Cramér–Rao Bounds and Monte Carlo Calculation of the Fisher Information Matrix in Difficult Problems[1]

James C. Spall
The Johns Hopkins University
Applied Physics Laboratory
Laurel, Maryland 20723-6099 U.S.A.
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.

2. FISHER INFORMATION MATRIX: DEFINITION AND NOTATION

Consider a collection of n random vectors $\boldsymbol{Z}^{(n)} \equiv [z_1, z_2, \ldots, z_n]^T$. 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 $\boldsymbol{Z}^{(n)}$ is known, but that this function depends on an unknown vector $\boldsymbol{\theta}$. Let the probability density/mass function for $\boldsymbol{Z}^{(n)}$ be $p_{\boldsymbol{Z}}(\boldsymbol{\zeta}|\boldsymbol{\theta})$ where $\boldsymbol{\zeta}$ ("zeta") is a dummy matrix representing the possible outcomes for the elements in $\boldsymbol{Z}^{(n)}$ (in $p_{\boldsymbol{Z}}(\boldsymbol{\zeta}|\boldsymbol{\theta})$, the index n on $\boldsymbol{Z}^{(n)}$ is being suppressed for notational convenience). The corresponding likelihood function, say $\ell(\boldsymbol{\theta}|\boldsymbol{\zeta})$, satisfies

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

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 $\boldsymbol{Z}^{(n)}$.

Let us assume that the Hessian matrix

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

exists. 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 $\{\boldsymbol{\zeta}: \ell(\boldsymbol{\theta}|\boldsymbol{\zeta}) > 0\}$ 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 $\boldsymbol{F}_n(\boldsymbol{\theta})$ is related to the Hessian matrix of $\log \ell$ through:

$$\boldsymbol{F}_n(\theta) = -E\left[\boldsymbol{H}\left(\theta \mid \boldsymbol{Z}^{(n)}\right)\middle|\theta\right] \tag{2.2}$$

Note that in some applications, the *observed* information matrix at a particular data set $\boldsymbol{Z}^{(n)}$ (i.e., $-\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}^{(n)})$) may be easier to compute and/or preferred from an inference point of view relative to the actual information matrix $\boldsymbol{F}_n(\boldsymbol{\theta})$ in (2.2) (e.g., Efron and Hinckley, 1978). Although the method in this paper is described for the determination of $\boldsymbol{F}_n(\boldsymbol{\theta})$, the efficient Hessian estimation described in Section 3 may also be used directly for the determination of $\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}^{(n)})$ when it is not easy to calculate the Hessian directly.

3. RESAMPLING-BASED CALCULATION OF THE INFORMATION MATRIX

The calculation of $\boldsymbol{F}_n(\boldsymbol{\theta})$ is often difficult or impossible in practical problems. Obtaining the required first or second

[1]**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. This paper extends results presented in preliminary form at the 2003 ACC to provide more theoretical justification for the Monte Carlo method (Sect. 4), to introduce the use of antithetic random numbers (Sect. 5), and to carry out a new numerical study (Sect. 6). A more complete version of this paper is available upon request.

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 $\boldsymbol{F}_n(\boldsymbol{\theta})$. This approach is useful when analytical methods for computing $\boldsymbol{F}_n(\boldsymbol{\theta})$ are infeasible. The approach makes use of an efficient method for Hessian estimation.

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 $\boldsymbol{F}_n(\boldsymbol{\theta})$. This approach is directly motivated by the definition of $\boldsymbol{F}_n(\boldsymbol{\theta})$ as the mean value of the negative Hessian matrix (eqn. (2.2)). 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 $\boldsymbol{Z}_n = \left[\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \ldots, \boldsymbol{z}_n^T\right]^T$ are jointly normally distributed, $N\left(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})\right)$, 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 ith pseudodata vector be $\boldsymbol{Z}_{\text{pseudo}}(i)$; the use of $\boldsymbol{Z}_{\text{pseudo}}$ 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 $\boldsymbol{Z}_n$) 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 $\log \ell(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ values (alone) or with the gradient $\boldsymbol{g}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}) \equiv \partial \log \ell(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}) \big/ \partial \boldsymbol{\theta}$ 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 $\boldsymbol{G}(\boldsymbol{\theta} | \boldsymbol{Z}_{\text{pseudo}})$ represent either a gradient *approximation* (based on $\log \ell(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ values) or the exact gradient $\boldsymbol{g}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$. Because of its efficiency, the simultaneous perturbation gradient approximation is recommended in the case where only $\log \ell(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ values are available (see Spall, 2000).

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

$$\hat{\boldsymbol{H}}_k = \frac{1}{2}\left\{ \frac{\delta \boldsymbol{G}_k}{2}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right] + \left(\frac{\delta \boldsymbol{G}_k}{2}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right] \right)^T \right\} \tag{3.1}$$

where $\delta \boldsymbol{G}_k \equiv \boldsymbol{G}(\boldsymbol{\theta} + \boldsymbol{\Delta}_k \mid \boldsymbol{Z}_{\text{pseudo}}) - \boldsymbol{G}(\boldsymbol{\theta} - \boldsymbol{\Delta}_k \mid \boldsymbol{Z}_{\text{pseudo}})$ and the perturbation vector $\boldsymbol{\Delta}_k \equiv [\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp}]^T$ is a mean-zero random vector such that the $\{\Delta_{kj}\}$ are "small" symmetrically distributed random variables that are uniformly bounded and satisfy $E\left(\left|1/\Delta_{kj}\right|\right) < \infty$ 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 Δ_{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 $\hat{\boldsymbol{H}}_k$ is a nearly unbiased estimator of the unknown $\boldsymbol{H}$. Spall (2000) gives conditions such that the Hessian estimate has an $O(c^2)$ bias (the main such condition is smoothness of $\log \ell(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i))$, as reflected in the assumption that $\boldsymbol{g}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i))$ 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 $\hat{\boldsymbol{H}}_k$ in (3.1) has (at most) rank two (and may not even be positive semi-definite). 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 $\boldsymbol{G}(\boldsymbol{\theta} + \boldsymbol{\Delta}_k \mid \boldsymbol{Z}_{\text{pseudo}})$ and

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

$$\boldsymbol{G}(\boldsymbol{\theta}\pm\boldsymbol{\Delta}_k \mid \boldsymbol{Z}_{\text{pseudo}}) = \frac{\log \ell(\boldsymbol{\theta}\pm\boldsymbol{\Delta}_k+\tilde{\boldsymbol{\Delta}}_k \mid \boldsymbol{Z}_{\text{pseudo}})-\log \ell(\boldsymbol{\theta}\pm\boldsymbol{\Delta}_k-\tilde{\boldsymbol{\Delta}}_k \mid \boldsymbol{Z}_{\text{pseudo}})}{2}\begin{bmatrix}\tilde{\Delta}_{k1}^{-1}\\ \tilde{\Delta}_{k2}^{-1}\\ \cdot\\ \cdot\\ \cdot\\ \tilde{\Delta}_{kp}^{-1}\end{bmatrix}, \quad (3.2)$$

with $\tilde{\boldsymbol{\Delta}}_k = [\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, ..., \tilde{\Delta}_{kp}]^T$ 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 $\boldsymbol{F}_n(\boldsymbol{\theta})$. Averaging Hessian estimates across many Zpseudo(i) yields an estimate of

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

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 $\hat{\boldsymbol{H}}_k$ values being averaged. The averaging of the $\hat{\boldsymbol{H}}_k$ 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 (2.2) (so the averaging must reflect many different values of $\boldsymbol{Z}_{\text{pseudo}}(i)$).

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

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

Step 0. (Initialization) Determine $\boldsymbol{\theta}$, 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 $\boldsymbol{g}(\cdot)$ will be used to form the $\hat{\boldsymbol{H}}_k$ estimates. Pick the small number c in the Bernoulli $\pm c$ distribution used to generate the perturbations Δ_{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 $\boldsymbol{\theta}$ given in step 0, generate by Monte Carlo the ith pseudodata vector of n pseudo-measurements $\boldsymbol{Z}_{\text{pseudo}}(i)$.

Step 2. (Hessian estimation) With the ith 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 $\bar{\boldsymbol{H}}^{(i)} = \bar{\boldsymbol{H}}^{(i)}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i))$. (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 $\bar{\boldsymbol{H}}^{(i)}$.)

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 $\bar{\boldsymbol{H}}^{(i)}$ produced in step 2; this is the estimate of $\boldsymbol{F}_n(\boldsymbol{\theta})$. (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 non-positive 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 $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$ represent the estimate of $\boldsymbol{F}_n(\boldsymbol{\theta})$ based on M Hessian estimates in step 2 and N pseudodata vectors.

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 the outer produce of gradients. 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 $\boldsymbol{F}(\boldsymbol{\theta})$ for $\boldsymbol{F}_n(\boldsymbol{\theta})$.

4.1 Lower Variability for Estimate Based on (3.1)

Let us consider the case where $\boldsymbol{g}(\cdot)$ values are directly available. The argument below is only a sketch of the reason that the form in (3.1) is preferred over a straightforward averaging of outer product values $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$ (across $\boldsymbol{Z}_{\text{pseudo}}(i)$), as it involves some "informal" (but very reasonable) approximations. 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 $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$ the variances depend on *fourth moments* of the same quantities. This leads to greater variability for a given number (N) of pseudodata. To illustrate the advantage, consider the special case where the point of evaluation $\boldsymbol{\theta}$ is close to a "true" value $\boldsymbol{\theta}^*$. Further, let us suppose that both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ are close to the maximum likelihood estimate for $\boldsymbol{\theta}$ at each data set $\boldsymbol{Z}_{\text{pseudo}}(i)$, say $\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))$ (i.e., n is large enough so that $\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i)) \approx \boldsymbol{\theta}^*$). Note that $\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))$ corresponds to a point where $\boldsymbol{g}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i)) = \boldsymbol{0}$. Let us compare the variance of the diagonal elements of the estimate of the information matrix using the average of the Hessian estimates (3.1) and the average of outer products (it is not assumed that the analyst knows that the information matrix is diagonal; hence, the full matrix is estimated).

In determining the variance based on (3.1), suppose that $M = 1$. The estimate $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$ is then formed from an average of N Hessian estimates of the form (3.1) (we see in Subsection 4.2 that $M = 1$ is an optimal solution in a certain sense). Hence, the variance of the jjth component of the estimate $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta}) = \bar{\boldsymbol{F}}_{1,N}(\boldsymbol{\theta})$ is

$$\operatorname{var}\left\{\left[\bar{\boldsymbol{F}}_{1,N}(\boldsymbol{\theta})\right]_{jj}\right\} = \frac{1}{N}\operatorname{var}\left(\hat{H}^2_{1;jj}\right) \tag{4.1}$$

Let $O_{(\cdot)}(c^2)$ denote a random "big-O" term, where the subscript denotes the relevant randomness; for example, $O_{\boldsymbol{Z},\boldsymbol{\Delta}_1}(c^2)$ denote a random "big-O" term dependent on $\boldsymbol{Z}_{\text{pseudo}}(i)$ and $\boldsymbol{\Delta}_1$ such that $O_{\boldsymbol{Z},\boldsymbol{\Delta}_1}(c^2)/c^2$ is bounded almost surely (a.s.) as $c \to 0$. Then, by Spall (2000), the jjth component of the estimate $\hat{\boldsymbol{H}}_1 = \hat{\boldsymbol{H}}_1(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i))$ is

$$\hat{H}_{1;jj} = H_{jj} + \sum_{\ell \neq j} H_{j\ell}\frac{\Delta_{1\ell}}{\Delta_{1j}} + O_{\boldsymbol{Z},\boldsymbol{\Delta}_1}(c^2),$$

where the pseudodata argument (and index i) and point of evaluation $\boldsymbol{\theta}$ have been suppressed. Let us now invoke one of the assumptions above in order to avoid a hopelessly messy variance expression. Namely, it is assumed that n is "large" and likewise that the points $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$, and $\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))$ are close to one another, implying that the Hessian matrix is nearly a constant independent of $\boldsymbol{Z}_{\text{pseudo}}(i)$ (i.e., $\log \ell(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i))$ is close to a quadratic function in the vicinity of $\boldsymbol{\theta}$); this is tantamount to assuming that n is large enough so that $\boldsymbol{H}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i)) \approx \boldsymbol{F}(\boldsymbol{\theta})$. Hence, ignoring the $O_{\boldsymbol{Z},\Delta_1}(c^2)$ error term,

$$\operatorname{var}\left(\hat{H}^2_{1;jj}\right) \approx \sum_{\ell \neq j} F^2_{j\ell} \tag{4.2}$$

where $F_{j\ell}$ denotes the $j\ell$th component of $\boldsymbol{F}(\boldsymbol{\theta})$.

Let us now analyze the form based on averages of $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$. Analogous to (4.1), the variance of the jjth component of the estimate of the information matrix is

$$\frac{1}{N}\operatorname{var}\left(g_j^2\right), \tag{4.3}$$

where g_j is the jth component of $\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}(i))$. From the mean value theorem,

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}(i)) &\approx \boldsymbol{g}(\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))|\boldsymbol{Z}_{\text{pseudo}}(i)) \\ &\quad + \boldsymbol{F}(\boldsymbol{\theta})\left[\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))\right] \\ &= \boldsymbol{F}(\boldsymbol{\theta})\left[\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))\right],\end{aligned}$$

where the approximation in the first line results from the assumption that $\boldsymbol{H}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i)) \approx \boldsymbol{F}(\boldsymbol{\theta})$. Hence, in analyzing the variance of the jjth component of $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$ according to (4.3), we have

$$\operatorname{var}\left(g_j^2\right) \approx \operatorname{var}\left\{\left[\sum_{\ell=1}^{p} F_{j\ell}\left(\theta_\ell - \hat{\theta}_{ML,\ell}\right)\right]^2\right\},$$

where θ_ℓ and $\hat{\theta}_{ML,\ell}$ are the ℓth components of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))$. From asymptotic distribution theory (assuming that the moments of $\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Z}_{\text{pseudo}}(i))$ correspond to the moments from the asymptotic distribution), we have, $E\left[\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML}\right)^T\right] \approx \boldsymbol{F}(\boldsymbol{\theta}^*)^{-1}$; further, $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML}$ is (at least approximately) asymptotically *normal with mean zero* since $\boldsymbol{\theta} \approx \boldsymbol{\theta}^*$. Because $E[\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T] = \boldsymbol{F}(\boldsymbol{\theta})$, the above implies

$$\begin{aligned}\operatorname{var}\left(g_j^2\right) &\approx \sum_{\ell=1}^{p}\sum_{m=1}^{p} F_{j\ell}F_{jm}E\left[\left(\theta_\ell - \hat{\theta}_{ML,\ell}\right)^2\left(\theta_m - \hat{\theta}_{ML,m}\right)^2\right] - [F_{jj}(\boldsymbol{\theta})]^2 \\ &\approx \sum_{\ell=1}^{p}\sum_{m\neq\ell} F^2_{j\ell}F^2_{jm}\left[E_{\ell\ell}(\boldsymbol{\theta})E_{mm}(\boldsymbol{\theta}) + 2E_{\ell m}(\boldsymbol{\theta})^2\right] \\ &\quad + 3\sum_{\ell=1}^{p} F^4_{j\ell}E_{\ell\ell}(\boldsymbol{\theta})^2 - [F_{jj}(\boldsymbol{\theta})]^2,\end{aligned} \tag{4.4}$$

where E_{jm} denotes the jmth component of $\boldsymbol{F}(\boldsymbol{\theta})^{-1}$ and the last equality follows by a result in Mardia, et al. (1979, p. 95) (which is a generalization of the relationship that $X \sim N(0, \sigma^2)$ implies $E(X^4) = 3\sigma^4$).

Unfortunately, the general expression in (4.4) is unwieldy. However, if we make the assumption that the off-diagonal elements in $\boldsymbol{F}(\boldsymbol{\theta})$ are small in magnitude relative to the diagonal elements, then for substitution into (4.3), $\operatorname{var}\left(g_j^2\right) \approx 2F^2_{jj}$. The corresponding expression for the (3.1)-based approach with substitution into (4.1) is $\operatorname{var}\left(\hat{H}^2_{1;jj}\right) \approx 0$ (of course, the exact variance will be slightly non-negative due to the approximations involved in (4.2)). So the Hessian estimate-based method of (3.1) provides a more precise estimate 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 $\boldsymbol{Z}_{\text{pseudo}}$. 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 $\boldsymbol{Z}_{\text{pseudo}}$ is relatively high, there may be advantages to averaging the Hessian estimates at each $\boldsymbol{Z}_{\text{pseudo}}$ (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 $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$. The two results below relate $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$ to the true matrix $\boldsymbol{F}(\boldsymbol{\theta})$. These results apply in both of the cases where

$\boldsymbol{G}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ in (3.1) represents a gradient *approximation* (based on $\log \ell(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ values) and where $\boldsymbol{G}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ represents the exact gradient $\boldsymbol{g}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$. Proofs of Propositions 1 and 2 are available in the more complete version of this paper.

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

Proposition 2. Suppose that the elements of $\left\{\boldsymbol{\Delta}_1^{(1)},...,\boldsymbol{\Delta}_M^{(1)}; \boldsymbol{\Delta}_1^{(2)},...,\boldsymbol{\Delta}_M^{(2)};...; \boldsymbol{\Delta}_1^{(N)},...,\boldsymbol{\Delta}_M^{(N)}; \boldsymbol{Z}_{\text{pseudo}}(1),...,\boldsymbol{Z}_{\text{pseudo}}(N)\right\}$ are mutually independent. For a fixed $B = MN$, the variance of each element in $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$ is minimized when $M = 1$.

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. Recall the basic formula for the variance of the sum of two random variables: $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$. It is apparent that the variance of the sum can be reduced over that in the independent X, Y case if the correlation between the two variables can be made negative. In the case of interest here, the sums will represent averages of Hessian estimates. Because ARNs are based on pairs of random variables, it is sufficient to consider $M = 2$ (although it is possible to implement ARNs based on multiple pairs, i.e., M being some multiple of two). ARNs are complementary to common random numbers, a standard tool in simulation for reducing variances associated with *differences* of random variables (e.g., Spall, 2003, Sect. 14.4).

Unfortunately, ARNs cannot be implemented blindly in the hope of improving the estimate; it is often difficult to know a priori if ARNs will lead to improved estimates. The practical implementation of ARNs often involves as much art as science. As noted in Law and Kelton (2000, p. 599), it is generally useful to conduct a small-scale pilot study to determine the value (if any) in a specific application. When ARNs are effective, they provide a "free" method of improving the estimates (e.g. Frigessi, et al., 2000, use them effectively to reduce the variance of Markov chain Monte Carlo schemes).

As shown in Proposition 2 of Section 4, the variance of each element in $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$ is minimized when $M = 1$ given a fixed "budget" of $B = MN$ Hessian estimates being produced (i.e., there is no averaging of Hessian estimates at each $\boldsymbol{Z}_{\text{pseudo}}(i)$). This result depends on the perturbation vectors $\boldsymbol{\Delta}_k^{(i)}$ being i.i.d. Suppose now that for a given i, we consider $M = 2$ and allow *dependence* between the perturbation vectors at $k = 1$ and $k = M = 2$, but otherwise retain all statistical properties for the perturbations mentioned below (3.1) (e.g., mean zero, symmetrically distributed, finite inverse moments, etc.). The complete version of this paper provides a sketch of how ARNs may be used towards reducing the variance of the information matrix estimate when $\boldsymbol{g}(\cdot)$ values are directly available.

6. Numerical Example

Suppose that the data $\boldsymbol{z}_i$ are independently distributed $N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{P}_i)$ for all i, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are to be estimated and the $\boldsymbol{P}_i$ 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 $N(\boldsymbol{0}, \boldsymbol{P}_i)$-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: $\dim(\boldsymbol{z}_i) = 4$, $n = 30$, and $\boldsymbol{P}_i = \sqrt{i}\,\boldsymbol{U}^T\boldsymbol{U}$, where $\boldsymbol{U}$ is generated according to a 4×4 matrix of uniform (0, 1) random variables (so the $\boldsymbol{P}_i$ 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 $\boldsymbol{F}_n(\boldsymbol{\theta})$ 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 $\boldsymbol{F}_n(\boldsymbol{\theta})$. This value corresponds to $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma}$ being a matrix with 1's on the diagonal and 0.5's on the off-diagonals.

This study illustrates three 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. Table 2 illustrates the value of ARNs (Section 5). 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 $\left|\hat{\lambda}_{\max} - \lambda_{\max}\right| / \lambda_{\max}$, 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 $\|\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta}) - \boldsymbol{F}_n(\boldsymbol{\theta})\| / \|\boldsymbol{F}_n(\boldsymbol{\theta})\|$ (the spectral norm of a square matrix $\boldsymbol{A}$ is $\|\boldsymbol{A}\|$ = [largest eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$]$^{1/2}$; 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 $\bar{\boldsymbol{F}}_{1,40000}$ with $\bar{\boldsymbol{F}}_{20,2000}$ (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 $\bar{\boldsymbol{F}}_{1,40000}$ and $\bar{\boldsymbol{F}}_{20,2000}$ are equally good in approximating $\boldsymbol{F}_n(\boldsymbol{\theta})$; the evidence is in favor of $\bar{\boldsymbol{F}}_{1,40000}$ being a better approximation. 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.

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.

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

Table 2 contains the results for the study of ARNs. In this study, ARNs are implemented for the first three (of four) elements for the $\boldsymbol{\mu}$ vector; the remaining element of $\boldsymbol{\mu}$ and all elements of $\boldsymbol{\Sigma}$ used the conventional independent sampling. The basis for this choice is prior information that the off-diagonal elements in the Hessian matrices for the first three elements are similar in magnitude. As in Table 1, we use the difference in maximum eigenvalues and the normed matrix deviation as the basis for comparison (both normalized by their true values). Because ARNs are implemented on only a subset of the $\boldsymbol{\mu}$ parameters, this study is restricted to the eigenvalues and norms of only the $\boldsymbol{\mu}$ portion of the information matrix (a 4×4 block of the 14×14 information matrix). Direct gradient evaluations are used in forming the Hessian estimates (3.1). Based on 100 independent experiments, we see relatively low *P*-values for both criteria, indicating that ARNs offer statistically significant improvement. However, this improvement is more restrictive than the overall improvement associated with Proposition 2 because it only applies to a subset of elements in $\boldsymbol{\theta}$. There is no statistical evidence of improved estimates for the $\boldsymbol{\Sigma}$ part of the information matrix. Of course, different implementations on this problem (i.e., to include some or all components of $\boldsymbol{\Sigma}$ in the modified generation of the perturbation vector) or implementations on other problems may yield broader improvement subject to conditions discussed in Section 5.

Table 2. Numerical assessment of ARNs. Comparisons via mean absolute deviations from maximum eigenvalue of $\boldsymbol{\mu}$ block of $\boldsymbol{F}_n(\boldsymbol{\theta})$ (n = 30) as a fraction of true value and mean spectral norm on $\boldsymbol{\mu}$ block as a fraction of true value. *P*-values based on two-sided *t*-test.

	$M = 1$ $N = 40{,}000$ **No ARNs**	$M = 2$ $N = 20000$ **ARNs**	*P*-value
Maximum eigenvalue	0.0037	0.0024	0.001
Norm	0.0084	0.0071	0.018

REFERENCES

[1] Bickel, P. J. and Doksum, K. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.

[2] 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.

[3] Frigessi, A., Gasemyr, J., and Rue, H. (2000), "Antithetic Coupling of Two Gibbs Sampling Chains," *Annals of Statistics*, vol. 28, pp. 1128–1149.

[4] Ghosh, M. and Rao, J. N. K. (1994), "Small Area Estimation: An Approach" (with discussion), *Statistical Science*, vol. 9, pp. 55–93.

[5] 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.

[6] Law, A. M. and Kelton, W. D. (2000), *Simulation Modeling and Analysis* (3rd ed.), McGraw-Hill, New York.

[7] 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.

[8] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, New York.

[9] 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.

[10] Šimandl, M., Královec, J., and Tichavský, P. (2001), "Filtering, Predictive, and Smoothing Cramér-Rao Bounds for Discrete-Time Nonlinear Dynamic Systems," *Automatica*, vol. 37, pp. 1703–1716.

[11] Spall, J. C. (1992), "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Transactions on Automatic Control*, vol. 37, pp. 332–341.

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

[13] Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization*, Wiley, Hoboken, NJ.

[14] 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.

[15] Wilks, S. S. (1962), *Mathematical Statistics*, Wiley, New York.