MPS/CO Computer 70-6
4 juin 1970

ETUDE DES CORRELATIONS ENTRE LES PARAMETRES DU PS

Première Partie

G. Benincasa

Introduction

La recherche des correlations existantes entre certains paramètres du PS a pour but principal une meilleure connaissance des phénomènes qui peuvent entraîner de l'instabilité pendant la marche de l'accélérateur et le déroulement des opérations.

L'argument a déjà été traité par d'autres auteurs [1] qui ont employé l'ordinateur CDC 6600 pour analyser les données enregistrées relatives à certains paramètres d'injection.

Aujourd'hui l'ordinateur IBM 1800 on-line avec le PS nous permet l'extension de l'étude à un nombre de paramètres beaucoup plus élevé et en même temps d'obtenir les résultats dans un délai très court.

Les résultats que nous avons déjà obtenus et que nous espérons obtenir à la fin de cette étude feront l'objet d'une deuxième partie de ce rapport.

Dans cette première partie nous donnons seulement la description de la méthode employée et des programmes par l'ordinateur. Les raisons qui nous ont poussés à diviser ce travail en deux parties sont surtout deux : la première est que cette méthode, étant donné sa simplicité et sa généralité, peut être employée en d'autres domaines par d'autres personnes. La deuxième raison, plus égoiste, est que nous souhaitons recevoir beaucoup de suggestions pour la continuation de notre travail.

/je

<u>Le coefficient de correlation</u>

Avant de parler de ce coefficient, rappelons brièvement quelques paramètres statistiques qui nous seront utiles par la suite [2,3].

Supposons d'avoir une variable aléatoire X avec une densité de probabilité f(x); cela signifie que f(x) dx représente la probabilité que X soit comprise entre x et x+dx

$$f(x)\ dx = p_r(x < X \leq x+dx) \qquad (1)$$

avec

$$\int_{-\infty}^{+\infty} f(x)\ dx = 1 \qquad (1a)$$

On peut alors définir les paramètres :

$$\text{Valeur moyenne} = \overline{X} = \int_{-\infty}^{+\infty} x f(x)\ dx \qquad (2)$$

$$\text{Variance} = \int_{-\infty}^{+\infty} (x - \overline{X})^2\ f(x)\ dx \qquad (3)$$

La "standard deviation" de X est donnée par la racine carrée de la variance.

Si nous avons deux variables aléatoires X et Y, on peut définir une densité de probabilité composée f(x,y) telle que f(x,y) dx dy représente la probabilité que X soit comprise entre x et x + dx et que Y soit comprise entre y et y + dy :

$$f(x,y)dx\ dy = Pr(x < X \leq x+dx \ \underline{\text{et}}\ y < Y \leq y+dy) \qquad (4)$$

Le paramètre qui nous intéresse maintenant est la covariance de X et Y :

$$\text{Covariance} = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} (x - \overline{X})\ (y - \overline{Y})\ f(x,y)\ dy \qquad (5)$$

Le coefficient de correlation ρ est défini comme le rapport entre la covariance et la racine carrée du produit des variances des deux variables.

$$\rho = \frac{\int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} (x - \overline{X})\ (y - \overline{Y})\ f(x,y)\ dy}{\sqrt{\int_{-\infty}^{+\infty} (x - \overline{X})^2 f(x)\ dx \cdot \int_{-\infty}^{+\infty} (y - \overline{Y})^2 f(y)\ dy}} \qquad (6)$$

Ce coefficient peut varier entre -1 et +1. La valeur absolue de ce coefficient indique la force de la correlation.

Le signe + indique que les variations des deux variables sont dans le même sens, le signe - qu'elles sont dans sens opposés. Dans les Appendices nous discutons un peu plus en détail sur certaines caractéristiques de ce coefficient.

Le programme

Il est composé d'un programme principal, CORRL, et de deux subroutines, CORRE et DATA.

a) CORRL est un petit programme en language FORTRAN qui sert à définir les dimensions et le format dans lequel on veut les résultats. Il définit aussi le nombre N des observations et le nombre M des paramètres à observer. Les coefficients de correlation seront imprimés sous forme de matrice symétrique par rapport à la diagonale. Cette diagonale est composée de 1 qui représentent en pratique une autocorrelation.

   Le programme CORRL appelle la subroutine CORRE.

b) CORRE : cette subroutine, en language FORTRAN, a été fournie par l'IBM [4]. Elle calcule les valeurs moyennes, les "standard deviations" et les coefficients de correlation selon les formules suivantes :

   N = nombre d'observations

   M = nombre de paramètres

$$\text{Covariance}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i-A)\,(Y_i-B) - \frac{1}{N^2}\sum_{i=1}^{N}(X_i-A)\sum_{i=1}^{N}(Y_i-B) \quad (7)$$

$$\text{Variance}(X) = \text{covariance}(X,X)$$

où

$$A = \frac{1}{M}\sum_{i=1}^{M}X_i \qquad B = \frac{1}{M}\sum_{i=1}^{M}Y_i \qquad (8)$$

Ces formules sont un peu différentes de celles que nous avions données avant pour la présence de A et B qui ont été introduites pour une plus grande précision de calcul. La subroutine CORRE appelle, à son tour, la subroutine DATA.

c) DATA : il s'agit d'une subroutine en language ASSEMBLER qui, via le système d'acquisition STAR, permet de lire à chaque cycle du PS les valeurs des paramètres choisis.

En Fig.1 nous avons un exemple de sortie avec $M = 10$ et $N = 1000$.

Si l'on veut changer le nombre de lectures N il suffit de changer la carte correspondante dans le programme CORRL, avec la nouvelle valeur. La même chose est valable pour le nombre de paramètres à lire M; seulement si on veut lire plus que 10 paramètres, il faudra aussi changer les cartes de FORMAT, parce que dans les conditions actuelles 10 paramètres occupent toute la feuille du "line printer".

Les adresses STAR, sous forme hexo-décimale, se trouvent dans le programme DATA, une carte pour chaque adresse.

Le programme complet est actuellement contenu dans le "test core load" TCL 03.

H. van der Beken, J.H.B. Madsen et D.J. Warner nous ont beaucoup aidés avec suggestions et remarques. E. Ratcliff nous a aidés pour la partie programmes. Nous les remercions tous.

G. Benincasa

Distribution

MPS Operation
Computer Section
MST

E. Asséo
S. Battisti
P. Germain
L. Henny
C.D. Johnson
J.H.B. Madsen
C. Mantakas
C. Steinbach
U. Tallgren
C.S. Taylor
D.J. Warner

Références

1) D.J. Warner, Computer analysis of data obtained with the CPS Linac Data Logging System, MPS/Int.LIN 66-6

2) Laming and Battin, Random processes in automatic control, McGraw-Hill

3) D.J. Hudson, CERN 63-29

4) IBM - 1130 Scientific Subroutine Package

APPENDICE A

Valeur physique du coefficient de correlation

Si nous regardons le numérateur de la (6), c'est-à-dire la covariance entre X et Y, nous apercevons que la valeur que ce paramètre aura à la fin de nos mesures dépendra de deux effets différents :

a) le premier est l'amplitude des variations de X et Y par rapport aux moyennes $\bar{X}$ et $\bar{Y}$

b) le deuxième est le sens relatif de ces variations de X par rapport à Y.

La seule indication de la covariance donc, avec un nombre limité de mesures, ne suffit pas à déceler une interdépendance entre deux variables avec un degré suffisant de confiance.

La seule chose que nous pouvons affirmer avec certitude est que si les deux variables X et Y sont statistiquement indépendantes, la covariance doit être nulle.

On peut le montrer facilement : dans ce cas on aura

$$f(x,y) = f(x)\, g(y) \tag{9}$$

où $f(x)$ et $g(y)$ sont les densités de probabilité pour X et Y considérées séparément.

La covariance alors devient :

$$\mathrm{COV}(X,Y) = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} (xy - x\bar{Y} - \bar{X}y + \bar{X}\bar{Y})\, f(x)\, g(y)\, dy \tag{10}$$

Le premier de ces quatre intégrales donne :

$$\int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} xy\ f(x)\ g(y)\ dy \ = \int_{-\infty}^{+\infty} xf(x)\ dx \ . \int_{-\infty}^{+\infty} yg(y)\ dy = \bar{X}.\bar{Y} \qquad (11)$$

On fait la même procédure pour les autres intégrales et, à la fin on obtient :

$$COV(X,Y) \ = \ \bar{X}\bar{Y} - \bar{X}\bar{Y} - \bar{X}\bar{Y} + \bar{X}\bar{Y} \ = \ 0$$

Si maintenant on fait la normalisation de la covariance par rapport au produit des "standard deviations" nous avons un coefficient qui résulte indépendent aux variations de X et de Y par rapport aux valeurs moyennes.

Le coefficient de correlation est donc sensible seulement à l'interdépendance de X et de Y.

APPENDICE B

Limites de validité du coefficient de correlation

Supposons d'avoir deux variables aléatoires X et Y et de vouloir rechercher, s'il y est, une interdépendance entre les deux. Plus précisément nous essayons de déterminer les coefficients de l'expression $aX + b$ de façon que cette expression puisse représenter, avec la meilleure approximation, la variable Y.

Une des techniques la plus employée est celle des moindres carrées qui, dans notre cas, consiste à rendre minime l'expression

$$E\,[(Y - aX - b)^2]$$

où E[ ] représente la moyenne de la quantité entre parenthèses. Si, par simplicité, nous supposons $\overline{X} = \overline{Y} = 0$, on peut facilement montrer que

$$E\,[(Y - aX - b)^2] = \overline{Y^2}(1 - \rho^2) \qquad (12)$$

cela signifie que le coefficient de correlation ρ est lié à la possibilité que Y puisse être exprimé comme fonction linéaire de X.

Si $\rho = \pm 1$ l'erreur de l'approximation linéaire sera 0. Si par contre $\rho \neq \pm 1$ on fera toujours une erreur dans l'approximation linéaire et cela même si les deux variables sont absolument dépendentes. Par exemple, l'expression $Y = X^2$ nous dit que chaque valeur de Y est déterminée par X mais le coefficient de correlation ne sera jamais égal à 1 parce que une parabole ne peut être approximée par une ligne droite sans erreur.

Dans le cas du PS, la liaison entre deux paramètres P_1 et P_2 ne sera pas en général du type linéaire; en plus elle peut changer selon le réglage d'autres paramètres. Cela n'empêche pas que, sous des conditions bien précises, le coefficient de correlation puisse donner des indications très intéressantes dans l'étude de certains problèmes.

Par exemple, si on se limite à considérer des variations assez petites, l'approximation linéaire peut devenir possible.

MEANS

172.15505	1178.93725	27.42100	69.74801	55.72400
543.82418	114.15601	1398.41830	604.35211	497.92907

STANDARD DEVIATION

3.98097	38.49494	1.00931	1.00471	9.66135
16.04267	44.67595	0.54366	1.38492	0.81400

Figure 1

CORRELATION COEFFICIENTS

ROW 1	1.00000	0.57843	0.46902	-0.19293	0.14769	0.69857	0.10170	0.02414	0.00661	0.06084
ROW 2	0.57843	1.00000	0.81506	-0.02059	-0.25631	0.74198	0.11242	0.00638	0.04187	0.02350
ROW 3	0.46902	0.81506	1.00000	0.00403	-0.08373	0.59333	0.07555	-0.04009	0.06788	0.00230
ROW 4	-0.19293	-0.02059	0.00403	1.00000	-0.00222	0.00618	0.16339	0.06109	0.00482	-0.05127
ROW 5	0.14769	-0.25631	-0.08373	-0.00222	1.00000	-0.22008	-0.05151	-0.24557	0.03389	-0.07275
ROW 6	0.69857	0.74198	0.59333	0.00618	-0.22008	1.00000	0.18452	0.14752	0.05161	0.08764
ROW 7	0.10170	0.11242	0.07555	0.16339	-0.05151	0.18452	1.00000	0.01659	0.08529	0.01279
ROW 8	0.02414	0.00638	-0.04009	0.06109	-0.24557	0.14752	0.01659	1.00000	-0.03208	0.13498
ROW 9	0.00661	0.04187	0.06788	0.00482	0.03389	0.05161	0.08529	-0.03208	1.00000	0.00620
ROW 10	0.06084	0.02350	0.00230	-0.05127	-0.07275	0.08764	0.01279	0.13498	0.00620	1.00000