Bulletin de la Société belge de Géologie Bulletin van de Belgische Vereniging voor Geologie	T. 92 V. 92	fasc. 4 deel 4	pp. 357-359 blz. 357-359	Bruxelles 1983 Brussel 1983

SIMULTANE VOORSTELLING VAN OBJECTEN EN VARIABELEN

door F. SYMONS en J. RIJMENAMS (★)

Uiteenzetting gehouden tijdens de zitting van 26 april, gewijd aan de INFORMATICA IN DE GEOLOGIE.

Weze ${}_nY_p$ een genormaliseerde datamatrix. De rijen beantwoorden aan objecten, de kolommen aan variabelen. De correlatie-matrix R kan in dit geval geschreven worden als R=Y'Y.

1. EEN VERANDERING VAN BASIS. DE VOORSTELLINGS-RUIMTE.

1. 1. Nemen wij aan dat de kolomvectoren $\vec{f}_i$ van ${}_nF_m$ een basis vormen voor de deelruimte gegenereerd door de kolomvectoren $\vec{y}_i$ van Y. Men bekomt dan volgende gelijkheid :

$$Y = F\,(F'F)^{-1}F'Y = FA'$$

of

$$\vec{y}_i = F\vec{w}_i$$

Nemen wij bovendien aan dat de voorgestelde basis F orthonormaal is : $(F'F=I) \Rightarrow Y = F\,(F'Y) = FA'$.

1. 2. Beschouwen wij de kolomvectoren $\vec{w}_i$ van A'. Deze genereren een m-dimensionele deelruimte. Geven wij t.o.v. de gewone basis dezer deelruimte simultaan de kolomvectoren $\vec{w}_i$ van A', en de rij-punten van Y weer.

Men merkt :

a) de afstand tussen 2 vectoren $\vec{w}_i$ en $\vec{w}_j$ (i, j, = 1,...,p) is gelijk aan deze tussen de overeenkomstige vectoren $\vec{y}_i$ en $\vec{y}_j$.
Inderdaad :

$$(\vec{y}_i-\vec{y}_j)'(\vec{y}_i-\vec{y}_j)=(\vec{w}_i-\vec{w}_j)'F'F(\vec{w}_i-\vec{w}_j)$$
$$=(\vec{w}_i-\vec{w}_j)'(\vec{w}_i-\vec{w}_j)$$

b) De scores van de oorspronkelijke variabelen y_i kunnen teruggevonden worden door de rij-punten van F loodrecht te projecteren op de assen gegenereerd door de vectoren $\vec{w}_i$ $(\vec{w}_i.\vec{w}_j=1)$. Het zijn de coördinaten dezer projecties :

$$\vec{w}_i(\vec{w}_i'\vec{w})^{-1}\vec{w}_i'F'=\vec{w}_i(\vec{w}_i'F')=\vec{w}_i(\vec{y}_i')$$

Om beide hier vermelde eigenschappen noemen wij deze m-dimensionele deelruimte de voorstellings-ruimte.
Noteren wij hierbij dat relevante informatie uitgaande van de beschrijving gegeven door Y, zal kunnen gewonnen worden uit de studie van de puntenwolk gevormd door de rij-punten van F (bv. het voorkomen van discontinuïteiten). De vectoren $\vec{w}_i$ geven dan aan in hoever de oorspronkelijke variabelen bij het tot stand komen dezer informatie betrokken zijn.

2. HET FACTOR-ANALYTISCH MODEL.

Gegeven dat de kolomvectoren $\vec{f}_i$ van $F=F_1;F_2$ een orthonormale basis vormen kan men schrijven :

$$Y=FA'=F_1A_1'+F_2A'_2 \quad \text{of ook}$$
$$\vec{y}_i=F\vec{w}_i=F_1\vec{u}_i+F_2\vec{v}_i$$

2. 1. De vectoren $F_1\vec{u}_i$ worden de orthogonale projecties genoemd der vectoren $\vec{y}_i$ op de deelruimte gegenereerd door de kolomvectoren van F_1.
Op lengte 1 gebracht
$F_1\vec{u}_i \lambda F_1\vec{u}_i(\vec{u}_i'\vec{u}_i)^{1/2}$, worden deze projecties de reconstructies genoemd van y_i, gegeven F_1. Inderdaad : van alle vectoren met lengte 1, die men zich in de

(★) Laboratorium voor beschrijvende Plantkunde, Kardinaal Mercierlaan, 92 - B-3030 Heverlee (België)

deelruimte, gegenereerd door de kolomvectoren F_1, indenken kan, zijn het deze vectoren $F_1\vec{u}_i/(\vec{u}_i'\vec{u}_i)^{1/2}$ die een maximale correlatie vertonen met de vectoren $\vec{y}_i$. Het is duidelijk : hoe beter geslaagd een reconstructie is, hoe dichter $(\vec{u}_i'\vec{u}_i)^{1/2}$ 1benaderen zal.

2. 2. Naar analogie met de benadering sub 1 wordt gesteld dat de kolomvectoren $\vec{u}_i$ de "gereduceerde" voorstellingsruimte genereren. Een simultane voorstelling van de vectoren $\vec{u}_i$ en de rij-punten van F_1 resulteert volgens bovenstaand schema in de reconstructies $F_1\vec{u}_i/(\vec{u}_i'\vec{u}_i)^{1/2}$ der vectoren $\vec{y}_i$.
Inderdaad : de coördinaten der projecties

$$(\vec{u}/\sqrt{\vec{u}_i'\vec{u}_i})\ \ (\vec{u}_i'F'/\sqrt{\vec{u}_i'\vec{u}_i})$$

leveren de scores $F_1\vec{u}_i/\sqrt{\vec{u}_i'\vec{u}_i}$.
Wil men nu, gegeven een bepaalde dim. q ($q \leq m$), trachten F_1 zo te kiezen dat het geheel der kolomvectoren van Y zo goed mogelijk gereconstrueerd wordt dan zal men moeten stellen dat

$$\sum_{i=1}^{n} \vec{u}_i'\vec{u}_i \quad \text{zo dicht mogelijk}$$

$\sum\vec{y}_i'\vec{y}_i=1$ benaderen moet.
Dit kan men bereiken door te stellen dat

a) een oplossing voor ${}_nF_m$ dient gevonden zo dat A'A een diagonale matrix weze, met als elementen de m eigenwaarden van Y'Y, groter dan 0;

b deze vectoren $\vec{f}_i$ (i=1,...,q) te beschouwen die beantwoorden aan de grootste q eigenwaarden.

Benadrukken wij echter dat door het optimaliseren der reconstructies men in dit geval er niet in lukt ook optimaal de afstanden tussen de oorspronkelijke vectoren $\vec{y}_i$ weer te geven.

3. CANONISCHE CORRELATIE-ANALYSE.

Onderstel dat men naast de beschrijving gegeven door de matrix Y, ook nog over een beschrijving beschikt, gegeven door een matrix X (men onderstelt dat ook de kolomvectoren $\vec{x}_i$ van X gestandardiseerd zijn).

Men kan ook hier trachten tot een voorstelling te komen. Dit keer echter zouden in deze voorstelling de relaties tussen beide beschrijvingen optimaal tot hun recht moeten komen.

3. 1. Geheel analoog met het factoranalytische model vertrekken wij ook hier van een orthonormale basis voor de deelruimte die gegenereerd wordt door de kolomvectoren van Y. Beschouwen wij nu de projecties van de vectoren in ${}_nX_1$ op deze deelruimte : F(F'X).
Waarschijnlijk, zal nu de dimensionaliteit van de deelruimte, gegenereerd door de kolomvectoren van F(F'X), kleiner zijn dan deze gegenereerd door de kolomvectoren van Y. Laten nu de kolomvectoren $F\vec{c}_i$ van ${}_nF_mC_r$ (i=1,...,r) ($r \leq m$) een orthonormale basis vormen voor deze beperkte ruimte (C'C=I).
In dit geval geldt :

$$F(F'X) = FC(C'F'X).$$

- Schrijven wij gemakkelijkheidshalve :
  FC=G
  C'F'X=H $\qquad (\vec{h}_j=C'F'\vec{x}_j)$
- De op lengte 1 gebrachte projecties
  $G\vec{h}_j/\sqrt{\vec{h}_j'G'G\vec{h}_j} = G\vec{h}_j/\sqrt{\vec{h}_j'\vec{h}_j}$
  zullen wij de reconstructies noemen van $\vec{x}_j$, gegeven Y.
- Als voorstellingsruimte kiezen wij deze gegenereerd door de kolomvectoren $\vec{h}_j$.
  Geven wij in deze ruimte simultaan de kolomvectoren $\vec{h}_j$ (kenmerkend voor de tweede set X) en de rij-punten van G (kenmerkend voor de eerste set G).
  De scores van de reconstructies $G\vec{h}_j/\sqrt{\vec{h}_j'\vec{h}_j}$ kunnen nu ook weer teruggevonden worden door de rijpunten van G loodrecht te projecteren op de assen gegenereerd door de op lengte 1 gebrachte vectoren $\vec{h}_j/\sqrt{\vec{h}_j'\vec{h}_j}$.
  Wijzen wij er op dat heel wat relevante informatie, resulterend uit de eerste beschrijving, zal kunnen gewonnen worden uit de studie van de pecularitеiten der wolk gevormd door de rijpunten van G. De richting van de vectoren $\vec{h}_j$ wijst dan aan in welke richting, gegeven de tweede set, een "verklaring" voor deze pecularitеiten kan gezocht worden.

3. 2. Ook hier is het natuurlijk mogelijk tot gereduceerde voorstellingen te komen.
Opdat deze echter zouden toelaten optimaal de relaties te bestuderen tussen de informatie van de set X enerzijds en deze van de set Y anderzijds moeten wij vooreerst nog beslissen welke kriteriumfunctie dient geoptimaliseerd. Een voor de hand liggende keuze lijkt te zijn

$$\text{trace } F'X(X'X)^{-1}X'F$$
$$= \text{trace } L'X'Y(Y'Y)^{-1}Y'XL$$

(de kolomvectoren van XL vormen in dit geval een orthonormale basis voor de deelruimte gegenereerd door de kolomvectoren van X). Geheel conform aan de vroegere werkwijze dient in dit geval een matrix ${}_mC_m$ gevonden, zo dat $C'F'X(X'X)^{-1}X'FC$ een diagonale matrix (zeg R^2) weze.
De diagonaal-elementen dezer matrix worden de gekwadrateerde canonische correlatie-coëfficiënten genoemd, de vectoren $F\vec{c}_i$

worden de canonische variabelen van de eerste set genoemd.

Gegeven dat slechts rekening gehouden wordt met de canonische variabelen $\vec{FC}_i$ van een set FC_1 zal men simultaan de rij-punten van FC_1 en de kolomvectoren van $C_1'F'X$ dienen te plotten.

Bovendien kunnen ook nog de kolomvectoren van $C_1'F'Y$ geplot worden. Voor de betekenis hiervan refereren wij naar het factor-analysch model.