THE GENERAL THEORY OF CANONICAL CORRELATION AND ITS RELATION TO FUNCTIONAL ANALYSIS

E. J. HANNAN [1]

(received 27 July 1960, revised 9 January 1961)

1. Introduction

The classical theory of canonical correlation is concerned with a standard description of the relationship between any linear combination of p random variables x_s and any linear combination of q random variables y_t insofar as this relation can be described in terms of correlation. Lancaster [1] has extended this theory, for $p = q = 1$, to include a description of the correlation of any function of a random variable x and any function of a random variable y (both functions having finite variance) for a class of joint distributions of x and y which is very general. It is the purpose of this paper to derive Lancaster's results from general theorems concerning the spectral decomposition of operators on a Hilbert space. These theorems lend themselves easily to the generalisation of the theory to situations where p and q are not finite. In the case of Gaussian, stationary, processes this generalisation is equivalent to the classical spectral theory and corresponds to a canonical reduction of a (finite) sample of data which is basic. The theory also then extends to any number of processes. In the Gaussian case, also, the present discussion is connected with the results of Gelfand and Yaglom [2] relating to the amount of information in one random process about another.

For the Gaussian case the theory can be presented in a very general form by commencing from assumptions concerning the symmetry of the covariance function of a vector-valued stochastic process and using the harmonic analysis for the corresponding group of symmetries to produce a spectral theory upon which the canonical correlation theory may be founded.

2. The Canonical Correlation of Stochastic Processes

We consider two stochastic processes, that is two families of random variables

[1] This research was mainly done at the University of North Carolina with support from the Office of Naval Research.

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

$$\{x_s, s \in \mathscr{S}\}; \quad \{y_t, t \in \mathscr{T}\}$$

where $\mathscr{S}$ and $\mathscr{T}$ are two index sets. We do not require that the two families be independent. We have mainly in mind the case where $\mathscr{S}$ and $\mathscr{T}$ are finite or are the set of all integers or all reals but the discussion is general.

We consider the space Ω which is the cartesian product of a family of copies of the real line, one copy for each point in $\mathscr{S}$ and one copy for each point in $\mathscr{T}$. It is a classic fact (Kolmogoroff [4]) that the joint distributions of finite sets of the x_s and y_t, provided they are consistent (Kolmogoroff [4], p. 29), may be used to institute a probability measure, μ, on a Borel field of sets in Ω which includes all cylinder sets having a Borel set in a finite dimensional Euclidean space as base. We call $\mathscr{H}$ the Hilbert space of all square integrable complex valued functions of $\omega \in \Omega$ and indicate the inner product in $\mathscr{H}$ by the square bracket,

$$[a, b] = \int_\Omega a(\omega)\overline{b(\omega)}\mathrm{d}\mu(\omega)$$
$$[a, a] = ||a||^2.$$

We shall use the letters $f, g, h, \cdots$ for those elements of $\mathscr{H}$ which are functions only of the coordinates of ω belonging to $\mathscr{S}$ and $u, v, w, \cdots$ for those depending only on the coordinates belonging to $\mathscr{T}$. The set of all such f forms a subspace (closed) of $\mathscr{H}$ which we call $\mathscr{M}$, and the set of all u forms a subspace which we call $\mathscr{N}$. We shall use the notation

$$[f, g] = (f, g)_{\mathscr{M}}, \qquad (f, f)_{\mathscr{M}} = ||f||^2_{\mathscr{M}}$$
$$[u, v] = (u, v)_{\mathscr{N}}, \qquad (u, u)_{\mathscr{N}} = ||u||^2_{\mathscr{N}}.$$

By a well known property of bounded linear functionals on a Hilbert space we know that

$$[f, u] = (Af, u)_{\mathscr{N}}, \tag{1}$$

where A is a linear operator from $\mathscr{M}$ to $\mathscr{N}$, and putting $u = Af$ in (1) we derive, from Schwartz's inequality, that

$$||Af||_{\mathscr{N}} \leqq ||f||_{\mathscr{M}},$$

so that A is bounded by 1. Of course we also have

$$[f, u] = (f, A^*u)_{\mathscr{M}}$$

where A^* is the adjoint of A.

In the classical canonical correlation theory we use the $x_s (s = 1, \cdots, p)$ and $y_t (t = 1, \cdots, q)$ as bases in the spaces $\mathscr{M}$ and $\mathscr{N}$ (respectively) so that the matricial form of A with respect to these bases, which we call $[A]$, is the matrix of regression coefficients of the x_s on the y_t, i.e.

$$x' = (x_1, \cdots, x_p) = (y_1, \cdots, y_q)[A] + (z_1, \cdots, z_q) = y'[A] + z',$$

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

where the cross product, $\mathscr{E}(z_s y_t)$, is zero for all s, t. Here we use the symbol $\mathscr{E}$ for 'expectation'. It is well known that we may find orthogonal matrices P and Q so that $Q[A]P'$ has non zero elements only in the diagonal commencing from the top left hand corner. Each element of Px will now depend linearly upon at most one element of Qy. Those that are dependent on an element of Qy (at most $\min(p, q)$ in number), after normalisation, are called the canonical x variables and the element of Py on which they depend, after normalisation, are called the canonical y variables. Thus the operator A may be thought of as first expressing each (linear) function of the x_s linearly in terms of the canonical variables of the x set and then replacing each of these by the corresponding members of the y set multiplied by the appropriate diagonal element, ρ_j, let us say, of $Q[A]P'$, adjusted to take account of the normalisations. That is we may write

$$A = W \sum \rho_j E_j$$

where E_j projects onto the subspace of $\mathscr{M}$ spanned by the canonical x variables corresponding to this ρ_j (there may be more than one for the same ρ_j) and W maps the space spanned by the canonical x variables onto the space spanned by the canonical y variables and is the null operator on the orthogonal complement, in $\mathscr{M}$, of the first-mentioned of these two spaces. The canonical decompositions of a (linear) function, f, of the x set and a linear function, u, of the y set are then given by

$$f = \sum E_j f, \quad u = W \sum E_j W^* u + u_1.$$

Here we have included among the E_j an operator which projects onto the orthogonal complement, in $\mathscr{M}$, of the space spanned by the canonical x variables. The operator W^* is the adjoint of W, mapping the y set onto the canonical x set and operating as the null operator on the orthogonal complement in $\mathscr{N}$ of the former set, while u_1 is orthogonal to all of the elements of the canonical set. Thus the second of these formulae expresses u in terms of the canonical y variables and a residual u_1. Finally

$$[f, u] = (Af, u)_{\mathscr{N}} = (W \sum \rho_j E_j f, W \sum E_j W^* u)_{\mathscr{N}} = \sum \rho_j (E_j f, E_j W^* u)_{\mathscr{M}}.$$

It is these last three displayed formulae that we seek to generalise. Before doing this we rewrite the first in the form

$$A = W \int_{0-}^{1} \rho \, \mathrm{d}E(\rho)$$

after defining

$$E(\rho) = \sum_{\rho_j \leqq \rho} E_j$$

where the summation is over those E_j for which $\rho_j \leqq \rho$. We have included a component corresponding to $\rho = 0$ so that $E(1)$ is the identity operator on $\mathscr{M}$. The other formulae are modified accordingly. Naturally in the general situa-

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

tion which we shall discuss below this representation of A will involve limiting processes and it is to be expected that the points where $E(\rho)$ increases need not be denumerable. We refer the reader who is not familiar with the spectral theory of operators in a Hilbert space to [7] pp. 261—277.

In the general situation we may write (Naimark [3], p. 284)

$$A = WB$$

where $B = (A^*A)^{1/2}$ (taking the positive square root) and W is 'partially isometric' in that it maps the closure of $B\mathscr{M}$ isometrically onto the closure of $A\mathscr{M}$ and is the null operator on $(B\mathscr{M})^\perp$ (the orthogonal complement of $B\mathscr{M}$ in $\mathscr{M}$). Thus

$$A = W\int_{0-}^{1} \rho \mathrm{d}E(\rho) \tag{2}$$

where $E(\rho)$ is a resolution of the identity in $\mathscr{M}$ relative to the Borel subsets of $[0, 1]$. It follows from (1) that unity is always a characteristic value of B corresponding to the characteristic function which is identically 1. Indeed

$$(A^*A1, g)_{\mathscr{M}} = (A1, Ag)_{\mathscr{N}} = [1, Ag] = (1, Ag]_{\mathscr{N}} = [1, g] = (1, g)_{\mathscr{M}}$$

for all $g \in \mathscr{M}$.

We now put

$$f = \int_{0-}^{1} \mathrm{d}E(\rho)f, \quad u = W\int_{0-}^{1} \mathrm{d}E(\rho)W^*u + u_1. \tag{3}$$

Here $u_1 \in (A\mathscr{M})^\perp$, since

$$\begin{aligned}(Af, u)_{\mathscr{N}} &= (Af, WW^*u)_{\mathscr{N}} + (Af, u_1)_{\mathscr{N}} \\ &= (W^*Af, W^*u)_{\mathscr{M}} + (Af, u_1)_{\mathscr{N}} \\ &= (Af, u)_{\mathscr{N}} + (Af, u_1)_{\mathscr{N}}.\end{aligned}$$

These formulae (3) constitute the generalisation of those given earlier which expressed f and u in terms of the canonical variables of the two sets.

Then, finally

$$\begin{aligned}[f, u] &= \int_{0-}^{1} \rho \mathrm{d}(WE(\rho)f, \quad WE(\rho)W^*u)_{\mathscr{N}} \\ &= \int_{0-}^{1} \rho \mathrm{d}(E(\rho)f, E(\rho)W^*u)_{\mathscr{M}}.\end{aligned} \tag{4}$$

3. The Canonical Correlation of Two Finite Sets

It is shown in [2] that the information (in Shannon's sense) in one set of p random variables about another set of q can be finite only if the probability distribution, in the $p + q$ dimensional space, induced by the joint distribution is absolutely continuous with respect to the product measure induced by the marginal distributions of the two sets. If we use $H(x, y)$, $M(x)$, $N(y)$ for these three distribution functions, then when this absolute continuity condition is satisfied, we shall put

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

$$A(x, y) = \frac{\partial H(x, y)}{\partial M(x) \partial N(y)}$$

for the Radon-Nikodym derivative of the H measure with respect to the product measure. This derivative is the (essentially) unique function for which

$$\int_{\mathscr{C}} \mathrm{d}H(x, y) = \int_{\mathscr{C}} \frac{\partial H(x, y)}{\partial M(x) \partial N(y)} \mathrm{d}M(x) \mathrm{d}N(y)$$

for every measurable set $\mathscr{C}$. (See Halmos [10] p. 128.)

Thus A in (1) is defined by

$$Af = \int A(x, y) f(x) \mathrm{d}M(x). \tag{5}$$

The typical case where this is not so is that for $p = q = 1$, where there is a concentration of mass along a line.

When $H(x, y)$ is Gaussian the amount of information becomes ([2], p. 217)

$$-\tfrac{1}{2} \sum \log(1-\rho_j^2)$$

where ρ_j is the j^{th} canonical correlation in the classical theory. Now

$$\tfrac{1}{2} \log \operatorname{tr} B^2 = -\tfrac{1}{2} \sum \log(1-\rho_j^2) \tag{6}$$

where the trace is defined as

$$\operatorname{tr} B^2 = \sum_t (B^2 \psi_t, \psi_t)_{\mathscr{M}} \tag{7}$$

and the ψ_t form any complete orthonormal sequence in the (separable) Hilbert space $\mathscr{M}$.

The validity of (6) follows (see Lancaster [1]) by considering the normalised Chebyshev-Hermite polynomials ([5], p. 133), $H_i(\xi_j)$, $i = 1, 2, \cdots$; $j = 1, \cdots, p$, in the variables ξ_j which are canonical in the classical theory. If η_j is the corresponding canonical y variable then the only non zero correlations between these polynomials are between $H_i(\xi_j)$ and $H_i(\eta_j)$. The finite products

$$\psi_t = \prod_{j=1}^{p} H_{i_t}(\xi_j), \prod_{j=1}^{p} H_{i_t}(\eta_j)$$

are dense in $\mathscr{M}$ and $\mathscr{N}$ respectively and the only non zero correlations are between corresponding pairs, a typical pair having correlation

$$\prod_j \rho_j^{i_t}.$$

Thus, from (7)

$$\begin{aligned} \tfrac{1}{2} \log \operatorname{tr} B^2 &= \tfrac{1}{2} \log \sum_t{}' \{\prod_j \rho_j^{2i_t}\} \\ &= \tfrac{1}{2} \log \prod \frac{1}{1-\rho_j^2} = -\tfrac{1}{2} \sum \log(1-\rho_j^2) \end{aligned}$$

the summation $\sum'$ being over all different products.

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

If we use 1/2 log tr B^2 as a measure of the information in one set of random variables about the other, even when the distributions are not Gaussian, and require this measure of information to be finite then the set of values ρ for which $E(\rho)$ increases (i.e. the spectrum of B and B^2) can have no limit point other than zero. For if $\rho_0 \neq 0$ is such a limit point we can find an ε, $0 < \varepsilon < \rho_0$, and an infinite sequence of orthonormal elements, ψ_t, which belong to the space on which $\{E(\rho_0+\varepsilon) - E(\rho_0-\varepsilon)\}$ projects ([7], p. 364). The contribution of these ψ_t to the right hand side of (7) would then be infinite. An operator which has at most one limit point in its spectrum, and this at the origin, is called compact (completely continuous) (see [7] sections 85 and 93). Since B is compact and B^2 has finite trace it follows (see for example the argument on pp. 242—3 of [7]) that $A(x, y)$ is square integrable with respect to $\mathrm{d}M(x)\mathrm{d}N(y)$. The analogy with the classical theory is now almost complete for B will be generated by the square integrable kernel

$$B(x, z) = \sum_0^\infty \rho_j\phi_j(x)\phi_j(z), \quad \sum_0^\infty \rho_j^2 < \infty$$

so that the $\phi_j(x)$ become the canonical x variables while

$$\psi_j(y) = W\phi_j(x)$$

are the canonical y variables. Finally

$$[f, u] = \sum \alpha_j\beta_j\rho_j$$

with

$$f = \sum_0^\infty \alpha_j\phi_j(x) + f_1, u = \sum_0^\infty \beta_j\psi_j(y) + u_1,$$

where f_1 is orthogonal to all ϕ_j, u_1 to all ψ_j.

This is the case discussed, for $p = q = 1$, by Lancaster [1]. Lancaster begins from Karl Pearson's coefficient of mean square contingency, Φ^2, which he generalises defining it (in our notation) by

$$\Phi^2 = \left[\int\int_{-\infty}^{\infty} \{\mathrm{d}F(x, y)\}^2 \Big/ \{\mathrm{d}M(x)\mathrm{d}N(y)\}\right] - 1$$

where the integral is taken as a Hellinger integral. This is (tr $B^2 - 1$) in our notation, the subtraction of unity being made so as to make Φ^2 zero when the only point of the spectrum of B is unity, which is a characteristic value of unit multiplicity, so that the correlation (after mean correction) is zero between any two functions. Lancaster commences from the assumption that Φ^2 is finite which is equivalent to our assumption that $\frac{1}{2}$ log tr B^2 is finite, and then develops the canonical correlation theory for this case. Of course $\frac{1}{2}$ log tr B^2 is not a wholly satisfactory measure of the information in one set of random variables about the other when the distributions are not Gaussian so that the justification, for assuming $A(x, y)$ square integrable,

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

given above is not very well founded. Nevertheless the assumption of a square integrable $A(x, y)$ would seem reasonable in practice.

4. The Canonical Correlation of Stationary Gaussian Processes

We consider first the case of two stationary Gaussian processes of the form

$$\text{(8)}\qquad \left\{\begin{aligned} x_s &= \sum_1^\infty a_j(\xi_{1j}\cos s\lambda_j + \xi_{2j}\sin s\lambda_j) \\ y_t &= \sum_1^\infty b_j(\eta_{1j}\cos t\lambda_j + \eta_{2j}\sin t\lambda_j) \qquad 0 \leqq \lambda_j \leqq \pi \end{aligned}\right.$$

$$\sum a_j^2 < \infty, \quad \sum b_j^2 < \infty$$

where the ξ_j and η_j are Gaussian and $\mathscr{S}$ and $\mathscr{T}$ are the integers. We also take these variables to have zero mean and unit variance and assume all of the covariances zero save for

$$[\xi_{ij}, \eta_{ij}] = \rho_{1j},$$
$$[\xi_{1j}, \eta_{2j}] = -[\xi_{2j}, \eta_{1j}] = \rho_{2j}.$$

These covariance assumptions are imposed so as to agree with those for a pair of stationary processes in discrete time (see formulae (11) below and the definitions following them).

If we form the random variables

$$\zeta_{1j} = (1/\sqrt{\rho_{1j}^2 + \rho_{2j}^2})(\rho_{1j}\eta_{1j} + \rho_{2j}\eta_{2j})$$
$$\zeta_{2j} = (1/\sqrt{\rho_{1j}^2 + \rho_{2j}^2})(\rho_{1j}\eta_{2j} - \rho_{2j}\eta_{1j})$$

then the only non zero correlations among the ξ_{ij} and ζ_{ij} are

$$[\xi_{ij}, \zeta_{ij}] = |\rho_{1j} + i\rho_{2j}| = \rho_j.$$

Now the functions of the form

$$\prod_t \xi_{i_t j_t},$$

where there are only a finite number of terms in the product, are dense in the space $\mathscr{M}$. This follows from the fact that the elements of $\mathscr{M}$ may be approximated (in the sense of mean square convergence, i.e. in the strong topology) arbitrarily closely by polynomials in the x_s and any x_s may be approximated strongly, uniformly in s, by a truncated sum of the form

$$x_{s,n} = \sum_1^n a_j(\xi_{1j}\cos t\lambda_j + \xi_{2j}\sin t\lambda_j).$$

Since $x_s - x_{s,n}$ is Gaussian the result follows.

To orthonormalise these products we replace the powers of ξ_{ij} by the standardised Chebyshev-Hermite polynomials, the p^{th} of which we indicate

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

by $H_p(\xi_{ij})$. We do the same with the ζ_{ij}. Then the products of the $H_p(\xi_{ij})$ form an orthonormal base for $\mathscr{M}$ and the $H_p(\zeta_{ij})$ perform the same function for $\mathscr{N}$. The space upon which $E(\rho)$, in (2), projects is thus spanned by the products

$$\prod_t H_{p_t}(\xi_{i_t j_t}) \tag{9}$$

for which

$$\prod_t \rho_{j_t}^{p_t} \leqq \rho. \tag{10}$$

This serves to describe the $E(\rho)$ for a general pair of Gaussian stationary processes. Taking $\mathscr{S}$ and $\mathscr{T}$ as the integers let x_s and y_t have the spectral representations ([6], p. 481)

$$\begin{cases} x_s = \int_0^\pi \cos s\lambda \, \mathrm{d}u_1(\lambda) + \int_0^\pi \sin s\lambda \, \mathrm{d}v_1(\lambda) \\ y_t = \int_0^\pi \cos t\lambda \, \mathrm{d}u_2(\lambda) + \int_0^\pi \sin t\lambda \, \mathrm{d}v_2(\lambda) \end{cases} \tag{11}$$

with

$$\begin{aligned}
&\mathscr{E}\{\mathrm{d}u_1(\lambda)^2\} = 2\mathrm{d}F_{11}(\lambda) = \mathscr{E}\{\mathrm{d}v_1(\lambda)^2\}\\
&\mathscr{E}\{\mathrm{d}u_2(\lambda)^2\} = 2\mathrm{d}F_{22}(\lambda) = \mathscr{E}\{\mathrm{d}v_2(\lambda)^2\}\\
&\mathscr{E}\{\mathrm{d}u_1(\lambda)\mathrm{d}u_2(\lambda)\} = \mathscr{E}\{\mathrm{d}v_1(\lambda)\mathrm{d}v_2(\lambda)\} = \mathscr{R}\{2\mathrm{d}F_{12}(\lambda)\}\\
&\mathscr{E}\{\mathrm{d}u_1(\lambda)\mathrm{d}v_2(\lambda)\} = -\mathscr{E}\{\mathrm{d}v_1(\lambda)\mathrm{d}u_2(\lambda)\} = \mathscr{I}\{2\mathrm{d}F_{12}(\lambda)\}
\end{aligned}$$

where $\mathscr{R}$ and $\mathscr{I}$ denote the real and imaginary parts and all other cross products have zero expectation. The $F_{ij}(\lambda)$ are the spectral distribution functions.

We may now approximate to the processes (11) by two sequences of processes $x_s^{(n)}$, $y_t^{(n)}$ of the form (8) with

$$\begin{aligned}
a_j^{(n)}\xi_{1j}^{(n)} &= u_1\left(\frac{j\pi}{n}\right) - u_1\left(\frac{(j-1)\pi}{n}\right)\\
a_j^{(n)}\xi_{2j}^{(n)} &= v_1\left(\frac{j\pi}{n}\right) - v_1\left(\frac{(j-1)\pi}{n}\right)\\
b_j^{(n)}\eta_{1j}^{(n)} &= u_2\left(\frac{j\pi}{n}\right) - u_2\left(\frac{(j-1)\pi}{n}\right)\\
b_j^{(n)}\eta_{2j}^{(n)} &= v_2\left(\frac{j\pi}{n}\right) - v_2\left(\frac{(j-1)\pi}{n}\right)
\end{aligned} \qquad j = 1, \cdots, n.$$

We call $\mathscr{M}_n$ the subspace of $\mathscr{M}$ spanned by the $\xi_{ij}^{(n)}$ and $\mathscr{N}_n$ the subspace of $\mathscr{N}$ spanned by the $\zeta_{ij}^{(n)}$. Correspondingly we use $E_n(\rho)$ for the spectral family of projections for the operator A_n, which is itself defined by (1) for the two sets of random variables $x_s^{(n)}$, $y_t^{(n)}$. If P_n projects onto $\mathscr{M}_n$ and Q_n onto $\mathscr{N}_n$ then

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

$$Q_n A P_n = A_n$$

on $\mathscr{M}_n$, and is the null operator on $\mathscr{M}_n^{\perp}$. Indeed if $f \in \mathscr{M}_n$ then

$$(Q_n A P_n f, u)_{\mathscr{N}} = (Af, Q_n u)_{\mathscr{N}} = [f, Q_n u] = (A_n f, Q_n u)_{\mathscr{N}} = (A_n f, u)_{\mathscr{N}}$$

since $A_n f \in \mathscr{N}_n$.

However, P_n and Q_n converge strongly to the identity operators on $\mathscr{M}$ and $\mathscr{N}$ (since $\{x_t^{(n)}\}^p$ converges strongly to x_t^p and similarly for$\{y_t^{(n)}\}^p$). Thus $Q_n A P_n$ converges strongly to A and it follows ([7], p. 369) that $E_n(\rho)$ converges strongly to $E(\rho)$ for each ρ not in the point spectrum of A.

If we put

$$\left[F_{ij}\left(\frac{k\pi}{n}\right) - F_{ij}\left(\frac{(k-1)\pi}{n}\right)\right] = \Delta_{(n)} F_{ij}(\lambda_k)$$

then the inequality (10) may be written in the form

$$\sum p_t \log \left\{ \frac{|\Delta_{(n)} F_{ij}(\lambda_{j_t})|}{\sqrt{\Delta_{(n)} F_{11}(\lambda_{j_t}) \Delta_{(n)} F_{22}(\lambda_{j_t})}} \right\} \leq \log \rho.$$

A characterization of $E(\rho)$ directly in terms of

$$\frac{|\mathrm{d}F_{12}(\lambda)|}{\sqrt{\mathrm{d}F_{11}(\lambda)\mathrm{d}F_{22}(\lambda)}}$$

would be preferable to the indirect one given above.

The situation for Gaussian stationary processes is simpler than that obtained in general, for the theory then extends (and only then) to any number of processes. Let $x_{i,t}$ be the i^{th} process. We use Ω, as before, for the space of all realizations of the vector process and $\mathscr{H}$ again for the Hilbert space of square integrable functions on Ω, but $\mathscr{M}_i$ for the subspace of functions of the realizations of the i^{th} process only.

Now

$$[f_i, f_j] = (A_{ij} f_i, f_j)_{\mathscr{M}_j}, \quad f_i \in \mathscr{M}_i,$$

where

$$A_{ij} = W_{ij} B_{ij}, \; B_{ij} = (A_{ij}^* A_{ij})^{\frac{1}{2}}.$$

It is not difficult to see that

$$W_{ji} = W_{ij}^*, \quad B_{ji} = W_{ij} B_{ij} W_{ij}^*.$$

If now, in addition, the $x_{i,t}$ are jointly Gaussian and stationary the B_{ij}, for fixed i and j varying, commute. This is easily seen to be so for the $B_{ij}^{(n)}$, corresponding to the $x_{i,t}^{(n)}$ as defined earlier in this section, and will be so also for their strong limits. Thus we may write ([7], p. 360)

$$B_{ij} = \int_{0-}^{1} \rho_{ij}(\lambda) \mathrm{d}E_i(\lambda)$$

where the $\rho_{ij}(\lambda)$ are defined, measurable and take values in [0, 1], almost everywhere with respect to all of the measures $(E_i(\lambda)f, f)_{\mathscr{M}_i}$, $f \in \mathscr{M}_i$.

Thus

$$[f_i, f_j] = \int_{0-}^{1} \rho_{ij}(\lambda) \mathrm{d}(E_i(\lambda) f_i, E_i(\lambda) W_{ji} f_j)_{\mathscr{M}_i}$$

and we have, speaking loosely, simultaneously diagonalised all of the A_{ij}.

For the present situation the information defined in one process about another can again be shown to be

$$\tfrac{1}{2} \log \operatorname{tr} (B^2).$$

This is infinite for a pair of processes with continuous spectra but for a process whose matrix of spectral distribution functions has only a denumerable sequence of jumps, at the points λ_j let us say, we have

$$\tfrac{1}{2} \log \operatorname{tr}(B^2) = \tfrac{1}{2} \log \sum \prod \rho_{j_t}^{2p_{j_t}},$$

where the summation is over all different products, each such product being repeated twice, however, since $H_p(\xi_{1j})$ and $H_p(\zeta_{1j})$ have the same correlation as $H_p(\xi_{2j})$ and $H_p(\zeta_{2j})$. Thus

$$\tfrac{1}{2} \log \operatorname{tr} (B^2) = \tfrac{1}{2} \log \textstyle\prod_j (1 - \rho_j^2)^{-2} = -\sum \log (1 - \rho_j^2)$$
$$= -\sum \log \left(1 - \frac{|\mathrm{d}F_{12}(\lambda_j)|^2}{\mathrm{d}F_{11}(\lambda_j) \mathrm{d}F_{22}(\lambda_j)}\right)$$

The methods used in this section extend to a number of other cases of which we mention two.

(a) We consider a vector valued process X_t having at most a denumerable number of components $x_{i,t}$ and covariance function

$$\gamma_{ij}(s, t) = \mathscr{E}\{x_{i,s} x_{j,t}\} \quad s, t \in \mathscr{T}.$$

Let there be given a priori a group of transformations, $\mathscr{G}$, of $\mathscr{T}$ such that

$$\gamma_{ij}(g \cdot s, g \cdot t) = \gamma_{ij}(s, t), \quad \text{all} \quad i, j, s, t \quad \text{and} \quad g \in \mathscr{G}.$$

We shall also assume that $\mathscr{G}$ acts transitively on $\mathscr{T}$. The situation where $\mathscr{G}$ does not act transitively but the set of transition equivalence classes of points of $\mathscr{T}$ (transition equivalence of two points being defined by the existence of an element of $\mathscr{G}$ carrying one into the other) is denumerable may be brought to the one we are considering by defining a new process obtained by considering vector processes obtained by taking one point from each equivalence class. Then the elements of $\mathscr{T}$ may be put into one to one correspondence with the cosets of $\mathscr{G}$ modulo the subgroup which leaves some fixed point, t_0, of $\mathscr{T}$ invariant. We indicate a representative element of the coset corresponding to t by g_t. We now topologise $\mathscr{G}$ by giving it the weakest topology in which all of the functions

$$f(g; i, j, s, t) = \gamma_{ij}(g \cdot s, t)$$

are continuous. Then $\mathscr{G}$ is a topological group in the usual sense. If $\mathscr{T}$ has a natural topology in which it is locally compact and in which the components

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

$x_{i,t}$ are mean square continuous while $\gamma_{ij}(s, t)$ tends to zero as s tends to infinity the topology of $\mathscr{G}$ will be the same as the initial topology (it being assumed that $x_{i,t}$ has non zero mean square for some i).

We form the Hilbert space $\mathscr{H}$ which is the linear closure (complex coefficients) of the $x_{i,t}$ with respect to the inner product defined by the covariance function, i.e.

$$(x_{j,s}, x_{k,t}) = \gamma_{jk}(s, t).$$

Then the group of transformations on $\mathscr{H}$ defined by

$$U_g x_{i,t} = x_{i,g^{-1}t}$$

constitutes a representation of $\mathscr{G}$ by means of unitary operators in $\mathscr{H}$.

The spectral theory for the process X_t now proceeds from the spectral decomposition of this representation of $\mathscr{G}$ (see Naimark [3], p. 519). The two simplest cases are discussed shortly below.

(i) If $\mathscr{G}$ is locally compact and commutative then

$$U_g = \int_{\hat{\mathscr{G}}} \overline{(\alpha, g)} \mathrm{d}E(\alpha)$$

where $\hat{\mathscr{G}}$ is the character group of $\mathscr{G}$ and $E(\alpha)$ is a resolution of the identity relative to the Borel sets of $\hat{\mathscr{G}}$. Then

$$(12) \qquad \begin{aligned} x_{i,t} = x_{i,g_t \cdot t_0} &= \int_{\hat{\mathscr{G}}} (\alpha, g_t) \mathrm{d}E(\alpha) x_{x_t, t_0} \\ &= \int_{\hat{\mathscr{G}}} (\alpha, g_t) \mathrm{d}z_i(\alpha). \end{aligned}$$

This result is due to Kampé de Fériet [8].

(ii) If $\mathscr{G}$ is compact but not abelian we may put (Naimark [3], pp. 484—6)

$$U_g = \sum_\alpha U_g^{(\alpha)}$$

where the $U_g^{(\alpha)}$ are irreducible representations of $\mathscr{G}$ in mutually orthogonal finite dimensional subspaces $\mathscr{H}^{(\alpha)}$ of $\mathscr{H}$. If the matricial form of $\mathrm{U}_g^{(\alpha)}$ is given by

$$U_g^{(\alpha)} \xi_j^{(\alpha)} = \sum_k c_{kj}^{(\alpha)}(g) \xi_k^{(\alpha)},$$

the $\xi_k^{(\alpha)}$ being an orthonormal basis for $\mathscr{H}^{(\alpha)}$, then putting

$$x_{i,t_0} = \sum_\alpha \sum_j z_{i,j}^{(\alpha)} \xi_j^{(\alpha)}$$

we obtain

$$(13) \qquad x_{i,t} = x_{i,g_t \cdot t_0} = \sum_\alpha \sum_k \xi^{(\alpha)} \{\sum_j z_{i,j}^{(\alpha)} c_{k,j}^{(\alpha)}(g^{-1})\}.$$

In both of these cases the canonical correlation theory then proceeds from the spectral representation (i.e. (12) or (13)) basically in the same way as before, provided X_t is Gaussian. For example if $\mathscr{T}$ is the surface of a sphere in 3 dimensional space and $\mathscr{G}$ is the rotation group (13) becomes

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

$$x_{i,t} = x_i(\theta, \phi) = \sum_{\alpha=1}^{\infty} \sum_{k=-\alpha}^{\alpha} z_{i,\alpha} \mathscr{Y}_{\alpha,k}(\theta, \phi) \xi_k^{(\alpha)}$$

where (θ, ϕ) are the spherical coordinates of the point t on the surface of the sphere and $\mathscr{Y}_{\alpha,k}(\theta, \phi)$ is a surface spherical harmonic. The canonical variables are now the normalised Chebyshev-Hermite polynomials in the $\xi_k^{(\alpha)}$, and ρ (in (2)) takes in the values 0 or 1 only, so that A_{ij} is itself a projection (onto the space spanned by the Chebyshev-Hermite polynomials in the $\xi_k^{(\alpha)}$ which occur in *both* $x_i(\theta, \phi)$ and $x_j(\theta, \phi)$ with non zero coefficients).

b) If only two processes are involved there is a range of other relevant cases of which we shall mention only that of two mean square continuous (but not necessarily stationary), Gaussian processes on a finite interval on the real line. It is now not very difficult to show that we may put

$$x_s = \sum_1^{\infty} \alpha_k(s) w_k \quad y_t = \sum_1^{\infty} \beta_k(s) z_k$$

where w_k and z_k are random variables with unit mean square and $\alpha_k(s)$, $\beta_k(s)$ are constants. The cross products between the random variables vanish save for those between w_k and z_k for the same k.

In this case the information, $\frac{1}{2}$ log tr B^2, is

$$-\tfrac{1}{2} \sum \log\,(1-\rho_k^2)$$

where

$$r_{12}(s, t) = \sum \rho_k \phi_k(s) \psi_i(t)$$

is the expansion (essentially by Mercer's theorem) of the continuous function

$$r_{12}(s, t) = \mathscr{E}\{x_s y_t\}.$$

5. The Canonical Analysis of Sample Sequences

For completeness we shall here briefly indicate the analysis of a sample of n consecutive observations on a vector stationary, Gaussian, process x_t of p components, $\mathscr{T}$ being the integers, which corresponds to the theoretical discussion given above.

We form the transforms

$$J_n(\lambda, x) = \frac{1}{\sqrt{n}} \sum_1^n x_t e^{it\lambda}$$

whose real and imaginary parts correspond to the $u(\lambda)$ and $v(\lambda)$ previously discussed. It is easy to show that

$$(14) \qquad \mathscr{E}\{J_n(\lambda, x)\overline{J_n(\mu, x)'}\} \to \text{null matrix}, \ \lambda \neq \mu,$$

if λ and μ are not points of jump of $F(\lambda)$, the matrix of spectral distribution functions, and that

$$(15) \qquad \sigma\{J_n(\lambda, x)\overline{J_n(\lambda, x)'}\} \to F'(\lambda) \text{ a.e.}$$

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press

However, we shall want to form $J_n(\lambda, x)$ for the n equispaced values

$$\lambda_j = \frac{2\pi j}{n} \quad j = 0, \cdots, \left[\frac{n}{2}\right]$$

and will want

$$\mathscr{E}\{J_n(\lambda_j, x)\overline{J_n(\lambda_k, x)}'\} \to 0 \quad j \neq k$$

for fixed j, k, uniformly in j and k. However,

$$\mathscr{E}\{J_n(\lambda_j, x)\overline{J_n(\lambda_k, x)}'\} = \frac{1}{n} \sum_{t} \sum_{s}{}_{=1}^{n} e^{i(s\lambda_j - t\lambda_k)} \Gamma(s-t) \tag{16}$$

where $\Gamma(s-t)$ is a matrix having $\gamma_{p,q}(s-t)$ in the $(p, q)^{\text{th}}$ place. Thus (16) equals

$$\sum_{\tau=-n+1}^{n-1} \Gamma(\tau) e^{i\tau\lambda_j} \left\{\frac{1}{n} \sum{}' e^{it(\lambda_j - \lambda_k)}\right\}$$

where the summation in $\sum'$ runs from 1 to $n - \tau$ if $\tau \geqq 0$ and from $l + \tau$ to n if $\tau \leqq 0$. The element of (16) in the p^{th} place in the main diagonal is

$$\sum_{-n+1}^{n-1} \gamma_{p \cdot p}(\tau) e^{i\tau\lambda_j} \left\{\frac{1}{n} \sum{}' e^{it(\lambda_j - \lambda_k)}\right\}$$

which becomes, after some elementary rearrangements,

$$a_n \sum_{\tau} \gamma_{p,p}(\tau) \frac{1}{r} \left\{\cos \tau \frac{\lambda_j + \lambda_k}{2} \sin \tau \frac{\lambda_j - \lambda_k}{2}\right\}$$

where $|a_n| = \{n/2r| \sin \pi r/2|\}^{-1}$ and $r = |j - k|$. Since $|a_n| < 1$ we see that the diagonal elements of (16) converge to zero, uniformly, for $|j - k| > 1$, provided

$$\sum_{\tau} |\gamma_{p,p}(\tau)| < \infty \quad \text{for all } p.$$

Thus (16) itself then converges to the null matrix, uniformly in j and k. It is also evident that

$$\mathscr{E}\{J_n(\lambda_j, x)\overline{J_n(\lambda_j, x)}'\} - F'(\lambda_j)$$

tends to zero uniformly in j, under the same conditions.

Thus the quantities $J_n(\lambda_j, x)$, or at least their real and imaginary parts, may take the part played by $u(\lambda)$ and $v(\lambda)$ in the previous section, at least asymptotically. Of course, the ρ_{ij} are not known unless $F(\lambda)$ is prescribed a priori, which is not likely, so that the effects of estimation need to be considered. It must be mentioned that $J_n(\lambda, x)\overline{J_n(\lambda, x)'}$ does not itself provide a useful estimator of $F'(\lambda)$ (see for example [9], chapter III).

Of course a considerable improvement in the rate at which (16) converges to zero can be obtained under suitable stronger conditions. For example if

$$x_t = \sum_{-\infty}^{\infty} A_j \varepsilon_{t-j}$$

where the ε_t form a sequence of random vectors with

$$\mathscr{E}(\varepsilon_s \varepsilon_t') = \delta_{s,t} G$$

and

$$\sum_{-\infty}^{\infty} ||A_j||\ |j|^{\frac{1}{2}} < \infty,$$

where $||A_j||$ is the norm of the matrix A_j, then ([9] section III.1)

$$J_n(\lambda, x) = \{\sum A_j e^{ij\lambda}\} J_n(\lambda, \varepsilon) + R_n(\lambda)$$

where $J_n(\lambda, \varepsilon)$ is formed from the ε_t in the same way as $J_n(\lambda, x)$ is formed from the x_t. Here

$$\mathscr{E}\{||R_n(\lambda)||^2\} \leqq Kn^{-1}$$

while

$$\mathscr{E}\{J_n(\lambda, \varepsilon)\overline{J_n(\mu, \varepsilon)'}\} = \delta_{\lambda,\mu} G.$$

References

[1] Lancaster, H. O., 'The Structure of Bivariate Distributions', *Ann. Math. Statist.*, 29, (1958) 719—736.

[2] Gelfand, I. M. and Yaglom, A. M., 'Calculation of the Amount of Information about a Random Function Contained in Another Such Function', *American Mathematical Society Translations*, Vol. 12 (1959) 199—246.

[3] Naimark, M. A., *Normed Rings*, Noordhoff, (1959).

[4] Kolmogoroff, A. N., *Foundations of Probability Theory*, Chelsea (1956).

[5] Cramér, H., *Mathematical Methods of Statistics*, Princeton (1946).

[6] Doob, J. L., *Stochastic Processes*, Wiley (1953).

[7] Riesz, F. R. and Sz-Nagy, B., *Functional Analysis*, Blackie (1956).

[8] Kampé de Fériet, J., 'Analyse harmonique des fonctions aléatoires stationnaires d'ordre 2 définies sur un groupe abélien localement compact', *C. R. Acad. Sci. Paris*, 226 (1948) 868—870

[9] Hannan, E. J., *Time Series Analysis*, Methuens (1960).

[10] Halmos, P. R., *Measure Theory*, Van Nostrand (1950).

School of General Studies
Australian National University
Canberra.

https://doi.org/10.1017/S1446788700026707 Published online by Cambridge University Press