta})$  in (2.3) (e.g., Efron and Hinckley, 1978). Although the method in this paper is described for the determination of  $\mathbf{F}_n(\boldsymbol{\theta})$ , the efficient Hessian estimation described in Section 3 may also be used directly for the determination of  $\mathbf{H}(\boldsymbol{\theta}|\mathbf{Z}^{(n)})$  when it is not easy to calculate the Hessian directly.

The above discussion focused on the definition of the information matrix and the equivalence of two representations for the matrix (the gradient-product form and the Hessian-based form). We now review two of the most important analytical properties of the matrix. Let  $\boldsymbol{\theta}^*$  denote the unknown “true” value of  $\boldsymbol{\theta}$ . The primary rationale for  $\mathbf{F}_n(\boldsymbol{\theta})$  as a measure of information about  $\boldsymbol{\theta}$  within the data  $\mathbf{Z}^{(n)}$  comes from its connection to the covariance matrix for the estimate of  $\boldsymbol{\theta}$  constructed from  $\mathbf{Z}^{(n)}$ . The first of the key properties makes this connection via an asymptotic normality result. In particular, for some common forms of estimates  $\hat{\boldsymbol{\theta}}_n$  (e.g., maximum likelihood and Bayesian maximum a posteriori), it is known that, under modest conditions,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{\text{dist}} N(\mathbf{0}, \mathbf{F}_*^{-1}) \quad (2.4)$$

where  $\xrightarrow{\text{dist}}$  denotes convergence in distribution and

$$\mathbf{F}_* \equiv \lim_{n \rightarrow \infty} \frac{\mathbf{F}_n(\boldsymbol{\theta}^*)}{n}$$

provided that the indicated limit exists and is invertible (e.g., Hoadley, 1971; Rao, 1973, pp. 415–417). Hence, in practice, for  $n$  reasonably large,  $\mathbf{F}_n(\boldsymbol{\theta})^{-1}$  can serve as an approximate covariance matrix of the estimate  $\hat{\boldsymbol{\theta}}_n$  when  $\boldsymbol{\theta}$  is chosen close to the unknown  $\boldsymbol{\theta}^*$  (since  $\hat{\boldsymbol{\theta}}_n$  is convergent to  $\boldsymbol{\theta}^*$  in some stochastic sense under the conditions in which (2.4) holds,  $\boldsymbol{\theta}$  is usually chosen to be  $\hat{\boldsymbol{\theta}}_n$  for the evaluation of  $\mathbf{F}_n(\boldsymbol{\theta})$ ).

Relationship (2.4) also holds for optimal implementations of some recursive algorithms where the data  $z_i$  are processed recursively instead of in a batch mode as is typical in

maximum likelihood. This includes optimal versions of gradient-based stochastic approximation algorithms (e.g., Kushner and Yang, 1995; Kushner and Yin, 1997, pp. 332–333; or Spall, 2003, pp. 356–357), which includes popular algorithms such as least mean-squares (LMS) and neural network backpropagation as special cases.

The second key property of the information matrix applies in finite samples. If  $\hat{\boldsymbol{\theta}}_n$  is *any unbiased* estimator for  $\boldsymbol{\theta}$  (not just one for which (2.4) holds),

$$\text{cov}(\hat{\boldsymbol{\theta}}_n) \geq \mathbf{F}_n(\boldsymbol{\theta}^*)^{-1} \quad \forall n \quad (2.5)$$

(i.e.,  $\text{cov}(\hat{\boldsymbol{\theta}}_n) - \mathbf{F}_n(\boldsymbol{\theta}^*)^{-1}$  is positive semidefinite). There is also an expression analogous to (2.5) for biased estimators, but it is not especially useful in practice because it requires knowledge of the gradient of the bias with respect to  $\boldsymbol{\theta}$  (Rao, 1973, pp. 323–327; Bickel and Doksum, 1977, pp. 127–128). Expression (2.5) is generally referred to as the Cramér-Rao inequality.

Expressions (2.4) and (2.5), taken together, point to the close connection between the inverse Fisher information matrix and the covariance matrix of the estimator. A larger  $\mathbf{F}_n(\boldsymbol{\theta})$  (in the matrix sense) is associated with a smaller covariance matrix (i.e., more information) while a smaller  $\mathbf{F}_n(\boldsymbol{\theta})$  is associated with a larger covariance matrix (i.e., less information). While (2.4) is an asymptotic result, (2.5) applies for all sample sizes subject to the unbiasedness requirement.

### 3. RESAMPLING-BASED CALCULATION OF THE INFORMATION MATRIX

The calculation of  $\mathbf{F}_n(\boldsymbol{\theta})$  is often difficult or impossible in practical problems. Obtaining the required first or second derivatives of the log-likelihood function may be a formidable task in some applications, and computing the required expectation of the generally nonlinear multivariate function is often impossible in problems of practical interest. For example, in the context of dynamic models, Šimandl et al. (2001) illustrate the difficulty in nonlinear state estimation problems and Levy (1995) shows how the information matrix may be very complex in even relatively benign parameter estimation problems (i.e., for the estimation of parameters in a *linear* state-space model, the information matrix contains 35 distinct sub-blocks and fills up a full page).

To address this difficulty, the subsection outlines a computer resampling approach to estimating  $\mathbf{F}_n(\boldsymbol{\theta})$ . This approach is useful when analytical methods for computing  $\mathbf{F}_n(\boldsymbol{\theta})$  are infeasible. The approach makes use of an efficient method for Hessian estimation.

The basis for the technique below is to use computational horsepower in lieu of traditional detailed theoretical analysis to determine  $\mathbf{F}_n(\boldsymbol{\theta})$ . Two other notable Monte Carlo techniques are the bootstrap method for determining statistical

distributions of estimates (e.g., Efron and Tibshirani, 1986; Lunneborg, 2000) and the Markov chain Monte Carlo method for producing pseudorandom numbers and related quantities (e.g., Gelfand and Smith, 1990). Part of the appeal of the Monte Carlo method here for estimating  $\mathbf{F}_n(\boldsymbol{\theta})$  is that it can be implemented with only evaluations of the log-likelihood (typically much easier to obtain than the customary gradient or second derivative information). Alternatively, if the gradient of the log-likelihood is available, that information can be used to enhance performance.

The essence of the method is to produce a large number of efficient “almost unbiased” estimates of the Hessian matrix of  $\log \ell(\cdot)$  and then average the negative of these estimates to obtain an approximation to  $\mathbf{F}_n(\boldsymbol{\theta})$ . This approach is directly motivated by the definition of  $\mathbf{F}_n(\boldsymbol{\theta})$  as the mean value of the negative Hessian matrix (eqn. (2.3)). To produce these estimates, we generate *pseudodata vectors* in a Monte Carlo manner analogous to the bootstrap method mentioned above. The pseudodata are generated according to a bootstrap resampling scheme treating the chosen  $\boldsymbol{\theta}$  as “truth.” The pseudodata are generated according to the probability model (2.1). So, for example, if it is assumed that the real data  $\mathbf{Z}_n = [z_1^T, z_2^T, \dots, z_n^T]^T$  are jointly normally distributed,  $N(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ , then the pseudodata are generated by Monte Carlo according to a normal distribution based on a mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$  evaluated at the chosen  $\boldsymbol{\theta}$ . Let the  $i$ th pseudodata vector be  $\mathbf{Z}_{\text{pseudo}}(i)$ ; the use of  $\mathbf{Z}_{\text{pseudo}}$  without the argument is a generic reference to a pseudodata vector. This data vector represents a sample of size  $n$  (analogous to the real data  $\mathbf{Z}_n$ ) from the assumed distribution for the set of data based on the unknown parameters taking on the chosen value of  $\boldsymbol{\theta}$ .

Given the aim to avoid the complex calculations usually needed to obtain second derivative information, the critical part of this conceptually simple scheme is the efficient Hessian estimation. Spall (2000) introduced an efficient scheme for estimating Hessian matrices in the context of optimization. While there is no optimization here per se, we use the same formula for Hessian estimation. This formula is based on the simultaneous perturbation principle (Spall, 1992).

The approach below can work with either  $\log \ell(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}})$  values (alone) or with the gradient  $\mathbf{g}(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}}) \equiv \partial \log \ell(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}}) / \partial \boldsymbol{\theta}$  if that is available. The former usually corresponds to cases where the likelihood function and associated nonlinear process are so complex that no gradients are available. To highlight the fundamental commonality of approach, let  $\mathbf{G}(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}})$  represent either a gradient *approximation* (based on  $\log \ell(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}})$  values) or

the exact gradient  $\mathbf{g}(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}})$ . Because of its efficiency, the simultaneous perturbation gradient approximation is recommended in the case where only  $\log \ell(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}})$  values are available (see Spall, 2000).

We now present the Hessian estimate. Let  $\hat{\mathbf{H}}_k$  denote the  $k$ th estimate of the Hessian  $\mathbf{H}(\cdot)$  in the Monte Carlo scheme. The formula for estimating the Hessian is:

$$\hat{\mathbf{H}}_k = \frac{1}{2} \left\{ \frac{\delta \mathbf{G}_k}{2} [\Delta_{k1}^{-1}, \dots, \Delta_{kp}^{-1}] + \left( \frac{\delta \mathbf{G}_k}{2} [\Delta_{k1}^{-1}, \dots, \Delta_{kp}^{-1}] \right)^T \right\}, \quad (3.1)$$

where  $\delta \mathbf{G}_k \equiv \mathbf{G}(\boldsymbol{\theta} + \boldsymbol{\Delta}_k | \mathbf{Z}_{\text{pseudo}}) - \mathbf{G}(\boldsymbol{\theta} - \boldsymbol{\Delta}_k | \mathbf{Z}_{\text{pseudo}})$  and the perturbation vector  $\boldsymbol{\Delta}_k \equiv [\Delta_{k1}, \Delta_{k2}, \dots, \Delta_{kp}]^T$  is a mean-zero random vector such that the  $\{\Delta_{kj}\}$  are ‘‘small’’ symmetrically distributed random variables that are uniformly bounded and satisfy  $E(|1/\Delta_{kj}|) < \infty$  uniformly in  $k, j$ . This latter condition *excludes* such commonly used Monte Carlo distributions as uniform and Gaussian. Assume that  $|\Delta_{kj}| \leq c$  for some small  $c > 0$ . In most implementations, the  $\{\Delta_{kj}\}$  are i.i.d. across  $k$  and  $j$ . In implementations involving antithetic random numbers (see Section 5),  $\boldsymbol{\Delta}_k$  and  $\boldsymbol{\Delta}_{k+1}$  may be dependent random vectors for some  $k$ , but at each  $k$  the  $\{\Delta_{kj}\}$  are i.i.d. (across  $j$ ). Note that the user has full control over the choice of the  $\Delta_{kj}$  distribution. A valid (and simple) choice is the Bernoulli  $\pm c$  distribution (it is not known at this time if this is the ‘‘best’’ distribution to choose for this application).

The prime rationale for (3.1) is that  $\hat{\mathbf{H}}_k$  is a nearly unbiased estimator of the unknown  $\mathbf{H}$ . Spall (2000) gives conditions such that the Hessian estimate has an  $O(c^2)$  bias (the main such condition is smoothness of  $\log \ell(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}}(i))$ , as reflected in the assumption that  $\mathbf{g}(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}}(i))$  is thrice continuously differentiable in  $\boldsymbol{\theta}$ ).

The symmetrizing operation in (3.1) (the multiple  $1/2$  and the indicated sum) is convenient to maintain a symmetric Hessian estimate. To illustrate how the *individual* Hessian estimates may be quite poor, note that  $\hat{\mathbf{H}}_k$  in (3.1) has (at most) rank two (and may not even be positive semidefinite). This low quality, however, does not prevent the information matrix estimate of interest from being accurate since it is not the Hessian per se that is of interest. The averaging process eliminates the inadequacies of the individual Hessian estimates.

The main source of efficiency for (3.1) is the fact that the estimate requires only a small (fixed) number of gradient or log-likelihood values for any dimension  $p$ . When gradient estimates are available, only two evaluations are needed. When only log-likelihood values are available, each of the

gradient approximations  $\mathbf{G}(\boldsymbol{\theta} + \boldsymbol{\Delta}_k | \mathbf{Z}_{\text{pseudo}})$  and  $\mathbf{G}(\boldsymbol{\theta} - \boldsymbol{\Delta}_k | \mathbf{Z}_{\text{pseudo}})$  require two evaluations of  $\log \ell(\cdot | \mathbf{Z}_{\text{pseudo}})$ . Hence, one approximation  $\hat{\mathbf{H}}_k$  uses four log-likelihood values. The gradient approximation at the two design levels is:

$$\mathbf{G}(\boldsymbol{\theta} \pm \boldsymbol{\Delta}_k | \mathbf{Z}_{\text{pseudo}}) = \left\{ \begin{array}{l} \frac{\log \ell(\boldsymbol{\theta} \pm \boldsymbol{\Delta}_k + \tilde{\boldsymbol{\Delta}}_k | \mathbf{Z}_{\text{pseudo}})}{2} \\ \frac{\log \ell(\boldsymbol{\theta} \pm \boldsymbol{\Delta}_k - \tilde{\boldsymbol{\Delta}}_k | \mathbf{Z}_{\text{pseudo}})}{2} \end{array} \right\} \begin{bmatrix} \tilde{\Delta}_{k1}^{-1} \\ \tilde{\Delta}_{k2}^{-1} \\ \vdots \\ \tilde{\Delta}_{kp}^{-1} \end{bmatrix}, \quad (3.2)$$

with  $\tilde{\boldsymbol{\Delta}}_k = [\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, \dots, \tilde{\Delta}_{kp}]^T$  generated in the same statistical manner as  $\boldsymbol{\Delta}_k$ , but independently of  $\boldsymbol{\Delta}_k$  (in particular, choosing  $\tilde{\Delta}_{ki}$  as independent Bernoulli  $\pm c$  random variables is a valid—but not necessary—choice).

Given the form for the Hessian estimate in (3.1), it is now relatively straightforward to estimate  $\mathbf{F}_n(\boldsymbol{\theta})$ . Averaging Hessian estimates across many  $\mathbf{Z}_{\text{pseudo}}(i)$  yields an estimate of

$$E[\mathbf{H}(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}}(i))] = -\mathbf{F}_n(\boldsymbol{\theta})$$

to within an  $O(c^2)$  bias (the expectation in the left-hand side above is with respect to the pseudodata). The resulting estimate can be made as accurate as desired through reducing  $c$  and increasing the number of  $\hat{\mathbf{H}}_k$  values being averaged.

The averaging of the  $\hat{\mathbf{H}}_k$  values may be done recursively to avoid having to store many matrices. Of course, the interest is not in the Hessian per se; rather the interest is in the (negative) *mean* of the Hessian, according to (2.3) (so the averaging must reflect many different values of  $\mathbf{Z}_{\text{pseudo}}(i)$ ).

Let us now present a step-by-step summary of the above Monte Carlo resampling approach for estimating  $\mathbf{F}_n(\boldsymbol{\theta})$ . Figure 1 is a schematic of the steps.

#### Monte Carlo Resampling Method for Estimating $\mathbf{F}_n(\boldsymbol{\theta})$

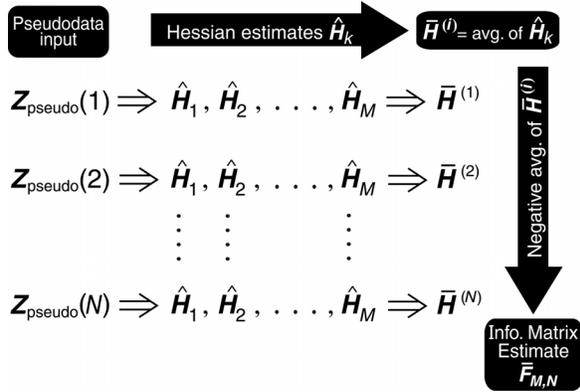
**Step 0 (Initialization)** Determine  $\theta$ , the sample size  $n$ , and the number of pseudodata vectors that will be generated ( $N$ ). Determine whether log-likelihood  $\log \ell(\cdot)$  or gradient information  $\mathbf{g}(\cdot)$  will be used to form the  $\hat{\mathbf{H}}_k$  estimates. Pick the small number  $c$  in the Bernoulli  $\pm c$  distribution used to generate the perturbations  $\Delta_{ki}$ ;  $c = 0.0001$  has been effective in the

author's experience (non-Bernoulli distributions may also be used subject to the conditions mentioned below (3.1)). Set  $i = 1$ .

**Step 1 (Generating pseudodata)** Based on  $\theta$  given in step 0, generate by Monte Carlo the  $i$ th pseudodata vector of  $n$  pseudo-measurements  $\mathbf{Z}_{\text{pseudo}}(i)$ .

**Step 2 (Hessian estimation)** With the  $i$ th pseudodata vector in step 1, compute  $M \geq 1$  Hessian estimates according to the formula (3.1). Let the sample mean of these  $M$  estimates be  $\bar{\mathbf{H}}^{(i)} = \bar{\mathbf{H}}^{(i)}(\theta | \mathbf{Z}_{\text{pseudo}}(i))$ . (As discussed in Section 4,  $M = 1$  has certain optimality properties, but  $M > 1$  is preferred if the pseudodata vectors are expensive to generate relative to the Hessian estimates forming the sample mean  $\bar{\mathbf{H}}^{(i)}$ .)

**Step 3 (Averaging Hessian estimates)** Repeat steps 1 and 2 until  $N$  pseudodata vectors have been processed. Take the negative of the average of the  $N$  Hessian estimates  $\bar{\mathbf{H}}^{(i)}$  produced in step 2; this is the estimate of  $F_n(\theta)$ . (In both steps 2 and 3, it is usually convenient to form the required averages using the standard recursive representation of a sample mean in contrast to storing the matrices and averaging later.) To avoid the possibility of having a nonpositive semidefinite estimate, it may be desirable to take the symmetric square root of the square of the estimate (the `sqrtm` function in MATLAB is useful here). Let  $\bar{\mathbf{F}}_{M,N}(\theta)$  represent the estimate of  $F_n(\theta)$  based on  $M$  Hessian estimates in step 2 and  $N$  pseudodata vectors.



**Figure 1** Schematic of method for forming estimate  $\bar{\mathbf{F}}_{M,N}(\theta)$ .

## 4. THEORETICAL BASIS FOR IMPLEMENTATION

There are several theoretical issues arising in the steps above. One is the question of whether to implement the Hessian estimate-based method from (3.1) rather than a

straightforward averaging based on (2.2). Another is the question of how much averaging to do in step 2 of the procedure in Section 3 (i.e., the choice of  $M$ ). We discuss these two questions, respectively, in Subsections 4.1 and 4.2. To streamline the notation associated with individual components of the information matrix, we generally write  $\mathbf{F}(\theta)$  for  $F_n(\theta)$ .

### 4.1 Lower Variability for Estimate Based on (3.1)

The defining expression for the information matrix in terms of the outer product of gradients (eqn. (2.2)) provides an alternative means of creating a Monte Carlo-based estimate. In particular, at the  $\theta$  of interest, one can simply average values of  $\mathbf{g}(\theta | \mathbf{Z}_{\text{pseudo}}(i))\mathbf{g}(\theta | \mathbf{Z}_{\text{pseudo}}(i))^T$  for a large number of  $\mathbf{Z}_{\text{pseudo}}(i)$ . Let us discuss why the Hessian-based method based on the alternative definition (2.3) is generally preferred. First, in the case where only  $\log \ell(\cdot)$  values are available (i.e., no gradients  $\mathbf{g}(\cdot)$ ), it is unclear how to create an unbiased (or nearly so) estimate of the integrand in (2.2). In particular, using the  $\log \ell(\cdot)$  values to create a near-unbiased estimate of  $\mathbf{g}(\cdot)$  does not generally provide a means of creating an unbiased estimate of the integrand  $\mathbf{g}(\cdot)\mathbf{g}(\cdot)^T$  (i.e., if  $X$  is an unbiased estimate of some quantity,  $X^2$  is not generally an unbiased estimate of the square of the quantity).

The full version of this paper considers the more subtle case where  $\mathbf{g}(\cdot)$  values are directly available. The fundamental advantage of (3.1) arises because the variances of the elements in the information matrix estimate depend on *second moments* of the relevant quantities in the Monte Carlo average, while with averages of  $\mathbf{g}(\cdot)\mathbf{g}(\cdot)^T$  the variances depend on *fourth moments* of the same quantities. This leads to greater variability for a given number ( $N$ ) of pseudodata.

### 4.2 Optimal Choice of $M$

It is mentioned in step 2 of the procedure in Section 3 that it may be desirable to average several Hessian estimates at each pseudodata vector  $\mathbf{Z}_{\text{pseudo}}$ . We now show that this averaging is only recommended if the cost of generating the pseudodata vectors is high. That is, if the computational "budget" allows for  $B$  Hessian estimates (irrespective of whether the estimates rely on new or reused pseudodata), the accuracy of the Fisher information matrix is maximized when each of the  $B$  estimates rely on a new pseudodata vector. On the other hand, if the cost of generating each pseudodata vector  $\mathbf{Z}_{\text{pseudo}}$  is relatively high, there may be advantages to averaging the Hessian estimates at each  $\mathbf{Z}_{\text{pseudo}}$  (see step 2). This must be considered on a case-by-case basis.

Note that  $B = MN$  represents the total number of Hessian estimates being produced (using (3.1)) to form  $\bar{\mathbf{F}}_{M,N}(\theta)$ . The two results below relate  $\bar{\mathbf{F}}_{M,N}(\theta)$  to the true matrix  $\mathbf{F}(\theta)$ .

These results apply in both of the cases where  $\mathbf{G}(\boldsymbol{\theta}|\mathbf{Z}_{\text{pseudo}})$  in (3.1) represents a gradient *approximation* (based on  $\log \ell(\boldsymbol{\theta}|\mathbf{Z}_{\text{pseudo}})$  values) and where  $\mathbf{G}(\boldsymbol{\theta}|\mathbf{Z}_{\text{pseudo}})$  represents the exact gradient  $\mathbf{g}(\boldsymbol{\theta}|\mathbf{Z}_{\text{pseudo}})$ .

**Proposition 1.** Suppose that  $\mathbf{g}(\boldsymbol{\theta}|\mathbf{Z}_{\text{pseudo}})$  is three times continuously differentiable in  $\boldsymbol{\theta}$  for almost all  $\mathbf{Z}_{\text{pseudo}}$ . Then, based on the structure and assumptions of (3.1),  $E[\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta})] = \mathbf{F}(\boldsymbol{\theta}) + O(c^2)$ .

**Proof.** Spall (2000) shows that  $E(\hat{\mathbf{H}}_k | \mathbf{Z}_{\text{pseudo}}) = \mathbf{H}(\boldsymbol{\theta} | \mathbf{Z}_{\text{pseudo}}) + O_Z(c^2)$  under the stated conditions on  $\mathbf{g}(\cdot)$  and  $\Delta_k$ . Because  $\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta})$  is simply a sample mean of  $\hat{\mathbf{H}}_k$  values, the result to be proved follows immediately. Q.E.D.

**Proposition 2.** Suppose that the elements of  $\{\Delta_1^{(1)}, \dots, \Delta_M^{(1)}; \Delta_1^{(2)}, \dots, \Delta_M^{(2)}; \dots; \Delta_1^{(N)}, \dots, \Delta_M^{(N)}; \mathbf{Z}_{\text{pseudo}}(1), \dots, \mathbf{Z}_{\text{pseudo}}(N)\}$  are mutually independent. For a fixed  $B = MN$ , the variance of each element in  $\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta})$  is minimized when  $M = 1$ .

**Proof.** From step 2 in Section 3,  $\bar{\mathbf{H}}^{(i)} = M^{-1} \sum_{k=1}^M \hat{\mathbf{H}}_k$ , where  $\hat{\mathbf{H}}_k = \hat{\mathbf{H}}_k(\mathbf{Z}_{\text{pseudo}}(i))$  for all  $k$ . The  $h_j$ th component of  $\hat{\mathbf{H}}_k$  can be represented in generic form as  $f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i))$ , where  $\Delta_k^{(i)}$  represents the  $p$ -dimensional perturbation vector used to form  $\hat{\mathbf{H}}_k$ . Note that

$$\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \bar{\mathbf{H}}^{(i)} = \frac{1}{MN} \sum_{i=1}^N \sum_{k=1}^M \hat{\mathbf{H}}_k(\mathbf{Z}_{\text{pseudo}}(i)). \quad (4.5)$$

Let  $[\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta})]_{hj}$  denote the  $h_j$ th element of  $\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta})$ . Because the elements of  $\{\Delta_1^{(1)}, \dots, \Delta_M^{(1)}; \Delta_1^{(2)}, \dots, \Delta_M^{(2)}; \dots; \Delta_1^{(N)}, \dots, \Delta_M^{(N)}; \mathbf{Z}_{\text{pseudo}}(1), \dots, \mathbf{Z}_{\text{pseudo}}(N)\}$  are mutually independent, (4.5) implies that the variance of the  $h_j$ th element is given by,

$$\begin{aligned} \text{var}\{[\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta})]_{hj}\} &= \frac{1}{M^2 N^2} \sum_{i=1}^N \sum_{k=1}^M \text{var}[f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i))] \\ &+ \frac{2}{M^2 N^2} \sum_{i=1}^N \sum_{m=1}^M \sum_{k < m} \text{cov}[f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i)), f_{hj}(\Delta_m^{(i)}, \mathbf{Z}_{\text{pseudo}}(i))] \end{aligned} \quad (4.6)$$

Because the  $\Delta_k^{(i)}$  are identically distributed and the  $\mathbf{Z}_{\text{pseudo}}(i)$  are identically distributed, the summands in the first double

sum of (4.6) are identical and the summands in the second double sum are identical. Further,

$$\begin{aligned} &\text{cov}[f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i)), f_{hj}(\Delta_m^{(i)}, \mathbf{Z}_{\text{pseudo}}(i))] \\ &= E[f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i)) f_{hj}(\Delta_m^{(i)}, \mathbf{Z}_{\text{pseudo}}(i))] - \bar{f}_{hj}^2 \\ &= E\{E[f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i)) f_{hj}(\Delta_m^{(i)}, \mathbf{Z}_{\text{pseudo}}(i)) | \mathbf{Z}_{\text{pseudo}}(i)]\} - \bar{f}_{hj}^2 \\ &= E\left\{E[f_{hj}(\Delta_m^{(i)}, \mathbf{Z}_{\text{pseudo}}(i)) | \mathbf{Z}_{\text{pseudo}}(i)]^2\right\} - \bar{f}_{hj}^2, \end{aligned} \quad (4.7)$$

where  $\bar{f}_{hj} \equiv E[f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i))]$ . Because  $E(X^2) \geq [E(X)]^2$  for any real-valued random variable  $X$ , and because  $\bar{f}_{hj} = E\{E[f_{hj}(\Delta_k^{(i)}, \mathbf{Z}_{\text{pseudo}}(i)) | \mathbf{Z}_{\text{pseudo}}(i)]\}$ , the right-hand side of (4.7) is non-negative. Hence, because  $MN$  is a constant ( $= B$ ), the variance of  $[\bar{\mathbf{F}}_{M,N}(\boldsymbol{\theta})]_{hj}$ , as given in (4.6), is minimized when the second double sum on the right-hand side of (4.6) is zero. This happens when  $M = 1$ . Q.E.D.

## 5. IMPLEMENTATION WITH ANTITHETIC RANDOM NUMBERS

*Antithetic random numbers* (ARNs) may sometimes be used in simulation to reduce the variance of sums of random variables. ARNs represent Monte Carlo-generated random numbers such that various pairs of random numbers are negatively correlated. The full version of the paper discusses the use of ARNs.

## 6. NUMERICAL EXAMPLE

Suppose that the data  $\mathbf{z}_i$  are independently distributed  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{P}_i)$  for all  $i$ , where  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are to be estimated and the  $\mathbf{P}_i$  are known. This corresponds to a signal-plus-noise setting where the  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ -distributed signal is observed in the presence of independent  $N(\mathbf{0}, \mathbf{P}_i)$ -distributed noise. The varying covariance matrix for the noise may reflect different quality measurements of the signal. Among other areas, this setting arises in estimating the initial mean vector and covariance matrix in a state-space model from a cross-section of realizations (Shumway, et al., 1981), in estimating parameters for random-coefficient linear models (Sun, 1982), or in small area estimation in survey sampling (Ghosh and Rao, 1994).

Let us consider the following scenario:  $\dim(\mathbf{z}_i) = 4$ ,  $n = 30$ , and  $\mathbf{P}_i = \sqrt{i} \mathbf{U}^T \mathbf{U}$ , where  $\mathbf{U}$  is generated according to a  $4 \times 4$  matrix of uniform (0, 1) random variables (so the  $\mathbf{P}_i$  are identical except for the scale factor  $\sqrt{i}$ ). Let  $\boldsymbol{\theta}$  represent the

unique elements in  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ ; hence,  $p = 4 + 4(4+1)/2 = 14$ . So, there are  $14(14+1)/2 = 105$  unique terms in  $F_n(\boldsymbol{\theta})$  that are to be estimated via the Monte Carlo scheme in Section 3. This is a problem where the analytical form of the information matrix is available (see Shumway, et al., 1981). Hence, the Monte Carlo resampling-based results can be compared with the analytical results. The value of  $\boldsymbol{\theta}$  used to generate the data is also used here as the value of interest in evaluating  $F_n(\boldsymbol{\theta})$ . This value corresponds to  $\boldsymbol{\mu} = \mathbf{0}$  and  $\boldsymbol{\Sigma}$  being a matrix with 1's on the diagonal and 0.5's on the off-diagonals.

This study illustrates two aspects of the resampling method. Table 1 presents results related to the optimality of  $M = 1$  when independent perturbations are used in the Hessian estimates (Subsection 4.2). This study is carried out using only log-likelihood values to construct the Hessian estimates (via using the simultaneous perturbation gradient estimate in (3.2)). The table also presents results related to the value of gradient information (when available) relative to using only log-likelihood values. All studies here are carried out in MATLAB (version 6) using the default random number generators (`rand` and `randn`). Note that there are many ways of comparing matrices; we use two convenient methods below. One is based on the maximum eigenvalue; the other is based on the norm of the difference. For the maximum eigenvalue, the two candidate estimates of the information matrix are compared based on the sample means of the quantity  $|\hat{\lambda}_{\max} - \lambda_{\max}|/\lambda_{\max}$ , where  $\hat{\lambda}_{\max}$  and  $\lambda_{\max}$  denote the maximum eigenvalues of the estimated and true information matrices, respectively. For the norm, the two matrices are compared based on the sample means of the standardized spectral norm of the deviations from the true (known) information matrix  $\|\bar{F}_{M,N}(\boldsymbol{\theta}) - F_n(\boldsymbol{\theta})\|/\|F_n(\boldsymbol{\theta})\|$  (the spectral norm of a square matrix  $A$  is  $\|A\| = [\text{largest eigenvalue of } A^T A]^{1/2}$ ; this appears to be the most commonly used form of matrix norm because of its compatibility with the standard Euclidean vector norm).

Table 1 shows that there is statistical evidence consistent with Proposition 2. Namely in the comparisons of  $\bar{F}_{1,40000}$

with  $\bar{F}_{20,2000}$  (column (a) versus (b)), the  $P$ -value (probability value) computed from a standard matched-pairs  $t$ -test, is 0.002 and 0.0009 for the maximum eigenvalue and norm comparison. These  $P$ -values are based on 50 independent experiments. Hence, there is strong evidence to reject the null hypothesis that  $\bar{F}_{1,40000}$  and  $\bar{F}_{20,2000}$  are equally good in approximating  $F_n(\boldsymbol{\theta})$ ; the evidence is in favor of  $\bar{F}_{1,40000}$  being a better approximation. (Note that computer run times for  $\bar{F}_{1,40000}$  are about 15 percent greater than for  $\bar{F}_{20,2000}$ , reflecting the additional cost of generating the greater number of pseudodata. This supports the comment in Section 4 that a small amount of averaging [ $M > 1$ ] may be desirable in practice even though  $M = 1$  is the optimal solution under the constraint of a fixed  $B = MN$ . Unfortunately, due to the problem-specific nature of the extra cost associated with generating pseudodata, it is not possible in general to determine a priori the optimal amount of averaging under the constraint of equalized run times.) At  $M = 1$  and  $N = 40,000$ , columns (a) and (c) of Table 1 also illustrate the value of gradient information, with both  $P$ -values being very small, indicating strong rejection of the null hypothesis of equality in the accuracy of the approximations. It is seen from the values in the table that the sample mean estimation error ranges from 0.5 to 1.5 percent for the maximum eigenvalue and 1.8 to 5.3 percent for the norm.

## 7. CONCLUDING REMARKS

The Fisher information matrix is widely used in the design and evaluation of systems. Important applications include uncertainty calculation (confidence intervals and prediction bounds), experimental design, and the determination of prior distributions for Bayesian analysis. However, in many realistic processes, analytical evaluation of the information matrix is difficult or impossible.

This paper has presented a relatively simple Monte Carlo means of obtaining the Fisher information matrix for use in complex estimation settings. In contrast to the conventional approach, there is no need to analytically compute the expected value of Hessian matrices or outer products of loss function gradients. The Monte Carlo approach can work with

**Table 1** Numerical assessment of Proposition 2 (column (a) vs. column (b)) and of value of gradient information (column (a) vs. column (c)). Comparisons via mean absolute deviations from maximum eigenvalues and mean spectral norm of difference as a fraction of true values (columns (a), (b), and (c)). Budget of SP Hessian estimates is constant ( $B = MN$ ).  $P$ -values based on two-sided  $t$ -test.

|                    | $M = 1$<br>$N = 40,000$<br>Likelihood values<br>(a) | $M = 20$<br>$N = 2000$<br>Likelihood values<br>(b) | $M = 1$<br>$N = 40,000$<br>Gradient values<br>(c) | $P$ -value<br>(Prop. 2)<br>(a) vs. (b) | $P$ -value<br>(gradient info.)<br>(a) vs. (c) |
|--------------------|---|--|---|--|---|
| Maximum eigenvalue | 0.0103  | 0.0150   | 0.0051  | 0.002                                  | 0.0002  |
| Norm               | 0.0502  | 0.0532   | 0.0183  | 0.0009                                 | $< 10^{-10}$                                  |

either evaluations of the log-likelihood function or the gradient, depending on what information is available. The required expected value in the definition of the information matrix is estimated via a Monte Carlo averaging combined with a simulation-based generation of “artificial” data. The averaging and generation of artificial data are similar to resampling in standard bootstrap methods in statistics. We also presented some theory that is useful in reducing the variability of the estimate through optimal forms of the required averaging and through the use of antithetic random numbers.

## 8. REFERENCES

- [1] Atkinson, A. C. and Donev, A. N. (1992), *Optimum Experimental Designs*, Oxford University Press, Oxford.
- [2] Bickel, P. J. and Doksum, K. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.
- [3] Efron, B. and Hinckley, D. V. (1978), “Assessing the Accuracy of the Maximum Likelihood Estimator: Observed versus Expected Fisher Information” (with discussion), *Biometrika*, vol. 65, pp. 457–487.
- [4] Efron, B. and Tibshirani, R. (1986), “Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy” (with discussion), *Statistical Science*, vol. 1, pp. 54–77.
- [5] Gelfand, A. E. and Smith, A. F. M. (1990), “Sampling-Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, vol. 85, pp. 399–409.
- [6] Ghosh, M. and Rao, J. N. K. (1994), “Small Area Estimation: An Approach” (with discussion), *Statistical Science*, vol. 9, pp. 55–93.
- [7] Hoadley, B. (1971), “Asymptotic Properties of Maximum Likelihood Estimates for the Independent Not Identically Distributed Case,” *Annals of Mathematical Statistics*, vol. 42, pp. 1977–1991.
- [8] Kushner, H. J. and Yang, J. (1995), “Stochastic Approximation with Averaging and Feedback: Rapidly Convergent On-Line Algorithms,” *IEEE Transactions on Automatic Control*, vol. 40, pp. 24–34.
- [9] Kushner, H. J. and Yin, G. G. (1997), *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York.
- [10] Levy, L. J. (1995), “Generic Maximum Likelihood Identification Algorithms for Linear State Space Models,” *Proceedings of the Conference on Information Sciences and Systems*, Baltimore, MD, pp. 659–667.
- [11] Ljung, L. (1999), *System Identification—Theory for the User* (2nd ed.), Prentice Hall PTR, Upper Saddle River, NJ.
- [12] Lunneborg, C. E. (2000), *Data Analysis by Resampling: Concepts and Applications*, Duxbury Press, Pacific Grove, CA.
- [13] Rao, C. R. (1973), *Linear Statistical Inference and its Applications* (2nd ed.), Wiley, New York.
- [14] Shumway, R. H., Olsen, D. E., and Levy, L. J. (1981), “Estimation and Tests of Hypotheses for the Initial Mean and Covariance in the Kalman Filter Model,” *Communications in Statistics—Theory and Methods*, vol. 10, pp. 1625–1641.
- [15] Šimandl, M., Kráľovec, J., and Tichavský, P. (2001), “Filtering, Predictive, and Smoothing Cramér-Rao Bounds for Discrete-Time Nonlinear Dynamic Systems,” *Automatica*, vol. 37, pp. 1703–1716.
- [16] Spall, J. C. (1992), “Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation,” *IEEE Transactions on Automatic Control*, vol. 37, pp. 332–341.
- [17] Spall, J. C. (2000), “Adaptive Stochastic Approximation by the Simultaneous Perturbation Method,” *IEEE Transactions on Automatic Control*, vol. 45, pp. 1839–1853.
- [18] Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization*, Wiley, Hoboken, NJ.
- [19] Sun, F. K. (1982), “A Maximum Likelihood Algorithm for the Mean and Covariance of Nonidentically Distributed Observations,” *IEEE Transactions on Automatic Control*, vol. AC-27, pp. 245–247.
- [20] Wilks, S. S. (1962), *Mathematical Statistics*, Wiley, New York.