Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 94515, 16 pages
doi:10.1155/2007/94515

Review Article

The Geometry of Statistical Efficiency and Matrix Statistics

K. Gustafson

Received 23 March 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

We will place certain parts of the theory of statistical efficiency into the author's operator trigonometry (1967), thereby providing new geometrical understanding of statistical efficiency. Important earlier results of Bloomfield and Watson, Durbin and Kendall, Rao and Rao, will be so interpreted. For example, worse case relative least squares efficiency corresponds to and is achieved by the maximal turning antieigenvectors of the covariance matrix. Some little-known historical perspectives will also be exposed. The overall view will be emphasized.

1. Introduction and Summary

Recently, Gustafson [1–3] was able to connect the theory of statistical efficiency to his operator trigonometry, which is a theory of antieigenvalues and antieigenvectors which he initiated in 1967 for a different purpose. The aim of this paper is to go beyond the [1–3] papers to provide a more overall view of these results and their implications. We will also use this opportunity to expose some historical perspectives that have been generally forgotten or which are otherwise little known.

The outline and summary of this paper are as follows. In Section 2, we obtain the statistical efficiency ratio of BLUE to OLSE covariance in terms of the geometry provided by the author's 1967 operator trigonometry. To fix ideas here, this result can be described as giving to the [4, 5] Bloomfield-Watson-Knott solution of the Durbin conjecture, that is, its geometrical meaning. In Section 3, we provide the reader with the basics of the operator trigonometry. This brief but adequate bibliographical citation is given from which further detail may be obtained. To augment the reader's intuition and appreciation for the operator trigonometry, and because we are writing here for an audience of statisticians, in Section 4 we recall the origin of the operator trigonometry, that is, operator semigroups, with application to Markov processes. This problem essentially induced both of the key elements of the operator trigonometry. In Section 5, we return to the topic of statistical efficiency and provide some lesser-known historical background. This is augmented in Section 6 with a look at an interesting early paper of von Neumann. From the latter, we are able to make here an interesting new connection of statistical efficiency to partial differential equations. In Section 7, we develop the interesting and useful distinction between what we call inefficiency vectors versus antieigenvectors. Both satisfy related variational equations. Through this link, we may then relate in Section 8 certain considerations of canonical correlations as treated in [6] by Rao-Rao to the general mathematical setting of statistical efficiency and operator trigonometry—all three are now combined. Section 9 concludes the paper with some further discussion of the historical view of statistical efficiency as viewed through the context of this paper.

2. The geometry of statistical efficiency

What follows was shown in Gustafson [1–3]. Considering the general linear model, we follow Wang and Chow [7] for convenience:

$$y = X\beta + e, \tag{2.1}$$

where y is an n-vector composed of n random samplings of a random variable Y, X is an $n \times p$ matrix usually called the design or model matrix, β is a z-vector composed of p unknown nonrandom parameters to be estimated, and e is an n-vector of random errors incurred in observing y. The elements x_{ij} of X may have different statistical meanings depending on the application. We assume for simplicity that the error or noise e has expected value 0 and has covariance matrix $\sigma^2 V$, where V is a symmetric positive definite $n \times n$ matrix. Of course one can generalize to singular V and to unknown V and so on by using singular value decomposition and generalized inverses throughout to develop a more general theory, but we shall not do so here. We absorb the σ^2 or nonidentical row-dependent variances into V. A customary assumption on X is that $n \geqq 2p$, that is, one often thinks of X as having only a few (regressor) columns available. In fact, it is useful to often think of p as just 1 or 2. Generally, it seems to be usually assumed that the columns of X are linearly independent, and often it is assumed that those columns form an orthonormal set $X^*X = I_p$.

The relative statistical efficiency for comparing an ordinary least-squares estimator OLSE $\hat{\beta}$ and the best linear unbiased estimator BLUE β^* is defined as

$$\mathrm{RE}\,(\hat{\beta}) = \frac{|\operatorname{Cov}(\beta^*)|}{|\operatorname{Cov}(\hat{\beta})|} = \frac{1}{|X^*VX|\,|X^*V^{-1}X|}, \tag{2.2}$$

where $|\cdot|$ denotes determinant. A fundamental lower bound for statistical efficiency is

$$\mathrm{RE}\,(\hat{\beta}) \geqq \prod_{i=1}^{p} \frac{4\lambda_i\lambda_{n-i+1}}{(\lambda_i + \lambda_{n-i+1})^2}, \tag{2.3}$$

where $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n > 0$ are the eigenvalues of V. This lower bound is sometimes called the Bloomfield-Watson-Knott lower bound; see Section 5 for more historical particulars. In Gustafson [1], the following new and geometrical interpretation of the lower bound (2.3) was obtained. More specifics of the operator trigonometry, antieigenvalues, and antieigenvectors will be given in Section 3. The essential meaning of Theorem 2.1 is that the linear model's statistical efficiency is limited by the maximal turning angles of the covariance matrix V.

Theorem 2.1. *For the general linear model (2.1) with SPD covariance matrix $V > 0$, for $p = 1$, the geometrical meaning of the relative efficiency (2.2) of an OLSE estimator $\widehat{\beta}$ against BLUE β^* is*

$$\mathrm{RE}\,(\widehat{\beta}) \geqq \cos^2 \phi(V), \tag{2.4}$$

where $\phi(V)$ is the operator angle of V. For $p \leqq n/2$, the geometrical meaning is

$$\mathrm{RE}\,(\widehat{\beta}) \geqq \prod_{i=1}^{p} \cos^2 \phi_i(V) = \prod_{i=1}^{p} \mu_i^2(V), \tag{2.5}$$

where the $\phi_i(V)$ are the successive decreasing critical turning angles of V, that is, corresponding to the higher antieigenvalues $\mu_i(V)$. The lower bound (2.3), as expressed geometrically in (2.4), is attained for $p = 1$ by either of the two first antieigenvectors of V:

$$x_{\pm} = \pm\left(\frac{\lambda_1}{\lambda_1 + \lambda_n}\right)^{1/2} x_n + \left(\frac{\lambda_n}{\lambda_1 + \lambda_n}\right)^{1/2} x_1. \tag{2.6}$$

For $p \leqq n/2$, the lower bound (2.3), as expressed geometrically in (2.5), is attained as

$$\prod_{i=1}^{p} \frac{\langle V x_{\pm}^i, x_{\pm}^i \rangle}{\|V x_{\pm}^i\| \|x_{\pm}^i\|}, \tag{2.7}$$

where $x_{\pm}^i$ denotes the ith higher antieigenvectors of V given by

$$x_{\pm}^i = \pm\left(\frac{\lambda_i}{\lambda_i + \lambda_{n-i+1}}\right)^{1/2} x_{n-i+1} + \left(\frac{\lambda_{n-i+1}}{\lambda_i + \lambda_{n-i+1}}\right)^{1/2} x_i. \tag{2.8}$$

In (2.6) and (2.8), x_i denotes the normalized ith eigenvector of V corresponding to the eigenvalue λ_i.

We remark that Theorem 2.1 follows rather immediately from (2.3) once one recognizes that the factors on the right-hand side of (2.3) are exactly the cosines of the critical turning angles of V. This connection was first pointed out in Gustafson [1]. In Gustafson [3], some related trace statistical efficiency bounds were also given an operator trigonometric interpretation.

3. The operator trigonometry: antieigenvalues and angles

For simplicity, let A be an $n\times n$ symmetric positive definite (SPD) matrix with eigenvalues $0<\lambda_n \leqq \lambda_2 \leqq \cdots \leqq \lambda_1$. Then, the first antieigenvalue of A was defined to be

$$\mu_1 = \min_{x\neq 0} \frac{\langle Ax, x\rangle}{\|Ax\|\,\|x\|} \tag{3.1}$$

and a related entity

$$\nu_1 = \min_{\epsilon>0} \|\epsilon A - I\| \tag{3.2}$$

also came naturally into the theory. How that came about will be described in Section 4. Because of the need for both μ_1 and ν_1, the author felt that ν_1 must also be trigonometric. Indeed it is. Gustafson [8] established the following key minmax result.

Theorem 3.1. *Given a strongly accretive operator B on a Hilbert space, then*

$$\sup_{\|x\|\leqq 1} \inf_{\epsilon} \|(\epsilon B - I)x\|^2 = \inf_{\epsilon>0} \sup_{\|x\|\leqq 1} \|(\epsilon B - I)x\|^2. \tag{3.3}$$

In particular for an SPD matrix A, one has

$$\mu_1^2 + \nu_1^2 = 1. \tag{3.4}$$

Originally, the minimum (3.1) was called $\cos A$ for obvious reasons, and after Theorem 3.1 was realized, the minimum (3.2) could be called $\sin A$. This is an essential critical point to understand about the operator trigonometry. One must have both a $\sin A$ and a $\cos A$ if one wants some kind of trigonometry. Later, the better notations $\cos\phi(A)$ and $\sin\phi(A)$ were introduced so as to avoid any unwarranted confusion with cosine and sine functions in an operator's functional calculus. Moreover, it is clear that A does have a meaningful operator angle $\phi(A)$ defined equivalently by either (3.1) or (3.2). This operator maximal turning angle $\phi(A)$ is a real tangible angle in n-dimensional Euclidean space. It is attained by A's two (here normalized to norm 1) antieigenvectors:

$$x_\pm = \pm\left(\frac{\lambda_1}{\lambda_1+\lambda_n}\right)^{1/2} x_n + \left(\frac{\lambda_n}{\lambda_1+\lambda_n}\right)^{1/2} x_1, \tag{3.5}$$

where x_1 and x_n are any (normalized) eigenvectors from the eigenspaces corresponding to λ_1 and λ_n, respectively. The antieigenvectors are those that are turned to the maximal amount when operated on by A, and they thus attain the minimums in (3.1) and (3.2).

A more general theory has been developed, and for that and further history and other ramifications of the operator trigonometry and antieigenvalue-antieigenvector theory, we just refer to the books of Gustafson [9], Gustafson and Rao [10], and the surveys of Gustafson [11, 2]. One more basic ingredient which should be mentioned here is the Euler equation

$$2\|Ax\|^2\|x\|^2(\operatorname{Re}A)x - \|x\|^2\operatorname{Re}\langle Ax,x\rangle A^*Ax - \|Ax\|^2\operatorname{Re}\langle Ax,x\rangle x = 0 \tag{3.6}$$

which is satisfied by the antieigenvectors of A, for any strongly accretive matrix A. When A is Hermitian or normal, this Euler equation is satisfied not only by the first antieigenvectors $x_\pm$ of A, but also by all eigenvectors of A. Thus, the expression (3.1) generalizes the usual Rayleigh quotient theory for SPD matrices A to now include antieigenvectors $x_\pm$, which minimize it, and all eigenvectors, which maximize it.

Higher antieigenvalues $\mu_i(A)$ and their corresponding higher antieigenvectors were originally defined, Gustafson [12], in a way analogous to that for higher eigenvalues in the Rayleigh-Ritz theory. That is okay for some applications but later, Gustafson [13], the author formulated a better general combinatorially based theory in which the higher antieigenvectors are those stated in (2.8). To each such pair, we obtain via (3.1) a sequence of decreasing-in-size maximal interior operator turning angles $\phi_i(V)$ as indicated in (2.5) (see Gustafson [14] for more details).

It is interesting to note that antieigenvectors, including the higher ones, always occur in pairs. In retrospect, this is a hint that there are connections of that fact to the fact that the usual analyses of statistical efficiency also often end up at a point where one needs to consider certain pairs of vectors. We will return to this point in Section 7.

4. The origin of the operator trigonometry: Markov processes

The author's creation of the operator trigonometry in 1967 came out of an abstract operator theoretic question. Let X be a Banach space and let A be the densely defined infinitesimal generator of a contraction semigroup e^{tA} on X. In other words, consider the initial value problem

$$\begin{gathered}\frac{du}{dt} = Au(t), \quad t > 0, \\ u(0) = u_0 \quad \text{given}\end{gathered} \tag{4.1}$$

and its solution $u(t) = U_t u_0 \equiv e^{tA} u_0$ with the contraction property $\|U_t\| \leqq 1$. So one can think of the heat equation, the Schrödinger equation, or a linear Markov process. In fact, it was a question of introducing a stochastic time change into a Markov process e^{tA}, which led to the following question. When can one multiplicatively perturb A to BA and still retain the contraction semigroup infinitesimal generator property in BA? The result was as follows, Gustafson [15], stated here in now familiar terms.

Theorem 4.1. *Let A be the infinitesimal generator of a contraction semigroup on a Banach space X. Then, BA is still an infinitesimal generator of a contraction semigroup if B is a strongly accretive operator satisfying*

$$\sin\phi(B) \leqq \cos\phi(A). \tag{4.2}$$

But the proof of Theorem 4.1 in Gustafson [15] did not originally involve any entity $\sin\phi(B)$ because such entities did not exist yet. The proof instead needed $\|\epsilon B - I\| \leqq \mu_1(A)$ for some positive ϵ. By the minmax Theorem 3.1, this requirement becomes (4.2).

Therefore, to better understand these now trigonometric entities, the author quickly computed them for some operator classes. For the most definitive and most useful class A a SPD matrix with eigenvalues $0 < \lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_1$, one has

$$\cos\phi(A) = \frac{2\sqrt{\lambda_1\lambda_n}}{\lambda_1+\lambda_n}, \qquad \sin\phi(A) = \frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}, \tag{4.3}$$

which are attained by the antieigenvector pair (3.5).

5. Some history of statistical efficiency

Although the theory of statistical efficiency is well documented in a number of books, and in the 1970's papers of Bloomfield-Watson [4], Knott [5], and others, in the writing of Gustafson [1] this author wanted to get some original feel of the history for himself. For one thing, it was wondered where the "Durbin conjecture" which led to the lower bound (2.3) was explicitly stated. This was not found. But some related historical perspectives were put into Gustafson [1, 3, Section 4]. There, for example, one finds a description of precursor work of Plackett [16], Aitken [17], and Durbin and Kendall [18]. The latter paper is quite explicitly geometrical, although, not operator theoretically. Plackett [16] takes the fundamental notions all the way back to Gauss.

A second more recent historical look has revealed some further interesting historical perspectives. In particular, the Watson [19] paper is probably the explicit source of the "Durbin conjecture." In fact, one finds it there, (3.5), with a footnote crediting it to J. Durbin. However, Watson [20] admits a flaw in his [19] argument and thus the verification of the Durbin conjecture remained an open problem until 1975.

Going back further to the two papers of Durbin-Watson [21, 22], one finds a more classical statistical analysis of (2.1) from the point of view of χ^2 distributions, which is of course of central importance to the theory of analysis of variance. In particular, the second paper is largely devoted to a study of the statistic

$$d = \frac{\sum(\Delta z)^2}{\sum z^2} \tag{5.1}$$

which is to be used for testing for serial correlation within error in terms of a regression model. We go back to the first paper (see [21, page 409]) and find that the principal issue is "the problem of testing the errors for independence forms the subject of this paper and its successor." Attribution is made to earlier papers by Anderson [23] and Anderson and Anderson [24], where possible serial correlations in least-squares residuals from Fourier regressions were tested. In Watson [20], which is a quite useful paper historically, study of the efficiency of least squares is said to follow that of Grenander [25] and Grenander and Rosenblatt [26]. In fact, we have traced efficiency explicitly back to Fisher [27]. See our further discussion in Section 9.

6. The von Neumann connection and a new connection to partial differential equations

In our historical search, tracing back through the two papers of Durbin and Watson [21, 22], one comes upon the interesting $n \times n$ matrix

$$A = \begin{bmatrix} 1 & -1 & 0 & \cdots & & 0 \\ -1 & 2 & -1 & \cdots & & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & & & \cdots & & 0 \\ & & & -1 & 2 & -1 \\ 0 & & & & -1 & 1 \end{bmatrix}. \tag{6.1}$$

It is stated there that this results from the statistic to be used to test for serial correlation

$$d = \frac{\sum(\Delta z)^2}{\sum z^2} = \frac{\langle Az, z\rangle}{\sum z^2}, \tag{6.2}$$

where z is the residual from linear regression. It was shown [22] that the mean and variance of the statistic d are given by

$$\begin{aligned} E(d) &= \frac{P}{n-k'-1}, \\ \operatorname{var}(d) &= \frac{2[Q-PE(d)]}{(n-k'-1)(n-k'+1)}, \end{aligned} \tag{6.3}$$

where

$$\begin{aligned} P &= \operatorname{tr} A - \operatorname{tr}\left(X'AX(X'X)^{-1}\right), \\ Q &= \operatorname{tr} A^2 - 2\operatorname{tr}\left(X'A^2X(X'X)^{-1}\right) + \operatorname{tr}\left((X'AX(X'X)^{-1}\right)^2, \end{aligned} \tag{6.4}$$

where k' is the number of columns of the matrix of observations of the independent variables

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k'1} \\ \vdots & & & \\ x_{1n} & x_{2n} & \cdots & c_{k'n} \end{bmatrix}. \tag{6.5}$$

One wonders, or at least this author wonders, about how A came about. It turns out that this query became quite interesting as we now explain.

A more careful reading of Durbin and Watson [21] leads to a paper of von Neumann [28], and one cannot resist looking at it. As it is well known, von Neumann was a polymath and this paper is not an exception. An in-depth study of the statistic

$$\eta = \frac{\delta^2}{s^2} \tag{6.6}$$

is carried out, where s^2 is the sample variance of a normally distributed random variable and $\delta^2 = \sum_{\mu=1}^{n-1}(x_{\mu+1} - x_\mu)^2/(n-1)$ is the mean square successive difference—the goal being to determine the independence or trend dependence of the observations $x_1,\dots,x_n$. Thus, we find this paper to be an early and key precedent to all the work done by Durbin, Watson, and others in the period of 1950–1975.

Von Neumann's analysis is extensive and he obtains a number of theoretical results which, if we might paraphrase see Durbin and Watson [21, page 418], are more or less beyond use by conventional statisticians. However, both Durbin-Watson papers [21, 22] go ahead and use the matrix A to illustrate their theory. So one looks further into von Neumann's paper to better understand the origin of the matrix A of (6.1). One finds there (see [28, page 367]) the statement: "the reasons for the study of the distribution of the mean square successive difference δ^2, in itself as well as in its relationship to the variance s^2, have been set forth in a previous publication, to which the reader is referred." However, it is made clear that comparing observed values of the statistic η will be used to determine "whether the observations $x_1,\dots,x_n$ are independent or whether a trend exists."

Since curiosity knows no bounds, we pushed the historical trace back to the previous publication of von Neumann, Kent, Bellison, and Hart [29]. The answer to our curiosity about why von Neumann became involved with this statistical regression problem is found there. To quote (see [29, page 154]), "the usefulness of the differences between successive observations only appears to be realized first by ballisticians, who faced the problem of minimizing effects due to wind variation, heat, and wear in measuring the dispersion of the distance traveled by shell." The 4-author paper originated from the Aberdeen Ballistic Research Laboratory, where von Neumann was consulting.

Returning to his analysis in von Neumann [28], we find that he begins with a now more or less classical multivariate analysis of normally distributed variables. By diagonalization, a quadratic form $\sum A_\mu x'_\mu$ is obtained where the A_μ, $\mu = 1,\dots,n$, are the eigenvalues of the form $(n-1)\delta^2$. The smallest eigenvalue $A_n = 0$ is found, with eigenvector $x_0 = (1,\dots,1)/\sqrt{n}$. A further analysis, using an interesting technique of assuming the $x'_1,\dots,x'_{n-1}$ to be uniformly distributed over an $n-1$ unit sphere, shows that the statistic η of (6.5) is then distributed according to

$$\eta = \frac{n}{n-1}\sum_{\mu=1}^{n-1} A_\mu x_\mu^2. \tag{6.7}$$

Thus, the sought eigenvalues A_μ, $\mu = 1,\dots,n$, are the eigenvalues of the quadratic form $(n-1)\delta^2$, which is then written as

$$(n-1)\delta^2 = x_1^2 + 2\sum_{\mu=2}^{n-1} x_\mu^2 + x_n^2 - 2\sum_{\mu=1}^{n-1} x_\mu x_{\mu+1}. \tag{6.8}$$

The matrix of this form is (6.1) and it is that matrix which is also borrowed and used in Durbin and Watson [21, 22]. Used as well are the eigenvalues

$$A_k = 4\sin^2\left(\frac{k\pi}{2n}\right), \quad k = 1,\dots,n-1 \tag{6.9}$$

which von Neumann computes from the determinant of A.

Commentary. When we first saw the matrix A in Durbin and Watson [21, 22], our take was completely different. As this author is a specialist in partial differential equations, for example, see Gustafson [30], we immediately see the matrix A in (6.1) as the discretized Poisson-Neumann boundary value problem

$$\begin{aligned} -\frac{d^2u(x)}{dx^2} &= f(x), \quad 0 < x < 1, \\ \frac{du}{dx} &= 0 \quad \text{at } x = 0, 1. \end{aligned} \tag{6.10}$$

In saying this, I am disregarding the exact interval and discrete Δx sizes.

This new connection between statistical efficiency and partial differential equations will be further explored elsewhere, especially as it will no doubt generalize to Dirichlet, Neumann, and Robin boundary value problems for the Laplacian operator $-\Delta = \sum \partial^2 u/\partial x^2$ in higher dimensions. The reverse implications for a more general context of statistical efficiency could also be interesting. Moreover, we have already worked out the complete operator trigonometry for the two-dimensional discretized Dirichlet problem in Gustafson [31].

We also comment in passing that a similar ballistic problem—that of control of rocket flight—was the motivating application in Japan during the Second World War that led Ito to develop his stochastic calculus, which is now so important in the theory of financial derivatives and elsewhere.

7. The inefficiency equation and the Euler equation

Following Wang and Chow [7], among others, one may apply a Lagrangian method to

$$\mathrm{RE}\,(\hat{\beta})^{-1} = |XV^{-1}X|\,|X'VX| \tag{7.1}$$

with the general case having been reduced to that of $X'X = I_p$. By a differentiation of $F(x,\lambda) = \ln|X'V^{-1}X| + \ln|X'VX| - 2\,\mathrm{tr}(X'X\Lambda)$ and subsequent minimization, the relation

$$X'X(\Lambda + \Lambda') = \Lambda + \Lambda' = 2I_p \tag{7.2}$$

is obtained. Here, Λ is a $p\times p$ upper triangular matrix which is the Lagrange multiplier with respect to the constraint $X'X = I_p$. From this and further work including the simultaneous diagonalization of $X'V^2X$, $X'VX$, and $X'V^{-1}X$, one arrives at the result

$$\mathrm{RE}\,(\hat{\beta})^{-1} = \prod_{i=1}^{p} x_i' V x_i x_i' V^{-1} x_i, \tag{7.3}$$

where X is now the $n\times p$ column matrix $X = [(x_1)\cdots(x_p)]$ whose columns go into the expression (7.3). The Lagrange multiplier minimization leading to (7.3) has also now yielded the equation for the x_i:

$$\frac{V^2 x_i}{x_i' V x_i} + \frac{x_i}{x_i' V^{-1} x_i} = 2Vx_i, \quad i = 1,\dots,p. \tag{7.4}$$

Clearly, the span $\{x_i, Vx_i\}$ is a two- (or one-)dimensional reducing subspace of V and it is spanned by two (or one) eigenvectors ψ_j and ψ_k of V. Writing each column $x_i = \sum_{j=1}^{n} \alpha_{ij}\psi_j$ in terms of the full eigenvector basis of V, (7.4) yields the quadratic equation

$$\frac{z^2}{x_i' V x_i} - 2z + \frac{1}{x_i' V^{-1} x_i} = 0 \tag{7.5}$$

for the two (or one) eigenvalues λ_j and λ_k associated to each x_i, $i = 1,\dots,p$. Substituting those eigenvalues as found from (7.5) into (7.3) brings (7.3) to the statistical efficiency lower bound (2.3).

On the other hand, the Euler equation (3.6) from the operator trigonometry, for $n\times n$ SPD matrices A, becomes

$$\frac{A^2x}{\langle A^2x, x\rangle} - \frac{2Ax}{\langle Ax, x\rangle} + x = 0. \tag{7.6}$$

Comparison of (7.5), which we call the inefficiency equation, and the Euler equation (7.6) yield the following result.

Theorem 7.1. *For any $n\times n$ SPD covariance matrix V or more generally any $n\times n$ SPD matrix A, all eigenvectors x_j satisfy the inefficiency equation (7.4) and the Euler equation (7.6). The only other vectors satisfying the inefficiency equation (7.4) are the "inefficiency vectors"*

$$x_{\pm}^{j+k} = \pm\frac{1}{\sqrt{2}}x_j + \frac{1}{\sqrt{2}}x_k, \tag{7.7}$$

where x_j and x_k are any eigenvectors corresponding to any distinct eigenvalues $\lambda_j \neq \lambda_k$. The only other vectors satisfying the Euler equation (7.6) are the antieigenvectors

$$x_{\pm}^{jk} = \pm\left(\frac{\lambda_k}{\lambda_j+\lambda_k}\right)^{1/2} x_j + \left(\frac{\lambda_j}{\lambda_j+\lambda_k}\right)^{1/2} x_k. \tag{7.8}$$

For details of the proof of Theorem 7.1, see Gustafson [1, 3].

Commentary. The statistical interpretation of relative statistical inefficiency of an OLSE estimator $\widehat{\beta}$ in terms of (2.2) is that the design matrix X chosen for (2.1) unfortunately contains columns of the form (7.7). That is why we called those the inefficiency vectors of V. The most critical are of course those with $j = 1$ and $k = n$. On the other hand, the new geometrical interpretation of relative statistical inefficiency of an OLSE estimator $\widehat{\beta}$, now in terms of the bound (2.3) as seen trigonometrically according to Theorem 2.1, is now in the worst-case situation; the matrix X under consideration unfortunately contains columns of the form (7.8). These antieigenvectors represent the critical turning angles of the covariance matrix V. The worst case is when $j = 1$ and $k = n$.

8. Canonical correlations and Rayleigh quotients

The Euler equation for the antieigenvectors can be placed (at least in the case of A symmetric positive definite) within a context of stationary values of products of Rayleigh quotients. To do so, we refer to the paper of Rao and Rao [6], and references therein. If one considers the problem of obtaining the stationary values of an expression

$$\frac{x'Cx}{(x'Ax)^{1/2}(x'Bx)^{1/2}} \tag{8.1}$$

with A and B being symmetric positive definite and C being symmetric, then squaring (8.1) gives the product of two Rayleigh quotients

$$\frac{\langle Cx,x\rangle}{\langle Ax,x\rangle}\cdot\frac{\langle Cx,x\rangle}{\langle Bx,x\rangle}. \tag{8.2}$$

Taking the functional derivative of (8.1) with respect to x yields the equation

$$\frac{x'Cx}{x'Ax}Ax+\frac{x'Cx}{x'Bx}Bx = 2Cx. \tag{8.3}$$

Note that if we let $C = T$, $A = T^2$, $B = 1$, then (8.1) becomes the antieigenvalue quotient (3.1). Similarly, (8.3), for the same operators and with x being normalized to $\|x\| = 1$, becomes the Euler equation (7.6). On the other hand, the full Euler equation (3.6) for any bounded accretive operator A on any Hilbert space is more general than (8.3) in the sense of operators treated. Moreover, one can easily put B and C operators into the coefficients by a similar derivation. Thus, a general theory encompassing statistical efficiency, operator trigonometry, and canonical correlations could be developed.

Commentary. In their analysis, Rao and Rao [6] arrive at two cases, the first case corresponds to stationary values equal to 1, and the second case corresponds to smaller stationary values. As regards the second case, they note that “there can be solutions of the form $x = ae_i + be_j$,” where the e_i and e_j are eigenvectors. But we now know from the operator trigonometry that these are the two cases covered by our Euler equation (3.6), and that the solutions in the second case are the antieigenvectors.

9. Concluding discussion

Who first formulated that the definition $\mathrm{RE}(\hat{\beta})$ of statistical efficiency was not clear to this author. Durbin and Kendall [18], certainly two great veterans in the field, specifically define E to be the efficiency of t' relative to t according to (see [18, page 151])

$$\rho(t,t') = \sqrt{\frac{\operatorname{var} t}{\operatorname{var} t'}} = \sqrt{E}. \tag{9.1}$$

Here $t = \sum_{j=1}^{n} \lambda_j x_j$ is a linear estimator of the mean. To be unbiased, the coefficients λ_j must satisfy $\sum \lambda_j = 1$. The variance of the estimator t is then $\sigma^2 \sum \lambda_j^2 = \sigma^2 (OP)^2$, where OP is the line segment from the origin to the $\sum \lambda_j = 1$ hyperplane in λ-space. Clearly, the smallest such variance arrives when one takes the point P to be the bottom of the line segment perpendicular to the hyperplane. Variance of t' is just $\sigma^2 (OP')^2$ for any other point P' in the hyperplane. So $E = \cos\phi$, where ϕ is the angle between the lines OP and OP'.

Durbin and Kendall [18] cite the book of Cramér [32] for statistical efficiency. There [32, Chapter 32, page 474], Cramér makes it clear that "in the sequel, we shall exclusively consider the measures of dispersion and concentration associated with the variance and its multidimensional generalizations." Then see [32, page 481], the efficiency $e(\alpha^*)$ is defined to be the ratio between the variance $D^2(\alpha^*)$ of an unbiased and regular estimate α^* and its smallest possible value

$$\frac{1}{n\int_{-\infty}^{\infty} \left(\frac{\partial \log f}{\partial \alpha}\right)^2 f\,dx}. \tag{9.2}$$

Here, $f(x,\alpha)$ is a continuous frequency function. The discrete case is also worked out in later pages. Cramér attributes the concept of efficient estimate to Fisher [27, 33]. Also mentioned (see [32, page 488]) are (later) papers by Neyman, Pearson, and Koopman. So the theory of statistical efficiency arises centrally out of the general theory of estimation of variance by maximum likelihood methods, and it seems, from the early days of that development.

In Freund's classic textbook (Miller and Miller [34]), one finds (see page 327) that the fact that $\operatorname{var}(\hat{\theta}) \geqq$ the quantity in (9.2) is called the Cramér-Rao inequality. The denominator of (9.2) is interpreted as the information about the estimator θ which is supplied by the sample. Smaller variance is interpreted to mean greater information. Thus, as Cramér already made clear (see our quote above and Chapter 32 of his book), we are looking at central tendency as measured by second moments.

We decided to bite the bullet and go back to Fisher [27, 33]. Indeed, in his first paper see [27, page 309], he clearly defines efficiency of a statistic as "the ratio whose intrinsic accuracy bears to that of the most efficient statistic possible; it expresses the proportion of the total available relevant information of which that statistic makes use." He carefully attributes, designates, or, in any case, cites in connection with that definition a 1908 paper by Student and a 1763 paper by Bayes. Then, we find (on page 315) that "in 1908, Student broke new ground by calculating the distribution of the ratio which the deviation of the mean from its population value bears to the standard deviation calculated from the sample." Of course, both papers [27, 33] also contain excellent discussions of the method of maximum likelihood and its pros and cons.

Here, this author must interject that in a classified naval intelligence task, in 1959, he first became aware of, and implemented, the χ^2 distribution for estimating goodness of fit for combinations of normally distributed random variables. The application was concerned with observations at several receiving sites of the bearings of received signal from a transmitting enemy submarine. For an unclassified account of this work, see the paper of Gustafson [35]. This author still remembers the genuine joy of operational naval personnel as they called out that "the χ^2 of the fit is...!" It is also perhaps an amusing irony that 45 years later this author, through the indirect and abstract path of his operator trigonometry, has arrived back at χ^2 testing.

A second point for discussion is that in this treatment, we have not gone into the more general theory of statistical efficiency utilizing generalized inverses. Certainly, it is natural and essential to do so for both theory and statistical applications. For example, when V is nonsingular, one has (e.g., see Puntanen and Styan [36]) in terms of generalized inverses

$$\begin{aligned}\text{BLUE}(X\beta) &= X\beta^* = X(X^*V^{-1}X)^-X^*V^{-1}y,\\ \text{OLSE}(X\beta) &= X\hat{\beta} = X(X^*X)^-X^*y.\end{aligned} \tag{9.3}$$

However, in this author's opinion, the essential points are first seen for $p = 1$, that is, in the case of X, a single regressor vector. In any case, the more general theory including generalized inverses is now so well worked out in the mathematical statistics literature that such a state of affairs should excuse the author from having to process it all. On the other hand, it is equally clear that the operator trigonometry of statistical efficiency should be extended to that setting including generalized inverses and, moreover, singular correlation matrices V. Possibly, we shall do that in the future, but such a comprehensive study is a task for another paper.

However, we here may "close the picture" from the other direction. From the usual assumption $X^*X = I_p$, where X is an $n \times p$ semi-unitary matrix, it is instructive to take its p orthonormal columns and conceptually add to them $n - p$ orthonormal columns. These may be thought of as "fictitious" additional regressors that one would like to have. How to do so is just the procedure in the proof of the classical Schur theorem. Call any one of these enlarged unitary regressor matrices X. Then, (9.3) is simplified to

$$\text{BLUE}\,(X\beta^*) = X^{-1}y, \qquad \text{OLSE}\,(X\hat{\beta}) = y. \tag{9.4}$$

Also, the efficiency (2.2) becomes 1, caused essentially by the unitarity of X. Although this exercise should not surprise anyone, still it seems to this author that the generalized inverse theory could be viewed as an "intermediate" theory dealing with how badly you have truncated and otherwise abused the fictitiously available large set of Schur unitaries. As a variation on this theme, for an arbitrary $n \times n$ matrix X written in its polar form $X = U|X|$, where U is the isometry from the range of the absolute value operator $|X|$ to the range of X, the operator trigonometry concerns itself only with the turning angles of the Hermitian polar factor $|X|$. See Gustafson [14] for more on this point. Thus, the essence of the minimization of the Durbin lower bound (2.3) by its attainment by antieigenvector regression vectors as described in Theorem 2.1 has to do with the polar Hermitian factor of X, and not with its isometric factor U. So our thought experiment exercise leading to (9.4) says that the unitary factor of the design matrix X has no effect on its statistical efficiency.

To conclude, in this paper we have placed the theory of statistical efficiency into the geometrical setting of the author's operator trigonometry. There are many remaining aspects of both together with their further interconnection, with which we have not dealt.

Addendum

In the intervening two years since the IWMS 2005 conference, on which the work herein was first presented, I have written two further related papers that should be mentioned: Gustafson [37, 38].

In [37], what follows are rendered trigonometric: Khatri-Rao inequality, Khatri-Rao-Ando bound, Bartmann-Bloomfield bound, and Hotelling correlation coefficient. In [38], I provide a complete survey of the various applications of my operator trigonometry, from 1966 to the present.

Acknowledgment

Finally, I would like to add that following the IWMS 2005 conference, I was led to a great appreciation of all of Jeff Hunter's pioneering contributions to the linear algebra of Markov chains and, in particular, his extremely fine and important results for the perturbation theory using generalized inverses and finite rank updates.

References

[1] K. Gustafson, "On geometry of statistical efficiency," preprint, 1999.

[2] K. Gustafson, "An unconventional computational linear algebra: operator trigonometry," in *Unconventional Models of Computation, UMC'2K*, I. Antoniou, C. Calude, and M. Dinneen, Eds., pp. 48–67, Springer, London, UK, 2001.

[3] K. Gustafson, "Operator trigonometry of statistics and econometrics," *Linear Algebra and Its Applications*, vol. 354, no. 1–3, pp. 141–158, 2002.

[4] P. Bloomfield and G. S. Watson, "The inefficiency of least squares," *Biometrika*, vol. 62, no. 1, pp. 121–128, 1975.

[5] M. Knott, "On the minimum efficiency of least squares," *Biometrika*, vol. 62, no. 1, pp. 129–132, 1975.

[6] C. R. Rao and C. V. Rao, "Stationary values of the product of two Raleigh quotients: homologous canonical correlations," *Sankhyā B*, vol. 49, no. 2, pp. 113–125, 1987.

[7] S. G. Wang and S.-C. Chow, *Advanced Linear Models: Theory and Applications*, vol. 141 of *Statistics: Textbooks and Monographs*, Marcel Dekker, New York, NY, USA, 1994.

[8] K. Gustafson, "A min-max theorem," *Notices of the American Mathematical Society*, vol. 15, p. 799, 1968.

[9] K. Gustafson, *Lectures on Computational Fluid Dynamics, Mathematical Physics, and Linear Algebra*, World Scientific, River Edge, NJ, USA, 1997.

[10] K. Gustafson and D. K. M. Rao, *Numerical Range: The Field of Values of Linear Operators and Matrices*, Universitext, Springer, New York, NY, USA, 1997.

[11] K. Gustafson, "Commentary on topics in the analytic theory of matrices," in *Collected Works of Helmut Wielandt 2*, B. Huppert and H. Schneider, Eds., pp. 356–367, DeGruyters, Berlin, Germaney, 1996.

[12] K. Gustafson, "Antieigenvalue inequalities in operator theory," in *Proceedings of the 3rd Symposium on Inequalities*, O. Shisha, Ed., pp. 115–119, Academic Press, Los Angeles, Calif, USA, September 1972.

[13] K. Gustafson, "Antieigenvalues," *Linear Algebra and Its Applications*, vol. 208-209, pp. 437–454, 1994.

[14] K. Gustafson, "An extended operator trigonometry," *Linear Algebra and Its Applications*, vol. 319, no. 1–3, pp. 117–135, 2000.

[15] K. Gustafson, "A note on left multiplication of semigroup generators," *Pacific Journal of Mathematics*, vol. 24, no. 3, pp. 463–465, 1968.

[16] R. L. Plackett, "A historical note on the method of least squares," *Biometrika*, vol. 36, no. 3-4, pp. 458–460, 1949.

[17] A. C. Aitken, "On least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh Section A*, vol. 55, pp. 42–48, 1935.

[18] J. Durbin and M. G. Kendall, "The geometry of estimation," *Biometrika*, vol. 38, no. 1-2, pp. 150–158, 1951.

[19] G. S. Watson, "Serial correlation in regression analysis—I," *Biometrika*, vol. 42, no. 3-4, pp. 327–341, 1955.

[20] G. S. Watson, "Linear least squares regression," *Annals of Mathematical Statistics*, vol. 38, no. 6, pp. 1679–1699, 1967.

[21] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression—I," *Biometrika*, vol. 37, no. 3-4, pp. 409–428, 1950.

[22] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression—II," *Biometrika*, vol. 38, no. 1-2, pp. 159–178, 1951.

[23] T. W. Anderson, "On the theory of testing serial correlation," *Skandinavisk Aktuarietidskrift*, vol. 31, pp. 88–116, 1948.

[24] R. L. Anderson and T. W. Anderson, "Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series," *Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 59–81, 1950.

[25] U. Grenander, "On the estimation of regression coefficients in the case of an autocorrelated disturbance," *Annals of Mathematical Statistics*, vol. 25, no. 2, pp. 252–272, 1954.

[26] U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, John Wiley & Sons, New York, NY, USA, 1957.

[27] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London A*, vol. 222, pp. 309–368, 1922.

[28] J. von Neumann, "Distribution of the ratio of the mean square successive difference to the variance," *Annals of Mathematical Statistics*, vol. 12, no. 4, pp. 367–395, 1941.

[29] J. von Neumann, R. H. Kent, H. R. Bellinson, and B. I. Hart, "The mean square successive difference," *Annals of Mathematical Statistics*, vol. 12, pp. 153–162, 1941.

[30] K. Gustafson, *Introduction to Partial Differential Equations and Hilbert Space Methods*, Dover, Mineola, NY, USA, 3rd edition, 1999.

[31] K. Gustafson, "Operator trigonometry of the model problem," *Numerical Linear Algebra with Applications*, vol. 5, no. 5, pp. 377–399, 1998.

[32] H. Cramér, *Mathematical Methods of Statistics*, vol. 9 of *Princeton Mathematical Series*, Princeton University Press, Princeton, NJ, USA, 1946.

[33] R. A. Fisher, "Theory of statistical estimation," *Proceedings of the Cambridge Philosophical Society*, vol. 22, pp. 700–725, 1925.

[34] I. Miller and M. Miller, *Freund's Mathematical Statistics*, Prentice-Hall, Upper Saddle River, NJ, USA, 6th edition, 1999.

[35] K. Gustafson, "Parallel computing forty years ago," *Mathematics and Computers in Simulation*, vol. 51, no. 1-2, pp. 47–62, 1999.

[36] S. Puntanen and G. P. H. Styan, "The equality of the ordinary least squares estimator and the best linear unbiased estimator," *The American Statistician*, vol. 43, no. 3, pp. 153–164, 1989.

[37] K. Gustafson, "The trigonometry of matrix statistics," *International Statistical Review*, vol. 74, no. 2, pp. 187–202, 2006.

[38] K. Gustafson, *Noncommutative trigonometry*, vol. 167 of *Operator Theory: Advances and Applications*, Birkhäuser Basel, Birkhäuser Verlag AG, Viaduktstraße 42, 4051 Basel, Switzerland, 2006.

K. Gustafson: Department of Mathematics, University of Colorado at Boulder, Boulder,
Colo 80309, USA
Email address: gustafs@euclid.colorado.edu