Res. Lett. Inf. Math. Sci., 2005, Vol.**8**, pp105-121
Available online at http://iims.massey.ac.nz/research/letters/

The geometry of statistical efficiency [1]

K. Gustafson

Department of Mathematics
University of Colorado at Boulder, USA

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 (1999, 2001, 2002) this author 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 (1999, 2001, 2002) 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 is 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 1975 Bloomfield–Watson–Knott solution of the Durbin conjecture, its geometrical meaning. In Section 3 we provide the reader with the basics of the operator trigonometry. This is 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

[1] This paper is an expanded version of a presentation given at the 14th International Workshop on Matrices and Statistics, Auckland, New Zealand, March 29–April 1, 2005

here for an audience of statisticians, in Section 4 we recall the origin of the operator trigonometry: 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 1987 by Rao–Rao to the general mathematical setting of statistical efficiency and operator trigonometry, all three 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

The following was shown in Gustafson (1999, 2002, see also 2001). Consider the general linear model, we follow Wang and Chow (1994) 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, 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 $n \stackrel{\geq}{=} 2p$, i.e., 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

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

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

is

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

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(1999) 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 the next 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 $\hat{\beta}$ against BLUE β^* is*

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

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

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

where the $\phi_i(V)$ are the successive decreasing critical turning angles of V, i.e., 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.

In Gustafson (2002) some related trace statistical efficiency bounds were also given operator trigonometric interpretation.

Commentary. The paper Gustafson (1999) was summarily rejected by one of the two referees when it was submitted to a journal. It was then published as the first four sections of Gustafson (2002). See also Gustafson (2001) where the result of Theorem 2.1 was summarized within the wider context of the operator trigonometry. Some more related historical perspective will be given in Section 5.

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 the next 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 (1968) 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 a 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 notation $\cos \phi(A)$ and $\sin \phi(A)$ was introduced so as to avoid any unwarranted confusion with cosine and sine functions in an operator's functional calculus. Moreover then 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 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 Gustafson (1997), Gustafson and Rao (1997), and the surveys Gustafson (1996, 2001). One more basic ingredient which should be mentioned here is the Euler equation

$$2||Ax||^2||x||^2(Re\, A)x - ||x||^2 Re\langle Ax, x\rangle A^*Ax - ||Ax||^2 Re\langle Ax, x\rangle x = 0 \qquad (3.6)$$

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 (1972), in a way analogous to that for higher eigenvalues in the Rayleigh–Ritz theory. That is okay for some applications but later, Gustafson (1994), 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 (2000) for more details.

Commentary. This "nested" operator turning angle theory for higher antieigenvalues occurred to the author advantagiously in the process of an application of the operator trigonometry to iterative solvers of linear systems $Ax = b$, in the early 1990s, and was first mentioned in Gustafson (1994). See also Gustafson (2000) for a discussion of this point.

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 below.

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 an Banach space and let A be the densely defined infinitesimal generator of a contraction semigroup e^{tA} on X. In other words,

consider the intial value problem

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

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, or 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 the following, Gustafson (1968a), 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 (1968a) 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

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

which are attained by the antieigenvector pair (3.5).

Commentary. It was very fortunate that the 1967 proof of Theorem 4.1 necessitated both entities μ_1 and ν_1, and hence gave rise to both $\cos \phi(A)$ and $\sin \phi(A)$, in a natural way. For more information and background on the operator trigonometry and the antieigenvalue–antieigenvector theory, see the not-so-old books Gustafson (1997), Gustafson–Rao (1997).

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 (1975), Knott (1975), and others, nonetheless in writing Gustafson (1999) 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 Section 4 of Gustafson (1999, 2002). There for example one finds a description of precursor work of Plackett (1949), Aiken (1934), and Durbin and Kendall (1951). The latter paper is quite explicitly geometrical, although, not operator theoretically. Plackett (1949) 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 (1955) paper is probably the explicit source of the "Durbin conjecture". In fact one finds it there, eqn (3.5), with a footnote crediting it to J. Durbin. However, Watson (1967) admits a flaw in his 1955 argument and thus the verification of the Durbin conjecture remained an open problem until 1975.

Going back further to the two papers Durbin–Watson (1950, 1951), 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 terms of a regression model. We go back to the first paper and find that (p. 409) the principal issue is "the problem of testing the errors for independence forms the subject of this paper and its sucessor." Attribution is made to earlier papers by T. W. Anderson (1948), R. L and T. W. Anderson (1950), where possible serial correlations in least squares residuals from Fourier regressions were tested. In Watson (1967) which is quite a useful paper historically, study of the efficiency of least squares is said to follow that of Grenander (1954), Grenander and Rosenblatt (1957). In fact we have traced efficiency explicitly back to Fisher (1922). See our further discussion in Section 9.

Commentary. We don't even want to touch the often discussed question of who first discovered what is now called the Kantorovich inequality. We ourselves did not know it and independently recreated it in obtaining the expressions (4.3). Watson (1955) attributes it to a proof given by Cassels in an appendix to that paper. Certainly Durbin in conjecturing his lower bound in that paper, unconsciously at least, stumbled upon it. In Watson (1967) it is attributed to the book Hardy, Littlewood, Polya (1934). An extremely complete and extensive history has been given in reviews by Styan and associates, see e.g. Watson, Alpargu, Styan (1997), who conclude that the pioneering credit for the inequality goes to Frucht in 1943.

However, to our knowledge, we were the first to see its natural and direct trigonometric content: that of maximal operator turning.

6 The Von Neumann Connection and a New Connection to Partial Differential Equations

In our historical search, tracing back through the two papers Durbin and Watson (1950, 1951), one comes upon the interesting $n \times n$ matrix

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

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

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

where z is the residual from linear regression. It was shown (1951) that the mean and variance of the statistic d are given by

$$\begin{array}{ll} E(d) & = \frac{P}{n-k'-1} \\ \text{var}(d) & = \frac{2[Q-PE(d)]}{(n-k'-1)(n-k'+1)} \end{array} \tag{6.2}$$

where

$$\begin{array}{ll} P & = trA - tr(X'AX(X'X)^{-1}) \\ Q & = trA^2 - 2tr(X'A^2X(X'X)^{-1}) + tr((X'AX(X'X)^{-1})^2 \end{array} \tag{6.3}$$

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

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k'1} \\ \vdots & & & \\ x_{1n} & x_{2n} & \cdots & c_{k'n} \end{bmatrix} \tag{6.4}$$

One wonders, or at least this author wondered, 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 (1950) leads to a paper of J. Von Neumann (1941) and one cannot resist looking at it. As is well-known, Von Neumann was a polymath and this paper is no exception. An in-depth study of the statistic

$$\eta = \frac{\delta^2}{s^2} \tag{6.5}$$

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

$x_1, \ldots, x_n$. Thus we find this paper to be an early and key precedent to all the work by Durbin, Watson, and others in the period 1950–1975.

Von Neumann's analysis is extensive and he obtains a number of theoretical results which, if we might paraphrase Durbin and Watson (1950), p. 418, are more or less beyond use by conventional statisticians. However, both Durbin–Watson papers (1950, 1951) 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 (p. 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 $x_1, \ldots, x_n$ are independent or whether a trend exists."

Curiousity knowing no bounds, we pushed the historical trace back to the previous publication V. Neumann, Kent, Bellison, Hart (1941). The answer to our curiousity about why Von Neumann became involved with this statistical regression problem is found there. To quote (p. 154): "The usefulness of the differences between successive observations only appears to have been 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 (1941), we find 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, $\mu = 1, \ldots, n$, are the eigenvalues of the form $(n-1)\delta^2$. A smallest eigenvalue $A_n = 0$ is found, with eigenvector $x_0 = (1, \ldots, 1)/\sqrt{n}$. A further analysis, using an interesting technique of assuming the $x'_1, \ldots, x'_{n-1}$ to be uniformly distributed over an $n-1$ unit sphere, shows that the statistic η of (6.5) is then distributed according to

$$\eta = \frac{n}{n-1} \sum_{\mu=1}^{n-1} A_\mu x_\mu^2. \tag{6.6}$$

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

$$(n-1)\delta^2 = x_1^2 + 2\sum_{\mu=2}^{n-1} x_\mu^2 + x_n^2 - 2\sum_{\mu=1}^{n-1} x_\mu x_{\mu+1}. \tag{6.7}$$

The matrix of this form is (6.1) and it is that matrix which is also borrowed and used in Durbin and Watson (1950, 1951). Used as well are the eigenvalues

$$A_k = 4\sin^2\left(\frac{k\pi}{2n}\right), \quad k = 1, \ldots, n-1 \tag{6.8}$$

which Von Neumann computes from the determinant of A.

Commentary. When we first saw the matrix A in Durbin and Watson (1950, 1951), our take was completely different. As this author is a specialist in partial differential equations, e.g. see Gustafson (1999b), we immediately saw the matrix A in (6.1) as the discretized Poisson–Neumann boundary value problem

$$\left\{ \begin{array}{ll} -\frac{d^2u(x)}{dx^2} = f(x), & 0 < x < 1 \\ \frac{du}{dx} = 0 \quad \text{at} & x = 0, 1. \end{array} \right\} \tag{6.9}$$

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 (1998).

We also comment in passing that a similar ballistic's 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 now so important in the theory of financial derivatives and elsewhere.

7 The Inefficiency Equation and the Euler Equation

Following Wang and Chow (1994), among others, one may apply a Lagrangian method to

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

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

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

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:

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

Clearly the span $\{x_i, Vx_i\}$ is a two (or one) dimensional reducing subspace of V and 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

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

for the two (or one) eigenvalues λ_j and λ_k associated to each x_i, $i = 1, \ldots, p$. 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) yields 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 (7.5) and the Euler equation (7.6). The only other vectors satisfying the Inefficiency equation (7.5) 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 (1999, 2002).

Commentary. The statistical interpretation of relative statistical inefficiency of an OLSE estimator $\hat{\beta}$ 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 $\hat{\beta}$, now in terms of the bound (2.3) as seen trigonometrically according to Theorem 2.1, is now that 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 Rao, Rao (1987). 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 symmetric positive definite and C 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 x 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, Rao (1987), they arrive at two cases, the first corresponding to stationary values equal to 1, the second corresponding to smaller stationary values. As concerns 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 the definition $RE(\hat{\beta})$ of statistical efficiency was not clear to this author. Durbin and Kendall (1951), certainly two great veterans in the field, specifically define E to be the efficiency of t' relative to t according to (p. 151):

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

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 (1951) cite the book of Cramér (1946) for statistical efficiency. There, Chapter 32, p. 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 (p. 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 R. A. Fisher (1922, 1923–25). Also mentioned (p. 488) are (later) papers by Neyman, Pearson, 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 (1999), one finds (p. 327) that the fact that $\text{var}(\hat{\theta}) \geq$ the quantity in (9.2), is called the Cramér–Rao inequality. The denominator of (9.2) is interpretted as the information about the estimator θ which is supplied by the sample. Smaller variance is interpretted 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 bit the bullet and go back to Fisher (1922, 1923–25). Indeed in his first paper on p. 309 he clearly defines Efficiency of a statistic as "the ratio which its 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, or designates, or in any case, cites in connection with that definition, a 1908 paper by Student and a 1763 paper by Bayes. Then on p. 315 we find "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 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 this author 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 Gustafson (1999a). This author still remembers the genuine joy of operational naval personnel as they called out "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 for statistical applications. For example when V is nonsingular one has, e.g. see Puntanen and Styan (1989), in terms of generalized inverses,

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

However in this author's opinion the essential points are first seen for $p = 1$, i.e., 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 matrixes 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 $n \times p$ semiunitary 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) simplifies to

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

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 (2000) 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, and their further interconnection, with which we have not dealt.

Acknowledgements

The author thanks Jeffrey Hunter and the organizers of IWMS2005 for the opportunity to speak at the conference. Also the author thanks George Styan and Simo Puntanen for their interest and communications in recent years which encouraged the author to present his new operator-theoretic geometrical view of statistical efficiency and statistical estimation and related matters to the matrix statistics community. There has been some recent concurrent related work which employs the author's theory of antieigenvalues by the school of C. R. Rao which has also been reported to this Workshop.

References

Aitken, A. C. (1934). On least squares and linear combination of observations. *Proc. Royal Soc. Edinburgh* A55, 42–48.

Anderson, R. L. and Anderson, T. W. (1950). Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series. *Annals of Math. Stat.*, 21, 59–81.

Anderson, T. W. (1948). On the theory of testing serial correlation. *Skand. Aktuarietidskr.*, 31, 88–116.

Bloomfield, P. and Watson, G. S. (1975). The inefficiency of least squares. *Biometrica*, 62, 121–128.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton:Princetion University Press.

Durbin, J. and Kendall, M. G. (1951). The geometry of estimation. *Biometrika*, 38, 150–158.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least square regression. I. *Biometrica*, 37, 409–428.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least square regression. II. *Biometrica*, 38, 159–177.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Royal Soc. London*, A 222, 309–368.

Fisher, R. A. (1923–25). Theory of statistical estimation. *Proc. Cambridge Phil. Soc.*, 22, 700–725.

Grenander, U. (1954). On the estimation of regression coefficients in the case of an autocorrelated disturbance. *Annals of Math. Stat.*, 25, 252–272.

Grenander, U. and Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series*. New York:Wiley.

Gustafson, K. (1968). A min-max theorem. *Notices Amer. Math. Soc.*, 15, 799.

Gustafson, K. (1968a). A note on left multiplication of semigroup generators. *Pacific J. Math.*, 24, 463–465.

Gustafson, K. (1972). Antieigenvalue inequalities in operator theory. *Inequalities III*, O. Shisha, ed., Academic Press, New York, 115–119.

Gustafson, K. (1994). Antieigenvalues. *Linear Algebra Appl.* 208/209, 437–454.

Gustafson, K. (1996) Commentary on topics in the analytic theory of matrices. *Collected Works of Helmut Wielandt 2*, B. Huppert and H. Schneider, eds., DeGruyters, Berlin, 356–367.

Gustafson, K. (1997). *Lectures on Computational Fluid Dynamics, Mathematical Physics, and Linear Algebra*. Singapore:World Scientific.

Gustafson, K. (1998). Operator trigonometry of the model problem. *Numer. Lin. Algebra Appl.*, 5, 377–399.

Gustafson, K. (1999). On geometry of statistical efficiency. (*preprint*).

Gustafson, K. (1999a). Parallel computing forty years ago. *Math. Comput. Simulation*, 51, 47–62.

Gustafson, K. (1999b). *Partial Differential Equations*, 3rd Edition, Dover, New York.

Gustafson, K. (2000). An extended operator trigonometry. *Linear Algebra Appl.*, 319, 117–135.

Gustafson, K. (2001). An unconventional computational linear algebra: operator trigonometry. *Unconventional Models of Computation, UMC'2K*, I. Antoniou, C. Calude, M. Dinneen, eds., Springer, London, 48–67.

Gustafson, K. (2002). Operator trigonometry of statistics and economics. *Linear Algebra Appl.*, 354, 141–158.

Gustafson, K. and Rao, D. (1997). *Numerical Range: The Field of Values of Linear Operators and Matrices*. Berlin:Springer.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1934). *Inequalities*. Cambridge:Cambridge University Press.

Knott, M. (1975). On the minimum efficiency of least squares. *Biometrika*, 62, 129–132.

Miller, I. and Miller, M. (1999). *John E. Freund's Mathematical Statistics*, Sixth Edition. New Jersey:Prentice Hall.

Plackett, R. I. (1949). A historical note on the method of least squares. *Biometrika*,36, 458–460.

Puntanen, S. and Styan, G. P. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, 43, 153–161.

Rao, C. R. and Rao, M. B. (1987). Stationary values of the product of two Rayleigh quotients: homologous canonical correlations. *Sankhya: Indian J. Statis.*, 49B, 113–125.

Von Neumann, J. (1941). Distribution of the ratio of mean square successive difference to the variance. *Annals of Math. Stat.*, 12, 367–395.

Von Neumann, J., Kent, R. H., Bellinson, H. R., and Hart, B. I. (1941). The mean square successive difference. *Annals of Math. Stat.*, 12, 153–162.

Wang, S. G. and Chow, S. C. (1994). *Advanced Linear Models*. New York:Marcel–Dekker.

Watson, G. S. (1955). Serial correlation in regression analysis I. *Biometrika*, 42, 327–341.

Watson, G. S. (1967). Linear least squares regression. *Annals of Math. Stat.*, 38, 1679–1699.

Watson, G. S., Alpargu, G., and Styan, G. P. (1997). Some comments on six inequalities associated with the inefficiency of ordinary least squares with one regressor. *Lienar Algebra Appl.*, 264, 13–54.