To appear: *Journal of Computational and Graphical Statistics*, December 2005, vol. 14(4)

MONTE CARLO COMPUTATION OF THE FISHER INFORMATION MATRIX IN NONSTANDARD SETTINGS[1]

James C. Spall
james.spall@jhuapl.edu

The Johns Hopkins University
Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, Maryland 20723-6099 U.S.A.

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 (SP); Hessian matrix estimation; 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

[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 N00024-03-D-6606, and the JHU/APL IRAD Program. I appreciate the helpful comments of the reviewer and Associate Editor.

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 NOTATION

Suppose that the *i*th measurement of a process is $\boldsymbol{z}_i$ and that a stacked vector of n such measurement vectors is $\boldsymbol{Z}_n \equiv [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, ..., \boldsymbol{z}_n^T]^T$. Let us assume that the *general form* for the joint probability density or probability mass (or hybrid density/mass) function for $\boldsymbol{Z}_n$ is known, but that this function depends on an unknown vector θ. Let the probability density/mass function for $\boldsymbol{Z}_n$ be $p_{\boldsymbol{Z}}(\zeta|\theta)$ where ζ ("zeta") is a dummy vector representing the possible outcomes for $\boldsymbol{Z}_n$ (in $p_{\boldsymbol{Z}}(\zeta|\theta)$, the index n on $\boldsymbol{Z}_n$ is being suppressed for notational convenience). The corresponding likelihood function, say $\ell(\theta|\zeta)$, satisfies

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

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 $\boldsymbol{Z}_n$.

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

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

In the case where the underlying data $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ are independent (and even in many cases where the data may be dependent), the magnitude of $\boldsymbol{F}_n(\boldsymbol{\theta})$ will grow at a rate proportional to n since $\log \ell(\cdot)$ will represent a sum of n random terms. Then, the bounded quantity $\boldsymbol{F}_n(\boldsymbol{\theta})/n$ is employed as an average information matrix over all measurements.

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 $\log \ell(\cdot)$ is twice differentiable in $\boldsymbol{\theta}$. That is, the Hessian matrix

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

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

$$\boldsymbol{F}_n(\boldsymbol{\theta}) = -E\left[\boldsymbol{H}(\boldsymbol{\theta} \mid \boldsymbol{Z}_n) \middle| \boldsymbol{\theta} \right]. \tag{2.3}$$

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 $\boldsymbol{Z}_n$ (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 (2.3) (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 $\boldsymbol{F}_n(\theta)$ 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).

This section outlines a computer resampling approach to estimating $\boldsymbol{F}_n(\theta)$ that is useful when analytical methods for computing $\boldsymbol{F}_n(\theta)$ are infeasible. The approach makes use of a computationally efficient and easy-to-implement method for Hessian estimation that was described in Spall (2000) in the context of optimization. The computational efficiency follows by the low number of log-likelihood or gradient values needed to produce each Hessian estimate. While there is no optimization here per se, we use the same basic simultaneous perturbation (SP) formula for Hessian estimation (this is the same SP principle given earlier in Spall, 1992, for *gradient* estimation). However, the way in which the individual Hessian estimates are averaged differs from Spall (2000) because of the distinction between the problem of recursive optimization and the problem of estimation of $\boldsymbol{F}_n(\theta)$.

The essence of the method is to produce a large number of SP estimates of the Hessian matrix of $\log \ell(\cdot)$ and then average the negative of these estimates to obtain an approximation to $\boldsymbol{F}_n(\theta)$. This approach is directly motivated by the definition of $\boldsymbol{F}_n(\theta)$ as the mean value of the negative Hessian matrix (eqn. (2.3)). To produce the SP Hessian estimates, we generate *pseudodata vectors* in a Monte Carlo manner. The pseudodata are generated according to a bootstrap resampling scheme treating the chosen θ as "truth." The pseudodata are generated according to the probability model $p_{\boldsymbol{Z}}(\zeta|\theta)$ given in (2.1). So, for example, if it is assumed that the real data $\boldsymbol{Z}_n$ are jointly normally distributed, $N(\mu(\theta), \Sigma(\theta))$, then the pseudodata are generated by Monte Carlo according to a normal distribution based on a mean μ and covariance matrix Σ

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 $\boldsymbol{\theta}$.

Hence, the basis for the technique is to use computational horsepower in lieu of traditional detailed theoretical analysis to determine $\boldsymbol{F}_n(\boldsymbol{\theta})$. 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 $\boldsymbol{F}_n(\boldsymbol{\theta})$ 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 approach below can work with either $\log \ell(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ values (alone) or with the gradient $\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}) \equiv \partial \log \ell(\boldsymbol{\theta}|\,\boldsymbol{Z}_{\text{pseudo}})/\partial\boldsymbol{\theta}$ 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 $\boldsymbol{G}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ represent either a gradient *approximation* (based on $\log \ell(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ values) or the exact gradient $\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$. Because of its efficiency, the SP gradient approximation is recommended in the case where only $\log \ell(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ 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 $\delta \boldsymbol{G}_k \equiv \boldsymbol{G}(\boldsymbol{\theta} + \boldsymbol{\Delta}_k \,|\, \boldsymbol{Z}_{\text{pseudo}}) - \boldsymbol{G}(\boldsymbol{\theta} - \boldsymbol{\Delta}_k \,|\, \boldsymbol{Z}_{\text{pseudo}})$ 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 $|\Delta_{kj}| \le c$ 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 $\log \ell(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}(i))$, as reflected in the assumption that $\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}(i))$ is thrice continuously differentiable in $\boldsymbol{\theta}$). Proposition 1 in Section 4 below considers this further in the context of the resulting (small) bias in the estimate of the information matrix.

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

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

with $\tilde{\Delta}_k = [\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, ..., \tilde{\Delta}_{kp}]^T$ generated in the same statistical manner as Δ_k, but independently of Δ_k. In particular, choosing $\tilde{\Delta}_{ki}$ as independent Bernoulli $\pm c$ random variables is a valid—but not necessary—choice. (With small $c > 0$, note that in the Bernoulli case, (3.2) has an easy interpretation as an approximate directional derivative of $\log \ell$ in the direction of a given vector of ± 1 elements at the point $\theta + \Delta_k$ or $\theta - \Delta_k$.)

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

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

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 (2.3) (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 $\boldsymbol{F}_n(\theta)$. Let $\Delta_k^{(i)}$ represent the kth perturbation vector for the ith realization (i.e., for $\boldsymbol{Z}_{\text{pseudo}}(i)$). Figure 1 is a schematic of the steps.

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

Step 0. **(Initialization)** Determine θ, 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 $\Delta_{kj}^{(i)}$; $c = 0.0001$ 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 θ 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 *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 $\bar{\boldsymbol{H}}^{(i)} = \bar{\boldsymbol{H}}^{(i)}(\theta \mid \boldsymbol{Z}_{\text{pseudo}}(i))$. (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 $\bar{\boldsymbol{H}}^{(i)}$.) Unless using antithetic random numbers (Section 4), the perturbation vectors $\left\{\Delta_k^{(i)}\right\}$ should be mutually independent *across* realizations *i* and *along* the realizations (along *k*). (In the case where only $\log \ell(\theta|\boldsymbol{Z}_{\text{pseudo}})$ values are available and SP gradient approximations are being used to form the $\boldsymbol{G}(\cdot)$ values, the perturbations forming the gradient approximations, say $\left\{\tilde{\Delta}_k^{(i)}\right\}$, should likewise be mutually independent.)

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 $\boldsymbol{F}_n(\theta)$. (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 $\bar{\boldsymbol{F}}_{M,N}(\theta)$ represent the estimate of $\boldsymbol{F}_n(\theta)$ based on M Hessian estimates in step 2 and N pseudodata vectors.

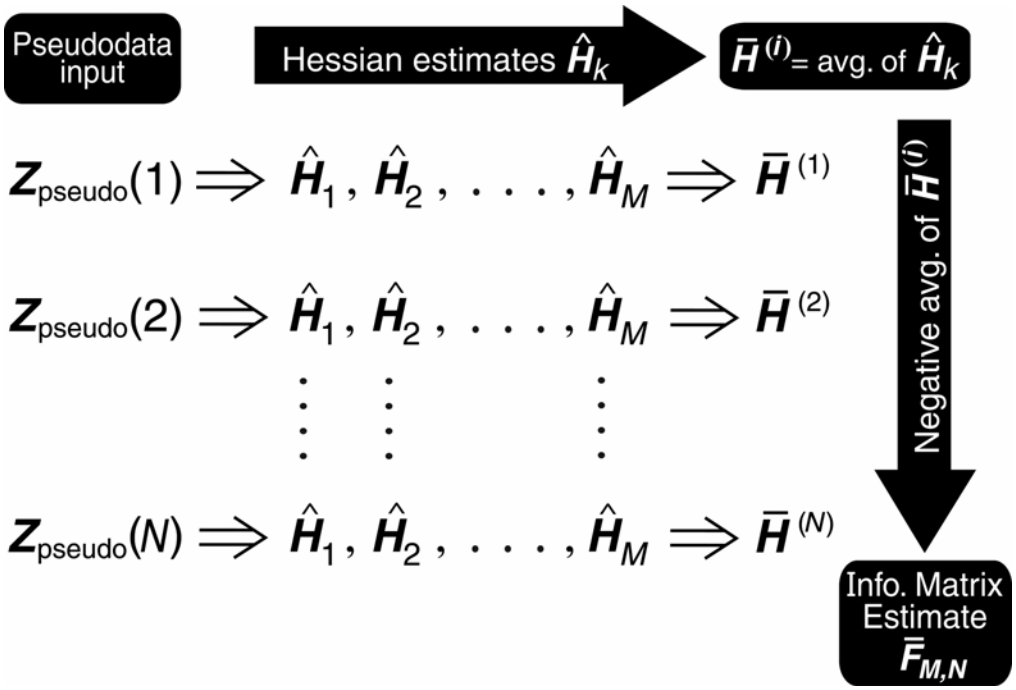

Figure 1. Schematic of method for forming estimate $\bar{F}_{M,N}(\theta)$.

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. A final question pertains to the choice of Hessian estimate, and whether there may be advantages to using a form other than the SP form above. This is discussed in Subsection 4.3. To streamline the notation associated with individual components of the information matrix, we generally write $\boldsymbol{F}(\theta)$ for $\boldsymbol{F}_n(\theta)$.

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 θ of interest, one can simply average values of $\boldsymbol{g}(\theta|\boldsymbol{Z}_{\text{pseudo}}(i))\boldsymbol{g}(\theta|\boldsymbol{Z}_{\text{pseudo}}(i))^T$ for a large number of $\boldsymbol{Z}_{\text{pseudo}}(i)$. 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 $\boldsymbol{g}(\cdot)$), 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 $\boldsymbol{g}(\cdot)$ does not generally provide a means of creating an unbiased estimate of the integrand $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$ (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).

Let us now consider the more subtle case where $\boldsymbol{g}(\cdot)$ values are directly available. The argument below is 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 $\boldsymbol{Z}_{\text{pseudo}}(i)$). A more rigorous analysis of the type below would involve several applications of the Lebesgue dominated convergence theorem and some very messy expansions and higher moment calculations (we have not pursued this). 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 $\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}(i)) = \boldsymbol{0}$. 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

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

where $\hat{H}_{1;jj}$ denotes the *jj*th element of $\hat{\boldsymbol{H}}_1 = \hat{\boldsymbol{H}}_1(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}(i))$. Let $O_{(\cdot)}(c^2)$ denote a random "big-*O*" term, where the subscript denotes the relevant randomness; for example, $O_{\boldsymbol{Z},\boldsymbol{\Delta}_1}(c^2)$ denotes a random "big-*O*" term dependent on $\boldsymbol{Z}_{\text{pseudo}}(i)$ and $\boldsymbol{\Delta}_1$ such that $O_{\boldsymbol{Z},\boldsymbol{\Delta}_1}(c^2)\big/c^2$ is bounded almost surely (a.s.) as $c \to 0$. Then, by Spall (2000), the *jj*th element of $\hat{\boldsymbol{H}}_1$ 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., $\log \ell(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}(i))$ is close to a quadratic function in the vicinity of $\boldsymbol{\theta}$); this is tantamount to assuming that *n* is large enough so that $\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}}(i)) \approx -\boldsymbol{F}(\boldsymbol{\theta})$. Hence, given the independence of the $\{\Delta_{1j}\}$ and assuming the dominated convergence theorem applies to the $O_{\boldsymbol{Z},\Delta_1}(c^2)$ error term,

$$\text{var}\left(\hat{H}_{1;jj}\right) \approx \sum_{\ell \neq j} F_{j\ell}^2 + O(c^2) \tag{4.2}$$

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 *jj*th component of the estimate of the information matrix is

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

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

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

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

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

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

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

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), $\text{var}\left(g_j^2\right) \approx 2F_{jj}^2$. The corresponding expression for the (3.1)-based approach with substitution into (4.1) is $\text{var}\left(\hat{H}_{1;jj}\right) \approx O(c^2)$. So, with small c, the

Hessian estimate-based method of (3.1) provides a more precise estimate for a given number (N) of pseudodata in the sense that variance of the jjth element of $\bar{\boldsymbol{F}}_{1,N}(\boldsymbol{\theta})$ is $O(c^2)/N$ while the corresponding variance of the jjth element of the method based on averages of $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$ is approximately $2F_{jj}^2/N$. (Note that each calculation of (3.1) requires two gradient values while each $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$ uses only one gradient. Equalizing the number of gradient values to $2N$ for each method reduces the $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$-based variance to F_{jj}^2/N at the expense of having the $\boldsymbol{g}(\cdot)\boldsymbol{g}(\cdot)^T$-based method take twice as many pseudodata as needed in $\bar{\boldsymbol{F}}_{1,N}(\boldsymbol{\theta})$.)

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

Proposition 1. Suppose that $\boldsymbol{g}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ 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), $E\left[\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})\right] = \boldsymbol{F}(\boldsymbol{\theta}) + O(c^2)$.

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

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 $M = 1$.

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

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

Let $\left[\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})\right]_{hj}$ denote the hjth element of $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$. Because the $\left\{\boldsymbol{\Delta}_1^{(1)},...,\boldsymbol{\Delta}_M^{(1)};\ \boldsymbol{\Delta}_1^{(2)},...,\boldsymbol{\Delta}_M^{(2)};....;\ \boldsymbol{\Delta}_1^{(N)},...,\boldsymbol{\Delta}_M^{(N)};\ \boldsymbol{Z}_{\text{pseudo}}(1),...,\boldsymbol{Z}_{\text{pseudo}}(N)\right\}$ are mutually independent, (4.5) implies that the variance of the hjth element is given by,

$$\begin{aligned}\operatorname{var}\left\{\left[\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})\right]_{hj}\right\} &= \frac{1}{M^2N^2}\sum_{i=1}^{N}\sum_{k=1}^{M}\operatorname{var}\left[\, f_{hj}(\boldsymbol{\Delta}_k^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i))\right] \\ &\quad + \frac{2}{M^2N^2}\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{k<m}\operatorname{cov}\left[\, f_{hj}(\boldsymbol{\Delta}_k^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i)),\ f_{hj}(\boldsymbol{\Delta}_m^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i))\right]. \end{aligned} \tag{4.6}$$

Because the $\boldsymbol{\Delta}_k^{(i)}$ are identically distributed and the $\boldsymbol{Z}_{\text{pseudo}}(i)$ are identically distributed, the summands in the first multiple sum of (4.6) are identical and the summands in the second multiple sum are identical. Further,

$$\begin{aligned}
&\operatorname{cov}\left[f_{hj}(\boldsymbol{\Delta}_k^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i)),\, f_{hj}(\boldsymbol{\Delta}_m^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i)) \right] \\
&\quad = E\left[f_{hj}(\boldsymbol{\Delta}_k^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i))\, f_{hj}(\boldsymbol{\Delta}_m^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i)) \right] - \bar{f}_{hj}^2 \\
&\quad = E\left\{ E\left[f_{hj}(\boldsymbol{\Delta}_k^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i))\, f_{hj}(\boldsymbol{\Delta}_m^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i)) \,\middle|\, \boldsymbol{Z}_{\text{pseudo}}(i) \right] \right\} - \bar{f}_{hj}^2 \\
&\quad = E\left(\left\{ E\left[f_{hj}(\boldsymbol{\Delta}_m^{(i)}, \boldsymbol{Z}_{\text{pseudo}}(i)) \,\middle|\, \boldsymbol{Z}_{\text{pseudo}}(i) \right] \right\}^2 \right) - \bar{f}_{hj}^2 , \qquad (4.7)
\end{aligned}$$

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

4.3 Comparison of SP-Based Approach with Finite-Difference-Based Approach

One issue related to the analysis above is whether other methods for Hessian estimation could be effectively used instead of the simultaneous perturbation method. It is clearly not possible to answer this question for all possible existing or future methods for Hessian estimation, but it is possible to carry out some analysis relative to the standard finite-difference (FD)-based method. The FD-based method is identical to the SP-based approach of Section 3 of the paper with the exception of using classical FD techniques for Hessian estimation; there is no need to consider $M > 1$ because all Hessian estimates along a realization at a given $\boldsymbol{Z}_{\text{pseudo}}(i)$ would be identical. For the analysis below, suppose that the variance of the hjth element of the deviation matrix, $\boldsymbol{F}(\boldsymbol{\theta}) - (-\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})) = \boldsymbol{F}(\boldsymbol{\theta}) + \boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ (the difference between the information matrix and the negative Hessian), is σ_{hj}^2. This analysis below represents a summary of the technical report, Spall (2005), available from the author upon request.

If direct gradient values $\partial \log \ell / \partial \boldsymbol{\theta}$ are available, the standard two-sided FD approximation requires two gradient values for each column of the Hessian; in contrast, the SP-based approximation uses two gradient values for the full matrix. In the case where only $\log \ell(\cdot)$ values are available, then the FD-based method uses $O(p^2)$ values in constructing one Hessian

estimate (a specific standard form based on double-differencing uses a total of $2p(p+1)$ function values to approximate the $p(p+1)/2$ unique entries in the symmetric Hessian matrix; this contrasts with a total of four function values in the SP-based method). Both of the FD- and SP-based Hessian estimates will be biased to within an $O(c^2)$ error, where c is the width of the difference intervals or the maximum magnitude of the $\Delta_{k\ell}^{(i)}$ perturbations. Because c can be chosen arbitrarily small, we ignore this bias in the analysis below.

First, suppose gradient values $\partial \log \ell / \partial \boldsymbol{\theta}$ are available. From Spall (2000 and 2003, p. 199), it is fairly straightforward to show that the variance of an arbitrary element of an individual SP-based Hessian estimate is $O(p)$ in general and $O(1)$ in the special case where only $O(p)$ of the elements in $\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ are non-zero (e.g., $\boldsymbol{H}$ is diagonal). In the standard case of $2p$ gradient values for each FD approximation, we know that an SP-based estimate of $\boldsymbol{F}(\boldsymbol{\theta})$ with $M = 1$ and $N = N'p$ uses the same number of $\partial \log \ell / \partial \boldsymbol{\theta}$ values as the FD-based estimate. Hence, when the same number of $\partial \log \ell / \partial \boldsymbol{\theta}$ values are used in both estimates, the ratio of variance for an arbitrary SP-based element over the corresponding variance for FD-based element is $O(1)$ in the general Hessian case or $O(1/p)$ in the special case where only $O(p)$ of the elements in $\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}_{\text{pseudo}})$ are non-zero. Similar analysis applies when only log-likelihood values (no gradients) are available, likewise leading to an $O(1)$ ratio in the general Hessian case or $O(1/p)$ in the special case.

Analysis of the $O(1)$ ratio of variances in the general Hessian case shows that the SP-based variance will be lower than the FD-based variance when σ_{hj}^2 is relatively large, M is small, and p is large. Stronger results apply in the $O(1/p)$ special case; these results indicate that the SP-based variance is guaranteed to be lower than the FD-based variance for *any* $\sigma_{hj}^2 > 0$ when M is small and p is sufficiently large. Further, because each σ_{hj}^2 reflects the difference between the hjth element of an estimate ($-\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{Z}^{(n)})$) and truth ($\boldsymbol{F}(\boldsymbol{\theta})$), it is conjectured that the "typical" σ_{hj}^2 will grow with increasing dimension. To the extent that this conjecture is true, the efficiency of the SP-based method *relative* to the FD-based method becomes greater. That is, the SP-based variance is guaranteed to be lower than the FD-based variance when M is small and p is sufficiently large for the general Hessian case.

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). Let us sketch below how ARNs may be used towards reducing the variance of the information matrix estimate when $\boldsymbol{g}(\cdot)$ values are directly available.

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.).

To emphasize that we are considering *dependent* random perturbation vectors for $k = 1$ and 2 and to simplify the subscript and superscript notation below, let us use the notation $\boldsymbol{r}$ and $\boldsymbol{s}$

to denote the two successive perturbation vectors and suppress the pseudodata index i in most of the discussion below (i.e., $\boldsymbol{r}$ is analogous to $\boldsymbol{\Delta}_1^{(i)}$ and $\boldsymbol{s}$ is analogous to $\boldsymbol{\Delta}_2^{(i)}$ at a given i). Let the hjth component of $\boldsymbol{H}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ be given by H_{hj} (recall $H_{hj} = H_{jh}$). Then, by Spall (2000 and 2003, p. 199), the hjth component of the estimate $\hat{\boldsymbol{H}}_1 = \hat{\boldsymbol{H}}_1(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}})$ when using direct gradient evaluations is

$$\hat{H}_{1;hj} = H_{hj} + \tfrac{1}{2}\sum_{\ell \neq j} H_{h\ell}\frac{r_\ell}{r_j} + \tfrac{1}{2}\sum_{\ell \neq h} H_{\ell j}\frac{r_\ell}{r_h} + O_{\boldsymbol{Z},\boldsymbol{r}}(c^2), \tag{5.1}$$

where the pseudodata argument (and index i) has been suppressed in the Hessian terms and (analogous to Section 4) $O_{\boldsymbol{Z},\boldsymbol{r}}(c^2)$ denotes a random term dependent on $\boldsymbol{Z}_{\text{pseudo}}$ and $\boldsymbol{r}$. The obvious analogue holds for $k = 2$ (i.e., for the element $\hat{H}_{2;hj}$) with elements of $\boldsymbol{s}$ replacing elements of $\boldsymbol{r}$. Hence, from (5.1), the average of the two elements needed in forming the hj element of $\bar{\boldsymbol{H}} = \bar{\boldsymbol{H}}^{(i)}$ (step 2 of the algorithm in Section 3) is

$$\bar{H}_{hj} = \frac{\hat{H}_{1;hj} + \hat{H}_{2;hj}}{2} = H_{hj} + \tfrac{1}{2}\sum_{\ell \neq j} H_{h\ell}\left(\frac{r_\ell}{r_j} + \frac{s_\ell}{s_j}\right) + \tfrac{1}{2}\sum_{\ell \neq h} H_{\ell j}\left(\frac{r_\ell}{r_h} + \frac{s_\ell}{s_h}\right) + O_{\boldsymbol{Z},\boldsymbol{r},\boldsymbol{s}}(c^2) \tag{5.2}$$

(suppressing the pseudodata argument, once again).

Given that the $O_{\boldsymbol{Z},\boldsymbol{r},\boldsymbol{s}}(c^2)$ term is negligible (recall that c is "small"), it is apparent from (5.2) that the variance of $\bar{H}_{hj}$ is driven by the middle two summation terms. In particular, using the fact that $\boldsymbol{r}$ and $\boldsymbol{s}$ have the same moment and inverse moment distributional properties as $\boldsymbol{\Delta}_k$, the arguments in Spall (2000, p. 1851) show that if $\log \ell(\cdot)$ has bounded third derivatives in the vicinity of $\boldsymbol{\theta}$, then $E\left(\bar{H}_{hj} \middle| \boldsymbol{Z}_{\text{pseudo}}\right) = H_{hj}(\boldsymbol{\theta} \mid \boldsymbol{Z}_{\text{pseudo}}) + O_{\boldsymbol{Z}}(c^2)$ (i.e., the dominated convergence theorem applies to the $\boldsymbol{r}$ and $\boldsymbol{s}$ contributions in the $O_{\boldsymbol{Z},\boldsymbol{r},\boldsymbol{s}}(c^2)$ term); the $O_{\boldsymbol{Z}}(c^2)$ error also holds for second moments of $\bar{H}_{hj}$. Hence,

$$\operatorname{var}\left(\bar{H}_{hj} \middle| \boldsymbol{Z}_{\text{pseudo}}\right) = \tfrac{1}{4}\operatorname{var}\left[\sum_{\ell \neq j} H_{h\ell}\left(\frac{r_\ell}{r_j} + \frac{s_\ell}{s_j}\right) + \sum_{\ell \neq h} H_{\ell j}\left(\frac{r_\ell}{r_h} + \frac{s_\ell}{s_h}\right) \middle| \boldsymbol{Z}_{\text{pseudo}}\right] + O_{\boldsymbol{Z}}(c^2). \tag{5.3}$$

Unfortunately, it is generally impossible to make the non-$O_{\boldsymbol{Z}}(c^2)$ expression on the right-hand side of (5.3) small for all hj. One reason is that the $H_{h\ell}$ terms are usually unknown (that is one of

the reasons for use of the Monte Carlo scheme!). Another reason is that a choice of $\boldsymbol{r}$ and $\boldsymbol{s}$ that makes $\mathrm{var}(\bar{H}_{hj}|\boldsymbol{Z}_{\text{pseudo}})$ small for one combination of hj may have a contrasting effect for another hj. For these reasons, some of the "art" associated with practical implementation of ARNs must be applied.

For motivation, note that in one special case ARNs provide near-perfect variance reduction (with only an inherent order c^2 bias remaining). In particular, consider $p = 2$. If $s_1 = -r_1$ and $s_2 = r_2$, then

$$\begin{aligned}
\mathrm{var}(\bar{H}_{11}|\boldsymbol{Z}_{\text{pseudo}}) &= \tfrac{1}{4}\mathrm{var}\left[\sum_{\ell\neq 1} H_{1\ell}\left(\frac{r_\ell}{r_1}+\frac{s_\ell}{s_1}\right)+\sum_{\ell\neq 1} H_{\ell 1}\left(\frac{r_\ell}{r_1}+\frac{s_\ell}{s_1}\right)\middle|\boldsymbol{Z}_{\text{pseudo}}\right] + O_{\boldsymbol{Z}}(c^2) \\
&= \tfrac{1}{2}\mathrm{var}\left[H_{12}\left(\frac{r_2}{r_1}-\frac{r_2}{r_1}\right)\middle|\boldsymbol{Z}_{\text{pseudo}}\right] + O_{\boldsymbol{Z}}(c^2) \\
&= O_{\boldsymbol{Z}}(c^2),
\end{aligned}$$

$$\mathrm{var}(\bar{H}_{22}|\boldsymbol{Z}_{\text{pseudo}}) = \tfrac{1}{2}\mathrm{var}\left[H_{21}\left(\frac{r_1}{r_2}-\frac{r_1}{r_2}\right)\middle|\boldsymbol{Z}_{\text{pseudo}}\right] + O_{\boldsymbol{Z}}(c^2) = O_{\boldsymbol{Z}}(c^2),$$

where the calculation for $\mathrm{var}(\bar{H}_{22}|\boldsymbol{Z}_{\text{pseudo}})$ follows in a manner analogous to the calculation of $\mathrm{var}(\bar{H}_{11}|\boldsymbol{Z}_{\text{pseudo}})$, and

$$\begin{aligned}
\mathrm{var}(\bar{H}_{12}|\boldsymbol{Z}_{\text{pseudo}}) &= \tfrac{1}{4}\mathrm{var}\left[\sum_{\ell\neq 2} H_{1\ell}\left(\frac{r_\ell}{r_2}+\frac{s_\ell}{s_2}\right)+\sum_{\ell\neq 1} H_{\ell 2}\left(\frac{r_\ell}{r_1}+\frac{s_\ell}{s_1}\right)\middle|\boldsymbol{Z}_{\text{pseudo}}\right] + O_{\boldsymbol{Z}}(c^2) \\
&= \tfrac{1}{4}\mathrm{var}\left[H_{11}\left(\frac{r_1}{r_2}-\frac{r_1}{r_2}\right)+H_{22}\left(\frac{r_2}{r_1}-\frac{r_2}{r_1}\right)\middle|\boldsymbol{Z}_{\text{pseudo}}\right] + O_{\boldsymbol{Z}}(c^2) \\
&= O_{\boldsymbol{Z}}(c^2).
\end{aligned}$$

Hence, from the above, one can construct "perfect" (to within $O_{\boldsymbol{Z}}(c^2)$) estimates of $\boldsymbol{H}(\theta|\boldsymbol{Z}_{\text{pseudo}})$ in the $p = 2$ case through use of ARNs. This result is consistent with the standard finite-difference method of estimating a Hessian matrix to within $O(c^2)$ (c governing the width of the difference interval in a deterministic method) by $2p$ gradient measurements (two for each column in the Hessian). For $p = 2$, both the ARN and deterministic methods take four gradient measurements. Of course, the primary advantage of SP-based methods arises with larger p, where ARNs provide the possibility of variance reduction in Hessian estimates taking far less than the standard $2p$ gradient approximations.

In the $p \geq 3$ case, the situation is not as easy or clean as the above for the reasons discussed below (5.3). However, variance reduction is possible under some conditions. Let us illustrate the approach when one is most interested in the accuracy of the diagonal elements of the information matrix and when it is known that the off-diagonal elements of the Hessian matrices have approximately similar (although unknown) magnitudes for varying $\boldsymbol{Z}_{\text{pseudo}}$. Let this unknown magnitude be $\bar{H}$ (i.e., $\bar{H} \approx |H_{j\ell}|$ for all $j \neq \ell$). This latter assumption is one of the ways to avoid having to know the values of the $H_{j\ell}$ terms in practice. The general reasoning in the sketch below may be followed if there is interest in other aspects of the information matrix and/or there are other assumptions on the $H_{j\ell}$ terms. From (5.3),

$$\begin{aligned}
\text{var}\left(\bar{H}_{jj} \,\middle|\, \boldsymbol{Z}_{\text{pseudo}}\right) &= \tfrac{1}{2}\text{var}\left[\sum_{\ell \neq j} H_{j\ell}\left(\frac{r_\ell}{r_j} + \frac{s_\ell}{s_j}\right)\middle|\, \boldsymbol{Z}_{\text{pseudo}}\right] + O_{\boldsymbol{Z}}(c^2) \\
&\approx \frac{\bar{H}^2}{2}\text{var}\left[\sum_{\ell \neq j}\left(\frac{r_\ell}{r_j} + \frac{s_\ell}{s_j}\right)\right] + O_{\boldsymbol{Z}}(c^2) \\
&= \frac{\bar{H}^2}{2}\sum_{\ell \neq j}\text{var}\left(\frac{r_\ell}{r_j} + \frac{s_\ell}{s_j}\right) + O_{\boldsymbol{Z}}(c^2), \qquad (5.4)
\end{aligned}$$

where the second line of (5.4) follows by the independence of $\boldsymbol{r}$ and $\boldsymbol{s}$ from $\boldsymbol{Z}_{\text{pseudo}}$ and the last line follows by the uncorrelatedness of the summands in the second line (the $\approx$ in the second line follows by $\bar{H} \approx |H_{j\ell}|$ for all $j \neq \ell$).

Let us consider the use of ARNs to minimize the sum of the variances of all or some of the diagonal elements to within the $O_{\boldsymbol{Z}}(c^2)$ error (i.e., minimize $\sum_j \text{var}\left(\bar{H}_{jj} \,\middle|\, \boldsymbol{Z}_{\text{pseudo}}\right)$, where the sum is over p or fewer elements). Let $\{1, 2, \ldots, q\}$ for $q \leq p$ represent the set of indices for the diagonal elements of interest. That is, without loss of generality, the relevant indices are the first q. If this is not the case, then the elements $\boldsymbol{\theta}$ should be reordered so that the first q indices correspond to the elements for which ARNs will be applied. Hence, given a perturbation distribution for the components of $\boldsymbol{r}$ (e.g., i.i.d. Bernoulli), we aim from (5.4) to pick $\boldsymbol{s}$ such that

$$\sum_{j=1}^{q}\sum_{\ell \neq j}\text{var}\left(\frac{r_\ell}{r_j} + \frac{s_\ell}{s_j}\right) \qquad (5.5)$$

is minimized. (Alternatively, a more general functional optimization problem may be posed where the distributions of both ***r*** and ***s*** are simultaneously chosen to minimize the expression above subject to meeting the basic requirements discussed below (3.1); it is, however, unclear how this problem would be solved in practice.)

One means of creating a solvable parametric optimization problem for the general case is to build on the pattern suggested by the $p = 2$ setting above. In particular, it is apparent that each of the summands in (5.5) has one of four possible forms: odd numerator/odd denominator, odd/even, even/odd, and even/even, where "odd" or "even" refers to the subscript of the numerator or denominator terms. Hence, for example, at $\ell = 6$ and $j = 3$ in (5.5), we have an even/odd contribution. Given the value of ***r***, the even-indexed elements of ***s*** may be determined according to $s_j = \gamma_{\text{even}} r_j + (1-\gamma_{\text{even}})\delta_j$, where δ_j is an independent random variable having the same distribution as r_j and $0 \le \gamma_{\text{even}} \le 1$. Analogously, for the odd-indexed elements of ***s***, we have $s_j = -\gamma_{\text{odd}} r_j + (1-\gamma_{\text{odd}})\delta_j$, where $0 \le \gamma_{\text{odd}} \le 1$. So, each of the even-indexed elements of ***s*** is a convex combination of an independent random variable and the corresponding element of ***r***; each of the odd-indexed elements is a convex combination of an independent random variable and the *negative* of the corresponding element of ***r***. (This division between odd and even elements is arbitrary and could equivalently be reversed.) There is now enough structure to formulate a two-variable optimization problem from (5.5) (i.e., optimize γ_{even} and γ_{odd}).

Suppose that the δ_j and the elements of ***r*** are i.i.d. Bernoulli $\pm c$. It is then straightforward to determine the four possible variance expressions appearing in (5.5). Because $E\left(r_\ell / r_j\right) = E\left(s_\ell / s_j\right) = 0$ in (5.5), the variance terms follow according to the formula

$$\operatorname{var}\left(\frac{r_\ell}{r_j} + \frac{s_\ell}{s_j}\right) = E\left(\frac{r_\ell^2}{r_j^2} + 2\frac{r_\ell}{r_j}\frac{s_\ell}{s_j} + \frac{s_\ell^2}{s_j^2}\right). \tag{5.6}$$

Following some algebra, we have from (5.6) the following four possible expressions in (5.5) for $\operatorname{var}\left(r_\ell / r_j + s_\ell / s_j\right)$:

Odd ℓ, odd j

$$1+2\frac{\gamma_{\text{odd}}^2}{2\gamma_{\text{odd}}-1}+\frac{\left[\gamma_{\text{odd}}^2+(1-\gamma_{\text{odd}})^2\right]^2}{\left[\gamma_{\text{odd}}^2-(1-\gamma_{\text{odd}})^2\right]^2}. \tag{5.7a}$$

Odd ℓ, even j

$$1-2\frac{\gamma_{\text{odd}}\gamma_{\text{even}}}{2\gamma_{\text{even}}-1}+\frac{\left[\gamma_{\text{odd}}^2+(1-\gamma_{\text{odd}})^2\right]\left[\gamma_{\text{even}}^2+(1-\gamma_{\text{even}})^2\right]}{\left[\gamma_{\text{even}}^2-(1-\gamma_{\text{even}})^2\right]^2}. \tag{5.7b}$$

Even ℓ, odd j

$$1-2\frac{\gamma_{\text{odd}}\gamma_{\text{even}}}{2\gamma_{\text{odd}}-1}+\frac{\left[\gamma_{\text{odd}}^2+(1-\gamma_{\text{odd}})^2\right]\left[\gamma_{\text{even}}^2+(1-\gamma_{\text{even}})^2\right]}{\left[\gamma_{\text{odd}}^2-(1-\gamma_{\text{odd}})^2\right]^2}. \tag{5.7c}$$

Even ℓ, even j

$$1+2\frac{\gamma_{\text{even}}^2}{2\gamma_{\text{even}}-1}+\frac{\left[\gamma_{\text{even}}^2+(1-\gamma_{\text{even}})^2\right]^2}{\left[\gamma_{\text{even}}^2-(1-\gamma_{\text{even}})^2\right]^2}. \tag{5.7d}$$

Hence, the criterion in (5.5) is minimized by choosing γ_{even} and γ_{odd} such that the appropriately weighted linear combination of terms in (5.7a, b, c, d) is minimized. The weighting is based on the value of q. For example, at $q = 4$, we have from (5.5) the weightings odd/odd: 1/6; odd/even: 1/3; even/odd: 1/3, and even/even: 1/6. The author uses a simple MATLAB code to carry out the optimization.

While the above discussion of ARNs is for a special case, it is clear that basic ideas may be used in other cases (e.g., where certain off-diagonal elements of the Hessian have magnitudes that are approximately a known factor times larger than other elements and/or where the prime interest is in improving accuracy for certain off-diagonal elements of the information matrix). Nevertheless, it is inevitable that any practical application of ARNs will involve some specialized treatment, as illustrated above and in the simulation literature.

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 SP gradient estimate in (3.2)). The second aspect pertains to the value of gradient information (when available) relative to using only log-likelihood values. Table 2 considers the third aspect, illustrating 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 in both Tables 1 and 2; a third method is used in Table 1 alone. The first two methods are based on the maximum eigenvalue and 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 $\|\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta}) - \boldsymbol{F}_n(\boldsymbol{\theta})\|/\|\boldsymbol{F}_n(\boldsymbol{\theta})\|$ (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).

The third way we compare the solutions—as shown in Table 1—is via a simulated chi-squared test statistic $\boldsymbol{x}^T\boldsymbol{F}\boldsymbol{x}$, where $\boldsymbol{F}$ represents either $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$ or $\boldsymbol{F}_n(\boldsymbol{\theta})$, as appropriate. Such test statistics are standard in multivariate problems where $\boldsymbol{x}$ represents the difference between an estimated quantity and some nominal mean value of the quantity (i.e., estimated $\boldsymbol{\theta}$ – nominal $\boldsymbol{\theta}$) and $\boldsymbol{F}$ represents the inverse of the covariance matrix for $\boldsymbol{x}$. The points $\boldsymbol{x}$ such that $\boldsymbol{x}^T\boldsymbol{F}\boldsymbol{x} \le$ constant define a p-dimensional confidence ellipse centered about $\mathbf{0}$. The values "Test statistic" in Table 1 represent the sample mean of 50 values of the normalized deviation

$$\frac{\sum_{i=1}^{20} \left| \boldsymbol{x}_i^T \left[\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta}) - \boldsymbol{F}_n(\boldsymbol{\theta}) \right] \boldsymbol{x}_i \right|}{\sum_{i=1}^{20} \boldsymbol{x}_i^T \boldsymbol{F}_n(\boldsymbol{\theta}) \boldsymbol{x}_i} \tag{6.1}$$

for a set of $\boldsymbol{x}_i$ generated according to a $N(\mathbf{0}, \boldsymbol{I})$ distribution. The same 20 values of $\boldsymbol{x}_i$ are used in all runs entering the sample means of Table 1.

Table 1 shows that there is statistical evidence consistent with Proposition 2. All statistical comparisons are based on 50 independent calculations of $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$. In the comparisons of $\bar{\boldsymbol{F}}_{1,40000}$ with $\bar{\boldsymbol{F}}_{20,2000}$ (column (a) versus (b)), the P-values (probability values) computed from a standard matched-pairs t-test are 0.002, 0.0009, and 0.0106 for the maximum eigenvalue, norm, and test statistic comparisons, respectively. 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. (Note that computer run times for $\bar{\boldsymbol{F}}_{1,40000}$ are about 15 percent greater than for $\bar{\boldsymbol{F}}_{20,2000}$, 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 all three 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, 1.8 to 5.3 percent for the norm, and 0.2 to 1.3 percent for the test statistic.

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, mean spectral norm of difference as a fraction of true values, and mean absolute deviation of chi-squared test statistics as given in (6.1) (columns (a), (b), and (c)). Budget of SP Hessian estimates is constant ($B = MN$). P-values based on two-sided t-test using 50 independent runs.

	$M = 1$ $N = 40{,}000$ 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	$<10^{-10}$
Test statistic	0.0097	0.0128	0.0021	0.0106	7.9×10^{-9}

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 μ vector; the remaining element of μ and all elements of Σ 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 the discussion of (5.4)). 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 μ parameters, this study is restricted to the eigenvalues and norms of only the μ 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 θ. Unsurprisingly, there is no statistical evidence of improved estimates for the Σ part of the information matrix. Of course, different implementations on this problem (i.e., to include some or all components of Σ 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 μ block of $\boldsymbol{F}_n(\theta)$ (n = 30) as a fraction of true value and mean spectral norm on μ 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

7. CONCLUDING REMARKS

In many realistic processes, analytical evaluation of the Fisher 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 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.

There are several issues remaining that would enhance the applicability of the approach. In practice, there may be instances when some blocks of $\boldsymbol{F}_n(\boldsymbol{\theta})$ are known while other blocks are unknown. In the author's work related to parameter estimation for state-space models, for example, certain blocks along the diagonal are sometimes known, while other off-diagonal blocks are unknown (and need to be estimated). The issue yet to be examined is whether there is a way of focusing the averaging process on the blocks of interest that is more effective than simply extracting the estimate for those blocks from the full estimate of the matrix. Another issue pertains to the choice of distribution for the elements of the perturbation vector ($\boldsymbol{\Delta}_k$). While Bernoulli is used in the numerical examples, other distributions meet the regularity conditions and may be more effective in certain instances. When accounting for the cost of pseudodata generation, the optimal choice of averaging (M and N) is likely to be highly problem dependent, but it would be useful to have some general method for determining the tradeoff (the optimal $M = 1$ solution in Subsection 4.2 ignores the cost of pseudodata generation). It would also be useful to formally analyze the conjecture in Subsection 4.3 pertaining to the potential dependence of σ_{hj}^2 on p (reflecting the hjth element of the difference between the information matrix and the negative Hessian). Recall that the conjecture is that the σ_{hj}^2, on average, will tend to *increase* with p subject to the underlying number of data points being constant. To the extent that this conjecture is true, the efficiency of the simultaneous perturbation-based method *relative* to the standard finite difference-based method becomes greater. Finally, although the use of antithetic random numbers was described in this paper, more work could be done to make the concept more readily applicable through the use of appropriate approximations to the infeasible optimal perturbation distributions. Nevertheless, despite the open issues above, the method as currently available provides a relatively easy Monte Carlo method for determining the information matrix in general problems.

REFERENCES

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

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.

Efron, B. and Tibshirini, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy" (with discussion), *Statistical Science*, vol. 1, pp. 54–77.

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

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.

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

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

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.

Lunneborg, C. E. (2000), *Data Analysis by Resampling: Concepts and Applications*, Duxbury Press, Pacific Grove, CA.

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

Rao, C. R. (1973), *Linear Statistical Inference and its Applications* (2nd ed.), Wiley, New York.

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.

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.

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

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

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

Spall, J. C. (2005), "On the Comparative Performance of Finite Difference and Simultaneous Perturbation Methods for Estimation of the Fisher Information Matrix," JHU/APL Technical Report PSA-05-006 (10 February 2005).

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.

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