Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2010, Article ID 508092, 19 pages
doi:10.1155/2010/508092

Review Article

Least Squares for Practitioners

J. A. Rod Blais

Department of Geomatics Engineering, Pacific Institute for the Mathematical Sciences, University of Calgary, 2500 University Drive N. W., Calgary, AB, Canada T2N 1N4

Correspondence should be addressed to J. A. Rod Blais, blais@ucalgary.ca

Received 13 May 2010; Accepted 16 August 2010

Academic Editor: Alois Steindl

In experimental science and engineering, least squares are ubiquitous in analysis and digital data processing applications. Minimizing sums of squares of some quantities can be interpreted in very different ways and confusion can arise in practice, especially concerning the optimality and reliability of the results. Interpretations of least squares in terms of norms and likelihoods need to be considered to provide guidelines for general users. Assuming minimal prerequisites, the following expository discussion is intended to elaborate on some of the mathematical characteristics of the least-squares methodology and some closely related questions in the analysis of the results, model identification, and reliability for practical applications. Examples of simple applications are included to illustrate some of the advantages, disadvantages, and limitations of least squares in practice. Concluding remarks summarize the situation and provide some indications of practical areas of current research and development.

1. Introduction

Least squares go back to Gauss and Legendre in the late 1790s. The first important publication on the topic was authored by Legendre in 1806 with the title "New Methods for Determination of a Comet's Orbit" and had a supplement entitled "On the method of least squares". Gauss's first publication on least squares appeared in 1809 at the end of his Theoria motus. He mentioned there in passing that Legendre had presented the method in his work of 1806, but that he himself had already discovered it in 1795. Gauss's correspondence and the papers found after his death proved that he was certainly the first to make the discovery, but since Legendre was first to publish it, priority rights belong to the latter. Obviously, as both of them reached the result independently of each other, both deserve the honour [1].

In these astronomical applications, "least squares" was a method of obtaining the best possible average value for a measured magnitude, given several observations of the

magnitude, when the measurements are found to be unavoidably different due to (random) errors. Of course, long before the theory of errors ever saw the light of day, common sense had chosen the arithmetic average value as the most probable value, hence the dichotomy in numerous situations: independently of the nature of the errors involved, a least-squares procedure gives the arithmetic mean, or more generally some weighted average value, which may or may not be the most likely value in the probabilistic sense (see, e.g., [2] for more general discussions). These two very different interpretations of least squares, now technically often referred to as the Best Approximation Estimate (BAE) in terms of quadratic norms, and the Maximum Likelihood Estimate (MLE) in terms of distributions, respectively, are widely used in functional analysis, inference analysis, and all kinds of application areas. Notice that in general, BAEs in terms of arbitrary norms can be very different from MLEs in terms of distributions. However, it is well known that in general, BAEs for p norms and MLEs for exponential distributions coincide with most interesting implications in practice (see, e.g., [3]). The following discussions will concentrate on the fundamental characteristics of the least-squares methodology and related implementation aspects.

Least-squares parameter estimation can be applied to underdetermined just as to overdetermined linear problems. In fact, underdetermined prediction problems are generally more common than overdetermined filtering and adjustment problems. With observations and unknown parameters of unequal weights modeled using some empirical or theoretical covariance (or correlation) functions, more sophisticated estimation methods such as Kriging and least-squares collocation are employed. Correspondingly, Radial Basis Functions (RBFs) and related strategies are used for interpolations of spatially scattered data and other approximations as BAEs.

The preceding implicit assumption of model linearity is essential for practical reasons. In practice, just about any engineering problem nonlinear in terms of its unknown parameters can be linearized as follows:

(a) Using a Taylor expansion about an appropriate point or parameter value, the linear term can be used to approximate the model in the neighborhood of the expansion point.

(b) Using differentiation in terms of the unknown parameters, the total derivative of the function is linear in terms of the differentials corresponding to the unknown parameters, and hence differential corrections to the unknown parameters can be evaluated as a linear problem.

These two strategies for nonlinear least squares then imply iterative procedures for differential correction estimates to the unknown parameters. Convergence of such iterative procedures is usually ensured for well-chosen parametrization, and the general situation has been discussed by Pope [4] and others in the estimation literature. Obviously, in general, there are numerous types of nonlinear complex problems that cannot be treated with such simplistic strategies but for the purposes of the following discussions, linearity in terms of the unknown parameters will be assumed from here on.

From an application perspective, one exceptional class of separable nonlinear least-squares problems deserves mention in this context. These are problems for which the mathematical model function is a linear combination of nonlinear functions. Specifically, one can assume that there are two sets of unknown parameters where one set is dependent on the other and can be explicitly eliminated. The method of variable projections has proven very appropriate for such nonlinear problems in several application areas [5]. General

nonlinear least-squares estimation is still the object of current research (see, e.g., [6] for further discussions and references).

The intrinsic linearity of least-squares computations implies that these can be done simultaneously or in a stepwise manner to obtain exactly the same estimation results. In other words, at the limit, one unknown parameter can be estimated at a time, or one observation or measurement can be processed at a time. This characteristic of linear computations is most useful in least-squares procedures and has led to numerous formulations such as summation of normals, and sequential adjustments. Furthermore, the quadratic computations can be avoided in critical numerically sensitive situations using orthogonal methods such as Givens rotations, Householder's reflections, and others.

2. Least Squares and Alternatives

Consider a system of M linear algebraic nonhomogeneous equations with N unknowns x_1, x_2, x_3,..., x_N, where M is not necessarily equal to N,

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1N}x_N &= f_1,\\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2N}x_N &= f_2,\\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3N}x_N &= f_3,\\
&\vdots\\
a_{M1}x_1 + a_{M2}x_2 + a_{M3}x_3 + \cdots + a_{MN}x_N &= f_M
\end{aligned} \tag{2.1}$$

with corresponding matrix representation

$$\mathbf{A}\mathbf{x} = \mathbf{f}, \tag{2.2}$$

and assuming $\mathbf{f} \neq \mathbf{0}$ for simplicity. When $M = N$ without any rank deficiencies in the matrix $\mathbf{A}$, that is, with $\mathbf{A}$ nonsingular, then the unique solution is simply

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{f}, \tag{2.3}$$

which can be evaluated in practice by Gaussian elimination or any other solution method for simultaneous linear equations. Notice that such solution methods are usually more efficient than the direct matrix inversion method, which is a consideration in numerous application contexts.

When the system is overdetermined with $M > N$, that is, more equations than unknowns, then one could use the first N equations, or some other selection of N equations, and assuming no rank deficiency, proceed as in the previous case of $M = N$. However, this is not appropriate for most applications as all the observations should somehow contribute to some "optimal" solution. Hence, rewriting the given system of equations with an error term $\mathbf{e}$, that is, $\mathbf{A}\mathbf{x} = \mathbf{f} + \mathbf{e}$, to emphasize that there may not exist one $\mathbf{x}$ value that would satisfy $\mathbf{A}\mathbf{x} = \mathbf{f}$ exactly, one obvious strategy is to minimize some norm of $\mathbf{e}$, that is, some acceptable

measure of the "length" of the vector $\mathbf{e}$. In practical terms, this norm of $\mathbf{e}$ can simply be its Euclidean length, that is,

$$\begin{aligned}\|\mathbf{e}\|_2 &= \left(e_1^2 + e_2^2 + e_3^2 + \cdots + e_M^2\right)^{1/2} \\ &= \left(|e_1|^2 + |e_2|^2 + |e_3|^2 + \cdots + |e_M|^2\right)^{1/2},\end{aligned} \tag{2.4}$$

but more generally, using p norms, denoted by L_p,

$$\|\mathbf{e}\|_p = \left(\sum_{i=1}^{M} |e_i|^p\right)^{1/p} \tag{2.5}$$

for $p = 1, 2, \ldots, \infty$. The solution $\mathbf{x}$ for a specified value of p, if one exists, is called an L_p BAE of $\mathbf{x}$. For $p = 2$, the L_2 estimate is the familiar least-squares estimate of $\mathbf{x}$, which is going to be discussed below. When $p = 1$, the L_1 estimate is a least-magnitude estimate of $\mathbf{x}$, a generalization of the median, and is well known in robust estimation. When $p = \infty$, the L_∞ estimate is a least-maximum or min-max estimate of $\mathbf{x}$. For other values of p, some BAEs are possible but not often used in practice, except perhaps for $1 < p < 2$ in multifacility location-allocation problems (e.g., [7]). Notice that for $p \neq 2$, the BAE does not necessarily exist and when it does exist, it is not necessarily unique, which can greatly complicate matters in applications.

When $p = 2$, the least-squares estimate always exists for a finite set of linear equations, assuming linearly independent columns, and its unique value is easily obtained using basic calculus:

$$\|\mathbf{e}\|_2^2 = \mathbf{e}^T\mathbf{e} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_M^2 \tag{2.6}$$

using matrix notation and to minimize $\mathbf{e}^T\mathbf{e} = (\mathbf{A}\mathbf{x} - \mathbf{f})^T(\mathbf{A}\mathbf{x} - \mathbf{f})$,

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{e}^T\mathbf{e}\right) = \frac{\partial}{\partial \mathbf{x}}\left((\mathbf{A}\mathbf{x} - \mathbf{f})^T(\mathbf{A}\mathbf{x} - \mathbf{f})\right) = \mathbf{0}, \tag{2.7}$$

which gives the familiar normal equations

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{f}, \tag{2.8}$$

where the square matrix $\mathbf{A}^T\mathbf{A}$ is easily seen to be symmetric and positive definite. The least-squares estimate is then written as

$$\widehat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{f} \tag{2.9}$$

which is easily verified to correspond to a minimum as

$$\frac{\partial^2}{\partial \mathbf{x}^2}\left(\mathbf{e}^T\mathbf{e}\right) = \frac{\partial^2}{\partial \mathbf{x}^2}\left((\mathbf{A}\mathbf{x}-\mathbf{f})^T(\mathbf{A}\mathbf{x}-\mathbf{f})\right) = 2\mathbf{A}^T\mathbf{A} > \mathbf{0}. \tag{2.10}$$

The previous matrix inequality simply means that, as the matrix $\mathbf{A}^T\mathbf{A}$ is symmetric and positive definite, it has positive real eigenvalues and hence the situation corresponds to a minimum of $\mathbf{e}^T\mathbf{e}$, as desired.

For the corresponding underdetermined system $\mathbf{A}\mathbf{x} = \mathbf{f}$, $\mathbf{f} \neq \mathbf{0}$, assuming linearly independent rows, there are obviously infinitely many solutions in general. For an optimal solution $\hat{\mathbf{x}}$ with minimum quadratic norm, the easiest approach is to use unknown correlates $\mathbf{x} = \mathbf{A}^T\mathbf{y}$ which imply by substitution

$$\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{A}^T\mathbf{y} = \mathbf{f} \tag{2.11}$$

and assuming no rank deficiency as before, $\mathbf{A}\mathbf{A}^T$ is nonsingular and hence

$$\mathbf{y} = \left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{f}, \tag{2.12}$$

which gives by substitution

$$\hat{\mathbf{x}} = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{f}. \tag{2.13}$$

To see the appropriateness of this estimate $\hat{\mathbf{x}}$, consider

$$\begin{aligned}
\mathbf{x} - \hat{\mathbf{x}} &= \mathbf{x} - \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{f} \\
&= \mathbf{x} - \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{A}\mathbf{x} \\
&= \left[\mathbf{I} - \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{A}\right]\mathbf{x} \\
&= \mathbf{N}_\mathbf{A}\mathbf{x},
\end{aligned} \tag{2.14}$$

with $\mathbf{N}_\mathbf{A}$ usually called the nullspace projector corresponding to $\mathbf{A}$. More explicitly,

(i) for a vector $\mathbf{z}$ with $\mathbf{A}\mathbf{z} = \mathbf{0}$, $\mathbf{N}_\mathbf{A}\mathbf{z} = \mathbf{z}$, and conversely;

(ii) $\mathbf{A}\mathbf{N}_\mathbf{A}\mathbf{z} = \mathbf{0}$ for all vector $\mathbf{z}$.

Therefore, for any vector $\mathbf{z}$,

$$\mathbf{A}(\hat{\mathbf{x}} + \mathbf{N}_\mathbf{A}\mathbf{z}) = \mathbf{A}\hat{\mathbf{x}} + \mathbf{A}\mathbf{N}_\mathbf{A}\mathbf{z} = \mathbf{f} \tag{2.15}$$

but for a minimum quadratic norm estimate $\hat{\mathbf{x}}$, only $\mathbf{z} \equiv \mathbf{0}$ is acceptable.

The previous result can readily be generalized to the quadratic norm with weight matrices for correlated observations or measurements of different quality as follows. Let $\mathbf{f}$ denote an M nonzero vector and $\mathbf{A}$ an $M \times N$ matrix with linearly independent columns. Then there is a unique N vector $\hat{\mathbf{x}}$ which minimizes $\{(\mathbf{f} - \mathbf{A}\mathbf{x})^T \mathbf{P}(\mathbf{f} - \mathbf{A}\mathbf{x})\}^{1/2}$ over all $\mathbf{x}$, for some appropriate weight matrix $\mathbf{P}$. Furthermore, $\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{P}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{P}\mathbf{f}$.

More generally, the unknowns themselves may have different relevance or other characteristics requiring weight matrices such as for regularization. Letting $\mathbf{f}$ denote an M nonzero vector and $\mathbf{A}$ an $M \times N$ matrix with linearly independent columns, then there is a unique N vector $\hat{\mathbf{x}}$ which minimizes $\{(\mathbf{f} - \mathbf{A}\mathbf{x})^T \mathbf{P}(\mathbf{f} - \mathbf{A}\mathbf{x}) + \mathbf{x}^T\mathbf{Q}\mathbf{x}\}^{1/2}$ over all $\mathbf{x}$, for some appropriate weight matrices $\mathbf{P}$ and $\mathbf{Q}$. Furthermore, $\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{P}\mathbf{A} + \mathbf{Q})^{-1}\mathbf{A}^T\mathbf{P}\mathbf{f}$ and when $\mathbf{P}$ and $\mathbf{Q}$ are nonsingular, $\hat{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T + \mathbf{P}^{-1})^{-1}\mathbf{f}$, by algebraic duality.

The proof of this last statement is a straightforward generalization of the previous situation involving the weighted observational errors and prior information about the unknown parameters. Using the Matrix Inversion Lemma (also called the Schur Identity), one can write

$$\begin{aligned}\hat{\mathbf{x}} &= \left(\mathbf{A}^T\mathbf{P}\mathbf{A} + \mathbf{Q}\right)^{-1}\mathbf{A}^T\mathbf{P}\mathbf{f} \\ &= \mathbf{Q}^{-1}\mathbf{A}^T\left(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T + \mathbf{P}^{-1}\right)^{-1}\mathbf{f},\end{aligned} \tag{2.16}$$

when the inverses of the weight matrices exist. Notice that the first RHS expression is in terms of the weight matrices while the second is in terms of their inverses, and that the preceding least-squares estimates are obtained with $\mathbf{P} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{0}$ for a simple overdetermined system, and with $\mathbf{Q}^{-1} = \mathbf{I}$ and $\mathbf{P}^{-1} = \mathbf{0}$ for a simple underdetermined system.

There is also an extensive theory dealing with the cases of rank deficiency in the matrix $\mathbf{A}$ implying a singular matrix $\mathbf{A}^T\mathbf{A}$ or $\mathbf{A}\mathbf{A}^T$ in the above expressions. In such cases, special precautions are required to reduce the number of parameters to be estimated or constrain their estimation. In general, the singular value decomposition discussed in the next section provides the best general strategy for linear problems with rank deficiencies.

In geometrical terms, the least-squares approach corresponds to an approximation using a normal (i.e., orthogonal) projection, and as a BAE, it does not necessarily involve any statistical information. In other words, as shown explicitly above, in all cases of underdetermined, determined, and overdetermined situations, even with weights associated with the observations and parameters, the least-squares solution is simply a weighted average of the observations for each unknown parameter. This is the reason for considering the least-squares approach basically as a mathematical approximation procedure which turns out to be most appropriate for statistical applications (see, e.g., [2] for more general discussions).

However, when interpreting the measurements or observations with finite first and second moments as a Gaussian sample, the average and hence the least-squares estimate becomes the unbiased minimum-variance estimate or the MLE. This is because any sample with finite first and second moments may be identified with a Gaussian sample as the Gaussian or normal distribution is fully specified by the first two moments. This statistical interpretation of least-squares estimates is really useful for error analysis and reliability considerations as in addition to Gaussian implications for the first moment, the second moment information behaves as a Chi-Square (χ^2) distribution. Gaussian statistics are widely

used in the analysis of least-squares estimates largely because of the well-developed theory and wide-ranging practical experience.

The previously introduced weight matrices $\mathbf{P}$ and $\mathbf{Q}$ are usually interpreted in the statistical sense as inversely proportional to the covariance matrices of the measurements and unknown parameters, respectively. Using unit proportionality factors, these are explictly

$$\mathbf{P} = \left\{\mathbf{E}\left[(\mathbf{e}-\mathbf{e}_0)(\mathbf{e}-\mathbf{e}_0)^T\right]\right\}^{-1}, \qquad \mathbf{Q} = \left\{\mathbf{E}\left[(\mathbf{x}-\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^T\right]\right\}^{-1}, \tag{2.17}$$

in which the zero subscript corresponds to the mean or expected value. Notice that $\mathbf{e}_0 = \mathbf{E}[\mathbf{e}] = \mathbf{0}$ for unbiasedness while $\mathbf{x}_0 = \mathbf{E}[\mathbf{x}]$ is not necessarily zero in general applications. Using the well-known *covariance propagation law*; that is, for any linear transformation $\mathbf{z} = \mathbf{R}\mathbf{x}$, one has for the corresponding second moment

$$\mathbf{E}\left[(\mathbf{z}-\mathbf{z}_o)(\mathbf{z}-\mathbf{z}_o)^T\right] = \mathbf{R}\mathbf{E}\left[(\mathbf{x}-\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^T\right]\mathbf{R}^T \tag{2.18}$$

then the variance of the estimated parameters is readily obtained as

$$\begin{aligned}\mathbf{E}\left[(\hat{\mathbf{x}}-\mathbf{x}_0)(\hat{\mathbf{x}}-\mathbf{x}_0)^T\right] &= \left(\mathbf{A}^T\mathbf{P}\mathbf{A}+\mathbf{Q}\right)^{-1} \\ &= \mathbf{Q}^{-1}-\mathbf{Q}^{-1}\mathbf{A}^T\left(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T+\mathbf{P}^{-1}\right)^{-1}\mathbf{A}\mathbf{Q}^{-1},\end{aligned} \tag{2.19}$$

which again shows a formulation in terms of the weights $\mathbf{P}$ and $\mathbf{Q}$, and a dual formulation in terms of their inverses. In practical applications, these two equivalent formulations can be exploited to minimize the computational efforts either in terms of weight (or information) matrices or in terms of covariance matrices. Notice that adding a diagonal term to an arbitrary matrix can be interpreted in several different ways to numerically stabilize the matrix inversion such as, for example, in ridge estimation [8], Tikhonov regularization [9], and variations thereof.

All nonzero covariance matrices and their inverses, the nonzero weight (or information) matrices, are symmetric and positive definite in least-squares estimation. The optimality of the estimates in the sense of minimum variances requires such symmetry and positive definiteness for all the nonzero weight and covariance matrices involved. In some applications, it may be required to control the dynamic range and spectral shape of the covariance of the estimation error and to that end, such methods as Covariance Shaping Least-Squares Estimation [10] can be used advantageously in practice.

For illustration purposes, consider the following situation. Given five measurements $\{(1,5),(2,7),(3,13),(4,8),(5,6)\}$ with some a priori information about a desired quadratic regression model, that is, $y = a+bx+cx^2$, with unknown parameters a, b, and c, the preceding least-squares formulations can be used to provide the estimates for the quadratic polynomial in the interior of the interval spanned by the observations used (see Table 1) for interpolation purposes. Notice that the mathematical model, namely, the quadratic polynomial, was assumed from the context of the experiment or exercise. In general, the model identification needs to be resolved and the situation will be briefly discussed in Section 6.

Table 1: Estimation of quadratic polynomials for interpolation purposes.

Observations Used	Estimated quadratic polynomial
(1, 5)	$\widehat{p}_1(x) = 1.66667 + 1.66667x + 1.66667x^2$ near $x = 1$
(1, 5), (2, 7)	$\widehat{p}_2(x) = 3.00000 + 2.00000x + 0.00000x^2$ for x in $(1,2)$
(1, 5), (2, 7), (3, 13)	$\widehat{p}_3(x) = 7.00000 - 4.00000x + 2.00000x^2$ for x in $(1,3)$
(1, 5), (2, 7), (3, 13), (4, 8)	$\widehat{p}_4(x) = -4.25000 + 10.25000x - 1.75000x^2$ for x in $(1,4)$
(1, 5), (2, 7), (3, 13), (4, 8), (5,6)	$\widehat{p}_5(x) = -2.60000 + 8.44286x - 1.35714x^2$ for x in $(1,5)$

3. Least-Squares Interpolation and Prediction

Simple linear interpolation consists in an arithmetic mean of N quantities, that is,

$$\widehat{u} = \frac{1}{N}\sum_{i=1}^{N} u_i = \sum_{i=1}^{N} \lambda_i u_i \tag{3.1}$$

with all the coefficients $\lambda_i \equiv 1/N$. In a more general context, the coefficients λ_i would be estimated to optimize the interpolation or prediction. For instance writing $f_0 = f(x_0)$ and for the observations, $f_i = f(x_i)$, $i = 1,N$, one has the general interpolation formula

$$f_0 = \sum_{i=1}^{N} p_i f_i, \quad \text{assuming } \sum_{i=1}^{N} p_i = 1, \tag{3.2}$$

that is, with normalized weights $p_1, \ldots, p_N$. In the presence of correlations between the observations, using the usual matrix notation,

$$\begin{aligned} f_0 &= \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} p_{11} & \cdots & p_{1N} \\ \vdots & \vdots & \vdots \\ p_{N1} & \cdots & p_{NN} \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix} \\ &= \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \vdots & \vdots \\ c_{N1} & \cdots & c_{NN} \end{pmatrix}^{-1} \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix}, \end{aligned} \tag{3.3}$$

assuming the weight matrix $\mathbf{P} = [p_{ij}]$ corresponds to the inverse of the covariance matrix $\mathbf{C} = [c_{ij}]$.

Correspondingly, the least-squares prediction formula

$$\widehat{f}_0 = \begin{pmatrix} c_{01} & \cdots & c_{0N} \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \vdots & \vdots \\ c_{N1} & \cdots & c_{NN} \end{pmatrix}^{-1} \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix}, \tag{3.4}$$

in which the quantities c_{ij} are defined in terms of the correlations between f_i and f_j, including f_0, assuming the expected value of f to be zero, $\mathbf{E}[f(x)] = 0$. Notice that this expression is of the form of the least-squares solution to an underdetermined linear problem as discussed before. For prediction applications, the correlation terms are often modelled empirically using correlation functions of the separation distance $d_{ij} = \|x_i - x_j\|$, $i, j = 0, \ldots, N$.

When $\mathbf{E}[f(x)]$ is an unknown constant α, the normal equations for the unknown coefficients $\lambda_1, \ldots, \lambda_N$ and α are written explicitly as

$$\begin{pmatrix} c_{11} & \cdots & c_{1N} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ c_{N1} & \cdots & c_{NN} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \alpha \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_N \\ 0 \end{pmatrix} \tag{3.5}$$

for some unknown Lagrange multiplier α. In general when $\mathbf{E}[f(x)]$ is modelled as a kth degree polynomial, these normal equations become

$$\left(\begin{array}{cccccc} c_{11} & \cdots & c_{1N} & 1 & \cdots & x^k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{N1} & \cdots & c_{NN} & 1 & & x^k \\ \hline 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x^k & \cdots & x^k & 0 & \cdots & 0 \end{array}\right) \left(\begin{array}{c} \lambda_1 \\ \vdots \\ \lambda_N \\ \hline \alpha_0 \\ \vdots \\ \alpha_k \end{array}\right) = \left(\begin{array}{c} f_1 \\ \vdots \\ f_N \\ \hline 0 \\ \vdots \\ 0 \end{array}\right) \tag{3.6}$$

for some unknown coefficients $\alpha_0, \ldots, \alpha_k$.

Least-squares prediction problems can be classified differently depending on application and other considerations. In different contexts, deterministic and probabilistic interpretations are used and hence the inferences are different. Three methodologies are mentioned here.

(a) Kriging methods with variograms or generalized covariance functions such as, $\mathrm{Cov}(d) = -d$ or d^3 for spatial distance d (see [11] for details).

(b) Radial Basis Function (RBF) methods with empirical RBFs as weighing functions, such as, $\mathrm{RBF}(r) = r^2 \log r$ or $(r^2 + c^2)^{1/2}$ for radial distance r and constant c (see, e.g., [12] for more details).

(c) Least-Squares Collocation (LSC) methods with ordinary covariance functions $\mathbf{C} \equiv \mathbf{C}_s + \mathbf{C}_n$ with $\mathbf{C}_s$ denoting the signal part and $\mathbf{C}_n$ denoting the noise part of the covariance matrix $\mathbf{C}$.

Furthermore, generalized covariance functions (that is with positive power spectra) include ordinary covariance functions and empirical RBFs that can often be interpreted as covariance or correlation functions. Notice that the nonsingularity of the normal equation matrices is always assumed to guarantee a solution without additional constraints.

4. Solution Methodology for Normal Equations

The normal equation matrices $\mathbf{A}^T\mathbf{A}$ or $\mathbf{A}\mathbf{A}^T$ and the like are positive definite symmetric matrices that lend themselves to $\mathbf{L}^T\mathbf{L}$ or $\mathbf{L}\mathbf{L}^T$ decompositions in terms of lower triangular matrices $\mathbf{L}$. The best known decomposition algorithm for such a matrix is the Cholesky square-root algorithm which is usually applied simultaneously to the normal equation matrix $\mathbf{C}$ and right-hand side vector (or column matrix) $\mathbf{F}$ as follows:

$$\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1N} & f_1 \\ c_{21} & c_{22} & \cdots & c_{2N} & f_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} & f_N \end{pmatrix} \xrightarrow[\text{square-root}]{\text{Cholesky's}} \begin{pmatrix} c'_{11} & c'_{12} & \cdots & c'_{1N} & f'_1 \\ 0 & c'_{22} & \cdots & c'_{2N} & f'_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c'_{NN} & f'_N \end{pmatrix}. \tag{4.1}$$

The solution is then obtained as a back substitution in the resulting upper triangular system of equations. The computational effort for N normal equations is approximately $1/3$ of the effort $O(N^3)$ required using inverse matrix strategies. Furthermore in the case of least-squares adjustments of blocks of stereomodels, photographs and networks of geodetic stations, the normal equation matrix is usually banded to $B \ll N$ and the Cholesky's algorithm only requires $O(NB^2)$ in such cases. This is really advantageous and has been used in large geodetic network, photogrammetric block adjustments, and most other similar least-squares applications.

Furthermore, in the Cholesky square-root reduction of a normal equation matrix to an upper (or lower) triangular matrix, the numerical conditioning is usually monitored by the magnitude of the computed diagonal elements as these should remain positive for a positive definite symmetric matrix. In some applications, the procedure of monitoring the magnitude of the diagonal elements of the triangular matrix is used to decide on an optimal order for a polynomial model which may be in terms of complex variables. Other similar strategies for numerical analysis of least squares problems based on the triangular decomposition of the normal equations matrix will be mentioned below.

However as mentioned in the previous section, the least-squares prediction problems have a "normal" matrix of the general form

$$\begin{pmatrix} \mathbf{C} & \mathbf{D} & \mathbf{F} \\ \mathbf{D}^T & \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{4.2}$$

with the matrix $\mathbf{C}$ symmetric and positive definite, $\mathbf{D}$ rectangular, and $\mathbf{0}$ denoting a zero matrix. Such a "normal" matrix is obviously nonpositive definite and hence is not really appropriate for the previous triangular matrix representation. However, the Cholesky's square-root procedure can be applied to the first N equations and then Givens rotations can be applied to transform the remaining M rows into the upper triangular form. Givens rotations are applied to two row vectors at any one time to eliminate the first nonzero element of the second row vector thus transforming the system of equations into an upper triangular

system for back substitution at any time. Explicit implementation details can be found in [13] and elsewhere. Graphically, the situation is as follows:

$$\left(\begin{array}{cccccccc} c_{11} & c_{12} & \cdots & c_{1N} & d_{11} & \cdots & d_{1M} & f_1 \\ c_{21} & c_{22} & \cdots & c_{2N} & d_{21} & \cdots & d_{2M} & f_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} & d_{N1} & \cdots & d_{NM} & f_N \\ \hline d_{11} & d_{21} & \cdots & d_{N1} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{1M} & d_{2M} & \cdots & d_{NM} & 0 & \cdots & 0 & 0 \end{array}\right)$$

$$\xrightarrow{\substack{\text{Cholesky's} \\ \text{square-root}}} \left(\begin{array}{cccccccc} c'_{11} & c'_{12} & \cdots & c'_{1N} & d'_{11} & \cdots & d'_{1M} & f'_1 \\ 0 & c'_{22} & \cdots & c'_{2N} & d'_{21} & \cdots & d'_{2M} & f'_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c'_{NN} & d'_{N1} & \cdots & d'_{NM} & f'_N \\ \hline d_{11} & d_{21} & \cdots & d_{N1} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{1M} & d_{2M} & \cdots & d_{NM} & 0 & \cdots & 0 & 0 \end{array}\right) \tag{4.3}$$

$$\xrightarrow{\substack{\text{Givens'} \\ \text{Rotations}}} \left(\begin{array}{cccccccc} c''_{11} & c''_{12} & \cdots & c''_{1N} & d''_{11} & \cdots & d''_{1M} & f''_1 \\ 0 & c''_{22} & \cdots & c''_{2N} & d''_{21} & \cdots & d''_{2M} & f''_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c''_{NN} & d''_{N1} & \cdots & d''_{NM} & f''_N \\ \hline 0 & 0 & \cdots & 0 & e''_{11} & \cdots & e''_{1M} & g''_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & e''_{MM} & g''_M \end{array}\right).$$

Notice that Givens rotations could be applied to the full $N + M$ equations but the preceding strategy is superior in terms of computational efficiency. Givens rotations have excellent numerical stability characteristics but often require slightly more computational efforts than the alternatives.

5. Singular Value Decomposition

For indepth analysis of least-squares results and other related applications, the Singular Value Decomposition (SVD) approach is essential in practice and a brief overview follows. Considering the preceding (rectangular $M \times N$) matrix $\mathbf{A}$, its SVD gives

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T, \tag{5.1}$$

where $\mathbf{U}$ is the matrix of (unit) eigenvectors of $\mathbf{A}\mathbf{A}^T$, $\mathbf{V}$ is the matrix of (unit) eigenvectors of $\mathbf{A}^T\mathbf{A}$, and $\mathbf{\Lambda}$ is an $M \times N$ matrix with diagonal elements (called the singular values) equal to the square roots of the nonzero eigenvalues of $\mathbf{A}^T\mathbf{A}$ or $\mathbf{A}\mathbf{A}^T$. By substitution, one readily obtains

$$\begin{aligned}
\mathbf{A}\mathbf{A}^T &= \left(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right)\left(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right)^T \\
&= \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T \\
&= \mathbf{U}\left(\mathbf{\Lambda}\mathbf{\Lambda}^T\right)\mathbf{U}^T, \\
\mathbf{A}^T\mathbf{A} &= \left(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right)^T\left(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right) \\
&= \mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \\
&= \mathbf{V}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}\right)\mathbf{V}^T,
\end{aligned} \tag{5.2}$$

where $\mathbf{\Lambda}\mathbf{\Lambda}^T$ and $\mathbf{\Lambda}^T\mathbf{\Lambda}$ denote the diagonal matrices of the squares of the singular values of $\mathbf{A}$ and dimensions $M \times M$ and $N \times N$, respectively. The last step in both derivations follows from the orthogonality of the (unit) eigenvectors in $\mathbf{U}$ and $\mathbf{V}$, respectively. Their inverses are, respectively,

$$\begin{aligned}
\left(\mathbf{A}\mathbf{A}^T\right)^{-1} &= \mathbf{U}\left(\mathbf{\Lambda}\mathbf{\Lambda}^T\right)^{-1}\mathbf{U}^T, \\
\left(\mathbf{A}^T\mathbf{A}\right)^{-1} &= \mathbf{V}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}\right)^{-1}\mathbf{V}^T,
\end{aligned} \tag{5.3}$$

as matrix inversion does not change the eigenvectors in the SVD of a symmetric matrix.

The previous least-squares solution for the overdetermined system $\mathbf{A}\mathbf{x} = \mathbf{f}$ is simply

$$\begin{aligned}
\widehat{\mathbf{x}} &= \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{f} \\
&= \mathbf{V}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}\right)^{-1}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{f} \\
&= \mathbf{V}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{f} \\
&= \mathbf{V}\mathbf{\Lambda}^{-}\mathbf{U}^T\mathbf{f}
\end{aligned} \tag{5.4}$$

with the special notation $\mathbf{\Lambda}^{-} = (\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T$ and for the underdetermined case

$$\begin{aligned}\widehat{\mathbf{x}} &= \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{f}\\ &= \mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{U}\left(\mathbf{\Lambda}\mathbf{\Lambda}^T\right)^{-1}\mathbf{U}^T\mathbf{f}\\ &= \mathbf{V}\mathbf{\Lambda}^T\left(\mathbf{\Lambda}\mathbf{\Lambda}^T\right)^{-1}\mathbf{U}^T\mathbf{f}\\ &= \mathbf{V}\mathbf{\Lambda}^{+}\mathbf{U}^T\mathbf{f}\end{aligned} \tag{5.5}$$

with the special notation $\mathbf{\Lambda}^{+} = \mathbf{\Lambda}^T(\mathbf{\Lambda}\mathbf{\Lambda}^T)^{-1}$. These $\mathbf{\Lambda}^{-}$ and $\mathbf{\Lambda}^{+}$ are usually called generalized inverses of $\mathbf{\Lambda}$.

From a computational perspective, for an overdetermined system with $M \gg N$, it may be more efficient to first perform a $\mathbf{QR}$ factorization of $\mathbf{A}$ with $\mathbf{Q}$ as an $M \times N$ matrix with orthogonal columns and an upper triangular matrix $\mathbf{R}$ of order N, and then compute the SVD of $\mathbf{R}$, since if $\mathbf{A} = \mathbf{QR}$ and $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, then the SVD of $\mathbf{A}$ is given by $\mathbf{A} = (\mathbf{QU})\mathbf{\Lambda}\mathbf{V}^T$. Similarly, for an underdetermined system with $M \ll N$, it may be more efficient to first perform an $\mathbf{LQ}$ factorization of $\mathbf{A}$ with a lower triangular matrix $\mathbf{L}$ of order M and an $M \times N$ matrix $\mathbf{Q}$ with orthogonal rows, and then compute the SVD of $\mathbf{L}$, since if $\mathbf{A} = \mathbf{LQ}$ and $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, then the SVD of $\mathbf{A}$ is given by $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}(\mathbf{Q}^T\mathbf{V})^T$. The SVD approach is often used in spectral analysis and computing a minimum norm solution for (possibly) rank-deficient linear least-squares and related problems. More discussion of the computational aspects can be found, for example, in [14].

In practice, the SVD of a matrix has been described as "one of the most elegant algorithms in numerical algebra for exposing quantitative information about the structure of a system of linear equations" [15]. In current data assimilation and prediction research using spatiotemporal processes such as in global change and other environmental applications, the SVD approach has become very important for at least three problem areas.

First, considering sequences of discrete data $\mathbf{x} = (x_1,\ x_2,\ x_3, \ldots, x_N)$ with zero mean for simplicity associated with discrete times $t_1,\ t_2, \ldots, t_M$, and written in matrix form as

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & \ldots & x_1^M \\ x_2^1 & x_2^2 & \ldots & x_2^M \\ \vdots & \vdots & \vdots & \vdots \\ x_N^1 & x_N^2 & \ldots & x_N^M \end{pmatrix} \tag{5.6}$$

in which the superscripts correspond to the times $t_1,\ t_2, \ldots, t_M$. Such a convention with columns corresponding to the data sequences at discrete times is quite common in environmental applications. Then a SVD of this data matrix $\mathbf{X}$ yields $\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T$, as discussed above. The columns of $\mathbf{U}$ are the *Empirical Orthogonal Functions* (EOFs) for the data matrix $\mathbf{X}$ while the columns of $\mathbf{V}$ are the corresponding *principal components*. The data transformation

$\mathbf{U}^T\mathbf{x}$ or more generally $\mathbf{U}^*\mathbf{x}$ for a (complex) data vector $\mathbf{x}$ is usually called a Karhunen-Loève transformation. Such resulting data sequences $\mathbf{z} = \mathbf{U}^T\mathbf{x}$ are easily seen to be uncorrelated as

$$\begin{aligned}\mathbf{E}\left[\mathbf{z}\mathbf{z}^T\right] &= \mathbf{E}\left[\mathbf{U}^T\mathbf{x}\mathbf{x}^T\mathbf{U}\right] = \mathbf{U}^T\mathbf{E}\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{U} \\ &= \mathbf{U}^T\ \mathbf{C}_x\mathbf{U} = \mathbf{U}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{U} = \boldsymbol{\Lambda}\end{aligned} \tag{5.7}$$

which is most useful in practical applications. It is also important to notice that the EOFs can be described as eigenvectors of the corresponding covariance matrix of the available data. These are often called *normal modes* of the measured spatiotemporal process [16]. Since the (power) spectrum of a data sequence is well known to correspond to the spectrum of its (auto) covariance matrix, such normal modes have interesting interpretations in the context of dynamical systems driven by noise (e.g., [17]). Further discussions can be found in [18] and the references therein. An example of simulated application is shown in Figure 1(a) with a spatial pattern of a box followed by two cones over a sinusoidal path, with the resulting time series in Figure 1(b) of 20 occurrences of the pattern of 36 observations. Figure 1(c) shows the first modal (spatial) pattern and the corresponding first singular values which are identical to the input information except for the different scaling in amplitude and sign convention of Mathematica 7 [19]. Analogous simulations can readily be done in two dimensions with various patterns (see [20]). Such simulations show the potential of this methodology in the analysis of environmental and other geophysical time series.

Second, in numerical conditioning analysis for any linear algebraic system of equations, the singular values of the matrix give most relevant information about the propagation of numerical errors from observations to estimated unknown parameters. For instance, considering some symmetric and positive definite matrix $\mathbf{B}$, consider the linear algebraic system

$$\mathbf{B}\mathbf{u} = \mathbf{v}, \tag{5.8}$$

and for some small perturbations $\boldsymbol{\delta}\mathbf{u}$ and $\boldsymbol{\delta}\mathbf{v}$ such that

$$\mathbf{B}\boldsymbol{\delta}\mathbf{u} = \boldsymbol{\delta}\mathbf{v}, \tag{5.9}$$

then using the spectral norm, it is well known that

$$\frac{\|\boldsymbol{\delta}\mathbf{u}\|}{\|\mathbf{u}\|} \leq \|\mathbf{B}\| \cdot \left\|\mathbf{B}^{-1}\right\| \cdot \frac{\|\boldsymbol{\delta}\mathbf{v}\|}{\|\mathbf{v}\|} = \frac{\lambda_{\max}(\mathbf{B})}{\lambda_{\min}(\mathbf{B})} \cdot \frac{\|\boldsymbol{\delta}\mathbf{v}\|}{\|\mathbf{v}\|}, \tag{5.10}$$

in which $\lambda_{\max}(\mathbf{B})/\lambda_{\min}(\mathbf{B}) \equiv \kappa(\mathbf{B})$ is the condition number of the matrix $\mathbf{B}$ in terms of its maximum and minimum eigenvalues, $\lambda_{\max}(\mathbf{B})$ and $\lambda_{\min}(\mathbf{B})$, respectively. This provides a powerful tool for the analysis of relative changes in the unknown parameters implied by some relative perturbations in the observations.

Third, as the nonzero eigenvalues of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are identical, computations, and numerical analysis in filtering and smoothing can take advantage of this fact in using the smaller matrix in the least-squares computations. In several areas of environmental research, enormous quantities of data and complex mathematical models lead to very large normal

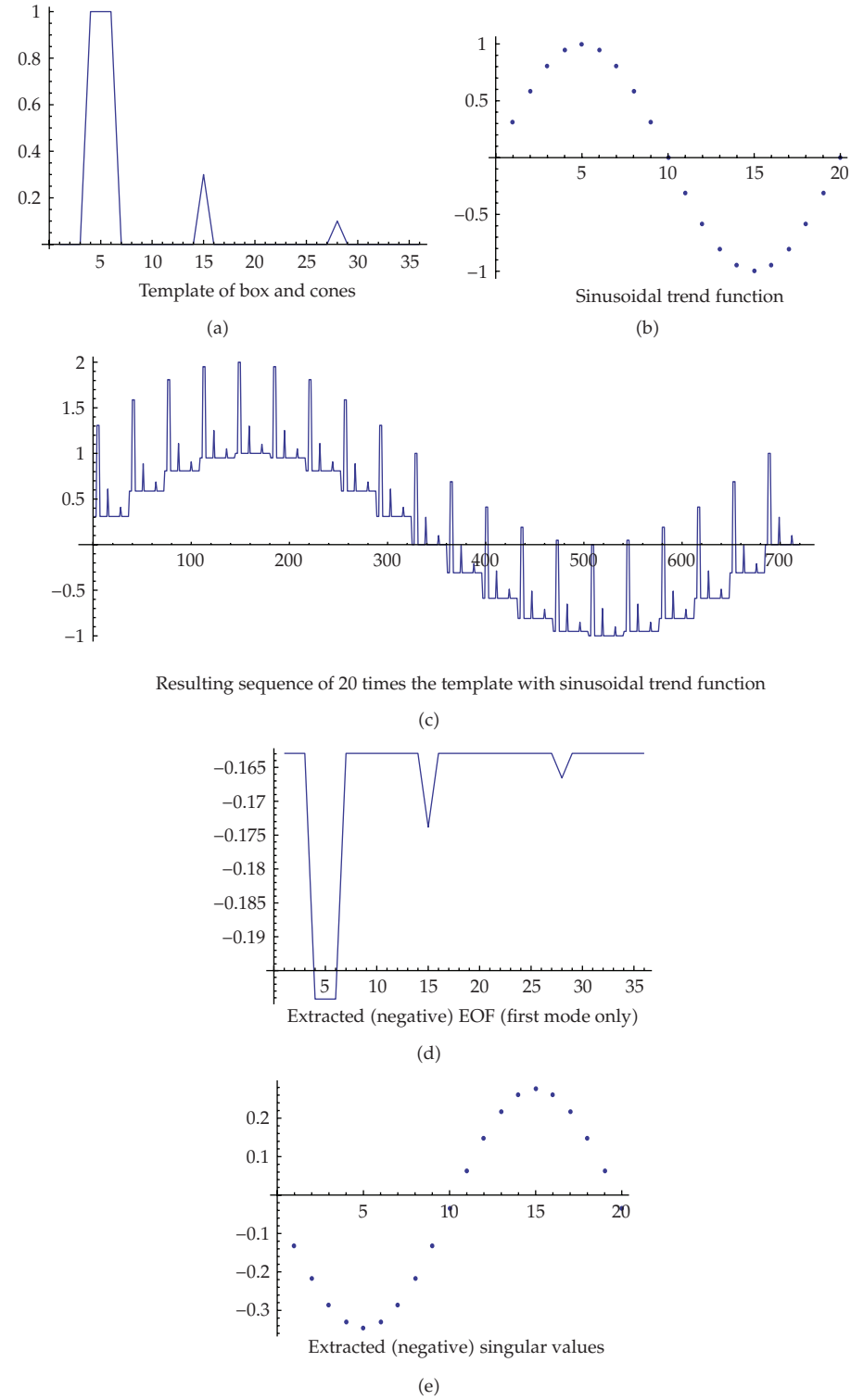

Figure 1: EOF application to template times sinusoidal trend function and extraction thereof.

equation systems that require much computational efforts. Substantial reductions in the Kalman filtering and data assimilation computations are implied by the proper choice of covariance/information formulations but for indepth error analysis of the results, SVD-based techniques become critical (see, e.g., [21]).

6. Model Identification and Reliability Considerations

As previously mentioned, alternatives to least squares (L_2), such as the least magnitude (L_1) and least maximum (L_∞), have advantages and disadvantages. First, the contributions of errors and especially outliers will increase in the estimation procedure from L_1 to L_∞. Therefore, in the absence of outliers, L_∞ would be best while in the presence of large errors, L_1 or robust estimation is more appropriate. However considering the linear normal equations with least squares, least squares are selected for most practical applications. Outliers are a well-known problem with least squares and much literature exists to mitigate the implications.

Furthermore, in numerous application contexts, assuming an algebraic formulation, one has to decide on degrees and orders in regression modeling such as in curve fitting and spectrum estimation. Considering a simple set of measurements or observations $\{(t_i, f(t_i));\ i = 1, 2, 3, \ldots, N\}$ to be modelled using an algebraic polynomial of the form $f(t) = a_0 + a_1 t + a_2 t^2 + \cdots$. If the degree of the polynomial is unknown, then the least-squares approach to estimating $a_0, a_1, a_2, \ldots$ can always achieve a perfect fit to the measurements by selecting the polynomial degree equal to $N - 1$. Actually, the variance of the residuals will decrease with higher and higher degrees to become zero for degree $N - 1$. Hence, the least-squares approach cannot be used to decide on the polynomial degree for such regression applications. Furthermore, the approach alone can hardly be used to decide on some other possible mathematical models such as $f(t) = b_0 + b_1 \cos(t) + b_2 \sin(t) + \cdots$ as additional information is necessary for model identification. Such model identification problems have been studied extensively by Akaike [22, 23] and others (see [24] for further discussions and references).

For example, the sample measurements $\{(1,5), (2,7), (3,13), (4,8), (5,6)\}$ with an algebraic polynomial of the form $y = a + bx + cx^2 + \cdots$ would lead to an exact fit for degree 4. However, such high degree would likely be unacceptable because of the oscillations between the data points which would imply uncertainty in any prediction. Considering the least-squares estimates for degrees 0, 1, 2, 3, and 4,

$$\begin{aligned}
\widehat{q}_0(x) &= 7.8,\\
\widehat{q}_1(x) &= 6.9 + 0.3x,\\
\widehat{q}_2(x) &= -2.6 + 8.44286x - 1.35714x^2,\\
\widehat{q}_3(x) &= -1.2 + 6.47619x - 0.60714x^2 - 0.08333x^3,\\
\widehat{q}_4(x) &= 51.0 - 91.9167x + 59.2917x^2 - 14.5833x^3 + 1.20833x^4
\end{aligned} \tag{6.1}$$

with corresponding error variances

$$\widehat{\sigma}_0^2 = 9.7, \quad \widehat{\sigma}_1^2 = 9.475, \quad \widehat{\sigma}_2^2 = 3.02857, \quad \widehat{\sigma}_3^2 = 3.00357, \quad \widehat{\sigma}_4^2 = 0.0. \tag{6.2}$$

However, considering the (normalized) Akaike Information Criterion (AIC), defined as

$$\mathrm{AIC} = \log \widehat{\sigma}^2 + \frac{2M + 3}{N} \tag{6.3}$$

for approximating M (model) parameters using N measurements, one obtains

$$3.27213,\quad 3.64866,\ 2.90809,\ 3.2998, \tag{6.4}$$

corresponding to each degree 0, 1, 2, and 3, respectively, implying an optimal degree 2 for the modeling as 2.90809 is the minimum AIC value. Other examples can be found in [24] and the references therein.

In this context of least squares, the assumption of finite first and second moments with Gaussian statistics interpretation have implications in terms of expected moments for the estimated parameters. Essentially, assuming a given mathematical model, the given variances for the measurements and/or observations and a priori variances for the unknown parameters can be propagated using the variance propagation law into the estimated parameters and interpreted at some confidence level such as 95%. This is the familiar approach in geomatics with error ellipses reflecting the accuracy of measurements and/or observations and the geometrical strength of a network in positioning.

For example, given positional information at two discrete points P_1 and P_2 with a variance σ^2, the mid point located by the arithmetic average of the coordinates of P_1 and P_2 has a predicted variance $\sigma^2/2$ when a linear model is known for any intermediate point. However, when the location and/or definition of the mid point is ambiguous or unknown, such as along some fuzzy line, its predicted uncertainty is likely to be much more than $\sigma^2/2$. In other words, the uncertainty in any estimation results is attributable to uncertainty in the assumed mathematical model and in the observational information used.

Another illustrative example is in the prediction of some quantity g, such as gravity, as a function of known data at the discrete points P_1 and P_2 with observational variance σ^2. At the mid point between P_1 and P_2, the average g value, that is, $[g(P_1) + g(P_2)]/2$, is usually an adequate prediction of the g value there but its variance is likely to be greater than $\sigma^2/2$. Otherwise, why bother with measurements! It should be noticed that in nonlinear and/or non-Gaussian situations, the error propagation is much more complex, and only numerical Monte Carlo simulations offer a general strategy for uncertainty modeling (see [25] for further details and references).

7. Concluding Remarks

Least squares are ubiquitous in applied science and engineering data processing. From a mathematical perspective, a least-squares estimate is a (weighted) mean solution which may be interpreted differently depending on the application context. Furthermore, any linear finite problem, even an ill-posed one, has a unique solution in the "average sense". Such a solution is a BAE using a minimum quadratic norm or minimum variance, with the flexibility of possible statistical interpretation as MLE for optimal and reliable predictions.

Advantages of the least-squares approach are essentially in the simple assumptions (i.e., finite first and second moments), the unique estimates from linear normal equations,

with excellent computational and applicability characteristics. Disadvantages of the least-squares approach are mainly in terms of oversmoothing properties (such as in curve or surface fitting) and relative overemphasis of outlier observations or measurements.

In terms of numerical computations, the least-squares approach has excellent characteristics in terms of stability and efficiency. This is best seen using the SVD approach for any indepth analysis of least-squares results. The readjustment of the geodetic networks in the North American Datum of 1983 has demonstrated that nearly one million unknowns can be handled reliably with only 32 bit arithmetic on conventional computer platforms (see [26, 27]). A better numerical approach would be difficult to find!

Furthermore, it is also important to emphasize that least squares are not appropriate for all types of estimation problems as there are numerous application contexts where a best observation or measurement value needs to be selected among the available ones (as the most frequent value or the one with minimum error). In other application contexts when dealing with observations or measurements likely to be affected by outliers, a more robust estimate such as a median value or L_1 estimate may be preferable. No single estimation method can be considered best or optimal for all applications as data characteristics and desired estimates need to be considered.

Finally, some areas of current research and development in least squares and computational analysis include multiresolution analysis and synthesis, data regularization and fusion, EOFs of multidimensional time sequences, RBFs, and related techniques for optimal data assimilation and prediction, especially in spatiotemporal processes. With scientific data generally considered to be increasing faster than computational power, real challenges in the analysis of current observations and measurements abound and strategies have to become more sophisticated.

Acknowledgments

The author would like to acknowledge the sponsorship of the Natural Science and Engineering Research Council in the form of a Research Grant on Computational Tools for the Geosciences. Comments and suggestions from the reviewers are also acknowledged with gratitude.

References

[1] T. Hall, *Carl Friedrich Gauss: A Biography*, Translated from the Swedish by Albert Froderberg, The MIT Press, Cambridge, Mass, USA, 1970.
[2] E. T. Jaynes, *Probability Theory*, Cambridge University Press, Cambridge, UK, 2003.
[3] A. Egger, "Maximum likelihood and best approximations," *The Rocky Mountain Journal of Mathematics*, vol. 20, no. 1, pp. 117–122, 1990.
[4] A. J. Pope, "Two approaches to nonlinear least squares adjustments," *The Canadian Surveyor*, vol. 28, no. 5, pp. 663–669, 1974.
[5] G. Golub and V. Pereyra, "Separable nonlinear least squares: the variable projection method and its applications," *Inverse Problems*, vol. 19, no. 2, pp. R1–R26, 2003.
[6] D. Pollard and P. Radchenko, "Nonlinear least-squares estimation," *Journal of Multivariate Analysis*, vol. 97, no. 2, pp. 548–562, 2006.
[7] H. D. Sherali and W. P. Adams, *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*, vol. 31 of *Nonconvex Optimization and Its Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.
[8] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[9] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, John Wiley & Sons, New York, NY, USA, 1977.
[10] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 686–697, 2003.
[11] J. A. R. Blais, "Generalized covariance functions and their applications in estimation," *Manuscripta Geodaetica*, vol. 9, no. 4, pp. 307–312, 1984.
[12] C. A. Micchelli, "Interpolation of scattered data: distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, no. 1, pp. 11–22, 1986.
[13] J. A. R. Blais, *Estimation and Spectral Analysis*, University of Calgary Press, Calgary, AB, Canada, 1988, http://www.netlibrary.com/.
[14] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, Pa, USA, 1996.
[15] V. C. Klema and A. J. Laub, "The singular value decomposition: its computation and some applications," *IEEE Transactions on Automatic Control*, vol. 25, no. 2, pp. 164–176, 1980.
[16] Gerald R. North, "Empirical orthogonal functions and normal modes," *Journal of the Atmospheric Sciences*, vol. 41, no. 5, pp. 879–887, 1984.
[17] R. W. Preisendorfer, *Principle Components and the Motions of Simple Dynamical Systems*, Scripps Institution of Oceanography, 1979, Ref. Ser. 70-11.
[18] K.-Y. Kim and Q. Wu, "A comparison study of EOF techniques: analysis of nonstationary data with periodic statistics," *Journal of Climate*, vol. 12, no. 1, pp. 185–199, 1999.
[19] Wolfram: Mathematica 7. Wolfram Research Inc., Champaign, Ill, USA, 2008.
[20] G. Eshel, "Geosci236: Empirical Orthogonal Functions," Technical Note, Department of the Geophysical Sciences, University of Chicago, 2005.
[21] D. Treebushny and H. Madsen, "A new reduced rank square root Kalman filter for data assimilation in mathematical models," in *Proceedings of the International Conference in Computational Science (ICCS '03)*, P. M. A. Sloot, D. Abramson, A. V. Bogdanov, J. C. Dongarra, A. Y. Zomaya, and Y. E. Gorbachev, Eds., vol. 2657 of *Lecture Notes in Computer Science*, pp. 482–491, 2003.
[22] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.
[23] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281, Akadémiai Kiadó, Budapest, Hungary, 1973.
[24] J. A. R. Blais, "On some model identification strategies using information theory," *Manuscripta Geodaetica*, vol. 16, no. 5, pp. 326–332, 1991.
[25] J. A. R. Blais, "Reliability considerations in geospatial information systems," *Geomatica*, vol. 56, no. 4, pp. 341–350, 2002.
[26] P. Meissl, "A Priori Prediction of Roundoff Error Accumulation in the Solution of a Super Large Geodetic Normal Equation System," NOAA Professional Paper 12, National Geodetic Information Branch, National Oceanic and Atmospheric Administration (NOAA), Rockville, Md, USA, 1980.
[27] C. R. Schwartz, Ed., North American Datum of 1983. NOAA Professional Paper NOS 2, National Oceanic and Atmospheric Administration (NOAA), US Department of Commerce, 1989.