Information on parameters of interest decreases under transformations

R.M. Fewster[a], P. E. Jupp[b,*]

[a]*Department of Statistics, The University of Auckland, Private Bag 92019, Auckland, New Zealand*

[b]*School of Mathematics and Statistics, University of St Andrews, St Andrews KY16 9SS, UK*

Abstract

An important property of Fisher information is that it decreases weakly under transformation of random variables. Kagan and Rao (2003) [A. Kagan, C.R. Rao, Some properties and applications of the efficient Fisher score, J. Statist. Plann. Inference 116 (2003) 343–352] showed that, in the presence of nuisance parameters, Fisher information on the interest parameters decreases similarly. We prove here a general algebraic result on partitioned positive-definite matrices, and use it to show that the decrease in Fisher information on the interest parameters is bounded below by the conditional Fisher information on the interest parameters. A consequence is that standard large-sample confidence regions for parameters of interest based on the deviance, score and Wald statistics become asymptotically 'wider' under transformations, both in the context of independent identically distributed random variables and for the binomial detectability models of Fewster and Jupp (2009) [R.M. Fewster, P.E. Jupp, Inference on population size in binomial detectability models, Biometrika 96 (2009) 805–820]. One implication is that models that combine different data sources for inference on the interest parameters are asymptotically more efficient than models for any of the individual data sources, despite the possible need for further nuisance parameters when combining the sources.

Keywords: Binomial detectability model, Confidence region, Fisher information, Observed information, Schur complement

[*]Corresponding author. Tel.: + 44 1334 46 3704; fax: + 44 1334 46 3748
Email address: pej@st-andrews.ac.uk (P. E. Jupp)

2010 MSC: 62B10, 62F25

1. Introduction

One of the key concepts in parametric statistics is Fisher information. For a random variable X on a sample space $\mathcal{X}$ with density $f(x;\theta)$ depending smoothly on a (vector) parameter θ, the Fisher information on θ given by X is

$$i_X(\theta) = E_\theta\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^{\mathrm{T}} \frac{\partial \log f(X;\theta)}{\partial\theta}\right].$$

An important property of Fisher information is that (under mild regularity conditions) it does not increase under transformation of random variables, i.e.

$$i_{t(X)} \leq i_X, \tag{1}$$

where $t : \mathcal{X} \to \mathcal{Y}$ is a transformation and $i_{t(X)}$ denotes the Fisher information on θ given by $t(X)$ (see, e.g. pp. 330–331 of [6]). Here we have followed the usual ordering on symmetric matrices, writing $A > B$ (or $B < A$) if $A - B$ is positive-definite and $A \geq B$ (or $B \leq A$) if $A - B$ is positive semi-definite.

Inequality (1) can be extended to parameters of interest. To describe this extension it is necessary to consider partitioned matrices. Let A be a $(p+q)\times(p+q)$ symmetric matrix partitioned as

$$A = \left(\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right),$$

where A_{11} and A_{22} are $p \times p$ and $q \times q$ matrices, respectively. The $p \times p$ matrix $A_{11\cdot 2}$, sometimes known as the *Schur complement of* A_{22}, is defined as

$$A_{11\cdot 2} = A_{11} - A_{12}A_{22}^{+}A_{21},$$

where A_{22}^{+} is the Moore–Penrose inverse of A_{22}, defined by the properties $A_{22}^{+}A_{22}A_{22}^{+} = A_{22}^{+}$, $A_{22}A_{22}^{+}A_{22} = A_{22}$ and $A_{22}A_{22}^{+}$ and $A_{22}^{+}A_{22}$ are symmetric.

Now suppose that the parameter θ parameterising the distribution of X can be decomposed as $\theta = (\psi, \nu)$, where ψ is a parameter of interest of dimension p and ν is a nuisance parameter of dimension q. Then i_X can be partitioned as

$$i_X = \left(\begin{array}{cc} i_{\psi\psi;X} & i_{\psi\nu;X} \\ i_{\nu\psi;X} & i_{\nu\nu;X} \end{array}\right).$$

An important role in inference on ψ is played by the quantity $i_{\psi\psi\cdot\nu;X}$, which is the Schur complement of $i_{\nu\nu;X}$ and is known variously as the *horizontal information on* ψ, *the profile information on* ψ or the *efficient matrix of Fisher information on* ψ. Lemma 1 of [5] shows that, if $i_{t(X)} > 0$ then (under mild regularity conditions)

$$i_{\psi\psi\cdot\nu;t(X)} \leq i_{\psi\psi\cdot\nu;X}. \tag{2}$$

The aims of this note are (i) to strengthen inequality (2) by placing it in a wider algebraic setting, (ii) to relate the strengthened inequality to conditional Fisher information, (iii) to use the strengthened inequality to compare the asymptotic 'widths' of confidence regions for interest parameters in the model for X and that for $t(X)$.

2. An algebraic inequality

We give the general algebraic result that underlies the information inequality (2).

Proposition 1. *Let*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

be $(p+q)\times(p+q)$ symmetric matrices partitioned in a conformable way and such that A_{22}, B_{22} and $A_{22}+B_{22}$ are invertible. Then

$$(A+B)_{11\cdot 2} = A_{11\cdot 2} + B_{11\cdot 2} + \Delta_{12}\Delta_{22}\Delta_{21}, \tag{3}$$

where

$$\begin{aligned} \Delta_{12} &= A_{12}{A_{22}}^{-1} - B_{12}{B_{22}}^{-1}, &(4)\\ \Delta_{22} &= A_{22}\left\{{A_{22}}^{-1} - (A_{22}+B_{22})^{-1}\right\}A_{22} &(5) \end{aligned}$$

and $\Delta_{21} = {\Delta_{12}}^T$.

If $A_{22} > 0$ and $B_{22} > 0$ then

$$(A+B)_{11\cdot 2} \geq A_{11\cdot 2} + B_{11\cdot 2} \tag{6}$$

and the following are equivalent

$$\begin{aligned} (A+B)_{11\cdot 2} &= A_{11\cdot 2} + B_{11\cdot 2} &(7)\\ (A_{22}+B_{22})^{-1}(A_{21}+B_{21}) &= {A_{22}}^{-1}A_{21} &(8)\\ {A_{22}}^{-1}A_{21} &= {B_{22}}^{-1}B_{21}. &(9) \end{aligned}$$

Proof. Simple manipulation gives

$$\begin{aligned}
A_{22}^{-1} - (A_{22}+B_{22})^{-1} &= (A_{22}+B_{22})^{-1}B_{22}A_{22}^{-1} && (10)\\
&= A_{22}^{-1}B_{22}(A_{22}+B_{22})^{-1}, && (11)\\
B_{22}^{-1} - (A_{22}+B_{22})^{-1} &= (A_{22}+B_{22})^{-1}A_{22}B_{22}^{-1} && (12)\\
&= B_{22}^{-1}A_{22}(A_{22}+B_{22})^{-1}, && (13)
\end{aligned}$$

and, using (12) and (13),

$$A_{21} - (A_{22}+B_{22})B_{22}^{-1}A_{22}(A_{22}+B_{22})^{-1}B_{21} = A_{22}\left(A_{22}^{-1}A_{21} - B_{22}^{-1}B_{21}\right). \quad (14)$$

Then

$$\begin{aligned}
& (A+B)_{11\cdot 2} - A_{11\cdot 2} - B_{11\cdot 2}\\
= \; & (A_{11}+B_{11}) - (A_{12}+B_{12})(A_{22}+B_{22})^{-1}(A_{21}+B_{21})\\
& \qquad - \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right) - \left(B_{11} - B_{12}B_{22}^{-1}B_{21}\right)\\
= \; & A_{12}A_{22}^{-1}A_{21} + B_{12}B_{22}^{-1}B_{21} - (A_{12}+B_{12})(A_{22}+B_{22})^{-1}(A_{21}+B_{21}). \quad (15)
\end{aligned}$$

Expanding the quadratic in (15) and making use of (11), (12), (14), (5) and (4) gives

$$\begin{aligned}
& (A+B)_{11\cdot 2} - A_{11\cdot 2} - B_{11\cdot 2}\\
= \; & A_{12}A_{22}^{-1}B_{22}(A_{22}+B_{22})^{-1}A_{21} - A_{12}(A_{22}+B_{22})^{-1}B_{21}\\
& \qquad -B_{12}(A_{22}+B_{22})^{-1}A_{21} + B_{12}(A_{22}+B_{22})^{-1}A_{22}B_{22}^{-1}B_{21}\\
= \; & \left\{A_{21} - (A_{22}+B_{22})B_{22}^{-1}A_{22}(A_{22}+B_{22})^{-1}B_{21}\right\}^{\mathrm{T}} A_{22}^{-1}B_{22}(A_{22}+B_{22})^{-1}\\
& \qquad \times \left\{A_{21} - (A_{22}+B_{22})B_{22}^{-1}A_{22}(A_{22}+B_{22})^{-1}B_{21}\right\}\\
= \; & \Delta_{12}\Delta_{22}\Delta_{21},
\end{aligned}$$

proving (3).

If $A_{22} > 0$ and $B_{22} > 0$ then $A_{22} + B_{22} > A_{22}$, so that ${A_{22}}^{-1} > (A_{22} + B_{22})^{-1}$. Then $\Delta_{22} > 0$, proving (6) and showing that (7) is equivalent to (9). Equivalence of (8) and (9) holds for any conformable matrices with A_{22}, B_{22} and $A_{22} + B_{22}$ invertible.

Equation (6) shows that the function that takes a partitioned symmetric matrix A to the Schur complement $A_{11\cdot 2}$ is superadditive on the space of positive-definite matrices. In the special case of information matrices of independent random variables, (6) was given in Lemma 2 of [5]. For general information matrices, the weaker result that $(A+B)_{11\cdot 2} \geq A_{11\cdot 2}$ was given in Lemma 1 of [5] and (implicitly), for the case $B_{22}=0, B_{12}=0, B_{21}=0$, in section 3 of [2]. The development in Proposition 1 unifies these results by giving a general algebraic formulation for all positive-definite matrices.

Remark 1. *Proposition 1 has been presented in the algebraic language of matrices. A geometrical interpretation can be gained by expressing Proposition 1 in coordinate-free geometric language. In this setting there are (a) two bilinear forms (represented in coordinate terms by matrices A and B) on a $(p+q)$-dimensional vector space (represented in coordinate terms as $\mathbb{R}^{p+q}$), (b) a q-dimensional subspace, the* vertical subspace *(represented in coordinate terms as $\{0\} \times \mathbb{R}^q = \{(0,\nu) : 0 \in \mathbb{R}^p,\ \nu \in \mathbb{R}^q\}$). The* horizontal subspace *of a bilinear form is the orthogonal complement of the vertical subspace with respect to that form. A simple calculation shows that the two horizontal subspaces are represented in coordinate terms as the subspaces $\left\{(\psi, -A_{22}^{-1}A_{21}\psi) : \psi \in \mathbb{R}^p\right\}$ and $\left\{(\psi, -B_{22}^{-1}B_{21}\psi) : \psi \in \mathbb{R}^p\right\}$ of $\mathbb{R}^{p+q}$. It follows that each of (7)–(9) is equivalent to the geometrical condition that the horizontal subspaces of the two bilinear forms are the same.*

The matrix Δ_{22} represents a bilinear form on the vertical subspace, Δ_{12} is a measure of discrepancy between the two horizontal subspaces, and $\Delta_{12}\Delta_{22}\Delta_{21}$ represents a type of squared distance between them.

Corollary 1. *Let A and B be symmetric with $A \geq 0$ and $B \geq 0$. Then*

$$(A+B)_{11\cdot 2} \geq A_{11\cdot 2} + B_{11\cdot 2}. \tag{16}$$

Proof. After change of basis of the parameter space, the matrices A_{22} and A_{12} can be written as

$$A_{22} = \begin{pmatrix} C_{11} & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad A_{12} = (D_{12}, E_{12}),$$

where C_{11} is an $r \times r$ matrix with $C_{11} > 0$, and D_{12} and E_{12} are $p \times r$ and $p \times (q-r)$ matrices, respectively. Since $A \geq 0$, we have $E_{12} = 0$. Then, for positive real λ, $(A + \lambda I_{p+q})_{22} > 0$ and

$$(A+\lambda I_{p+q})_{11\cdot 2} = A_{11} + \lambda I_p - D_{12}\left(C_{11} + \lambda I_r\right)^{-1} {D_{12}}^{\mathrm{T}} \to A_{11} - A_{12}A_{22}^{+}A_{12} = A_{11\cdot 2}$$

as $\lambda \to 0$. Similarly, $(B+\lambda I_{p+q})_{22} > 0$ and $(B+\lambda I_{p+q})_{11\cdot 2} \to B_{11\cdot 2}$ as $\lambda \to 0$. Replacing A and B in (6) by $A + \lambda I_{p+q}$ and $B + \lambda I_{p+q}$, respectively, and taking the limit as $\lambda \to 0$ gives (16).

3. Applications to statistics

3.1. The information inequality for interest parameters

The conditional information $i_{X|t(X)}$ on θ is defined as

$$i_{X|t(X)}(\theta) = E_\theta \left[\left(\frac{\partial \log f(X|t(X);\theta)}{\partial \theta} \right)^{\mathrm{T}} \frac{\partial \log f(X|t(X);\theta)}{\partial \theta} \right],$$

where $f(x|t(x);\theta)$ is the conditional density of X given $t(X)$. The Fisher information i_X can be decomposed as

$$i_X = i_{t(X)} + i_{X|t(X)} \tag{17}$$

(see Section 7 of [1]). Taking $A = i_{t(X)}$ and $B = i_{X|t(X)}$ in Proposition 1 and the Corollary gives the following strengthening of (2) in the spirit of (17).

Proposition 2. *The inequality*

$$i_{\psi\psi\cdot\nu;X} \geq i_{\psi\psi\cdot\nu;t(X)} + i_{\psi\psi\cdot\nu;X|t(X)} \tag{18}$$

holds. If $i_{\nu\nu;t(X)} > 0$ *and* $i_{\nu\nu;X|t(X)} > 0$ *then equality holds in (18) if and only if*

$$i_{\nu\nu;X}^{-1}\, i_{\nu\psi;X} = i_{\nu\nu;t(X)}^{-1}\, i_{\nu\psi;t(X)}.$$

3.2. Application to confidence regions

The standard likelihood-based large-sample approximate confidence regions for ψ are derived by inverting the likelihood ratio, Wald, and score tests, using either Fisher information or observed information. The large-sample approximate $100(1-\alpha)\%$ confidence regions based on observations on random variables $X_1, \ldots, X_n$ are

(i) Wald regions based on Fisher information,

$$\mathrm{CR}_{\alpha,X,e} = \left\{ \psi : \left(\psi - \hat{\psi}_{X_1,\ldots,X_n} \right)^{\mathrm{T}} \hat{i}_{\psi\psi\cdot\nu;X_1,\ldots,X_n} \left(\psi - \hat{\psi}_{X_1,\ldots,X_n} \right) < \chi^2_{p;\alpha} \right\}, \tag{19}$$

where $\hat{\psi}_{X_1,\ldots,X_n}$ is the maximum likelihood estimate of ψ, $\hat{i}_{\psi\psi\cdot\nu;X_1,\ldots,X_n}$ is the horizontal Fisher information on ψ based on $X_1, \ldots, X_n$ and evaluated at $\hat{\theta}_{X_1,\ldots,X_n}$, p is the dimension of ψ, and $\chi^2_{p;\alpha}$ denotes the upper α quantile of the χ^2 distribution with p degrees of freedom;

(ii) Wald regions based on observed information,

$$\mathrm{CR}_{\alpha,X,o} = \left\{ \psi : \left(\psi - \hat{\psi}_{X_1,\ldots,X_n} \right)^{\mathrm{T}} \hat{\jmath}_{\psi\psi\cdot\nu;X_1,\ldots,X_n} \left(\psi - \hat{\psi}_{X_1,\ldots,X_n} \right) < \chi^2_{p;\alpha} \right\}, \tag{20}$$

where

$$\hat{\jmath}_{\psi\psi\cdot\nu;X_1,\ldots,X_n} = \hat{\jmath}_{\psi\psi;X_1,\ldots,X_n} - \hat{\jmath}_{\psi\nu;X_1,\ldots,X_n} \hat{\jmath}^{-1}_{\nu\nu;X_1,\ldots,X_n} \hat{\jmath}_{\nu\psi;X_1,\ldots,X_n}$$

with

$$\hat{\jmath}_{\theta\theta;X_1,\ldots,X_n} = -\frac{\partial^2 l}{\partial\theta^{\mathrm{T}}\partial\theta}(\hat{\theta}) = \begin{pmatrix} \hat{\jmath}_{\psi\psi;X_1,\ldots,X_n} & \hat{\jmath}_{\psi\nu;X_1,\ldots,X_n} \\ \hat{\jmath}_{\nu\psi;X_1,\ldots,X_n} & \hat{\jmath}_{\nu\nu;X_1,\ldots,X_n} \end{pmatrix}$$

being the observed information on θ at $\hat{\theta}_{X_1,\ldots,X_n}$ and $l(\psi, \nu; x_1, \ldots, x_n)$ denoting the log-likelihood;

(iii) deviance regions (also called profile likelihood confidence regions),

$$\mathrm{CR}_{\alpha,X,d} = \left\{ \psi : 2 \left(l_p(\hat{\psi}_{X_1,\ldots,X_n}) - l_p(\psi) \right) < \chi^2_{p;\alpha} \right\}, \tag{21}$$

with l_p denoting the profile log-likelihood, defined by $l_p(\psi; x_1, \ldots, x_n) = \sup_\nu l(\psi, \nu; x_1, \ldots, x_n)$;

(iv) 'expected' score regions,

$$\mathrm{CR}_{\alpha,X,es} = \left\{ \psi : \frac{\partial l_p}{\partial\psi}(\psi) \hat{\imath}^{-1}_{\psi\psi\cdot\nu;X} \frac{\partial l_p}{\partial\psi}(\psi)^{\mathrm{T}} < \chi^2_{p;\alpha} \right\}; \tag{22}$$

(v) 'observed' score regions,

$$\mathrm{CR}_{\alpha,X,os} = \left\{ \psi : \frac{\partial l_p}{\partial\psi}(\psi) \hat{\jmath}^{-1}_{\psi\psi\cdot\nu;X} \frac{\partial l_p}{\partial\psi}(\psi)^{\mathrm{T}} < \chi^2_{p;\alpha} \right\}. \tag{23}$$

We now show that the confidence regions for ψ based on X are asymptotically 'shorter' than those based on $t(X)$, for each of these standard types of confidence region.

Remark 2. *If all instances of $\hat{\imath}$ and $\hat{\jmath}$ in (i), (ii), (iv), and (v) above are replaced by i and j evaluated at the maximum likelihood estimate of θ given ψ, for each candidate value of ψ in the confidence region, then Propositions 3 and 4 below hold without change.*

3.2.1. The i.i.d. case

Proposition 3. *Let $t : \mathcal{X} \to \mathcal{Y}$ be a transformation. Denote by $\mathrm{CR}_{\alpha,X,m}$ and $\mathrm{CR}_{\alpha,t(X),m}$ the confidence regions (19)–(23) for ψ based on observations on $X_1, \ldots, X_n$ and $t(X_1), \ldots, t(X_n)$, respectively. If $i_{\psi\psi\cdot\nu;X|t(X)} > 0$ then, using any coordinate system for ψ,*

$$P\left(\mathrm{CR}_{\alpha,X,m} - \hat{\psi}_X \subset \mathrm{CR}_{\alpha,t(X),m} - \hat{\psi}_{t(X)}\right) \to 1 \quad \text{as } n \to \infty, \tag{24}$$

for $m = e, o, d, es, os$, where $\hat{\psi}_X$ and $\hat{\psi}_{t(X)}$ are the maximum likelihood estimates of ψ based on $X_1, \ldots, X_n$ and $t(X_1), \ldots, t(X_n)$, respectively, $\subset$ in (24) denotes strict inclusion, and $\mathrm{CR}_{\alpha,X,m} - \hat{\psi}_X = \left\{\psi - \hat{\psi}_X : \psi \in \mathrm{CR}_{\alpha,X,m}\right\}$, etc.

Proof. From Proposition 2, $i_{\psi\psi\cdot\nu;X} > i_{\psi\psi\cdot\nu;t(X)}$. It follows from consistency of $\hat{\theta}_{X_1,\ldots,X_n}$ and $\hat{\theta}_{t(X_1),\ldots,t(X_n)}$ that, with probability tending to 1 as $n \to \infty$, $i_{\psi\psi\cdot\nu;X}(\hat{\theta}_{X_1,\ldots,X_n}) > i_{\psi\psi\cdot\nu;t(X)}(\hat{\theta}_{t(X_1),\ldots,t(X_n)})$. Then (24) follows in the case $m = e$. The case $m = o$ follows using $n^{-1}\hat{j}_{\theta\theta;X_1,\ldots,X_n} - \hat{i}_{\theta\theta;X} \to 0$ as $n \to \infty$. Standard second-order expansions of log-likelihood (e.g. Chapter 3 of [3]) yield the results in the cases $m = d$, $m = es$ and $m = os$.

3.2.2. Binomial detectability models

Many models used in the estimation of population size are binomial detectability models in the sense of [4]. These models involve a binomially-distributed number, n, of independent identically distributed observations $x_1, \ldots, x_n$ and the probability density functions have the form

$$f(n, x_1, \ldots, x_n; N, \theta) = \binom{N}{n} p(\theta)^n \{1 - p(\theta)\}^{N-n} \prod_{i=1}^{n} k(x_i; \theta),$$

where the interest parameter N is the size of the population, θ is a nuisance parameter, p is a specified function with $0 < p(\theta) < 1$ and k is a probability density function. Although binomial detectability models lie outside the strict context of independent identically distributed random variables, there are results analogous to those of Proposition 3. These are given in Proposition 4 below.

It follows from Theorem 1 of [4] that, under mild regularity conditions (such as those of Section 4.2.2 of [7]), the asymptotic distribution of the maximum likelihood estimate $\log \hat{N}$ of $\log N$ is

$$N^{1/2}\left(\log \hat{N} - \log N\right) \sim N\left(0, i_{NN\cdot\theta}(\theta)^{-1}\right),$$

asymptotically, as $N \to \infty$, where

$$i_{NN\cdot\theta} = i_{NN} - i_{N\theta} i_{\theta\theta}^{-1} i_{\theta N}$$

with

$$\begin{pmatrix} i_{NN} & i_{N\theta} \\ i_{\theta N} & i_{\theta\theta} \end{pmatrix}$$

being the Fisher information on (N, θ). Thus, $\log \hat{N}$ is a consistent estimator of $\log N$, although $\hat{N}$ is not a consistent estimator of N.

Arguments similar to those used in the proof of Proposition 3 give the following result on confidence regions for N in binomial detectability models.

Proposition 4. *Let $t : \mathcal{X} \to \mathcal{Y}$ be a transformation. Let $\mathrm{CR}_{\alpha,X,m}$ and $\mathrm{CR}_{\alpha,t(X),m}$ be confidence regions (19)–(23) for $\log N$ based on a binomial detectability model with population size N. If $i_{NN\cdot\theta;X|t(X)} > 0$ then*

$$P\left(\mathrm{CR}_{\alpha,X,m} - \log \hat{N}_X \subset \mathrm{CR}_{\alpha,t(X),m} - \log \hat{N}_{t(X)}\right) \to 1 \quad \text{as } N \to \infty, \qquad (25)$$

for $m = e, o, d, es, os$, where $\hat{N}_X$ and $\hat{N}_{t(X)}$ are the maximum likelihood estimates of N based on $n, X_1, \ldots, X_n$ and $n, t(X_1), \ldots, t(X_n)$, respectively.

Proposition 4 applies to confidence regions for $\log N$. Simulations showed a pattern of increasing inclusion of $\mathrm{CR}_{\alpha,X,m} - \log \hat{N}_X$ in $\mathrm{CR}_{\alpha,t(X),m} - \log \hat{N}_{t(X)}$ as N was increased, where $t : \mathcal{X} \to \mathcal{Y}$ mapped observation $X = (Y, Z)$ to Y. In contrast, no pattern of increasing inclusion of analogous confidence regions for N emerged. This makes us suspect that the analogue of (25) for N does not hold.

3.3. Implication for modelling

Propositions 3 and 4 have an important implication for model-building. Given a parametric statistical model parameterised by (ψ, ν) for a random variable Y, where ψ is a parameter of interest and ν is a nuisance parameter, one might ask when it is worth 'enlarging' the model to gain more information about ψ. A more precise version of the question considers parametric statistical models parameterised by (ψ, ν, ω) for (Y, Z), where Z is an additional random variable, such that the marginal model for Y is the given model parameterised by (ψ, ν). Then inference on ψ can be carried out either using the full likelihood on (ψ, ν, ω) based on observations of (Y, Z) or using the marginal likelihood of (ψ, ν) based on observations of Y alone. The question is then 'Under what conditions does the model for (Y, Z) yield shorter confidence regions for ψ than the model for Y does?'. Taking $X = (Y, Z)$ and $t(Y, Z) = Y$ in Propositions 3 and 4 shows that, provided that the horizontal conditional information on ψ from Z given Y is non-singular, the use of (Y, Z) produces asymptotically 'shorter' confidence regions than are given by the use of Y alone. This formalises the intuitive idea that (if there is negligible cost in taking observations on additional variables or handling a more complicated model) it is (asymptotically) always worth observing additional variables, despite the addition of extra nuisance parameters, provided that the additional variables are informative about ψ, i.e. $i_{\psi\psi\cdot(\nu,\omega);Z|Y} > 0$.

Acknowledgement

PEJ gratefully acknowledges the Department of Statistics of The University of Auckland for hospitality during the completion of this work.

References

[1] S-I. Amari, Geometrical theory of asymptotic ancillarity and conditional inference, Biometrika 69 (1982) 1–17.

[2] R.J. Barker, L. Kavalieris, Efficiency gain from auxiliary data requiring additional nuisance parameters, Biometrics 57 (2001) 563–566.

[3] O.E. Barndorff-Nielsen, D.R. Cox, Inference and Asymptotics, Chapman and Hall, London, 1994.

[4] R.M. Fewster, P.E. Jupp, Inference on population size in binomial detectability models, Biometrika 96 (2009) 805–820.

[5] A. Kagan, C.R. Rao, Some properties and applications of the efficient Fisher score, J. Statist. Plann. Inference 116 (2003) 343–352.

[6] C.R. Rao, Linear Statistical Inference and its Applications, 2nd Edition. Wiley, New York, 1973.

[7] R.J. Serfling, Approximation Theorems of Mathematical Statistics, New York, Wiley, 1980.