2008 American Control Conference
Westin Seattle Hotel, Seattle, Washington, USA
June 11-13, 2008

Improved Methods for Monte Carlo Estimation of the Fisher Information Matrix

James C. Spall (james.spall@jhuapl.edu)

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

Summary: The Fisher information matrix summarizes the amount of information in a set of data relative to the quantities of interest and forms the basis for the Cramér-Rao (lower) bound on the uncertainty in an estimate. There are many applications of the information matrix in modeling, systems analysis, and estimation. This paper presents a resampling-based method for computing the information matrix together with some new theory related to efficient implementation. We show how certain properties associated with the likelihood function and the error in the estimates of the Hessian matrix can be exploited to improve the accuracy of the Monte Carlo-based estimate of the information matrix.

Key words: System identification; Monte Carlo simulation; Cramér-Rao bound; simultaneous perturbation (SPSA); likelihood function.

1. Problem Setting

The Fisher information matrix has long played an important role in parameter estimation and system identification (e.g., Ljung, 1999, pp. 215–218). In many practical problems, however, the information matrix is difficult or impossible to obtain analytically. This includes nonlinear and/or non-Gaussian models as well as some linear problems (e.g., Segal and Weinstein, 1988; Levy, 1995).

In previous work (Spall, 2005), the author 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 either the Hessian matrix of the log-likelihood function or the outer product of the gradient of the log-likelihood function. The Monte Carlo approach can work with either evaluations of the log-likelihood function or evaluations of the gradient of the log-likelihood function, 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. An extension to this basic Monte Carlo method is given in Das et al. (2007), where it is shown that prior knowledge of *some* of the elements in in the information matrix can lead to improved estimates for *all* elements.

This paper introduces two simple modifications to the approach of Spall (2005) that improve the accuracy of the Monte Carlo estimate. These modifications may be implemented with little more difficulty than the original approach. These modifications may be used with or without the approach of Das et al. (2007) for handling prior information.

2. The Fisher Information Matrix and Associated Approximations

Consider a collection of n random vectors $\boldsymbol{Z} \equiv [z_1, z_2, \ldots, z_n]^T$; these vectors are not necessarily i.i.d. 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 $\boldsymbol{Z}$ is known, but that this function depends on an unknown vector $\boldsymbol{\theta}$. Let the probability density/mass function for $\boldsymbol{Z}$ be $p_{\boldsymbol{Z}}(\boldsymbol{\zeta}|\boldsymbol{\theta})$ where $\boldsymbol{\zeta}$ ("zeta") is a dummy matrix representing the possible outcomes for the elements in $\boldsymbol{Z}$. 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 $\boldsymbol{Z}$. Let $L(\boldsymbol{\theta}) = -\log \ell(\boldsymbol{\theta}|\boldsymbol{Z})$ (so we are suppressing the data dependence in L).

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

$$\boldsymbol{F}(\boldsymbol{\theta}) \equiv E\left(\frac{\partial L}{\partial \boldsymbol{\theta}} \frac{\partial L}{\partial \boldsymbol{\theta}^T} \right). \tag{2.2}$$

In the case where the underlying data $\{z_1, z_2, \ldots, z_n\}$ are independent, the magnitude of $\boldsymbol{F}(\boldsymbol{\theta})$ will grow at a rate proportional to n since L will represent a sum of n random terms. The bounded quantity $\boldsymbol{F}(\boldsymbol{\theta})/n$ is employed as an average information matrix over all measurements. Note also that when the data depend on some analyst-specified inputs , then $\boldsymbol{F}(\boldsymbol{\theta})$ also depends on these inputs. For notational convenience—and since many applications depend on cases (such as i.i.d. data) where there are no

inputs—we suppress this dependence and write $\boldsymbol{F}(\boldsymbol{\theta})$ for the information matrix.

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 L is twice continuously differentiable in $\boldsymbol{\theta}$ (a.s. in $\boldsymbol{Z}$). That is, the Hessian matrix

$$\boldsymbol{H}(\boldsymbol{\theta}) \equiv \frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

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 $\{\boldsymbol{\zeta}\colon \ell(\boldsymbol{\theta}|\boldsymbol{\zeta}) > 0\}$ 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 L through:

$$\boldsymbol{F}(\boldsymbol{\theta}) \equiv E\left(\frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) \tag{2.3}$$

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 $\boldsymbol{Z}$ (i.e., $\boldsymbol{H}(\boldsymbol{\theta})$) may be easier to compute and/or preferred from an inference point of view relative to the actual information matrix $\boldsymbol{F}(\boldsymbol{\theta})$ in (2.3) (e.g., Efron and Hinckley, 1978). Although the method in this paper is described for the determination of $\boldsymbol{F}(\boldsymbol{\theta})$, the efficient Hessian estimation described in Section 4 may also be used directly for the determination of $\boldsymbol{H}(\boldsymbol{\theta})$ when it is not easy to calculate the Hessian directly.

Expression (2.3) directly motivates a Monte Carlo simulation-based approach, as given in Spall (2005). Let $\boldsymbol{Z}_{\text{pseudo}}(i)$ be a Monte-Carlo generated random matrix from the assumed distribution for the actual data based on the parameters $\boldsymbol{\theta}$ taking on some specified value (typically an estimated value). Note that $\boldsymbol{Z}_{\text{pseudo}}(i)$ represents a sample of size n, analogous to the real data $\boldsymbol{Z}$, and that $\dim(\boldsymbol{Z}_{\text{pseudo}}(i)) = \dim(\boldsymbol{Z})$. Further, let $\hat{\boldsymbol{H}}_{k|i}$ represent the kth estimate of $\boldsymbol{H}(\boldsymbol{\theta})$ at the data vector $\boldsymbol{Z}_{\text{pseudo}}(i)$; $\hat{\boldsymbol{H}}_{k|i}$ is to be used in an averaging process as described below. As described below, the estimate $\hat{\boldsymbol{H}}_{k|i}$ is generated via efficient simultaneous perturbation (SPSA) principles (Spall, 1992) using either log-likelihood $L(\boldsymbol{\theta})$ values (alone) or the gradient (score vector) $\boldsymbol{g}(\boldsymbol{\theta}) \equiv \partial L(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ 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 $\boldsymbol{G}_{k|i}(\boldsymbol{\theta})$ represent either a gradient *approximation* (based on $L(\boldsymbol{\theta})$ values) or the exact gradient $\boldsymbol{g}(\boldsymbol{\theta})$, as used in the kth Hessian estimate at the given $\boldsymbol{Z}_{\text{pseudo}}(i)$.

Let $\boldsymbol{\Delta}_{k|i} \equiv [\Delta_{k1|i}, \Delta_{k2|i}, \ldots, \Delta_{kp|i}]^T$ be a mean-zero random vector such that the scalar elements $\{\Delta_{kj|i}\}$ are independent, identically distributed, symmetrically distributed random variables that are uniformly bounded and satisfy $E\left(\left|1/\Delta_{kj|i}\right|\right) < \infty$. The latter condition *excludes* such commonly used Monte Carlo distributions as uniform and Gaussian. Further, assume that the $\Delta_{kj|i}$ are bounded in magnitude. Note that the user has full control over the choice of the $\Delta_{kj|i}$ distribution. A valid (and simple) choice is the Bernoulli ± 1 distribution (it is not known at this time if this is the "best" distribution to choose for this application).

The formula for estimating the Hessian at the point $\boldsymbol{\theta}$ is:

$$\hat{\boldsymbol{H}}_{k|i} = \frac{1}{2}\left\{ \frac{\delta \boldsymbol{G}_{k|i}}{2c}(\boldsymbol{\Delta}_{k|i}^{-1})^T + \left(\frac{\delta \boldsymbol{G}_{k|i}}{2c}(\boldsymbol{\Delta}_{k|i}^{-1})^T \right)^T \right\}, \tag{2.4}$$

where $\delta \boldsymbol{G}_{k|i} \equiv \boldsymbol{G}_{k|i}(\boldsymbol{\theta} + c\boldsymbol{\Delta}_{k|i}) - \boldsymbol{G}_{k|i}(\boldsymbol{\theta} - c\boldsymbol{\Delta}_{k|i})$, $\boldsymbol{\Delta}_{k|i}^{-1}$ denotes the vector of inverses of the p individual elements of $\boldsymbol{\Delta}_{k|i}$, and $c > 0$ is a "small" constant. The prime rationale for (2.4) is that $\hat{\boldsymbol{H}}_{k|i}$ is a nearly unbiased estimator of the unknown $\boldsymbol{H}(\boldsymbol{\theta})$. Spall (2000) gives conditions such that the Hessian estimate has an $O(c^2)$ bias (the main such condition is smoothness of $L(\boldsymbol{\theta})$, as reflected in the assumption that $\boldsymbol{g}(\boldsymbol{\theta})$ is thrice continuously differentiable in $\boldsymbol{\theta}$ near the nominal value of interest).

The symmetrizing operation in (2.4) (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 $\hat{\boldsymbol{H}}_{k|i}$ in (2.4) 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 (2.4) 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 (i.e., the two values $\boldsymbol{G}_{k|i}(\boldsymbol{\theta} \pm c\boldsymbol{\Delta}_{k|i}) = \boldsymbol{g}(\boldsymbol{\theta} \pm c\boldsymbol{\Delta}_k)$ evaluated at $\boldsymbol{Z}_{\text{pseudo}}(i)$ are used to form the Hessian estimate). When only log-likelihood values are available, each of the gradient approximations $\boldsymbol{G}_{k|i}(\boldsymbol{\theta} \pm c\boldsymbol{\Delta}_{k|i})$ require two evaluations of $L(\cdot)$. Hence, one approximation $\hat{\boldsymbol{H}}_{k|i}$ uses four log-likelihood values. The gradient approximations at the two design levels are:

$$\boldsymbol{G}_{k|i}(\boldsymbol{\theta}\pm c\boldsymbol{\Delta}_k)=\frac{L(\boldsymbol{\theta}\pm c\boldsymbol{\Delta}_{k|i}+\tilde{c}\tilde{\boldsymbol{\Delta}}_{k|i})-L(\boldsymbol{\theta}\pm c\boldsymbol{\Delta}_{k|i}-\tilde{c}\tilde{\boldsymbol{\Delta}}_{k|i})}{2\tilde{c}}\tilde{\boldsymbol{\Delta}}_{k|i}^{-1} \tag{2.5}$$

where $\tilde{\boldsymbol{\Delta}}_{k|i} = [\tilde{\Delta}_{k1|i}, \tilde{\Delta}_{k2|i}, ..., \tilde{\Delta}_{kp|i}]^T$ is generated in the same statistical manner as $\boldsymbol{\Delta}_{k|i}$, but independently of $\boldsymbol{\Delta}_{k|i}$ (in particular, choosing $\tilde{\Delta}_{kj|i}$ as independent Bernoulli ± 1 random variables is a valid—but not necessary—choice), $\tilde{\boldsymbol{\Delta}}_{k|i}^{-1}$ denotes the vector of inverses of the p elements of $\tilde{\boldsymbol{\Delta}}_{k|i}$, and $\tilde{c} > 0$ (like c) is a small constant.

The Monte Carlo approach of Spall (2005) is based on a double averaging scheme. The first "inner" average forms Hessian estimates at a given $\boldsymbol{Z}_{\text{pseudo}}(i)$ (i = 1, 2,..., N) from k = 1, 2, ..., M values of $\hat{\boldsymbol{H}}_{k|i}$ and the second "outer" average combines these sample mean Hessian estimates across the N values of pseudo data. Therefore, the Monte Carlo-based estimate of $\boldsymbol{F}(\boldsymbol{\theta})$ in Spall (2005), denoted $\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta})$, is:

$$\bar{\boldsymbol{F}}_{M,N}(\boldsymbol{\theta}) \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{H}}_{k|i}\ , \tag{2.6}$$

or, in equivalent recursive (in i = 1, 2,…, N) form:

$$\bar{\boldsymbol{F}}_{M,i}(\boldsymbol{\theta}) = \frac{i-1}{i}\bar{\boldsymbol{F}}_{M,i-1}(\boldsymbol{\theta}) + \frac{1}{iM}\sum_{k=1}^{M}\hat{\boldsymbol{H}}_{k|i} \tag{2.7}$$

($\bar{\boldsymbol{F}}_{M,0} = \boldsymbol{0}$). The aim of this paper is to introduce two modifications to the "basic" form in (2.6) and (2.7).

3. Characterization of Error in Hessian Estimate

This section provides a few key facts about $\hat{\boldsymbol{H}}_{k|i}$, as used in the averaging process for the estimation of the information matrix (eqn. (2.6)). We will use these facts in Sections 4 and 5 to show how the accuracy can be improved relative to the basic averaging process in (2.6) and (2.7). The probabilistic big-O terms appearing below are to be interpreted in the almost surely (a.s.) sense (e.g., $O(c^2)$ implies a function that is a.s. bounded when divided by c^2, $c \to 0$); all associated equalities hold a.s.

One of the key ways in which the error in the Fisher estimate will be reduced is through the use of feedback. We now present an expression for the error in the Hessian estimates that is useful in creating the feedback term. From Spall (2006), it is known that $\hat{\boldsymbol{H}}_{k|i}$ in (2.4) can be decomposed into three parts:

$$\hat{\boldsymbol{H}}_{k|i} = \boldsymbol{H}(\boldsymbol{\theta}) + \boldsymbol{\Psi}_{k|i} + O(c^2)\,, \tag{3.1}$$

where $\boldsymbol{\Psi}_{k|i}$ is a $p \times p$ matrix of terms dependent on $\boldsymbol{H}(\boldsymbol{\theta})$, $\boldsymbol{\Delta}_{k|i}$, and, when only L values are available, the additional perturbation vector $\tilde{\boldsymbol{\Delta}}_{k|i}$ (see Spall, 2006). Note that $\boldsymbol{\Psi}_{k|i}$ represents the error due to the simultaneous perturbations ($\boldsymbol{\Delta}_{k|i}$ and, if relevant, $\tilde{\boldsymbol{\Delta}}_{k|i}$). The specific form and notation for the $\boldsymbol{\Psi}_{k|i}$ term depends on whether L ($\boldsymbol{\theta}$) values or $\boldsymbol{g}(\boldsymbol{\theta})$ values are available, represented as $\boldsymbol{\Psi}_{k|i} = \boldsymbol{\Psi}_{k|i}^{(L)}$ or $\boldsymbol{\Psi}_{k|i} = \boldsymbol{\Psi}_{k|i}^{(\boldsymbol{g})}$, respectively, in the notation below. Finally, the big-O error is a reflection of the bias in the Hessian estimate; in the case where only L values are available, the $O(c^2)$ bias assumes that $\tilde{c}/c$ is $O(1)$ ($c \to 0$).

Let us define

$$\boldsymbol{D}_{k|i} = \boldsymbol{\Delta}_{k|i}(\boldsymbol{\Delta}_{k|i}^{-1})^T - \boldsymbol{I}_p \text{ and } \tilde{\boldsymbol{D}}_{k|i} = \tilde{\boldsymbol{\Delta}}_{k|i}(\tilde{\boldsymbol{\Delta}}_{k|i}^{-1})^T - \boldsymbol{I}_p$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix (note that $\boldsymbol{D}_{k|i}$ and $\tilde{\boldsymbol{D}}_{k|i}$ are symmetric when the perturbations are i.i.d. Bernoulli distributed). Given $\boldsymbol{D}_{k|i}$ and $\tilde{\boldsymbol{D}}_{k|i}$ above, it is shown in Spall (2006) that with the $\boldsymbol{G}_{k|i}(\boldsymbol{\theta}\pm c\boldsymbol{\Delta}_{k|i})$ formed from L measurements only (see (2.5)):

$$\boldsymbol{\Psi}_{k|i}^{(L)}(\boldsymbol{H}) = \tfrac{1}{2}\Big[\tilde{\boldsymbol{D}}_{k|i}^T\boldsymbol{H}\boldsymbol{D}_{k|i} + \tilde{\boldsymbol{D}}_{k|i}^T\boldsymbol{H} + \boldsymbol{H}\boldsymbol{D}_{k|i}\Big] + \tfrac{1}{2}\Big[\tilde{\boldsymbol{D}}_{k|i}^T\boldsymbol{H}\boldsymbol{D}_{k|i} + \tilde{\boldsymbol{D}}_{k|i}^T\boldsymbol{H} + \boldsymbol{H}\boldsymbol{D}_{k|i}\Big]^T, \tag{3.2}$$

while with the $\boldsymbol{G}_{k|i}(\boldsymbol{\theta}\pm c\boldsymbol{\Delta}_{k|i})$ formed from direct $\boldsymbol{g}$ values:

$$\boldsymbol{\Psi}_{k|i}^{(\boldsymbol{g})}(\boldsymbol{H}) \equiv \tfrac{1}{2}\boldsymbol{H}\boldsymbol{D}_{k|i} + \tfrac{1}{2}\boldsymbol{D}_{k|i}^T\boldsymbol{H}\,. \tag{3.3}$$

4. Implementation with Independent Perturbation per Measurement

Let us assume in this section that the n data vectors entering each $\boldsymbol{Z}_{\text{pseudo}}(i)$ are mutually independent (analogous to the real data $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ being mutually independent). The "basic" structure in Section 2 can be improved by exploiting this independence. In particular, the variance of the elements of the individual Hessian estimates $\hat{\boldsymbol{H}}_{k|i}$ can be reduced by decomposing $\hat{\boldsymbol{H}}_{k|i}$ into a sum of n independent estimates, each corresponding to one of the data vectors. A separate perturbation vector can then be applied to each of the independent estimates, which produces variance reduction in the resulting estimate $\bar{\boldsymbol{F}}'_{M,N}$. the independent perturbations above reduce the variance of the elements in the estimate of $\boldsymbol{F}(\boldsymbol{\theta})$ from $O(1/N)$ in Spall (2005) to $O(1/(nN))$. The complete version of the paper available upon request includes these results.

5. Feedback-Based Method for $\boldsymbol{F}(\boldsymbol{\theta})$ Estimation

Aside from the independent perturbation idea of Section 4, variance reduction is possible by using the current estimate of $\boldsymbol{F}(\boldsymbol{\theta})$ to reduce the variance of the next Hessian estimate. This, in turn, reduces the variance of the next estimate for $\boldsymbol{F}(\boldsymbol{\theta})$. This feedback idea applies whether or not the independent perturbations of Section 4 are used. The idea here is in the spirit of Spall (2006), but the context is different in that the primary quantity of interest is $\boldsymbol{F}(\boldsymbol{\theta})$, not the Hessian matrix. In essence, the variance reduction below follows from the fundamental property that var(X) ≤ $E(X^2)$ for any random variable X having a finite second moment.

Let $\bar{\boldsymbol{F}}'_{M,N}$ denote the feedback-based form of this section. So $\bar{\boldsymbol{F}}'_{M,N}$ is a direct replacement for $\bar{\boldsymbol{F}}_{M,N}$ in Sections 2 – 4. Using the basic recursive form in (2.7) as the starting point, the feedback-based form for the estimate of the information matrix in recursive (in i) form is

$$\bar{\boldsymbol{F}}'_{M,i}(\boldsymbol{\theta}) = \frac{i-1}{i}\bar{\boldsymbol{F}}'_{M,i-1}(\boldsymbol{\theta}) + \frac{1}{iM}\sum_{k=1}^{M}\left[\hat{\boldsymbol{H}}_{k|i} - \boldsymbol{\Psi}_{k|i}\left(\bar{\boldsymbol{F}}'_{M,i-1}(\boldsymbol{\theta})\right)\right], \qquad (5.1)$$

where $\boldsymbol{\Psi}_{k|i} = \boldsymbol{\Psi}^{(L)}_{k|i}$ or $\boldsymbol{\Psi}_{k|i} = \boldsymbol{\Psi}^{(g)}_{k|i}$ from (3.2) or (3.3), as appropriate ($\bar{\boldsymbol{F}}'_{M,0} = \boldsymbol{0}$). Note that the current best estimate of $\boldsymbol{F}(\boldsymbol{\theta})$ (i.e., $\bar{\boldsymbol{F}}'_{M,i-1}$) is used in place of $\boldsymbol{H}$ when evaluating $\boldsymbol{\Psi}_{k|i}$).

The main result related to improved accuracy is Theorem 1 (and its Corollary 1) below. Because of the frequent need to refer to the specific $\boldsymbol{\theta}$ of interest, as well as expansions of functions around that point of interest, we let $\boldsymbol{\theta}^*$ represent the particular $\boldsymbol{\theta}$ at which we wish to determine $\boldsymbol{F}(\boldsymbol{\theta})$; also, for convenience, let $\boldsymbol{F}^* = \boldsymbol{F}(\boldsymbol{\theta}^*)$. As a first step towards analyzing the variance reduction due to feedback, we show in the Lemma below that $\bar{\boldsymbol{F}}'_{M,N} \to \boldsymbol{F}^* + O(c^2)$ in mean square as $N \to \infty$ (fixed M). The results below apply when $\hat{\boldsymbol{H}}_{k|i}$ is formed from either $L(\boldsymbol{\theta})$ values (alone) or from gradient (score) values $\boldsymbol{g}(\boldsymbol{\theta})$ and/or when the independent perturbation idea of Section 4 is used. Following the previously established notation for the third derivatives of L, let $L''''(\boldsymbol{\theta})$ denote the $1 \times p^4$ row vector of all possible fourth derivatives of L. Vector and matrix norms are the standard Euclidean norm; in the matrix case, this corresponds to the Frobenius norm: $\|\boldsymbol{A}\|^2 = \sum_i \sum_j a_{ij}^2$, where the a_{ij} are the components of $\boldsymbol{A}$.

Lemma. For some open neighborhood of $\boldsymbol{\theta}^*$, suppose $L''''(\boldsymbol{\theta})$ exists continuously (a.s. in $\boldsymbol{Z}$) and that $E\left(\left\|L''''(\boldsymbol{\theta})\right\|^2\right)$ is bounded in magnitude. Further, let $E\left(\left\|\hat{\boldsymbol{H}}_{k|i}\right\|^2\right) < \infty$ (recall that the $\hat{\boldsymbol{H}}_{k|i}$ are identically distributed for all k and i). Then, given the basic conditions on $\boldsymbol{\Delta}_{k|i}$ in Section 2 (applying also to $\tilde{\boldsymbol{\Delta}}_{k|i}$ when only L values are used for $\hat{\boldsymbol{H}}_{k|i}$), $E\left(\left\|\bar{\boldsymbol{F}}'_{M,N} - \boldsymbol{F}^* - \boldsymbol{B}(\boldsymbol{\theta}^*)\right\|^2\right) \to 0$ as $N \to \infty$ for any fixed $M \geq 1$ and all c sufficiently small, where $\boldsymbol{B}(\boldsymbol{\theta}^*)$ is a bias matrix satisfying $\boldsymbol{B}(\boldsymbol{\theta}^*) = O(c^2)$.

Proof. Due to space limitations, we do not include the proof of the Lemma here; the proof is available upon request. A similar proof is in Spall (2006).

We are now in a position to establish that feedback reduces the asymptotic mean-squared error of the estimate for the information matrix. In the proofs of Theorem 1 and Corollary 1, we use the mean-squared convergence result of the Lemma to establish the main result on improved accuracy. Note that Theorem 1 and Corollary 1 only consider the $p \geq 2$ case because at $p = 1$, $\bar{\boldsymbol{F}}'_{M,N} = \bar{\boldsymbol{F}}_{M,N}$ due to the errors in (3.2) and (3.3) being identically zero.

Theorem 1. Suppose that the conditions of the Lemma hold, $p \geq 2$, $E\left(\left\|\boldsymbol{H}(\boldsymbol{\theta}^*)\right\|^2\right) < \infty$, $\boldsymbol{F}^* \geq \boldsymbol{0}$, and $\boldsymbol{F}^* \neq \boldsymbol{0}$. Further, suppose that for some $\delta > 0$ and $\delta' > 0$ such that $(1+\delta)^{-1} + (1+\delta')^{-1} = 1$, $E\left(\left\|L''''(\boldsymbol{\theta})\right\|^{2+2\delta'}\right)$ is uniformly bounded in magnitude for all $\boldsymbol{\theta}$ in an open neighborhood of $\boldsymbol{\theta}^*$, $E\left(\left|1/\Delta_{kj|i}^{2+2\delta}\right|\right) < \infty$ and, when only L values are used, $E\left(\left|1/\tilde{\Delta}_{kj|i}^{2+2\delta}\right|\right) < \infty$ (arbitrary i, j, and k by the i.i.d. assumption for the perturbation elements). Then the accuracy of $\bar{\boldsymbol{F}}'_{M,N}$ is greater than the accuracy of $\bar{\boldsymbol{F}}_{M,N}$ in the sense that

$$\lim_{N\to\infty} \frac{E\left[\left\|\bar{\boldsymbol{F}}'_{M,N} - \boldsymbol{F}^*\right\|^2\right]}{E\left[\left\|\bar{\boldsymbol{F}}_{M,N} - \boldsymbol{F}^*\right\|^2\right]} \leq 1 + O(c^2). \qquad (5.2)$$

Proof. Let $\bar{f}_N$, $\bar{f}'_N$, and f^* denote corresponding (arbitrary) scalar elements of $\bar{\boldsymbol{F}}_{M,N}$ (no feedback estimate), $\bar{\boldsymbol{F}}'_{M,N}$, and $\boldsymbol{F}^*$, respectively. That is, $\bar{f}_N$ and $\bar{f}'_N$ are estimates for one of the $p(p+1)/2$ unique elements in the information matrix. Further, let ψ_i, $\psi_i^{\boldsymbol{H}}$, and $\psi_i^{\boldsymbol{F}}$ be the corresponding scalar elements of the error matrices $\boldsymbol{\Psi}_{1|i}\left(\bar{\boldsymbol{F}}'_{i-1}\right)$, $\boldsymbol{\Psi}_{1|i}\left(\boldsymbol{H}(\boldsymbol{\theta}^*)\right)$, and $\boldsymbol{\Psi}_{1|i}(\boldsymbol{F}^*)$ (corresponding to the component of the information matrix being estimated by $\bar{f}_N$ and $\bar{f}'_N$). Only $\boldsymbol{\Psi}_{1|i}\left(\bar{\boldsymbol{F}}'_{i-1}\right)$ is actually computed in the algorithm; however, the other two errors, $\boldsymbol{\Psi}_{1|i}\left(\boldsymbol{H}(\boldsymbol{\theta}^*)\right)$ and $\boldsymbol{\Psi}_{1|i}(\boldsymbol{F}^*)$, play a key role in the proof below (recall that $\boldsymbol{H}(\boldsymbol{\theta}^*)$, in general, depends on $\boldsymbol{Z}_{\text{pseudo}}(i)$). Also, let $\hat{h}_i$ be the scalar element of $\hat{\boldsymbol{H}}_{1|i}$ corresponding to the component of $\bar{\boldsymbol{F}}_{M,N}$ and $\bar{\boldsymbol{F}}'_{M,N}$ being estimated by $\bar{f}_N$ and $\bar{f}'_N$, respectively.

Because (5.2) is based on the Frobenius norm, it is sufficient to show that

$$\lim_{N\to\infty} E\left[\left(\bar{f}'_N - f^*\right)^2\right] \Big/ E\left[\left(\bar{f}_N - f^*\right)^2\right] \leq 1 + O(c^2)$$

for all of the $p(p+1)/2$ unique elements. As in the proof of the Lemma, without loss of generality, take $M = 1$. From (5.1),

$$\bar{f}'_N = \frac{1}{N}\sum_{i=1}^{N}(\hat{h}_i - \psi_i).$$

Then, from the fact that $E\left(\left\|\hat{\boldsymbol{H}}_{1|i}\right\|^2\right) < \infty$ (with $E\left(\left\|\hat{\boldsymbol{H}}_{1|i}\right\|^2\right) = E\left(\left\|\hat{\boldsymbol{H}}_{1|j}\right\|^2\right)$ for all i and j) and fact that $\bar{\boldsymbol{F}}'_{1,i}$ converges in

mean square to $\boldsymbol{F}^* + \boldsymbol{B}(\boldsymbol{\theta}^*)$ (Lemma), it is known that the mean-squared error $E\left[(\bar{f}'_N - f^*)^2\right]$ exists for all N. Thus, by the standard bias-variance decomposition of mean-squared error (e.g., Spall, 2003, p. 333),

$$E\left[(\bar{f}'_N - f^*)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{var}(\hat{h}_i - \psi_i) + O(c^4), \quad (5.3)$$

where the summation follows from the uncorrelatedness of the summands $\hat{h}_i - \psi_i$, which follows by the independence across i of the $\boldsymbol{\Delta}_{1|i}$ (and $\tilde{\boldsymbol{\Delta}}_{1|i}$ when the estimates are based only on measurements of L), and the $O(c^4)$ term follows from the fact that the elements of the bias matrix $\boldsymbol{B}(\boldsymbol{\theta}^*)$ are $O(c^2)$ (the Lemma). Analogously, in the no-feedback case (eqns. (2.6) and (2.7))

$$E\left[(\bar{f}_N - f^*)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{var}(\hat{h}_i) + O(c^4) \quad (5.4)$$

(the specific analytical form of the squared bias contribution $O(c^4)$ is identical in both the feedback and non-feedback cases).

From expressions (5.3) and (5.4) for the mean-squared errors, it is clear that the relative errors in the feedback and non-feedback cases are determined by analyzing the relative behavior of $\mathrm{var}(\hat{h}_i - \psi_i)$ and $\mathrm{var}(\hat{h}_i)$. It can then be shown that,

$$\mathrm{var}(\hat{h}_i - \psi_i) = \mathrm{var}(h_i + \psi_i^{\boldsymbol{H}} - \psi_i) + O(c^2), \quad (5.5)$$

$$\mathrm{var}(\hat{h}_i) = \mathrm{var}(h_i + \psi_i^{\boldsymbol{H}}) + O(c^2). \quad (5.6)$$

In the case of having direct $\boldsymbol{g}$ values, e_i is simpler: e_i is a numerator of sums of elements in $L'''(\boldsymbol{\theta})$ times terms within $\boldsymbol{\Delta}_{1|i}$ over a denominator of one term from $\boldsymbol{\Delta}_{1|i}$. The arguments above then follow analogously, leading to the relationships in (5.5) and (5.6).

Note that h_i and $\psi_i^{\boldsymbol{H}} - \psi_i$ are uncorrelated and h_i and ψ_i are uncorrelated (the uncorrelatedness at each i follows by the form for $\boldsymbol{\Psi}_{1|i}$ and the independence of the $\boldsymbol{\Delta}_{1|i}$, $\boldsymbol{Z}_{\mathrm{pseudo}}(i)$, and, when the estimates are based only on measurements of L, $\tilde{\boldsymbol{\Delta}}_{1|i}$). Then, the variance terms on the right-hand side of (5.5) and (5.6) satisfy

$$\mathrm{var}(h_i + \psi_i^{\boldsymbol{H}} - \psi_i) = \mathrm{var}(h_i) + \mathrm{var}(\psi_i^{\boldsymbol{H}} - \psi_i), \quad (5.7)$$

$$\mathrm{var}(h_i + \psi_i^{\boldsymbol{H}}) = \mathrm{var}(h_i) + \mathrm{var}(\psi_i^{\boldsymbol{H}}). \quad (5.8)$$

As a vehicle towards characterizing the relative values of $\mathrm{var}(\hat{h}_i - \psi_i)$ and $\mathrm{var}(\hat{h}_i)$ via (5.5) – (5.8), we now show that the second term on the right-hand side of (5.7) satisfies $\mathrm{var}(\psi_i^{\boldsymbol{H}} - \psi_i) - \mathrm{var}(\psi_i^{\boldsymbol{H}} - \psi_i^{\boldsymbol{F}}) \to O(c^2)$ as $i \to \infty$. Note that,

$$\mathrm{var}(\psi_i^{\boldsymbol{H}} - \psi_i^{\boldsymbol{F}}) = E\left[(\psi_i^{\boldsymbol{H}} - \psi_i^{\boldsymbol{F}})^2\right] = E\left[(\psi_i^{\boldsymbol{H}})^2 - (\psi_i^{\boldsymbol{F}})^2\right], \quad (5.9)$$

where the first equality follows from $E(\psi_i^{\boldsymbol{H}}) = E(\psi_i^{\boldsymbol{F}}) = 0$ and the second equality follows by the independence of the $\boldsymbol{\Delta}_{1|i}$, $\boldsymbol{Z}_{\mathrm{pseudo}}(i)$, and (when the estimates are based only on measurements of L) $\tilde{\boldsymbol{\Delta}}_{1|i}$ at each i. Hence, from the Lemma,

$$\lim_{i\to\infty}\left[\mathrm{var}(\psi_i^{\boldsymbol{H}} - \psi_i) - \mathrm{var}(\psi_i^{\boldsymbol{H}} - \psi_i^{\boldsymbol{F}})\right] = O(c^2), \quad (5.10)$$

as desired. Relative to the right-hand side of (5.8),

$$\mathrm{var}(\psi_i^{\boldsymbol{H}}) = E\left[(\psi_i^{\boldsymbol{H}})^2\right], \quad (5.11)$$

implying from (5.9) that $\mathrm{var}(\psi_i^{\boldsymbol{H}} - \psi_i^{\boldsymbol{F}}) \le \mathrm{var}(\psi_i^{\boldsymbol{H}})$ (the inequality is strict when $E\left[(\psi_i^{\boldsymbol{F}})^2\right] \ne 0$; see Corollary 1).

Therefore, by the principle of Cesàro summability (e.g., Apostol, 1974, Theorem 8.48), it is known from (5.5) – (5.11) that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{var}(\hat{h}_i) - \mathrm{var}(\hat{h}_i - \psi_i)\right]$$
$$= \mathrm{var}(\psi_1^{\boldsymbol{H}}) - \mathrm{var}(\psi_1^{\boldsymbol{H}} - \psi_1^{\boldsymbol{F}}) \ge 0 \quad (5.12)$$

(it is sufficient to consider only first values, $\psi_1^{\boldsymbol{H}}$ and $\psi_1^{\boldsymbol{F}}$, in the limiting variance due to the identical distribution of $\psi_i^{\boldsymbol{H}}$ and $\psi_i^{\boldsymbol{F}}$).

To establish the result to be proved, we need to ensure that the indicated limit in (5.2) exists. From (5.3), (5.4), and (5.12) (which show that the numerator and denominator have the same $O(1/N)$ convergence rate to unique limits), the limit in (5.2) exists if $\mathrm{var}(\hat{h}_1) > 0$ for at least one of the $p(p+1)/2$ elements being estimated (for all sufficiently small c). From (5.6) and (5.8), $\mathrm{var}(\hat{h}_1) = \mathrm{var}(h_1) + \mathrm{var}(\psi_1^{\boldsymbol{H}}) + O(c^2)$. In the case where $\boldsymbol{H}(\boldsymbol{\theta}^*)$ is deterministic, we know from the error decompositions in (3.2) and (3.3), that $\mathrm{var}(\psi_1^{\boldsymbol{H}}) > 0$ for at least one element because $\boldsymbol{H}(\boldsymbol{\theta}^*) = \boldsymbol{F}^*$ and it is assumed that $\boldsymbol{F}^* \ge \boldsymbol{0}$ and $\boldsymbol{F}^* \ne \boldsymbol{0}$. In the case where $\boldsymbol{H}(\boldsymbol{\theta}^*)$ is random (i.e., at least one element is non-degenerate random), $\mathrm{var}(h_1)$ for the non-degenerate elements exist (by the assumption $E\left(\left\|\boldsymbol{H}(\boldsymbol{\theta}^*)\right\|^2\right) < \infty$) and satisfy $\mathrm{var}(h_1) > 0$. Hence, by (5.6) and (5.8), $\mathrm{var}(\hat{h}_1) > 0$ for at least one element.

Finally, for any element such that $\mathrm{var}(\hat{h}_1) > 0$, we know by (5.3), (5.4), (5.5), (5.6), and (5.12) that

$$\lim_{N\to\infty}\frac{E\left[\left(\bar{f}'_N - f^*\right)^2\right]}{E\left[\left(\bar{f}_N - f^*\right)^2\right]} = \frac{\mathrm{var}(h_1) + \mathrm{var}(\psi_1^{\boldsymbol{H}} - \psi_1^{\boldsymbol{F}}) + O(c^2)}{\mathrm{var}(h_1) + \mathrm{var}(\psi_1^{\boldsymbol{H}}) + O(c^2)}$$
$$\le 1 + O(c^2). \quad (5.13)$$

The result to be proved in (5.2) then follows by the fact that the results in (5.5) – (5.12) show that any elements with $\text{var}(\hat{h}_1) = 0$ do not contribute to the norms in either the numerator or denominator. *Q.E.D.*

Corollary 1 establishes conditions for strict inequality in (5.2). The proof rests on the following: The solution matrix $\boldsymbol{X}$ to the equation $\boldsymbol{AX} + \boldsymbol{XB} = \boldsymbol{0}$ is unique ($\boldsymbol{X} = \boldsymbol{0}$) if and only if the square matrices $\boldsymbol{A}$ and $-\boldsymbol{B}$ have no eigenvalues in common (Lancaster and Tismenetsky, 1985, p. 414).

Corollary 1 to Theorem 1. Suppose that the conditions of Theorem 1 hold, $\text{rank}(\boldsymbol{F}^*) \geq 2$, and that the elements of $\boldsymbol{\Delta}_{k|i}$ and $\tilde{\boldsymbol{\Delta}}_{k|i}$ are generated according to the Bernoulli ± 1 distribution. Then the inequality in (5.2) is strict (i.e., $<$ instead of $\leq$).

Proof. From (5.2) and the definition of Frobenius norm, it is sufficient to show that the inequality in (5.13) is strict for at least one scalar element. From (5.12), this strict inequality follows if $\text{var}(\psi_i^{\boldsymbol{H}} - \psi_i^{\boldsymbol{F}}) < \text{var}(\psi_i^{\boldsymbol{H}})$, which, as noted below (5.11), holds when $E\left[(\psi_i^{\boldsymbol{F}})^2\right] \neq 0$, requiring that $\psi_i^{\boldsymbol{F}}$ cannot be 0 a.s. For both the case of L measurements and $\boldsymbol{g}$ measurements, we now establish that $\psi_i^{\boldsymbol{F}}$ cannot be 0 a.s. for at least one scalar element.

Let $\text{rank}(\boldsymbol{F}^*) = r \leq p$, and without loss of generality assume that the elements of $\boldsymbol{\theta}$ are ordered such that the upper left $r \times r$ block of $\boldsymbol{F}^*$ is full rank; in the arguments below, a subscript $r \times r$ denotes the upper left $r \times r$ block of the indicated matrix. For the case of L measurements, we know from (3.2),

$$
\begin{aligned}
&E\left(\boldsymbol{\Psi}_{1|i}^{(L)}(\boldsymbol{F}^*)\middle|\Delta_{1\,1|i}, \Delta_{12|i}, ..., \Delta_{1r|i}\right) \\
&\quad = \tfrac{1}{2} E\left(\boldsymbol{F}^* \boldsymbol{D}_{1|i} + \boldsymbol{D}_{1|i} \boldsymbol{F}^* \middle|\Delta_{1\,1|i}, \Delta_{12|i}, ..., \Delta_{1r|i}\right) \\
&\quad = \tfrac{1}{2} \boldsymbol{F}^* \begin{bmatrix} \boldsymbol{D}_{1|i;r\times r} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \tfrac{1}{2} \begin{bmatrix} \boldsymbol{D}_{1|i;r\times r} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{F}^*,
\end{aligned}
$$

where the first equality follows by the symmetry of $\boldsymbol{F}^*$ and $\boldsymbol{D}_{1|i}$ (the latter due to the assumed Bernoulli distribution for the perturbations) and the fact that $\tilde{\boldsymbol{D}}_{1|i}$ has mean $\boldsymbol{0}$ while the second equality follows from $\boldsymbol{D}_{1|i}$ having mean $\boldsymbol{0}$ (the $\boldsymbol{D}_{1|i;r\times r}$ submatrix that remains is from the indicated conditioning). Because $\boldsymbol{F}_{r\times r}^*$ and $-\boldsymbol{F}_{r\times r}^*$ have no eigenvalues in common, the above-mentioned result in Lancaster and Tismenetsky (1985, p. 414) indicates that $E\left(\boldsymbol{\Psi}_{1|i}^{(L)}(\boldsymbol{F}^*)\middle|\Delta_{1\,1|i}, \Delta_{12|i}, ..., \Delta_{1r|i}\right) = \boldsymbol{0}$ (a necessary condition for $\boldsymbol{\Psi}_{1|i}^{(L)}(\boldsymbol{F}^*) = \boldsymbol{0}$) if and only if $\boldsymbol{D}_{1|i;r\times r} = \boldsymbol{0}$. This cannot happen since $\boldsymbol{D}_{1|i;r\times r}$ takes on 2^{r-1} unique values ($r \geq 2$), none of which are $\boldsymbol{0}$.

Likewise, for the case of $\boldsymbol{g}$ measurements, we know from (3.3) that $E\left(\boldsymbol{\Psi}_{1|i}^{(\boldsymbol{g})}(\boldsymbol{F}^*)\middle|\Delta_{1\,1|i}, \Delta_{12|i}, ..., \Delta_{1r|i}\right)$ is equal to the right-most expression above for $E\left(\boldsymbol{\Psi}_{1|i}^{(L)}(\boldsymbol{F}^*)\middle|\Delta_{1\,1|i}, \Delta_{12|i}, ..., \Delta_{1r|i}\right)$. Hence, the same reasoning applies using the result in Lancaster and Tismenetsky (1985, p. 414). We thus know that $\psi_i^{\boldsymbol{F}}$ cannot be 0 a.s. for at least one element of $\boldsymbol{\Psi}_{1|i}^{(L)}(\boldsymbol{F}^*)$ or $\boldsymbol{\Psi}_{k|i}^{(\boldsymbol{g})}(\boldsymbol{F}^*)$ (as relevant), completing the proof. *Q.E.D.*

Final remarks and acknowledgment: A numerical study has been performed on the feedback idea in Section 5, showing strong improvement in the quality of the estimate on the problem setting considered in Spall (2005); details are available upon request. This work was partially supported by U.S. Navy Contract N00024-03-D-6606.

References

Apostol, T. M. (1974), *Mathematical Analysis* (2nd ed.), Addison-Wesley, Reading, MA.

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

Das, S., Spall, J. C., and Ghanem, R. (2007), "An Efficient Calculation of Fisher Information Matrix: Monte Carlo Approach Using Prior Information," *Proceedings of the IEEE Conference on Decision and Control*, 12–14 December 2007, New Orleans, LA, USA, paper WeB11.4.

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.

Lancaster, P. and Tismenetsky, M. (1985), *The Theory of Matrices* (2nd ed.), Academic, 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*, March 1995, Baltimore, MD, pp. 659–667.

Ljung, L. (1999), *System Identification—Theory for the User* (2nd ed.), Prentice Hall PTR, Upper Saddle River, NJ.

Segal, M. and Weinstein, E. (1988), "A New Method for Evaluating the Log-Likelihood Gradient (Score) of Linear Dynamic Systems," *IEEE Transactions on Automatic Control*, vol. 33, pp. 763–766.

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), "Monte Carlo Computation of the Fisher Information Matrix in Nonstandard Settings," *Journal of Computational and Graphical Statistics* (American Statistical Assoc.), vol. 14, pp. 889–909.

Spall, J. C. (2006), Convergence Analysis for Feedback- and Weighting-Based Jacobian Estimates in the Adaptive Simultaneous Perturbation Algorithm," *Proceedings of the IEEE Conference on Decision and Control*, 13–15 December 2006, San Diego, CA, pp. 5669–5674 (paper FrB13.2).

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