KM 16(1984)
pag 53 - 61

# HET VULLEN VAN GATEN IN EEN MATRIX

## (Een toepassing van statistiek in de chemie)

J.C. van Houwelingen

Samenvatting

Aan de hand van gegevens uit chemisch onderzoek wordt gedemonstreerd hoe met behulp van statistische technieken ontbrekende getallen kunnen worden gereconstrueerd. De gegevens in kwestie kunnen worden opgevat als elementen van een matrix. Om de lege plekken op te vullen moet een model opgesteld worden van de samenhang tussen de elementen van de matrix. Uitgaande van een model kan dan een voorspellingsinterval geconstrueerd worden voor elke lege plek. Bij toenemende complexiteit van het model neemt de nauwkeurigheid van de voorspellingen eerst toe en dan weer af. Empirisch wordt het beste model bepaald.
Tenslotte wordt besproken wat te doen als het niet om een matrix maar een driedimensionaal blok van gegevens gaat.
De extra-informatie aanwezig in het blok maakt het mogelijk om (nog) betere voorspellingen te maken. De techniek hiervoor is recentelijk op ons instituut ontwikkeld.

---

Voordracht gegeven op 29 februari 1984 in de serie Kaleidoscoop II van het Mathematisch Instituut te Utrecht.

Adres : Instituut voor Mathematische Statistiek
Budapestlaan 6
3508 TA UTRECHT
tel. 030-531459

Deze voordracht kwam voort uit gezamenlijk werk met Prof.dr. C.L. de Ligny, dr. M.C. Spanjer (Anal. Chem. Lab Utrecht) en drs. H.M. Weesie (destijds Inst. voor Math. Stat. Utrecht).

Onderstaande tabel toont een gedeelte van de gegevens van Spanjer (1984). Het betreft hier bepaalde chemische evenwichtsconstantes die gerubriceerd worden naar substituent en reactietype.

| | | reactietype | | | | | |
|---|---|---|---|---|---|---|---|
| | | 16 | 17 | 18 | 19 | 20 | 21 |
| substituent | F | 158 | 144 | 120 | 28 | 71 | -5 |
| | CL | 162 | 181 | 154 | 126 | 117 | 85 |
| | Br | 100 | 134 | 99 | 86 | 81 | 76 |
| | I | 34 | 148 | 73 | 80 | 51 | 60 |
| | $CH_3$ | -24 | -26 | -107 | -96 | -74 | -65 |
| | $NO_2$ | 229 | 295 | 363 | 314 | 317 | 176 |
| | CN | 258 | 223 | 368 | 314 | 351 | 274 |
| | $OCH_3$ | 150 | 118 | -35 | -158 | -45 | -125 |
| | $COCH_3$ | 104 | 151 | 210 | 146 | 161 | 105 |
| | $COOCH_3$ | - | - | 170 | 186 | - | 110 |

Deze gevens zijn door Spanjer verzameld uit de chemische literatuur. Daar, waar in de tabel streepjes (-) staan ontbreken de literatuurgegevens. Om deze gegevens experimenteel te bepalen is niet eenvoudig en de vraag rijst daarom of het mogelijk is om deze getallen te reconstrueren uitgaande van de wel aanwezige gegevens. M.a.w. de vraag is "Wat staat/stond er op de lege plekken?"
Om die vraag te beantwoorden maken wij een statistisch model.
Wij schrijven

$$X = M + E \text{ met } X,M,E \quad n \times m\text{-matrices},$$

X : gedeeltelijk waargenomen matrix van gegevens
M : matrix van onbekenden
E : matrix van toevalsfluctuaties, t.g.v. meetfout, modelfout etc.

Wij veronderstellen de $E_{ij}$'s onderling onafhankelijk $N(0,\sigma^2)$ verdeeld.
Wij maken nu een model $M(\theta)$ $(\theta \in \mathbb{R}^p)$ zodanig dat als M op de ontbrekende plekken na bekend is, de ontbrekende plekken via dit model te berekenen zijn.
Laat $\hat{\theta}$ een schatter van $\theta$ zijn, gebaseerd op de waarneembare $X_{ij}$'s en laat voor een "gat" (i,j) $\hat{M}_{ij} = M_{ij}(\hat{\theta})$ een schatter zijn van onbekende $X_{ij}$, dan geldt dat

$$E_\theta(\hat{M}_{ij} - X_{ij})^2 = \sigma^2 + E_\theta(\hat{M}_{ij} - M_{ij})^2 = \sigma^2_{pred,i,j}.$$

Hierbij is gebruik gemaakt van de onafhankelijkheid van $\hat{M}_{ij}$ en $X_{ij}$. Dit in tegenstelling tot de modellen voor het opvullen van migratietabellen in De Jong (1982), waar de ontbrekende gegevens afhankelijk zijn van de aanwezige gegevens.

De algemene gang van zaken bij een bepaald model is de volgende

i) Schat $\theta$ door $\sum_{i,j} [X_{ij}-M_{ij}(\theta)]^2$ te minimaliseren. Hierbij gaat de sommatie over de combinaties (i,j) waarvoor $X_{ij}$ is waargenomen.

ii) Schat $\sigma^2$ d.m.v. $\hat{\sigma}^2 = \sum_{i,j} [X_{ij}-M_{ij}(\hat{\theta})]^2/\nu$, waarbij $\nu$ = N-p, met N = # waarnemingen = nm - # "gaten", en p = # vrije parameters.

iii) Bepaal voor de ontbrekende waarnemingen de (benaderde) variantie $\hat{\sigma}^2(M_{ij}(\hat{\theta})) = \hat{\sigma}^2 C_{ij}(\hat{\theta})$, waarbij $C_{ij}$ verkregen wordt door linearisatie van het model.

iv) Bepaal voor de "gaten" het 95% predictie-interval $M_{ij}(\theta) \pm k_\nu \hat{\sigma}\sqrt{1+C_{ij}(\hat{\theta})}$ voor $X_{ij}$ met $P(|T|>k_\nu) = .05$. (T heeft de $t_{[\nu]}$-verdeling.)

In het geval het model lineair is in $\theta$ (model I,II), is bovenstaande procedure exact. In geval van andere modellen (III e.v.) is het de vraag of de stappen iii) en iv) inderdaad bij benadering een 95% predictie-interval opleveren en hoe je in de praktijk de coëfficiënten $C_{ij}(\hat{\theta})$ bepaalt. De lezer die niet vertrouwd is met linearisatie wordt verwezen naar de appendix voor een korte toelichting.

Bij toenemende complexiteit van het model (d.w.z. een grotere waarde van p, het aantal vrije parameters) zal $\sigma^2$ dalen, maar $C_{ij}$ toenemen omdat er meer parameters geschat moeten worden. Bovendien wordt $k_\nu$ groter omdat $\nu$ kleiner wordt. Dit alles heeft tot gevolg dat wij mogen verwachten dat de lengte van de predictie-intervallen bij toenemende p eerst zal dalen en daarna zal stijgen. De kunst is om een model te vinden met zo klein mogelijke predictie-intervallen.

Model I. $M_{ij} = \mu$.

Dit allersimpelste model levert

$$\hat{\mu} = \overline{X} = \sum_{i,j} X_{ij}/N, \quad \sigma^2(\hat{\mu}) = \sigma^2/N, \quad \sigma^2_{pred} = \sigma^2(1+N^{-1}).$$

Het blijkt dat voor onze gegevens

$$\hat{\mu} = 118, \ \hat{\sigma} = 125, \text{ predictie-interval voor alle 3 gaten: } 118 \pm 250.$$

Dit model is vermoedelijk te simpel. Een beter model is waarschijnlijk het bekende additieve

Model II. $M_{ij} = \mu + \alpha_i + \beta_j, \quad \sum_{i=1}^{n} \alpha_i = 0, \quad \sum_{j=1}^{m} \beta_j = 0.$

Omdat er gegevens ontbreken zijn er geen pasklare formules voor $\hat{\mu}$, $\hat{\alpha}_i$ en $\hat{\beta}_j$. M.b.v. regressie analyse vinden wij

$\hat{\mu}$ : 121
$\hat{\alpha}$ : -35, 16, -25, -47, -186, 161, 177, -137, 25, 51
$\hat{\beta}$ : 15, 37, 20, -19, -1, -52
$\hat{\sigma}$ : 54.

De constantes $C_{ij}$ in de $\sigma_{pred}$ kunnen m.b.v. regressie-analyse bepaald worden. Omdat in ons voorbeeld alle ontbrekende gegevens in dezelfde rij zitten kan het hier nog direct.
Na enig puzzelen blijkt dat

$$\hat{M}_{10,1} = \sum_{i=1}^{9} X_{i,1}/9 + \sum_{j=3,4,6} X_{10,j}/3 - \sum_{i=1}^{9} \sum_{j=3,4,6} X_{ij}/27.$$

Omdat de termen onafhankelijk zijn, is

$$\sigma^2(\hat{M}_{10,1}) = \sigma^2(1/9 + 1/3 + 1/27) = 13/27\ \sigma^2.$$

(Analoog voor de andere gaten). Wij krijgen de volgende predictie-intervallen

$X_{10,1}$ : 187 ± 132
$X_{10,2}$ : 209 ± 132
$X_{10,5}$ : 171 ± 132.

De mogelijkheden van een lineair model zijn hiermee uitgeput. De volgende stap zou zijn een model met interactie, maar daarin zijn de ontbrekende cellen niet meer voorspelbaar uit de waargenomen cellen. Wij stappen daarom over op een multiplicatief model. Wij zullen straks zien dat dat wel op een natuurlijke wijze verder uit te breiden is.

Model III. $M_{ij} = \alpha_i\beta_j$.

Hierin zijn de $\alpha_i$'s en $\beta_j$'s op een vermenigvuldigingsfactor na bepaald. Het aantal vrije parameters is dus $p = n + m - 1$. Wij normeren de vectors $\alpha$ en $\beta$ op lengte 1 en schrijven $M_{ij} = \lambda\alpha_i\beta_j$, $\sum_i \alpha_i^2 = 1$, $\sum_j \beta_j^2 = 1$.
Merk op dat dit model niet terug te brengen is tot een additief model door de logarithme te nemen, omdat sommige van de $X_{ij}$'s negatief zijn.
De parameters kunnen worden geschat m.b.v. herhaalde regressie door op te merken dat bij gegeven $\beta_j$'s $\sum_{i,j} (X_{ij}-\lambda\alpha_i\beta_j)^2$ wordt geminimaliseerd door

$$\lambda\alpha_i = \sum_j X_{ij}\beta_j / \Sigma_j \beta_j^2$$

waarbij in teller en noemer gesommeerd wordt over de j's waarvan $X_{ij}$ is waargenomen. Als de $\alpha_i$'s gegeven zijn wordt het minimum bereikt voor

$$\lambda\beta_j = \sum_i X_{ij}\alpha_i / \sum_i \alpha_i^2.$$

Uitgaande van startwaarden voor de $\beta_j$'s kan nu $\sum_{i,j} (X_{ij}-\lambda\alpha_i\beta_j)^2$ geminimaliseerd worden door eerst de $\alpha_i$'s m.b.v. bovenstaande formule te bepalen, vervolgens de $\beta_j$'s, daarna weer de $\alpha_i$'s enzovoorts totdat $\sum_{i,j} (X_{ij}-\lambda\alpha_i\beta_j)^2$ niet meer afneemt. Deze procedure levert een locaal minimum voor de som van de gekwadrateerde residuen. Door verschillende startwaarden te kiezen kan gecontroleerd worden dat inderdaad het absolute minimum is bereikt. In ons voorbeeld levert dat

$\hat{\lambda}$ : 1269
$\hat{\alpha}$ : 0.170, 0.265, 0.183, 0.143, -0.129, 0.554, 0.575, -0.038, 0.287, 0.331
$\hat{\beta}$ : 0.359, 0.409, 0.497, 0.424, 0.436, 0.296
$\hat{\sigma}$ : 56

De bijbehorende voorspellingen zijn $\hat{M}_{10,1} = 151$, $\hat{M}_{10,2} = 172$, $\hat{M}_{10,5} = 183$. Men kan zelf aan de hand van de residuen nagaan dat dit model ongeveer even goed is als het additieve model II. De vraag is hoe hier predictie-intervallen geconstrueerd kunnen worden. In Van Houwelingen (1983) is via de omweg van de asymptotiek aangetoond dat voor $n \to \infty$ en/of $m \to \infty$ de eerder geschetste aanpak (linearisatie + t-verdeling) geoorloofd is bij de constructie van predictie-intervallen voor de ontbrekende $X_{ij}$'s (niet voor de parameters zelf!). Deze methode leverde de volgende predictie-intervallen

$X_{10,1}$ : 151 ± 133
$X_{10,2}$ : 172 ± 136
$X_{10,5}$ : 183 ± 139

Merk op dat deze predictie-intervallen ongeveer dezelfde lengte hebben als bij model II. Dat demonstreert opnieuw dat model II en III ongeveer even goed zijn. (II is iets beter.) Beide modellen zijn echter nog niet ideaal omdat rij 8 ($OCH_3$) nogal grote residuen heeft. Het grote voordeel van het multiplicatieve model is dat het verder uit te breiden is.

Model IV. $M_{ij} = \lambda_1\alpha_{1i}\beta_{1j} + \lambda_2\alpha_{2i}\beta_{2j}$.

Met restricties $\sum_i \alpha_{1i}^2 = \sum_i \alpha_{2i}^2 = \sum_j \beta_{1j}^2 = \sum_j \beta_{2j}^2 = 1$, $\sum_i \alpha_{1i}\alpha_{2i} = \sum_j \beta_{1j}\beta_{2j} = 0$. Het aantal vrije parameters is $p = 2(m+n) + 2 - 6 = 2(m+n-2)$. Een andere manier om het model te beschrijven is te zeggen dat M een matrix van rang 2 is. (In model III is rang(M) = 1). De methoden geschetst onder model III zijn hier weer van toepassing en wij vinden:

$\hat{\lambda}_1$ : 1250

$\hat{\alpha}_1$ : 0.170, 0.269, 0.186, 0.145, -0.133, 0.565, 0.589, -0.038, 0.294, 0.267

$\hat{\beta}_1$ : 0.340, 0.390, 0.508, 0.432, 0.442, 0.303

$\hat{\lambda}_2$ : 340

$\hat{\alpha}_2$ : 0.415, 0.220, 0.128, 0.083, 0.168, 0.044, -0.152, 0.819, 0.000, -0.179

$\hat{\beta}_2$ : 0.506, 0.538, -0.043, -458, -0.138, 0.399

$\hat{\sigma}$ : 26

Aan de coëfficiënten van $\hat{\alpha}_2$ zien wij dat de tweede "factor" vooral veroorzaakt wordt door het afwijkend gedrag van de 8e rij ($OCH_3$).
De predictie-intervallen worden

$$X_{10,1} = 79 \pm 129$$
$$X_{10,2} = 97 \pm 131$$
$$X_{10,5} = 156 \pm 70.$$

Vooral het laatste interval is aanzienlijk kleiner dan bij model II of III.
Wij kunnen nog verder gaan.

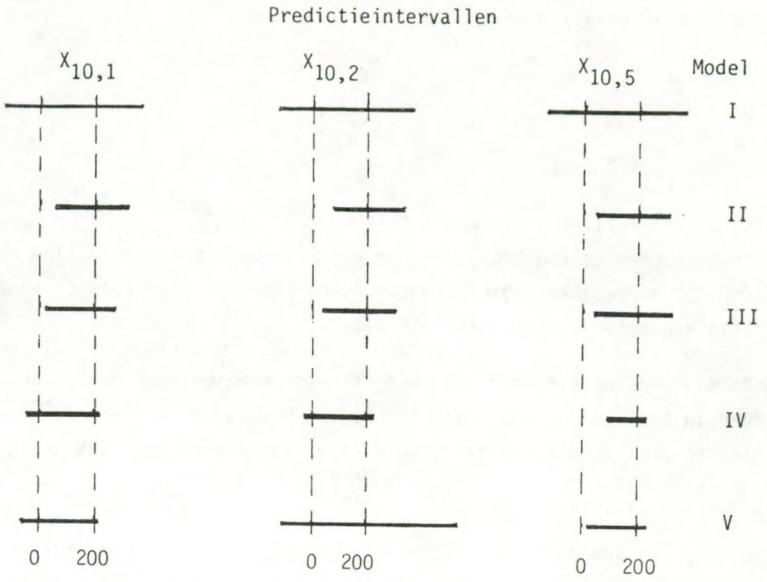
Model V. Het rang = 3 model.

$$M_{ij} = \sum_{f=1}^{3} \lambda_f \alpha_{fi} \beta_{fj}.$$

Het rekenwerk loopt volledig analoog. Wij vinden $\hat{\lambda}_1 = 1250$, $\hat{\lambda}_2 = 334$, $\hat{\lambda}_3 = 150$, $\hat{\sigma} = 20$. Dat lijkt goed, maar de predictie-intervallen worden

$$X_{10,1} : 69 \pm 137$$
$$X_{10,2} : 217 \pm 327$$
$$X_{10,5} : 127 \pm 104.$$

Wij zien dat deze weer groter worden. Klaarblijkelijk wordt de derde "factor" erg onnauwkeurig bepaald.
De verschillende predictie-intervallen zijn in nevenstaande figuur weergegeven. Wij zien hieraan nogmaals dat model IV, het multiplicatieve model met 2 "factoren" de smalste predictie-intervallen oplevert.
Opmerking.
Het gebruik van het woord "factor" suggereert een verwantschap met factor-analyse. Deze is ook aanwezig. Om precieser te zijn is bovenstaande nauw verwant met principale-componenten-analyse. In het geval van volledige gegevens worden de




α-vectoren verkregen als eigenvectoren van XX' en de β-vectoren als eigenvectoren van X'X. Rekentechnisch is er dus een grote verwantschap.
De vraagstelling is echter een geheel andere.

Hiermee zou het verhaal afgerond zijn, ware het niet dat ..... .
In werkelijkheid beschikt Spanjer over drie vergelijkbare matrices, dus over een systeem van gegevens $X_{ijk}(k = 1,2,3)$. Wat wij tot nu toe hebben laten zien was $X_{ij1}$. Deze extra gegevens kunnen gebruikt worden om een model te maken van het type

$$M_{ijk} = \sum_{pqr} L_{pqr} A_{pi} B_{qj} C_{rk}, \qquad p = 1,\ldots,P,\ q = 1,\ldots,Q,\ r = 1,\ldots,R.$$

Dit soort modellen is ook bestudeerd door Kroonenberg (1983), echter met een ander oogmerk dan het onze. Hij bestudeert samenhangen. Wij willen gaten vullen. Om dit probleem aan te kunnen is in Utrecht het programma GEPCAM ontwikkeld dat via geitereerde regressie de parameters schat en puntpredicties maakt. Zie Weesie en Van Houwelingen (1983). Later is ook een programma gemaakt dat m.b.v. de schattingen uit GEPCAM via linearisatie de coëfficiënten $C_{ijk}(\hat{\theta})$ berekent die nodig zijn om predictie-intervallen te maken. Het ligt in de bedoeling om beide programma's te integreren. Toepassing van deze programmatuur leverde de bes-

te predictie-intervallen bij P = Q = R = 3, en wel

$$X_{10,1,1} : 102 \pm 93$$

$$X_{10,2,1} : 133 \pm 97$$

$$X_{10,5,1} : 170 \pm 81.$$

Elk van de intervallen heeft een lengte $\leqslant 200$.
De betrokken chemici (De Ligny en Spanjer) zijn met deze intervallen redelijk tevreden. Op dergelijke wijze geconstrueerde predictie-intervallen nemen een belangrijke plaats in in Spanjer (1984).

Resumé: Voor het construeren van predictie-intervallen voor ontbrekende waarnemingen in toepassingen als deze is het van belang om modellen te construeren met niet te veel parameters die geldig zijn voor zoveel mogelijk waarnemingen.

Appendix Linearisatie

Om aan te geven hoe de linearisatie in zijn werk gaat, veronderstellen wij eerst dat $\theta \in \mathbf{R}^p$ de matrix M éénduidig bepaalt. Laat $d^{ij}(\theta)$ de vector zijn met als coördinaten

$$d_u^{ij}(\theta) = \frac{\partial}{\partial \theta_u} M_{ij}(\theta).$$

De Fisher-informatiematrix $I(\theta)$ wordt nu gegeven door $I_{u,v}(\theta) = \sum_{i,j} d_u^{ij}(\theta) d_v^{ij}(\theta)$, waarbij gesommeerd wordt over de paren (i,j) waarvoor $X_{ij}$ is waargenomen.
De (geschatte)variantie van $\hat{M}_{ij} = M_{ij}(\hat{\theta})$ voor een gat (i,j) wordt gegeven door

$$\hat{\sigma}^2 C_{ij}(\hat{\theta}) \text{ met } C_{ij}(\hat{\theta}) = \sum_{u=1}^{p} \sum_{v=1}^{p} d_u^{ij}(\hat{\theta}) d_v^{ij}(\hat{\theta}) (I(\hat{\theta})^{-1})_{u,v}.$$

In Van Houwelingen (1983) is aangetoond dat voor model III e.v. dit inderdaad de (benaderde) variantie van $\hat{M}_{ij}$ geeft, ondanks het feit dat $I(\hat{\theta})^{-1}$ niet de goede benadering geeft voor de covariantiematrix van $\hat{\theta}$.

De voorwaarde dat $\theta$ de matrix M éénduidig bepaalt, is in de praktijk niet handig om mee te werken. Het opstellen van bijvoorwaarden om de eenduidigheid af te dwingen maakt de zaak soms nodeloos gecompliceerd. In navolging van Rao (1973) kunnen wij echter de bijvoorwaarden laten vallen, als wij in de bepaling van $C_{ij}$, $I^{-1}$ vervangen door een gegeneraliseerde inverse $I^-$. Bijv. in model III is het het eenvoudigst om $\theta = (\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_m)$ te nemen en $M_{ij} = \alpha_i\beta_j$ te schrijven. Dit geeft dan $d^{11}(\theta) = (\beta_1,0,\ldots,0,\alpha_1,0,\ldots,0)$, $d^{12}(\theta) = (\beta_2,0,\ldots,0,0,\alpha_1,\ldots,0)$ etc. M.b.v. een dergelijke aanpak zijn de $C_{ij}$'s voor model III e.v. berekend.

Dankwoord. De referee heeft het artikel zeer zorgvuldig gelezen en bijgedragen aan een verbeterde presentatie van het geheel.

Referenties

J.C. van Houwelingen (1983), Principal components of large matrices with missing elements, preprint 310, Math. Inst. Utrecht.
(verschijnt in Proceedings of 3rd Prague Symposium of Asymptotic Statistics).

P.M. de Jong (1982), The reliability of methodes for predicting missing figures in migration tables, Metron 40, 171-182.

P.M. Kroonenberg (1983), Three-mode principal component analysis, Dissertatie, Rijksuniversteit Leiden.

C.R. Rao (1973), Linear Statistical Inference and its applications, Wiley.

M.C. Spanjer (1984), Substituent interaction effects and mathematical-statistical description of retention in liquid chromatography, Dissertatie Rijksuniversiteit Utrecht.

H.M. Weesie en J.C. van Houwelingen (1983), GEPCAM User's manual, Instituut voor Mathematische Statistiek, Utrecht.