Lineaire algebra in de statistiek

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

COSOR Memorandum 78-03

Lineaire Algebra in de Statistiek

Voordracht Wintersymposium van het wiskundig genootschap, op 7 januari 1978 te Roermond

A.J. Bosch

februari 1978

Nederland

Lineaire Algebra in de Statistiek
Voordracht Wintersymposium van het wiskundig genootschap, op 7 januari 1978
te Roermond

A.J. Bosch

1. Inleiding

Lineaire Algebra en Statistiek werden op de Nederlandse Universiteiten rond 1946 voor het eerst gedoceerd. Veel later zijn beide vakken opgenomen in het programma van de leraarsopleidingen, en wel is Statistiek pas in 1975 in het programma gekomen. Nu is het reeds zover dat beide vakken op middelbare scholen onderwezen worden. De reden is dat beiden op vele uiteenlopende gebieden worden toegepast, zelfs in de α- en γ-wetenschappen.
De multivariate analyse (hieronder vallen o.a.: discriminantanalyse, componenten- en factoranalyse, clusteranalyse, canonieke correlatie e.d.) wordt nl. zeer veel toegepast in de antropologie, sociologie, psychologie, planologie, medicijnen, landbouw, econometrie e.d. Echter zonder de lineaire algebra zou het onmogelijk zijn multivariate analyse te bedrijven.
Uit de vele toepassingen van de lineaire algebra in de multivariate analyse, wil ik er slechts één behandelen: de meetkundige interpretatie van correlaties. Hierdoor wordt het geheel veel duidelijker.
Uiteindelijk zullen we het toepassen op het volgende probleem:
U hebt ongetwijfeld gehoord van de bewering: "Het aantal pakjes sigaretten dat men per dag rookt is hoog gecorreleerd met de kans op een hartinfarct (of longkanker)".
Wat moet men met deze uitspraak? Is het dan werkelijk zo dat, indien men minder gaat roken ook de kans op een hartinfarct afneemt? Zoals in § 6 zal blijken moet men zeer zorgvuldig tewerk gaan wil men zo'n conclusie mogen trekken.

2. Benodigde basiskennis matrixrekening

1. $\mathrm{sp}\, A_{n\times n} := \sum_{1}^{n} a_{ii}$; $\mathrm{expsp}\, A := \exp(\mathrm{sp}\, A) = e^{\mathrm{sp}\, A}$;
   Er geldt $\mathrm{sp}(ABC) = \mathrm{sp}(CAB) = \mathrm{sp}(BCA)$.

2. A' $(= A^T)$ is de getransponeerde van A.
   Er geldt $(AB)' = B'A'$; $A^{-T} := (A^{-1})' = (A')^{-1}$.

3. $|A| := \det A$; $I := \{\delta_{ij}\}$ is de eenheidsmatrix; δ_{ij} het Kronecker-symbool.

4. a_{i*} is de i-de rij van A als kolom geschreven
   a'_{i*} is de i-de rij van A als rij geschreven
   a_{*j} is de j-de kolom van A als kolom geschreven
   a'_{*j} is de j-de kolom van A als rij geschreven.

   $A_{p\times q} = (a_{*1}\dots a_{*q})$; $\langle A\rangle$ is de kolommenruimte van $A_{p\times q}$, dat is de lineaire deelruimte in $\mathbb{R}^p$ opgespannen door de kolommen van A.

   Er geldt $Ax \in \langle A\rangle$.

5. $A > 0$ (A heet positief definiet) d.w.z. $A' = A$ en $\forall x \neq 0$: $x'Ax > 0$;
   $A \geq 0$ (A heet niet negatief definiet) d.w.z. $A' = A$ en $\forall x$: $x'Ax \geq 0$;
   $A > B$ d.w.z. $A - B > 0$.

6. $A^{\underline{2}} := A'A$; $A^{-\underline{2}} := (A^{\underline{2}})^{-1}$; $a^{\underline{2}} = a'a$
   $(A^{\underline{2}})_{ij} = (A'A)_{ij} = a'_{*i}a_{*j}$
   Er geldt $A^{\underline{2}} \geq 0$.

3. Benodigde basiskennis statistiek

	a) 1-dimensionaal $\underline{x}$	b) p-dimensionaal $\underline{x}' = (\underline{x}_1,\dots,\underline{x}_p)$
1.	$\mu := \mathcal{E}\underline{x}$	$\mu' := \mathcal{E}\underline{x}' = (\mu_1,\dots,\mu_p)$
2.	$\mathrm{cov}(\underline{x},\underline{y}) := \mathcal{E}(\underline{x}-\mu_x)(\underline{y}-\mu_y)$	$\mathrm{cov}(\underline{x},\underline{y}) := \mathcal{E}(\underline{x}-\mu_x)(\underline{y}-\mu_y)'$
3.	$\mathrm{var}\ \underline{x} := \mathcal{E}(\underline{x}-\mu)^2 = \sigma^2$	$\mathrm{var}\ \underline{x} := \mathcal{E}(\underline{x}'-\mu')^{\underline{2}} = \Sigma^{\underline{2}}$
4.	$\mathrm{cov}(a\underline{x},b\underline{y}) = ab\ \mathrm{cov}(\underline{x},\underline{y})$	$\mathrm{cov}(A\underline{x},B\underline{y}) = \mathcal{E}[(A\underline{x}-A\mu_x)(B\underline{y}-B\mu_y)'] = A\,\mathrm{cov}(\underline{x},\underline{y})B'$
5.	$\mathrm{var}(a\underline{x}) = a^2\sigma^2$	$\mathrm{var}(A\underline{x}) = A\Sigma^{\underline{2}}A'$
6.	$\mathrm{cov}(\underline{x}+b\underline{y},\underline{z}) = \mathrm{cov}(\underline{x},\underline{z}) + b\ \mathrm{cov}(\underline{y},\underline{z})$	$\mathrm{cov}(\underline{x}+B\underline{y},\underline{z}) = \mathrm{cov}(\underline{x},\underline{z}) + B\ \mathrm{cov}(\underline{y},\underline{z})$
7.	$\underline{\chi} \sim N(0,1)$ standaard normaal	$\underline{\chi}_p \sim N_p(0,I)$ standaard normaal $\underline{\chi}_p = (\underline{\chi}_{(1)},\dots,\underline{\chi}_{(p)})'$ alle onderl. onafh.
8.	$f(\chi) = (2\pi)^{-\frac{1}{2}}\exp(-\frac{1}{2}\chi^2)$	$f(\chi_p) = (2\pi)^{-\frac{1}{2}p}\exp(-\frac{1}{2}\chi_p^{\underline{2}})$
9.	$\underline{x} = \mu + \underline{\chi}\sigma \sim N(\mu,\sigma^2)$	$\underline{x}' = \mu' + \underline{\chi}_p'\Sigma \sim N_p(\mu',\Sigma^{\underline{2}})$
10.	$f(x) = \frac{1}{\sqrt{\sigma^2}(2\pi)^{\frac{1}{2}}}\exp[-\frac{1}{2}(x-\mu)^2\sigma^{-2}]$	$f(x) = \frac{1}{\sqrt{\lvert\Sigma^2\rvert}(2\pi)^{\frac{1}{2}p}}\ \mathrm{expsp}[-\frac{1}{2}(x'-\mu')^{\underline{2}}\Sigma^{-\underline{2}}]$

Toelichting bij het overzicht:

Ad 1b. De verwachting van een stochastische vector is de vector van verwachtingen der componenten.

2b. Stel $\underline{y} = (\underline{y}_1,\dots,\underline{y}_q)'$, dan is $\mathrm{cov}(\underline{x},\underline{y})$ een $p \times q$ matrix;
$\mathrm{cov}(\underline{y},\underline{x}) = \mathrm{cov}'(\underline{x},\underline{y})$.

Ad 3b. $\underline{\Sigma}^2$ heet de variantie-covariantie matrix van $\underline{x}$. Dat deze notatie geoorloofd is blijkt uit:

5b. $\mathrm{var}(A\underline{x}) = A\underline{\Sigma}^2A'$ volgt uit 4b. met $B = A$ en $\underline{y} = \underline{x}$. Neem $A = \alpha' = (\alpha_1,\ldots,\alpha_p)$. Nu is $0 \le \mathrm{var}\ \alpha'\underline{x} = \alpha'\underline{\Sigma}^2\alpha$. Inderdaad dus $\underline{\Sigma}^2 \ge 0$. $\underline{\Sigma}^2 > 0$ als de $\underline{x}_i$ lineair onafhankelijk zijn.

7b. $\underline{\chi}_p$ is een vector van p o.o. stand. normaal verdeelde variabelen $\underline{\chi}_{(i)}$.

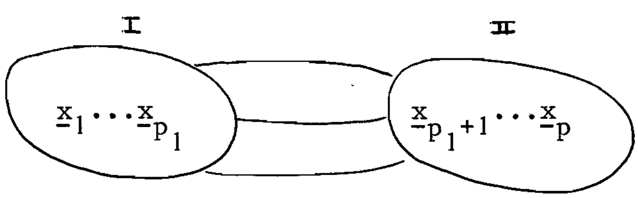
8b. $f(\chi_p) = \prod_1^p f(\chi_{(i)})$ wegens de onafhankelijkheid der $\underline{\chi}_{(i)}$.

9b. $\underline{x}$ is p-dimensionaal normaal verdeeld. Hier is Σ een $p \times p$ matrix. Is deze singulier, dan ook $\underline{\Sigma}^2$ en heet $\underline{x}$ singulier normaal verdeeld.

10b. $\underline{x} = \mu + \Sigma'\underline{\chi}_p$; $\mathrm{var}\ \underline{x} = \Sigma' I \Sigma = \underline{\Sigma}^2$. Ondersteld is $\underline{\Sigma}^2 > 0$, anders bestaat $\underline{\Sigma}^{-2}$ niet en hiermee ook niet de kansdichtheid $f(x)$.
$\mathrm{expsp}[-\frac{1}{2}(x'-\mu')^{\underline{2}}\underline{\Sigma}^{-2}] = \exp[-\frac{1}{2}(x-\mu)'\underline{\Sigma}^{-2}(x-\mu)]$. Dit volgt uit 2.1.

4. Correlatie

We beschouwen 2 verzamelingen stochasten die onderling samenhangen:



Noteer $\underline{x}^{(1)} := (\underline{x}_1,\ldots,\underline{x}_{p_1})'$; $\underline{x}^{(2)} := (\underline{x}_{p_1+1},\ldots,\underline{x}_p)'$; $\underline{x} = (\underline{x}^{(1)'},\underline{x}^{(2)'})'$

$\mathrm{var}\ \underline{x} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} > 0$ ondersteld. Dus $\mathrm{var}\ \underline{x}^{(1)} = \Sigma_{11}$, $\mathrm{var}\ \underline{x}^{(2)} = \Sigma_{22}$,
$\mathrm{cov}(\underline{x}^{(1)},\underline{x}^{(2)}) = \Sigma_{12} = \Sigma_{21}'$. Zij $\underline{x} \sim N_p(\mu,\underline{\Sigma}^2)$.
Nu poneren we, zonder afleiding, de voorwaardelijke verdeling van $\underline{x}^{(1)}$ onder de voorwaarde $\underline{x}^{(2)} = x^{(2)}$

$$f(x^{(1)}|x^{(2)}) = N_{p_1}(\mu^{(1)} + B(x^{(1)} - \mu^{(2)});\Sigma_{11.2}) .$$

Hierin is $B := \Sigma_{12}\Sigma_{22}^{-1}$ de matrix der regressiecoëfficiënten
$\Sigma_{11.2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ de matrix der partiële varianties en covarianties.

Opmerking.

1. Eenvoudig is te bewijzen dat $\underline{\Sigma}^2 > 0 \Rightarrow \Sigma_{11},\Sigma_{22},\Sigma_{11.2} > 0$.
2. Noteer $\Sigma_{1.2} := \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ dan is dus $\Sigma_{11} = \Sigma_{11.2} + \Sigma_{1.2}$.

3. Is $\Sigma_{12} = 0$, d.w.z. $\underline{x}_i \in$ I ongecorreleerd met $\underline{x}_j \in$ II voor alle i en j, dan is $B = 0$ en $\Sigma_{11.2} = \Sigma_{11}$ zodat $f(x^{(1)}\ x^{(2)}) = f(x^{(1)})$ oftewel $\underline{x}^{(1)}$ en $\underline{x}^{(2)}$ zijn stoch. onafhankelijk.

Zonder commentaar (de betekenis volgt in § 6) definiëren we de volgende correlatiecoëfficiënten:

1. $\rho_{ij} := \dfrac{(\Sigma_{11})_{ij}}{\sqrt{(\Sigma_{11})_{ii}(\Sigma_{11})_{jj}}}$, de <u>enkelvoudige</u> correl. coef. tussen $\underline{x}_i$ en $\underline{x}_j \in$ I.

2. $\rho_{ij.2} := \dfrac{(\Sigma_{11.2})_{ij}}{\sqrt{(\Sigma_{11.2})_{ii}(\Sigma_{11.2})_{jj}}}$, de <u>partiële</u> correl. coef., dat is de enkelvoudige correl. coef. tussen $\underline{x}_i$ en $\underline{x}_j \in$ I onder de voorwaarde $\underline{x}^{(2)} = x^{(2)}$.

3. $\rho_{i.2} := \dfrac{(\Sigma_{1.2})_{ii}}{\sqrt{(\Sigma_{11})_{ii}(\Sigma_{1.2})_{ii}}}$, de <u>multiple</u> correl. coef., dat is de enkelvoudige correl. coef. tussen $\underline{x}_i \in$ I en de lineaire combinatie $\in$ II waarmee $\underline{x}_i$ maximaal gecorreleerd is.

5. <u>Een meetkundige interpretatie van</u> Σ^2

Voor de bestudering van de samenhang van stochasten is het geen beperking aan te nemen dat $\mathcal{E}\underline{x} = 0$. Verder onderstellen we var $\underline{x}_i < \infty$, $i = 1,\ldots,p$.

a) <u>Lineaire Algebra</u>	b) <u>Multivariate Analyse</u>
1. Zij V een vectorruimte.	Zij $V = \mathcal{L}^2$ vectorr. van stochasten (= vectoren).
2. Zij $D \subset V$ een p dim. deelruimte opgespannen door $x_1,\ldots,x_p$.	$D \subset V$ een p dim. deelruimte opgespannen door $\underline{x}_1,\ldots,\underline{x}_p$.
3. Definieer een inproduct (x,y) en norm $\Vert x\Vert^2 = (x,x)$.	$(\underline{x},\underline{y}) := \mathcal{E}(\underline{x},\underline{y}) = \mathrm{cov}(\underline{x},\underline{y})$ $\Vert\underline{x}\Vert^2 = \mathrm{var}\ \underline{x}$.
4. $\cos\varphi_{x,y} := \dfrac{(x,y)}{\Vert x\Vert\ \Vert y\Vert}$	$\cos\varphi_{x,y} = \dfrac{\mathrm{cov}(\underline{x},\underline{y})}{\sigma_x\sigma_y} = \rho(\underline{x},\underline{y})$.
5. $\cos\varphi = 0$ als $x \perp y$ $\cos\varphi = \pm 1$ als $x = \alpha y$.	$\rho = 0$ als $\underline{x}$ en $\underline{y}$ ongecorreleerd $\rho = \pm 1$ als $\underline{x}$ en $\underline{y}$ lineair afhankelijk: $\underline{x} = \alpha\underline{y}$.
6. Zij $e_1,\ldots,e_p$ orthon. basis van D, dus $(e_i,e_j) = \delta_{ij}$.	$\underline{\chi}_{(1)},\ldots,\underline{\chi}_{(p)}$ basis van D van ongecorreleerde gestandaarde stochasten, nl. $(\underline{\chi}_{(i)},\underline{\chi}_{(j)}) = \delta_{ij}$.

7. Zij $x_i = x_{1i}e_1 + \ldots + x_{pi}e_p$.
Noteer $(x_{1i},\ldots,x_{pi}) = x'_{*i}$.
Dan is $(x_i,x_j) = x'_{*i}x_{*j}$ en
$\|x_i\|^2 = x^{\underline{2}}_{*i}$.
Zij $X := (x_{*1}\ldots x_{*p})$ dan is dus
$(x_i,x_j) = (X^{\underline{2}})_{ij}$ oftewel $X^{\underline{2}}$ is de inproduktmatrix en
$|X^{\underline{2}}| = \mathrm{vol}^2(x_{*1}\ldots x_{*p})$.

$\underline{x}' = \underline{\chi}'_p\Sigma$ geeft
$\underline{x}_i = \sigma_{1i}\underline{\chi}_{(1)} + \ldots + \sigma_{pi}\underline{\chi}_{(p)}$
$(\sigma_{1i},\ldots,\sigma_{pi}) = \sigma'_{*i}$
$(\underline{x}_i,\underline{x}_j) = \sigma'_{*i}\sigma_{*j}$
$\|\underline{x}_i\|^2 = \mathrm{var}\ \underline{x}_i = \sigma^{\underline{2}}_{*i}$
$\Sigma = (\sigma_{*1},\ldots,\sigma_{*p})$
$(\underline{x}_i,\underline{x}_j) = (\Sigma^{\underline{2}})_{ij}$
$\Sigma^{\underline{2}}$ is de inproduktmatrix van $\underline{x}_1,\ldots,\underline{x}_p$
$|\Sigma^{\underline{2}}| = \mathrm{vol}^2(\sigma_{*1}\ldots\sigma_{*p})$.

Toelichting bij het overzicht:

Ad 1b. Stochasten vormen een vectorruimte. Om dit precies te bewijzen moet men maattheoretisch te werk gaan.
Zo is de "inverse" van $\underline{x}$ xnl $-\underline{x}$; het "eenheidselement" is de stochast $\underline{0}$, dat is een stochast die overal 0 is, op een verzameling van de maat nul na.

2b. De $\underline{x}_i$ zijn hier dus lineair onafhankelijk ondersteld.
Pas op: stochastisch onafhankelijke $\underline{x}_i$ is een veel zwaardere eis, nl. zijn $\underline{x}_i$ en $\underline{x}_j$ stoch. onafhankelijk, dan is dus $\rho(\underline{x}_i,\underline{x}_j) = 0$ oftewel $\underline{x}_i \perp \underline{x}_j$, dus zeker lineair onafhankelijk.

3b. Eenvoudig is na te gaan dat dit een inprodukt definieert.

4b. Hieruit volgt direkt dat $|\rho| \leq 1$, nl. $\rho = \cos\varphi$.

6b. In het algemeen is het niet noodzakelijk dat de basisvectoren normaal verdeeld zijn.

7b. Σ bevat dus de componenten der $\underline{x}_i$ t.o.v. basis der $\underline{\chi}_{(i)}$.
Merk op dat $\cos\varphi$ tussen $\underline{x}_i$ en $\underline{x}_j$ gelijk is aan $\cos\varphi$ tussen σ_{*i} en $\sigma_{*j} \in \mathbb{R}^p$. Analoog is de norm van $\underline{x}_i$ gelijk aan de norm van $\sigma_{*i} \in \mathbb{R}^p$. Een goed beeld van de samenhang der $\underline{x}_i$ wordt zodoende gevormd door de σ_{*i}.

6. Meetkundige interpretatie van de verschillende correlatiecoëfficiënten

Allereerst bepalen we enkele varianties en covarianties. Noteer $B\underline{x}^{(2)} =: \hat{\underline{x}}^{(1)}$

$$\mathrm{cov}(\underline{x}^{(1)} - \hat{\underline{x}}^{(1)}, \underline{x}^{(2)}) = \mathrm{cov}(\underline{x}^{(1)},\underline{x}^{(2)}) - B\ \mathrm{cov}(\underline{x}^{(2)},\underline{x}^{(2)}) = \Sigma_{12} - B\Sigma_{22} = 0.$$

$$\mathrm{cov}(\underline{x}^{(1)},\hat{\underline{x}}^{(1)}) = \mathrm{cov}(\underline{x}^{(1)},B\underline{x}^{(2)}) = \Sigma_{12}B' = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{1.2}.$$

$$\mathrm{var}\ \hat{\underline{x}}^{(1)} = \mathrm{var}(B\underline{x}^{(2)}) = B\Sigma_{22}B' = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{1.2}.$$

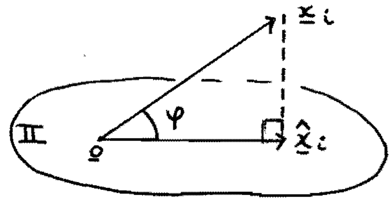
$$\mathrm{var}(\underline{x}^{(1)} - \hat{\underline{x}}^{(1)}) = \mathrm{var}\ \underline{x}^{(1)} + \mathrm{var}\ \hat{\underline{x}}^{(1)} - 2\ \mathrm{cov}(\underline{x}^{(1)},\hat{\underline{x}}^{(1)}) = \Sigma_{11} + \Sigma_{1.2} - 2\Sigma_{1.2} = \Sigma_{11.2}.$$

Wat betekent dit alles nu meetkundig?
Voor $\underline{x}_i \in I$ is dus $\mathrm{cov}(\underline{x}_i - \hat{\underline{x}}_i, \underline{x}^{(2)}) = 0$ d.w.z. $x_i - \hat{x}_i \perp$ ruimte opgespannen door $\underline{x}_{p_1+1}, \ldots, \underline{x}_p \in II$. Nu is $\hat{\underline{x}}_i = (B\underline{x}^{(2)})_i = b_{i*}^{\prime} x^{(2)} \in II$.

Conclusie: $\hat{\underline{x}}_i$ is de <u>orthogonale projectie</u> van $\underline{x}_i$ op de deelruimte II.
Er geldt dus $\|\underline{x}_i\|^2 = \|\underline{x}_i - \hat{\underline{x}}_i\|^2 + \|\hat{\underline{x}}_i\|^2$ oftewel in statistische termen $\mathrm{var}\ \underline{x}_i = \mathrm{var}(\underline{x}_i | x^{(2)}) + \mathrm{var}\ \hat{\underline{x}}_i$. Dit zijn de diagonaalelementen van $\Sigma_{11} = \Sigma_{11.2} + \Sigma_{1.2}$.
We zagen reeds dat ρ_{ij} = cosinus van de hoek tussen $\underline{x}_i$ en $\underline{x}_j \in I$.
Wat stelt $\rho_{i.2}$ nu voor?

$$\cos(\underline{x}_i, \hat{\underline{x}}_i) = \cos\varphi = \frac{\|\hat{\underline{x}}_i\|}{\|\underline{x}_i\|} = \frac{\|\hat{\underline{x}}_i\|^2}{\|\underline{x}_i\|\,\|\hat{\underline{x}}_i\|} = \frac{(\Sigma_{1.2})_{ii}}{\sqrt{(\Sigma_{11})_{ii}(\Sigma_{1.2})_{ii}}} = \rho_{i.2}\ .$$

Dit klopt inderdaad met de definitie van multiple corr. coef. Deze is dus gelijk aan de cosinus van de hoek tussen $\underline{x}_i$ en zijn projectie. Inderdaad is $\hat{\underline{x}}_i$ maximaal met $\underline{x}_i$ gecorreleerd.

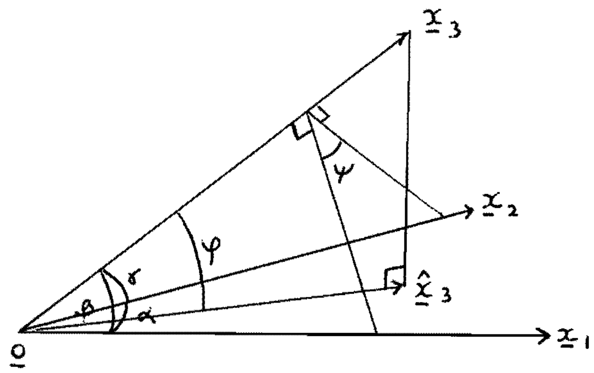


Er geldt $\mathrm{var}(\underline{x}_i | x^{(2)}) = \mathrm{var}\ \underline{x}_i - \mathrm{var}\ \hat{\underline{x}}_i = \mathrm{var}\ \underline{x}_i(1 - \rho_{i.2}^2) \le \mathrm{var}\ \underline{x}_i$. Eenvoudig is te bewijzen dat zelfs $\Sigma_{11.2} \le \Sigma_{11}$.
Is $\underline{x}_i \perp II$ dan is dus $\rho_{i.2} = 0$. Dit klopt, nl. $(\Sigma_{1.2})_{ii} = (\Sigma_{12})_{i*}^{\prime}\Sigma_{22}^{-1}(\Sigma_{21})_{*i} = 0$ daar $\mathrm{cov}(\underline{x}_i, \underline{x}^{(2)}) = (\Sigma_{12})_{i*}^{\prime} = 0$.
Wat stelt $\rho_{ij.2}$ voor?
Uit de formule voor $\rho_{ij.2}$ blijkt direct dat deze gelijk is aan de cosinus van de hoek tussen $\underline{x}_i - \hat{\underline{x}}_i$ en $\underline{x}_j - \hat{\underline{x}}_j$ waarbij $\underline{x}_i$ en $\underline{x}_j \in I$.
Nu zijn $\underline{x}_i - \hat{\underline{x}}_i$ en $\underline{x}_j - \hat{\underline{x}}_j$ beide $\perp$ II, d.w.z. beide zijn ongecorreleerd met $\underline{x}^{(2)}$. Dus de partiële correl. coef. is de correl. coef. tussen $\underline{x}_i$ en $\underline{x}_j$ na "eliminatie" van het effect van $\underline{x}^{(2)}$. Waren $\underline{x}_i$ en $\underline{x}_j$ reeds ongecorreleerd met $\underline{x}^{(2)}$, d.w.z. $\underline{x}_i$ en $\underline{x}_j \perp$ II, dan zijn $\hat{\underline{x}}_i = \hat{\underline{x}}_j = 0$ en dus $\rho_{ij.2} = \rho_{ij}$.

Voor $p = 3$ geeft de volgende figuur een verhelderend beeld:

$\cos\alpha = \rho_{12}$; $\cos\beta = \rho_{13}$; $\cos\gamma = \rho_{23}$; $\cos\varphi = \rho_{3.12}$;

$\cos\psi = \rho_{12.3}$.



Zijn $\cos\alpha$, $\cos\beta$, $\cos\gamma$ gegeven, dan zijn $\cos\varphi$ en $\cos\psi$ te berekenen. De onderlinge ligging (samenhang) der vectoren ligt volkomen vast door de <u>correlatiematrix</u> $R := D^{-1}\underline{\Sigma}^2 D^{-1}$ met

$$D := \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}.$$

Te bewijzen is dat geldt $|R| = \prod_2^p (1-\rho^2_{i\cdot 1\ldots(i-1)})$ en

$|\underline{\Sigma}^2| = \operatorname{var} \underline{x}_1 \prod_2^p \operatorname{var}(\underline{x}_i | x_1 \ldots x_{i-1})$.

We zien dat $|R| = 0$ zodra er een $\rho_{i\cdot 1\ldots(i-1)} = 1$.

$|R| = 1$ als alle $\rho_{i\cdot 1\ldots(i-1)} = 0$ d.w.z. alle $\underline{x}_i$ onderling ongecorreleerd oftewel $R = I$.

Analoog is $|\underline{\Sigma}^2| = \prod_1^p \operatorname{var} \underline{x}_i$ d.e.s.d. als $\underline{\Sigma}^2 = D^2$ (dus diagonaal).

Voor $p = 3$ krijgen we: $|R| = (\text{volume})^2 = (\text{hoogte} \times \text{opp. grondvlak})^2 =$

$= \sin^2\varphi \sin^2\alpha = (1-\rho^2_{3.12})(1-\rho^2_{12})$.

Hieruit volgt $\rho^2_{3.12} = \dfrac{\rho^2_{13} + \rho^2_{23} - 2\rho_{12}\rho_{13}\rho_{23}}{1-\rho^2_{12}}$.

Analoog is meetkundig te bewijzen dat $\rho_{12.3} = \dfrac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho^2_{13})(1-\rho^2_{23})}}$.

We zien dat als $\rho_{13} = \rho_{23} = 0$, $\rho_{3.12} = 0$ en $\rho_{12.3} = \rho_{12}$.

Bekijken we nu nog eens de bewering betreffende het roken. Laten we aannemen dat uit waarnemingen is gebleken dat inderdaad het aantal pakjes sigaretten dat men per dag rookt hooggecorreleerd is met de kans op een hartinfarct. Is het sigarettenroken dan wel de boosdoener? M.a.w. als men minder gaat roken neemt de kans op een hartinfarct dan ook af of anders gezegd is er een oorzakelijk verband? Noem de stochast $\underline{x}_1$ = het aantal pakjes dat iemand per dag rookt; $\underline{x}_2$ = de kans op een hartinfarct.

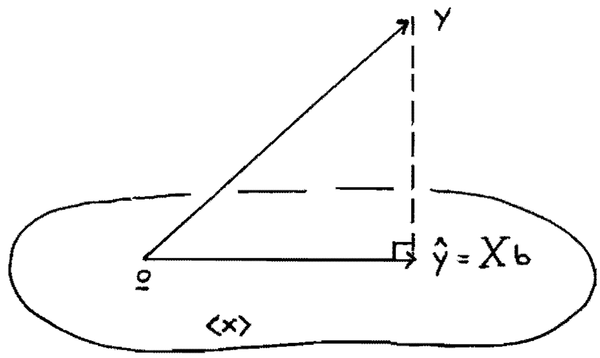
Gemeten is dus dat ρ_{12} hoog is, dus de hoek tussen $\underline{x}_1$ en $\underline{x}_2$ vrij klein. Onderstel nu dat er een andere stochast $\underline{x}$ is (ons nog onbekend, of zelfs meerdere) die met beide hooggecorreleerd is. Dat betekent dat $\rho_{3.12}$ groot is oftewel dat $\underline{x}_3$ dicht bij het vlak ligt opgespannen door $\underline{x}_1$ en $\underline{x}_2$. Bekijken we nu eens $\rho_{12.3}$ d.w.z. de partiële correlatie tussen $\underline{x}_1$ en $\underline{x}_2$ nà eliminatie van de stochast $\underline{x}_3$. Dan kan het zo zijn (zie fig.) dat cos $\psi = 0$ (of zelfs < 0) m.a.w. dat $\underline{x}_1$ en $\underline{x}_2$ ongecorreleerd zijn nà eliminatie van $\underline{x}_3$. In dat geval heeft een wijziging in de rookgewoonten (mits men dus $\underline{x}_3$ constant houdt) gèèn invloed op de kans op een hartinfarct. Zo'n stochast $\underline{x}_3$ zou in dit voorbeeld best eens de mate van stress kunnen zijn. Stress verhoogt de bloeddruk (en dus de kans op een infarct) en mensen onder stress zijn wellicht geneigd eerder naar een sigaret te grijpen.
Dit voorbeeld toont dus aan dat, wil men een causaal verband tussen twee stochasten aantonen, de onbekende stochasten waarvan $\underline{x}_1$ en $\underline{x}_2$ afhankelijk zijn, gedurende de waarnemingen constant (of geëlimineerd) moeten zijn.
Is de proefopzet dus onnauwkeurig geschied, dan kan men gemakkelijk tot foute conclusies komen.

7. Samenhang met het lineaire model

Onderstel weer $\mathcal{E}\underline{x} = 0$; $f(x^{(1)}|x^{(2)}) = N_{p_1}(Bx^{(2)};\Sigma_{11.2})$.
Vaak is $\Sigma_{11.2}$ en dus ook B onbekend.
We voeren nu een andere notatie in, meer gebruikelijk in lineaire modellen:

Noteer $\underline{x}_i \in I$: $\underline{y}$
$\hat{\underline{x}}_i$: $\hat{\underline{y}}$
$(B)'_{i*} = b'_{i*}$: $(\beta_0,\beta_1,\ldots,\beta_p) = \beta'$
$x^{(2)}$: $(x_0,x_1,\ldots,x_p) = x'$.

Dan wordt $\mathcal{E}(\underline{x}_i|x^{(2)}) = b'_{i*}x^{(2)}$: $\mathcal{E}(\underline{y}|x) = \beta_0x_0 + \ldots + \beta_px_p = \beta'x$ oftewel
$\underline{y} = \beta_0x_0 + \ldots + \beta_px_p + \underline{e}$ met $\mathcal{E}\underline{e} = 0$; var $\underline{e} = \sigma^2$.
Gevraagd β te schatten uit een steekproef.
Het steekproefmodel wordt nu: $\underline{y} = X\beta + \underline{e}$ met $\mathcal{E}\underline{e} = 0$ en var $\underline{e} = \sigma^2 I$.
Zij nu $y = (y_1,\ldots,y_n) \in \mathbb{R}^n$ een realisatie van $\underline{y}$. Als y zonder fout behept was, zou $y = X\beta \in \langle X\rangle$. De beste schatting voor $X\beta$ (d.i. de kleinste kwadraten schatting) is dus de loodrechte projectie van y op $\langle X\rangle$.
$X'(Y - Xb) = 0$, $X'y = X'Xb$; $b = (X'X)^{-1}X'y$ en $b' = y'X(X'X)^{-1} = (S_{12})'_{i*}S_{22}^{-1}$, waarbij S_{12} en S_{22} als schattingen optreden voor resp. Σ_{12} en Σ_{22}.



Evenals $\hat{\underline{x}}_i$ de projectie is van $\underline{x}_i$ op de ruimte opgespannen door de componenten van $x^{(2)}$, zo is $\hat{y}$ de projectie van y op de ruimte opgespannen door de kolommen van X.

$b'_{i*} = (\Sigma_{12}\Sigma_{22}^{-1})'_{i*} = (\Sigma_{12})'_{i*}\Sigma_{22}^{-1}$; b' is dus 'n schatting van b'_{i*}.