Ann. Inst. Statist. Math.
Vol. 41, No. 4, 623–648 (1989)

FISHER INFORMATION UNDER RESTRICTION OF SHANNON INFORMATION IN MULTI-TERMINAL SITUATIONS*

SHUN-ICHI AMARI

Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan

(Received December 2, 1988; revised April 10, 1989)

Abstract. Fisher information generally decreases by summarizing observed data into encoded messages. The present paper studies the amount of Fisher information included in independently summarized messages from correlated information sources; that is, the amount of Fisher information when sequences x^N and y^N of N independent observations of random variables x and y are encoded (summarized) independently of each other into messages m_X and m_Y. The problem is to obtain the maximal amount of Fisher information when the size of the summarized data or Shannon message information is limited. The problem is solved in the case of completely compressed symmetric data summarization. An achievable bound is given in the general case. Information geometry, which is a powerful new differential geometrical method applicable to statistics and systems theory, is applied to this problem, proving its usefulness in information theory as well.

Key words and phrases: Shannon information, Fisher information, multi-terminal information theory, information geometry, information loss, data compression, asymptotic theory.

1. Introduction

Let X and Y be two mutually correlated information sources subject to a joint probability distribution $p(x,y)$. Let us consider a situation where N independent observations $x^N = x_1 \cdots x_N$ are obtained at one location and $y^N = y_1 \cdots y_N$ are obtained at another location, where (x_i, y_i), $i = 1, 2, \ldots, N$, are independent pairs of correlated random variables. A usual statistical problem is to make a statistical inference concerning the unknown probability distribution $p(x,y)$ from N independent pairs of observations

*The present work is supported in part by Grant-in-Aid for Scientific Research #61030014, Ministry of Education, Science and Culture of Japan.

$(x_1, y_1), \ldots, (x_N, y_N)$. When the possible candidates for probability distributions are parameterized by a parameter t, we have a statistical model $M = \{p(x, y; t)\}$. Statistical estimation is the problem of obtaining the estimated value $\hat{t}$ of t based on x^N and y^N. The statistical test is the problem of deciding whether a hypothesis H_0: $t = t_0$ is acceptable or not, where the alternative is H_1: $t \neq t_0$. It is known that, when N is large, the performance of the asymptotically best estimator or best test is uniquely characterized by the amount of Fisher information $g(t)$ at the true t or $t = t_0$. Fisher information indeed represents the expected amount of statistical information which is included in observed data (x^N, y^N).

We are forced, in a multi-terminal situation, to encode or summarize x^N and y^N into messages $m_X(x^N)$ and $m_Y(y^N)$ independently and send them separately to a common location. When the transmission rates are restricted, the amounts of Shannon information included in m_X and m_Y are compressed. This reduction of Shannon information gives rise to a reduction of the amount of Fisher information which is utilized for statistical inference. In the present paper we study the amount of Fisher information included in the encoded messages m_X and m_Y under the restriction of the amounts of Shannon information. This is a typical problem of multi-terminal information theory, because the loss of Fisher information is caused by encoding x^N and y^N separately, instead of encoding the pairs (x^N, y^N).

This problem was proposed by T. Berger, and has recently been studied intensively by many researchers. Amari (1986) studied the maximum Fisher information in the case of complete data compression. Achievable bounds are given by Zhang and Berger (1988) and by Ahlswede and Burnashev (1989) in the general case.

The problem can be studied from another point of view, where we evaluate, instead of Fisher information, the asymptotic power exponent of a test H_0: $t = t_0$ against an alternative H_1: $t = t_1$. Ahlswede and Csiszár (1986) gave an achievable bound. Han (1987) gave an improved bound and obtained the optimal power exponent in the case of complete data compression. Amari and Han (1989) applied a new differential geometrical method called information geometry (Amari (1985, 1987*a*, 1987*b*)). They not only elucidated the geometrical structure of the present problem but also gave an explicit solution in the symmetric complete data compression case.

The present paper explains how the differential geometrical notions are connected with Fisher information. By using the geometrical method, we give the maximum amount of Fisher information included in symmetrically encoded, completely compressed data. A good achievable bound in the general case is also given by this approach.

This paper elucidates the intrinsic structure of the present problem from the geometrical point of view. It also demonstrates that the new geometrical method, which has already been proven to be important in statistics (Amari (1982*a*, 1982*b*, 1985, 1987*a*, 1987*b*), Nagaoka and Amari (1982), Amari and Kumon (1983), Kumon and Amari (1983), etc.) is useful in information theory, too (see also Campbell (1985)).

2. Statement of the problem

2.1 *Statistical model*

Let X and Y be two mutually correlated information sources with finite alphabets $A_X = \{0, 1, \ldots, n\}$ and $A_Y = \{0, 1, \ldots, m\}$, respectively. Let x and y be random variables taking values on A_X and A_Y, respectively. Then, the joint probability distribution of (x, y) is specified by a matrix $P = (p_{ij})$,

$$p_{ij} = \text{Prob}\,\{x = i, y = j\}, \qquad i = 0, 1, \ldots, n; \quad j = 0, 1, \ldots, m\,.$$

A pair of the correlated information sources (X, Y) is characterized by this matrix.

Let S_{XY} be the set of all the pairs of information sources, or their joint probability distributions which characterize the pairs,

$$S_{XY} = \{P \mid 1 > p_{ij} > 0, \Sigma p_{ij} = 1\}\,.$$

The set S_{XY} is an open simplex in an $\{(n+1)(m+1) - 1\}$-dimensional Euclidean space, because $\Sigma p_{ij} = 1$ holds. We exclude distributions P whose entries p_{ij} include 0. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ be N pairs of letters, which are independently emitted from a pair of fixed correlated information sources. The sequences of letters are abbreviated as

$$x^N = x_1 x_2 \cdots x_N, \qquad y^N = y_1 y_2 \cdots y_N\,.$$

Statisticians are interested in estimating, or testing, the true joint distribution P from which the data (x^N, y^N) are produced. To this end, a statistical model

$$M = \{P(t)\}, \qquad P \in S_{XY}$$

is sometimes assumed, which is a parameterized family of probability distributions. When t is a scalar parameter, M forms a curve in S_{XY}. When the true distribution P belongs to M, it is specified by the value of the parameter t. Hence, an estimator $\hat{t}$ of t is used for estimating the probability distribution, $\hat{P} = P(\hat{t})$. In the case of a test, a hypothesis of the form H_0: $P = P(t_0)$ is tested against the alternative H_1: $P \neq P(t_0)$.

A statistical model M may be higher-dimensional, where the parameter t is a vector. It can be identical even with S_{XY} itself, if we parameterize S_{XY} by an $(nm + n + m)$-dimensional parameter, say $t = (p_{ij};$

$i \neq 0$ or $j \neq 0$), because p_{00} is calculated from the others. However, we mainly treat the scalar parameter case for the sake of simplicity of presentation. The vector parameter case is studied in a quite similar manner, so that we state only results.

2.2 *Fisher information*

Let us denote a probability distribution P by

$$p(x,y) = \sum p_{ij}\delta_i(x)\delta_j(y)\,, \tag{2.1}$$

where $\delta_i(x) = 1$ when $x = i$ and $\delta_i(x) = 0$ when $x \neq i$. Given a statistical model $M = \{P(t)\}$, let us put

$$l(x,y;t) = \log p(x,y;t)\,. \tag{2.2}$$

Then, the Fisher information $g(t)$ at point $P(t)$ is given by

$$g(t) = E[\{\dot{l}(x,y;t)\}^2]\,, \tag{2.3}$$

where "$\cdot$" implies d/dt and E denotes the expectation with respect to $p(x,y;t)$, i.e.,

$$E[a(x,y)] = \sum_{x,y} p(x,y;t)a(x,y)\,.$$

The Fisher information, when N independent repeated observations (x^N, y^N) are available, is given by using the probability distribution

$$p(x^N, y^N; t) = \prod_{i=1}^{N} p(x_i, y_i; t)\,.$$

The result is just N times the Fisher information $g(t)$ in one observation, showing additivity of Fisher information.

Let f and h be mappings, or encoders, f: $X^N \to M_X$, h: $Y^N \to M_Y$. That is, $m_X = f(x^N)$ and $m_Y = h(y^N)$, $m_X \in M_X$, $m_Y \in M_Y$, are encoded messages of x^N and y^N, respectively. One may say that data x^N and y^N are compressed and encoded into the messages m_X and m_Y, respectively. The joint probability distribution $p(m_X, m_Y; t)$ of m_X and m_Y, when the original distribution is given by $p(x,y;t)$, is easily calculated by using the functions f and h. The amount of Fisher information per letter which the encoded data (m_X, m_Y) carry, is then defined by

$$g_M(t) = N^{-1}E\left[\left\{\frac{d}{dt}\log p(m_X, m_Y; t)\right\}^2\right]. \tag{2.4}$$

It is easy to prove

$$g(t) \geq g_M(t) . \tag{2.5}$$

Fisher information $g(t)$ in the vector parameter case, is a matrix whose (i,j) entry is given by

$$g_{ij}(t) = E\left[\left\{ \frac{\partial}{\partial t^i} l(x,y;t) \right\}\left\{ \frac{\partial}{\partial t^j} l(x,y;t) \right\}\right], \tag{2.6}$$

where $t = (t^1,\dots,t^r)$ is the parameter.

Fisher information represents the amount of statistical information which observed data are expected to carry. It plays a fundamental role in the asymptotic theory of statistical inference (Amari (1985)), as is shown in the following theorems, which hold under some mild regularity conditions. Let $\hat{t} = \hat{t}(m_X, m_Y)$ be an unbiased estimator based on messages m_X and m_Y. Here, an estimator is said to be unbiased, when $E[\hat{t}] = t$ holds for any t.

THEOREM 2.1. *The mean square error of an unbiased estimator is bounded by*

$$N^{-1}E[(\hat{t} - t)^2] \geq g_M(t)^{-1} . \tag{2.7}$$

The equality holds asymptotically (i.e., for large N) for the maximum likelihood estimator $\hat{t}_{\text{m.l.e.}}$ (which is asymptotically unbiased).

Let us consider the problem of testing hypothesis H_0: $t = t_0$ against H_1: $t \neq t_0$. The power function usually approaches 1 as N tends to infinity. To evaluate the power of the test more accurately, we put

$$t_u = t_0 + \frac{u}{\sqrt{N}} , \tag{2.8}$$

and study the power at t_u, which is very close to t_0 when N is large. The power function in a neighborhood of t_0 is then defined by

$$P_N(u) = \text{Prob}\,\{H_0 \text{ is rejected, when the true distribution is } P(t_u)\} ,$$

where N denotes the number of observations. The function

$$P(u) = \lim_{N\to\infty} P_N(u) \tag{2.9}$$

is said to be the (first-order) asymptotic power function.

The significance level α of a test is the probability that H_0 is erroneously rejected when the true probability is t_0. A level α test satisfies

$$P(0) \leq \alpha .$$

Among all the level α tests, a test is said to be asymptotically uniformly most powerful, or efficient, when its asymptotic power function satisfies

$$P(u) \geq \bar{P}(u)$$

at all u compared with the power function $\bar{P}(u)$ of any level α test. We search for the efficient test based on the encoded messages. The following theorem shows that Fisher information g_M represents the characteristic of the efficient test.

THEOREM 2.2. *The asymptotic power function of the efficient test is given by*

$$P(u) = \Phi(u_1 - \sqrt{g_M}\, u) \tag{2.10}$$

in the one-sided case and

$$P(u) = \Phi(u_2 - \sqrt{g_M}\, u) + \Phi(u_2 + \sqrt{g_M}\, u) \tag{2.11}$$

in the two-sided case, where $g_M = g_M(t_0)$ *is Fisher information at* t_0, u_1 *is the one-sided* $100\alpha\%$ *point* (u_2 *is the two-sided* $100\alpha\%$ *point*) *of the unit normal distribution, and* $\Phi(t)$ *is*

$$\Phi(t) = \int_t^{\infty} (2\pi)^{-1/2} \exp\left\{ -\frac{1}{2} u^2 \right\} du .$$

The above two theorems show that Fisher information g_M represents the amount of statistical information involved in the encoded messages m_X and m_Y. The characteristics of the best statistical inference is determined by Fisher information g_M, at least locally (a geometrical theory of the global testing problem is studied in Amari and Han (1989)).

2.3 *Restriction of Shannon information*

Let us consider the following situation where x^N and y^N are encoded into messages $m_X = f(x^N)$, $m_Y = h(y^N)$, and the cardinalities $|M_X|$ and $|M_Y|$ of the message signals are bounded above by 2^{NR_X} and 2^{NR_Y}, respectively. In other words, data x^N and y^N are compressed into m_X and m_Y, whose transmission rates are R_X bits and R_Y bits per letter, respectively. This can be rewritten as

$$\frac{1}{N} I(X^N: M_X) \le R_X, \quad \frac{1}{N} I(Y^N: M_Y) \le R_Y , \tag{2.12}$$

where I is the Shannon mutual information.

Problem. To find the (asymptotic) maximum amount of Fisher information

$$\bar{g}_M(t; R_X, R_Y) = \overline{\lim_{N \to \infty}} \sup g_M(t) , \tag{2.13}$$

where the supremum is taken over all the encoders satisfying the rate constraint (2.12) of Shannon information.

When the cardinalities satisfy

$$\log |M_X| = O(\log N), \quad \log |M_Y| = O(\log N) , \tag{2.14}$$

we have $\lim_{N \to \infty} R_X = \lim_{N \to \infty} R_Y = 0$. This special case is called the complete data compression. We mainly treat this case in this paper.

3. Geometrical preliminaries

3.1 *Tangent space and dual bases*

We present here differential geometry of the set of all the probability distributions on a fixed (finite) number of atoms. Its global characteristics are shown in Amari and Han (1989). This is a special example of "information geometry", which is constructed upon differential geometry of a general family of probability distributions (Amari (1985, 1987*a*)). It gives a powerful new method for studying statistics, systems theory, information theory, etc.

Let x be a random variable taking on a finite number of values $\{0, 1, \dots, n\}$. A probability distribution is written as

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x), \quad p_i > 0 , \tag{3.1}$$

where $p_i = \text{Prob}\,\{x = i\} = p(i)$. The set S_n of all these probability distributions forms an open n-simplex. We introduce two special coordinate systems $\theta = (\theta^1, \theta^2, \dots, \theta^n)$ and $\eta = (\eta_1, \eta_2, \dots, \eta_n)$ to specify points in S_n. The coordinate system η is simply given by

$$\eta_i = p_i, \quad i = 1, 2, \dots, n , \tag{3.2}$$

i.e., we use the last n elements of $(p_0,\dots,p_n)$, where p_0 is determined from

$$p_0(\eta) = 1 - \sum_{i=1}^{n} p_i = 1 - \Sigma\,\eta_i\,. \tag{3.3}$$

Here, p_0 is regarded as a function of η.

The other coordinate system θ is defined by

$$\theta^i = \log\,(p_i/p_0), \qquad i = 1,\dots,n\,. \tag{3.4}$$

Conversely, the probabilities are given by

$$p_i(\theta) = p_0 \exp\,(\theta^i)\,,$$
$$p_0(\theta) = \{1 + \Sigma \exp\,(\theta^j)\}^{-1}\,.$$

The probability distribution specified by θ is written as

$$p(x,\theta) = \sum_{i=0}^{n} p_i(\theta)\delta_i(x)\,.$$

It is known that S_n is an exponential family, and θ is called the canonical parameter (coordinate system) of S_n, and η is called the expectation parameter (coordinate system) of S_n.

Let T_P be the tangent space at point P of S_n. It is an n-dimensional vector space spanned by n vectors $\{e_1, e_2,\dots, e_n\}$, where e_i is the tangent vector along the coordinate curve θ^i; i.e., it represents the direction in which θ^i increases but all the other θ^j are fixed. Mathematicians traditionally denote this tangent vector e_i by

$$\partial_i = \partial/\partial\theta^i\,.$$

Any vector $A \in T_P$ is written by its linear combination,

$$A = \Sigma\,A^i e_i\,.$$

Let us consider the following random variable (a function of x)

$$\partial_i l(x,\theta) = \frac{\partial}{\partial\theta^i}\log p(x,\theta)\,, \tag{3.5}$$

defined at point $P = (p(x,\theta))$. This represents how the log probability changes as θ changes in the direction of the coordinate curve θ^i. Since $\partial_i l$'s $(i = 1,\dots,n)$ are linearly independent, we can identify the tangent space T_P with the vector space spanned by the n random variables $\partial_i l$. Then $\partial_i l$ is regarded as the random variable representation of the tangent vector e_i. Any tangent vector $A = \Sigma A^i e_i$ can be represented by a random variable

$$A(x) = \Sigma A^i \partial_i l(x, \theta)$$

and vice versa. The basis vector $\partial_i l$ is explicitly given by

$$e_i = \partial_i l(x, \theta) = \delta_i(x) - p_i \,. \tag{3.6}$$

Let e^{*i} be the tangent vector along the coordinate curve η_i of the η-system. Then, $\{e^{*1}, \ldots, e^{*n}\}$ forms another basis of T_P. Its random variable representation is given by $\partial^i l(x, \eta)$, where $\partial^i = \partial / \partial \eta_i$. Therefore,

$$e^{*i} = (\partial / \partial \eta_i) \log p(x, \eta) = \frac{\delta_i(x)}{\eta_i} - \frac{\delta_0(x)}{p_0} \,. \tag{3.7}$$

Let us introduce an inner product in T_P by the usual way,

$$\langle A, B \rangle = E[A(x) B(x)] \,, \tag{3.8}$$

where $A(x)$ and $B(x)$ are the random variable representations of $A \in T_P$ and $B \in T_P$, respectively. Then, the matrix $g = (g_{ij})$ defined by

$$g_{ij} = \langle e_i, e_j \rangle = E[\partial_i l \partial_j l] \tag{3.9}$$

is called the metric tensor. Since this is the Fisher information matrix, it is called the Fisher metric. The inner product of two vectors is written by the bilinear form

$$\langle A, B \rangle = \Sigma g_{ij} A^i B^j, \qquad A = \Sigma A^i e_i, \quad B = \Sigma B^i e_i$$

by using their components A^i and B^j. The metric tensor g_{ij} is calculated as

$$g_{ij}(\theta) = p_i(\theta) \delta_{ij} - p_i(\theta) p_j(\theta) \,, \tag{3.10}$$

where δ_{ij} is the Kronecker delta (i.e., the unit matrix).

The metric tensor g^{ij} in the basis $\{e^{*i}\}$ is defined by

$$g^{ij}(\eta) = \langle e^{*i}, e^{*j} \rangle = E[\partial^i l \partial^j l] = \frac{1}{p_i(\eta)} \delta_{ij} + \frac{1}{p_0(\eta)} \,. \tag{3.11}$$

Let $M = \{p(x, t)\}$ be a statistical model. Since $p(x) \in S_n$ is parameterized by θ or η in the whole S_n, the model is represented by the curve

$$\theta = \theta(t) \qquad \text{or} \qquad \eta = \eta(t)$$

in the respective coordinate systems. The tangent vector e_t of the model curve M is given by

$$e_t = \frac{d}{dt}\, l(x, t)\,,$$

where $l(x, t) = \log p(x, \theta(t)) = \log p(x, \eta(t))$. It is rewritten as

$$e_t = \sum \dot{\theta}^i(t) e_i = \sum \dot{\eta}_i(t) e^{*i}\,,$$

where "$\cdot$" denotes d/dt. The Fisher information $g(t)$ of the model is the magnitude of the tangent vector.

$$g(t) = E[\{\dot{l}(x, t)\}^2] = \langle e_t, e_t \rangle = \sum g_{ij} \dot{\theta}^i \dot{\theta}^j = \sum g^{ij} \dot{\eta}_i \dot{\eta}_j\,. \tag{3.12}$$

We now study the dualistic properties of the manifold S_n, which can be understood from the general theory of information geometry (Amari (1985)). The following is a consequence of the e- and m-flatness of S_n.

THEOREM 3.1. *The two bases $\{e_i\}$ and $\{e^{*i}\}$ are mutually dual or reciprocal systems*:

$$\langle e_i, e^{*j} \rangle = \delta_i^j\,. \tag{3.13}$$

The Fisher matrix (g^{ij}) is the inverse of (g_{ij}), and the two bases are related by

$$e_i = \sum g_{ij} e^{*j}, \qquad e^{*j} = \sum g^{ij} e_i\,. \tag{3.14}$$

PROOF. We give here a direct proof. For $i \neq j$, calculations give

$$\begin{aligned}\langle e_i, e^{*j} \rangle &= E[\partial_i l \partial^j l] \\ &= E\left[\{\delta_i(x) - p_i\}\left\{\frac{1}{p_j}\delta_j(x) - \frac{1}{p_0}\delta_0(x)\right\}\right] = 0\,,\end{aligned}$$

and $\langle e_i, e^{*i} \rangle = 1$, proving (3.13). By multiplying g_{jk} with both sides of (3.13) and summing up with respect to j, we have

$$\left\langle e_i, \sum_j g_{jk} e^{*j} \right\rangle = \sum \delta_i^j g_{jk} = g_{ik}\,,$$

which together with (3.9) proves (3.14). It is easy to prove that (g^{jk}) is the inverse of (g_{jk}).

We give some results from the general theory given by Nagaoka and Amari (1982) (see also Amari (1985)). This theory guarantees that there exist two potential functions $\psi(\theta)$ and $\varphi(\eta)$ such that the metric tensors are given by their second derivatives. We have indeed

$$\psi(\theta) = -\log p_0(\theta)\,, \tag{3.15}$$

which is the logarithm of the cumulant generating function, and

$$\varphi(\eta) = -H(\eta) = \sum_{i=0}^{n} p_i(\eta) \log p_i(\eta)\,, \tag{3.16}$$

which is the negentropy. The metric tensors are given by

$$g_{ij} = \partial_i \partial_j \psi(\theta), \qquad g^{ij} = \partial^i \partial^j \varphi(\eta)\,, \tag{3.17}$$

where

$$\partial_i = \partial / \partial \theta^i, \qquad \partial^i = \partial / \partial \eta_i\,.$$

The coordinate transformation between θ and η is given by

$$\theta^i = \partial^i \varphi(\eta), \qquad \eta_i = \partial_i \psi(\theta)\,. \tag{3.18}$$

This is a Legendre transformation, and

$$\psi(\theta) + \varphi(\eta) - \Sigma\, \theta^i \eta_i = 0 \tag{3.19}$$

holds.

We can introduce two mutually dual affine connections, the e- and m-connections. The manifold S_n is flat with respect to these connections, although it is curved with respect to the Riemannian connection. There exists an invariant divergence function in such a dually flat manifold; the divergence function reduces in the present case to the Kullback-Leibler divergence

$$D(P_1, P_2) = \Sigma\, p_{1i} \log \frac{p_{1i}}{p_{2i}}\,, \qquad P_1 = (p_{1i}), \quad P_2 = (p_{2i})\,, \tag{3.20}$$

and the generalized Pythagorian theorem holds in S_n (see Nagaoka and Amari (1982), Amari (1985), Amari and Han (1989)). This plays a fundamental role in the global theory of hypothesis testing in the multi-terminal information-restricted situation. It suffices to note that the divergence reduces to the square of the Riemannian metric in the present local case,

$$D(P, P+dP) = \frac{1}{2}\sum g_{ij}d\theta^i d\theta^j = \frac{1}{2}\,\|d\theta\|^2\,, \tag{3.21}$$

where the coordinates of P and $P+dP$ are θ and $\theta+d\theta$, respectively.

3.2 *Projection*

It is useful to divide the base $\{e_i\} = \{e_1,\dots,e_n\}$ into two parts, say $\{e_1,\dots,e_k; e_{k+1},\dots,e_n\}$. We use indices a, b, c to denote the former part, $\{e_a\}$, $a = 1,2,\dots,k$; and we use indices κ, λ, μ to denote the latter part $\{e_\kappa\}$, $\kappa = k+1,\dots,n$. The dual base $\{e^{*i}\}$, $i = 1,\dots,n$; is also divided into two parts, $\{e^{*i}\} = \{e^{*a}; e^{*\kappa}\}$. The Fisher metric g_{ij} is accordingly partitioned as

$$g_{ij} = \begin{bmatrix} g_{ab} & g_{a\kappa} \\ g_{\lambda b} & g_{\lambda\kappa} \end{bmatrix}, \tag{3.22}$$

where

$$g_{ab} = \langle e_a, e_b\rangle, \qquad g_{a\kappa} = \langle e_a, e_\kappa\rangle\,,$$

$$g_{\lambda b} = \langle e_\lambda, e_b\rangle, \qquad g_{\lambda\kappa} = \langle e_\lambda, e_\kappa\rangle$$

are partitioned minor matrices. Similarly, we have

$$g^{ij} = \begin{bmatrix} g^{ab} & g^{a\kappa} \\ g^{\lambda b} & g^{\lambda\kappa} \end{bmatrix}. \tag{3.23}$$

It is useful to adopt the mixed base $\{e_a; e^{*\kappa}\}$ or

$$\{e_1, e_2,\dots, e_k; e^{*k+1},\dots, e^{*n}\}\,.$$

Let T_1 be the subspace spanned by $\{e_a\} = \{e_1,\dots,e_k\}$, and let T_2 be the subspace spanned by $\{e^{*\kappa}\} = \{e^{*k+1},\dots,e^{*n}\}$. Then, T_1 and T_2 are the orthogonal complements of each other at every point of S_n, and the tangent space T_P is decomposed into the orthogonal direct sum,

$$T_P = T_1 \oplus T_2\,.$$

Let us decompose the tangent vector $e_t = \dot{l}(x,t)$ of the statistical model into the sum of its T_1- and T_2-parts. To this end, we define two matrices

$(\bar{g}^{ab})$ and $(\bar{g}_{\kappa\lambda})$ which are the inverses of the minor matrices (g_{ab}) and $(g^{\kappa\lambda})$, respectively. It should be noted that $(\bar{g}^{ab})$ is different from the (g^{ab}) which is the minor matrix of the entire inverse (g^{ij}) of (g_{ij}), and $(\bar{g}_{\kappa\lambda})$ is different from $(g_{\kappa\lambda})$.

LEMMA 3.1. *Let*

$$X = \Sigma X^a e_a + \Sigma X^\kappa e_\kappa = \Sigma X_a e^{*a} + \Sigma X_\kappa e^{*\kappa}$$

be the representations of a vector X *in the basis* $\{e_i\} = \{e_a; e_\kappa\}$ *and* $\{e^{*i}\} = \{e^{*a}; e^{*\kappa}\}$. *Then, its mixed representation in the basis* $\{e_a; e^{*\kappa}\}$ *is given by*

$$X = \Sigma(\Sigma X_a \bar{g}^{ab}) e_b + \Sigma(\Sigma X^\lambda \bar{g}_{\lambda\kappa}) e^{*\kappa} . \tag{3.24}$$

The square of the magnitude of X *is decomposed as*

$$\|X\|^2 = \Sigma X_a X_b \bar{g}^{ab} + \Sigma X^\kappa X^\lambda \bar{g}_{\kappa\lambda} . \tag{3.25}$$

PROOF. We put

$$X = \Sigma Y^b e_b + \Sigma Y_\lambda e^{*\lambda} .$$

By taking the inner product of X and e_a, we have

$$\langle e_a, X\rangle = \Sigma Y^b \langle e_a, e_b\rangle = \Sigma Y^b g_{ab} ,$$

because of $\langle e_a, e^{*\lambda}\rangle = 0$. On the other hand, because of $\langle e_a, e^{*b}\rangle = \delta_a^b$, we have

$$\langle e_a, X\rangle = X_a .$$

Therefore, we have

$$Y^b = \Sigma X_a \bar{g}^{ab} .$$

Similarly, we have

$$Y_\lambda = \Sigma X^\kappa \bar{g}_{\kappa\lambda} .$$

The orthogonality of T_1 and T_2 yields

$$\|X\|^2 = \Sigma Y^a Y^b g_{ab} + \Sigma Y_\lambda Y_\kappa g^{\lambda\kappa} ,$$

proving (3.25).

In the case of the tangent vector

$$e_t = \Sigma\, \dot{\theta}^a e_a + \Sigma\, \dot{\theta}^\kappa e_\kappa = \Sigma\, \dot{\eta}_a e^{*a} + \Sigma\, \dot{\eta}_\kappa e^{*\kappa} ,$$

the decomposition yields

$$e_t = \Sigma\,(\Sigma\, \dot{\eta}_a \bar{g}^{ab})\, e_b + \Sigma\,(\Sigma\, \dot{\theta}^\lambda \bar{g}_{\lambda\kappa})\, e^{*\kappa} . \tag{3.26}$$

Hence, the Fisher information is decomposed into the sum

$$g(t) = \Sigma\, \dot{\eta}_a \dot{\eta}_b \bar{g}^{ab} + \Sigma\, \dot{\theta}^\kappa \dot{\theta}^\lambda \bar{g}_{\lambda\kappa} . \tag{3.27}$$

3.3 *Multi-terminal source*

We now return to the multi-terminal situation. The manifold S_{XY} of all the probability distributions is identical with S_{mn+m+n}, if we renumber the pairs (x, y) from 0 to $mn + m + n$. However, taking the multi-terminal situation into account, it is better to use the following η- and θ-coordinate systems. The η coordinates are defined in this case by

$$\eta = (\eta_i^X, \eta_j^Y; \eta_{ij}^{XY}), \qquad i = 1, \ldots, n; \quad j = 1, \ldots, m ,$$

where

$$\eta_i^X = p_{i\cdot} = \sum_{j=0}^{m} p_{ij} = \text{Prob}\,\{x = i\} ,$$

$$\eta_j^Y = p_{\cdot j} = \sum_{i=0}^{n} p_{ij} = \text{Prob}\,\{y = j\} ,$$

$$\eta_{ij}^{XY} = p_{ij} .$$

Obviously, the first part (η_i^X, η_j^Y) represents the marginal distributions of P, while η_{ij}^{XY} is partly responsible for their correlations.

The θ-coordinates $\theta = (\theta_X^i, \theta_Y^j; \theta_{XY}^{ij})$ are defined by

$$\theta_X^i = \log\,(p_{i0}/p_{00}) ,$$

$$\theta_Y^j = \log\,(p_{0j}/p_{00}) ,$$

$$\theta_{XY}^{ij} = \log\,(p_{ij}p_{00}/p_{i0}p_{0j}) .$$

Then, we can prove that the basis

$$\{e_X^{*i}, e_Y^{*j}; e_{XY}^{*ij}\}$$

which are the tangent vectors of the coordinate curves of η, and the basis

$$\{e_i^X, e_j^Y; e_{ij}^{XY}\}$$

which are the tangent vectors of the coordinate curves of θ, are mutually dual or reciprocal.

The mixed coordinate system

$$\{\eta_i^X, \eta_j^Y; \theta_{XY}^{ij}\}$$

is adequate for the analysis of the multi-terminal situation. The first two, i.e., η_i^X, η_j^Y, represent the marginal distributions, while the last, θ_{XY}^{ij}, represents the purely correlational properties. When $\theta_{XY}^{ij}=0$, the two random variables x and y are independent. The mixed coordinate system has nice global properties given in Amari and Han (1989).

Let us consider a small change of probability distribution from P to $P+dP$. It is represented by a vector

$$d \log P = \Sigma\, d\eta_i^X e_X^{*i} + \Sigma\, d\eta_j^Y e_Y^{*j} + \Sigma\, d\eta_{ij}^{XY} e_{XY}^{*ij}$$

in terms of the η-coordinate system. The directions in which the marginal distributions do not change (i.e., $d\eta_i^X = d\eta_j^Y = 0$) but the correlations change are hence represented by the vectors e_{XY}^{*ij}. Dually to this, the directions in which the correlations do not change (i.e., $d\theta_{XY}^{ij}=0$) but the marginal distributions do change are represented by the vectors e_i^X and e_j^Y. The subspace T_1 spanned by $\{e_i^X, e_j^Y\}$ is orthogonal to the subspace T_2 spanned by $\{e_{XY}^{*ij}\}$. The subspace T_1 represents the directions in which only the marginal distributions change, and the subspace T_2 represents the directions in which only the correlations change. In this sense, it is very convenient for investigating the intrinsic structure of correlations to use the mixed basis

$$\{e_i^X, e_j^Y; e_{XY}^{*ij}\}\,.$$

We denote the $\{e_i^X, e_j^Y\}$ part by $\{e_a\}$, and the $\{e_{XY}^{*ij}\}$ part by $\{e^{*\kappa}\}$. Then, the decomposition (3.26) of the tangent vector of the model M, as well as the decomposition (3.27) of the Fisher information, holds without any change.

4. Statistical preliminaries

4.1 *Loss of information*

Given a statistical model $M=\{p(x,t)\}$ in S_n, the Fisher information $g(t)$ is the squared norm of the tangent vector e_t or $\dot{l}(x,t)$. When x is encoded in m by a function

$$m = f(x)\,,$$

the Fisher information $g_M(t)$ carried by the statistic or message m is, in general, smaller than $g(t)$ because of the data compression. Its amount is calculated from the probability distribution $p_M(m,t)$ for m:

$$p_M(m,t) = \sum_{x:m(x)=m} p(x,t) ,$$

or its logarithm

$$l_M(m,t) = \log p_M(m,t) .$$

Let $\mathscr{M}$ be the σ-algebra generated by the random variable $m(x)$, i.e., the set of all the random variables which are functions of m. Then, the conditional expectation of a random variable $r(x)$ conditioned on m is a function of m defined by

$$E[r(x)|m] = \Sigma\, r(x)p(x|m) , \tag{4.1}$$

where $p(x|m)$ is the conditional probability of x given m. This is the projection of $r(x)$ to $\mathscr{M}$. A simple calculation verifies

$$\dot{l}_M(m,t) = E[\dot{l}(x,t)|m] . \tag{4.2}$$

Its squared norm $\|\dot{l}_M\|^2$ is the expectation of $\dot{l}_M^2$.

THEOREM 4.1. *Fisher information $g(t) = \|\dot{l}\|^2$ is decomposed into the sum*

$$g(t) = \|\dot{l}\|^2 = \|\dot{l}_M\|^2 + \|\dot{l} - \dot{l}_M\|^2 ,$$

where

$$g_M(t) = \|\dot{l}_M\|^2 = \|E[\dot{l}|m]\|^2$$

represents the amount of Fisher information carried by the encoded message, and the latter is the amount of loss of information.

4.2 *Repeated observations and asymptotics*

Let $x^N = x_1 \cdots x_N$ be N independent random variables subject to the same probability distribution $p(x,\theta)$. Then, their joint probability is given by

$$p_N(x^N;\theta) = \prod_{s=1}^{N} p(x_s,\theta) , \tag{4.3}$$

so that its logarithm is

$$l_N(x^N;\theta) = \sum_{s=1}^{N} l(x_s,\theta) \tag{4.4}$$

$$= \sum_{s=1}^{N}\left[\sum_{i=1}^{n}\theta^i\delta_i(x_s) - \psi(\theta)\right] = N\sum_{i=1}^{n}\{\theta^i\bar{x}_i - \psi(\theta)\}\,,$$

where

$$\bar{x}_i = \frac{1}{N}\sum_{s=1}^{N}\delta_i(x_s) \tag{4.5}$$

is the relative frequency that the letter i is observed (i.e., $x = i$) in N observations x^N. The vector $\bar{x} = (\bar{x}_1,\ldots,\bar{x}_n)$ is called, in information theory, the type of the observed sequence x^N. Equation (4.4) shows that the probability distribution of x^N depends only on the type vector $\bar{x}$. This implies, in terms of statistics, that $\bar{x}$ is a sufficient statistic and that all Fisher information is included in $\bar{x}$.

It is known that the geometrical structure of the manifold S_n is the same as that of the manifold based on N sequence x^N, except that the Fisher information or metric is enlarged N times in the latter space. However, the probability distribution of the random variable $\partial_i l_N(\bar{x},\theta)$ has a simple asymptotic form in the latter case, because the central limit theorem can be applied.

We use the normalized tangent vector e_i^N

$$e_i^N = \frac{1}{\sqrt{N}}\,\partial_i l_N(x^N,\theta) \tag{4.6}$$

in the case of N sequences x^N. Then, its squared norm gives the Fisher information g_{ij} per letter,

$$g_{ij}(\theta) = \langle e_i^N, e_j^N\rangle = \langle e_i, e_j\rangle\,,$$

which is exactly equal to the Fisher information of a single letter. The normalized basis vector can be explicitly written from (4.4) as

$$e_i^N = \sqrt{N}\,(\bar{x}_i - p_i) \overset{\text{def}}{=} \tilde{x}_i\,. \tag{4.7}$$

It is easy to prove that

$$E[\tilde{x}_i] = 0, \qquad E[\tilde{x}_i\tilde{x}_j] = g_{ij}(\theta)\,.$$

Because of the central limit theorem, the vector

$$\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n) \tag{4.8}$$

is (asymptotically) subject to the multivariate normal distribution with mean 0 and covariance matrix $g = (g_{ij})$.

The dual basis

$$e_N^{*i} = \frac{1}{\sqrt{N}} \partial^i \log p(x^N, \eta) \tag{4.9}$$

can be written as

$$e_N^{*i} = \Sigma g^{ij} e_j^N = \Sigma g^{ij} \tilde{x}_j . \tag{4.10}$$

The tangent vector of a statistical model $M = \{p(x^N, t)\}$ based on N observations is given by

$$e_t^N = \Sigma \dot{\theta}^i \tilde{x}_i$$

when the model is specified by $\theta = \theta(t)$.

Let us consider the multi-terminal situation S_{XY} based on N observations. The tangent vectors are given by

$$\begin{aligned}
e_i^X &= \tilde{x}_i = \sqrt{N} (\bar{x}_i - p_{i\cdot}) , \\
e_j^Y &= \tilde{y}_j = \sqrt{N} (\bar{y}_j - p_{\cdot j}) , \\
e_{ij}^{XY} &= \tilde{w}_{ij} = \frac{1}{\sqrt{N}} \sum_{s=1}^{N} (\delta_i(x_s)\delta_j(y_s) - p_{ij}) \\
&= \sqrt{N} (\bar{w}_{ij} - p_{ij}) ,
\end{aligned}$$

where we neglected the suffix N. The term

$$\bar{w}_{ij} = \frac{1}{N} \sum_{s=1}^{N} \delta_i(x_s)\delta_j(y_s)$$

represents the relative frequency of jointly occurring $x = i$ and $y = j$. The quantity $\bar{w} = (\bar{w}_{ij})$ is called the joint type.

The triplet $(\tilde{x}, \tilde{y}, \tilde{w})$ is jointly asymptotically normally distributed. The dual basis vectors e_X^{*i}, e_Y^{*j}, e_{XY}^{*ij} are defined similarly, and are also normally distributed.

5. Fisher information under complete data compression

We study the case where two sequences x^N and y^N are summarized or encoded separately into the respective type vectors $\bar{x}$ and $\bar{y}$. Since the cardinalities of the sets of the type vectors are

$$|\bar{X}| \leq (N+1)^{n+1}, \qquad |\bar{Y}| \leq (N+1)^{m+1},$$

when these type vectors are used as messages to be sent,

$$I(X^N: M_X) = O(\log N), \qquad I(Y^N: M_Y) = O(\log N).$$

Therefore, this is a typical example of complete data compression. Let $g_m(t)$ be the Fisher information included in the compressed data $\bar{x}$ and $\bar{y}$. This is called the marginal Fisher information, because it represents the Fisher information included in the marginal data $\bar{x}$ and $\bar{y}$.

It is easy to show that any symmetric function $f(x^N)$ of $x_1, \ldots, x_N$ can be expressed as a function of $\bar{x}$. Therefore,

$$g_m(t) \geq g_M(t)$$

holds for any symmetric encoding with complete data compression.

The marginal Fisher information $g_m(t)$ is given by

$$g_m(t) = \| E[\dot{l} \mid \tilde{x}, \tilde{y}] \|^2 . \tag{5.1}$$

When the number N of observations is large, the random variables $\dot{l}$, $\tilde{x}_i$, and $\tilde{y}_j$ are all asymptotically jointly normally distributed. The following lemma gives us a good means of calculating $g_m(t)$ in the asymptotic situation.

LEMMA 5.1. *Let $s, t_1, \ldots, t_k$ be jointly normal random variables with zero means. Then, the conditional expectation of s conditioned on t_i is given by a linear combination of t_i,*

$$E[s \mid t_1, \ldots, t_k] = \Sigma c_i t_i . \tag{5.2}$$

The conditional expectation in this normally distributed case is given by the projection of s to the linear subspace spanned by $t_1, \ldots, t_k$. We note that $\dot{l}$, $\tilde{x}_i$, $\tilde{y}_j$ and $\tilde{w}_{ij}$ are asymptotically jointly normally distributed and are tangent vectors of T_P. Let T_m be the subspace spanned by $e_i^X = \tilde{x}_i$ and $e_j^Y = \tilde{y}_j$, which we call the marginal subspace. Let T_c be the subspace spanned by $e_{XY}^{*ij} = \tilde{w}^{ij}$, which we call the correlational subspace. Then, the tangent space T_P is decomposed into the orthogonal sum,

$$T_P = T_m \oplus T_c \,. \tag{5.3}$$

The score function $\dot{l}(x^N, y^N; t) \in T_P$ is decomposed as

$$\dot{l} = E[\dot{l} \mid T_m] + E[\dot{l} \mid T_c] \,, \tag{5.4}$$

where $\dot{l}_m = E[\dot{l} \mid T_m]$ is the projection of $\dot{l}$ to T_m and is given by the conditional expectation. We have from (3.24)

$$\dot{l}_m = E[\dot{l} \mid T_m] = \Sigma(\Sigma \dot{\eta}_a \bar{g}^{ab}) e_b \,,$$
$$\dot{l}_c = E[\dot{l} \mid T_c] = \Sigma(\Sigma \dot{\theta}^\lambda \bar{g}_{\lambda\kappa}) e^{*\kappa} \,,$$

where Roman indices a, b, etc. stand for the indices of the basis vectors $\{e_i^X, e_j^Y\}$ of the marginal subspace T_m, and the Greek indices κ, λ, etc. stand for the indices of the basis vectors $\{e_{XY}^{*ij}\}$ of the correlational subspace. The relations (3.27) or Theorem 4.1 give the main theorem.

THEOREM 5.1. *The maximal Fisher information under symmetric complete data compression is given by*

$$g_m(t) = \Sigma \dot{\eta}_a \dot{\eta}_b \bar{g}^{ab} \tag{5.5}$$

and the loss of Fisher information is given by

$$g_c(t) = \Sigma \dot{\theta}^\kappa \dot{\theta}^\lambda \bar{g}_{\lambda\kappa} \,. \tag{5.6}$$

The asymptotically best estimator $\hat{t}$ based on $\bar{x}$ and $\bar{y}$ is obtained by solving the projected likelihood equation

$$\dot{l}_m(\bar{x}, \bar{y}; t) = E[\dot{l} \mid T_m] = 0 \,. \tag{5.7}$$

In order to study the characteristic of $\hat{t}$, we put

$$\dot{\bar{\theta}}^a = \Sigma \bar{g}^{ab} \dot{\eta}_b \,, \tag{5.8}$$

$$\hat{t} = t + \varepsilon \,, \tag{5.9}$$

where t is the true parameter. Then, the likelihood equation is rewritten as

$$\dot{l}_m = \Sigma \dot{\bar{\theta}}^a (t + \varepsilon)[\bar{x}_a - \eta_a(t + \varepsilon)] \,,$$

where $\bar{x}_a$ represents $(\bar{x}_i, \bar{y}_j)$. By expanding this and neglecting the higher order terms, the error ε is obtained with $\tilde{x}_a = (\tilde{x}_i, \tilde{y}_j)$ as

$$\varepsilon = (1/\sqrt{N}) g_m^{-1} \Sigma (\dot{\bar{\theta}}^a \tilde{x}_a) , \tag{5.10}$$

because of

$$g_m = \Sigma \dot{\bar{\theta}}^a \dot{\eta}_a = \Sigma \dot{\eta}_a \dot{\eta}_b \bar{g}^{ab} .$$

The mean square error of the estimator $\hat{t}$ is easily calculated as

$$\|\varepsilon\|^2 = N^{-1} g_m^{-1} ,$$

proving that $\hat{t}$ is indeed the best estimator satisfying the Cramér-Rao bound.

6. Composition of the best estimator $\hat{t}$ from partial data

Let us consider the best estimator $\hat{t}_X$ based only on $\bar{x}$. It is obtained from the marginal model, and so is the best estimator $\hat{t}_Y$ based only on $\bar{y}$. Let g_X and g_Y be Fisher information of the marginal models $\{p_X(x;t)\}$ and $\{p_Y(y;t)\}$, respectively,

$$p_X(x;t) = \sum_y p(x,y;t) ,$$

$$p_Y(y;t) = \sum_x p(x,y;t) .$$

They are given by

$$g_X(t) = \|\dot{l}_X\|^2, \qquad g_Y(t) = \|\dot{l}_Y\|^2 ,$$

where $l_X = \log p_X$ and $l_Y = \log p_Y$. The Fisher information included in $\bar{x}$ ($\bar{y}$) only is given by g_X (g_Y). It is easy to show

$$\dot{l}_X = E[\dot{l} \mid \tilde{x}], \qquad \dot{l}_Y = E[\dot{l} \mid \tilde{y}] . \tag{6.1}$$

The estimators $\hat{t}_X$ and $\hat{t}_Y$ are the maximum likelihood estimators of the marginal models, and their errors ε_X and ε_Y

$$\hat{t}_X = t + \varepsilon_X, \qquad \hat{t}_Y = t + \varepsilon_Y \tag{6.2}$$

are written as

$$\sqrt{N}\, \varepsilon_X = g_X^{-1} \Sigma \dot{\bar{\theta}}_X^i \tilde{x}_i, \qquad \sqrt{N}\, \varepsilon_Y = g_Y^{-1} \Sigma \dot{\bar{\theta}}_Y^j y_j , \tag{6.3}$$

where

$$\mathring{\theta}^i_X = \sum g^{ij}_X \dot{\eta}^X_i, \qquad \mathring{\theta}^j_Y = \sum g^{jk}_Y \dot{\eta}^Y_k$$

are the θ-coordinates of the marginal models. The matrix g^{ij}_X (g^{jk}_Y) is the inverse of the minor matrix

$$g^X_{ij} = E[\tilde{x}_i \tilde{x}_j], \qquad g^Y_{jk} = E[\tilde{y}_j \tilde{y}_k].$$

Although $\hat{t}_X$ and $\hat{t}_Y$ are the best estimators based on $\bar{x}$ only and $\bar{y}$ only, respectively, their combination does not give the best one, $\hat{t}$, based on both $\bar{x}$ and $\bar{y}$. This implies some information is lost by separately summarizing $\bar{x}$ and $\bar{y}$ into the best estimators $\hat{t}_X$ and $\hat{t}_Y$, respectively.

It should be noted that $\bar{x}$ and $\hat{t}_X$ have the same amount of Fisher information. Hence, the lost information is included in some asymptotically ancillary statistic.

We first show the best estimator $\check{t}$ obtained from $\hat{t}_X$ and $\hat{t}_Y$. Let $c(t)$ be the correlation of $\dot{l}_X$ and $\dot{l}_Y$,

$$c(t) = E[\dot{l}_X(x;t)\dot{l}_Y(y;t)]. \tag{6.4}$$

Let G be a 2×2 matrix

$$G = \begin{bmatrix} g_X & c \\ c & g_Y \end{bmatrix}. \tag{6.5}$$

We define a_X and a_Y by

$$\begin{bmatrix} a_X \\ a_Y \end{bmatrix} = G^{-1} \begin{bmatrix} g_X \\ g_Y \end{bmatrix}. \tag{6.6}$$

THEOREM 6.1. *The best estimator obtained from $\hat{t}_X$ and $\hat{t}_Y$ is their weighted sum,*

$$\check{t} = (a_X g_X \hat{t}_X + a_Y g_Y \hat{t}_Y)/(a_X g_X + a_Y g_Y), \tag{6.7}$$

where

$$g_s = a_X g_X + a_Y g_Y = [g_X, g_Y] G^{-1} \begin{bmatrix} g_X \\ g_Y \end{bmatrix} \tag{6.8}$$

is the amount of Fisher information included in $(\hat{t}_X, \hat{t}_Y)$.

The proof is not difficult and is omitted. It is also not difficult to prove

$$g_s \leq g_m \, .$$

This is because $\dot{l}_m$ cannot in general be represented by a linear combination of $\dot{l}_X$ and $\dot{l}_Y$. However, in the binary case where $n = m = 1$, i.e., x and y takes on $\{0, 1\}$, T_m is two-dimensional, and is spanned by $\dot{l}_X$ and $\dot{l}_Y$. Hence, $\check{t}$ gives $\hat{t}$ and $g_s = g_m$ in this case. This shows that it is inadequate to summarize the data $\bar{x}$ and $\bar{y}$ into $\hat{t}_X$ and $\hat{t}_Y$.

We have another method of data summarization, which causes no loss of information. Let us rewrite $\dot{l}_m = \Sigma \dot{\bar{\theta}}^a \tilde{x}_a$ by decomposing e_a into the $\bar{x}$ part and $\bar{y}$ part explicitly. We then have

$$\dot{l}_m = \tilde{l}_X + \tilde{l}_Y \, , \tag{6.9}$$

where

$$\tilde{l}_X = \Sigma \dot{\bar{\theta}}^i_X \tilde{x}_i, \qquad \tilde{l}_Y = \Sigma \dot{\bar{\theta}}^j_Y \tilde{y}_j \, . \tag{6.10}$$

Here, $\dot{\bar{\theta}}^i_X$ can be written as

$$\dot{\bar{\theta}}^i_X = \Sigma \bar{g}^{ij}_X \dot{\eta}^X_j + \Sigma \bar{g}^{ik}_{XY} \dot{\eta}^Y_k \, ,$$

etc., where

$$\bar{g}^{ab} = \begin{bmatrix} \bar{g}^{ij}_X & \bar{g}^{ik}_{XY} \\ \bar{g}^{mj}_{YX} & \bar{g}^{mk}_Y \end{bmatrix}$$

is the partitioned form of the inverse $\bar{g}^{ab}$ of g_{ab}. We call $\tilde{l}_X$ and $\tilde{l}_Y$ the quasi marginal likelihoods.

The quasi marginal likelihood equations

$$\tilde{l}_X(\bar{x}, \tilde{t}_X) = 0, \qquad \tilde{l}_Y(\bar{y}, \tilde{t}_Y) = 0 \tag{6.11}$$

give two estimators $\tilde{t}_X$ and $\tilde{t}_Y$ which are determined from $\bar{x}$ only and $\bar{y}$ only, respectively.

THEOREM 6.2. *The two estimators $\tilde{t}_X$ and $\tilde{t}_Y$ together include the full amount g_m of Fisher information. The efficient estimator $\hat{t}$ is reconstructed from them by*

$$\hat{t} = (\tilde{g}_X \tilde{t}_X + \tilde{g}_Y \tilde{t}_Y)/g_m \, , \tag{6.12}$$

where

$$\tilde{g}_X = \Sigma\, \dot{\bar{\theta}}^i_X \dot{\eta}^X_i, \qquad \tilde{g}_Y = \Sigma\, \dot{\bar{\theta}}^j_Y \dot{\eta}^Y_j\,. \tag{6.13}$$

PROOF. It is easy to prove that the respective errors, defined by

$$\tilde{t}_X = t + \tilde{\varepsilon}_X, \qquad \tilde{t}_Y = t + \tilde{\varepsilon}_Y \tag{6.14}$$

are asymptotically written as

$$\sqrt{N}\, \tilde{\varepsilon}_X = \tilde{g}_X^{-1}\, \Sigma\, \dot{\bar{\theta}}^i_X \tilde{x}_i, \qquad \sqrt{N}\, \tilde{\varepsilon}_Y = g_Y^{-1}\, \Sigma\, \dot{\bar{\theta}}^j_Y \tilde{y}_j\,, \tag{6.15}$$

and

$$g_m = \tilde{g}_X + \tilde{g}_Y \tag{6.16}$$

hold. The result (6.12) is easily obtained from the relations (6.13) ~ (6.16). Since the variance of $\hat{t}$ is asymptotically equal to g_m^{-1}, they together include an amount g_m of Fisher information.

It should be noted that $\|\hat{\varepsilon}_X\|^2 \le \|\tilde{\varepsilon}_X\|^2$, so that $\tilde{t}_X$ and $\tilde{t}_Y$ are worse than $\hat{t}_X$ and $\hat{t}_Y$, respectively. However, they together include more information than $\hat{t}_X$ and $\hat{t}_Y$ do. This is included in the statistics $\hat{t}_X - \tilde{t}_X$ and $\hat{t}_Y - \tilde{t}_Y$, which are asymptotically ancillary, including no Fisher information by themselves. However, they include conditional Fisher information conditioned on $\hat{t}_X$ and $\hat{t}_Y$.

It is obvious that we can construct many efficient tests, such as the likelihood ratio test, the Wald test, the Rao test, etc. by utilizing the full amount g_m of Fisher information from $\bar{x}$ and $\bar{y}$. In some cases, $\hat{t}$ or $\tilde{t}_X$ and $\tilde{t}_Y$ are again sufficient to construct such a test.

7. The Fisher information based on noisy data

Let us consider two noisy memoryless channels C_X and C_Y, with input alphabets X and Y, and output alphabets U and V, respectively. The channels are specified by the conditional probability distributions $p(u|x)$ and $p(v|y)$. When data x and y are transmitted letterwise through these channels, respectively, the amounts of Shannon information included in the output letters u and v are given by the transmission rates

$$R_X = I(X; U), \qquad R_Y = I(Y; V)\,,$$

where $I(X; U)$ is Shannon's mutual information between X and U, and so on. We study the amount of Fisher information involved in the transmitted noisy data $u^N = u_1 \cdots u_N$ and $v^N = v_1 \cdots v_N$. This problem is interesting not only in its own right, but because its solution gives an achievable bound of the Fisher information under the rate restriction within R_X and R_Y of the Shannon information.

The geometrical method is applicable to this problem. A probability distribution $P = (p(x,y))$ naturally induces a joint probability distribution $Q = (q(x,y,u,v))$ over four random variables X, Y, U, V as

$$q(x,y,u,v) = p(x,y)p(u|x)p(v|y) . \tag{7.1}$$

(The four random variables satisfy the Markovian condition

$$U—X—Y—V$$

in the above case. It is important to study the case with

$$U—X—Y, \qquad X—Y—V$$

in order to obtain a good achievable bound.) A statistical model $p(x,y;t)$ induces an enlarged model $q(x,y,u,v;t)$.

We can study the geometrical structure of the manifold consisting of all the Q's in a similar manner. Refer to Amari and Han (1989) in more detail. We define the observable space T_0 at each point of Q, which is a subspace of the tangent space T_Q. It is spanned by the vectors e_U, e_V, e_{UV},

$$T_0 = \{\text{vectors spanned by } e_U, e_V, e_{UV}\} ,$$

where e_U etc. stand for vectors

$$e_U = \partial / \partial \theta_U^i, \qquad e_V = \partial / \partial \theta_V^j, \qquad e_{UV} = \partial / \partial \theta_{UV}^{ij} .$$

Since we have type vectors $\bar{u}$, $\bar{v}$, and a joint type vector $\overline{uv}$ from the transmitted messages u^N and v^N, we have the following theorem.

THEOREM 7.1. *The Fisher information based on u^N and v^N is given by*

$$g_0 = \| E[\dot{l} \mid T_0] \|^2 . \tag{7.2}$$

Let T_0' be the subspace spanned by T_0, e_X and e_Y. Since $\bar{x}$ and $\bar{y}$ can be sent with asymptotically zero rates when coding is admitted, we have the following achievable bound.

THEOREM 7.2. *An achievable bound of Fisher information under* *the rate restriction is given by*

$$\bar{g}(R_X, R_Y) = \sup \| E[\dot{l} \mid T_0'] \|^2 , \tag{7.3}$$

where the supremum is taken over all the channels with given rates R_X and R_Y.

Acknowledgements

The author would like to thank Professor K. Murota and Professor T. S. Han whose criticism and suggestions were very useful for improving the present results.

REFERENCES

Ahlswede, R. and Burnashev, M. V. (1989). On minimax estimation in the presence of side information about remote data, *Ann. Probab.* (to appear).

Ahlswede, R. and Csiszár, I. (1986). Hypothesis testing with communication, *IEEE Trans. Inform. Theory*, **32**, 533–542.

Amari, S. (1982*a*). Differential geometry of curved exponential families—curvatures and information loss, *Ann. Statist.*, **10**, 357–385.

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

Amari, S. (1985). Differential geometrical methods in statistics, *Lecture Notes in Statistics*, **28**, Springer, Berlin-Heidelberg.

Amari, S. (1986). Data compression and statistical inference, *J. Inst. Electr. & Comm. Eng. Japan*, **J69-A**, 757–763 (in Japanese).

Amari, S. (1987*a*). Differential geometrical theory of statistics, *Differential Geometry in Statistical Inference*, *IMS Monograph Series*, **10**, 19–94, IMS, California.

Amari, S. (1987*b*). Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections, and divergence, *Math. Systems Theory*, **20**, 53–82.

Amari, S. and Han, T. S. (1989). Statistical inference under multi-terminal rate restrictions—a differential geometrical approach, *IEEE Trans. Inform. Theory*, **35**, 217–227.

Amari, S. and Kumon, M. (1983). Differential geometry of Edgeworth expansions in curved exponential family, *Ann. Inst. Statist. Math.*, **35**, 1–24.

Campbell, L. L. (1985). The relation between information theory and the differential geometry approach to statistics, *Inform. Sci.*, **35**, 199–210.

Han, T. S. (1987). Hypothesis testing with multi-terminal data compression, *IEEE Trans. Inform. Theory*, **33**, 759–772.

Kumon, M. and Amari, S. (1983). Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference, *Proc. Roy. Soc. London Ser. A*, **387**, 429–458.

Nagaoka, H. and Amari, S. (1982). Differential geometry of smooth families of probability distributions, Tech. Report, METR 82-7, University of Tokyo.

Zhang, Z. and Berger, T. (1988). Estimation via compressed information, *IEEE Trans. Inform. Theory*, **34**, 198–211.