AUSTRIAN JOURNAL OF STATISTICS
Volume 32 (2003), Number 1&2, 71–83

Statistical Approach to Some Mathematical Problems

Abram Kagan
University of Maryland, College Park, U.S.A.

Dedicated to Abram A. Zinger on the occasion of his 75^{th} birthday

Abstract: A number of examples illustrate the following thesis. Mathematical problems involving the Fisher information have statistical roots. The list of examples includes (i) singularity of product measures with respect to their shifts, (ii) Carlen's superadditivity, (iii) Stam inequality, (iv) Milne's inequality, (v) efficient score and partitioning systems of normal equations, (vi) convexity of the Fisher information matrix.

Keywords: Efficient Score, Fisher Information, Linear Regression, Milne's Inequality, Shifts of Product Measures, Stam Inequality, Superadditivity.

1 Distinguishing a Sequence of Independent Random Variables from its Shifts

Let

$$\mathbf{X} = (X_1, X_2, \ldots) \tag{1}$$

be a sequence of independent random variables. Denote by $\mu_{\mathbf{X}}$ the measure in the space $I\!R^{\infty}$ of all sequences generated in the standard way by $\mathbf{X}$. For a sequence $\mathbf{a} = (a_1, a_2, \ldots)$ of numbers we denote by $\mu_{\mathbf{X}+\mathbf{a}}$ the measure in $I\!R^{\infty}$ generated by the shifted sequence

$$\mathbf{X} + \mathbf{a} = (X_1 + a_1, X_2 + a_2, \ldots). \tag{2}$$

For independent identically distributed $X_1, X_2, \ldots$ the following results were proved in Shepp (1965).
(i) If $||\mathbf{a}||^2 = \sum_1^{\infty} a_j^2 = \infty$, then the measures $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{X}+\mathbf{a}}$ are mutually singular whatever the distribution of X_j is.
(ii) If $I_{X_j} < \infty$, the condition $||\mathbf{a}|| = \infty$ becomes also necessary for mutual singularity of $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{X}+\mathbf{a}}$.

In this section statistical meaning of Shepp's results is discussed. It may be of a general interest since, as noticed in Shepp (1965), "it was unexpected that the Fisher information plays such a central role".

For independent (not necessarily identically distributed) $X_1, X_2, \ldots$ consider an infinite sequence of observations $\mathbf{Y} = (Y_1, Y_2, \ldots)$,

$$Y_j = X_j + a_j\theta,\ j = 1, 2, \ldots \tag{3}$$

with $\theta \in I\!R$ as a parameter.

Distinguishing (1) from (2) means testing the null hypothesis $H_0 :\ \theta = 0$ versus the alternative $H_1 : \theta = 1$ with zero probabilities of type 1 and type 2 errors. Let us deal with a more general problem of estimating θ.

Assuming the variances of $X_1, X_2, \ldots$ finite and positive (the latter assumption eliminate trivial cases when a single observation suffices for distinguishing $\mathbf{X}$ from $\mathbf{X} + \mathbf{a}$),

$$0 < \sigma_j^2 = \mathrm{var}(X_j) < \infty,\ j = 1, 2, \ldots, \tag{4}$$

consider the (generalized) least squares estimator of θ from $Y_1, \ldots, Y_n$:

$$\hat{\theta}_n^{LS} = \arg\min_{\theta} \sum_1^n \sigma_j^{-2}(Y_j - \theta a_j)^2 = \frac{\sum_1^n a_j \sigma_j^{-2} Y_j}{\sum_1^n a_j^2 \sigma_j^{-2}}. \tag{5}$$

The estimator (5) is unbiased with the variance

$$\mathrm{var}(\hat{\theta}_n^{LS}) = \frac{1}{\sum_1^n a_j^2 \sigma_j^{-2}}. \tag{6}$$

Thus, if

$$\sum_1^\infty a_j^2 \sigma_j^{-2} = \infty, \tag{7}$$

the least squares estimator $\hat{\theta}_n^{LS}$ converges to θ in mean square as $n \to \infty$.
The following version of the strong law of large numbers ensures the convergence of $\hat{\theta}_n^{LS}$ to θ with probability one.

Lemma 1.1 *Let $\zeta_1, \zeta_2, \ldots$ be a sequence of independent random variables with $E(\zeta_j) = 0$, $E(\zeta_j^2) = v_j$, $j = 1, 2, \ldots$ and let a sequence of numbers $b_1, b_2, \ldots$ satisfy the conditions*

$$b_1 < b_2 < \ldots,\ \lim_{j\to\infty} b_j = +\infty. \tag{8}$$

If

$$\sum_1^\infty v_j b_j^{-2} < \infty, \tag{9}$$

then with probability one

$$(1/b_n) \sum_1^n \zeta_j \to 0 \ \text{ as } \ n \to \infty. \tag{10}$$

Proof of Lemma 1. See, e.g., Shiryaev (1996), Chapter 4, Theorem 2. □

On setting $\zeta_j = a_j \sigma_j^{-2}(Y_j - a_j\theta)$ one has $E(\zeta_j) = 0,\ E(\zeta_j^2) = v_j = a_j^2 \sigma_j^{-2}$.
For $b_n = \sum_1^n v_j$ one has $b_{j-1} < b_j$ and

$$\begin{aligned}\sum_2^n v_j b_j^{-2} &= \sum_2^n (b_j - b_{j-1})/b_j^2 < \sum_2^n (b_j - b_{j-1})/(b_j b_{j-1}) \\ &= \sum_2^n (1/b_{j-1} - 1/b_j) = 1/b_1 - 1/b_n < 1/b_1\end{aligned}$$

so that (9) holds.

By virtue of Lemma 1, with probability one

$$\hat{\theta}_n^{LS} - \theta = \frac{\sum_1^n a_j \sigma_j^{-2}(Y_j - a_j\theta)}{\sum_1^n a_j \sigma_j^{-2}} = (1/b_n)\sum_1^n \zeta_j \to 0,\ n \to \infty. \tag{11}$$

The relation (11) implies that under condition (7) any two measures μ_1, μ_2 in $\mathbb{R}^\infty$ generated by the random variables (3) with $\theta = \theta_1$ and $\theta = \theta_2$, respectively, $\theta_1 \neq \theta_2$ are mutually singular. The above measures $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{X}+\mathbf{a}}$ correspond to $\theta = 0$ and $\theta = 1$. Hence, the following result holds.

Theorem 1.1 *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with finite positive variances $\sigma_1^2, \sigma_2^2, \ldots$ and let $a_1, a_2, \ldots$ be a sequence of numbers. If*

$$\sum_1^\infty a_j^2 \sigma_j^{-2} = +\infty,$$

the sequence $X_1, X_2, \ldots$ is distinguishable from $X_1 + a_1, X_2 + a_2, \ldots$, i.e., there is a set $A \subset \mathbb{R}^\infty$ such that $P\{(X_1, X_2, \ldots) \in A\} = 1$ while
$P\{(X_1 + a_1, X_2 + a_2, \ldots) \in A\} = 0$.

Note that the construction of a discriminating set A requires the knowledge of the variances $\sigma_1^2, \sigma_2^2, \ldots$ (but not of the distributions of $X_1, X_2, \ldots$) and of $a_1, a_2, \ldots$. The discrimination can be based on the tail σ-algebra $\sigma = \bigcap_n \sigma^{(n)}$ where $\sigma^{(n)} = \sigma(Y_n, Y_{n+1}, \ldots)$ is the σ-algebra generated by $Y_n, Y_{n+1}, \ldots$.

Assume now that X_j has a density f_j such the Fisher information in X_j

$$I_j = \int (f_j'(x)/f_j(x))^2 f_j(x)dx$$

is finite, $j = 1, 2, \ldots$. The Fisher information on θ contained in $Y_j = X_j + a_j\theta$ equals

$$I_{Y_j} = a_j^2 I_j$$

and due to additivity of the Fisher information, the information on θ in $\mathbf{Y} = (Y_1, Y_2, \ldots)$ is

$$I_{\mathbf{Y}} = \sum_1^\infty a_j^2 I_j. \tag{12}$$

For independent identically distributed X_j one has $I_j = I$ and $I_{\mathbf{Y}} = \sum_1^\infty a_j^2$, the last relation explaining one of Shepp's results:

$$\text{if } I < \infty, \quad \text{then } I_{\mathbf{Y}} = \infty \Longleftrightarrow \sum_1^\infty a_j^2 = \infty.$$

The condition (7) also has informational meaning. Remind that if Y is a random variable whose mean and variance are functions of a (scalar) parameter θ,

$$E_\theta(Y) = m(\theta), \mathrm{var}_\theta(Y) = v(\theta),$$

the linear score in Y is defined as $J_{\text{lin}}(Y;\theta) = m'(\theta)(Y - m(\theta))/v(\theta)$ and the linear information as

$$I_{\text{lin},Y}(\theta) = E_\theta\{(J(Y;\theta)^2\} = \{m'(\theta)\}^2/v(\theta).$$

If $I_Y(\theta) < \infty$ (so that the Fisher score $J(Y;\theta)$ is well defined), the linear score equals

$$J_{\text{lin}}(Y;\theta) = \hat{E}_\theta\{J(Y;\theta)|Y\}$$

where the mathematical expectation in the wide sense $\hat{E}_\theta$ simply means the minimum variance linear approximation to $J(Y;\theta)$.

The linear information in (3) is $\sum_1^\infty a_j^2\sigma_j^{-2}$ and the condition (7) simply means that $I_{\text{lin},\mathbf{Y}}(\theta) = \infty$.

A basic statistical principle supported by the Cramér-Rao inequality is that the inequality $I_{\mathbf{Y}}(\theta) < \infty$ is an obstruction to consistent estimation of θ from the data $\mathbf{Y}$. We shall show that due to the special character of the parameter in the setup (3), the finiteness of $I_{\mathbf{Y}}$ is also an obstruction to testing $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ with zero probabilities of type 1 and type 2 errors. The idea of the proof is known; see, Shepp (1965) and Ibragimov and Hasminskii (1981).

Let f_j be the density of X_j and I_j the Fisher information. Write

$$|\sqrt{f_j(x)} - \sqrt{f_j(x-a)}| = |\int_{-a}^0 \left(\sqrt{f_j(x-u)}\right)' du|,$$

then square both sides and integrate the result over x:

$$\int |\sqrt{f_j(x)} - \sqrt{f_j(x-a)}|^2 dx = \int |\int_{-a}^0 \left(\sqrt{f_j(x-u}\right)' du|^2 dx \le (a^2/4)I_j, \quad (13)$$

since

$$\int \left(\left(\sqrt{f_j(x-u)}\right)'\right)^2 dx = I_j/4.$$

If H_j is the Hellinger distance between the distributions of X_j and $X_j + a_j$,

$$H_j = \int \sqrt{f_j(x)f_j(x-a_j)}dx = 1 - (1/2)\int \left(\sqrt{f_j(x)} - \sqrt{f_j(x-a_j)}\right)^2 dx,$$

then due to (13),

$$H_j \ge 1 - a_j^2 I_j/8.$$

The Hellinger distance $H(\mu_{\mathbf{X}}, \mu_{\mathbf{X}+\mathbf{a}})$ between product measures is the product of the Hellinger distances between the factors. Hence,

$$H(\mu_{\mathbf{X}}, \mu_{\mathbf{X}+\mathbf{a}}) = \prod_1^\infty H_j \ge \prod_1^\infty (1 - a_j^2 I_j/8) > 0$$

due to $\sum_1^\infty a_j^2 I_j < \infty$ and the measures $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{X}+\mathbf{a}}$ are not mutually singular.
Thus, if all $I_j < \infty$, the condition $\sum_1^\infty a_j^2 I_j = \infty$ is necessary for mutual singularity of $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{X}+\mathbf{a}}$. In case of independent identically distributed $X_1, X_2, \ldots$, $I_j = I$ and the condition becomes Shepp's $\sum_1^\infty a_j^2 = \infty$ so that in this case it is necessary and sufficient for mutual singularity of $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{X}+\mathbf{a}}$.

2 Carlen's Superadditivity

Turn now to the phenomenon observed by Carlen (1991). Let

$$X = (X_1, X_2) \tag{14}$$

be an $s = (p + r)$-variate random vector. Assume that $s \times s$, $p \times p$ and $r \times r$ matrices I_X, I_1, I_2 of Fisher information about θ, θ_1 and θ_2, $\theta = (\theta_1, \theta_2)$ contained in observations of $X+\theta$, $X_1+\theta_1$ and $X_2+\theta_2$, respectively, are well defined. They are symmetric positive semi-definite matrices that do not depend on (location) parameters.

Let

$$I_X = \begin{pmatrix} I_{X,11}\, I_{X,12} \\ I_{X,21}\, I_{X,22} \end{pmatrix}$$

be the partition of I_X corresponding to (14). Using logarithmic Sobolev inequalities, Carlen proved the following property he called superadditivity of the Fisher information:

$$\operatorname{trace} I_X \geq \operatorname{trace} I_1 + \operatorname{trace} I_2. \tag{15}$$

He says in passing that usefulness of the (15) in statistics is unlikely.

However, it turns out that (15) has a simple and nice statistical interpretation that, besides being of interest in its own, leads to a few lines proof.

Consider two setups. In the first, the statistician observes the random vector $(X_1 + \theta_1, X_2)$ with θ_1 as a parameter. In the second, the statistician observes only $X_1 + \theta_1$. Simple calculations show that the matrix of Fisher information on θ_1 in $(X_1 + \theta_1, X_2)$ equals $I_{X,11}$ and due to monotonicity of the information matrix, $I_{X,11} \geq I_1$ with respect to the standard partial ordering of square matrices ($A \geq B$ if the matrix $A - B$ is positive semi-definite.) Hence,

$$\operatorname{trace} I_{X,11} \geq \operatorname{trace} I_1. \tag{16}$$

Similarly, comparing the matrices of Fisher information on θ_2 contained in $(X_1, X_2 + \theta_2)$ and $X_2 + \theta_2$, one gets

$$\operatorname{trace} I_{X,22} \geq \operatorname{trace} I_2. \tag{17}$$

Since

$$\operatorname{trace} I_X = \operatorname{trace} I_{X,11} + \operatorname{trace} I_{X,22},$$

the inequalities (16) and (17) prove Carlen's superadditivity.

Referring for the detail to Kagan and Landsman (1997), let us consider here a special case of Carlen's superadditivity.

Let

$$V = \begin{pmatrix} V_{11}\, V_{12} \\ V_{21}\, V_{22} \end{pmatrix}$$

be a symmetric positive definite $s \times s$ matrix partitioned as shown with $p \times p$ and $r \times r$ matrices V_{11} and V_{22}, $p + r = s$. If X_1, X_2, $X = (X_1, X_2)$ are Gaussian p-,r- and s-variate Gaussian random vectors with zero means and covariance matrices V_{11}, V_{22}, V, then $I_X = V^{-1}, I_1 = V_{11}^{-1}, I_2 = V_{22}^{-1}$. Carlen's superadditivity implies

$$\operatorname{trace}(V^{-1}) \geq \operatorname{trace}(V_{11}^{-1}) + \operatorname{trace}(V_{22}^{-1}).$$

3 Superadditivity of the Efficient Information Matrix

In this section another form of superadditivity is discussed.

Let X be a random element of arbitrary nature whose distribution depends on an s-dimensional parameter θ partitioned into p- and r-dimensional parameters

$$\theta = (\theta_1, \theta_2). \tag{18}$$

We assume that the (row) vector score

$$J(X;\theta) = J = (J_1, J_2),$$

partitioned according to (18) is well defined and that all the components are square integrable so that the matrix of Fisher information on θ in X ,

$$I_X(\theta) = \begin{pmatrix} I_{X,11}(\theta)\, I_{X,12}(\theta) \\ I_{X,21}(\theta)\, I_{X,22}(\theta), \end{pmatrix}$$

partitioned according to (18) is also well defined. We assume $I_X(\theta)$ positive definite. The efficient score for θ_1 is defined as

$$\hat{J}_1^{(1)}(X;\theta) = J_1 - \hat{E}_\theta(J_1|J_2) = J_1 - I_{12}I_{22}^{-1}J_2$$

(we skipped the arguments X and θ on the right hand side). As usual, $\hat{E}_\theta$ stands for the minimum variance approximation of the components of J_1 by linear combinations of the components of J_2 (see, e.g., Bickel et al., 1998, Chapter 2).

The efficient information matrix on θ_1 is, by definition,

$$\hat{I}_X^{(1)}(\theta) = E_\theta\left(\{(\hat{J}^{(1)}(X;\theta)\}^{\mathrm{T}}\{(\hat{J}^{(1)}(X;\theta)\}\right) = I_{11} - I_{12}I_{22}^{-1}I_{21}.$$

The symbol T stands for transposition; note that $\hat{I}_X^{(1)}(\theta)$ is a $p \times p$ matrix depending on an s-dimensional parameter. We refer to Bickel et al. (1998), Chapter 2 for the role of the efficient score in estimation (see also Section 1 of Kagan and Rao, 2003). In Klebanov and Melamed (1976) and Klebanov and Melamed (1978), the concept of efficient score was introduced and studied under the name "informant in presence of a nuisance parameter".

Similar to the Fisher information matrix, the efficient information matrix is monotone: *if $S = S(X)$ is a statistic, then (under standard regularity conditions) $\hat{I}_S^{(1)}(\theta) \le \hat{I}_X^{(1)}(\theta)$.*

However, unlike the Fisher information matrix, the efficient information matrix is not additive but superadditive: *if X, Y are independent random elements, then*

$$\hat{I}_{X,Y}^{(1)}(\theta) \ge \hat{I}_X^{(1)}(\theta) + \hat{I}_Y^{(1)}(\theta). \tag{19}$$

If X, Y are independent and identically distributed, the equality sign holds in (19).

In Milne (1925) (see also Hardy et al., 1952, Problem 67), the following inequality was proved. For any positive numbers $w_1, \ldots, w_n$ such that $w_1 + \ldots + w_n = 1$ and any ρ_j with $|\rho_j| < 1,\ j = 1, \ldots, n$

$$\left(\sum_1^n \frac{w_j}{1-\rho_j}\right)\left(\sum_1^n \frac{w_j}{1+\rho_j}\right) \ge \sum_1^n \frac{w_j}{1-\rho_j^2}. \tag{20}$$

(Milne needed (20) for a problem in astrophysics).

It turns out that (20) is a straightforward corollary of monotonicity of the efficient information in the setup of independent bivariate normal vectors $X_j = (X'_j, X''_j)$, $j = 1, \dots, n$ with $E_\theta(X_j) = \theta = (\theta_1, \theta_2)$ as a parameter and known covariance matrix $\mathrm{var}(X_j) = v_j^{-1} \begin{pmatrix} 1 & \rho_j \\ \rho_j & 1 \end{pmatrix}$.

The efficient information on θ_1 in $\mathbf{X} = (X_1, \dots, X_n)$ is

$$\hat{I}^{(1)}_X = \sum_1^n \frac{v_j}{1-\rho_j^2} - \left(\sum_1^n \frac{\rho_j v_j}{1-\rho_j^2} \right)^2 \left(\sum_1^n \frac{v_j}{1-\rho_j^2} \right)^{-1} \tag{21}$$

and since the distribution of the vector $\mathbf{X}' = (X'_1, \dots, X'_n)$ of the first components depends only on θ_1,

$$\hat{I}^{(1)}_{\mathbf{X}'} = I_{\mathbf{X}'}(\theta_1) = \sum_1^n v_j. \tag{22}$$

Since $\mathbf{X}' = \mathbf{X}'(\mathbf{X})$ is a statistic, monotonicity of the efficient information together with (21) and (22) results in

$$\sum_1^n \frac{v_j}{1-\rho_j^2} - \left(\sum_1^n \frac{\rho_j v_j}{1-\rho_j^2} \right)^2 \left(\sum_1^n \frac{v_j}{1-\rho_j^2} \right)^{-1} \geq \sum_1^n v_j$$

which, on setting $w_j = v_j / \sum_1^n v_j$, becomes Milne's inequality (20).

See Kagan and Rao (2003) and Rao (2000) for the detail and other applications of the efficient information.

Of general statistical interest is the fact that in the standard linear regression setup, the efficient score partitions the system of normal equations into two subsystems of lower total computational complexity. Namely, consider the standard linear regression

$$Y = A\theta + X \tag{23}$$

where Y is an $n \times 1$ observable vector, A is a known $n\times$ design matrix of full rank $s \leq n$, θ is an s-dimensional (column) vector parameter partitioned into p- and s-dimensional subvectors, $\theta^{\mathrm{T}} = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})$ and X is an $n \times 1$ vector of independent identically distributed errors with zero mean and finite variance σ^2.

The least squares estimator of θ

$$\hat{\theta}^{LS} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}Y \tag{24}$$

is the best linear unbiased estimator of θ and is obtained as the solution of the system of normal equations

$$A^{\mathrm{T}}(Y - A\theta) = 0. \tag{25}$$

The linear (column) vector score for θ (see Section 1) in Y is

$$J_{\mathrm{lin}}(Y; \theta) = (1/\sigma^2)A^{\mathrm{T}}(Y - A\theta).$$

Let us partition A and $J_{\rm lin}^{\rm T}$ as

$$A = (A_1 \vdots A_2),\ J^{\rm T} = (J_1^{\rm T}, J_2),$$

where A_1 is an $n \times p$ matrix, A_2 is an $n \times r$ matrix, J_1 is a $p \times 1$ vector, etc.

(For notational convenience, we shall skip the subindices $\rm lin$ and Y). The linear information matrix on θ in Y is

$$I = (1/\sigma^2) \begin{pmatrix} I_{11}\ I_{12} \\ I_{21}\ I_{22} \end{pmatrix}$$

where

$$I_{11} = A_1^{\rm T} A_1,\ I_{22} = A_2^{\rm T} A_2,\ A_{12} = I_{21}^{\rm T} = A_1^{\rm T} A_2.$$

The efficient linear score for θ_1,

$$\hat{J}^{(1)} = J_1 - \hat{E}_\theta(J_1|J_2) = (1/\sigma^2)(A_1^{\rm T} - I_{12} I_{22}^{-1} A_2^{\rm T})(Y - A_1^{\rm T}\theta_1)$$

does not depend on θ_2.

Similarly, the efficient linear score for θ_2,

$$\hat{J}^{(2)} = J_2 - \hat{E}_\theta(J_2|J_1) = (1/\sigma^2)(A_2^{\rm T} - I_{21} I_{11}^{-1} A_1^{\rm T})(Y - A_2^{\rm T}\theta_2)$$

does not depend on θ_1. The solution $\hat{\theta}_1$ of the system of p linear equations

$$\hat{J}_{\rm lin,1} = 0 \tag{26}$$

and $\hat{\theta}_2$ of the system of r equations

$$\hat{J}_{\rm lin,2} = 0 \tag{27}$$

are easily seen to be a partition of the least squares estimator (24), $\hat{\theta}^{LS} = (\hat{\theta}_1, \hat{\theta}_2)$. Computing $\hat{\theta}^{LS}$ from (24) requires inverting one $s \times s$ matrix while computing it from (26) and (27) requires inverting four matrices, two of order $p \times\ p$ and two of order $r \times r$. The computational complexity of inverting an $m \times m$ matrix (for large m) is $Cm^{2+\delta}$ for some constants $C, \delta,\ 0 < \delta < 1$. Thus, if $p = r = s/2$, the computational complexity of obtaining the least squares estimator directly from (24) is 2^δ times higher than from the systems (26), (27) generated by the efficient scores.

This phenomenon is well known to computer scientists, at least (see, e.g., Aho et al., 1974, Chapter 6). For a statistician it is encouraging to know that a modification of the classical Fisher score turns a useful computational tool.

4 The Pitman Estimator and Stam Inequality

Let X, Y, Z be random variables with distribution functions $F_1(x-\theta)$, $F_2(x-\theta)$, $F(x-\theta),\ F(x) = (F_1 * F_2)(x)$, respectively, depending on a parameter $\theta \in I\!R$. If the Fisher information I_1 on θ in X and I_2 on θ in Y are finite (in this case a fortiori the distributions

of X and Y are given by densities), the information I on θ is also finite. Back in 1959 Stam proved the following classical by appearance inequality:

$$1/I \geq 1/I_1 + 1/I_2 \tag{28}$$

which is much stronger than trivial $I < \min(I_1, I_2)$, see Stam (1959) or the monograph Blahut (1987), Chapter 7.

The Stam inequality has a counterpart in terms of the Pitman estimators. The only assumption is

$$\int x^2 dF_1 < \infty, \int x^2 dF_2 < \infty. \tag{29}$$

The finiteness of the Fisher information is not assumed.

Let $X_1, \ldots, X_n$; $Y_1, \ldots, Y_n$; $Z_1, \ldots, Z_n$ be samples of size $n \geq 2$ from populations $F_1(x-\theta)$, $F_2(x-\theta)$, $F(x-\theta)$, respectively. The Pitman estimator of θ (i.e., the minimum variance equivariant estimator) from $X_1, \ldots, X_n$ is

$$t'_n = \bar{X} - E(\bar{X}|X_1 - \bar{X}, \ldots, X_n - \bar{X}) \tag{30}$$

where the conditional expectation here and in what follows is taken when $\theta = 0$. Denote by t''_n and t_n the Pitman estimators of θ from $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_n$.

The variance of the Pitman estimator is constant (i.e., does not depend on θ) so that in calculating the variance one may assume $\theta = 0$.

From (30),

$$\mathrm{var}(t'_n) = \mathrm{var}(\bar{X}) - E\left(E(\bar{X}|X_1 - \bar{X}, \ldots, X_n - \bar{X})\right)^2, \tag{31}$$

$$\mathrm{var}(t''_n) = \mathrm{var}(\bar{Y}) - E\left(E(\bar{Y}|Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})\right)^2, \tag{32}$$

$$\mathrm{var}(t_n) = \mathrm{var}(\bar{Z}) - E\left(E(\bar{Z}|Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})\right)^2. \tag{33}$$

Since $F = F_1 * F_2$, the vector $(\bar{Z}, Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})$ is equidistributed with $(\bar{X} + \bar{Y}, X_1 - \bar{X} + Y_1 - \bar{Y}, \ldots, X_n - \bar{X} + Y_n - \bar{Y})$ where $X_1, \ldots, Y_n$ are assumed independent. In view of this, the σ-algebra $\sigma(Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}) = \sigma$ generated by the Z-residuals can be considered a subalgebra of a larger σ-algebra $\sigma(X_1 - \bar{X}, \ldots, Y_n - \bar{Y}) = \tilde{\sigma}$ so that

$$E(\bar{Z}|\sigma) = E\{E(\bar{Z}|\tilde{\sigma})|\sigma\}$$

and thus,

$$E\{E(\bar{Z}|\sigma)\}^2 \leq E\{E(\bar{Z}|\tilde{\sigma})\}^2.$$

To proceed further one needs a lemma.

Lemma 4.1 *Let ξ be a random variable with $E(|\xi|) < \infty$ and η, ζ arbitrary random elements. If the pair (ξ, η) is independent of ζ, then*

$$E(\xi|\eta, \zeta) = E(\xi|\eta).$$

Proof. See Meyer (1966), Theorem 51 of which Lemma 2 is a special case. □

On applying Lemma 2 first with $\xi = \bar{X},\ \eta = (X_1 - \bar{X}, \ldots, X_n - \bar{X}),\ \zeta = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})$ and second with $\xi = \bar{Y},\ \eta = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}),\ \zeta = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$ one gets

$$E\{E(\bar{Z}|Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})\}^2 \leq$$
$$E\{E(\bar{X}|X_1 - \bar{X}, \ldots, X_n - \bar{X}) + E(\bar{Y}|Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})\}^2 =$$
$$E\{E(\bar{X}|X_1 - \bar{X}, \ldots, X_n - \bar{X}))\}^2 + E\{E(\bar{Y}|Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})\}^2. \quad (34)$$

Combining (31)-(34) results in

$$\mathrm{var}(t_n) \geq \mathrm{var}(t'_n) + \mathrm{var}(t''_n). \quad (35)$$

The inequality (35) can be called a small sample version of the Stam inequality. Its proof is completely different from the original proof of the Stam inequality.

Under regularity type conditions on F_1, F_2 (that include the finiteness of the Fisher information)

$$\mathrm{var}(t'_n) = \frac{1}{nI_1}(1 + o(1)),\ \mathrm{var}(t''_n) = \frac{1}{nI_2}(1 + o(1)),\ \mathrm{var}(t_n) = \frac{1}{nI}(1 + o(1)) \quad (36)$$

as $n \to \infty$ (see Ibragimov and Hasminskii, 1981, Example to Theorem 3.1). Hence, (35) is stronger than (28) modulo these conditions plus (29).

In Kagan (2002), one can find inequalities similar to (35) for polynomial Pitman estimators and a version of (28).

5 Convexity of the Fisher Information Matrix

For the sake of the material of this section it is convenient to use the concept of a statistical experiment $\mathcal{E}_j, j = 1, \ldots, n$ consisting in an observation of a random element X_j taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ and having distribution $P_\theta^{(j)}$ depending on a parameter $\theta \in \Theta,\ \Theta$ being an open set in $\mathbb{R}^s$. We shall call an experiment $\mathcal{E}_{\mathrm{mix}}$ a mixture of $\mathcal{E}_1, \ldots, \mathcal{E}_n$ with weights $w_1 > 0, \ldots, w_n > 0,\ w_1 + \ldots + w_n = 1$, if $\mathcal{E}_{\mathrm{mix}}$ consists in observing a random element X taking values in $(\mathcal{X}, \mathcal{A})$ and having distribution

$$P_\theta = w_1 P_\theta^{(1)} + \ldots + P_\theta^{(n)} \quad (37)$$

and shall write in this case

$$\mathcal{E}_{\mathrm{mix}} = w_1 \mathcal{E}_1 + \ldots + w_n \mathcal{E}_n.$$

Let us assume the experiments *regular* (see Ibragimov and Hasminskii, 1981, Chapter 1), so that the matrices $I_j(\theta)$ of Fisher information on θ in X_j (or, equivalently, in $\mathcal{E}_j$) and $I_{\mathrm{mix}}(\theta)$ in X (or in $\mathcal{E}_{\mathrm{mix}}$) are positive definite $s \times s$ matrices.

Lemma 5.1 *The matrix of Fisher information is convex with respect to the mixtures,*

$$I_{\mathrm{mix}}(\theta) \leq w_1 I_1(\theta) + \ldots + w_n I_n(\theta). \quad (38)$$

Proof. The proof of (38) that follows is purely statistical based on monotonicity of the information matrix.

Consider the experiment $\mathcal{E} = \mathcal{E}(w_1, \ldots, w_n)$ consisting in an observation of a random pair (Δ, X) where the distribution

$$P(\Delta = i) = w_i,\ i = 1, \ldots, n$$

of the discrete component Δ does not involve θ and the general component has the conditional distribution

$$P_\theta(X \in A | \Delta = i) = P_\theta^{(i)}(A),\ A \in \mathcal{A}.$$

The joint distribution of (Δ, X) is

$$P_\theta(\Delta = i,\ X \in A) = w_i P_\theta^{(i)}(A),\ i = 1, \ldots, n;\ A \in \mathcal{A}$$

whence the matrix of Fisher information on θ in the pair (Δ, X) (or in $\mathcal{E}$) is

$$I_{\Delta,X}(\theta) = w_1 I_1(\theta) + \ldots + w_n I_n(\theta).$$

Since the marginal distribution of X is (37), $\mathcal{E}_{\text{mix}}$ is a subexperiment of $\mathcal{E}$ and monotonicity of the Fisher information matrix leads to (38). □

For Gaussian experiments, when $X_j \sim N(\theta, V_j)$ the information matrices $I_j = V_j^{-1}$ and I_{mix} do not depend on θ and (38) becomes

$$I_{\text{mix}} \le w_1 V_1^{-1} + \ldots + w_n V_n^{-1}.$$

Using the fact that the Gaussian distribution minimizes the matrix of Fisher information on a location parameter (see, e.g. Kagan, 2001), the last inequality can be amended in the following:

$$\left(w_1 V_1^{-1} + \ldots + w_n V_n^{-1}\right)^{-1} \le I_{\text{mix}} \le w_1 V_1^{-1} + \ldots + w_n V_n^{-1}. \tag{39}$$

Without the middle term, (39) is a multivariate version of the classical inequality between the arithmetic and harmonic means. In the univariate case, one may put in the middle the geometric mean. In the multivariate case, $V_1, \ldots, V_n$ do not commute in general and (to the best of author's knowledge) there is no analog of the geometric mean. Maybe, in some sense the matrix of Fisher information on a location parameter contained in the mixture of Gaussian experiments is a candidate for the role of the geometric mean?

References

A.V. Aho, J.E. Hopcroft, and J.D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, New York, 1974.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A.Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, 1998.

R.E. Blahut. *Principles and Practice of Information Theory*. Addison-Wesley, New York, 1987.

E.A. Carlen. Superadditivity of Fisher's information and logarithmic Sobolev inequalities. *J. Funct. Anal.*, 101:194–211, 1991.

G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, Cambridge, 2nd edition, 1952.

I.A. Ibragimov and R.Z. Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer, 1981.

A. Kagan. Another look at the Cramér-Rao inequality. *The Amer. Statistician*, 55:211–212, 2001.

A. Kagan. An inequality for the Pitman estimators related to the Stam inequality. *Sankhyā Ser. A*, 64:281–292, 2002.

A. Kagan and Z. Landsman. Statistical meaning of Carlen's superadditivity of the Fisher information. *Statist. Probab. Letters*, 32:175–179, 1997.

A. Kagan and C.R. Rao. Some properties and applications of the efficient Fisher score. *J. Statist. Plan. Inference*, to appear, 2003.

L.B. Klebanov and I.A. Melamed. The Fisher information in presence of nuisance parameters and its application in statistical theory of estimation (in Russian). *Transac. of Acad. of Sci. of Georgian SSR*, 81:21–23, 1976.

L.B. Klebanov and I.A. Melamed. Several notes on Fisher information in presence of nuisance parameters. *Math. Operationsforschung und Statistik*, 9:85–90, 1978.

P.A. Meyer. *Probability and Potentials*. Blaisdell, 1966.

E.A. Milne. Note on Rosseland's integral for absorption coefficient. Monthly Notices Royal Astron. Soc., 85:979–984, 1925.

C.R. Rao. Statistical proofs of some matrix inequalities. *Linear Algebra and Applications*, 321:307–320, 2000.

L.A. Shepp. Distinguishing a sequence of random variables from a translate of itself. *Ann. Math. Statist.*, 36:1107–1112, 1965.

A.N. Shiryaev. *Probability*. Springer, New York, 2nd edition, 1996.

A. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2:101–112, 1959.

Author's address:

Abram Kagan
Department of Mathematics
University of Maryland
College Park, MD 20742
USA
E-mail: amk@math.umd.edu