A Recursive Algorithm for Computing CR-Type Bounds on Estimator Covariance[1]

Alfred Hero (Corresponding Author)

Dept. of EECS
The University of Michigan
Ann Arbor, MI 48109-2122
Tel: (313)-763-0564

Jeffrey A. Fessler

Division of Nuclear Medicine
The University of Michigan
Ann Arbor, MI 48109-0028
Tel: (313)-763-1434

KEYWORDS: Multi-dimensional parameter estimation, estimator covariance bounds, complete-incomplete data problem, image reconstruction.

Abstract

We derive an iterative algorithm that calculates submatrices of the Cramér-Rao (CR) matrix bound on the covariance of any unbiased estimator of a vector parameter $\underline{\theta}$. Our algorithm computes a sequence of lower bounds that converges monotonically to the CR bound with exponential speed of convergence. The recursive algorithm uses an invertable "splitting matrix," and we present a statistical approach to selecting this matrix based on a "complete data - incomplete data" formulation similar to that of the well known EM parameter estimation algorithm. As a concrete illustration we consider image reconstruction from projections for emission computed tomography.

[1]This research was supported in part by the National Science Foundation under grant BCS-9024370, a DOE Alexander Hollaender Postdoctoral Fellowship, and DOE Grant DE-FG02-87ER65061.

1 Introduction

The Cramer-Rao (CR) bound on estimator covariance is an important tool for predicting fundamental limits on best achievable parameter estimation performance. For a vector parameter $\underline{\theta} \in \Theta \subset \mathbb{R}^n$, an observation $\mathbf{Y}$, and a p.d.f. $f_Y(y;\theta)$, one seeks a lower bound on the minimum achievable variance of an unbiased estimator $\hat{\theta}_1 = \hat{\theta}_1(\mathbf{Y})$ of a scalar parameter θ_1 of interest. More generally, if, without loss in generality, the p parameters $\theta_1, \ldots, \theta_p$ are of interest, $p \leq n$, one may want to specify a $p \times p$ matrix which lower bounds the error covariance matrix for unbiased estimators $\hat{\theta}_1, \ldots, \hat{\theta}_p$. The upper left hand $p \times p$ submatrix of the $n \times n$ inverse Fisher information matrix F_Y^{-1} provides the CR lower bound for these parameter estimates. Equivalently, the first p columns of F_Y^{-1} provide this CR bound. The method of sequential partitioning [11] for computing the upper left $p \times p$ submatrix of F_Y^{-1} and Cholesky based Gaussian elimination techniques [3] for computing the p first columns of F_Y^{-1} are efficient direct methods for obtaining the CR bound but require $O(n^3)$ floating point operations. Unfortunately, in many practical cases of interest, e.g. when there are a large number of nuisance parameters, high computation and memory requirements make direct implementation of the CR bound impractical.

In this correspondence we give an iterative algorithm for computing columns of the CR bound which requires only $O(pn^2)$ per iteration. This algorithm falls into the class of "splitting matrix iterations" [3] with the imposition of an additional requirement: the splitting matrix must be chosen to ensure that a valid lower bound results at each iteration of the algorithm. While a purely algebraic approach to this restricted class of iterations can easily be adopted [16], the CR bound setting allows us to exploit additional properties of Fisher information matrices arising from the statistical model. Specifically, we formulate the parameter estimation problem in a complete data - incomplete data setting and apply a version of the "data processing theorem" [1] for Fisher information matrices. This setting is similar to that which underlies the classical formulation of the Maximum Likelihood Expectation Maximization (ML-EM) parameter estimation algorithm. The ML-EM algorithm generates a sequence of estimates $\{\hat{\underline{\theta}}^k\}_k$ for $\underline{\theta}$ which successively increase the likelihood function and converge to the maximum likelihood estimator. In a similar manner, our algorithm generates a sequence of tighter and tighter lower bounds on estimator covariance which converge to the actual CR matrix bound.

The algorithms given here converge monotonically with exponential rate where the speed of convergence increases as the spectral radius $\rho(I - F_X^{-1} F_Y)$ decreases. Here I is the $n \times n$ identity matrix and F_X and F_Y are the complete and incomplete data Fisher information matrices, respectively. Thus when the complete data is only moderately more informative than the incomplete data, F_Y is close to F_X so that $\rho(I - F_X^{-1} F_Y)$ is close to 0 and the algorithm converges very quickly. To implement the algorithm, one must 1) precompute the first p columns of F_X^{-1}, and 2) provide a subroutine that can multiply $F_X^{-1} F_Y$ or $F_X^{-1} E_{\underline{\theta}}[\nabla^{11} Q(\underline{\theta}; \underline{\theta})]$ by a column vector (see (17)). By appropriately choosing the complete data space this precomputation can be quite simple, e.g. $\mathbf{X}$ can be chosen to make F_X sparse or even diagonal. If the complete data space is chosen intelligently only a few iterations may be required to produce a bound which closely approximates the CR bound. In this case the proposed algorithm gives an order of magnitude computational savings as compared to conventional exact methods of computing the CR bound.

The paper concludes with an implementation of the recursive algorithm for bounding the minimum achievable error of reconstruction for a small region of interest (ROI) in an image reconstruction problem arising in emission computed tomography. By using the complete data of the standard EM algorithm for PET reconstruction [15], F_X is diagonal and the implementation of the CR bound algorithm is very simple. As in the PET reconstruction algorithm, the rate of convergence of the iterative CR bound algorithm depends on the image intensity and the tomographic system response matrix.

2 CR Bound and Iterative Algorithm

2.1 Background and General Assumptions

Let Θ_i be an open subset of the real line $\mathbb{R}$. Define $\underline{\theta} = [\theta_1, \ldots, \theta_n]^T$ a real, non-random parameter vector residing in $\Theta = \Theta_1 \times \cdots \times \Theta_n$. Let $\{P_{\underline{\theta}}\}_{\underline{\theta} \in \Theta}$ be a family of probability measures for a certain random variable $\mathbf{Y}$ taking values in a set $\mathcal{Y}$. Assume that for each $\underline{\theta} \in \Theta$, $P_{\underline{\theta}}$ is absolutely continuous with respect to a dominating measure μ so that for each $\underline{\theta}$ there exists a density function $f(y; \underline{\theta}) = dP_{\underline{\theta}}(y)/d\mu$ for $\mathbf{Y}$.

The family of densities $\{f_Y(y; \underline{\theta})\}_{\underline{\theta} \in \Theta}$ is said to be a *regular family* [9] if Θ is an open subset of $\mathbb{R}^p$ and: 1) $f_Y(y; \underline{\theta})$ is a continuous function on Θ for μ-almost all $\mathbf{y}$; 2) $\ln f(\mathbf{Y}; \underline{\theta})$ is mean-square differentiable in $\underline{\theta}$; and 3) $\nabla_{\underline{\theta}} \ln f(\mathbf{Y}; \underline{\theta})$ is mean-square continuous in $\underline{\theta}$. These three conditions guarantee that the Fisher information matrix $F_Y(\underline{\theta}) = E_{\underline{\theta}}[\nabla^T_{\underline{\theta}} \ln f(\mathbf{Y}; \underline{\theta}) \nabla_{\underline{\theta}} \ln f(\mathbf{Y}; \underline{\theta})]$ exists and is finite, where $\nabla_{\underline{\theta}} = [\partial/\partial\theta_1, \ldots, \partial/\partial\theta_n]$ is the (row) gradient operator.

Finally we recall convergence results for linear recursions of the form

$$\underline{v}^{i+1} = A\underline{v}^i, \qquad i = 1, 2, \ldots$$

where $\underline{v}^i$ is a vector and A is a matrix. Let $\rho(A)$ denote the spectral radius, i.e. the maximum magnitude eigenvalue, of A. If $\rho(A) < 1$ then $\underline{v}^i$ converges to zero and the asymptotic rate of convergence increases as the *root convergence factor* $\rho(A)$ decreases [14].

2.2 The CR Lower Bound

Let $\hat{\underline{\theta}} = \hat{\underline{\theta}}(\mathbf{Y})$ be an unbiased estimator of $\underline{\theta} \in \Theta$, and assume that the densities $\{f_Y(y; \underline{\theta}\}_{\underline{\theta} \in \Theta}$ are a regular family. Then the covariance matrix of $\hat{\underline{\theta}}$ satisfies the matrix CR lower bound [9]:

$$\mathrm{cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \geq B(\underline{\theta}) = F_Y^{-1}(\underline{\theta}). \tag{1}$$

In (1) $F_Y(\underline{\theta})$ is the assumed non-singular $n \times n$ Fisher information matrix associated with the measurements $\mathbf{Y}$:

$$F_Y(\underline{\theta}) \stackrel{\text{def}}{=} E_{\underline{\theta}}[\nabla_{\underline{u}} \ln f_Y(\mathbf{Y}; \underline{u})|_{\underline{u}=\underline{\theta}}]^T [\nabla_{\underline{u}} \ln f_Y(\mathbf{Y}; \underline{u})|_{\underline{u}=\underline{\theta}}]. \tag{2}$$

Under the additional assumption that the mixed partials $\frac{\partial^2}{\theta_i \theta_j} f_Y(\mathbf{Y}; \underline{\theta})$, $i, j = 1 \ldots, n$, exist, are continuous in $\underline{\theta}$, and are absolutely integrable in $\mathbf{Y}$, the Fisher information matrix is equivalent to the Hessian, or "curvature matrix," of the mean $\ln f_Y(\mathbf{Y}; \underline{\theta})$:

$$F_Y(\underline{\theta}) = -E_{\underline{\theta}}[\nabla^T_{\underline{u}} \nabla_{\underline{u}} \ln f_Y(\mathbf{Y}; \underline{u})|_{\underline{u}=\underline{\theta}}] = -\nabla^T_{\underline{u}} \nabla_{\underline{u}} E_{\underline{\theta}}[\ln f_Y(\mathbf{Y}; \underline{u})]|_{\underline{u}=\underline{\theta}}. \tag{3}$$

Assume that among the n unknown quantities $\underline{\theta} = [\theta_1, \ldots, \theta_n]^T$ only a small number $p \ll n$ of parameters $\underline{\theta}^I = [\theta_1, \ldots, \theta_p]^T$ are directly of interest, the remaining $n - p$ parameters being considered "nuisance parameters." Partition the Fisher information matrix F_Y as:

$$F_Y = \begin{bmatrix} F_{11} & F_{12}^T \\ F_{12} & F_{22} \end{bmatrix}, \tag{4}$$

where F_{11} is the $p \times p$ Fisher information matrix for the parameters $\underline{\theta}^I$ of interest, F_{22} is the $(n-p) \times (n-p)$ Fisher information matrix for the nuisance parameters, and F_{12} is the $(n-p) \times p$ information coupling matrix. The CR bound on the covariance of any unbiased estimator $\underline{\hat{\theta}}^I = [\hat{\theta}_1, \ldots, \hat{\theta}_p]^T$ of the parameters of interest is simply the $p \times p$ submatrix in the upper left hand corner of F_Y^{-1}:

$$\mathrm{cov}_{\underline{\theta}}(\underline{\hat{\theta}}^I) \geq \mathcal{E}^T F_Y^{-1} \mathcal{E} \tag{5}$$

where $\mathcal{E}$ is the $n \times p$ matrix consisting of the p first columns of the $n \times n$ identity matrix, i.e. $\mathcal{E} = [\underline{e}_1, \ldots, \underline{e}_p]$ and $\underline{e}_j$ is the j-th unit column vector in $\mathbb{R}^n$. Using a standard identity for the partitioned matrix inverse [3] this submatrix can be expressed in terms of the partition elements of F_Y yielding the following equivalent form for the unbiased CR bound:

$$\mathrm{cov}_{\underline{\theta}}(\underline{\hat{\theta}}^I) \quad \geq \quad \left[F_{11} - F_{12}^T F_{22}^{-1} F_{12}\right]^{-1}. \tag{6}$$

By using the method of sequential partitioning [11], the right hand side of (6) could be computed with $O(n^3)$ floating point operations. Alternatively, the CR bound (5) is specified by the first p columns $F_Y^{-1}\mathcal{E}$ of F_Y^{-1}. These p columns are given by the columns of the $n \times p$ matrix solution U to $F_Y U = \mathcal{E}$. The topmost $p \times p$ block $\mathcal{E}^T U$ of U is equal to the right hand side of the CR bound inequality (5). By using the Cholesky decomposition of F_Y and Gaussian elimination [3] the solution U to $F_Y U = \mathcal{E}$ could be computed with $O(n^3)$ floating point operations.

Even if the number of parameters of interest is small, for large n the feasibility of directly computing the CR bound (5) is limited by the high number $O(n^3)$ of floating point operations. For example, in the case of image reconstruction for a moderate sized 256×256 pixelated image F_Y is $256^2 \times 256^2$ so that direct computation of the CR bound on estimation errors in a small region of the image requires on the order of 256^6 or 10^{19} floating point operations!

2.3 A Recursive CR Bound Algorithm

The basic idea of the algorithm is to replace the difficult inversion of F_Y with an easily inverted matrix F. To simplify notation, we drop the dependence on θ. Let F be a $n \times n$ matrix. Assume that F_Y is positive definite and that $F \geq F_Y$, i.e. $F - F_Y$ is nonnegative definite. It follows that F is positive definite, so let $F^{\frac{1}{2}}$ be the positive definite matrix-square-root-factor of F. Then $F^{-\frac{1}{2}} F_Y F^{-\frac{1}{2}}$ is positive definite, $F^{-\frac{1}{2}}(F - F_Y)F^{-\frac{1}{2}}$ is non-negative definite, and therefore:

$$F^{-\frac{1}{2}} F_Y F^{-\frac{1}{2}} \quad = \quad [I - F^{-\frac{1}{2}}(F - F_Y)F^{-\frac{1}{2}}] > 0.$$

Hence $0 \leq I - F^{-\frac{1}{2}} F_Y F^{-\frac{1}{2}} < I$ so that all of the eigenvalues of $I - F^{-\frac{1}{2}} F_Y F^{-\frac{1}{2}}$ are non-negative and strictly less than one. Since $I - F^{-\frac{1}{2}} F_Y F^{-\frac{1}{2}}$ is similar to $I - F^{-1}F_Y$, it follows that the eigenvalues of $I - F^{-1}F_Y$ lie in $[0, 1)$ [8, Corollary 1.3.4]. Thus, applying the matrix form of the geometric series [8, Corollary 5.6.16]:

$$\begin{aligned} B &= [F_Y]^{-1} = [F - (F - F_Y)]^{-1} \\ &= [I - F^{-1}(F - F_Y)]^{-1} F^{-1} \\ &= \left(\sum_{k=0}^{\infty} [I - F^{-1}F_Y]^k \right) F^{-1}. \end{aligned} \tag{7}$$

This infinite series expression for the unbiased $n \times n$ CR bound B is the basis for the matrix recursion given in the following theorem.

Theorem 1 *Assume that F_Y is positive definite and $F \geq F_Y$. When initialized with the $n \times n$ matrix of zeros $B^0 = 0$, the following recursion yields a sequence of matrix lower bounds $B^k = B^k(\underline{\theta})$ on the $n \times n$ covariance of unbiased estimators $\underline{\hat{\theta}}$ of $\underline{\theta}$. This sequence asymptotically converges to the $n \times n$ unbiased CR bound F_Y^{-1} with root convergence factor $\rho(A)$.*

Recursive Algorithm: For $k = 0, 1, 2, \ldots$

$$B^{k+1} = A \cdot B^k + F^{-1}, \tag{8}$$

where $A = I - F^{-1}F_Y$ has eigenvalues in $[0,1)$. Furthermore, the convergence is monotone in the sense that $B^k \leq B^{k+1} \leq B = F_Y^{-1}$, for $k = 0, 1, 2, \ldots$.

Proof

Since all eigenvalues of $I - F^{-1}F_Y$ are in the range $[0,1)$ we obviously have $\rho(I - F^{-1}F_Y) < 1$. Now consider:

$$\begin{aligned} B^{k+1} - F_Y^{-1} &= (I - F^{-1}F_Y)B^k + F^{-1} - F_Y^{-1} \\ &= (I - F^{-1}F_Y)(B^k - F_Y^{-1}) \end{aligned} \tag{9}$$

Since the eigenvalues of $I - F^{-1}F_Y$ are in $[0,1)$, this establishes that $B^{k+1} \to F_Y^{-1}$ as $k \to \infty$ with root convergence factor $\rho(I - F^{-1}F_Y)$. Similarly:

$$B^{k+1} - B^k = (I - F^{-1}F_Y)[B^k - B^{k-1}], \qquad k = 1, 2, \ldots,$$

with initial condition $B^1 - B^0 = F^{-1}$. By induction we have

$$B^{k+1} - B^k = F^{-\frac{1}{2}} \left[F^{-\frac{1}{2}}(I - F_Y)F^{-\frac{1}{2}} \right]^k F^{-\frac{1}{2}}$$

which is non-negative definite for all $k \geq 0$. Hence the convergence is monotone. □

By right multiplying each side of the equality (8) by the matrix $\mathcal{E} = [\underline{e}_1, \ldots, \underline{e}_p]$, where $\underline{e}_j$ is the j-th unit vector in $\mathbb{R}^n$ we obtain a recursion for the first p columns $B^k\mathcal{E} = [\underline{b}_1^k, \ldots, \underline{b}_p^k]$. Furthermore, the first p rows $\mathcal{E}^T B^k \mathcal{E}$ of $B^k\mathcal{E}$ correspond to the upper left hand corner $p \times p$ submatrix of B^k and, since $\mathcal{E}^T[B^{k+1} - B^k]\mathcal{E}$ is non-negative definite, by Theorem 1 $\mathcal{E}^T B^k \mathcal{E}$ converges monotonically to $\mathcal{E}^T F_Y^{-1}\mathcal{E}$. Thus we have the following Corollary to Theorem 1.

Corollary 1 *Assume that F_Y is positive definite and $F \geq F_Y$ and let $\mathcal{E} = [\underline{e}_1, \ldots, \underline{e}_p]$ be the $n \times p$ matrix whose columns are the first p unit vectors in $\mathbb{R}^n$. When initialized with the $n \times p$ matrix of zeros $\beta^0 = 0$, the top $p \times p$ block $\mathcal{E}^T\beta^k$ of β^k in the following recursive algorithm yields a sequence of lower bounds on the covariance of any unbiased estimator of $\underline{\theta}^I = [\theta_1, \ldots, \theta_p]^T$ which asymptotically converges to the $p \times p$ CR bound $\mathcal{E}^T F_Y^{-1}\mathcal{E}$ with root convergence factor $\rho(A)$:*

Recursive Algorithm: For $k = 0, 1, 2, \ldots$

$$\beta^{(k+1)} = A \cdot \beta^{(k)} + \mathcal{F}^{-1}, \tag{10}$$

where $A = I - F^{-1}F_Y$ has eigenvalues in $[0,1)$ and $\mathcal{F}^{-1} = F^{-1}\mathcal{E}$ is the $n \times p$ matrix consisting of the first p columns of F^{-1}. Furthermore, the convergence is monotone in the sense that $\mathcal{E}^T\underline{\beta}^{(k)} \leq \mathcal{E}^T\underline{\beta}^{(k+1)} \leq \mathcal{E}^T F_Y^{-1}\mathcal{E}$, for $k = 0, 1, 2, \ldots$.

The $n \times n$ times $n \times p$ matrix multiplication $A \cdot \beta^k$ requires only $O(pn^2)$ floating point operations. Hence, for $p \ll n$ the recursion (10) requires only $O(n^2)$ floating point operations per iteration.

2.4 Discussion

We make the following comments on the recursive algorithms of Theorem 1 and Corollary 1.

1. For the algorithm (10) for computing columns of F_Y^{-1} to have significant computational advantages relative to the direct approaches discussed in Section 2, the precomputation of the matrix inverse F^{-1} and multiplication by the matrix product $A = I - F^{-1}F_Y$ must be simple, and the iterations must converge reasonably quickly. By choosing an F that is is sparse or diagonal the computation of F^{-1} and A requires only $O(n^2)$ floating point operations. If in addition F is chosen so that $\rho(I - F^{-1}F_Y)$ is small, then the algorithm (10) will converge to within a small fraction of the corresponding column of F_Y^{-1} with only a few iterations and thus will be an order of magnitude less costly than direct methods requiring $O(n^3)$ operations.

2. One can use the recursion $I - F_Y B^{k+1} = A \cdot [I - F_Y B^k]$, obtained similar to (9) of the proof of Theorem 1, to monitor the progress of the j-th column $\underline{b}_j^k - F_Y B^k$ towards zero. This recursion can be implemented alongside of the bound recursion (10).

3. For $p = 1$ the iteration of Corollary 1 is related to the "matrix splitting" method [3] for iteratively approximating the solution $\underline{u}$ to a linear equation $C\underline{u} = \underline{c}$. In this method, a decomposition $C = F - N$ is found for the non-singular matrix C such that F is non-singular and $\rho(F^{-1}N) < 1$. Once this decomposition is found the algorithm below produces a sequence of vectors $\underline{u}^k$ which converge to the solution $\underline{u} = C^{-1}\underline{c}$ as $k \to \infty$:

$$\underline{u}^{k+1} = F^{-1}N\underline{u}^k + F^{-1}\underline{c}. \tag{11}$$

   Identifying C as the incomplete data Fisher information F_Y, N as the difference $F - F_Y$, $\underline{u}$ as the j-th column of F_Y^{-1}, and $\underline{c}$ as the j-th unit vector $\underline{e}_j$ in $\mathbb{R}^n$, the splitting algorithm (11) is equivalent to the column recursion of Corollary 1. The novelty of the recursion of Corollary 1 is that based on statistical considerations presented in the following section we have identified a particular class of matrix decompositions for F_Y that guarantee *monotone convergence* of the j-th component of $\underline{b}_j^k$ to the scalar CR bound on $\mathrm{var}_{\underline{\theta}}(\hat{\theta}_j)$. Moreover, for general $p \geq 1$ the recursion of Corollary 1 implies that when p parallel versions of (11) are implemented with $\underline{c} = \underline{e}_j$ and $\underline{u}^k = \underline{u}_j^k$, $j = 1, \ldots, p$, respectively, the first p rows of the concatenated sequence $[\underline{u}_1^k, \ldots, \underline{u}_p^k]$ converge monotonically to the $p \times p$ CR bound on $\mathrm{cov}_{\underline{\theta}}(\hat{\underline{\theta}}^I)$, $\hat{\theta}^I = [\hat{\theta}_1, \ldots, \hat{\theta}_p]^T$. Monotone convergence is important in the statistical estimation context since it ensures that no matter when the iterative algorithm is stopped a valid lower bound is obtained.

4. The basis for the matrix recursion of Theorem 1 is the geometric series (7). A geometric series approach was also employed in [12, Section 5] to develop a method to speed up the asymptotic convergence of the EM parameter estimation algorithm. This method is a special case of Aitken's acceleration which requires computation of the inverse of the *observed Fisher information* $\hat{F}_Y(\underline{\theta}) = -\nabla^2 \ln f_Y(\mathbf{Y};\underline{\theta})$ evaluated at successive EM iterates, $\underline{\theta} = \underline{\theta}^k$, $k = m, m+1, m+2, \ldots$ where m is a large positive integer. If $\hat{F}(\underline{\theta}^k)$ is positive definite then Theorem 1 of this paper can be applied to iteratively compute this inverse. Unfortunately, $\hat{F}_Y(\underline{\theta})$ is not guaranteed to be positive definite except within a small neighborhood $\{\underline{\theta} : ||\underline{\theta} - \hat{\underline{\theta}}|| \leq \delta\}$ of the MLE, so that in practice such an approach may fail to produce a converging algorithm.

3 Statistical Choice for Splitting Matrix

The matrix F must satisfy $F \geq F_Y$ and must also be easily inverted. For an arbitrary matrix F, verifying that $F \geq F_Y$ could be quite difficult. In this section we present a statistical approach to choosing the matrix F; the matrix F is chosen to be the Fisher information matrix of the complete-data that is intrinsic to an EM algorithm. This approach guarantees that $F \geq F_Y$ through a Fisher information version of the data-processing inequality.

3.1 Incomplete Data Formulation

Many estimation problems can be conveniently formulated as an incomplete data - complete data problem. The setup is the following. Imagine that there exists a different set of measurements $\mathbf{X}$ taking values in a set $\mathcal{X}$ whose probability density $f_X(x;\underline{\theta})$ is also a function of $\underline{\theta}$. Further assume that this hypothetical set of measurements $\mathbf{X}$ is larger and more informative as compared to $\mathbf{Y}$ in the sense that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is functionally independent of $\underline{\theta}$. $\mathbf{X}$ and $\mathcal{X}$ are called the complete data and complete data space while $\mathbf{Y}$ and $\mathcal{Y}$ are called the incomplete data and incomplete data space, respectively. This definition of incomplete - complete data is equivalent to defining $\mathbf{Y}$ as the output of a $\underline{\theta}$-independent possibly noisy channel having input $\mathbf{X}$. Note that our definition contains as a special case the standard definition [2] whereby $\mathbf{X}$ and $\mathbf{Y}$ must be related via a deterministic functional transformation $\mathbf{Y} = h(\mathbf{X})$, where $h : \mathcal{X} \to \mathcal{Y}$ is many-to-one.

Assume that a complete data set $\mathbf{X}$ has been specified. For regular probability densities $f_X(x;\underline{\theta})$, $f_Y(y;\underline{\theta})$, $f_{X|Y}(x|y;\underline{\theta})$ we define the associated Fisher information matrices $F_X(\underline{\theta}) = -E_{\underline{\theta}}\left[\nabla_{\underline{\theta}}\nabla^T_{\underline{\theta}} \ln f_X(\mathbf{X};\underline{\theta})\right]$, $F_Y(\underline{\theta}) = -E_{\underline{\theta}}\left[\nabla_{\underline{\theta}}\nabla^T_{\underline{\theta}} \ln f_Y(\mathbf{Y};\underline{\theta})\right]$, $F_{X|Y}(\underline{\theta}) = -E_{\underline{\theta}}\left[\nabla_{\underline{\theta}}\nabla^T_{\underline{\theta}} \ln f_{X|Y}(\mathbf{X}|\mathbf{Y};\underline{\theta})\right]$, respectively. First we give a decomposition for $F_Y(\underline{\theta})$ in terms of $F_X(\underline{\theta})$ and $F_{X|Y}(\underline{\theta})$..

Lemma 1 *Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables which have a joint probability density $f_{X,Y}(x,y;\underline{\theta})$ relative to some product measure $\mu_X \times \mu_Y$. Assume that $\mathbf{X}$ is more informative than $\mathbf{Y}$ in the sense that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is functionally independent of $\underline{\theta}$. Assume also that $\{f_X(x;\underline{\theta})\}_{\underline{\theta}\in\Theta}$ is a regular family of densities with mixed partials $\frac{\partial^2}{\theta_i\theta_j} f_X(x;\theta)$ which are continuous in $\underline{\theta}$ and absolutely integrable in x. Then $\{f_Y(x;\underline{\theta})\}_{\underline{\theta}\in\Theta}$ is a regular family of densities with continuous and absolutely integrable mixed partials, the above defined Fisher information matrices $F_X(\underline{\theta})$, $F_Y(\underline{\theta})$, and $F_{X|Y}(\underline{\theta})$ exist are finite and*

$$F_Y(\underline{\theta}) \quad = \quad F_X(\underline{\theta}) - F_{X|Y}(\underline{\theta}) \tag{12}$$

Proof of Lemma 1

Since $\mathbf{X}, \mathbf{Y}$ has the density $f_{X,Y}(x,y;\underline{\theta})$ with respect to the measure $\mu_X \times \mu_Y$ there exist versions $f_{Y|X}(y|x;\underline{\theta})$ and $f_{X|Y}(x|y;\underline{\theta})$ of the conditional densities. Furthermore, by assumption $f_{Y|X}(y|x;\underline{\theta}) = f_{Y|X}(y|x)$ does not depend on $\underline{\theta}$. Since $f_Y(y;\underline{\theta}) = \int_{\mathcal{X}} f_{Y|X}(y|x) f_X(x;\underline{\theta}) d\mu_X$ it is straightforward to show that the family $\{f_Y(y;\underline{\theta})_{\underline{\theta}\in\Theta}$ inherits the regularity properties of the family $\{f_X(x;\underline{\theta})_{\underline{\theta}\in\Theta}$. Now for any y such that $f_Y(y;\underline{\theta}) > 0$ we have from Bayes' rule

$$f_{X|Y}(x|y;\underline{\theta}) \quad = \quad \frac{f_{Y|X}(y|x) f_X(x;\underline{\theta})}{f_Y(y;\underline{\theta})}. \tag{13}$$

Note that $f_{X,Y}(x,y;\underline{\theta}) > 0$ implies that $f_Y(y;\underline{\theta}) > 0$, $f_X(x;\underline{\theta}) > 0$, $f_{Y|X}(y|x) > 0$ and $f_{X|Y}(x|y;\underline{\theta}) > 0$. Hence, we can use (13) to express:

$$\log f_{X|Y}(x|y;\underline{\theta}) \quad = \quad \log f_X(x;\underline{\theta}) - \log f_Y(y;\underline{\theta}) + \log f_{Y|X}(y|x), \tag{14}$$

whenever $f_{X,Y}(x,y;\underline{\theta}) > 0$. From this relation it is seen that $f_{X|Y}(x|y;\underline{\theta})$ inherits the regularity properties of the $\mathbf{X}$ and $\mathbf{Y}$ densities. Therefore, since the set $\{(x,y) : f_{X,Y}(x,y;\underline{\theta}) > 0\}$ has probability one we obtain from (14):

$$E_{\underline{\theta}}[-\nabla^2_{\underline{\theta}} \log f_{X|Y}(\mathbf{X}|\mathbf{Y};\underline{\theta})] = E_{\underline{\theta}}[-\nabla^2_{\underline{\theta}} \log f_X(\mathbf{Y};\underline{\theta})] - E_{\underline{\theta}}[-\nabla^2_{\underline{\theta}} \log f_Y(\mathbf{X};\underline{\theta})]$$

This establishes the lemma. $\square$

Since the Fisher information matrix $F_{X|Y}$ is non-negative definite, an important consequence of the decomposition of Lemma 1 is the matrix inequality:

$$F_X(\underline{\theta}) \geq F_Y(\underline{\theta}). \tag{15}$$

The inequality (15) can be interpreted as a version of the "data processing theorem" of information theory [1] which asserts that any irreversible processing of data $\mathbf{X}$ entails a loss in information in the resulting data $\mathbf{Y}$.

3.2 Remarks

1. The inequality (15) is precisely the condition required of the splitting matrix F by the recursive CR bound algorithm (10). Furthermore, in many applications of the EM algorithm, the complete-data space is chosen such that the dependence of X on θ is "uncoupled," so that F_X is diagonal or very sparse. Since many of the problems in which F_Y is difficult to invert are problems for which the EM algorithm has been applied, the Fisher information of the corresponding complete-data space is thus a natural choice for F.

2. If the incomplete data Fisher matrix F_Y is available the matrix A in the recursion (8) can be precomputed as:

$$A = I - F_X^{-1} F_Y. \tag{16}$$

   On the other hand, if the Fisher matrix F_Y is not available, the matrix A in the recursion (8) can be computed directly from $Q(\underline{u},\underline{v}) = E\{\log f(\mathbf{X};\theta)|\mathbf{Y},\bar{\theta}\}$ arising from the E step of the EM parameter estimation algorithm [2]. Note that, under the assumption that exchange of order of differentiation and expectation is justified [10, Sec. 2.6]:

$$\begin{aligned} F_{X|Y}(\underline{\theta}) &= E_{\underline{\theta}}\left[-\nabla^2_{\underline{u}} E_{\underline{\theta}}\left[\ln f_{X|Y}(\mathbf{X}|\mathbf{Y};\underline{u})|\mathbf{Y}\right] |_{\underline{u}=\underline{\theta}}\right] \\ &= E_{\underline{\theta}}\left[-\nabla^{20} H(\underline{\theta};\underline{\theta})\right], \end{aligned}$$

   where $H(\underline{u};\underline{v}) \stackrel{\text{def}}{=} E_{\underline{v}}\{\log f_{X|Y}(\mathbf{X}|\mathbf{Y};\underline{u})|\mathbf{Y}=y\}$. We can make use of an identity [2, Lemma 2]:

$$\nabla^{20} H(\underline{\theta};\underline{\theta}) = -\nabla^{11} H(\underline{\theta};\underline{\theta})$$

   Furthermore,

$$\nabla^{11} H(\underline{\theta};\underline{\theta}) = \nabla^{11} Q(\underline{\theta};\underline{\theta}).$$

   This gives the identity:

$$F_{X|Y}(\underline{\theta}) \quad = \quad E_{\underline{\theta}}[\nabla^{11} Q(\underline{\theta};\underline{\theta})].$$

   Giving an alternative expression to (16) for precomputing A:

$$A = F_X^{-1} E_{\underline{\theta}}[\nabla^{11} Q(\underline{\theta};\underline{\theta})]. \tag{17}$$

3. The form $\rho(I - F_X^{-1} F_Y)$ for the rate of convergence of the algorithms (8) and (10) implies that for rapid convergence the complete data space $\mathbf{X}$ should be chosen such that $\mathbf{X}$ is not significantly more informative than $\mathbf{Y}$ relative to the parameter $\underline{\theta}$.

4. The matrix recursion of Theorem 1 can be related to the following Frobenius normalization method for inverting a sparse matrix C:

$$B^{k+1} = B^k[I - \alpha C] + \alpha I, \tag{18}$$

where $\alpha = 1/\|C\|_2$ is the inverse of the Frobenius norm of C. When initialized with $B^0 = I$ the above algorithm converges to C^{-1} as $k \to \infty$. For the case that C is the Fisher matrix F_Y the matrix recursion (18) can be interpreted as a special case of the algorithm of Theorem 1 for a particular choice of complete data $\mathbf{X}$. Specifically, let the complete data be defined as the concatenation $\mathbf{X} = [\mathbf{Y}^T, \mathbf{S}^T]^T$ of the incomplete data $\mathbf{Y}$ and a hypothetical data set $\mathbf{S} = [\mathbf{S}_1, \ldots, \mathbf{S}_m]^T$ defined by the following:

$$\mathbf{S} = \underline{c}(\underline{\theta}) + \mathbf{W}, \tag{19}$$

where $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_m]^T$ are i.i.d. standard Gaussian random varibles independent of $\mathbf{Y}$, and $\underline{c} = [c_1, \ldots, c_m]^T$ is a vector function of $\underline{\theta}$. It is readily verified that the Fisher matrix F_S for $\underline{\theta}$ based on observing $\mathbf{S}$ is of the form: $F_S = \sum_{j=1}^m \nabla^T c_j(\underline{\theta}) \nabla c_j(\underline{\theta})$, where $\nabla = [\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_n}]$. Now since $\mathbf{S}$ and $\mathbf{Y}$ are independent $F_X = F_S + F_Y$ so that if we could choose $c(\underline{\theta})$ such that $F_S = \|F_Y\| \cdot I - F_Y$ the recursion of Theorem 1 would be equivalent to (18) with $F_Y = C$, $F_X^{-1} = \alpha I$, $A = I - \alpha F_Y$. In particular, for the special case that F_Y is functionally independent of $\underline{\theta}$, we can take m equal to n and take the hypothetical data $\mathbf{S} = [\mathbf{S}_1, \ldots, \mathbf{S}_n]^T$ as the n-dimensional linear Gaussian model:

$$\mathbf{S}_j = \underline{c}_j \underline{\theta}^T + \mathbf{W}_j, \quad j = 1, \ldots, n,$$

where

$$\underline{c}_j = [\|F_Y\|_2 - \sqrt{\lambda_j}]\underline{\nu}_j, \quad j = 1, \ldots, n,$$

and $\{\underline{\nu}_1, \ldots, \underline{\nu}_n\}$ are the eigenvectors and $\{\lambda_1, \ldots, \lambda_n\}$ are the eigenvalues of F_Y. With this definition of $\mathbf{S}$, $F_Y = \sum_{j=1}^n \lambda_j \underline{\nu}_j \underline{\nu}_j^T$ is simply the eigendecomposition of the matrix $\|F_Y\| \cdot I - F_Y$ so that $F_X = \|F_Y\| \cdot I = \alpha I$ as required.

4 Application to ECT Image Reconstruction

We consider the case of positron emission tomography (PET) where a set of m detectors is placed about an object to measure positions of emitted gamma-rays. The mathematical formulation of PET is as follows. Over a specified time interval a number $\mathbf{N}_b$ of gamma-rays are randomly emitted from pixel b, $b = 1, \ldots, n$, and a number $\mathbf{Y}_d$ of these gamma-rays are detected at detector d, $d = 1, \ldots, m$. The average number of emissions at pixels $1, \ldots, n$ is an unknown vector $\underline{\theta} = [\theta_1, \ldots, \theta_n]^T$, called the object intensity. It is assumed that the $\mathbf{N}_b$'s are independent Poisson random variables with rates θ_b, $b = 1, \ldots, n$, and the $\mathbf{Y}_d$'s are independent Poisson distributed with rates $\mu_d = \sum_{b=1}^n P_{d|b}\theta_b$, where $P_{d|b}$ is the transition probability corresponding to emitter location b and detector location d. For simplicity we assume that $\mu_d > 0$ $\forall d$. The objective is to estimate a subset $[\theta_1, \ldots, \theta_p]^T$, $p \ll n$, of the object intensities within a p-pixel region of interest (ROI). In this section we develop the recursive CR bound for this estimation problem.

The log-likelihood function for $\underline{\theta}$ based on $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_m]^T$ is simply

$$\begin{aligned} \ln f_Y(\mathbf{Y}; \underline{\theta}) &= \ln \prod_{d=1}^m \frac{[\mu_d]^{\mathbf{Y}_d}}{\mathbf{Y}_d!} e^{-\mu_d} && (20) \\ &= -\sum_{d=1}^m \mu_d + \sum_{d=1}^m \mathbf{Y}_d \ln \mu_d + \text{constant}. && (21) \end{aligned}$$

From this the Hessian matrix with respect to $\underline{\theta}$ is simply calculated and, using the fact that $E_{\underline{\theta}}[\mathbf{Y}_d] = \mu_d$, the $n \times n$ Fisher information matrix F_Y is obtained:

$$\begin{aligned} F_Y &= \sum_{d=1}^{m} \frac{1}{\mu_d} \underline{P}_{d|*}^T \underline{P}_{d|*} \\ &= \left(\left(\sum_{d=1}^{m} \frac{P_{d|i} P_{d|j}}{\mu_d} \right) \right)_{i,j=1,\ldots,n}, \end{aligned} \tag{22}$$

where $\underline{P}_{d|*} = [P_{d|1}, \ldots, P_{d|n}]$ is the d-th row of the $m \times n$ system matrix $((P_{j|i}))$. If $m \geq n$, and the linear span of $\{P_{d|*}\}_{d=1}^{n}$ is $\mathbb{R}^n$, then F_Y is invertible and the CR bound exists. However, even for an imaging system of moderate resolution, e.g. a 256×256 pixel plane, direct computation of the $p \times p$ ROI submatrix $\mathcal{E}^T F_Y^{-1} \mathcal{E}$, $\mathcal{E} = [\underline{e}_1, \ldots, \underline{e}_p]$, of the $(256)^2 \times (256)^2$ Fisher matrix F_Y is impractical.

The standard choice of complete data for estimation of $\underline{\lambda}$ via the EM algorithm is the set $\mathbf{N}_{db}$, $d = 1, \ldots, m$, $b = 1, \ldots, n$, where $\mathbf{N}_{db}$ denotes the number of emissions in pixel b which are detected at detector d. $\{\mathbf{N}_{db}\}$ are independent Poisson random variables with intensity $E_{\underline{\lambda}}[\mathbf{N}_{db}] = P_{d|b} \lambda_b$, $d = 1, \ldots, m$, $b = 1, \ldots, n$. By Lemma 1 we know that, with F_X the Fisher information matrix associated with the complete data, $F_X - F_Y$ is non-negative definite. Thus F_X can be used in Theorem 1 to obtain a montonically convergent CR bound recursion.

The log-likelihood function associated with the complete data set $\mathbf{X} = [\mathbf{N}_{db}]_{d=1,b=1}^{m,n}$ is of similar form to (21):

$$\ln f_X(\mathbf{X}; \underline{\theta}) = -\sum_{d=1}^{m} \sum_{b=1}^{n} P_{d|b} \underline{\theta}_b + \sum_{d=1}^{m} \sum_{b=1}^{n} \mathbf{N}_{db} \ln \theta_b + \text{constant}.$$

The Hessian $\nabla_{\underline{\theta}}^2 \ln f_X(\mathbf{X}; \underline{\theta})$ is easily calculated, and, assuming $\theta_b > 0$ $forallb$, the Fisher information matrix F_X is obtained as:

$$F_X = \text{diag}_b \left(\frac{\sum_{d=1}^{m} P_{d|b}}{\underline{\theta}_b} \right), \tag{23}$$

where $\text{diag}_b(a_b)$ denotes a diagonal $n \times n$ matrix with the a_b's indexed successively along the diagonal.

Using the results (23) and (22) above we obtain:

$$\begin{aligned} A &= I - F_X^{-1} F_Y \\ &= I - \left(\left(\sum_{l=1}^{d} \frac{\underline{\theta}_i}{\sum_{d=1}^{m} P_{d|i}} \frac{P_{l|i} P_{l|j}}{\mu_l} \right) \right)_{i,j=1,\ldots,n}. \end{aligned} \tag{24}$$

The recursive CR bound algorithm of Corollary 1 can now be implemented using the matrices (23) and (24).

The rate of convergence of the recursive CR bound algorithm is determined by the maximum eigenvalue $\rho(A)$ of A specified by (24). For a fixed system matrix $((P_{j|i}))$ the magnitude of this eigenvalue will depend on the image intensity $\underline{\theta}$. Assume for simplicity that with probability one any emitted gamma-ray is detected at some detector, i.e. $\sum_{d=1}^{m} P_{d|b} = 1$ for all b. Since $\text{trace}(A) = \sum_{i=1}^{n} \lambda_i$, where $\{\lambda_i\}_{i=1}^{n}$ are the eigenvalues of A, using (24) it is seen that the maximum eigenvalue $\rho(A)$ must satisfy:

$$\frac{1}{n} \text{trace}(A) = 1 - \frac{1}{n} \sum_{l=1}^{d} \frac{\sum_{i=1}^{n} P_{l|i}^2 \underline{\theta}_i}{\sum_{j=1}^{n} P_{l|j} \underline{\theta}_j} \leq \rho(A) < 1. \tag{25}$$

Now applying Jensen's inequality to $\sum_{i=1}^{n} P_{l|i}^2 \underline{\theta}_i$ it can be shown that

$$\frac{1}{n} \text{trace}(A) \leq 1 - \frac{1}{n}, \tag{26}$$

where equality occurs if $P_{l|j}$ independent of j. On the other hand, it can easily be seen that equality is approached in (26) as the intensity $\underline{\theta}$ concentrates an increasing proportion of its mass on a single pixel k_o, e.g.:

$$\underline{\theta}_i = \begin{cases} (1-\epsilon)\frac{n-1}{n} \sum_{b=1}^{n} \underline{\theta}_b, & i = k_o \\ \\ \epsilon \frac{1}{n} \sum_{b=1}^{n} \underline{\theta}_b & i \neq k_o \end{cases},$$

and ϵ approaches zero. Thus for this case we have from (25): $1 - 1/n < \rho(A) < 1$. Since n is typically very large, this implies that the asymptotic convergence rate of the recursive algorithm will suffer for image intensities which approach that of an ideal point source, at least for this particular splitting matrix.

5 Conclusion and Future Work

We have given a recursive algorithm which can be used to compute submatrices of the CR lower bound F_Y^{-1} on unbiased multi-dimensional parameter estimation error covariance. The algorithm sucessively approximates the inverse Fisher information matrix F_Y^{-1} via a monotonically convergent splitting matrix iteration. We have given a statistical methodology for selecting an appropriate splitting matrix F which involves application of a data processing theorem to a complete-data incomplete-data formulation of the estimation problem. We are currently investigating a purely algebraic methodology in which a splitting matrix F is selected from sets of diagonal, tridiagonal, or circulant matrices to optimize the norm difference $||F - F_Y||$ subject to $F \geq F_Y$. We are also developing analogous recursive algorithms to sucessively approximate generalized matrix CR bounds, such as those developed in [7], for biased estimation.

References

[1] R. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, 1987.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society, Ser. B*, vol. 39, pp. 1–38, 1977.

[3] G. H. Golub and C. F. Van Loan, *Matrix Computations* (2nd Edition), The Johns Hopkins University Press, Baltimore, 1989.

[4] J. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. on Inform. Theory*, Nov. 1990.

[5] A. O. Hero and L. Shao, "Information analysis of single photon computed tomography with count losses," *IEEE Trans. on Medical Imaging*, vol. 9, no. 2, pp. 117–127, June 1990.

[6] A. O. Hero and J. A. Fessler, "On the convergence of the EM algorithm," Technical Report in prep., Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, April 1992.

[7] A. O. Hero, "A Cramer-Rao type lower bound for essentially unbiased parameter estimation," MIT Lincoln Laboratory, Lexington, Mass., Technical Rep. 890, (3 January 1992). DTIC AD-A246666.

[8] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge, 1985.

[9] I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: Asymptotic theory*, Springer-Verlag, New York, 1981.

[10] S. Kullback, *Information Theory and Statistics*, Dover, 1978.

[11] A. R. Kuruc, "Lower bounds on multiple-source direction finding in the presence of direction-dependent antenna-array-calibration errors," Technical Report 799, M.I.T. Lincoln Laboratory, Oct., 1989.

[12] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. Royal Statistical Society, Ser. B*, vol. 44, no. 2, pp. 226–233, 1982.

[13] M. I. Miller, D. L. Snyder, and T. R. Miller, "Maximum-likelihood reconstruction for single photon emission computed tomography," *IEEE Trans. Nuclear Science*, vol. 32, pp. 769–778, Feb. 1985.

[14] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[15] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. on Medical Imaging*, vol. MI-1, No. 2, pp. 113–122, Oct. 1982.

[16] M. Usman and A. O. Hero, "Algebraic versus statistical optimization of convergence rates for recursive CR bound algorithms," Technical Report 298, Communications and Signal Processing Laboratory, Dept. of EECS, The University of Michigan, Ann Arbor, 48109-2122, Jan. 1993.