THE CANONICAL CORRELATION OF FUNCTIONS OF A RANDOM VECTOR

by

E. J. Hannan

University of North Carolina

and

Canberra University College

This research was supported by the Office of Naval Research under Contract No. Nonr 855 (09) for research in probability and statistics at Chapel Hill. Reproduction in whole or in part for any purpose of the United States Government is permitted

Institute of Statistics
Mimeograph Series No. 250
April 1960

THE CANONICAL CORRELATION OF FUNCTIONS OF A RANDOM VECTOR [1]

by

E. J. Hannan

The University of North Carolina

and The Canberra University College

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_t 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 functions of x and y (which have finite variances) 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 in another fashion which lends itself easily to generalizations to the case where p and q are not finite. In the case of Gaussian, stationary, processes this generalization is equivalent to the classical spectral theory and corresponds to a canonical reduction of a (finite) sample of data

---

[1] This research was supported by the Office of Naval Research under Contract No. Nonr-855 (09) for research in probability and statistics at Chapel Hill. Reproduction in whole or in part for any purpose of the United States Government is permitted.

which is basic. The theory also then extends to any number of processes. In the Gaussian case, also, the present discussion is intimately connected with the results of Gelfand and Yaglom [2] relating to the amount of information in one random process about another.

2. The Canonical Correlation of Stochastic Processes.

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

$$x_s \quad , s\varepsilon S \qquad ; \quad y_t \quad , t\varepsilon T$$

where S and T are two index sets. We do not require that the two sets be independent. We have mainly in mind the case where S and 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 S and T. It is a classic fact (Kolmogoroff [3]) that the joint distributions of finite sets of the x_s and y_t 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 $\mathcal{H}$ the Hilbert space of all square integrable complex valued functions of $\omega \varepsilon \Omega$ and indicate the inner product in $\mathcal{H}$ by the square bracket,

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

$$[a, a] = \| a \|^2.$$

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

$$[f, g] = (f, g)_M , \quad (f, f)_M = \|f\|^2_M$$
$$[u, v] = (u, v)_N , \quad (u, u)_N = \|u\|^2_N .$$

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

(1) $$[f, u] = (Af, u)_N ,$$

where A is a linear operator from M to N, and putting u = Af in (1) we derive, from Schwarz's inequality, that

$$\|Af\|_N \leq \|f\|_M ,$$

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

$$[f, u] = (f, A^*u)_M$$

where A* is the adjoint of A.

We may write (Riesz and Nagy [4] p. 286)

$$A = WB$$

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

(2) $$A = W \int_{0-}^{1} \rho \, dE(\rho)$$

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

Now putting

$$(3)\quad f = \int_{0-}^{1} dE(p) f\ ,\quad u = W \int_{0-}^{1} dE(p)\, W^{*} u + u_1,$$

where $u_1 \in (A\mathcal{M})^{\perp}$, we have

$$[f, u] = \int_{0-}^{1} \rho\, d\,(WE(\rho)f,\ WE(\rho)\, W^{*} u)_N$$

$$(4)\qquad = \int_{0-}^{1} \rho\, d\,(E(\rho)f, E(\rho)W^{*}u)_M$$

The relations (1), (2), (3), (4) are the generalizations of the classic relations of the theory of canonical correlation. There $\mathcal{H}$ is the space of all (real) linear combinations of the $x_1, \ldots x_p, y_1, \ldots y_q$ while $\mathcal{M}$ and $\mathcal{N}$ are similarly constructed from the x and y sets respectively. The inner product $[f, u]$ is usually written out as a p x q matrix and A corresponds to the matrix of regressions of the x set on the y set.

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 demensional 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 continiuty condition is satisfied, we shall put

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

for the Radon-Nikodym derivative of the H measure with respect to the product measure. Thus A in (1) is defined by

(5) $\quad Af = \int A(x,y) f(x) dM(x).$

The typical case where this is not so is that forp= 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).

$$-\frac{1}{2} \Sigma \log (1 - r_j^2)$$

where r_j is the j^{th} canonical correlation in the classical theory. This last expression is

(6) $\quad \frac{1}{2} \log \operatorname{tr} B^2$

where the trace is defined as

(7) $\quad \Sigma_k (B^2 \xi_k, \xi_k)_M$

and the ξ_k from any complete orthonormal sequence in the (separable) Helbert space $\mathcal{m}$. The expression (6) follows from the fact that the spectrum of B^2 is now discrete with jumps at all points

$$\rho = \prod_j r_j^{2p_j} \quad , p_j \text{ integral},$$

the typical jump corresponding to a characteristic function which is the product of the p_j -th standardized Cheybeyshev-Hermite polynomial (see for example [5] p. 133) in the j-th x variable canonical in the classical theory.

If (7) is required to be finite in any case then evidently B^2 must be a compact (completely continuous) operator with finite trace so that A (x,y) must be square integrable with respect to d M (x) dN(y). This is the case studied (under a slightly different guise)

by Lancaster [1]. Though (6) gives no longer a satisfactory measure of information when the distributions are not Gaussian (since it may be infinite when the maximum correlation between an f and a u is very small) the restriction to square integrable A (x,y) appears not unreasonable.

The analogy with the classical theory is now almost complete for B will be generated by the square intergrable kernel

$$B(x,z) = \Sigma_0^\infty \rho_j \phi_j(x)\, \phi_j(z), \quad \Sigma_0^\infty \rho_j^2 < \infty$$

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

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

with $f = \Sigma_0^\infty \alpha_j \phi_j(x) + f_1$, $u = \Sigma_0^\infty \beta_j \psi_j(y) + u_1$,

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

4. The Canonical Correlation of Stationary Gaussian Processes.

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

$$(8)\qquad \begin{aligned} x_s &= \Sigma_1^\infty a_j (\xi_{1j} \cos t\lambda_j + \xi_{2j} \sin t\lambda_j) \\ y_t &= \Sigma_1^\infty b_j (\eta_{1j} \cos t\lambda_j + \eta_{2j} \sin t\lambda_j) \quad 0 \leq \lambda_j \leq \pi \\ \Sigma a_j^2 &< \infty, \quad \Sigma b_j^2 < \infty \end{aligned}$$

where the ξ_j and η_j are Gaussian and S and T are the integers. We also take these variables to have zero mean and unit variance and assume all of the variances zero save for

$$[\xi_{ij}, \eta_{ij}] = \rho_{ij},$$

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

If we form the random variables

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

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

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, span the space $\mathcal{M}$. This follows from the fact that the elements of $\mathcal{M}$ may be approximated (in the strong topology) arbitrarily closely by polynomials in the $\mathbf{x}_s$ and any x_s may be approximated strongly, uniformly in s, by a truncated sum of the form

$$x_{s,n} = \Sigma_1^n a_j (\xi_{1j} \cos t\lambda_j + \xi_{2j} \sin t\lambda_j).$$

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

To orthonomalize these products we replace the powers of ξ_{1j} by the standardized Cheybeyshev-Hermite polynomials, the p^{th} of

which we indicate 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 $\mathcal{M}$ and the $H(\zeta_{ij})$ perform the same function for $\mathcal{N}$. 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

$$\prod_t \rho^{p_t}_{j_t} < \rho . \tag{10}$$

This serves to describe, in a somewhat unsatisfactory fashion, the $E(\rho)$ for a general pair of Gaussian stationary processes. Let x_t and y_t have the spectral representations ([6] p. 481)

$$x_t = \int_0^\pi \cos t\lambda \, du_1(\lambda) + \int_0^\pi \sin t\lambda \, dv_1(\lambda) \tag{11}$$

$$y_t = \int_0^\pi \cos t\lambda \, du_2(\lambda) + \int_0^\pi \sin t\lambda \, dv_2(\lambda)$$

with

$$\mathcal{E}\{du_1(\lambda)^2\} = 2dF_{11}(\lambda) = \mathcal{E}\{dv_1(\lambda)^2\}$$

$$\mathcal{E}\{du_2(\lambda)^2\} = 2dF_{22}(\lambda) = \mathcal{E}\{dv_2(\lambda)^2\}$$

$$\mathcal{E}\{du_1(\lambda)\, du_2(\lambda)\} = \mathcal{E}\{dv_1(\lambda)\, dv_2(\lambda)\} = \mathcal{R}\{2dF_{12}(\lambda)\}$$

$$\mathcal{E}\{du_1(\lambda)\, dv_2(\lambda)\} = -\mathcal{E}\{dv_1(\lambda)\, du_2(\lambda)\} = \mathcal{I}\{2dF_{12}(\lambda)\}$$

where $\mathcal{R}$ and $\mathcal{J}$ 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 $x_t^{(n)}$, $y_t^{(n)}$ of the form (8) with

$$a_j^{(n)} \xi_{1j}^{(n)} = u_1\left(\frac{j\pi}{n}\right) - u_1\left(\frac{(j-1)\pi}{n}\right)$$

$$a_j^{(n)} \xi_{2j}^{(n)} = v_1\left(\frac{j\pi}{n}\right) - v_1\left(\frac{(j-1)\pi}{n}\right)$$

$$j = 1, \ldots, n$$

$$b_j^{(n)} \eta_{1j}^{(n)} = u_2\left(\frac{j\pi}{n}\right) - u_2\left(\frac{(j-1)\pi}{n}\right)$$

$$b_j^{(n)} \eta_{2j}^{(n)} = v_2\left(\frac{j\pi}{n}\right) - v_2\left(\frac{(j-1)\pi}{n}\right)$$

If $E_n(\rho)$ projects upon the subspace $\mathcal{M}_n$ of $\mathcal{M}$ spanned by the elements defined by (9) and (10) for the $\xi_{ij}^{(n)}$ then $E_n(\rho)$ converges strongly to $E(\rho)$ for every ρ which does not belong to the point spectrum of A (in (2)). Indeed, if we put

$$A_n = Q_n A P_n ,$$

where Q_n projects onto the subspace $\mathcal{N}_n$ of $\mathcal{N}$ spanned by the $H_p(\zeta_{ij}^{(n)})$ and P_n projects onto $\mathcal{M}_n$ then A_n is the operator defined by (1) for the $x_s^{(n)}$ and $y_t^{(n)}$ processes, on the subspace $\mathcal{M}_n$, since if

$$f = f_1 + f_2,\ f_1 \in \mathcal{M}_n,\ f_2 \in \mathcal{M}_n^{\perp}$$

$$u = u_1 + u_2,\ u_1 \in \mathcal{N}_n,\ u_2 \in \mathcal{N}_n^{\perp}$$

then $$(Q_n A P_n f, u)_N = (A f_1, u_1)_N = [f_1, u_1].$$

However, P_n and Q_n converge strongly to the identity operators on $\mathcal{M}$ and $\mathcal{N}$ (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 ([4] p 369) that $E_n(\rho)$ converges strongly to $E(\rho)$ for each ρ not in the point spectrum of A.

If we put

$$[F_{ij}(\tfrac{k\pi}{n}) - F(\tfrac{(k-1)\pi}{n})] = \Delta_{(n)} F_{ij}(\lambda_k)$$

then the inequality (1o) may be written in the form

$$\Sigma p_t \log \frac{|\Delta_{(n)} F_{12}(\lambda_{j_t})|}{\Delta_{(n)} F_{11}(\lambda_{j_t}) \Delta_{(n)} F_{22}(\lambda_{j_t})} < \rho.$$

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

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

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 extends 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 $\mathcal{H}$ again for the Hilbert space of square integrable functions on Ω , but $\mathcal{M}_i$ for the subspace of functions of the realizations of the i^{th} process only.

Now

$$\left[f_i, f_j \right] = (A_{ij} f_i, f_j)_{M_j} \qquad f_\varepsilon \varepsilon \mathcal{M}_i$$

where

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

It is not difficult to show 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 ([4] p. 360)

$$B_{ij} = \int_{0-}^{1} \rho_{ij} (\lambda) \, dE_i (\lambda)$$

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

Thus

$$[f_i, f_j] = \int_{0-}^{\prime} \rho_{ij}(\lambda)\, d\,(E_i(\lambda) f_i, E_i(\lambda) W_{ji} f_j)_{M_i}$$

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

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

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

This is infinite in general for a pair of processes with continuous spectrum 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

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

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 $H_p(\zeta_{2j})$,

$$= \frac{1}{2} \log \prod_j (1 - \rho_j^2)^{-2} = -\Sigma \log (1-\rho_j^2)$$

$$= - \Sigma_j \log \left(1 - \frac{| dF_{12}(\lambda_j) |^2}{dF_{11}(\lambda_j)\, dF_{22}(\lambda_j)}\right)$$

agreeing with the formula given in [2] p. 226.

The methods used in this section extend to a number of other cases, in all of which the processes are Gaussian, of course.

(a) A mean-square continuous vector valued stationary process on the real line. Indeed the theory may be extended to any mean square continuous process (vector valued) defined on a separable, locally compact, Abelian group, $\mathcal{G}$, for which the covariance function

$$\mathcal{E}\{x_{is}\ x_{jt}\} \qquad , s, t \in \mathcal{G}$$

is a function only of the translation required to take s into t.

Indeed we may write the vector $\mathbf{x}_t$, having x_{it} in the i^{th} place, in the form

$$x_t = \int_{\hat{\mathcal{G}}} (\alpha, t)\, d\, z\, (\alpha)$$

where $\hat{\mathcal{G}}$ is the character group of $\mathcal{G}$, (α,t) is a character and $z\ (\alpha)$ is a vector valued function defined on $\hat{\mathcal{G}}$ for which

$$\mathcal{E}\left\{z(S_1)\ \ \overline{z(S_2)}\right\} = 0 \qquad S_1 \cap S_2 = \phi\ .$$

The development then follows the same line as that given above.

(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) processes on a finite interval on the real line. It is now not very difficult to show that we may put

$$x_s = \sum_1^\infty \alpha_k(s)\, w_k \qquad y_t = \sum_1^\infty \beta_k(s)\, z_k$$

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, 1/2 log tr B^2, is

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

where

$$r_{12}(s, t) = \Sigma\ \rho_k \phi_k(s)\ \psi_k(t)$$

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

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

5. The Canomical Analysis of Sample Sequences

For completeness we shall here briefly indicate the analysis of a sample of n consecutive observations on a vector stationary process x_t of p components.

We form the transforms

$$J_n(\lambda, x) = \frac{1}{\sqrt{n}} \Sigma_1^n\ x_t\, e^{it\lambda}$$

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

$$(12) \qquad \mathcal{E}\left\{J_n(\lambda, x)\ \overline{J_n(\mu,x)}'\right\} \to \text{null matrix}\ ,\ \lambda \neq \mu,$$

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

$$(13) \qquad \mathcal{E}\left\{J_n(\lambda, x)\ \overline{J_n(\lambda,x)}'\right\} \to F'(\lambda) \quad \text{a.e.}$$

However, we shall want to form $J_n(\lambda, x)$ for n equi-spaced values, say

$$\lambda_j = \frac{2\pi j}{n} \qquad j = o, \cdots, n-1$$

Now it is easy to show that

$$(14) \qquad \left\| \mathcal{E}\left\{J_n(\lambda_j,x)\ \overline{J_n(\lambda_k,x)}'\right\} \right\| \leq K \left|\sin \frac{\pi p}{n}\right|^{-1} \sum_{1}^{n-1} \|\Gamma(h)\| \left|\sin \frac{\pi p h}{n}\right|$$

where $p = |j - k|$

$$\leq K_1 \sum_{1}^{n-1} \|\Gamma(h)\| \frac{h}{n}\ ,\ p \geq 1,$$

which converges to zero as n increases if

$$\sum_{o}^{\infty} \|\Gamma(h)\| < \infty$$

which does not appear to be an over strong condition.

Thus the real and imaginary parts of $J_n(\lambda, x)$, for the case where x_t is Gaussian can take the part played by u (λ) and v (λ)

in the previous section, at least asymptotically. Of course, the ρ_{ij} are not known unless $F(\lambda)$ is prescribed $\underset{\wedge}{a}$ priori, which is not likely, so that the effects of estimation need to be considered.

Of course a considerable improvement in the rate at which (14) 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

$$\mathcal{E}(\varepsilon_s \varepsilon_t) = \delta_{s,t} G$$

and

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

then ([7] section III.1)

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

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

$$\mathcal{E}\left\{ \| R_n(\lambda) \|^2 \right\} \leq K n^{-1}$$

while

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

References

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

[2] Gelfan , 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] Kolmogoroff, A. N. Foundations of Probability Theory, Chelsea (1956).

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

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

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

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