

# *Codages, recodages et régression monotone*



**Gilbert Saporta**

Chaire de Statistique Appliquée & CEDRIC

CNAM

292 rue Saint Martin, F-75003 Paris

[gilbert.saporta@cnam.fr](mailto:gilbert.saporta@cnam.fr)

<http://cedric.cnam.fr/~saporta>

*novembre 2011*

# 1. Codages et recodages d'une variable

- 1.1 Quantification d'une variable nominale à  $m$  catégories

$$x \in \{1, 2, 3 \dots, m\}$$

$$\tilde{x} \in \{a_1, a_2, a_3 \dots, a_m\}$$

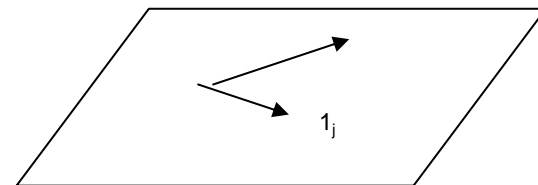
$$\tilde{x} = \sum_{j=1}^m a_j 1_j$$

Formulation matricielle

$$x = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ \cdot \\ m \end{pmatrix} \quad \tilde{x} = \begin{pmatrix} a_1 \\ a_2 \\ a_2 \\ a_1 \\ \cdot \\ a_m \end{pmatrix} = \begin{pmatrix} 1000..0 \\ 0100..0 \\ 0100..0 \\ 1000..0 \\ \cdot \\ 0000..1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ a_m \end{pmatrix}$$

Tableau disjonctif

$\tilde{x} \in W$  sous-espace à  $m$  dimensions



## ■ 1.2 Quantification d'une variable ordinale à m catégories

$$\tilde{x} = \sum_{j=1}^m a_j 1_j \text{ avec } a_m \geq a_{m-1} \geq \dots \geq a_1$$

Reparamétrisation:

$$a_1 = b_1$$

$$a_2 = b_1 + b_2$$

..

$$a_m = b_1 + b_2 + \dots + b_m$$

$$\begin{aligned} \tilde{x} &= \sum_{j=1}^m a_j 1_j = b_1 1_1 + (b_1 + b_2) 1_2 + \dots + (b_1 + b_2 + \dots + b_m) 1_m \\ &= b_1 (1_1 + 1_2 + \dots + 1_m) + b_2 (1_2 + \dots + 1_m) + \dots + b_m 1_m \\ &= b_1 + b_2 (1_2 + \dots + 1_m) + \dots + b_m 1_m \end{aligned}$$

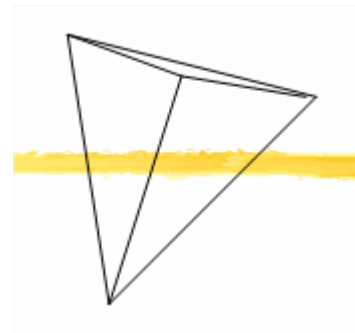
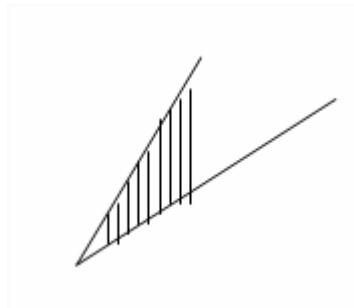
$$b_2, \dots, b_m \geq 0$$

$$b_1 \in \mathbb{R}$$

# combinaison linéaire à coefficients positifs ou nuls de $m-1$ variables + constante

$$\begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ a_m \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ b_m \end{pmatrix}$$

$\tilde{x} \in C$  cône polyédrique



- 1.3 Recodage d'une variable numérique

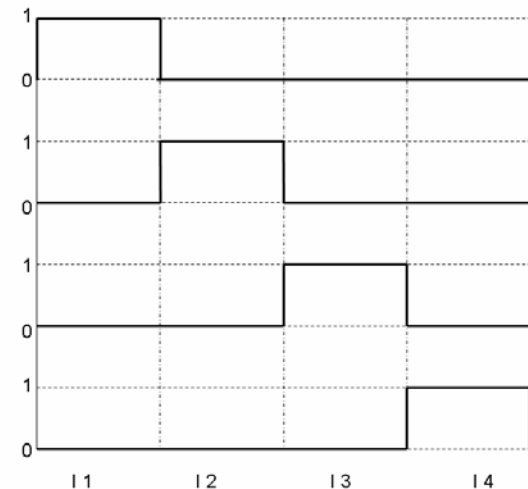
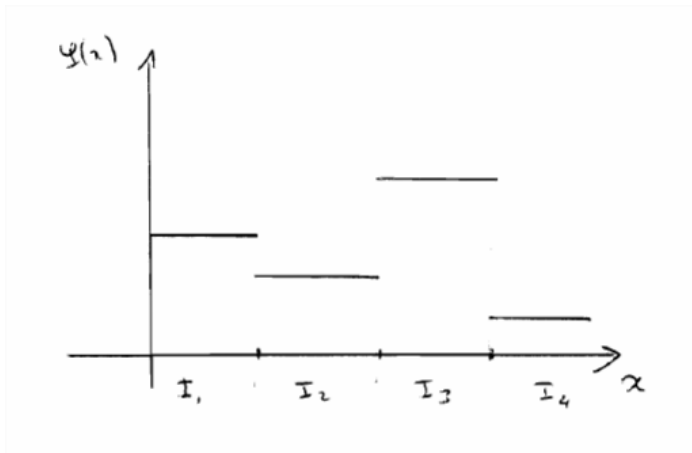
$$Y = \varphi(x)$$

- 1.3.1 recodage discret

découpage des valeurs de  $x$  en  $m$  intervalles

$\varphi$  constante par morceaux

$$\varphi(x) = \sum_{j=1}^m a_j 1(I_j)$$

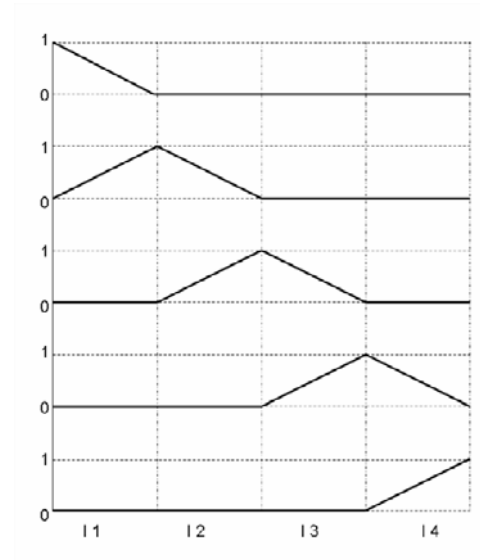
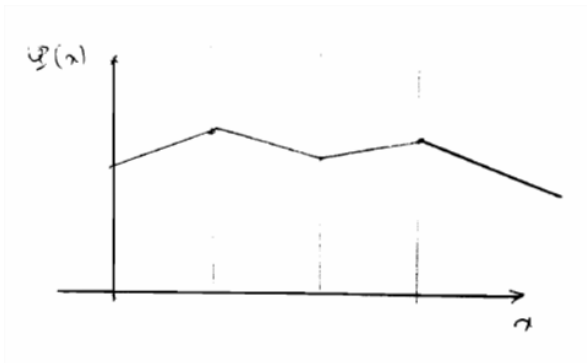


Splines de degré 0

- 1.3.2 recodage continu

$\varphi$  continue et linéaire par morceaux

splines de degré 1



$$\begin{cases} B_1(x) = 1 - (k+1)x & \text{si } x \in I_1 \\ B_1(x) = 0 & \text{sinon} \end{cases}$$

$$\begin{cases} B_2(x) = (k+1)x & \text{si } x \in I_1 \\ B_2(x) = 2 - (k+1)x & \text{si } x \in I_2 \\ B_2(x) = 0 & \text{sinon} \end{cases}$$

$$\begin{cases} B_{j+1}(x) = (k+1)x - (j-1) & \text{si } x \in I_j \\ B_{j+1}(x) = j+1 - (k+1)x & \text{si } x \in I_{j+1} \\ B_{j+1}(x) = 0 & \text{sinon} \end{cases}$$

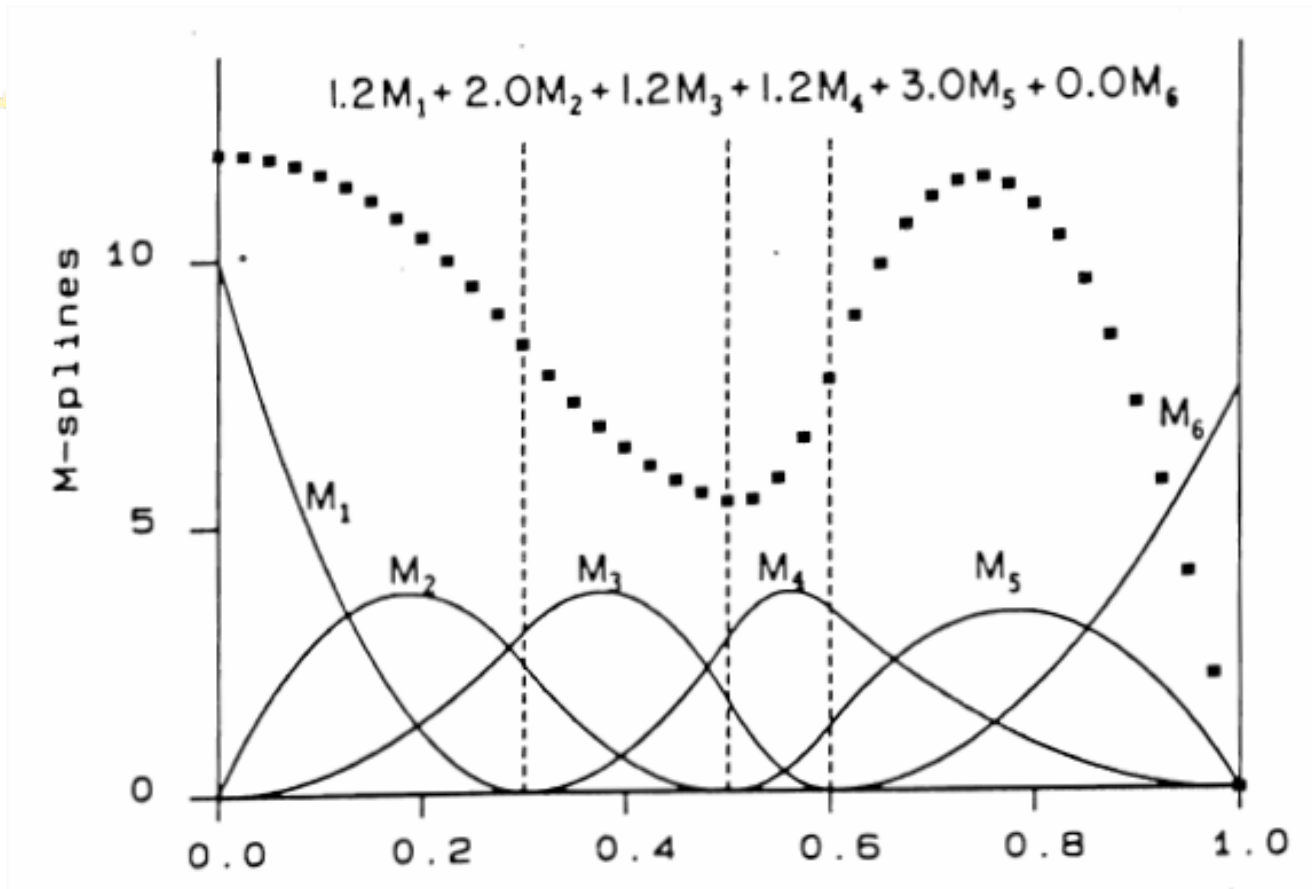
$$\begin{cases} B_{k+2}(x) = (k+1)x - k & \text{si } x \in I_{k+1} \\ B_{k+2}(x) = 0 & \text{sinon} \end{cases}$$

# Généralités sur les splines

- Soit  $x$  une variable définie sur  $[a,b]$  et  $k$  points intérieurs régulièrement espacés ou non, on appelle spline de degré  $d$  à  $k$  nœuds une fonction  $S(x)$  qui sur chacun des  $k+1$  intervalles est un polynôme de degré  $d$  et est  $d-1$  fois dérivable si  $d > 1$ , ou seulement continue si  $d=1$  (linéaire par morceau).
- Les splines de degré  $d$  à  $k$  nœuds forment un espace vectoriel de dimension  $d+k+1$ . Les combinaisons linéaires de splines de degré  $d$  à  $k$  nœuds sont encore des splines de degré  $d$  à  $k$  nœuds.

## Généralités sur les splines (2)

- Sur l'intervalle  $I_1$  le polynôme est libre et dépend de  $d+1$  paramètres, mais sur chacun des  $k$  intervalles suivants, les conditions de raccordement (continuité et dérivabilité  $d-1$  fois) ne laissent plus qu'un paramètre libre, d'où le résultat.
- Puisque l'ensemble des transformations spline est un espace vectoriel, on peut exprimer toute fonction  $S(x)$  comme une combinaison linéaire de  $d+k+1$  éléments d'une base de B-splines, ce qui revient dans un tableau de données  $X$  à remplacer chaque colonne-variable par  $d+k+1$  colonnes.
- Les B-splines de degré  $d$  sont non nulles sur  $d+1$  intervalles au plus

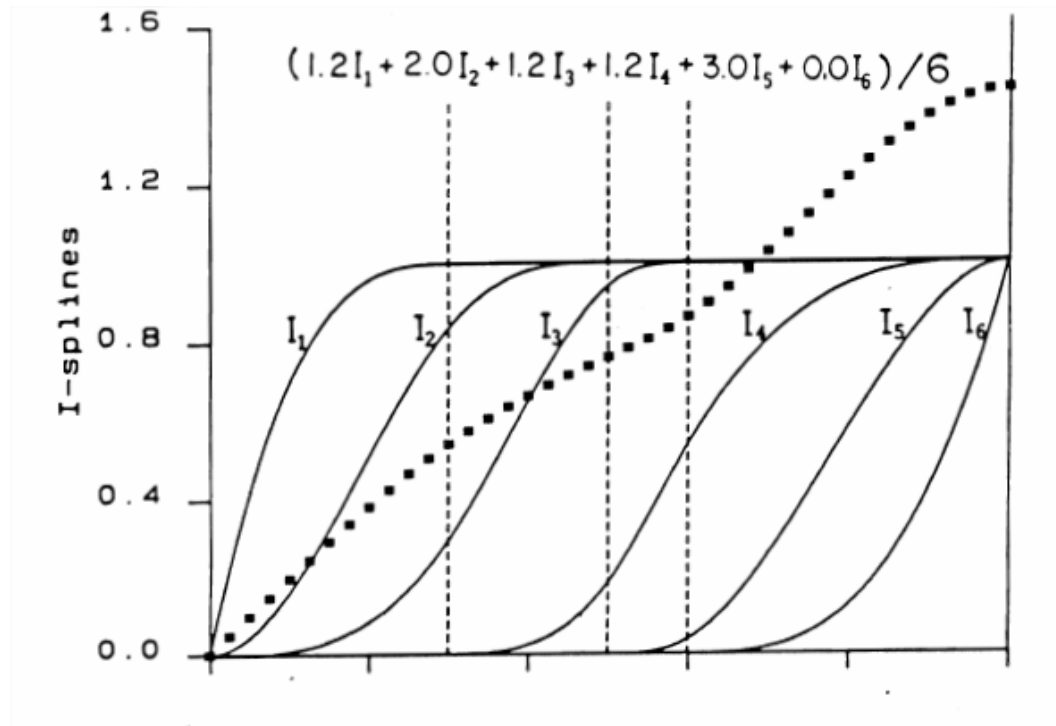


Ramsay, 1988

Splines de degré 2 sur 4 intervalles

### ■ 1.3.3 Transformations monotones polynomiales par morceaux

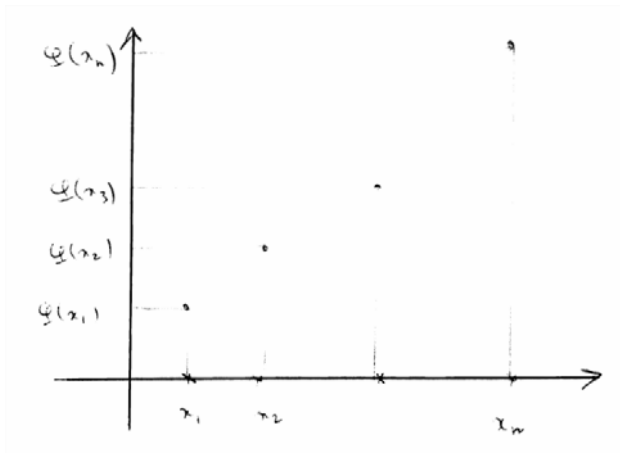
I-splines ou splines intégrés à partir de splines positifs, et combinaison linéaire à coefficients positifs



Ramsay, 1988

- 1.3.4 Transformations monotones (croissantes)  
point par point

Identique au cas de recodage d'une variable ordinale à n modalités



$$\left. \begin{array}{l} T(x_1) \\ T(x_2) \\ \cdot \\ T(x_n) \end{array} \right) = \left( \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array} \right) \left( \begin{array}{c} b_1 \\ b_2 \\ \cdot \\ b_n \end{array} \right) \right\} \geq 0$$

## 2. *ACP avec recodage*

- **2.1 Exemple:** enquête d'opinion BVA 1997 avec questions notées de 1 (très défavorable) à 5 (très favorable) :
  - Alco: les marques d'alcool et de tabac ne devraient pas avoir le droit de faire de la publicité
  - Porn: les magazines et films X offrent des distractions inoffensives à ceux et celles qui aiment cela
  - Femm: les femmes devraient revenir à leur rôle traditionnel dans la société
  - Homo: l'homosexualité est un mode de vie que la société devrait accepter
  - Mari: les mariages entre gens de races différentes sont à éviter
  - Eurv: les étrangers en provenance de l'UE et vivant en France depuis plus de 5 ans devraient avoir le droit de vote aux élections municipales
  - Immi: les immigrants renforcent notre pays grâce à leur travail

# ■ Proc Prinqual de SAS

```
PROC PRINQUAL < options > ;  
  TRANSFORM transform(variables < / t-options >)  
    < ... transform(variables < / t-options > ) > ;  
  BY variables ;  
  FREQ variable ;  
  ID variables ;  
  WEIGHT variable ;
```

```
proc prinqual;  
  transform spline(_all_ / degree=2);  
  transform opscore(N1-N10) monotone(M1-M10);  
run;
```

## MONOTONE MON

finds a monotonic transformation of each variable, with the restriction that ties are preserved. The Kruskal (1964) secondary least-squares monotonic transformation is used. This transformation weakly preserves order and category membership (ties). Variables following MONOTONE must be numeric, and they are typically discrete.

## MSPLINE MSP

finds a monotonically increasing B-spline transformation with monotonic coefficients (de Boor 1978; de Leeuw 1986) of each variable. You can specify the DEGREE=, KNOTS=, NKNOTS=, and EVENLY *t-options* with MSPLINE. By default, PROC PRINQUAL uses a quadratic spline. Variables following MSPLINE must be numeric, and they are typically continuous.

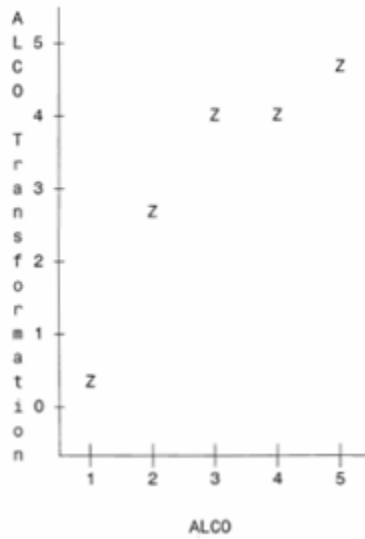
## OPSCORE OPS

finds an optimal scoring of each variable. The OPSCORE transformation assigns scores to each class (level) of the variable. Fisher's (1938) optimal scoring method is used. Variables following OPSCORE can be either character or numeric; numeric variables should be discrete.

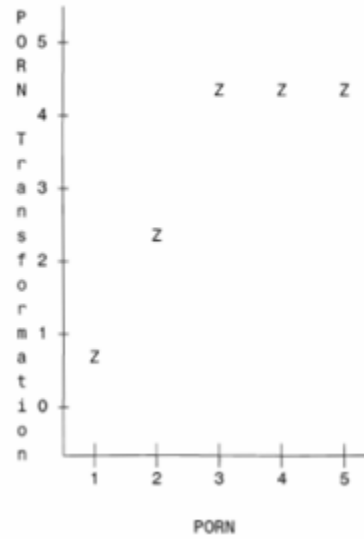
## SPLINE SPL

finds a B-spline transformation (de Boor 1978) of each variable. By default, PROC PRINQUAL uses a cubic polynomial transformation. You can specify the DEGREE=, KNOTS=, NKNOTS=, and EVENLY *t-options* with SPLINE. Variables following SPLINE must be numeric, and they are typically continuous.

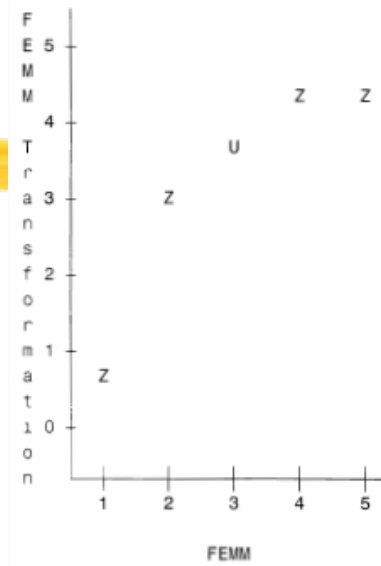
A=1, B=2, etc. Plot of TALCO\*ALCO.



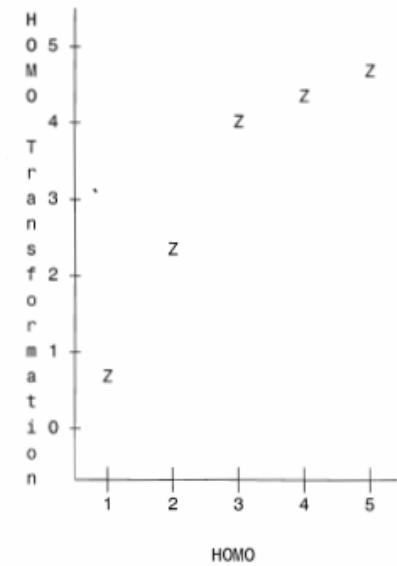
A=1, B=2, etc. Plot of TPORN\*PORN.



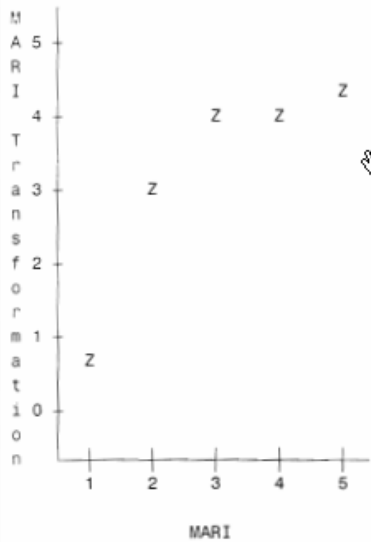
A=1, B=2, etc. Plot of TFEMM\*FEMM.



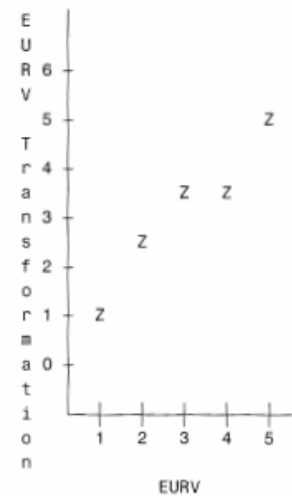
A=1, B=2, etc. Plot of THOMO\*HOMO.



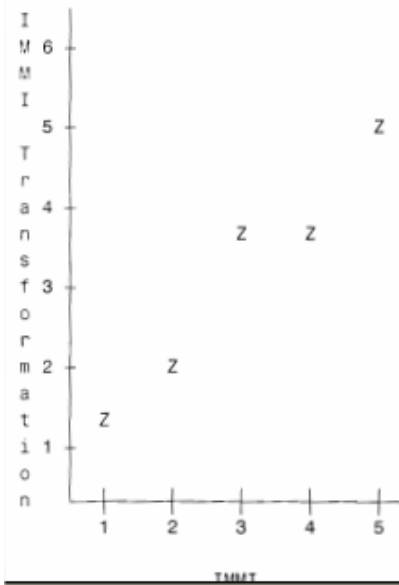
A=1, B=2, etc. Plot of TMARI\*MARI.

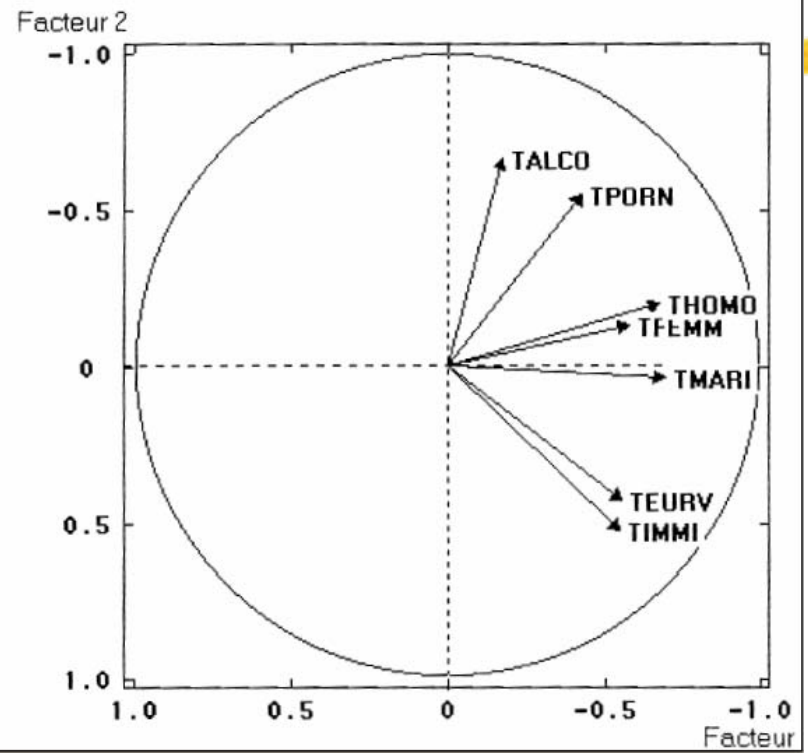
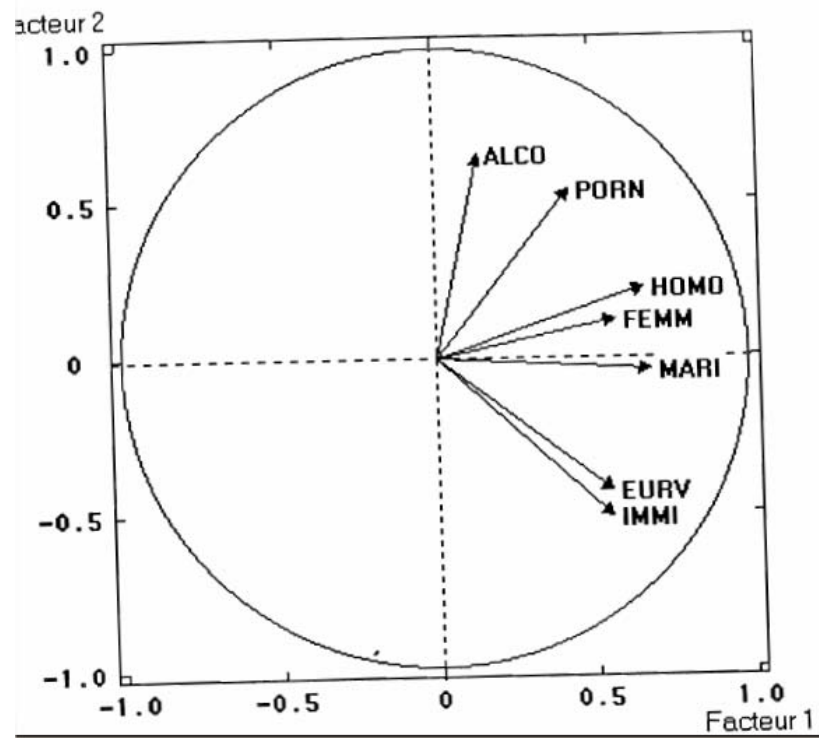



A=1, B=2, etc. Plot of TEURV\*EURV.



A=1, B=2, etc. Plot of TIMMI\*IMMI.





- 
- Principe: chercher les transformations telles que l'ACP sur variables transformées donne un % d'inertie expliqué maximal sur  $k$  axes ( $k=2$  par défaut).
  - Sur l'exemple : justification de l'ACP sur données 1-5

## 2.2 ACP avec splines

Ramsay, 1988

MONOTONE REGRESSION SPLINES IN ACTION

100

TABLE 1  
Automobile data from 1986 Consumer Reports

Automobile	Plot symbol in Fig. 9	Price	Displacement	City gas	Expwy gas	Weight
Chevrolet Chevette	A	56	1.6	13.3	6.8	10.0
Chevrolet Nova *	B	74	1.6	10.5	5.5	10.2
Chevrolet Spectrum	C	67	1.5	10.1	5.3	8.7
Dodge Colt	D	56	1.5	11.0	5.6	9.9
Dodge Omni	E	68	1.6	11.5	5.6	9.5
Ford Escort	F	61	1.9	12.0	6.2	10.9
Honda Civic	G	56	1.5	11.5	6.5	9.2
Mitsubishi Tredia	H	75	2.0	12.0	6.3	10.9
Nissan Sentra	I	56	1.6	10.5	5.6	9.5
Renault Alliance	J	60	1.4	12.0	5.6	9.1
Subaru	K	80	1.8	12.0	5.6	10.4
Toyota Corolla	L	73	1.6	11.0	5.3	10.3
Toyota Tercel	M	56	1.5	11.0	5.5	9.7
Volkswagen Golf	N	74	1.8	12.0	6.3	10.1
Volkswagen Jetta	O	83	1.8	12.6	6.3	10.5
Chrysler Laser	P	94	2.2	15.8	7.4	12.7
Honda Civic CRX	Q	73	1.5	9.4	5.6	9.0
Honda Prelude	R	110	1.8	10.5	6.5	10.6
Isuzu Impulse	S	109	1.9	14.9	6.8	12.4
Mitsubishi Cordia	T	89	2.0	12.0	6.3	11.2
Nissan 200SX	U	95	1.8	12.6	6.5	12.2
Nissan Pulsar NX	V	87	1.6	9.7	5.1	9.2
Pontiac Fiero	W	89	2.5	12.6	6.7	11.5
Audi 4000S	X	142	2.2	14.1	8.2	10.7
Chevrolet Cavalier	Y	67	2.0	15.8	7.2	11.6
Ford Tempo	Z	74	2.3	13.3	6.2	11.8
Honda Accord	1	88	2.0	13.3	6.3	11.8
Mazda 626	2	92	2.0	12.6	6.7	11.8
Mitsubishi Galant	3	132	2.4	14.9	6.7	12.9
Nissan Stanza	4	101	2.0	11.0	5.6	11.1
Oldsmobile Calais	5	93	2.5	14.1	6.7	12.0
Saab 900	6	126	2.0	14.1	7.9	12.9
Toyota Camry	7	97	2.0	11.0	5.5	12.2
Volvo	8	144	2.3	14.1	7.9	13.3
Buick Century	9	101	2.5	16.9	6.5	12.6
Chrysler Fifth Avenue	0	149	5.2	23.0	9.7	16.2
Chrysler Le Baron	@	100	2.2	13.3	7.9	11.7
Dodge Aries	#	72	2.2	13.3	7.9	11.5
Dodge Lancer	\$	94	2.2	15.8	8.7	12.7
Mercury Cougar	%	114	3.8	15.8	8.7	14.1
Oldsmobile Cutlass	/	107	3.8	19.5	9.0	15.2
Pontiac 6000	&	95	2.8	16.9	8.7	12.6
Buick Electra	<	154	3.8	18.1	7.7	15.0
Chevrolet Caprice	>	106	5.0	18.1	8.2	16.5

Notes: Price is measured in 100 U. S. dollars, displacement in liters, gasoline consumption in liters/100 km, and weight in 100 kg.

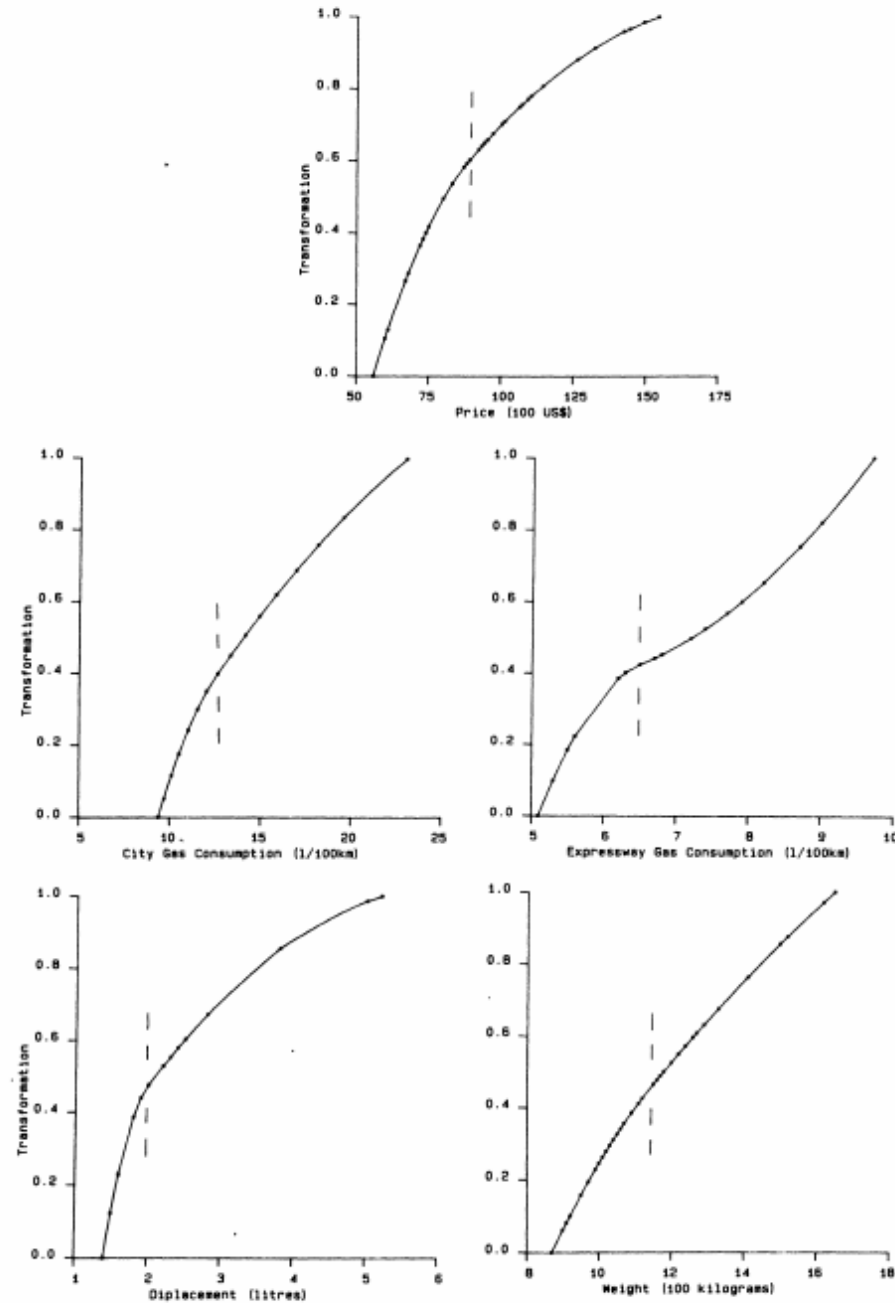


FIG. 8. Estimated monotone spline transformations resulting from an approximation of five transformed variables by the first two principal components.

### *3. L'AFC comme recherche de codages optimaux*



- 3.1 Le cas de deux variables qualitatives nominales

# *Un peu de (pré)histoire*

- Fisher (1940)
  - Equations de l'AFC
  - Introduction du vocable « Scores »

## THE PRECISION OF DISCRIMINANT FUNCTIONS \*

\* See Author's Note, Paper 155.

### I. INTRODUCTORY

IN a paper (1938*a*) on "The statistical utilization of multiple measurements" the author considered the general procedure of the establishment of discriminant functions, or sets of scores, based on an analysis of covariance, for a battery of different experimental determinations. In general, these functions are those giving stationary values to the ratio of

For example, in a contingency table individuals are cross classified in two categories, such as eye colour and hair colour, as in the following example (Tocher's data for Caithness compiled by K. Maung of the Galton Laboratory).

Eye colour	Hair colour					Total
	Fair	Red	Medium	Dark	Black	
Blue	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Variation among the four eye colours may be regarded as due to variations in three variates defined conveniently in some such way as the following:

Eye colour	$x_1$	$x_2$	$x_3$
Blue	0	0	0
Light	1	0	0
Medium	0	1	0
Dark	0	0	1

We may then ask for what eye colour scores, i.e. for what linear function of  $x_1, x_2, x_3$ , are the five hair colour classes most distinct. The answer may be found in a variety of ways. For example, by starting with arbitrarily chosen scores for eye colour, determining from these average scores for hair colour, and using these latter to find new scores for eye colour.

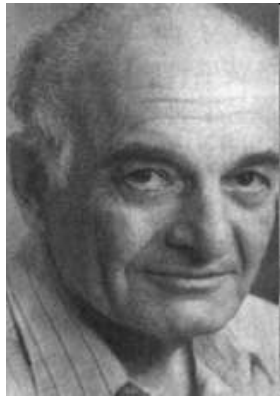
Apart from a contraction of scale by a factor  $R^2$  for each completed cycle, this form tends to a limit, and yields scores such as the following:

Eye colour	$x$	Hair colour	$y$
Light	-0.9873	Fair	-1.2187
Blue	-0.8968	Red	-0.5226
Medium	0.0753	Medium	-0.0941
Dark	1.5743	Dark	1.3189
		Black	2.4518

The particular values given above have been standardized so as to have mean values zero, and mean square deviations unity. In the sample from which they are derived each score has a linear regression on the other, the regression coefficient being 0.44627; this is, of course, equal to the correlation coefficient between the two scores regarded as variates. Hotelling has called pairs of functions of this kind canonical components. It may be noticed that no assumption is introduced as to the order of the classes of each category. In Tocher's schedule Light eyes come between Blue and Medium, but the discriminant function puts Blue between Medium and Light, though near the latter.

- L'algorithme des moyennes réciproques
- L'analyse canonique (canonical correlation analysis)
  - Proposée par H. Hotelling (1933)
  - Recherche de combinaisons linéaires de variables de  $X_1$  et  $X_2$  en corrélation maximale
  - AFC si  $X_1$  et  $X_2$  tableaux d'indicatrices de 2 variables qualitatives

- 3.2 L'ACM comme codage optimal simultané de p variables qualitatives
- Premiers travaux dans les années 1930-40
  - Louis Guttman (1916-1987)



Professeur de sociologie à Cornell University  
Fondateur de l'Israel Institute of Applied Social  
Research, devenu Guttman Center de l' Israel  
Democracy Institute

"The Quantification of a Class of Attributes: A  
Theory and Method of Scale Construction," in P.  
Horst et al., *The Prediction of Personal*  
*Adjustment*, SSRC, 1941, pp. 321-348

'Effet Guttman' ou 'horseshoe effect'

# The Quantification of a Class of Attributes: A Theory and Method of Scale Construction

## 1 THE PROBLEM

In the social sciences we are often confronted with a set of acts of a population of individuals that we would like to consider as a single class of behavior. Examples would be that totality of acts which is considered to constitute attitude toward war, or that totality of acts which is considered to constitute competence in a job, or that totality of acts which is considered to constitute marital adjustment.

We are interested in the case where the acts are attributes recorded in the form of items with mutually exclusive categories.

Thus, we are given the responses of a population of  $U$  individuals to a set of  $m$  items which have a common content that is desired to be thought as a single class of behavior. These responses can be represented by check marks as in the following table (with hypothetical entries):

SUBCATEGORY	INDIVIDUAL					
	1	2	3	4	...	$U$
$A_1$	✓		✓		...	✓
$A_2$		✓			...	
$A_3$				✓	...	
<hr style="border-top: 1px dashed black;"/>						
$B_1$			✓		...	✓
$B_2$	✓	✓		✓	...	
<hr style="border-top: 1px dashed black;"/>						
<hr style="border-top: 1px dashed black;"/>						
$Z_1$		✓			...	✓
$Z_2$	✓		✓		...	
$Z_3$				✓	...	
$Z_4$					...	

Principe de cohérence interne:  
 Attribuer à chaque modalité  
 une valeur numérique telle  
 que les variables  $\xi_i$  ainsi créées  
 soient aussi proches que possible  
 et la note moyenne la plus dispersée  
 Possible.  
 A l'origine des travaux hollandais.

Considérons le tableau  $n, p$  des variables  $\xi_j$  :

$$\begin{array}{c} \xi_1 \dots \dots \xi_p \\ 1 \left[ \begin{array}{cccc} \xi_{11} & \dots & \dots & \xi_{1p} \\ 2 & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ n & \cdot & \dots & \xi_{np} \end{array} \right. \end{array}$$

notons  $\bar{\xi}_1, \bar{\xi}_2, \dots, \bar{\xi}_p$  les moyennes des différentes lignes :

$$\bar{\xi}_i = \frac{1}{p} \sum_{j=1}^p \xi_{ij}$$

Supposons, ce qui ne nuit pas à la généralité que chaque  $\xi_j$  est une variable de moyenne nulle.

On cherche alors à avoir des mesures les plus homogènes possible en minimisant en moyenne la dispersion intra-individuelle.

Pour chaque observation celle-ci vaut  $\frac{1}{p} \sum_{j=1}^p (\xi_{ij} - \bar{\xi}_i)^2$  donc en moyenne elle vaut :

$$\frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (\xi_{ij} - \bar{\xi}_i)^2$$

La variance totale du tableau  $(\xi_{ij})$  étant égale à la moyenne des variances plus la variance des moyennes :

$$\frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (\xi_{ij})^2 = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (\xi_{ij} - \bar{\xi}_i)^2 + \frac{1}{n} \sum_{i=1}^n (\bar{\xi}_i)^2$$

Il revient au même de maximiser :

$$\frac{\frac{1}{n} \sum_{i=1}^n (\bar{\xi}_i)^2}{\frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (\xi_{ij})^2}$$

Or :

$$\xi_j = \mathbf{X}_j \mathbf{a}_j \quad \text{et} \quad \bar{\xi} = \frac{1}{p} \sum_{j=1}^p \mathbf{X}_j \mathbf{a}_j = \frac{1}{p} \mathbf{X} \mathbf{a}$$

donc :

$$\frac{1}{n} \sum_{i=1}^n (\bar{\xi}_i)^2 = \frac{1}{n} \left( \frac{1}{p} \mathbf{X} \mathbf{a} \right)' \left( \frac{1}{p} \mathbf{X} \mathbf{a} \right) = \frac{1}{np^2} \mathbf{a}' \mathbf{X}' \mathbf{X} \mathbf{a}$$

$$\sum_{i=1}^n (\xi_{ij})^2 = \xi_j' \xi_j = (\mathbf{X}_j \mathbf{a}_j)' (\mathbf{X}_j \mathbf{a}_j) = \mathbf{a}_j' \mathbf{D}_j \mathbf{a}_j$$

d'où :

$$\frac{1}{np} \sum_{j=1}^p \sum_{i=1}^n (\xi_{ij})^2 = \frac{1}{np} \sum_{j=1}^p \mathbf{a}_j' \mathbf{D}_j \mathbf{a}_j = \frac{1}{np} \mathbf{a}' \mathbf{D} \mathbf{a}$$

La quantité critère vaut donc :

$$\frac{\frac{1}{np^2} \mathbf{a}' \mathbf{X}' \mathbf{X} \mathbf{a}}{\frac{1}{np} \mathbf{a}' \mathbf{D} \mathbf{a}} = \frac{1}{p} \frac{\mathbf{a}' \mathbf{X}' \mathbf{X} \mathbf{a}}{\mathbf{a}' \mathbf{D} \mathbf{a}}$$

Son maximum est atteint pour  $\mathbf{a}$  vecteur propre associé à la plus grande valeur propre  $\lambda_1$  de  $\frac{1}{p} \mathbf{D}^{-1} \mathbf{X}' \mathbf{X}$ . On retrouve bien le premier facteur de l'ACM de  $\mathbf{X}$ .

## 5 THE "CHI-SQUARE" METRIC

The multiplicity of solutions in Section 4 shows that we are dealing with a problem akin to that of factor analysis in psychology. We can, as a matter of fact, throw our solution into the form of a principal axis solution, as we shall do in this section.

There is an essential difference, however, between the present problem of quantifying a class of attributes and the problem of "factoring" a set of quantitative variates. The principal axis solution for a set of quantitative variates depends on the preliminary units of measurement of those variates. In the present problem, the question of preliminary units does not arise since we limit ourselves to considering the presence or absence of behavior. But we shall now see that in a sense a metric has arisen out of our analysis, a metric that we shall call the "chi-square" metric.

*Expressing the form of a principal axis solution for*

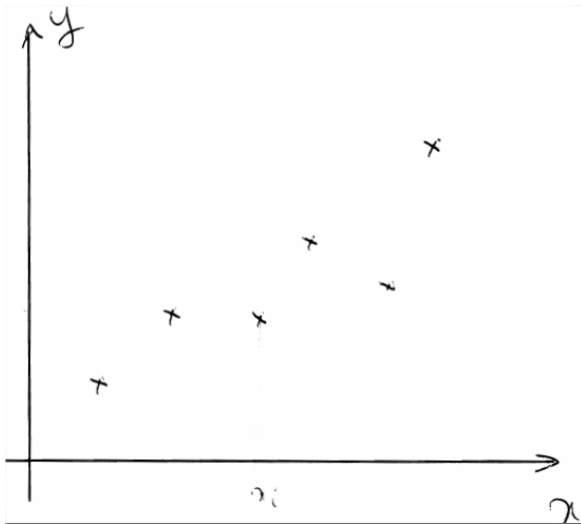
## ■ 3.3 Autres approches

- Chercher les codages qui maximisent la première valeur propre de l'ACP réduite des variables quantifiées: PRINQUAL à une dimension

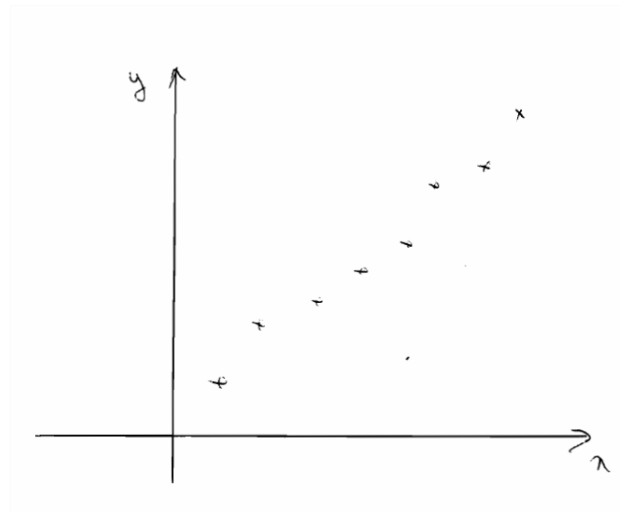
# 4. Quelques problèmes de régression monotone

- 4.1 Régression simple sur variable transformée

$$\min_T \sum_{i=1}^n (y_i - T(x_i))^2$$



Si relation monotone  $T(x)=y$



- Algorithme de Kruskal:

- Si  $y_2 < y_1$  on pose  $T(x_1) = T(x_2) = (y_1 + y_2)/2$  et on continue avec  $T(x_3) = y_3$  si  $y_3 > (y_1 + y_2)/2$
  - Sinon on pose  $T(x_1) = T(x_2) = T(x_3) = (y_1 + y_2 + y_3)/3$
  - Etc.

- Présentation matricielle:

$$\mathbf{y} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ b_n \end{pmatrix} + \mathbf{e} = \begin{pmatrix} T(x_1) \\ T(x_2) \\ \cdot \\ T(x_n) \end{pmatrix} + \mathbf{e}$$

Régression multiple à coefficients positifs  
sauf la constante

$$\hat{\mathbf{y}} = b_1 \mathbf{1} + \sum_{j=2} b_j \mathbf{z}_j$$

- Un algorithme général de régression sous contraintes de positivité:

Régression descendante avec élimination des variables à coefficients négatifs et itération.

## 4.2 Proc Transreg de SAS

```
PROC TRANSREG < DATA=SAS-data-set >  
    < OUTTEST=SAS-data-set > < a-options > < o-options > ;  
MODEL < transform(dependents < / t-options > )  
    < transform(dependents < / t-options > )...> = >  
    transform(independents < / t-options > )  
    < transform(independents < / t-options > )...> < / a-options > ;  
OUTPUT < OUT=SAS-data-set > < o-options > ;  
ID variables ;  
FREQ variable ;  
WEIGHT variable ;  
BY variables ;
```

**Table 90.2 Transformation Families**

<b>Transformation Description</b>	
<b>Variable Expansions</b>	
<a href="#">BSPLINE</a>	B-spline basis
<a href="#">CLASS</a>	set of coded variables
<a href="#">EPOINT</a>	elliptical response surface
<a href="#">POINT</a>	circular response surface & PREFMAP
<a href="#">PSPLINE</a>	piecewise polynomial basis
<a href="#">QPOINT</a>	quadratic response surface
<b>Nonoptimal Transformations</b>	
<a href="#">ARSIN</a>	inverse trigonometric sine
<a href="#">EXP</a>	exponential
<a href="#">LOG</a>	logarithm
<a href="#">LOGIT</a>	logit
<a href="#">POWER</a>	raises variables to specified power
<a href="#">RANK</a>	transforms to ranks
<b>Nonlinear Fit Transformations</b>	
<a href="#">BOXCOX</a>	Box-Cox
<a href="#">PBSPLINE</a>	penalized B-splines
<a href="#">SMOOTH</a>	noniterative smoothing spline
<b>Optimal Transformations</b>	
<a href="#">LINEAR</a>	linear
<a href="#">MONOTONE</a>	monotonic, ties preserved
<a href="#">MSPLINE</a>	monotonic B-spline
<a href="#">OPSCORE</a>	optimal scoring
<a href="#">SPLINE</a>	B-spline
<a href="#">UNTIE</a>	monotonic, ties not preserved
<b>Other Transformations</b>	
<a href="#">IDENTITY</a>	identity, no transformation
<a href="#">SSPLINE</a>	iterative smoothing spline

# Références



- Desbois, D. (2008) L'analyse des correspondances multiples " à la hollandaise" : introduction à l'analyse d'homogénéité. *La revue MODULAD*, 38, 194-244
- Ramsay, J.O. (1988): Monotone regression splines in action, *Statistical Science*, Volume 3, Number 4, 425-441.
- Tenenhaus, M.: Canonical analysis of two convex polyhedral convex cones and applications, *Psychometrika*, 1988, vol 53, 503-524
- Young, F. W., de Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: alternating least squares method with optimal scaling features. *Psychometrika*, 41, 505-529.
- Young, F. W., Takane, Y. & de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, 279-281.