A KRYLOV SUBSPACE METHOD FOR LARGE ESTIMATION PROBLEMS

Michael K. Schneider and Alan S. Willsky

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA

ABSTRACT

Computing the linear least-squares estimate of a high-dimensional random quantity given noisy data requires solving a large system of linear equations. In many situations, one can solve this system efficiently using the conjugate gradient (CG) algorithm. Computing the estimation error variances is a more intricate task. It is difficult because the error variances are the diagonal elements of a complicated matrix. This paper presents a method for using the conjugate search directions generated by the CG algorithm to obtain a converging approximation to the estimation error variances. The algorithm for computing the error variances falls out naturally from a novel estimation-theoretic interpretation of the CG algorithm. The paper discusses this interpretation and convergence issues and presents numerical examples.

1. INTRODUCTION

For certain large linear least-squares estimation problems, especially in medical imaging and remote sensing, one is interested in computing not only estimates but also estimation error variances. The error variances provide important quantitative information concerning the quality of the estimates that can be used in subsequent data analysis and fusion. This paper presents a method for computing both the estimates and the error variances. The method is efficient for a significant number of large estimation problems.

The estimation algorithm presented in this paper has connections to a variety of algorithms for solving linear algebra problems. In particular, one can view the estimation algorithm as a variant of the conjugate gradient (CG) method for solving linear systems of equations. Paige and Saunders have also discussed a variant of the CG algorithm, LSQR, that is capable of computing an approximation to the error variances [4]. Unlike the algorithm proposed here, however, that approximation often does not converge. Other work to which our algorithm is related includes that on Krylov subspace model reduction [3,8]. These algorithms generate a reduced-order, deterministic model of a dynamic system. Our algorithm also generates a reduced-order model using a Krylov subspace method, but the model is static and stochastic, in nature.

Section 2 derives the algorithm, Section 3 discusses convergence issues, and Section 4 presents numerical examples.

This material is based upon work supported in part by a National Science Foundation Graduate Fellowship, ONR grant N00014-91-J-1004, and AFOSR Grant F49620-98-1-0349.

2. ALGORITHM DERIVATION

Consider the problem of forming the linear least-squares estimate (LLSE) of a zero-mean n-dimensional random vector x given an m-dimensional linear measurement $y = Cx + v$ where v is a zero-mean random variable uncorrelated with x and C is a deterministic matrix. Denote the covariance of x by Λ_x and that of v by R. The LLSE of x and associated error covariance are given by $\hat{x}(y) = \Lambda_x C^T \Lambda_y^{-1} y$ and $\Lambda_e(y) = \Lambda_x - \Lambda_x C^T \Lambda_y^{-1} C \Lambda_x$, respectively, where $\Lambda_y = C\Lambda_x C^T + R$ is the covariance of y. If one assumes that multiplication by Λ_x and C is efficient, then the computation of $\hat{x}$ and Λ_e is dominated by matrix-vector and matrix-matrix products involving the inverse of of Λ_y.

The work of performing these multiplies by a matrix inverse could be reduced if one had available a set of linearly independent vectors p_i that whiten the data. In other words, one desires p_i such that $\mathrm{E}\left[(p_i^T y)(p_j^T y)\right] = p_i^T \Lambda_y p_j$ equals one if $i = j$ and zero otherwise. Then, the estimate of x based on y is the same as that based on $p_1^T y, \ldots, p_m^T y$. Furthermore, one can use the following recursion to compute, for each k, the LLSE of x based on $p_1^T y, \ldots, p_k^T y$, $\hat{x}(p_1^T y, \ldots, p_k^T y)$, and the associated error variances, which are the diagonal elements of the estimation error covariance, $\Lambda_e(p_1^T y, \ldots, p_k^T y)$:

$$\hat{x}(p_1^T y, p_2^T y, \ldots, p_{k+1}^T y) = \hat{x}(p_1^T y, p_2^T y, \ldots, p_k^T y) + \Lambda_x C^T p_{k+1} p_{k+1}^T y \quad (1)$$

$$\left(\Lambda_e(p_1^T y, p_2^T y, \ldots, p_{k+1}^T y)\right)_{ii} = \left(\Lambda_e(p_1^T y, p_2^T y, \ldots, p_k^T y)\right)_{ii} - (\Lambda_x C^T p_{k+1})_i^2 \quad (2)$$

where $(\cdot)_{ii}$ denotes the ith element of a matrix, $(\cdot)_i$ denotes the ith element of a vector, and i runs from one to n. The recursion is initialized by setting $\hat{x}(p_1^T y) = \Lambda_x C^T p_1 p_1^T y$ and $\left(\Lambda_e(p_1^T y)\right)_{ii} = (\Lambda_x)_{ii} - (\Lambda_x C^T p_1)_i^2$ for $i = 1, \ldots, n$.

One method for recursively choosing p_i that whiten the data is as follows:

$$p_1 = \frac{y}{\sqrt{y^T \Lambda_y y}} \quad (3)$$

$$r_k = y - \hat{y}(p_1^T y, \ldots, p_k^T y) \quad (4)$$

$$\nu_{k+1} = r_k - (r_k^T \Lambda_y p_k) p_k \quad (5)$$

$$p_{k+1} = \frac{\nu_{k+1}}{\sqrt{\nu_{k+1}^T \Lambda_y \nu_{k+1}}} \quad (6)$$

where $\hat{y}(p_1^T y, \ldots, p_k^T y) = \Lambda_y (p_1 p_1^T + \cdots + p_k p_k^T) y$ is the best linear estimate of y based on the linear functionals of y, $p_1^T y, \ldots, p_k^T y$ (for the purposes of forming $\hat{y}$, the $p_1, \ldots, p_k$ are viewed as deterministic vectors). The idea here is to first choose $p_1 \propto y$ and such that $\mathrm{Var}(p_1^T y) = 1$. The remaining p_k are defined by a recursion. First, the error, r_k, in estimating y based on $p_1^T y, \ldots p_k^T y$ is computed. Then, $r_k^T y$ is made uncorrelated with $p_k^T y$ to form $\nu_{k+1}^T y$. Finally, $\nu_{k+1}^T y$ is normalized to have unit variance.

That the p_i chosen according to (3)-(6) whiten the data follows from standard results concerning the CG algorithm. This method for picking the p_i is, in fact, the CG algorithm's method for picking conjugate search directions when computing $\Lambda_y^{-1} y$. In the context of estimation, Λ_y-conjugate means white; so, the standard theorems for demonstrating that the p_i are Λ_y-conjugate imply that the $p_i^T y$ are white [1,2].

Much of the theory regarding the CG algorithm exploits the fact that

$$\mathrm{span}(p_1, \ldots, p_k) = \mathrm{span}(y, \Lambda_y y, \ldots, \Lambda_y^{k-1} y), \quad (7)$$

which is the Krylov subspace of dimension k associated with the vector y and matrix Λ_y. Thus, the proposed estimation algorithm is computing estimates and error variances for the problem of estimating x based on the projection of the measurements y onto a Krylov subspace. Note that the novelty of the Krylov subspace estimation algorithm is its ability to exploit an estimation-theoretic interpretation of CG to compute estimation error variances.

3. CONVERGENCE ISSUES

Assuming that one can efficiently multiply vectors by Λ_y and $\Lambda_x C^T$, the proposed method for computing estimates and error variances is efficient provided that one can stop the recursion in (1) and (2) after a few number of steps k such that $\hat{x}(p_1^T y, \ldots, p_k^T y) \approx \hat{x}(y)$ and $\left(\Lambda_e(p_1^T y, \ldots, p_k^T y)\right)_{ii} \approx \left(\Lambda_e(y)\right)_{ii}$ for $i = 1, \ldots, n$. The standard convergence results for CG imply that one can stop after a few number of steps and obtain a good approximation to $\hat{x}(y)$ [1,2]. However, the convergence of the computed error variances does not immediately follow from these standard results.

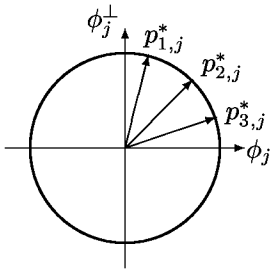

Figure 1: This picture illustrates the statement of Theorem 1 in [6]. The vector ϕ_j is an eigenvector of Λ_y, and $\phi_j^\perp$ represents the directions in the orthogonal complement of $\mathrm{span}(\phi_j)$. The vectors $p_{1,j}^*$, $p_{2,j}^*$, and $p_{3,j}^*$ are the closest unit length vectors in $\mathrm{span}(p_1)$, $\mathrm{span}(p_1, p_2)$, and $\mathrm{span}(p_1, p_2, p_3)$, respectively, to ϕ_j. The theorem establishes the rate at which the $p_{i,j}^*$ are approaching the eigenvectors ϕ_j as i tends to infinity.

Analyzing the convergence of the computed error variances is difficult, in general, but the analysis can be carried out for certain special cases. In particular, one can analyze the situation in which x, v, and y are jointly Gaussian random vectors, $C = I$, $R = I$, and Λ_x has eigenvalues that decrease geometrically. A description of the three major pieces of the analysis is provided here. More details can be found in [7].

The first piece of the analysis bounds the angle between the span of the first k conjugate search directions,

$$\mathrm{span}(p_1, \ldots, p_k), \quad (8)$$

and the dominant eigenvectors of Λ_y. By Theorem 1 in [6], this angle is rapidly decreasing provided that the data y has significant components in the directions of all its eigenvectors (see Figure 1). The second piece of the analysis establishes that, with probability one, y has significant components in the directions of all its eigenvectors. Specifically, the components of y in the directions of its eigenvectors, divided by the corresponding eigenvalues, are uniformly bounded away from zero. The third piece of the analysis consists of noting that Λ_x and Λ_y have the same eigenvectors and that the corresponding eigenvalues differ by one. These three facts imply that the p_i are tending to the direction of the small eigenvectors of Λ_x; so, only the first few p_i are significant in the recursion (1) and (2).

4. NUMERICAL EXAMPLES

Results from running the Krylov subspace estimation algorithm on two synthetic examples are presented here. In both cases, a computer generates a realization of the random vector to be estimated, x, and noisy observations of the vector, y. The estimation algorithm is then run for a certain number of steps to obtain estimates of x and estimation error variances. The number of steps needed is determined by trial and error. Numerical issues surrounding the Λ_y-conjugacy of the p_i are addressed using a non-standard implementation of the CG algorithm that incorporates techniques developed by Parlett and Scott for maintaining the orthogonality of the Lanczos vectors in the Lanczos algorithm [5]. Further implementation details can be found in [7].

Figure 2 depicts results for estimating 1024 samples of a fractional Brownian motion (fBm). The fBm has a Hurst parameter of 0.75 and is scaled to have unit variance at time one. The measurements of the fBm consist of the samples in the intervals $[0, 0.25]$ and $[0.75, 1]$ embedded in independent zero-mean white Gaussian noise with variance 3.2. The solid line in Figure 2a is the path of the fBm, and the dashed lines are the computed estimate and the estimate plus and minus the square root of the computed error variances, *i.e.*, the error standard deviations. Figure 2b depicts the error standard deviations by themselves, and Figure 2c depicts the difference between the error variances computed using the Krylov subspace algorithm and the optimal ones computed using direct methods in MATLAB on a machine with a floating point precision of 2.2×10^{-16}. The results of Figure 2 were generated using nine steps of the algorithm. That only nine steps were needed indicates that the estimation problem is solved by reducing the 512-dimensional measurement vector to a nine-dimensional one.

Figure 3 depicts results for estimating a stationary Gaussian random field on a 32×32 toroidal grid. The power spectral density (collection of eigenvalues) of the field is given by

$$\frac{0.3^{\sqrt{i^2+j^2}}}{\sum_{-15 \le k,l \le 16} 0.3^{\sqrt{k^2+l^2}}} \tag{9}$$

where $-15 \le i, j \le 16$. The sum in the denominator sets the variance of the field to one. Measurements are made of those random field elements whose coordinates (i, j) are such that $-15 \le j \le 16$ and $-15 \le i \le -8$ or $9 \le i \le 16$. The measurements contain independent zero-mean white Gaussian noise with variance 16. Figure 3a depicts the random field to be estimated; 3b, the computed estimates; 3c, the computed error standard deviations; and 3d, the difference between the error variances computed using the Krylov subspace estimation algorithm and the optimal ones computed using direct methods in MATLAB. The results shown in Figure 3 were generated using 50 steps of the algorithm. That 50 steps were used indicates that the estimation problem was solved by reducing the 512-dimensional measurement vector to a 50-dimensional one.

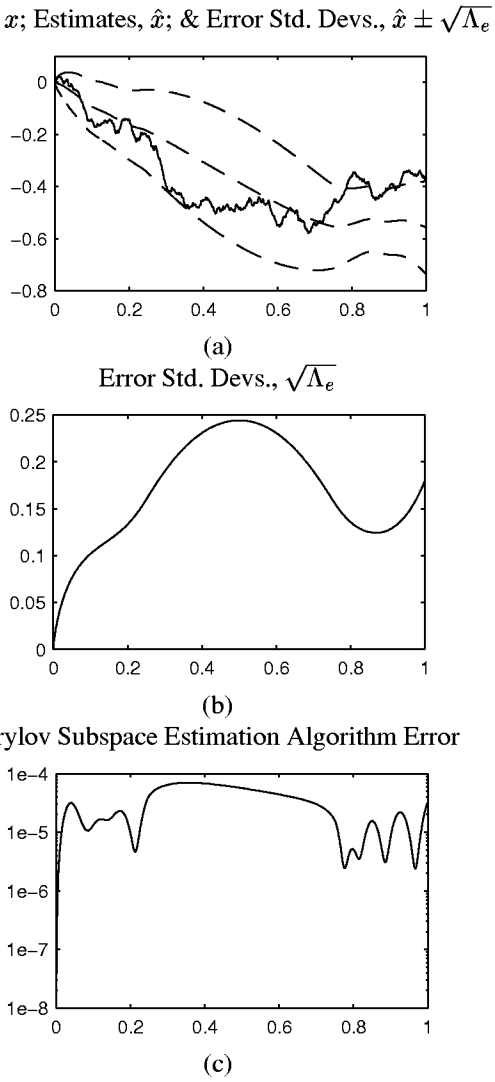

Figure 2: The solid line in (a) represents a fBm with Hurst parameter 0.75 and a variance at time one of one. The dotted lines in (a) represent the computed LLSE based on noisy sparse measurements of the fBm and the LLSE plus and minus the computed error standard deviations. The computed error standard deviations are depicted by themselves in (b), and the difference between the error variances computed using the Krylov subspace estimation algorithm and the optimal ones computed using direct methods in MATLAB are depicted in (c).

For both of these examples, the Krylov subspace estimation algorithm has efficiently computed estimates and estimation error variances. The error variances are close to the optimal estimation error variances relative to the maximum

a priori variance over the domain of the problem. The algorithm has also been tested on estimation problems involving other prior covariances and measurement structures [7].

5. CONCLUSION

This paper presents an estimation-theoretic interpretation of the CG algorithm that has led to a novel method for computing estimation error variances for linear least-squares estimation problems. Analysis and numerical examples establish that the algorithm works and is efficient for certain problems. These promising results encourage further investigation. Potential topics of research include an extensive examination of the range of applicability of the algorithm and the development of an automatic stopping criterion for selecting the number of steps needed to obtain good accuracy.

6. REFERENCES

[1] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

[2] Gene Golub and Charles Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

[3] Inmad Jaimoukha and Ebrahim Kasenally. Oblique projection methods for large scale model reduction. *SIAM Journal on Matrix Analysis and Applications*, 16(2):602–627, April 1995.

[4] Christopher C. Paige and Michael A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 1982.

[5] B. N. Parlett and D. S. Scott. The Lanczos algorithm with selective orthogonalization. *Mathematics of Computation*, 33(145):217–238, January 1979.

[6] Y. Saad. On the rates of convergence of the Lanczos and the block-Lanczos method. *SIAM Journal of Numerical Analysis*, 17(5):687–706, October 1980.

[7] Michael K. Schneider. A Krylov subspace estimation algorithm. Doctoral Thesis Proposal, September 1998.

[8] L. Miguel Silveira, Mattan Kamon, Ibrahim Elfadel, and Jacob White. A coordinate-transformed Arnoldi algorithm for generating stable reduced-order models of RLC circuits. In *IEEE International Conference on Computer-Aided Design*, November 1996.

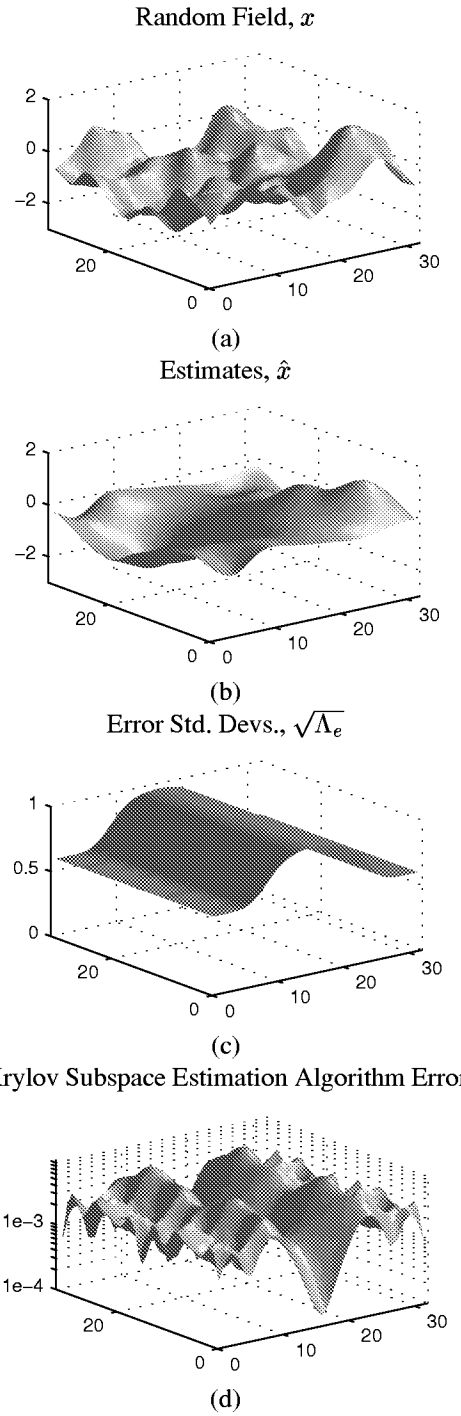

Figure 3: The surface in (a) depicts a stationary random field on a 32×32 toroidal grid. The computed LLSE based on noisy sparse measurements of the field is depicted in (b), the computed error standard deviations are depicted in (c), and the difference between the error variances computed using the Krylov subspace estimation algorithm and the optimal ones computed using direct methods in MATLAB are depicted in (d).