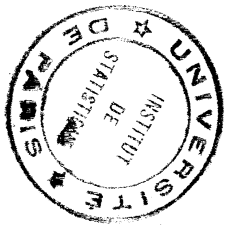
PUBLICATIONS DE L'INSTITUT DE STATISTIQUE DE L'UNIVERSITÉ DE PARIS

MÉMOIRES ET CONFÉRENCES SUR LE CALCUL DES PROBABILITÉS, LA STATISTIQUE THÉORIQUE ET APPLIQUÉE, L'ÉCONOMÉTRIE



VOL. II

PARIS
INSTITUT HENRI POINCARÉ
11, Rue Pierre Curie

REMARQUES
SUR LE
PROBLÈME DU CODAGE BINAIRE

par

M. P. SCHÜTZENBERGER

Hôpital Saint-Louis, Paris

INTRODUCTION

Soit le codage binaire optimum du point de vue de la fréquence totale des erreurs d'un ensemble de m messages de longueur l fixée à l'avance. Sa recherche pose un problème combinatoire pour lequel les méthodes de la théorie des communications sont d'assez peu d'efficacité. Le but de cette brève note est de montrer que ce problème se trouve déjà résolu en partie par les travaux des statisticiens qui ont été amenés à construire, pour un but tout différent d'ailleurs, des objets mathématiques, les "balanced incomplete block designs", dont il est possible de montrer qu'ils réalisent précisément ces codes optimum, tout au moins pour certaine valeur des paramètres.

DÉFINITIONS

Par définition, un code sera un ensemble de M messages $m_i = \{ X_i^j \}$ constitués chacun par une séquence de L symboles 0 ou 1. Il est naturel de supposer que des considérations de coût relatif imposent à priori une valeur déterminée au nombre N des symboles 1 dans le code et on posera :

$$N = \sum_i r_i = \sum_J k^J$$

où r_i et k^J désignent respectivement le nombre de symboles 1 dans le ième message et à la jème position sur l'ensemble de m messages.

Pour deux messages quelconques ou plus généralement pour deux séquences $\{y_a^J\}$ et $\{y_b^{*J}\}$ (on posera (cf. 6) :

$$2\,D_{ab} = \sum_J \left[y_a^J \ (1-y_b^{*J}) + (1-y_a^J) y_b^{*J} \right]$$

D sera une distance entre $\{y_a^J\}$ et $\{y_b^J\}$ puisque $D_{ab} = D_{ba}$ et que $D_{ab} = o$ entraine l'identité des deux séquences.

STRUCTURE DU BRUIT DE FOND ET DÉCODAGE

Nous supposerons que chaque symbole est transmis indépendamment et avec une probabilité constante p d'être reçu correctement.

Dans ces conditions il est évident que le vrai problème du décodage est exactement celui du choix entre plusieurs hypothèses tel qu'il est étudié en statistique mathématique.

Il est donc normal d'adopter une stratégie Mini Max consistant à interpréter la séquence reçue $y_a = \{y_a^J\}$ comme provenant de l'émission de celui des messages m_i tel que $Pr\,(y_a^J \mid m_i)$ soit maximum et en effectuant un tirage au sort avec des probabilités égales si plusieurs messages m_i se trouvaient vérifier cette condition.

Par conséquent, la discussion de l'optimalité d'un code devrait se faire sur la base d'une "information de Wald" c'est-à-dire en considérant des variables de la forme

$$Z_{ij} = \sum_a Pr\ (y_a)\ \ Log\ \frac{Pr_a\,(y_a \mid m_i)}{Pr\ (y_a \mid m_j)}$$

qui permettent cette discrimination entre les messages émis.

De fait, sous cette forme le problème semble inextricable et nous le remplacerons par le problème approché de trouver des codes tel que la valeur minimum de D_{ij} sur l'ensemble de toutes les paires de messages soit la plus grande possible. Nous appelerons pour abréger ces codes "codes optimaux".

LES CODES OPTIMAUX COMME « BALANCED BLOCK DESIGNS »

Théorème : Pour des valeurs données de M, L et N, il existe un code optimal si l'on peut construire une M x L matrice formée de 0 et de 1 telle que :

1°) Ses vecteurs lignes ont tous la même longueur,

2°) Ses vecteurs colonnes ont tous la même longueur,

3°) Le produit scalaire de deux secteurs lignes a une valeur constante.

Démonstration :

Calculons la variance de k^J sur l'ensemble des L positions

$$\text{var}\ (k^J) = \sum_J (k^J)^2 - \frac{1}{l} (\Sigma k^J)^2$$

$$= \sum_J (\sum_i x_i^J)^2 - l^{-1} N^2$$

par développement de $(\sum_i x_i^J)^2$ et permutations des deux sommations il vient :

$$\text{var}\ (k^J) = 2 L^{-1} (M^2 L^2 - N^2 - (ML - N)^2) - 2 \sum_{i,i'} D_{ii'}$$

comme var (k^J) est nécessairement non négatif, la valeur maxima de $\Sigma_{i,i'} D_{ii'}$ n'est atteinte que si tous les k^J sont égaux à une certaine valeur constante K.

Si cela est arithmétiquement possible, la valeur minimum de $D_{ii'}$ sur l'ensemble des couples de messages sera la plus grande quand tous les $D_{ii'}$ seront égaux. Mais pour un message m fixe, on a :

$$2 \sum_{ii'} D_{ii'} = \sum_{i \neq i'} \sum_J (x_i^J (1-x_{i'}^J) + (1-x_i^J)(x_{i'}^J)) = r_i (M-K) + K (L-r_i)$$

u qui implique cette fois-ci $K \neq M - K$:

$r_i = R$ pour tout i.

Dans le cas où l'on aurait K = M-K, le raisonnement ne s'appliquerait pas, mais la conclusion reste sensiblement la même :

Prenons un message quelconque m_o. En permutant les symboles 0,1 entre eux, dans tous les messages pour certaines positions, on peut faire en sorte que $\{X_o^J\}$ soit toujours 1 sans affecter les $D_{ii'}$, et le raisonnement subsiste pour l'ensemble de ces m - 1 messages.

Les conditions énoncées par le théorème sont précisément celles qui définissent les "balanced incomplete block designs", tels qu'ils ont été introduits par F. Y. Yates pour les besoins de l'expérimentation statistique.

Sans entrer dans l'historique de cette théorie, nous rappelerons que le problème général de leur construction n'est pas résolu quoique l'on connaisse à la fois les solutions pour les faibles valeurs de M et de L (tables dans : (3)) et diverses méthodes plus ou moins générales de construction(Gf.en particulier (1) (8) et 10).

On sait d'autre part que les 5 paramètres sont liés par les deux relations diophantiennes :

$$L K = M R$$

et $\lambda (m - 1) = L (K - 1)$

où $\lambda = R - D$

et que l'on a toujours (R.A. Fisher 2)

$L \geqslant M$

Que ces conditions ne sont pas suffisantes a été montré pour la première fois dans (9) (et indépendamment peu après dans (12) pour une infinité de valeur des paramètres (si M = L = un nombre paire et si D n'est pas un carré parfait). D'autres résultats plus profonds ont été publiés, récemment (f.bibliographie dans (4) et (7)).

RÉFÉRENCES

1 Bose R. C. (1939) Ann. of Eug. IX 353-399

2 Fisher R.A. (1940) Ann. of Eug. X 52-75

3 " and F. Yates. Statistical tables for biological etc.. 3e edition. London 1948

4 Hall M. and Ryser H. J. (1951) Can. J. Math. (4) 495-502

5 Hamming R. W. (1950) Bell Syst. Tech. J. (26) 147-161

6 Laemmel A. E. (1952) Symp. on Appl. of comm.Theor.London

7 Mann. R. M. (1952) Can. J. Math. (4) 222-226

8 Rao. C. R. (1946) Proc. Nat. Ins. Ind. 123-135.

9 Schutzenberger M. P. (1949) Ann. Of Eug. XIV 286-287

10 " (1951) J. Roy. Stat.Soc. (B)XIII.120-125

11 " (1953) - Thèse à paraître.

12 Shrikande S. S. (1950) Ann. Math. Stat. (2) 106-111.

13 Yates F. (1936) Ann of Eug. VII 121-140.