Noter til

Multivariate Statistiske Metoder med SAS

Erling B. Andersen

Juli 2002

Forord.

I dette notesæt er den nødvendige statistiske teori for at kunne benytte SAS-procedurerne Proc Factor, Proc Calis og Proc Corresp gennemgået. Desuden er det vist, hvordan man benytter SAS-proceduren Proc IML til mest almindelige matrixberegninger.

Den største vægt er dog lagt på at illustrere anvendelsen af de nævnte SAS-procedurer. Her er for Proc Factor og Proc Calis benyttet den meget store tværnatonale database ISSP, der administreres af ZUMA i Köln. For Proc Corresp's vedkommende er der benyttet en række variable fra den store danske Velfærdsundersøgelse, professor Erik Jørgen Hansen gennemførte i 1974, og publicerede resultaterne i 1976.

Notesættet, der første gang blev brugt i foråret 2001, er belvet revideret of udvidet i forbindelse med mit kursus i 'Multivariat Statistik med SAS', som bliver gennemført for polit-studerende i efterårssemestret 2002.

Rungsted. Juli 2002. Erling B. Andersen

Indhold

1. ISSP databasen.

Gennemgangen af metoderne i disse noter eksemplificeres ved data fra den såkaldte ISSP database. Den omfatter et stort antal spørgsmål, stillet gennem en en årrække fra 1985 og fremefter, i et varierende antal lande.

```
ISSP 89 - Module - Work Orientations - ZA No. 1840

V3    COUNTRY

Country
-------
                                                          Unweighted
                                                          Abs.      %

02. West Germany                  (D)                     1575  10.66
03. Great Britain                 (GB)                    1297   8.78
04. USA                           (USA)                   1453   9.84
05. Austria                       (A)                     1997  13.52
06. Hungary                       (H)                     1000   6.77
07. Netherlands                   (NL)                    1690  11.44
08. Italy                         (I)                     1028   6.96
09. Ireland                       (IRL)                    972   6.58
10. Northern Ireland              (NIRL)                   780   5.28
11. Norway                        (N)                     1848  12.51
12. Israel                        (IL)                    1133   7.67
                                                         ----- ------
                                                         14773 100.00


V24   IMPORTANT: JOB SECURITY

Q.6  From the following list, please tick one box for each
item to show how important you personally think it is in
a job?

Q.6a  How important is: Job security?
(Please tick one box on each line)
-------------------------------------------------------------

    1. Very important
    2. Important
    3. Neither important nor unimportant
    4. Not important
    5. Not important at all

    8. Can't choose, don't know
    9. NA


V25   IMPORTANT: HIGH INCOME

Q.6b  How important is: High income?
-----------------------------------

    1. Very important
    2. Important
    3. Neither important nor unimportant
    4. Not important
    5. Not important at all

    8. Can't choose, don't know
    9. NA
```

Der foreligger en kodebog for de variable, der er udtrukket til undervisningsbrug. Ovenstående boks og nedenstående boks viser kodebogen for de variable V3, V24, V25 V26 og V27 for 1989 undersøgelsen, der, sammen med V28 - V32, især benyttes i noterne.

De variable der er vist her er dels V3: De deltagende lande, med antal respondenter og deres procentdel af hele stikprøven, dels de første 4 variable, V24 - V27, der drejer sig om hvilke forhold, respondenterne finder af betydning for deres job.

```
V26   IMPORTANT: ADVANCEMENT

Q.6c  How important is: Good opportunities for advancement?
-------------------------------------------------------------
    1. Very important
    2. Important
    3. Neither important nor unimportant
    4. Not important
    5. Not important at all

    8. Can't choose, don't know
    9. NA

V27   IMPORTANT: LEISURE TIME

Q.6d  How important is: A job that leaves a lot of leisure
time?

    1. Very important
    2. Important
    3. Neither important nor unimportant
    4. Not important
    5. Not important at all

    8. Can't choose, don't know
    9. NA
```

Den næste boks viser nogle af de baggrundsvariable, vi skal bruge i forbindelse med stianalyserne siden hen.

V90 WORKING HOURS

How many hours a week do you normally work in your main job?

---

02. Two hours
.. ..
95. 95 hours and more

98. Don't know
99. NA, no hours
00. Not applicable

V100 EDUCATION

Years in school

01. 1 year
.. ..

22. 22 years

95. Still at school
96. Still at college, university
98. Don't know
99. NA
00. No formal schooling

V110 SUBJECTIVE SOCIAL CLASS

Most people see themselves as belonging to a particular class. Please tell me which social you would say you belong to?

1. Lower class: Poor
2. Working class
3. Lower middle/ Upper working class
4. Middle class
5. Upper middle class
6. Upper class

8. Don't know
9. NA

2. Lidt matrixregning.

En matrix **A** af dimension n x m kan skrives:

$$\begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} & \dots & a_{im} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nm} \end{bmatrix}$$

Vektorer vil altid i disse noter være søjlevektorer, dvs. matricer af dimension n x 1. For eksempel er vektoren **b** lig med

$$\mathbf{b} = \begin{bmatrix} b_1 \\ \dots \\ b_i \\ \dots \\ b_n \end{bmatrix}.$$

Den transponerede matrix $\mathbf{A}^T$ er **A**, hvor rækker og søjler er byttet om, dvs. matricen

$$\begin{bmatrix} a_{11} & \dots & a_{i1} & \dots & a_{n1} \\ \dots & \dots & \dots & \dots & \dots \\ a_{1j} & \dots & a_{ij} & \dots & a_{nj} \\ \dots & \dots & \dots & \dots & \dots \\ a_{1m} & \dots & a_{im} & \dots & a_{nm} \end{bmatrix}$$

For en ikke-singulær, kvadratisk matrix **A** af dimension n er den inverse matrix $\mathbf{A}^{-1}$ en matrix, der opfylder

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I},$$

hvor **I** er enhedsmatricen af dimension n. Ikke-singulær betyder, at **A** har fuld rang, så

$$\det(\mathbf{A}) \neq 0 ,$$

hvor det(**A**) er **A**'s determinant.

En matrix af dimension n x m siges at være ortogonal, hvis

$$\mathbf{A} \cdot \mathbf{A}^T = \mathbf{I},$$

dvs. hvis

$$\sum_{j=1}^{m} a_{ij}^2 = 1 \quad , \quad \text{for alle rækker i}$$

og

$$\sum_{j=1}^{m} a_{ij} a_{kj} = 0 \quad , \quad \text{for alle par af rækker i og k .}$$

Man siger, at to vektorer **a** og **b** er ortogonale, hvis de begge har længden 1, dvs.

$$\sum_{i=1}^{n} a_i^2 = 1$$

og

$$\sum_{i=1}^{n} b_i^2 = 1 \ ,$$

samtidigt med, at produktsummen af deres elementer er 0, dvs.

$$\sum_{i=1}^{n} a_i b_i = 0 \ .$$

For en kvadratisk, symmetrisk og ikke-negativ definit matrix **A** af dimension n (sådan som variansmatricer er det), gælder, at der findes en kvadratisk matrix **B** af dimension n, så $\mathbf{B}^T$ er ortogonal (dvs. søjlerne i **B** er ortogonale), og en diagonal matrix Λ også af dimension n, så **A** kan skrives

$$\mathbf{A} = \mathbf{B} \, \Lambda \, \mathbf{B}^T.$$

Elementerne i Λ kaldes da egenværdierne, mens søjlerne i **B** kaldes egenvektorerne for **A**. Alle diagonalelementerne $\lambda_1, \ldots, \lambda_n$ i Λ er ikke-negative, og antal positive λ'er er lig med **A**'s rang.

Denne opspaltning af en kvadratisk matrix går under forskellige navne. På engelsk siges ofte: 'Single value decomposition'. Jeg skal blot referere til resultatet, som en egenværdiopspaltning eller en egenværdi-dekomponering.

I algebraen er resultatet noget bredere. Under visse betingelser kan en ikke kvadratisk matrix **C** opspaltes på tilsvarende måde, nemlig som

$$\mathbf{C} = \mathbf{G} \, \Lambda \, \mathbf{H}^T \ ,$$

hvor $\mathbf{G}^T$ og $\mathbf{H}^T$ er ortogonale. Man kan let bestemme **G**, **H** og Λ fra denne ligning. Hvis vi f.eks. ganger bagved med **H** og $\mathbf{G}^T$, får man idet $\mathbf{H}^T\,\mathbf{H} = \mathbf{I}$, hvis $\mathbf{H}^T$ er ortogonal,

$$\mathbf{C}\,\mathbf{H}\,\mathbf{G}^T = \mathbf{G}\,\Lambda\,\mathbf{G}^T\ ,$$

hvorfra **G** og Λ kan findes ved en egenværdiopspaltning. **H** bestemmes ved omvendt at gange foran med $\mathbf{G}^T$ og **H**, så man får

$$\mathbf{H}\,\mathbf{G}^T\,\mathbf{C} = \mathbf{H}\,\Lambda\,\mathbf{H}^T\ .$$

3. Matrixregning i SAS.

Matrixregning i SAS foregår i PROC IML. Her kan man definere en variabel som en matrix. Det foregår således, hvis vi som eksempel tager følgende korrelationsmatrix:

1.000	0.546	0.244	0.149
0.546	1.000	0.398	0.182
0.244	0.398	1.000	0.271
0.149	0.182	0.271	1.000

Kalder vi denne matrix for **V**, men i SAS-sprog blot for "v", kan man indlæse **V**, og finde den inverse $\mathbf{V}^{-1}$ på følgende simple måde, hvor inv(v) er lig med $\mathbf{V}^{-1}$:

```
proc iml;
v = {1.000   0.546   0.244   0.149,
     0.546   1.000   0.398   0.182,
     0.244   0.398   1.000   0.271,
     0.149   0.182   0.271   1.000};
vi = inv(v);
vny = v*vi;
print v vi vny;
run;
```

Læg mærke til, hvordan matricer indlæses i SAS. Man skriver blot matricens rækker adskilt ved et komma, mens hele matricen omkranses med 2 'Tuborg'-paranteser. Husk, som altid i SAS, at afslutte med et semikolon efter hvor ordre. I dette program kontrolleres også, ved at beregne $\mathbf{V} \cdot \mathbf{V}^{-1}$ som matricen 'vny', at $\mathbf{V} \cdot \mathbf{V}^{-1}$ rent faktisk er lig med enhedsmatricen **I** af dimension 4.

SAS-udskriften fra denne programstump bliver:

```
          V                                        VI
          1     0.546     0.244     0.149 1.4306219 -0.757125 -0.029477 -0.067378
      0.546         1     0.398     0.182 -0.757125 1.5973461 -0.434719 -0.060096
      0.244     0.398         1     0.271 -0.029477 -0.434719 1.2493311 -0.255058
      0.149     0.182     0.271         1 -0.067378 -0.060096 -0.255058 1.0900975


        VNY
          1 -2.48E-17 -2.17E-19 -6.48E-18
1.019E-17           1  -2.8E-17 -1.17E-17
-7.08E-18 1.547E-17           1 1.144E-17
1.385E-17 -1.95E-18 1.824E-17           1
```

Læg mærke til at SAS har en tendens til at skrive 0'er som et tal opløftet til en høj negativ potens (ofte fra 17 til 20).

Matrixmultiplikation i IML foregår sådan: Lad **A** være en matrix, der kan ganges forfra med **V**, f.eks. 2x4 matricen

```
0.7  0.5  0.4  0.3
0.3  0.4 -0.5 -0.7
```

Man danner da produktet **A** · **V** i IML som:

```
proc iml;
v = {1.000   0.546   0.244   0.149,
     0.546   1.000   0.398   0.182,
     0.436   0.398   1.000   0.271,
     0.149   0.182   0.271   1.000};
a = {0.7  0.5  0.4  0.3,
     0.3  0.4 -0.5 -0.7};
vp = a*v;
print v a vp;
run;
```

Udskriften fra en SAS-kørsel, hvor matricen **A** kaldes 'a', **V** 'v' og **A** · **V** kaldes 'vp', ser sådan ud:

```
          V                                                A
          1       0.546       0.244       0.149         0.7         0.5         0.4         0.3
      0.546           1       0.398       0.182         0.3         0.4        -0.5        -0.7
      0.244       0.398           1       0.271
      0.149       0.182       0.271           1


         VP
    1.1153       1.096      0.8511      0.6037
    0.2921      0.2374     -0.4573      -0.718
```

Man transponerer en matrix i PROC IML ved at skrive **A**T som t(a), hvor 'a' er den matrix, der skal transponeres. Følgende programstump viser beregningen af **A**T.

```
proc iml;
v = {1.000   0.546   0.244   0.149,
     0.546   1.000   0.398   0.182,
     0.436   0.398   1.000   0.271,
     0.149   0.182   0.271   1.000};
a = {0.7  0.5  0.4  0.3,
     0.3  0.4 -0.5 -0.7};
at = t(a);
print a at;
run;
quit;
```

Udskriften fra en SAS-kørsel, ser sådan ud

```
   A                                        AT
0.7        0.5        0.4        0.3       0.7        0.3
0.3        0.4       -0.5       -0.7       0.5        0.4
                                           0.4       -0.5
                                           0.3       -0.7
```

Hvis man ønsker at undersøge, om rækkerne i **A** er ortogonale, dvs. om $\mathbf{A} \cdot \mathbf{A}^T = \mathbf{I}$, skal man skrive programstumpen, hvor $\mathbf{A}\ \mathbf{A}^T$ kaldes 'a1'.

```
proc iml;
v = {1.000   0.546   0.244   0.149,
     0.546   1.000   0.398   0.182,
     0.436   0.398   1.000   0.271,
     0.149   0.182   0.271   1.000};
a = {0.7  0.5  0.4  0.3,
     0.3  0.4 -0.5 -0.7};
at = t(a);
a1 = a*t(a);
print a1;
run;
quit;
```

SAS ouput'et fra dette program bliver.

```
       A1
     0.99 -1.36E-20
-1.36E-20      0.99
```

Bemærk igen, at SAS har en tendens til ikke at afrunde fornuftigt, så man får 0.99 i stedet for 1 og -1.36 opløftet til 10^{-20} i stedet for 0.

Man kan også udføre en egenværdidekomponering i SAS. Det sker ved hjælp af kaldet EIGEN. Input til kaldet af EIGEN skal være en være en symmetrisk matrix, her kaldet 'v'. Output er en søjlevektor bestående af egenværdierne i aftagende størrelsesorden, som vi betegner 'lambda', og en matrix, 'e', af samme dimension som 'v', der har egenvektorerne som søjler. Programmet for en egenværdidekomponering af 'v' kan se sådan ud, hvor også størrelsen 'b = e*diag(lambda)*t(e)' beregnes for at kontrollere om SAS udfører sine beregninger rigtigt. Her er 'diag(lambda') en matrixoperation, der laver lambda om til en diagonalmatrix med egenværdierne i diagonalen og 0'er uden for diagonalen.

```
proc iml;
v = {1.000   0.546   0.244   0.149,
     0.546   1.000   0.398   0.182,
     0.436   0.398   1.000   0.271,
     0.149   0.182   0.271   1.000};
call eigen(lambda,e,v);
print lambda e;
dl = diag(lambda);
b=e*diag(lambda)*t(e);
print v b;
run;
```

Output fra dette program ser sådan ud:

```
   LAMBDA          E
1.9295776 0.5333939  -0.44712 0.3970508 0.5982689
 0.946497 0.5927437 -0.302554 -0.012485 -0.746298
0.7014503  0.494584 0.2932501 -0.766266 0.2867543
0.4224751 0.3457348 0.7890181 0.5050067 -0.053723


        V                                       B
        1     0.546     0.244     0.149         1     0.546     0.244     0.149
    0.546         1     0.398     0.182     0.546         1     0.398     0.182
    0.244     0.398         1     0.271     0.244     0.398         1     0.271
    0.149     0.182     0.271         1     0.149     0.182     0.271         1
```

En sammenligning af matricerne 'v' og 'b'= 'e*diag(lambda)*t(e)', der skal være ens, hvis SAS har regnet rigtigt, godkender beregningerne.

4. Transformationer af vektorer af stokastiske variable.

Lad **X** være søjlevektoren dannet af de stokastiske variable X_1, ... ,X_n og betragt transformationen

$$\mathbf{Y} = \mathbf{A}\ \mathbf{X}\ ,$$

hvor **A** er en n x n matrix af konstanter, så **Y** bliver en søjlevektor af dimension n dannet af de stokastiske variable Y_1, ... ,Y_n.

Kovariansmatricen $\mathbf{V}_Y$ for **Y** bliver da

$$\mathbf{V}_Y = \mathbf{A}\ \mathbf{V}_X\ \mathbf{A}^T\ ,$$

hvor $\mathbf{V}_X$ er kovariansmatricen for **X.**

Dette variansresultat gælder også, såfremt vektoren af Y-variable er kortere end vektoren af X-variable. Hvis **Y** er en søjlevektor af dimension m < n, hvor n er dimensionen af søjlevektoren **X**, og transformationsmatricen **A** har dimension m x n, skal ligningen

$$\mathbf{Y} = \mathbf{A}\ \mathbf{X}$$

derfor opfattes som en transformation, hvor der er færre Y'er efter, end der var X'er før transformationen. Altså en projektion på et lavere Euclidisk rum. Men det gælder stadig, at kovariansmatricen $\mathbf{V}_Y$ for **Y** er givet ved

$$\mathbf{V}_Y = \mathbf{A}\ \mathbf{V}_X\ \mathbf{A}^T\ ,$$

hvor $\mathbf{V}_X$ er kovariansmatricen for **X**, idet **A** og $\mathbf{A}^T$ bringer dimensionen af **X** 'ned fra' n til m.

Disse resultater, og resultaterne fra Kapitel 2, kan kombineres til:

(1) En transformation **Y** = **A X** kaldes ortogonal, hvis $\mathbf{A}^T$ er ortogonal, dvs. hvis rækkerne i **A** er ortogonale. Hvis således **X** af dimension n transformeres til **Y** af dimension m < n, har **A** dimensionen m x n. Ortogonaliteten betyder, at

$$\sum_{j=1}^{n} a_{ij}^2 = 1 \quad , \quad \text{for alle } i$$

og

$$\sum_{k=1}^{n} a_{ik} a_{jk} = 0 \quad , \quad \text{for alle } i \text{ og } j\ .$$

(2) Der findes en ortogonal transformation $\mathbf{Y} = \mathbf{A}\ \mathbf{X}$ af $\mathbf{X}$, der gør Y'erne uafhængige.

Betragt nemlig egenværdiopspaltningen

$$\mathbf{V}_X = \mathbf{A}^T\ \Lambda\ \mathbf{A}.$$

Ganger vi foran og bagved med $\mathbf{A}$ og $\mathbf{A}^T$, får vi

$$\mathbf{V}_Y = \Lambda = \mathbf{A}\ \mathbf{V}_X\ \mathbf{A}^T.$$

Det betyder, at hvis vi transformerer $\mathbf{X}$ som $\mathbf{Y} = \mathbf{A} \cdot \mathbf{X}$, kan X'erne transformeres til uafhængige variable via den ortogonale transformation $\mathbf{A}$. Desuden fremgår det, varianserne for Y'erne er egenværdierne for $\mathbf{V}_X$, og at transformationen består i en matrix $\mathbf{A}$ med egenvektorerne for $\mathbf{V}_X$ som rækker.

Disse simple resultater fra matrixalgebraen har haft enorm betydning for den matematisk statistik. Allerede tidligt vurderede ledende statistikere egenværdi-dekomponeringer af kovariansmatricer for centrale statistiske principper. Begreber som 'kanonisk korrelation', 'principalkomponentmetoden' og først og fremmest 'faktoranalyse' er udløbere af denne simple algebraiske teknik. Senest er den "skole" inden for området kategoriserede data, der kaldes deres metoder **korrespondance analyse,** endnu en anvendelse af den for matematikere så gamle, velkendte dekomponering af symmetriske, positivt definite matricer i deres komponenter.

5. Hotellings T^2 - test.

Det simple t-test for middelværdien i en enkelt stikprøve kan generaliseres til et test for middelværdierne i p korrelerede stikprøver.

For p = 1, dvs. for én stikprøve, har vi n observationer x_1, ... ,x_n, hvor de tilsvarende stokastiske variable X_1, ... ,X_n antages at være uafhængige, ensfordelte med middelværdi μ og varians σ^2. For at teste hypotesen

$$H_0 : \mu = \mu_0 ,$$

når variansen σ^2 er ukendt, benyttes teststørrelsen

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s} ,$$

hvor

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Denne teststørrelse er, som bekendt, t-fordelt med n - 1 frihedsgrader, og man kalder testet for et t-test.

For p > 1 stikprøver, kan observationerne opstilles i observationsmatricen

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \dots & \dots & \dots \\ x_{p1} & \dots & x_{pn} \end{bmatrix}$$

Det antages nu, at observationerne i den i'te række er uafhængige, ensfordelte med fælles middelværdi μ_i og varians σ_i^2. Observationerne i den i'te række og den j'te række er derimod korrelerede med korrelationskoefficienten σ_{ij}.

Under hypotesen

$$H_0 : \mu_i = \mu_{i0} ,$$

kan de hypotetiske middelværdier samles i vektoren

$$\mu_0 = \begin{bmatrix} \mu_{10} \\ \dots \\ \mu_{p0} \end{bmatrix}$$

således at $\boldsymbol{\mu}_0$ er en søjlevektor af længden p. Estimatet for $\boldsymbol{\mu}_0$ er søjlevektoren

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \dots \\ \bar{x}_p \end{bmatrix} ,$$

mens kovariansmatricen

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & \sigma_{1p} \\ \dots & \dots & \dots \\ \sigma_{p1} & \dots & \sigma_p^2 \end{bmatrix}$$

estimeres ved

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \dots & s_{1p} \\ \dots & \dots & \dots \\ s_{p1} & \dots & s_p^2 \end{bmatrix} .$$

Her er

$$s_i^2 = \frac{1}{n-1} \sum_{k=1}^{n} \left(x_{ik} - \bar{x}_i \right)^2$$

og

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \left(x_{ik} - \bar{x}_i \right)\left(x_{jk} - \bar{x}_j \right)$$

Hotelling foreslog, at generalisere t-teststørrelsen for én stikprøve til teststørrelsen

$$T^2 = n\left(\bar{\mathbf{x}} - \mu_0\right)^T \mathbf{S}^{-1}\left(\bar{\mathbf{x}} - \mu_0\right) .$$

Hvis man derfor ved hjælp af SAS har beregnet alle gennemsnittene for de p stikprøver og den tilsvarende empiriske pxp kovariansmatrix **S**, kan man i IML beregne Hotellings T^2 - teststørrelse.

Fordelingen af Hotellings T^2 - teststørrelse kan, viste Hotelling, karakteriseres ved at

$$\frac{(n-1)T^2}{p(n-p)} \sim F(p, n-p)$$

Dette resultat svarer til for én stikprøve, dvs. p = 1, at

$$\frac{(n-1)t^2}{1(n-1)} = t^2 \sim F(1, n-1)$$

I SAS proceduren IML er det uhyre nemt at beregne Hotellings teststørrelse og den forbundne F-størrelse. Lad os betragte de variable V27 og v29 til V32 fra ISSP-databasen 1987 for de 377 mænd fra UK, der har fuldtidsarbejde. De variable drejer sig om forskellige forhold, de skal bedømme betydningen af for deres daglige abejde. Alle 5 spørgsmål har samme svarmuligheder, som vi scorer 1 til 5. Vi skal teste den hypotese, at middelværdien for alle 5 variable er den gennemsnitlige score 3. Her bliver vektoren $\boldsymbol{\mu}_0$ altså lig med

$$\mu_0 = \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix}.$$

Hvis vi kalder denne vektor ’x0’, mens ’xbar og ’x’ er vektorerne bestående af differenserne mellem gennemsnittene og de hypotetiske middelværdiergennemsnit, bliver SAS-programmet til beregning af T^2 og F.

```
proc iml;
d ={  0.791523224      0.078580902      0.162353970      0.138509510      0.315854450,
      0.078580902      0.753414414      0.356432361      0.332221062      0.245358090,
      0.162353970      0.356432361      0.807494780      0.577212314      0.313575823,
      0.138509510      0.332221062      0.577212314      0.864354648      0.289914781,
      0.315854450      0.245358090      0.313575823      0.289914781      1.002596083};
x = {2.5411,
    2.2228 ,
    2.4164 ,
    2.3820 ,
    2.8541};
x0 = {3,
      3,
      3,
      3,
      3};
xbar = x - x0;
di = inv(d);
print d xbar di;
t2 = 377*t(xbar)*di*xbar;
f = (377-1)*t2/5/(377-5);
print t2 f;
run;
quit;
```

Endelig er 'd' den empiriske kovariansmatrix. Der er 377 observationer, så n = 377.

SAS-udskriften fra dette program bliver:

```
        D                                                       XBAR
0.7915232 0.0785809  0.162354 0.1385095 0.3158545   -0.4589
0.0785809 0.7534144 0.3564324 0.3322211 0.2453581   -0.7772
 0.162354 0.3564324 0.8074948 0.5772123 0.3135758   -0.5836
0.1385095 0.3322211 0.5772123 0.8643546 0.2899148    -0.618
0.3158545 0.2453581 0.3135758 0.2899148 1.0025961   -0.1459


       DI
550.15738 23.392082 -53.00132 -8.044186 -160.1413
23.392082 659.73529 -191.7469 -102.7288 -79.14462
-53.00132 -191.7469 987.87406 -548.4014 -86.77112
-8.044186 -102.7288 -548.4014 859.72943 -49.40846
-160.1413 -79.14462 -86.77112 -49.40846 487.26869


       T2         F
441.99675  89.34988
```

Da 95% fraktilen for en F-fordeling med (5,372) frihedsgrader er lig med 2.25, er dette resultat endog meget signifikant.

Hvis vi ændrer de 5 hypotetiske værdier af middelværdierne til 2.5, bliver SAS-programmet:

```
proc iml;
d ={  0.791523224     0.078580902     0.162353970     0.138509510     0.315854450,
      0.078580902     0.753414414     0.356432361     0.332221062     0.245358090,
      0.162353970     0.356432361     0.807494780     0.577212314     0.313575823,
      0.138509510     0.332221062     0.577212314     0.864354648     0.289914781,
      0.315854450     0.245358090     0.313575823     0.289914781     1.002596083};
x = {2.5411,
    2.2228 ,
    2.4164 ,
    2.3820 ,
    2.8541};
x0 = {2.5,
      2.5,
      2.5,
      2.5,
      2.5};
xbar = x - x0;
di = inv(d);
t2 = 377*t(xbar)*di*xbar;
print d xbar di;
f = (377-1)*t2/5/(377-5);
print t2 f;
run;
quit;
```

SAS output fra dette program bliver:

```
        D                                                       XBAR
0.7915232 0.0785809  0.162354 0.1385095 0.3158545    0.0411
0.0785809 0.7534144 0.3564324 0.3322211 0.2453581   -0.2772
 0.162354 0.3564324 0.8074948 0.5772123 0.3135758   -0.0836
0.1385095 0.3322211 0.5772123 0.8643546 0.2899148    -0.118
0.3158545 0.2453581 0.3135758 0.2899148 1.0025961    0.3541

       DI
550.15738 23.392082 -53.00132 -8.044186 -160.1413
23.392082 659.73529 -191.7469 -102.7288 -79.14462
-53.00132 -191.7469 987.87406 -548.4014 -86.77112
-8.044186 -102.7288 -548.4014 859.72943 -49.40846
-160.1413 -79.14462 -86.77112 -49.40846 487.26869

       T2         F
125.21961 25.313211
```

F-teststørrelsen er stadig meget signifikant i forhold til 95% fraktilen, der er 2.25 for en F-fordeling med (5,372) frihedsgrader. Selv hvis vi vælger 99% fraktilen for en F-fordeling med (5,372) frihedsgrader, svarende til en signifikanssandsynlighed mindre end 0.005, får vi 3.09, så resultatet er stadig meget signifikant.

6. Principalkomponentmetoden.

6.1 Principalkomponentmetoden.

Principalkomponentmetoden, som den benyttes i økonometrien for at løse bl.a. multikollinearitetsproblemer, er måske det bedste eksempel både på egenværdidekomponeringens enkelthed og dens styrke som analyseredskab.

Antag, at man i en regressionsanalyse med mange forklarende variable, som af natur er stokastiske, er løbet ind i et tilfælde af rigtig sur multikollinearitet. Hvis antal forklarende variable er n, er problemet jo, at kovariansmatricen mellem de forklarende variable - X'erne - er singulær eller næsten singulær. Her ville det være af betydning for analysen, hvis man kunne finde et antal forklarende variable: Y_1, ... ,Y_k, hvor k < n, der dels afhang lineært af X'erne, dels ikke var korrelerede, så var multikollinearitetsproblemet jo bragt ud af verden. Det man her søger er altså en lineær transformation

$$\mathbf{Y} = \mathbf{A}\,\mathbf{X}\ ,$$

hvor **A** er en matrix af dimension k x n, med k < n, og hvor Y'erne er ukorrellerede. Åbenbart er problemet, ifølge Kapitel 2, løst ved at vælge **A** som k rækker af egenvektorer for $\mathbf{V}_X$ = var[**X**]. Y'ernes varianser bliver da egenværdierne svarende til de valgte egenvektorer. Så ønsker man, at Y'erne har så stor variation som muligt, skal man vælge egenvektorerne svarende til de største egenværdier. Denne procedure kaldes principalkomponentmetoden.

Principalkomponentmetoden kan også motiveres ved følgende resultater, der viser at metoden succesivt giver de normerede linearkombinationer af X'erne, der har størst mulig varians og dermed forklarer mest muligt af variationen i lineære forudsigelser baseret på X'erne. Det første resultat er:

Koefficienterne b_{11}, ... ,b_{n1}, der opfylder

$$\max_{\sum_i b_{i1}^2 = 1} \left(\operatorname{var}\left[\sum_{i=1}^{n} b_{i1} X_i \right] \right)$$

dvs. maksimaliserer variansen på $\mathbf{b}_1^{\,T}\,\mathbf{X}$ fås netop, når søjlevektoren

$$\mathbf{b_1} = \begin{bmatrix} b_{11} \\ \dots \\ b_{n1} \end{bmatrix}$$

er lig med egenvektoren $\mathbf{a}_1$ for $\mathbf{V}_X$, svarende til den største egenværdi.

Det andet resultat er, at givet $\mathbf{b}_2$ er normeret til længden 1 og ortogonal med $\mathbf{b}_1$, dvs.

$$\sum_{i=1}^{n} b_{i2}^2 = 1$$

og

$$\sum_{i=1}^{n} b_{i1} b_{i2} = 0 \ ,$$

da opnås

$$\max\left(\text{var}\left[\sum_{i=1}^{n} b_{i2} X_i \right] \right)$$

netop når $\mathbf{b}_2$ er lig med egenvektoren $\mathbf{a}_2$ for $\mathbf{V}_X$, svarende til den næststørste egenværdi.

Sådan fortsættes med, at givet $\mathbf{b}_3$ er normeret til længden 1 og ortogonal med både $\mathbf{b}_1$ og $\mathbf{b}_2$, da opnås

$$\max\left(\text{var}\left[\sum_{i=1}^{n} b_{i3} X_i \right] \right)$$

når $\mathbf{b}_3$ er lig med egenvektoren $\mathbf{a}_3$ for $\mathbf{V}_X$, svarende til den trediestørste egenværdi. Etc.

Principalkomponentmetoden kan tillige motiveres som en mindste kvadraters løsning. Det gælder nemlig, at hvis vi forsøger for fastkoldt værdi af k at minimalisere Q, givet ved

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(s_{ij} - \sum_{q=1}^{k} b_{iq} b_{jq} \lambda_q \right)^2 \quad , \quad s_{ii} = s_i^2 \ ,$$

da er $\lambda_1, \ldots, \lambda_k$ de k største egenværdier, og $(b_{1q}, \ldots, b_{nq})$ den til λ_q svarende egenværdi. Dette resultat betyder simpelthen, at man får egenværdi-dekomponeringen, hvis man minimaliserer kvadratsumsafvigelsen mellem elementerne i estimatet for $\mathbf{V}_X$ og elementerne i $\mathbf{B}\ \Lambda\ \mathbf{B}^T$.

6.2 Forklaret variation. Kommunaliteter.

Da **A** som rækker har de k egenvektorer for $\mathbf{V}_X$, der svarer til de k største egenværdierne for $\mathbf{V}_X$, og disse k egenværdier netop er varianserne for de ukorrelerede Y'er, dvs. diagonalelementerne i diagonalmatricen $\mathbf{V}_Y$, gælder

$$\mathbf{V}_X \simeq \mathbf{V}_X^* = \mathbf{A}^T\ \mathbf{V}_Y\ \mathbf{A} \ .$$

Hvis k=n, dvs. vi medtager alle egenvektorer og egenværdier, vil naturligvis gælde et lighedstegn, dvs.

$$\mathbf{V}_X = \mathbf{A}^T \mathbf{V}_Y \mathbf{A} .$$

Med tr(**M**) betegner man sporet (Engelsk 'trace'), dvs. summen $m_{11} + ...+ m_{nn}$ af diagonalelementerne i **M**. Hvis man transformerer ortogonalt, gælder at sporet af kovariansmatricerne er ens. Det vil sige

$$tr(\mathbf{V}_X^*) = tr(\mathbf{V}_Y)$$

og hvis alle egenvektorer medtages

$$tr(\mathbf{V}_X) = tr(\mathbf{V}_Y) .$$

Man kan derfor altid bedømme, hvor tæt man er på at have forklaret hele variationen ved at summere de egenværdier, der indgår i egenværdi-dekomponeringen. Diagonalelementerne i approksimationen $\mathbf{V}_X^*$ til $\mathbf{V}_X$ kaldes kommunaliteterne. Hvis $\mathbf{V}_X$ er en korrelationsmatrix, er kommunaliteterne alle mindre end 1, og deres indbyrdes størrelsesorden viser hvilke X'er, der kommer nærmest til at forklare hele variationen.

6.3 Rotation.

Lad **x** og **y** være søjlevektorer af dimension n. En ortogonal transformation af fuld rang, dvs.

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} ,$$

hvor **A** er en n x n matrix med ortogonale rækker, er en rotation i det n-dimensionale rum. Det betyder, at hvis **x** er koordianterne til et punkt, er **y** koordianterne til det samme punkt, men i et koordinatsystem, der er drejet vinkelret i forhold til det gamle.

For n = 2, tager det sig sådan ud: Punktet med koordinaterne (x_1,x_2) føres ved transformationen

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos v & \sin v \\ -\sin v & \cos v \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

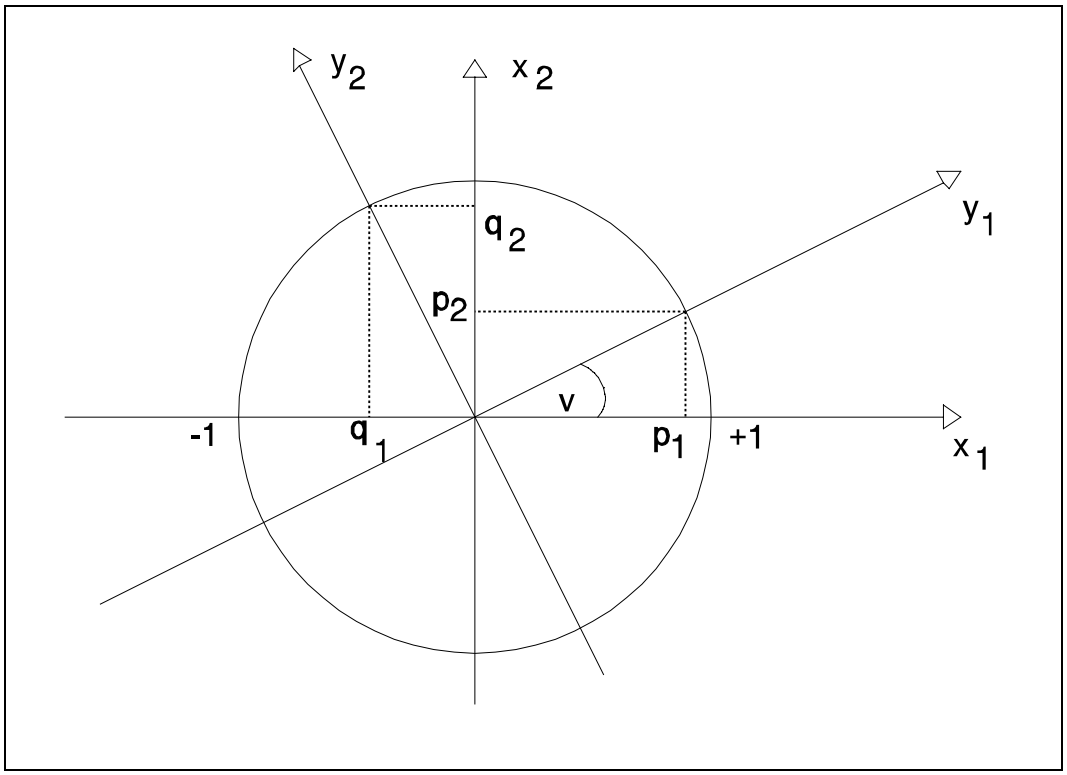
over i punktet (y_1,y_2). Denne transformation er ortogonal, idet de to rækkevektorer har længder 1 ifølge 'idiotformlen'

$$\cos^2 v + \sin^2 v = 1 ,$$

og produktsummen af deres komponenter er:

$$-\cos v \cdot \sin v + \sin v \cdot \cos v = 0 ,$$

På figuren neden for er vist, hvordan de to punkter (p_1,p_2) og (q_1,q_2) beliggende på enhedscirklen transformeres over i punkterne $(y_1,y_2) = (1,0)$ og $(0,1)$.



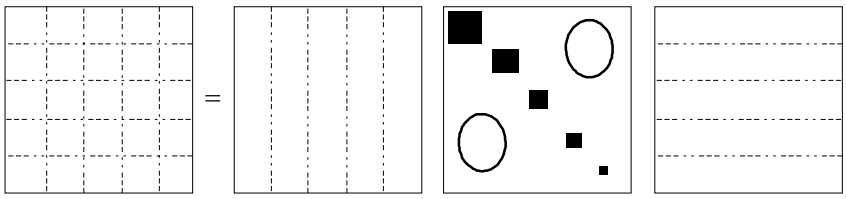
Som figuren ret tydeligt viser, kan man altid dreje et koordinatsystem, eller som vi her skal sige 'rotere koordinatsystemet', så specielle punkter ligger særlig attraktivt, f.eks. på akserne i det nye koordinatsystem.

Ved principalkomponentmetoden - og senere i faktoranalysen - skal vi benytte os af sådanne ortogonale transformationer.

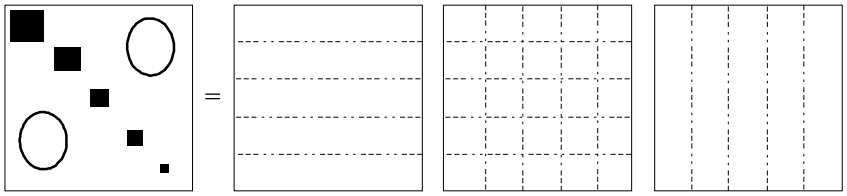
Principalkomponentmetoden sigter mod at gøre de nye variable ukorrelerede. Men det betyder ikke nødvendigvis, at de nye variable er dem, der bedst forklarer afhængigheden mellem de oprindelige variable (X'erne) og de nye (Y'erne). **Varimax-rotation** er en metode, der roterer Y'ernes kooordinatsystem (der, som vi erindrede, normalt er af lavere dimension end X'ernes koordiantsystem), så der er flest mulige af de oprindelige X'er, der korrelerer højt med Y'erne.

Antag, for at 'skære det ud i pap', at vi skal analysere V24 - V32 fra ISSP 1989. Her er 9 V'er, og vi kun ønsker 2 Y'er. Lad os derfor antage, at vi kan rotere Y-koordinatsystemet, så de første 4 V'er korrelerer med en korrelationskoefficient ret tæt ved 1 med Y_1, mens de sidste 5 V'er korrelerer med en korrelationskoefficient ret tæt ved 1 med Y_2. De to principale komponenter Y_1 og Y_2 er derfor et udtryk for, at en meget stor del af variationen i X'erne kan forklares ved, at de variable V24, ... ,V27 varierer tæt omkring en y_1 -akse, mens de variable V28, ... ,V32 varierer tæt omkring en y_2 - akse. Eksemplet i næste kapitel illustrerer dette forhold udmærket.

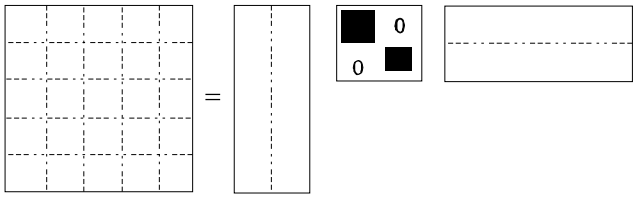
Man kan anskueliggøre matrixoperationerne i dette Kapitel på følgende måde:



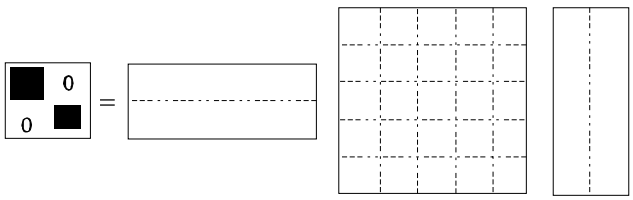
$$\mathbf{A} = \mathbf{B}\Lambda\mathbf{B}^T$$



$$\Lambda = \mathbf{B}^T\mathbf{A}\mathbf{B}$$



$$\mathbf{A} = \mathbf{B}\Lambda\mathbf{B}^T$$

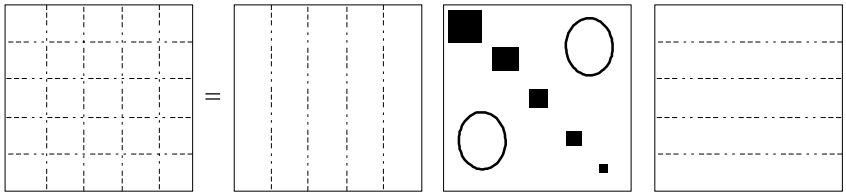


$$\Lambda = \mathbf{B}^T\mathbf{A}\mathbf{B}$$

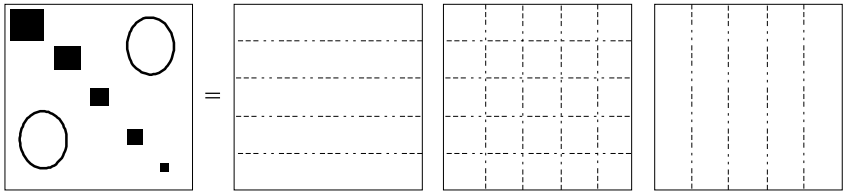
I tilfældet med kovariansmatricer $\mathbf{V}_Y$ for $\mathbf{Y}$ og $\mathbf{V}_X$ for $\mathbf{X}$ ved transformationen

$$\mathbf{Y} = \mathbf{AX}$$

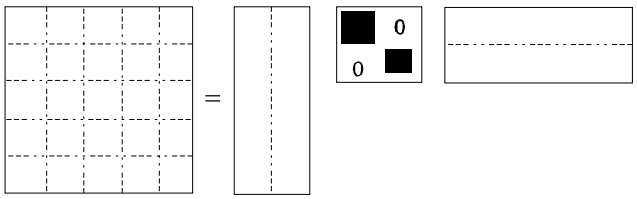
bliver billederne



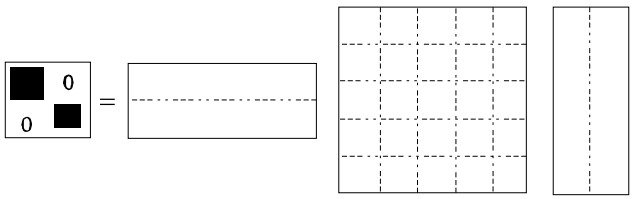
$$\mathbf{V}_X = \mathbf{A}^T\Lambda\mathbf{A}$$



$$\mathbf{V}_Y = \Lambda = \mathbf{A}\mathbf{V}_X\mathbf{A}^T$$



$$\mathbf{V}_X \simeq \mathbf{A}^T\Lambda\mathbf{A}$$



$$\mathbf{V}_Y \simeq \Lambda = \mathbf{A}\mathbf{V}_X\mathbf{A}^T$$

7. Principalkomponentmetoden i SAS.

Til illustration af principalkomponetmetoden i SAS betragter vi de variable V24 til V32 fra ISSP 1989, som beskriver hvilke forhold ved jobbet, respondenterne selv anser for betydningsfulde. De enkelte spørgsmål og de mulige svarkategorier fremgår af udsnittet fra kodebogen for 1989-undersøgelsen, der er vist i de to næste bokse.

V24 IMPORTANT: JOB SECURITY

Q.6a How important is: Job security?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

V25 IMPORTANT: HIGH INCOME

Q.6b How important is: High income?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

V26 IMPORTANT: ADVANCEMENT

Q.6c How important is: Good opportunities for advancement?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all
8. Can't choose, don't know
9. NA

V27 IMPORTANT: LEISURE TIME

Q.6d How important is: A job that leaves a lot of leisure time?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

V28 IMPORTANT: INTERESTG JOB

Q.6e How important is: An interesting job?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

V29 IMP.: INDEPENDENT WORK

Q.6f How important is: A job that allows someone to work independently?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

V30 IMP.: JOB HELPS PEOPLE

Q.6g How important is: A job that allows someone to help other people?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

V31 IMP.: JOB USEFUL F. SOC.

Q.6h How important is: A job that is useful to society?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

V32 IMP.:FLEXIBLE WRKG HOURS

Q.6i How important is: A job with flexible working hours?

1. Very important
2. Important
3. Neither important nor unimportant
4. Not important
5. Not important at all

For at udføre en principalkomponentanalyse, inklusive en VARIMAX-rotation, skal man skrive følgende program:

```
1. libname eba 'S:\eba-data';
2. title 'Principal component analysis on ISSP 1989. All countries';
3. data a;
4. set eba.d89um;
5. keep v24-v32;
6. data b;
7. set a;
8. if v24>5 then delete;
9. if v25>5 then delete;
10. if v26>5 then delete;
11. if v27>5 then delete;
12. if v28>5 then delete;
13. if v29>5 then delete;
14. if v30>5 then delete;
15. if v31>5 then delete;
16. if v32>5 then delete;
17. proc factor data=b
18.   method = prin
19.   priors = one
20.   mineigen = 1
21.   flag = 0.5;
22. run;
23. quit;
```

Da det er første gang vi viser et eksempel på en større analyse med udgangspunkt i ISSP databasen, skal vi gennemgå SAS-programmet i passende dele. Den første del består af programlinie 1 til 5. De ser sådan ud:

```
1. libname eba 'S:\eba-data';
2. title 'Principal component analysis on ISSP 1989. All countries';
3. data a;
4. set eba.d89um;
5. keep v24-v32;
```

Linie 1 knytter programmet til det område på S-drevet, hvor ISSP databasen ligger. Linie 2 sikrer, at alle sider i output får en overskrift, der fortæller, hvilket dataset og analyseform det handler om. I linie 3 og 4 oprettes dataset 'a', som i første omgang sættes til alle de variable i 1989-delen af ISSP databasen. Linie 5 sørger herefter for, at kun de de 9 variable V24 til V32 kommer med i dataset 'a'.

Linie 6 til 16 ser sådan ud.

```
6. data b;
7. set a;
8. if v24>5 then delete;
9. if v25>5 then delete;
10. if v26>5 then delete;
11. if v27>5 then delete;
12. if v28>5 then delete;
13. if v29>5 then delete;
14. if v30>5 then delete;
15. if v31>5 then delete;
16. if v32>5 then delete;
```

Her sættes i første omgang dataset ’b’ lig med de 9 variable i dataset ’a’. Men herefter fjernes alle personer, der har ubesvarede spørgsmål. De 5 svarmuligheder er kodet 1 til 5, som vist oven for, mens ubesvarede er kodet 8 eller 9. Dataset ’b’ består derfor nu af svarene på V24 til V32 for alle de personer, der har besvaret alle de 9 spørgsmål med en af de 5 givne svarkategorier.

Den første eksplorative analyse foregår ved hjælp af PROC FACTOR. Denne del består af linie 17 til 21, hvor programmet ser sådan ud:

```
17. proc factor data=b
18.   method = prin
19.   priors = one
20.   mineigen = 1
21.   flag = 1;
22. run;
23. quit;
```

Inden det første semikolon skal vi angive en række ’options’ for at sikre, at vi får den analyse, vi ønsker:

’data = b’ : Analysen udføres på dataset ’b’.
’method = prin’ : Der udføres en principal komponent analyse.
’mineigen = 1’: SAS udskriver egenvektorerne svarende til egenværdier med værdier over 1.
’flag = 0.5’: Egenvektorer angives ganget med 100 og rundet af, samtidig med at værdier større end 0.5 forsynes med en ’*’.
(På engelsk ’to be flaged’.)

Herefter sættes et semikolon og der skrives ’run’ og ’quit’, så programmet bliver kørt.

De vigtigste dele af output fra en SAS kørsel af dette program ser således ud: (Når jeg skriver 'de vigtigste dele', betyder det, at SAS output'et viser en del flere resultatet, som du i første omgang ikke skal bekymre dig om!)

```
Initial Factor Method: Principal Components

                        1           2           3           4           5
Eigenvalue          2.6688      1.3339      1.1094      0.9521      0.7761

                        6           7           8           9
Eigenvalue          0.6705      0.5970      0.5325      0.3599

3 factors will be retained by the MINEIGEN criterion.

Factor Pattern

             FACTOR1    FACTOR2    FACTOR3

V24            38         43        -45        IMPORTANT: JOB SECURITY
V25            41         70 *        2        IMPORTANT: HIGH INCOME
V26            54 *       44        -23        IMPORTANT: ADVANCEMENT
V27            40         21         66 *      IMPORTANT: LEISURE TIME
V28            59 *       -6         -6        IMPORTANT: INTERESTG JOB
V29            62 *      -17         15        IMP.: INDEPENDENT WORK
V30            68 *      -45        -22        IMP.: JOB HELPS PEOPLE
V31            66 *      -42        -25        IMP.: JOB USEFUL F. SOC.
V32            52 *       -8         53 *      IMP.:FLEXIBLE WRKG HOURS

NOTE: Printed values are multiplied by 100 and rounded to the nearest integer.
      Values greater than 0.5 have been flagged by an '*'.

Variance explained by each factor

   FACTOR1    FACTOR2    FACTOR3
  2.668801   1.333858   1.109400
```

Som det fremgår, er der 3 egenværdier større end 1, nemlig 2.67, 1.33 og 1.11. Egenvektorerne til disse tre egenværdier kaldes her FACTOR1, FACTOR2 og FACTOR3, så FACTOR1 svarer til den største egenværdi 2.67, osv. I output'et noteres også, at antal egenvektorer (faktorer) er baseret på, at den mindste medtagne egenværdi (MINEIGEN) er 1. Det nævnes også, (NOTE:) at egenvektorer angives ganget med 100 og rundet af, samtidig med, at værdier større end 0.5 forsynes med en '*'. Sidste linie er blot de 3 største egenværdier gentaget.

Der er to grunde til, at vi nu vælger kun at medtage 2 faktorer. For det første er den tredie største egenværdi ganske tæt ved 1. For det andet ser billedet lidt rodet ud for FACTOR2 og FACTOR3, hvad angår hvilke variable, der bidrager til forklaringen af variationen i kovariansmatricen. Også efter en VARIMAX-rotation, der ikke er vist her, ser det ikke tilfredsstillende ud.

Hvis der kun skal medtages 2 egenvektorer, kan vi godt bruge MINEIGEN optionen ved at skrive 'MINEIGEN = 1.2'. Men det virker lidt kunstigt. I stedet kan man direkte forlange, at der medtages 2 faktorer ved at skrive 'NFACT = 2'. For at se hvilke egenvektorer, der bidrager mest til at forklare variationen i de variable, må vi rotere koordinatsystemet ved en

ortogonal rotation. Benyttes VARIMAX metoden, skal der stå 'ROTATE = VARIMAX'. Ellers er programmet det samme, så det kommer til at se sådan ud:

```
1. libname eba 'S:\eba-data';
2. title 'Principal component analysis on ISSP 1989. All countries';
3. data a;
4. set eba.d89um;
5. keep v24-v32;
6. data b;
7. set a;
8. if v24>5 then delete;
9. if v25>5 then delete;
10. if v26>5 then delete;
11. if v27>5 then delete;
12. if v28>5 then delete;
13. if v29>5 then delete;
14. if v30>5 then delete;
15. if v31>5 then delete;
16. if v32>5 then delete;
17. proc factor data=b
18.   method = prin
19.   priors = one
20.   nfact = 2
21.   rotate = varimax
22.   flag = 0.5;
23. run;
24. quit;
```

Optionen 'priors = one' er rent teknisk, idet den bruges for at starte varimax rotationen.

De vigtigste dele af output fra dette program ser således ud:

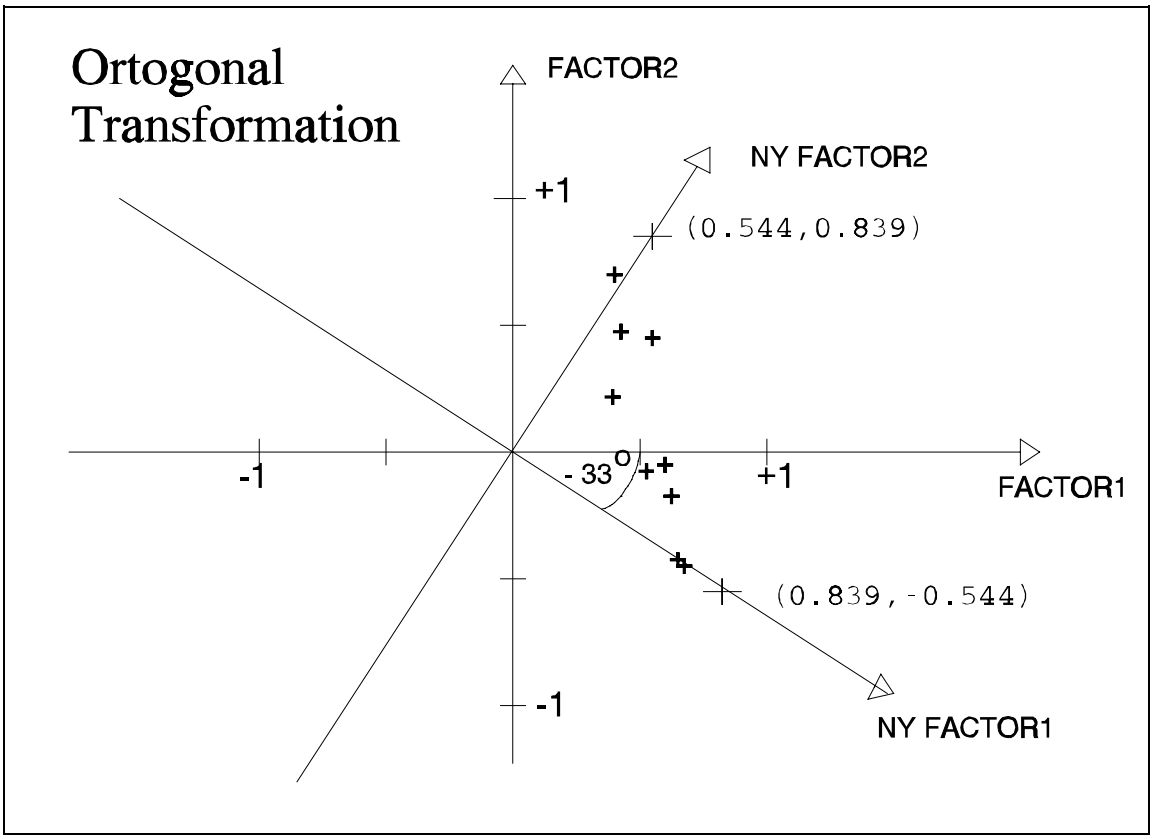
```
2 factors will be retained by the NFACTOR criterion.

Factor Pattern

           FACTOR1   FACTOR2

V24          38        43       IMPORTANT: JOB SECURITY
V25          41        70 *     IMPORTANT: HIGH INCOME
V26          54 *      44       IMPORTANT: ADVANCEMENT
V27          40        21       IMPORTANT: LEISURE TIME
V28          59 *      -6       IMPORTANT: INTERESTG JOB
V29          62 *     -17       IMP.: INDEPENDENT WORK
V30          68 *     -45       IMP.: JOB HELPS PEOPLE
V31          66 *     -42       IMP.: JOB USEFUL F. SOC.
V32          52 *      -8       IMP.:FLEXIBLE WRKG HOURS

Rotation Method: Varimax

Orthogonal Transformation Matrix

              1         2

    1      0.83882   0.54441
    2     -0.54441   0.83882

Rotated Factor Pattern

           FACTOR1   FACTOR2

V24           8        56 *     IMPORTANT: JOB SECURITY
V25          -4        82 *     IMPORTANT: HIGH INCOME
V26          21        66 *     IMPORTANT: ADVANCEMENT
V27          22        40       IMPORTANT: LEISURE TIME
V28          53 *      27       IMPORTANT: INTERESTG JOB
V29          62 *      19       IMP.: INDEPENDENT WORK
V30          82 *      -1       IMP.: JOB HELPS PEOPLE
V31          78 *       1       IMP.: JOB USEFUL F. SOC.
V32          48        22       IMP.:FLEXIBLE WRKG HOURS
```

Af den første linie output fremgår, at kun 2 faktorer vises ifølge NFACTOR kriteriet ('nfact = 2'). I de næste linie vises egenvektorerne, der altid i SAS kaldes 'FACTOR1', FACTOR2, etc. Herefter noteres det, at man roterer med VARIMAX metoden, og den ortogonale rotations matrix vises. Det fremgår, at den vinkel v, koordinalsystemet drejes er bestemt af cos v = 0.839 og sin v = - 0.544. Det svarer til en vinkel på v = - 33^0.

Her skal man lægge mærke til, at SAS definerer vektorer som matricer af dimension 1 x n, dvs. rækkevektorer, mens jeg i Kapitel 6, som sædvane er, opfatter vektorer som matricer af orden n x 1, dvs. som søjlevektorer. Det betyder, at den ortogonale transformation, der udskrives af SAS under overskriften 'Orthogonal Transformation Marix' er:

$$\begin{bmatrix} \cos v & -\sin v \\ \sin v & \cos v \end{bmatrix}.$$

På grafisk form ser rotationen sådan ud:



Efter rotationen vises igen de to egenvektorer: FACTOR1 og FACTOR2.

I den sidste del vises faktorerne efter VARIMAX rotationen. Det fremgår ved at betragte de faktorer, der er forsynet med en '*', at V28, V29, V30 og V31 korrelerer højt med FACTOR1. Fortolkningen af denne faktor er oplagt: Der er tale om aspekter ved job'et, der gør det interessant og betydningsfuldt. Omvendt synes FACTOR2 at korrellere højt med variable, der har at gøre med, hvad job'et gør for dig selv, bl.a. høj løn og gode avancementsmuligheder. Om job'et er betydningsfuldt, fordi det giver dig megen fritid, tenderer mod FACTOR2, lige som dét, at have et job med fleksible arbejdstider tenderer mod at betyde mest for FACTOR1.

Vi kan altså konstatere, at principal komponent metoden her deler de variable op i to 'naturlige grupper', som man måske kunne kalde 'Interessant Job' og 'Godt Job'.

I forbindelse med stianalyserne i Kapitel 14 og 15, skal vi bruge endnu en sammensat variabel fra 1989 undersøgelsen. De variable V68 til V74 drejer sig om forhold vedrørende arbejdsbetingelserne ved jobbet, som respondenten selv oplever dem, både fysiske og psykiske. Udsnittet, vist i boksen neden for, beskriver spørgsmålene og svarkategorierne. (De 5 svarkategorier er kun vist ved V68, da de er ens for alle 7 spørgsmål.)

V68 HOME FROM WORK EXHAUSTED

Q.17 Now some more questions about your working conditions. Please tick one box for each item below to show how often it applies to your work.

Q.17a How often do you come home from work exhausted?

1. Always
2. Often
3. Sometimes
4. Hardly ever
5. Never

V69 HARD PHYSICAL WORK

Q.17b How often do you have to do hard physical work?

V70 STRESSFUL WORK

Q.17b How often do you have to do stressful work?

V71 BORED AT WORK

Q.17d How often are you bored at work?

V72 WRK IN DANGRS CONDITIONS

Q.17e How often do you work in dangerous conditions?

V73 UNHEALTHY CONDITIONS

Q.17f How often do you work in unhealthy conditions?

V74 PHYS. UNPLEASANT COND.

Q.17g How often do you work in physically unpleasant conditions?

En principalkomponentanalyse af disse 7 spørgsmål giver følgende egenvektorer, med NFACT lig med 2, og efter en VARIMAX rotation.

```
Rotated Factor Pattern

           FACTOR1    FACTOR2

V68          10          85 *      HOME FROM WORK EXHAUSTED
V69          62 *        30        HARD PHYSICAL WORK
V70           7          80 *      STRESSFUL WORK
V71          30          -2        BORED AT WORK
V72          80 *        11        WRK IN DANGRS CONDITIONS
V73          81 *        11        UNHEALTHY CONDITIONS
V74          73 *         6        PHYS. UNPLEASANT COND.
```

FACTOR1, som korrelerer højt med V69, V72, V73 og V74, har åbenbart at gøre med jobbets fysiske belastning, mens FACTOR2 har at gøre med jobbets psykiske belastning. Generelt har man erfaring for, at der skal mindst tre variable, helst flere, før man kan tale om en egentlig faktor. Så fra de 7 variable, der vedrører arbejdsmiljøet, kan vi kun bruge FACTOR1, som man kan kalde ’Hard Job’ eller ’Dangerous Job’.

8. Om normeringer.

I PROC FACTOR sker der nogle lidt andre normeringer end i mine noter. Det, der vises som FACTOR1, FACTOR2, etc., er således ikke egenvektorerne, der jo er normerede til længden 1, dvs. at kvadratsummen af elementerne er 1. I PROC FACTOR er der normeret så længden er kvadratroden af egenværdien, dvs. kvadratsummen af elementerne er lig med egenværdien. Lidt senere står der "Standardized Scoring Coefficients", som til gengæld har længden 1 divideret med kvadratroden af egenværdien, dvs. kvadratsummen af elementerne er lig med 1 divideret med egenværdien. Det er de sidste, SAS skriver som FACTOR1, FACTOR2, etc. i output-sættet, dvs. med ordren "out=d".

Der er den fordel ved Standardized Scoring Coefficients som faktorer, at de også er uafhængige efter en ortogonal transformation, hvad egenvektorerne ikke nødvendigvis er. Det ser man således:

Hvis transformationen er $\mathbf{Y} = \mathbf{AX}$, og $\mathbf{A}$ er den ortogonale matrix, der optræder i egenværdi-dekomponeringen

$$\mathbf{V}_X = \mathbf{A}^T\Lambda\mathbf{A} ,$$

har vi

$$\mathbf{V}_Y = \mathbf{A}\mathbf{V}_X\mathbf{A}^T = \Lambda \text{ (diagonal)} .$$

Men definerer vi transformationen vi som $\mathbf{Z} = \Lambda^{-½}\mathbf{AX}$, hvor $\mathbf{A}$ er den samme ortogonale matrix, som oven for, da gælder

$$\mathbf{V}_Z = \Lambda^{-½}\mathbf{A}\mathbf{V}_X\mathbf{A}^T\Lambda^{-½} = \Lambda^{-½}\mathbf{A}\mathbf{A}^T\Lambda\mathbf{A}\mathbf{A}^T\Lambda^{-½} = \mathbf{I} .$$

idet $\mathbf{A}\mathbf{A}^T = \mathbf{I}$, når $\mathbf{A}$ er ortogonal. Hvis vi nu transformerer med en ny ortogonal matrix $\mathbf{B}$, dvs. $\mathbf{Q} = \mathbf{BZ}$, får vi

$$\mathbf{V}_Q = \mathbf{B}\mathbf{I}\mathbf{B}^T = \mathbf{I} ,$$

så Q'erne, ligesom Z'erne er uafhængige, men nu alle har varians 1.

9. Asymptotiske resultater for ML-estimatorer og kvotienttesstørrelser.

Resultaterne i dette afsnit bygger på **den centrale grænseværdisætning.** Den siger, at hvis Y_1, ... ,Y_n er uafhængige og ensfordelte stokastiske variable med endelig middelværdi μ og endelig varians σ^2, da gælder for store værdier af n

$$\sum_{i=1}^{n} Y_i \sim N(n\mu, n\sigma^2) .$$

Som sådan er resultatet ret meningsløst, idet både middelværdi og varians går mod uendelig. Men det skal opfattes sådan, at hvis summen af Y'er standardiseres, så nærmer fordelingen sig til den standardiserede normalfordeling, dvs. med

$$S_Y = \sum_{i=1}^{n} Y_i ,$$

så vil

$$\frac{S_Y - n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \sim N(0,1) ,$$

ML-estimatoren fås som løsningen til

$$\frac{\partial \ln L(\hat{\theta})}{\partial \theta} = 0 .$$

Hvis der er tale om n uafhængige, ensfordelte stokastiske variable X_i, i=1, ... ,n, hvor

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta) ,$$

fås ML-estimatorerne af ligningen

(9.1)
$$\frac{\partial \ln L(\hat{\theta})}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(X_i|\hat{\theta})}{\partial \theta} = \sum_{i=1}^{n} \frac{f'(X_i|\hat{\theta})}{f(X_i|\hat{\theta})} = 0 .$$

Rækkeudvikles den partielt afledede af ln L mht. til θ i punktet θ_0, som er den sande værdi af θ, fås approksimationen

$$0 = \frac{\partial \ln L(\hat{\theta})}{\partial \theta} \simeq \frac{\partial \ln L(\theta_0)}{\partial \theta} + (\hat{\theta} - \theta_0)\frac{\partial^2 \ln L(\theta_0)}{\partial \theta^2} ,$$

så

$$(\hat{\theta} - \theta_0) \simeq \frac{\partial \ln L(\theta_0)}{\partial \theta} \cdot \left(- \frac{\partial^2 \ln L(\theta_0)}{\partial \theta^2} \right)^{-1} . \tag{9.2}$$

Men af (9.1) med

$$Y_i = \frac{\partial \ln f(X_i|\theta_0)}{\partial \theta} = \frac{f'(X_i|\theta_0)}{f(X_i|\theta_0)}$$

følger, at vi kan anvende den centrale grænseværdisætning. Middelværdi og varians af Y_i får man af identiteten

$$\int f(x|\theta_0)dx = 1 ,$$

hvoraf følger, at

$$\int f'(x|\theta_0)dx = \int f''(x|\theta_0)dx = 0 .$$

Først ser man, at

$$E[Y_i] = E\left[\frac{\partial \ln f(X_i|\theta_0)}{\partial \theta}\right] = \int \frac{f'(x|\theta_0)}{f(x|\theta_0)} \cdot f(x|\theta_0)dx = \int f'(x|\theta_0) = 0 ,$$

så $E[Y_i] = 0$. For at finde variansen på Y_i, betragtes den anden afledede af ln f(x|θ):

$$\frac{\partial^2 \ln f(x|\theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left(\frac{f'(x|\theta)}{f(x|\theta)}\right) = \frac{f''(x|\theta) \cdot f(x|\theta) - (f'(x|\theta)^2}{(f(x|\theta))^2}$$

$$= \frac{f''(x|\theta)}{f(x|\theta)} - \left(\frac{f'(x|\theta}{f(x|\theta)}\right)^2 .$$

Det betyder, at

$$E\left[\frac{\partial^2 \ln f(X_i|\theta)}{\partial\theta^2}\right] = \int f''(x|\theta)\,dx - \int\left(\frac{f'(x|\theta)}{f(x|\theta)}\right)^2 f(x|\theta)\,dx = 0 - \mathrm{var}\left[Y_i\right] .$$

Vi tillader os yderligere i (9.2) at erstatte det led, der inverteres, med sin middelværdi, der jo bliver

$$E\left[-\frac{\partial^2 \ln L(\theta_0)}{\partial\theta^2}\right] = \sum_{i=1}^{n} E\left[-\frac{\partial^2 \ln f(X_i|\theta)}{\partial\theta^2}\right] = n\cdot\mathrm{var}\left[Y_i\right] .$$

Den første faktor på højresiden i (9.2) er summen af Y'erne. Så en anvendelse af den centrale grænseværdisætning giver, at for store værdier af n har vi approximationen

$$\left(\hat\theta - \theta_0\right) \sim N\left(0\,,\ \frac{1}{\left(n\cdot\mathrm{var}\left[Y_i\right]\right)^2}\cdot n\cdot\mathrm{var}\left[Y_i\right]\right) = N\left(0, \frac{1}{n\,\sigma^2}\right),$$

hvor

$$\sigma^2 = \left(E\left[-\frac{\partial^2 \ln f(X_i|\theta_0)}{\partial\theta^2}\right]\right)^{-1} .$$

For store værdier af n er ML-estimatoren altså approksimativt normalfordelt med middelværdi θ_0, dvs. den sande værdi af θ, og en varians, der nemt beregnes ud fra fordelingen af X_i.

Kvotienttesstørrelsen for hypotesen $\theta = \theta_0$ er givet som

$$Z = -2\ln\frac{L(\theta_0)}{L(\hat\theta)} .$$

Rækkeudvikler man ln $L(\theta_0)$ omkring $\hat\theta$, får man

$$\ln L(\theta_0) \simeq \ln L(\hat\theta) + \left(\theta_0 - \hat\theta\right)\frac{\partial \ln L(\hat\theta)}{\partial\theta} + \frac{1}{2}\left(\theta_0 - \hat\theta\right)^2\left(\frac{\partial^2 \ln L(\hat\theta)}{\partial\theta^2}\right).$$

Da

$$\frac{\partial \ln L(\hat\theta)}{\partial\theta} = 0 \quad ,$$

får man derfor
For $\hat\theta$ tæt ved θ_0 gælder
Om kvotienttesstørrelsens asymptotiske fordeling gælder altså

$$Z = -2\ln L(\theta_0) + 2\ln L(\hat{\theta}) \simeq (\theta_0 - \hat{\theta})^2 \left(- \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta^2} \right).$$

$$- \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta^2} \simeq E\left[- \frac{\partial^2 \ln L(\theta_0)}{\partial \theta^2} \right] \simeq \frac{1}{\mathrm{var}[\hat{\theta}]} .$$

$$Z \simeq \left(\frac{\hat{\theta} - \theta_0}{\sqrt{\mathrm{var}[\hat{\theta}]}} \right)^2 \sim U^2 \sim \chi^2(1)$$

Også hvis θ er en vektor af længde m, og alle de m komponenter er specificerede under hypotesen, vil Z være χ^2-fordelt, men nu med m frihedsgrader.

Hvis hypotesen - i vektor tilfældet - består af k < m bånd mellem θ'erne, vil Z være χ^2-fordelt, men nu med k frihedsgrader. Det er dette tilfælde, man i statistikkens teori kalder sammensatte hypoteser (Engelsk: Composite hypotheses). Men man skal for sammensatte hypoteser være forsigtig med identifikationen af modellens parametre. Meget generelt kan man sige, at hvis de her viste resultater skal holde, skal der være tale om en reparametrisering af $(\theta_1, \ldots, \theta_m)$ over i $(\tau_1, \ldots, \tau_m)$, så hypotesen består i:

$$\tau_1 = \tau_{10}$$

$$\tau_2 = \tau_{20}$$

$$\ldots$$

$$\tau_k = \tau_{k0} ,$$

mens $\tau_{k+1}, \ldots, \tau_m$ ikke er specificerede under hypotesen.

Reparametriseringen kan godt være θ'erne selv, så hypotesen består i at specificere de k første komponenter af θ-vektoren.

Et typisk eksempel på en reparametrisering er hypotesen

$$\theta_1 + \ldots + \theta_m = 0.$$

Her skal man reparametrisere med

$$\tau_1 = \theta_1 + \ldots + \theta_m ,$$

og $\tau_j = \theta_j$, j = 2, ... , m. Hypotesen er herefter $\tau_1 = 0$, svarende til at der lægges ét bånd mellem θ'erne, og antal frihedsgrader er dermed lig med 1.

10. Faktoranalyse.

Faktoranalyse er egentlig blot en multipel regressionsanalyse med forklarende variable, der ikke er observerbare. Det vil sige, at man skal estimere både regressionskoefficienterne og de forklarende variable, udover restledsvariansen. Formelt set defineres faktoranalyse ved matrixligningen

$$\mathbf{X} = \mathbf{B}\ \mathbf{F} + \mathbf{E},$$

Her er **X** en søjlevektor af dimension n af stokastiske variable X_1, ... ,X_n, **B** er en n x k-dimensional matrix af ukendte konstanter, som man ofte kalder 'factor loadings', **F** er en søjlevektor af dimension k af ikke-observerbare stokastiske variable F_1, ... , F_k, mens **E** er en søjlevektor af stokastiske restled E_1, ... E_n. Vi skal senere kalde ikke-observerbare variable for latente variable.

Denne matrixligning kan i form af observerede variable blive skrevet fuldt ud som

$$x_i = \sum_{j=1}^{k} b_{ij} f_j + e_i$$

sådan, at x_i'erne (i = 1, ... ,n) er de observerede variable, f_j, j = 1, ... ,k er værdierne af ikke-observerbare (latente) faktorer og e_i'erne (i = 1, ... ,n) på sædvanlig vis er restled. I forhold til regressionsanalyse er forskellen, at de forklarende variable her er erstattet af ikke-observerbare 'konstruerede' variable f_1, ... ,f_k, mens elementerne i **B** er parametre, som de er i vanlig regresssionsanalyse.

Det antages at F'erne og E'erne er ukorrelerede. Normalt antages endvidere, at alle variable er centrerede, så de har middelværdi 0. I så fald bliver kovariansmatricen $\mathbf{V}_X$ for **X** lig med

$$\mathbf{V}_X = E\left[\ \mathbf{X}\ \mathbf{X}^T\ \right] = \mathbf{B}\ E\left[\ \mathbf{F}\ \mathbf{F}^T\ \right] \mathbf{B}^T + E\left[\ \mathbf{E}\ \mathbf{E}^T\ \right],$$

eller

$$\mathbf{V}_X = \mathbf{B}\ \mathbf{V}_F\ \mathbf{B}^T + \mathbf{V}_E\ .$$

Ofte analyseres korrelationsmatricen, dvs. kovariansmatricen omskaleret så matricen har 1-taller i diagonalen, i stedet for selv kovariansmatricen. Men for det følgende spiller dette ikke den store rolle. Forskellen på de to udgangspunkter for analysen, vender vi tilbage til.

Hvis man medtager lige så mange faktorer, som der er X'er, dvs. n = k, kan man nemt finde en løsning, hvor alle fejlleddene E har varians 0, og hvor faktorerene F_1, ... , F_k er ukorrelerede. Læsningen er, at $\mathbf{V}_F$ er en diagonalmatrix dannet af egenværdierne for $\mathbf{V}_X$ og **B** er en matrix med egenvektorerne for $\mathbf{V}_X$ som søjler, idet man blot, jvf. Kapitel 2, skal foretage en egenværdi-dekomponering af $\mathbf{V}_X$.

Hvis k < n er én løsning på faktoranalysen at danne $\mathbf{V}_F$ som en diagonalmatrix af de k største egenværdier, og vælge matricen **B** af factor loadings som dannet af de tilsvarende k egenvektorer som søjler.

I dette tilfælde bliver elementerne $\mathbf{V}_E$ således forskellene på diagonalelementerne i $\mathbf{V}_X$ og diagonalelementerne i $\mathbf{B}\,\mathbf{V}_F\,\mathbf{B}^T$. Som ved principalkomponentmetoden kaldes diagonalelementerne i $\mathbf{B}\,\mathbf{V}_F\,\mathbf{B}^T$ for kommunaliteterne, og beskriver hvor stor en andel af variationen faktormodellen bidrager til hver af X'erne.

Man kan også vælge nogle initialværdier for kommunaliteterne, dvs. diagonalelementerne i $\mathbf{B}\,\mathbf{V}_F\,\mathbf{B}^T$. De øvrige initialparametre fås de ved at 'regne baglæns' i beregningen oven for. Denne metode bruges generelt af SAS.

For at udlede ML-estimatorerne for parametrene i faktor modellen, skal man kan opstille log-likelihoodfunktionen for den empiriske kovariansmatrix $\mathbf{V}_X$ (eller den tilsvarende korrelationsmatrix), som en funktion af de ukendte parametre, og maksimalisere denne funktion.

Her er det vigtigt at bemærke 3 ting.

(1) Matricen $\mathbf{S}_X^2$ af dimension n x n, der som elementer har de empiriske varaianser og kovarianser mellem X'erne er en sufficient stikprøvefunktion for den teoretiske kovariansmatrix $\mathbf{V}_X$. Man kan derfor opstille log-likelihoodfunktionen for estimation af faktor-modellens parametre, ved at betragte den simultane tæthedsfunktion for elementerne i $\mathbf{S}_X^2$. Denne fordeling er kendt som Wishart-fordlingen, jvf. Kapitel 4.

(2) I Faktormodellen

$$\mathbf{V}_X = \mathbf{B}\,\mathbf{V}_F\,\mathbf{B}^T + \mathbf{V}_E$$

er alle elementerne på højresiden parametre, så vi kan godt skrive ligningen som

$$\mathbf{V}_X = \mathbf{V}_X(\theta_1, \ldots, \theta_m) ,$$

hvor $\theta_1, \ldots, \theta_m$ er modellens parametre. Det er med denne parametrisering, ML-estimationen kan udføres.

(3) Faktormodellen skal være identificerbar, dvs. der må ikke være flere parametre m, end der er datapunkter i den empiriske kovariansmatrix. Optællingen af m er meget lettere, end man skulle tro. Og sker iøvrigt også i PROC FACTOR. Vi får, når der medtages k egenværdier

(i) k egenværdier, dvs. varianser for de ukorrelerede faktorer.

(ii) antal frit varierende elementer i **B**, som er

$$k \cdot n - k - \frac{k(k-1)}{2} ,$$

idet der er $k \cdot n$ elementer i **B**, men det lægger k bånd på egenvektorerne, at de skal have længden 1 og k(k-1)/2 bånd på egenvektorerne, at produktsummen af enhver kombination dem skal være 0.

(iii) n varianser af fejlleddene.

Det giver ialt

$$m = k + (k \cdot n - k - \frac{k(k-1)}{2}) + n = n(k+1) - \frac{k(k-1)}{2} .$$

På den anden side er der

$$\frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}$$

frit varierende elementer i en kovariansmatrix $\mathbf{V}_X$ af dimension n x n.

I en korrelationsmatrix er der selvfølgelig n færre parametre end i en kovariansmatrix, idet alle diagonal leddene er lig med 1. Til gengæld skal summen af diagonalleddene i $\mathbf{B}\,\mathbf{V}_F\,\mathbf{B}^T$ og $\mathbf{V}_E$ være 1, så kriteriet for identificerbarhed er derfor det samme for de to tilfælde.

I begge tilfælde er modellen, som sagt, identificerbar, hvis der er flere 'datapunkter', dvs frit varierende elementer i kovarians/korrelations-matricen, end der er frit varierende parametre i modellen. Vælger vi at betagte tilfældet med en kovariansmatrix, bliver betingelsen

$$\frac{n(n+1)}{2} > n(k+1) - \frac{k(k-1)}{2} .$$

Hvis der er lighedstegn, er modellen lige akkurat identificerbar, så der er tale om en reparametrisering.

Antag for eksempel, at der er 6 variable, dvs. 6 · 7/2 = 21 datapunkter. Her kan man bestemme k = 2 faktorer, idet antal parametre for k = 2 er

$$6 \cdot 3 - 2 \cdot 1/2 = 18 - 1 = 17.$$

Men hvis vi forsøger os med k = 3, er der tale om en reparametrisering, idet kriteriet her bliver

$$6 \cdot 7/2 = 21 > 6 \cdot 4 - 3 \cdot 2/2 = 24 - 3 = 21 ,$$

Da der er lighedstegn, er modellen ganske vist identificerbar, men vi kan ikke teste modellens tilpasningsevne, da der jo bliver perfekt tilpasning mellem data og model.

Selve likelihoodligningerne kan ikke løses eksplicit. Man er nødt til at 'gætte' på nogle udgangsværdier for parametrene, og så gradvis forbedre parameterværdierne, indtil de partielt afledede af log-likelihoodligningen er 0, eller - hvad der kommer ud på det samme - indtil log-likelihood ligningen når sin maksimale værdi. Det viser sig, at de vigtigste parametre, som skal sættes til fornuftige udgangsværdier, er kommunaliterne, dvs. diagonalelementerne i $\mathbf{V}_F$.

Som ved principalkomponentmetoden kan man rotere det koordinatsystem, hvori faktorerne er repræsenteret. Faktormodellen er nemlig kun identificeret, hvis vi forlanger, at faktorerne er ukorrelerede. Dette krav giver ikke nødvendigvis den bedste fortolkning af faktorernes relative indflydelse på X'erne. Lad således **A** være en ortogonal matrix af dimension k x k, dvs. $\mathbf{A}\,\mathbf{A}^T = \mathbf{I}$. Hvis **F** transformeres over i $\mathbf{F}^*$ som $\mathbf{F}^* = \mathbf{A}^T\,\mathbf{F}$, eller $\mathbf{F} = \mathbf{A}\,\mathbf{F}^*$, får vi ifølge faktoranalysemodellen

$$\mathbf{X} = \mathbf{B}\,\mathbf{A}\,\mathbf{F}^* + \mathbf{E}\ ,$$

og dermed, når $\mathbf{B}^* = \mathbf{B}\,\mathbf{A}$

$$\mathbf{V}_X = \mathbf{B}^*\,\mathbf{V}_F{}^*\,\mathbf{B}^{*T} + \mathbf{V}_E\ .$$

Denne faktormodel er ækvivalent med den første, blot kræver vi ikke længere, at faktorerne F_1^*, ... , F_k^* er ukorrelerede.

Man kan også rotere F-koordiantsystemet, så de nye akser ikke er vinkelrette på hinanden, dvs. transformationsmatricen **A** ikke er ortogonal. Det kaldes på engelsk en 'oblique' (skæv) transformation. SAS-proceduren VARIMAX roterer ortogonalt, mens SAS-optionen CORMAX i PROC FACTOR roterer både ortogonalt og skævt.

11. Eksplorativ Faktor Analyse i SAS.

11.1. Regulære tilfælde.

Vi skal nu se, hvorledes man udfører en faktor analyse i SAS. Det er stadig PROC FACTOR, vi benytter, men med lidt andre options. SAS-programmet eksemplificeres stadig med de variable V24 - V32 fra ISSP 1989. I det første program benytter vi ML-estimation af parametrene og foretager en VARIMAX rotation, dvs. roterer koordinatsystemet ortogonalt. Programmet kan f.eks. se sådan ud:

```
1. libname eba 's:\eba-data';
2. title 'Factor analysis on ISSP 1989. All countries. ML-estimation. Varimax rotation';
3. data a;
4. set eba.d89um;
5. keep v24-v32;
6. data b;
7. set a;
8. if v24>5 then delete;
9. if v25>5 then delete;
10. if v26>5 then delete;
11. if v27>5 then delete;
12. if v28>5 then delete;
13. if v29>5 then delete;
14. if v30>5 then delete;
15. if v31>5 then delete;
16. if v32>5 then delete;
17. proc factor data=b
18.   method = ml
19.   priors = smc
20.   nfact = 2
21.   rotate = varimax
22.   flag = 0.5;
23. run;
24. quit;
```

Den første del af programmet, linie 1 - 16, er identisk med programmerne i Kapitel 7. Men i kaldet af PROC FACTOR er optionerne følgende:

(1) Option'en 'data = b', har samme betydning som i Kapitel 7.

(2) Optionen 'method = ml' i linie 18 betyder, at der benyttes ML-estimation. Denne estimationsmetode er iterativ, dvs. med udgangsværdi i nogle beregnede **udgangsværdier**, itererer SAS sig frem til estimerede parameterværdier, der tilfredsstiller likelihood ligningerne.

(3) Optionen 'prior = smc' i linie 19 betyder, at udgangsværdierne fastsættes ved, at udgangsværdierne for kommunaliterne sættes lig med de kvadrerede multiple korrelationskoefficienter mellem de enkelte variable og de øvrige 8 variable. Det har vist sig, at disse udgangsværdier,

bortset fra helt 'skæve' tilfælde, giver en hurtig konvergens til løsningerne af likelihood ligningerne.

(4) Optionen 'nfact = 2' i linie 20 betyder, som før, at antal ønskede faktorer sættes til 2.

(5) Optionen 'rotate = varimax' betyder, at koordinatsystemet roteres ortogonalt for af få så store korrelationer som muligt mellem de variable og den ene eller den anden faktor.

(6) 'flag = 0.5' har samme betydning, som i Kapitel 7, nemlig, at der sættes en '*' ved alle korrelationer større end 0.5. Samtidigt rundes faktorerne (automatisk) op til 100 gange deres størrelse, dvs. skal 'læses' som procenter.

Med optionen 'method = ml', kan man ikke i SAS analysere en kovariansmatrix, så i det følgende er det altid korrelationsmatricen, der analyseres.

De vigtigste dele af SAS-out'et fra dette program ser sådan ud:

```
Factor analysis on ISSP 1989. All countries. ML-estimation. Varimax rotation        Initial
Factor Method: Maximum Likelihood

Prior Communality Estimates: SMC

       V24        V25        V26        V27        V28        V29        V30        V31        V32
  0.117728   0.219241   0.220309   0.134460   0.222785   0.249457   0.439657   0.418069   0.176114
```

I den første boks vises den del af SAS output'et, hvor de valgte udgangsværdier for kommunaliteteterne, betegnet 'Prior Communality Estimates' er vist.

Den næste boks viser de egenværdier, man får ved en opspaltning af kovariansmatricen baseret på udgangsværdierne for kommunaliteterne.

```
                      1          2          3          4          5
Eigenvalue       2.8883     0.8526     0.3753     0.2472    -0.0674
Difference       2.0357     0.4773     0.1281     0.3146     0.1206
Proportion       0.9060     0.2674     0.1177     0.0775    -0.0211
Cumulative       0.9060     1.1735     1.2912     1.3687     1.3476

                      6          7          8          9
Eigenvalue      -0.1880    -0.2217    -0.3078    -0.3905
Difference       0.0337     0.0860     0.0827
Proportion      -0.0590    -0.0696    -0.0965    -0.1225
Cumulative       1.2886     1.2190     1.1225     1.0000

2 factors will be retained by the NFACTOR criterion.
```

De to første egenværdier er klart større end de øvrige, så optionen 'NFACT = 2' er ganske velvalgt.

I den sidste del af SAS output'et vises nogle nøgletal fra de nødvendige iterationer for at bestemme ML-estimaterne.

```
Iter Criterion    Ridge   Change   Communalities

  1    0.20755    0.000  0.24780   0.12181 0.40187 0.30749 0.11444 0.22337
                                    0.24893 0.68746 0.52223 0.16570

  2    0.20488    0.000  0.03526   0.13812 0.39694 0.32852 0.11435 0.20106
                                    0.22794 0.68968 0.55749 0.15675

  3    0.20479    0.000  0.00930   0.13872 0.40559 0.32702 0.11093 0.19964
                                    0.22405 0.69899 0.55337 0.15486

  4    0.20478    0.000  0.00282   0.13958 0.40693 0.32769 0.11111 0.19811
                                    0.22330 0.69617 0.55591 0.15442

  5    0.20478    0.000  0.00093   0.13975 0.40786 0.32762 0.11097 0.19793
                                    0.22295 0.69707 0.55509 0.15435

Convergence criterion satisfied.
```

Det som står under 'Criterion' er ikke selve værdien af log-likelihoodfunktioenen, men et numerisk mål (kriterie) for, hvor tæt man er ved maksimum for likelihood funktionen.

Kriteriet 'Criterion' for, hvor tæt man er ved likelihoodfunktionens maksimum er imidlertid tæt forbundet med selve værdien af likelihoodfunktion.

Tæthedsfunktionen for matricen S af empiriske varianser og kovarianser blev udledt af den amerikanske statistiker Wishart, og bærer derfor for hans navn. Wishart-fordelingens tæthedsfunktion kan skrives

$$f(S) = K \cdot \det(S)^{\frac{1}{2}(n-p-1)} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}S\right)\right) \cdot \det(\Sigma)^{-\frac{1}{2}n},$$

hvor K er en konstant, der er uafhængig af modellens parametre, n er antal variable, p er antal parametre, Σ er kovariansmatricen $\mathbf{V}_X$ med de sande værdier af varianserne og kovarianserne under modellen indsat og tr(A) (for en symmetrisk matrix) er summen af leddene i matricens diagonal, kaldet 'sporet' eller 'trace' på engelsk. Jeg har her benyttet betegnelsen det(A) for determinanten af en matrix A.

Beregner man transformationen -2 lnL af likelihoodfunktionen, hvor L = f(S), får man fra Wishart fordelingens tæthedsfunktion

$$-\frac{1}{2}\ln L = -\frac{1}{2}\ln f(S) = \mathrm{tr}\left(\Sigma^{-1}S\right) + \ln\left(\det(\Sigma)\right) - (n-p-1)\ln\left(\det(S)\right)$$

Den kriteriefunktion, som SAS benytter er

$$\text{Criterion} = \text{tr}\left(\Sigma^{-1}S\right) + \ln\left(\det(\Sigma)\right) - \ln\left(\det(S)\right) - n \ .$$

Der er altså tale om en 'omskalering' af -2 lnL, som har det samme minimum som -2 lnL. Faktoren n - p - 1 har at gøre med at ML-estimatoren for σ^2 i regressionsanalyse med p forklarende variable har nævneren n, mens den middelrette estimator har nævneren n - p - 1. Grunden til at algoritmen søger efter et minimum for 'Criterion' er, at logaritmen til likelihoodfunktionen er ganget med -2, så et maksimum bliver til et minimum.

Når SAS bruger en omskaleret værdi af -2 lnL er det fordi, som man let ser, at SAS-kriteriet 'Criterion' er lig med 0, hvis S = Σ, idet $\text{tr}(\Sigma^{-1}S)$ i så fald er lig med n.

'Change' viser hvor store ændringerne er i konvergenskriteriet. Desuden vises ændringerne i kommunaliteterne. Det fremgår, at kriteriet nogenlunde svarer til, at kommunaliteterne er bestemt med 3 korrekte decimaler.

I den næste boks vises den del af output, som vedrører testene for modellen og faktorværdierne før og efter en VARIMAX rotation.

Der testes først

$$H_0 : \text{Ingen fælles faktorer}$$

mod

$$H_A : \text{Mindste én fælles faktor.}$$

Her får man en transformeret kvotienttesstørrelse på z = 19953.2 med 36 frihedsgrader, som selvfølgelig er vildt signifikant. Nul-hypotesen er her, at alle de variable er uafhængige, og alternativet, at der en vis afhængighed, som kan forklares ved én eller flere faktorer.

Dette test har selvfølgelig ingen betydning i praksis. Desværre er det en svaghed ved SAS, at det alt for ofte udskriver disse 'nonsens test'.

Der testes dernæst

$$H_0 : \text{2 faktorer er tilstrækkeligt}$$

mod

$$H_A : \text{Flere faktor er nødvendige.}$$

Her bliver den transformerede kvotienttesstørrelse på z = 2729.3 med 19 frihedsgrader, som også signifikant, men graden af tilpasning er åbenbart blevet stærkt forbedret. Nul-hypotesen er her, at en faktor model med 2 faktorer beskriver data, mod at der skal flere faktorer til for at beskrive data med en faktor model.

```
Significance tests based on 13334 observations:

   Test of H0: No common factors.
        vs HA: At least one common factor.

   Chi-square = 19953.154   df = 36   Prob>chi**2 = 0.0001

   Test of H0: 2 Factors are sufficient.
        vs HA: More factors are needed.

   Chi-square = 2729.292   df = 19   Prob>chi**2 = 0.0001
```

Dernæst vises faktorværdierne svarende til de 2 faktorer med de største egenværdier, både før og efter en VARIMAX-rotation. Her svarer rotationen til cos v = 0.937 og sin v = - 0.349, eller v = - 20.4^0, jvf. Kapitel 7.

```
Factor Pattern

           FACTOR1   FACTOR2

V24          25        28       IMPORTANT: JOB SECURITY
V25          21        60 *     IMPORTANT: HIGH INCOME
V26          35        45       IMPORTANT: ADVANCEMENT
V27          22        25       IMPORTANT: LEISURE TIME
V28          41        18       IMPORTANT: INTERESTG JOB
V29          45        14       IMP.: INDEPENDENT WORK
V30          81 *     -21       IMP.: JOB HELPS PEOPLE
V31          73 *     -16       IMP.: JOB USEFUL F. SOC.
V32          37        14       IMP.:FLEXIBLE WRKG HOURS

Rotation Method: Varimax

Orthogonal Transformation Matrix

              1         2
    1      0.93731   0.34851
    2     -0.34851   0.93731

Rotated Factor Pattern

           FACTOR1   FACTOR2

V24          14        35       IMPORTANT: JOB SECURITY
V25          -1        64 *     IMPORTANT: HIGH INCOME
V26          17        55 *     IMPORTANT: ADVANCEMENT
V27          12        31       IMPORTANT: LEISURE TIME
V28          32        31       IMPORTANT: INTERESTG JOB
V29          38        29       IMP.: INDEPENDENT WORK
V30          83 *       8       IMP.: JOB HELPS PEOPLE
V31          74 *      11       IMP.: JOB USEFUL F. SOC.
V32          30        26       IMP.:FLEXIBLE WRKG HOURS
```

Lige som ved principal komponent metoden finder også faktor analysen frem til, at spørgsmålene om, hvad jobbet gør for dig: Især V25 og V26, men også til dels V24 og V27 vægter højt med faktoren FACTOR2. På den anden side vægter de spørgsmål, som udtrykker, hvad godt, du kan gøre ved hjælp af dit jobbet, nemlig V30 og V31, højt med FACTOR1. Til gengæld vægter V28, V29 og V32 nogenlunde lige meget på de to faktorer, så disse variable er svære at placere med kun to faktorer i modellen.

Læg mærke til, at vi får et lidt andet resultat, end vi fandt i Kapitel 7, hvor vi analyserede de samme spørgsmål.

Hvis vi i stedet for benytter PROMAX rotations metoden, skal optionerne for PROC FACTOR skrives om på følgende måde:

```
proc factor data=b
  method = ml
  priors = smc
  nfact = 2
  rotate = promax
  flag = 0.5;
run;
```

Her er 'rotate = varimax' byttet ud med 'rotate = promax'. Det giver en lidt længere udskrift for rotationsdelen af output. Den første boks viser igen værdierne af FACTOR1 og FACTOR2 efter VARIMAX rotationen.

```
Prerotation Method: Varimax

Rotated Factor Pattern

           FACTOR1   FACTOR2

V24          14         35       IMPORTANT: JOB SECURITY
V25          -1         64 *     IMPORTANT: HIGH INCOME
V26          17         55 *     IMPORTANT: ADVANCEMENT
V27          12         31       IMPORTANT: LEISURE TIME
V28          32         31       IMPORTANT: INTERESTG JOB
V29          38         29       IMP.: INDEPENDENT WORK
V30          83 *        8       IMP.: JOB HELPS PEOPLE
V31          74 *       11       IMP.: JOB USEFUL F. SOC.
V32          30         26       IMP.:FLEXIBLE WRKG HOURS
```

De 2 næste bokse viser en række tekniske mellemtrin vedr. de forskellige rotationer, f.eks. 'Procrustean Transformation Matrix' og ' Normalized Oblique Transformation Matrix', samt forskellige normaliseringer af faktorvægtene.

Rotation Method: Promax

Procrustean Transformation Matrix

	1	2
1	1.28687	-0.25041
2	-0.13815	1.71708

Normalized Oblique Transformation Matrix

	1	2
1	0.92388	0.21642
2	-0.46110	1.00961

Inter-factor Correlations

	FACTOR1	FACTOR2
FACTOR1	100 *	25
FACTOR2	25	100 *

Rotated Factor Pattern (Std Reg Coefs)

	FACTOR1	FACTOR2	
V24	10	34	IMPORTANT: JOB SECURITY
V25	-8	65 *	IMPORTANT: HIGH INCOME
V26	12	53 *	IMPORTANT: ADVANCEMENT
V27	9	30	IMPORTANT: LEISURE TIME
V28	29	27	IMPORTANT: INTERESTG JOB
V29	35	24	IMP.: INDEPENDENT WORK
V30	84 *	-4	IMP.: JOB HELPS PEOPLE
V31	74 *	0	IMP.: JOB USEFUL F. SOC.
V32	28	22	IMP.:FLEXIBLE WRKG HOURS

Reference Axis Correlations

	FACTOR1	FACTOR2
FACTOR1	100 *	-25
FACTOR2	-25	100 *

Rotation Method: Promax

Reference Structure (Semipartial Correlations)

	FACTOR1	FACTOR2	
V24	10	33	IMPORTANT: JOB SECURITY
V25	-8	63 *	IMPORTANT: HIGH INCOME
V26	11	51 *	IMPORTANT: ADVANCEMENT
V27	9	29	IMPORTANT: LEISURE TIME
V28	28	26	IMPORTANT: INTERESTG JOB
V29	34	23	IMP.: INDEPENDENT WORK
V30	82 *	-4	IMP.: JOB HELPS PEOPLE
V31	72 *	0	IMP.: JOB USEFUL F. SOC.
V32	27	21	IMP.:FLEXIBLE WRKG HOURS

```
Factor Structure (Correlations)

           FACTOR1   FACTOR2

V24          18        36       IMPORTANT: JOB SECURITY
V25           8        63 *     IMPORTANT: HIGH INCOME
V26          25        56 *     IMPORTANT: ADVANCEMENT
V27          17        32       IMPORTANT: LEISURE TIME
V28          36        34       IMPORTANT: INTERESTG JOB
V29          41        32       IMP.: INDEPENDENT WORK
V30          83 *      17       IMP.: JOB HELPS PEOPLE
V31          75 *      19       IMP.: JOB USEFUL F. SOC.
V32          33        29       IMP.:FLEXIBLE WRKG HOURS
```

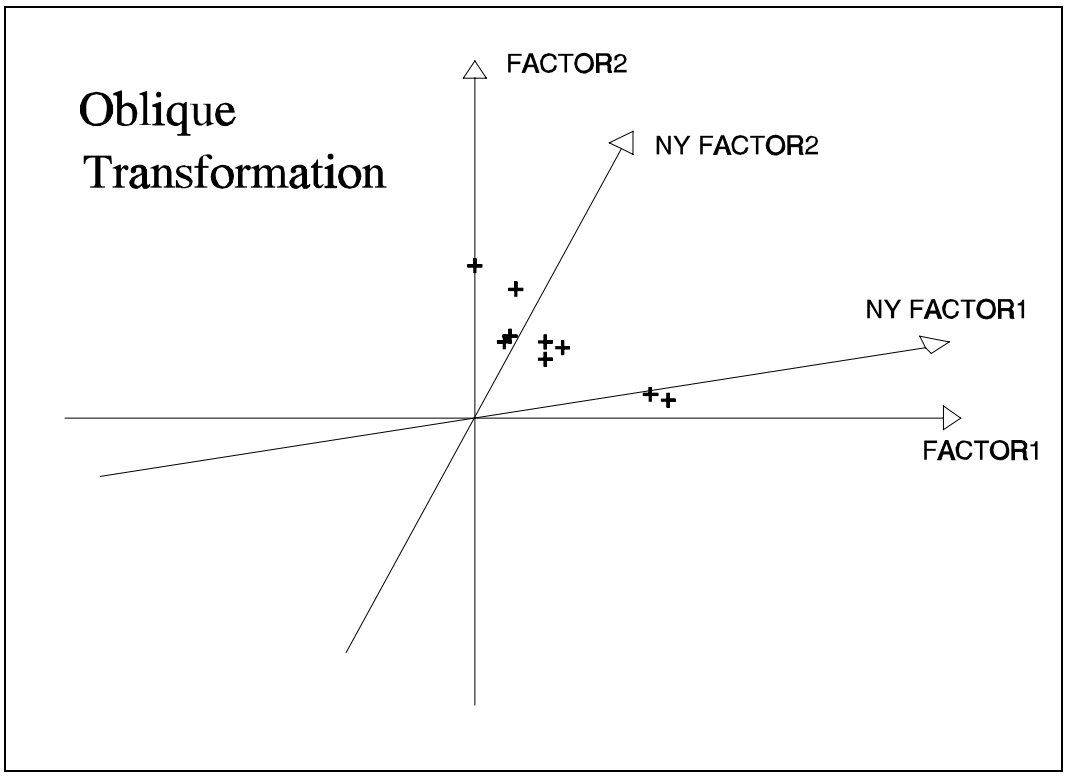
De roterede faktorvægte er dog nogenlunde ens, og de afsluttende faktorvægte 'Faktor Structure (Correlations) afviger ikke meget fra, hvad vi fandt ved VARIMAX-rotations-metoden. Den væsentligste forskel er mellem de egenværdier (FACTOR's), der er vist under overskriften 'Semipartial Correlations' og under overskriften 'Correlations'. Der er tale om en omnormering, hvor vi tager hensyn til variansen på de variable V24 til V32, idet egen-værdierne under 'Semipartial Correlations' er næsten identiske med de FACTOR-værdier, der er vist under 'Std Reg Coefs', dvs. standardiserede regressions koefficienter. Disse sidste er simpelthen de koefficienter SAS beregner, når der normeres, så variansen på FACTOR's er 1. De værdier, der vises under 'Correlations' er desuden normeret, så variansen på V'erne alle er 1. Derved bliver de viste værdier af FACTOR1 og FACTOR2 rent faktisk lig med korrelationskefficienterne mellem V'erne og de to faktorer.

Mest interessant er måske de nye værdier af FACTOR1 og FACTOR2 ved den skæve transformation (Normalized Oblique Transformation). Figuren neden for viser hvor de 9 punkter, svarende til V24 - V32, ligger i forhold til akserne (FACTOR1, FACTOR2) før den skæve transformation og efter den skæve transformation (NY FACTOR1, NY FACTOR2).

Man får de nye koordinater udfra den 2 x 2 matrix som står under 'Normalized Oblique Transformation Matrix' i SAS output'et, nemlig som

$$\text{NY FACTOR1} = 0.92388\ \text{FACTOR1} + 0.21642\ \text{FACTOR2}$$

$$\text{NY FACTOR2} = -0.46110\ \text{FACTOR1} + 1.00961\ \text{FACTOR2}$$



Som det fremgår af figuren, kan man få den endnu bedre samling omkring akserne, dvs. en højere korrelation mellem de observerede variable og faktorerne, ved at rotere til akser, der ikke står vinkelret (ortogonalt) på hinanden. Husk her, at der allerede (ved en VARIMAX rotation) er roteret ortogonalt, så de to akser FACTOR1 og FACTOR2 giver en så god vinkelret tilpasning til datapunkterne, som muligt. Så den endnu bedre tilpasning til akserne NY FACTOR1 og NY FACTOR2, skyldes alene, at vi tillader en ikke-ortogonal rotering.

Man kan også prøve at medtage 3 faktorer, så skal programmet med PROMAX-rotation skrives:

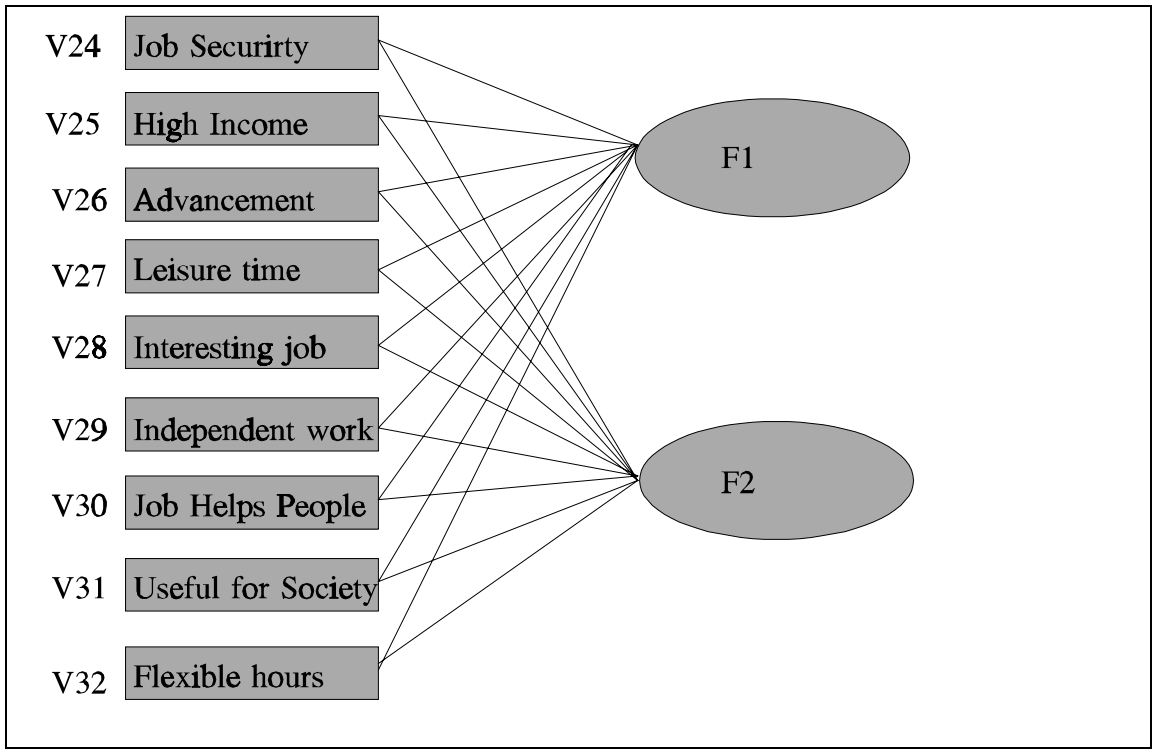
```
proc factor data=b
  method = ml
  priors = smc
  nfact = 3
  rotate = promax
  flag = 0.5;
run;
```

I SAS-output'et neden for er vist den del, der vedrører teststørrelserne. Der er sket en væsentlig forbedring af modeltilpasningen, men modellen er stadig ikke i stand til at beskrive data tilfredsstillende.

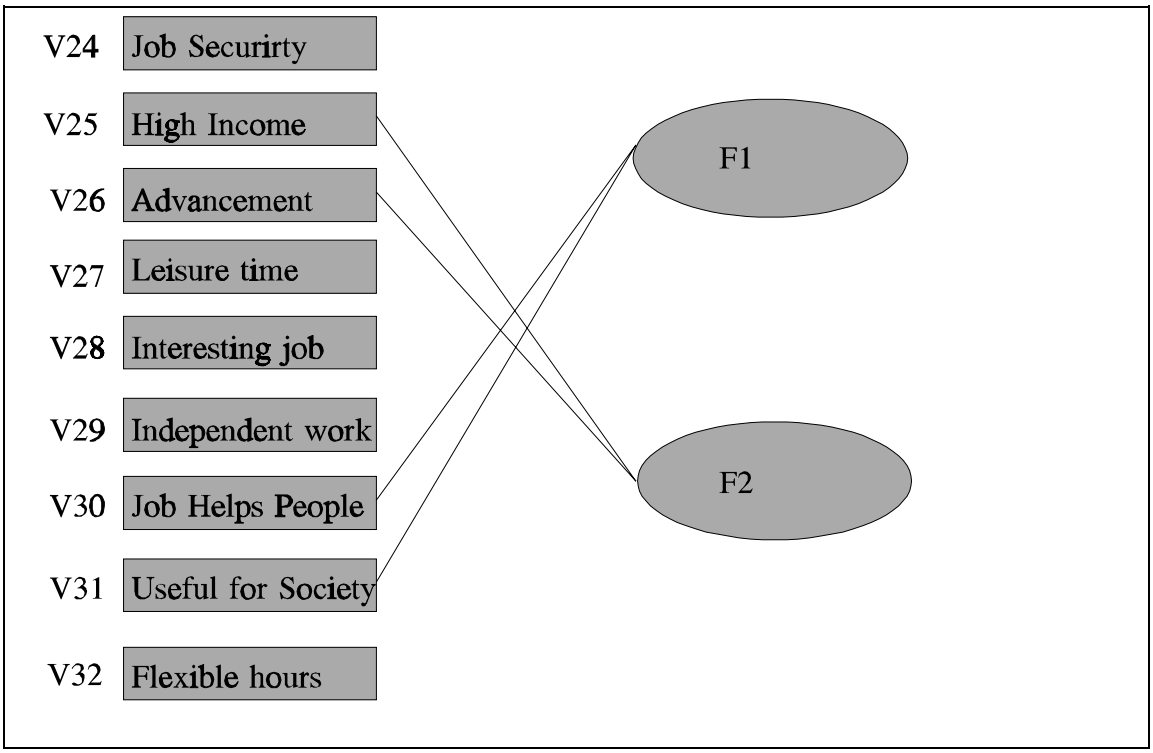
```
Significance tests based on 13334 observations:

   Test of H0: No common factors.
        vs HA: At least one common factor.

   Chi-square = 19953.154   df = 36   Prob>chi**2 = 0.0001

   Test of H0: 3 Factors are sufficient.
        vs HA: More factors are needed.

   Chi-square = 1134.183   df = 12   Prob>chi**2 = 0.0001
```

Af denne boks fremgår det, at medtagelse af tre faktorer udvælger 3 sæt á 2 variable (hvis V26 regnes til FACTOR2), der korrelerer højt med hver sin factor. V25 og V26, der beskriver de egoistiske aspekter af jobbet. V28 og V29 beskriver de interessante aspekter ved jobbet, mens V30 og V31 beskriver de altruistske aspekter.

Skal man beskrive resultatet af en faktor analyse af spørgsmål V24 til V32 grafisk, får man følgende graf, hvis vi betragter 2 faktorer og alt korrelerer med alt:

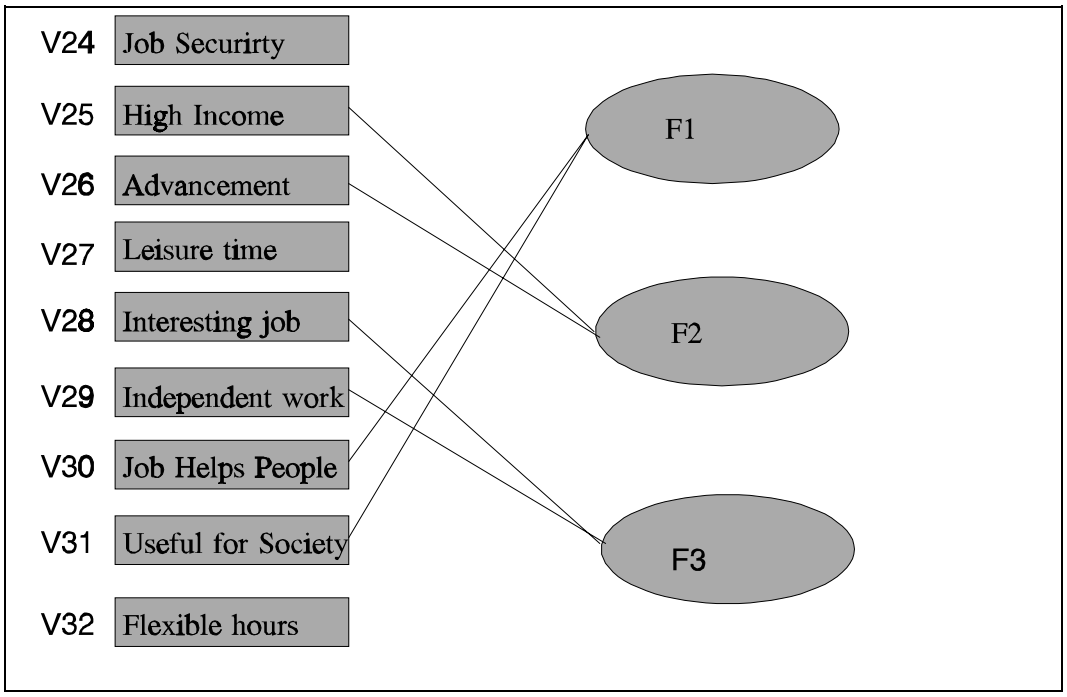


Hvis vi kun medtager de sammenhænge mellem de 9 observerede variable og de 2 signifikante faktorer, SAS-programmet peger på, dvs. hvor korrelationen er større end 0.5, får man det lidt mere enkle billede, vist i den næste graf.



På denne graf kan man se, at den latente variabel F1 har at gøre med, hvad jobbet kan gøre for dig, dvs. høj indkomst og gode avancementsforhold, mens den latente variable F2 beskriver hvad respondenten ønsker at opnå via sit job, nemlig at hjælpe andre mennesker og at værer nyttig for samfundet.

Hvis man medtager 3 faktorer, får man følgende graf, med den fortolkning, vi gav oven for:



11.2. Heywood tilfælde.

I visse ekstreme tilfælde, vil ML-estimaterne for faktormodellens parametre ikke ligge inden for de specificerede parameterområder. Sædvanligvis er det kommunaliteterne - og dermed varianserne på restleddene - der falder uden for parameterrummet. Disse tilfælde kaldes Heywood-tilfælde efter den amerikanske statistiker, der første gang obeserverede og studerede fænomenet. For teorien for ML-estimater betyder det, at nogle ML-estimater falder på randen af det parameterområde, modellen foreskriver. Hermed er alle resultater vedr. ML-estimaters og kvotienttests asymptotiske egenskaber ikke længere gyldige, så vi kan ikke vurdere en models tilpasningevne til data, eller standard fejlen på ML-estimaterne udfra de sædvanlige formler.

For at belyse, hvordan man i paraksis identificerer et Heywood-tilfælde, betragter vi variabel V40 til V48 fra ISSP-undersøgelsen i 1985. Disse spørgsmål drejer sig om, hvilke emner de adspurgte mener er væsentlige for deres skoleuddannelse. Et uddrag af kodebogen er vist i den næste boks.

V40 School: Read with math

Q.15 And now a few questions about education. Here are some things that might be taught i school. How important is it that schools teach each of theses to 15 years old?

Q.15a Importance of things taught in school: Reading, writing and mathematics.

1. Essential, must be taught
2. Very important
3. Fairly important
4. Not very important
5. Not needed, should not be taught

V41 School: Sex education

Q.15b Importance of things taught in school: Sex education.

V42 School: Respect authority.

Q.15c Importance of things taught in school: Respect authority.

V43 School: History, literature, art.

Q.15d Importance of things taught in school: History, literature and the arts.

V44 School: Good justment.

Q.15e Importance of things taught in school: Ability to make one's own judgements.

V45 School: Job training

Q.15f Importance of things taught in school: Job training.

V46 School: Technology

Q.15g Importance of things taught in school: Science and technology.

V47 School: Concerm for others

Q.15h Importance of things taught in school: Concern for minorities and the poor.

V48 School: Disciplin, order

Q.15i Importance of things taught in school: Disciplin and orderliness.

SAS-programmet for en faktoranalyse af disse 9 variable ser sådan ud:

```
libname eba 's:\eba-date';
title 'ISSP 1885 data faktoranalyse på variabel v40 til v48';
data a;
keep v40-v48;
data b;
set a;
if v40>5 then delete;
if v41>5 then delete;
if v42>5 then delete;
if v43>5 then delete;
if v44>5 then delete;
if v45>5 then delete;
if v46>5 then delete;
if v47>5 then delete;
if v48>5 then delete;
if v40=0 then delete;
if v41=0 then delete;
if v42=0 then delete;
if v43=0 then delete;
if v44=0 then delete;
if v45=0 then delete;
if v46=0 then delete;
if v47=0 then delete;
if v48=0 then delete;
proc factor data=b
   method = ml
   priors = smc
   nfact = 2
   rotate = varimax
   round
   flag = 0.5
;
var v40-v48;
run;
quit;
```

Det væsentlige SAS-output fra denne SAS-kørsel bliver

```
Initial Factor Method: Maximum Likelihood

Prior Communality Estimates: SMC

   v40         v41         v42         v43         v44         v45         v46
0.06336876  0.08598355  0.45092857  0.26429743  0.23169534  0.19495692  0.25860247
   v47         v48
0.27575655  0.47225748

Preliminary Eigenvalues: Total = 3.5103851  Average = 0.39004279

         Eigenvalue    Difference    Proportion    Cumulative

   1    3.13829840    2.03851395        0.8940        0.8940
   2    1.09978444    0.90726152        0.3133        1.2073
   3    0.19252292    0.06655921        0.0548        1.2621
   4    0.12596371    0.17095506        0.0359        1.2980
   5    -.04499135    0.04233448       -0.0128        1.2852
   6    -.08732583    0.09705310       -0.0249        1.2603
   7    -.18437893    0.11077581       -0.0525        1.2078
   8    -.29515475    0.13917876       -0.0841        1.1237
   9    -.43433351                     -0.1237        1.0000

2 factors will be retained by the NFACTOR criterion.

Iteration   Criterion   Ridge  Change Communalities
    1       0.1281866  0.0000  0.3884  0.05848 0.07951  0.44675  0.35650  0.31479  0.22508
0.34111  0.34220 0.86069
    2       0.1206839  0.0000  0.2189  0.05773  0.08570  0.38828  0.35598  0.31132  0.21675
.34198  0.34044  1.07958


ERROR: Communality greater than 1.0.
```

Som det fremgår er estimatetet for den sidste Kommunalatet over 1 så SAS stoppper her, idet vi har identficeret et Heywood-tilfælde.

I SAS har man nu to options. Den første består i at anvende optionen ’Heywood’, som vist i den næste program stump. Ved denne option begrænser man kommunaliteterne til højst at have værdien 1. Det svarer til, at vi søger et maximum for log-likelihood funktionen på randen af det tilladte parameterområde. I programmet, vist neden for, er således tilføjet ordet ’heywood’ blandt optionerne i kaldet af Proc Factor.

```
libname eba 's:\eba-date';
title 'ISSP 1885 data faktoranalyse på variabel v40 til v48';
data a;
keep v40-v48;
data b;
set a;
if v40>5 then delete;
if v41>5 then delete;
if v42>5 then delete;
if v43>5 then delete;
if v44>5 then delete;
if v45>5 then delete;
if v46>5 then delete;
if v47>5 then delete;
if v48>5 then delete;
if v40=0 then delete;
if v41=0 then delete;
if v42=0 then delete;
if v43=0 then delete;
if v44=0 then delete;
if v45=0 then delete;
if v46=0 then delete;
if v47=0 then delete;
if v48=0 then delete;
proc factor data=b
   method = ml
   priors = smc
   nfact = 2
   heywood
   rotate = varimax
   round
   flag = 0.5
;
var v40-v48;
run;
quit;
```

Dette giver SAS-outputtet, vist i de to næste bokse:

2 factors will be retained by the NFACTOR criterion.

Iteration	Criterion	Ridge	Change	Communalities					
1	0.1281866	0.0000	0.3884	0.05848	0.07951	0.44675	0.35650	0.31479	0.22508
				0.34111	0.34220	0.86069			
2	0.1224055	0.0000	0.1393	0.05800	0.08345	0.40953	0.35617	0.31258	0.21978
				0.34166	0.34108	1.00000			
3	0.1211278	0.0000	0.0292	0.05805	0.09099	0.43872	0.35649	0.31285	0.21757
				0.34031	0.34065	1.00000			
4	0.1211272	0.0000	0.0006	0.05823	0.09140	0.43877	0.35631	0.31314	0.21774
				0.34011	0.34002	1.00000			

Convergence criterion satisfied.

Initial Factor Method: Maximum Likelihood

Significance Tests Based on 5848 Observations

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	36	9390.9273	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	19	707.6051	<.0001
HA: More factors are needed			

Initial Factor Method: Maximum Likelihood

Factor Pattern

	Factor1	Factor2
v40	13	20
v41	3	30
v42	65 *	11
v43	13	58 *
v44	20	52 *
v45	28	37
v46	18	55 *
v47	29	51 *
v48	100 *	0

```
     Orthogonal Transformation Matrix

                         1               2

        1          0.10533         0.99444
        2          0.99444        -0.10533


    Rotated Factor Pattern

        Factor1      Factor2

v40          22           11
v41          30            0
v42          18           64 *
v43          59 *          6
v44          54 *         14
v45          40           24
v46          57 *         12
v47          53 *         23
v48          11           99 *
```

Vi kan også bruge optionen 'ultra', som vist i næste programstump. Her lægges ingen begrænsninger på værdien af kommunaliteterne. Det svarer til at bestemme maximum for log-likelihood funktionen, selv om maximum ligger uden for det tilladte parameter område. Det giver nogle mærkelige egenværdier og desuden én eller flere negative varianser på restleddene. Og dette er jo en hel del for viderekommende at fortolke.

```
libname eba 's:\eba-date';
title 'ISSP 1885 data faktoranalyse på variabel v40 til v48';
data a;
keep v40-v48;
data b;
set a;
if v40>5 then delete;
if v41>5 then delete;
if v42>5 then delete;
if v43>5 then delete;
if v44>5 then delete;
if v45>5 then delete;
if v46>5 then delete;
if v47>5 then delete;
if v48>5 then delete;
if v40=0 then delete;
if v41=0 then delete;
if v42=0 then delete;
if v43=0 then delete;
if v44=0 then delete;
if v45=0 then delete;
if v46=0 then delete;
if v47=0 then delete;
if v48=0 then delete;
proc factor data=b
   method = ml
   priors = smc
   nfact = 2
   ultra
   rotate = varimax
   round
   flag = 0.5
;
var v40-v48;
run;
quit;
```

Dette program giver følgende SAS-output.

2 factors will be retained by the NFACTOR criterion.

Iteration	Criterion	Ridge	Change	Communalities					
1	0.1281866	0.0000	0.3884	0.05848	0.07951	0.44675	0.35650	0.31479	0.22508
				0.34111	0.34220	0.86069			
2	0.1206839	0.0000	0.2189	0.05773	0.08570	0.38828	0.35598	0.31132	0.21675
				0.34198	0.34044	1.07958			
3	0.1178987	0.0000	0.2234	0.05778	0.08986	0.34156	0.34783	0.31445	0.21365
				0.33868	0.33982	1.30297			
4	0.1165473	0.0000	0.2471	0.05785	0.09187	0.30472	0.34073	0.31751	0.21367
				0.33429	0.34018	1.55007			
5	0.1158802	0.0000	0.2663	0.05784	0.09260	0.27713	0.33535	0.31977	0.21467
				0.33090	0.34062	1.81632			
6	0.1155707	0.0000	0.2680	0.05782	0.09276	0.25722	0.33163	0.32126	0.21583
				0.32850	0.34092	2.08428			
7	0.1154385	0.0000	0.2457	0.05780	0.09269	0.24331	0.32913	0.32220	0.21681
				0.32688	0.34111	2.32994			
8	0.1153867	0.0000	0.2024	0.05778	0.09256	0.23393	0.32748	0.32278	0.21755
				0.32581	0.34122	2.53230			
9	0.1153679	0.0000	0.1497	0.05777	0.09245	0.22784	0.32643	0.32314	0.21806
				0.32513	0.34129	2.68202			
10	0.1153615	0.0000	0.1007	0.05776	0.09237	0.22404	0.32578	0.32336	0.21839
				0.32470	0.34134	2.78276			
11	0.1153594	0.0000	0.0630	0.05775	0.09231	0.22176	0.32538	0.32349	0.21859
				0.32445	0.34137	2.84572			
12	0.1153587	0.0000	0.0374	0.05775	0.09228	0.22043	0.32516	0.32357	0.21871
				0.32429	0.34139	2.88311			
13	0.1153585	0.0000	0.0215	0.05774	0.09226	0.21968	0.32503	0.32361	0.21878
				0.32421	0.34140	2.90461			
14	0.1153584	0.0000	0.0121	0.05774	0.09225	0.21925	0.32495	0.32364	0.21882
				0.32416	0.34141	2.91674			
15	0.1153584	0.0000	0.0068	0.05774	0.09224	0.21902	0.32491	0.32365	0.21884
				0.32413	0.34141	2.92350			
16	0.1153584	0.0000	0.0037	0.05774	0.09224	0.21889	0.32489	0.32366	0.21885
				0.32411	0.34142	2.92725			
17	0.1153584	0.0000	0.0021	0.05774	0.09224	0.21881	0.32488	0.32367	0.21886
				0.32411	0.34142	2.92932			
18	0.1153584	0.0000	0.0011	0.05774	0.09224	0.21877	0.32487	0.32367	0.21886
				0.32410	0.34142	2.93046			
19	0.1153584	0.0000	0.0006	0.05774	0.09224	0.21875	0.32486	0.32367	0.21886
				0.32410	0.34142	2.93109			

Convergence criterion satisfied.

Significance Tests Based on 5848 Observations

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors HA: At least one common factor	36	9390.9273	<.0001
H0: 2 Factors are sufficient HA: More factors are needed	19	673.9044	<.0001

Factor Pattern

	Factor1	Factor2
v40	1	24
v41	-2	30
v42	16	33
v43	-1	57 *
v44	1	57 *
v45	3	46
v46	1	57 *
v47	4	58 *
v48	83 *	29

```
Rotation Method: Varimax

     Orthogonal Transformation Matrix

                       1               2

        1         -0.19707         0.98039
        2          0.98039         0.19707


    Rotated Factor Pattern

         Factor1      Factor2

v40          23            6
v41          30            3
v42          29           22
v43          56 *         11
v44          56 *         12
v45          45           12
v46          56 *         12
v47          56 *         15
v48          12           87 *
```

Fortolkningen af resultater ved at bruge optionerne: ’heywood’ og ’ultra’ kræver megen erfaring med analyser af factor analyse output. Dette ligger derfor uden for rammerne af dette kursus.

12. Om skaleringer i Linear Structural Models.

En ligning i PROC FACTOR eller i PROC CALIS kan f.eks. se sådan ud:

$$V1 = 0.871 \cdot F1 + E1 \ .$$

Det betyder i SAS-sprog, at den observérbare eller manifeste variabel V1 er knyttet sammen med den latente variabel F1 som en lineær sammenhæng, mens den del af variationen i V1, der ikke forklares af F1, betegnes restledsvariablen E1. Hvis F1 og E1 er ukorrelerede, og alle variable har middelværdi 0, (som man i reglen antager, eller **kan** antage), kan man direkte beregne denne del af kovariansstrukturen som

$$var[V1] = (0.871)^2 var[F1] + var[E1] \ ,$$

idet cov(F1,E1) = 0. Variansen af V1 er således ganske enkelt opdelt i var[F1] ganget med kvadratet på koefficienten i den lineære ligning, og restledsvariansen var[E1].

Koefficenten i den lineære ligning kan fortolkes på følgende vis. Gentager vi ligningen oven for med en vilkårlig konstant b, dvs.

$$V1 = b \cdot F1 + E1 \ ,$$

er det spørgsmålet om b kan fortolkes i en covarians-sammenhæng. Ved direkte beregning får man

$$cov(V1,F1) = cov(bF1 + E1,F1) = bvar[F1] + 0 \ ,$$

da F1 og E1 er ukorrelerede. Heraf følger, at

$$b = \frac{cov(V1,F1)}{var[F1]} \ .$$

Skal dette udtrykkes ved en korrelationskoefficient, får man

$$b = \frac{cov(V1,F1)}{\sqrt{var[F1] \cdot var[V1]}} \cdot \frac{\sqrt{var[V1]}}{\sqrt{var[F1]}} \ ,$$

eller
Koefficenten b i den strukturelle ligning er altså en
skaleret udgave af korrelationskoefficienten mellem de to variable V1 og F1. Skaleringen skal naturligvis blot tage højde for, hvordan den observérbare variabel V1 er skaleret, og hvordan

$$b = \rho_{V1F1}\frac{\sigma_{V1}}{\sigma_{F1}} \quad .$$

vi har valgt at skalere den latente variabel F1. Hvis man skalerer den latente variabel sådan, at var[F1] = 1, får vi

$$b = \rho_{V1F1} \cdot \sigma_{V1} \ .$$

I målingsmodellen i Kapitel 18 er det denne skalering, der benyttes. Hvis man skalerer så var[V1] = 1 bliver $\sigma_{V1} = 1$, og koefficienten b ændres til koefficienten b^*, defineret som

$$b^* = \rho_{V1F1} \ ,$$

dvs. korrelationskoefficienten mellem den manifeste og den latente variable.

I SAS-output fra PROC CALIS benævnes koefficienterne b^* 'Standardized Coefficients'.

Det er vigtigt at bemærke, at hvis man standardiserer så variansen på de manifeste variable bliver 1, skal koefficienten til restleddet E ganges med en konstant, svarende til at ligningen 'ganges igennem' med $1/\sigma_{V1}$ i eksemplet her.

Man kan sammenligne dette med en almindelig lineær regressionsanalyse, hvor både den forklarende variabel X og responsvariablen Y er stokastiske. Det svarer til modellen

$$Y = \beta X + e \quad , \quad e \sim N(0, \sigma^2) \ ,$$

hvor regressionskoefficienten β har formen

$$\beta = \rho_{XY}\frac{\sigma_Y}{\sigma_X} \ .$$

Så det forekommer ret klart, hvordan en koefficient i en strukturel SAS-ligning skal fortolkes, nemlig enten som regressionsparameteren i den tilsvarende regressionsmodel mellem de variable V1 og F1, eller som en skaleret korrelationskoefficient, på samme måde, som β i regressionsanalysen er en skaleret version af korrelationskoefficienten. Til gengæld vises selve korrelationskoefficienten (b^*) også i SAS-output'et.

Det er vigtigt at holde rede på disse skaleringer. Korrelationskoeffienter er tal mellem 0 og 1, dvs. i princippet skalauafhængige. Regressionskoeffienter er udtryk for én variabels relative ændring i forhold til en anden, og afhænger derfor på afgørende vis af skaleringen af de indgående variable.

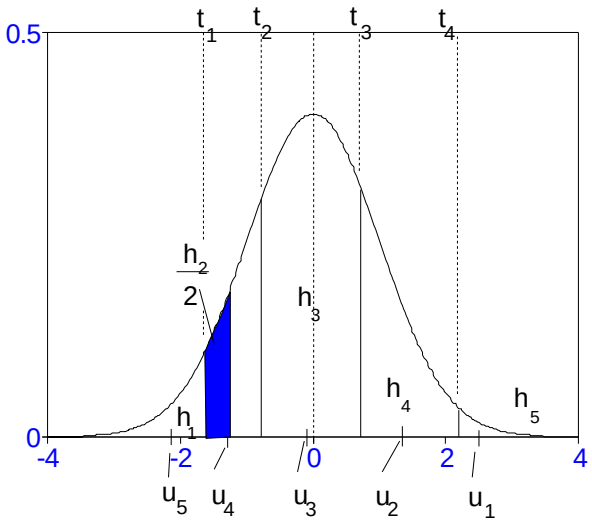
13. Polyseriale korrelationskoefficienter.

Mange variable har som responskategorier tallene fra 1 til 4, eller fra 1 til 5, svarende til f.eks. svarmulighederne:

1. Meget enig.
2. Enig.
3. Hverken eller.
4. Uenig.
5. Meget uenig.

Her kan man så blot benytte de hele tal, svarende til svarkategorierne, til at danne korrelationer. Ofte får man imidlertid et bedre resultat, hvis man forestiller sig, at de 5 svar er udtryk for, at en latent variabel u, der udtrykker ens holdning til spørgsmålet, overskrider visse tærskelværdier t_1, t_2, t_3 og t_4. Da korrelationskoefficienter jo mest meningsfyldt er knyttet til normalfordelte stokastiske variable, skal vi antage, at u er standardiseret normalfordelt. Svaret 5 gives derfor, hvis $u < t_1$, svaret 4, hvis $t_1 < u < t_2$, osv op til $u > t_4$, der giver svaret 1. Man kunne derfor måske få en bedre scoring af svarkategorierne ved at referere til en sådan bagvedliggende standardiseret normalfordeling.

Hvis h_1, h_2, h_3, h_4 og h_5 er de marginale hyppigheder af 5, 4, 3, 2 og 1 svar i datamaterialet, og $H_1 = h_1$, H_2, H_3, H_4 og $H_5 = 1$, de tilsvarende kumulerede hyppigheder, kan man lave tegningen



Det fremgår, at en rimelig måde at score svarmulighederne på som

$$u_5 = \Phi^{-1}\left(\frac{h_1}{2}\right),$$

$$u_4 = \Phi^{-1}\left(H_1 + \frac{h_2}{2}\right),$$

$$u_3 = \Phi^{-1}\left(H_2 + \frac{h_3}{2}\right),$$

$$u_2 = \Phi^{-1}\left(H_3 + \frac{h_4}{2}\right)$$

og

$$u_1 = \Phi^{-1}\left(H_4 + \frac{h_5}{2}\right)$$

Korrelationskoefficienter, dannet ud fra de således konstruerede u'er, kaldes polyseriale korrelationskoefficienter.

Metoden, vist oven for, er egentlig ikke den helt rigtige. Men den kommer dels tæt på det rigtige resultat, og viser i det mindste grundidéen. For at danne de rigtige polyseriale korrelationskoefficienter for to variable skal man opstille hele likelihoodfunktionen for den kontingenstabel, der dannes af de fem svarkategorier, dvs. hvor antallet a_{ij} i celle (i,j) er dem, der har svaret i på det første spørgsmål og j på det andet spørgsmål. Denne skal så udtrykkes ved den to-dimensionale normalfordeling for de latente u'er for de to variable, og herfra kan så de polyseriale korrelationskoefficienter bestemmes ved ML-estimation i multinomialfordelingen for kontingenstabellen.

14. Principalkomponentanalyse og faktor analyse med polyseriale korrelationskoefficienter.

PROC FACTOR i SAS kan ikke direkte beregne de polyseriale korrelationskoefficienter i forbindelse med kaldet af proceduren. Man kan imidlertid lave denne, ganske vist lidt omstændelige, men dog praktisk mulige, beregningsmåde i SAS. Følger vi metoden, foreslået i Kapitel 13, skal man skrive programmet, vist i boksen neden for.

```
1. libname eba 'S:\eba-data';
2. title 'ISSP 1989. Polyseriale koeefficenter. Principal component analyse.';
3. data a;
4. set eba.d89um;
5. keep v24-v32;
6. data b;
7. set a;
8.  if v24>5 then delete;
9.  if v25>5 then delete;
10. if v26>5 then delete;
11. if v27>5 then delete;
12. if v28>5 then delete;
13. if v29>5 then delete;
14. if v30>5 then delete;
15. if v31>5 then delete;
16. if v32>5 then delete;
17. proc freq data=b noprint;
18. table v24/out=d24;
19. data dn24;
20. set d24;
21. kum+percent/100;
22. if v24=1 then kum1=kum/2;
23. kum2=(kum+lag(kum))/2;
24. k=sum(kum1,kum2);
25. u=probit(k);
26. proc print data=dn24;

.....

27. proc freq data=b noprint;
28. table v32/out=d32;
29. data dn32;
30. set d32;
31. kum+percent/100;
32. if v32=1 then kum1=kum/2;
33. kum2=(kum+lag(kum))/2;
34. k=sum(kum1,kum2);
35. u=probit(k);
36. proc print data=dn32;
37. run;
38. quit;
```

Data set 'a', baseret på ISSP-databasen for 1989, indlæses, som vist i Kapitel 1, i programlinierne 1 til 5. I linierne 6 til 16 renses data på sædvanligvis for ubesvarede spørgsmål. Det rensede data set kaldes 'b'.

I programlinerne 17 og 18 dannes et output sæt, kaldet 'd24'. Dette dataset indeholder, gennem kaldet af PROC FREQ og optionen 'out = d24' antal procenter, kumulerede antal og kumulerede procenter for hver af de 5 svarkategorier på spørgsmål V24. I de næste programlinier, linie 19 til 23 beregnes de kumulerede andele 'kum+percent/100' og midtpunkterne mellem de kumulerede andele: 'kum1=kum/2' for V24 = 1 og 'kum2=(kum+lag(kum))/2' for V24 = 2 til V24 = 5. (De tilsvarende beregninger for V32 sker i linierne 29 til 33.) Endelig

beregnes i programlinie 24 (34) for den værdi k, som den inverse af Φ skal beregnes af. Den inverse af den normerede normalfordelings fordelingskurve, dvs. tranformationen Φ^{-1}, betegnes i reglen i matematisk statistik som 'probit', så i SAS-sprog skal vi blot skrive 'u = probit(k)'. Dette sker i linie 25 for V24 og i linie 35 for V32. De beregnede størrelser udskrives i linie 26 for V24 og i linie 36 for V34.

SAS output fra dette program ser sådan ud:

ISSP 1989. Polyseriale koefficenter. Principal component analyse.

OBS	V24	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	7894	59.2020	0.59202	0.29601	.	0.29601	-0.53591
2	2	4575	34.3108	0.93513	.	0.76357	0.76357	0.71785
3	3	590	4.4248	0.97938	.	0.95725	0.95725	1.71965
4	4	214	1.6049	0.99543	.	0.98740	0.98740	2.23834
5	5	61	0.4575	1.00000	.	0.99771	0.99771	2.83554

OBS	V25	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	3458	25.9337	0.25934	0.12967	.	0.12967	-1.12796
2	2	7321	54.9048	0.80838	.	0.53386	0.53386	0.08498
3	3	2003	15.0217	0.95860	.	0.88349	0.88349	1.19263
4	4	505	3.7873	0.99648	.	0.97754	0.97754	2.00538
5	5	47	0.3525	1.00000	.	0.99824	0.99824	2.91782

OBS	V26	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	4299	32.2409	0.32241	0.16120	.	0.16120	-0.98952
2	2	6651	49.8800	0.82121	.	0.57181	0.57181	0.18098
3	3	1650	12.3744	0.94495	.	0.88308	0.88308	1.19053
4	4	634	4.7548	0.99250	.	0.96873	0.96873	1.86240
5	5	100	0.7500	1.00000	.	0.99625	0.99625	2.67380

OBS	V27	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	1585	11.8869	0.11887	0.059435	.	0.05943	-1.55954
2	2	4930	36.9732	0.48860	.	0.30373	0.30373	-0.51369
3	3	4116	30.8685	0.79729	.	0.64294	0.64294	0.36634
4	4	2377	17.8266	0.97555	.	0.88642	0.88642	1.20770
5	5	326	2.4449	1.00000	.	0.98778	0.98778	2.25000

OBS	V28	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	6565	49.2350	0.49235	0.24618	.	0.24618	-0.68658
2	2	5952	44.6378	0.93873	.	0.71554	0.71554	0.56964
3	3	609	4.5673	0.98440	.	0.96156	0.96156	1.76914
4	4	178	1.3349	0.99775	.	0.99108	0.99108	2.36873
5	5	30	0.2250	1.00000	.	0.99888	0.99888	3.05510

OBS	V29	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	4445	33.3358	0.33336	0.16668	.	0.16668	-0.96737
2	2	6106	45.7927	0.79129	.	0.56232	0.56232	0.15686
3	3	1954	14.6543	0.93783	.	0.86456	0.86456	1.10102
4	4	750	5.6247	0.99408	.	0.96595	0.96595	1.82437
5	5	79	0.5925	1.00000	.	0.99704	0.99704	2.75192

ISSP 1989. Polyseriale koeefficenter. Principal component analyse.

OBS	V30	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	3084	23.1288	0.23129	0.11564	.	0.11564	-1.19705
2	2	6384	47.8776	0.71006	.	0.47068	0.47068	-0.07357
3	3	2856	21.4189	0.92425	.	0.81716	0.81716	0.90459
4	4	850	6.3747	0.98800	.	0.95613	0.95613	1.70741
5	5	160	1.1999	1.00000	.	0.99400	0.99400	2.51216

OBS	V31	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	3419	25.6412	0.25641	0.12821	.	0.12821	-1.13491
2	2	6408	48.0576	0.73699	.	0.49670	0.49670	-0.00827
3	3	2576	19.3190	0.93018	.	0.83358	0.83358	0.96842
4	4	767	5.7522	0.98770	.	0.95894	0.95894	1.73851
5	5	164	1.2299	1.00000	.	0.99385	0.99385	2.50344

OBS	V32	COUNT	PERCENT	KUM	KUM1	KUM2	K	U
1	1	2385	17.8866	0.17887	0.089433	.	0.08943	-1.34425
2	2	4922	36.9132	0.54800	.	0.36343	0.36343	-0.34930
3	3	3464	25.9787	0.80778	.	0.67789	0.67789	0.46181
4	4	2145	16.0867	0.96865	.	0.88822	0.88822	1.21711
5	5	418	3.1348	1.00000	.	0.98433	0.98433	2.15262

Søjlerne 'KUM' og 'K' viser for hver af de 5 mulige svarkategorier 'KUM' = den kumulerede andel af svar i denne kategori, og 'K' lig med værdierne af de ved formlerne

$$h_1/2 \ , \ H_1 + h_2/2, \ ... \ ,H_4 + h_5/2 \ ,$$

beregnede værdier. I søjlen 'U' er vist værdierne af Φ (K), dvs. de værdier de polyseriale korrelationskoefficienter skal beregnes udfra.

Desværre er man nødt til at kopiere de fundne u-værdier 'i hånden' til et nyt data sæt, hvis man ønsker en principal komponentanalyse af de 'nye' variable, u'erne. Man skal her skrive programmet:

```
1. libname eba 'S:\eba-data';
2. title 'ISSP 1989. Polyseriale koefficienter. Principal komponent metode.';
3. data a;
4. set eba.d89um;
5. keep v24-v32;
6. data b;
7. set a;
8.  if v24>5 then delete;
9.  if v25>5 then delete;
10. if v26>5 then delete;
11. if v27>5 then delete;
12. if v28>5 then delete;
13. if v29>5 then delete;
14. if v30>5 then delete;
15. if v31>5 then delete;
16. if v32>5 then delete;
17. data bb;
18. set b;
19. if v24=1 then u24=-0.536;
20. if v24=2 then u24=0.718;
21. if v24=3 then u24=1.720;
22. if v24=4 then u24=2.238;
23. if v24=5 then u24=2.836;
24. if v25=1 then u25=-0.128;
25. if v25=2 then u25=0.085;
26. if v25=3 then u25=1.192;
27. if v25=4 then u25=2.005;
28. if v25=5 then u25=2.918;

 .....

29. if v32=1 then u32=-1.344;
30. if v32=2 then u32=-0.349;
31. if v32=3 then u32=0.462;
32. if v32=4 then u32=1.217;
33. if v32=5 then u32=2.153;
34. proc factor data=bb
35.   method = prin
36.   priors = one
37.   nfact = 2
38.   rotate = varimax
39.   flag= 0.5;
40.  var u24-u32;
41. run;
42. quit;
```

I programlinierne 19 til 33 omdannes værdierne af V24 til V32 til de u-værdier, vi fandt i det forrige program.

I programlinie 40 får SAS ordre til at bruge u-værdierne, defineret i programlinierne 19 til 33, til egenværdi-dekomponeringen af korrelationsmatricen.

Output fra kørslen af dette program ser således ud:

```
ISSP 1989. Polyseriale koefficienter. Principal komponent metode.

1           2           3           4           5
Eigenvalue        2.6821      1.3259      1.0991      0.9474      0.7637

2 factors will be retained by the NFACTOR criterion.

Rotation Method: Varimax

Orthogonal Transformation Matrix

              1         2

    1      0.87197   0.48956
    2     -0.48956   0.87197

Rotated Factor Pattern

            FACTOR1   FACTOR2

U24           9         59 *
U25          -8         80 *
U26          25         65 *
U27          26         34
U28          55 *       25
U29          63 *       18
U30          80 *       -1
U31          77 *        1
U32          51 *       16

Variance explained by each factor

   FACTOR1   FACTOR2
  2.357061  1.650944


Final Communality Estimates: Total = 4.008004

       U24       U25       U26       U27       U28       U29       U30       U31       U32
  0.354253  0.654101  0.481343  0.187233  0.365662  0.432521  0.647573  0.595310  0.290009
```

Hvis man sammenligner med resultatet af at bruge de direkte korrelationskoefficienter, som beskrevet i Kapitel 7, kan man se, at resultaterne ikke har ændret sig mærkbart. Ved en 5-punkts svarskala kan man altså i mange tilfælde nå fornuftige resultater, uden at benytte polyseriale korrelationskoefficienter. Det svarer nogenlunde til almindelige erfaringer. Men det er vigtigt at huske på, at de 5 svarkategorier skal være konstrueret verbalt, så de danner en skala fra det meget positive/enige til det meget negative/uenige alternativ. Desuden skal respondenterne naturligvis også opfatte de 5 svarkategorier som en skala.

Også for en eksplorativ faktor analyse, får man tilsvarende omtrent ens resultater ved at en analyse udført direkte på korrelationsmatricen mellem de oprindelige variable og ved at analysere matricen af polyseriale korrelationskoefficienter. Her ser programmet således ud:

```
libname eba 'S:\eba-data';
title 'ISSP 1989. Polyseriale koefficienter. Principal komponent metode.';
data a;
set eba.d89um;
keep v24-v32;
data b;
set a;
if v24>5 then delete;

.....

data bb;
set b;
if v24=1 then u24=-0.536;
if v24=2 then u24=0.718;
if v24=3 then u24=1.720;
if v24=4 then u24=2.238;
if v24=5 then u24=2.836;

.....

if v32=1 then u32=-1.344;
if v32=2 then u32=-0.349;
if v32=3 then u32=0.462;
if v32=4 then u32=1.217;
if v32=5 then u32=2.153;
proc factor data=bb
  method = ml
  priors = smc
  nfact = 2
  rotate = promax
  flag= 0.5
;
var u24-u32;
run;
quit;
```

De vigtigste dele af SAS-output'et bliver

```
Significance tests based on 13334 observations:

   Test of H0: No common factors.
        vs HA: At least one common factor.

   Chi-square = 19932.709   df = 36   Prob>chi**2 = 0.0001

   Test of H0: 2 Factors are sufficient.
        vs HA: More factors are needed.

   Chi-square = 2703.916   df = 19   Prob>chi**2 = 0.0001

Rotation Method: Promax

Factor Structure (Correlations)

           FACTOR1   FACTOR2

U24          19        35
U25           4        51 *
U26          26        57 *
U27          18        31
U28          37        41
U29          43        39
U30          84 *      20
U31          75 *      22
U32          34        30
```

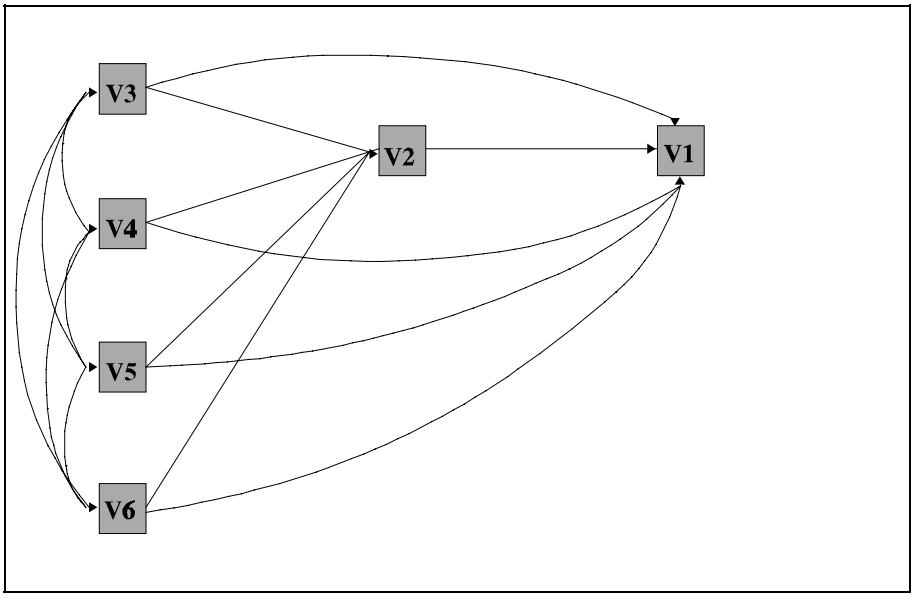
Også her får vi stort set de samme resultater som ved at regne direkte på de observerede variable.

15. Stianalyse

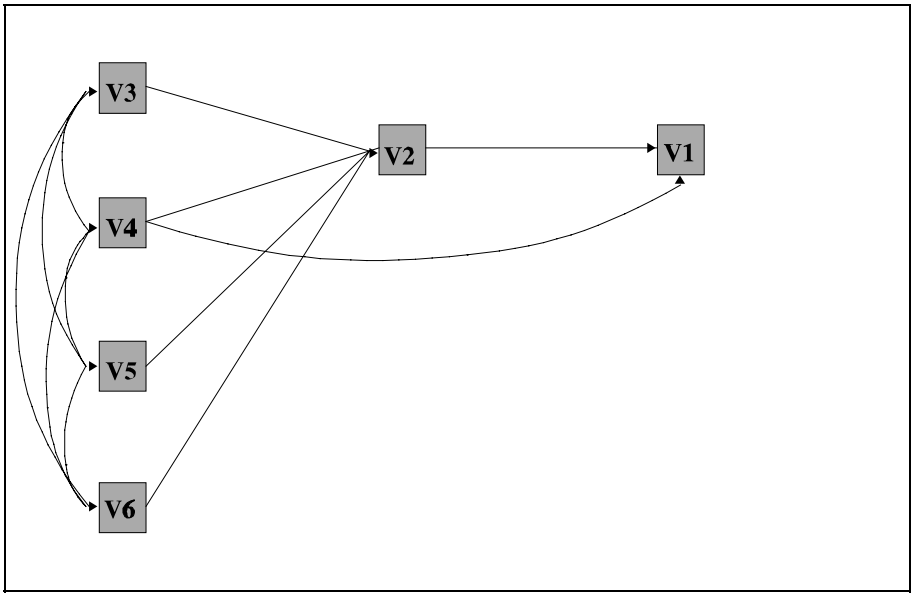
Stianalyse er en statistisk metode, hvor man forsøger at fortolke korrelationerne, dvs. afhængighederne mellem en række variable, ved hjælp af et stidiagram, hvor nogle variable antages at have indflydelse fremad i diagrammet på andre variable. Jvf. diagrammet neden for.

Det er ikke usædvanligt, at man fortolker disse sammenhænge fremad, specielt hvor de variable, fremad i diagrammet, er observeret til senere tidspunkter, end de første, som en **kausal** sammenhæng, dvs en årsagssammenhæng. Begrebet kausalitet er imidlertid et meget kompliceret begreb, som man skal være uhyre varsom med at anvende. Der er utallige eksempler i statistikkens brogede historie, hvor korrelationer er blevet fortolket som årsagssammenhænge. Det mest berømte eksempel er fra starten af dette århundrede, hvor man fandt en overraskende stærk korrelation mellem antal børnefødsler og antal observerede storke i Skåne. Nu er der jo mere mellem himmel og jord, end man går og tror, men alligevel .....

For at gøre det hele mere konkret: Antag at vi har fundet en række statistisk signifikante korrelationer mellem de nogle typiske 'baggrundsvariable' V3 til V6, som køn, alder, urbanisering (dvs. om man bor i København, Københavns forstæder, de tre store byer - Odense, Århus, Aalborg -, Byer iøvrigt eller på Landet) og uddannelse. V1 og V2 er på den anden siden 'afledte' variable, som f.eks. Holdning til EU eller Miljøbevisthed. Hvis der er korrelationer/afhængighed mellem alle de 6 variable, ville et diagram til beskrivelse af afhængighederne se således ud.



Man kunne også tænke sig, at korrelationer mellem V3 og V1, V5 og V1 samt mellem V6 og V1 er insignifikante. Det ville give stidiagrammet vist den næste figur.



Læg mærke til, at der er linier mellem alle de fire baggrundsvariable. Det betyder, at vi ikke interesserer os for, om der evt. er sammenhænge mellem disse 4 variable. Den væsentligste oplysning, man kan aflæse af diagrammet, er, at V3, V5 og V6 kun påvirker V1 via V2, mens V4 både påvirker V1 via V2 og desuden direkte.

Det er denne type stidiagrammer, vi skal interessere os for i det næste kapitel og i Kapitel 19.

16. Stianalyse med manifeste variable

Vi skal anvende stianalyse på nogle af de variabelgrupper, vi allerede har arbejdet med samt nogle spørgsmål fra gruppen V68 til V74, der drejer sig om respondenternes arbejdsbetingelser.

Fra spørgsmålene vedrørende respondenternes vurdering af de forskellige aspekter af deres job, fandt vi to faktorer 'Good Job' og 'Interesting Job', der var baseret på henholdsvis V24 til V26 og V28 til V32. I denne omgang giver vi alle de variable, som bidrager til en faktor samme vægt. Det betyder, at vi kan konstruere den manifeste variabel

$$V1 = \text{'Good Job'} = V24 + V25 + V26,$$

dvs. for hver person summen af scoreværdien på de tre spørgsmål. Den maksimale scoreværdi bliver 15 - med 5 svarkategorier, scoret 1 til 5 - , mens den mindste scoreværdi bliver 3. En score på 15 betyder, at man har svaret 3 gange 'Not important at all', svarende til en person, der mener hans eller hendes job er meget lidt 'væsentligt'. En score på 3 betyder på den anden side, at den pågældende respondent har svaret 'Very important' på alle 3 spørgsmål, og således mener, at hans eller hendes job er særdeles væsentligt. Tilsvarende kan vi konstruere den manifeste variable

$$V2 = \text{'Interesting Job'} = V28 + V29 + V30 + V31,$$

som på tilsvarende måde bliver et mål for hvor interessant jobbet er. Fra analysen af de 7 spørgsmål vedrørende arbejdsmiljøet kan vi danne en manifest variabel, der beskriver, hvor hårdt arbejdet er ('Dangerous Job'), som

$$V3 = \text{'Dangerous Job'} = V69 + V72 + V73 + V74.$$

Her betyder en lav score, at arbejdet er fysisk meget belastende, mens en høj score betyder, at arbejdet ikke er særligt fysisk belastende.

Desuden inddrages i analysen baggrundsvariablene V4, der er opgivne antal timer, respondenten arbejder om ugen, her kaldet 'Hours a Week', V5, der er social klasse ('Social Class') med 6 sociale klasser, og V6, der er antal år, respondentens har gået i skole, her benævnt 'School Years'. Baggrundsvariablene er altså

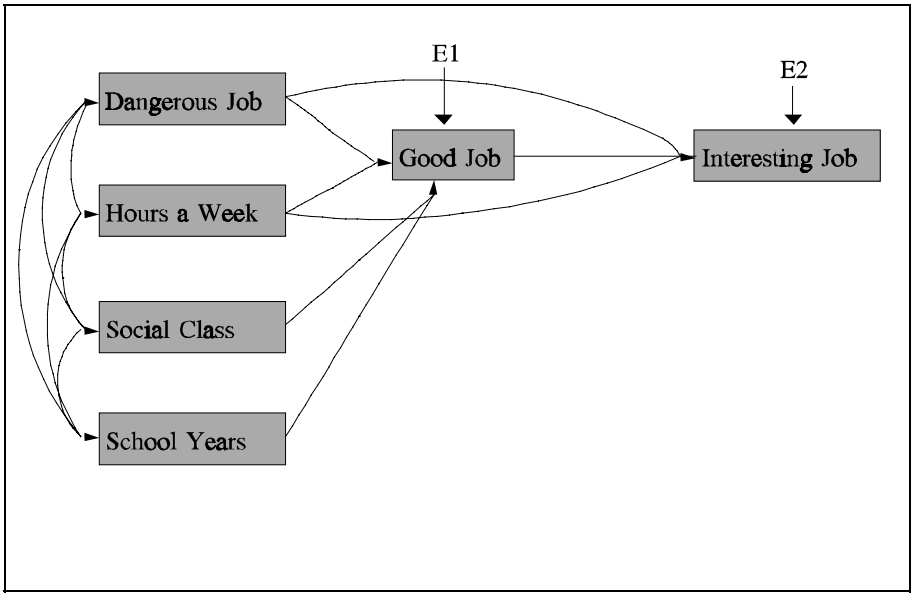
$$V4 = \text{'Hours a Week'} = V90,$$

$$V5 = \text{'Social Class'} = V110$$

og

$$V6 = \text{'School Years'} = V100.$$

Med disse variable, starter vi med stidiagrammet, vist i nedenstående figur.



Hvorfor der står E1 og E2 ved 'Good Job' og 'Interesting Job', vender jeg tilbage til.

Dette stidiagram analyseres ved SAS-programmet vist i de to næste bokse.

Den første boks, på næste side, viser i linierne 1 til 5 indlæsning af de nødvendige variable fra ISSP-databasen. Herefter etableres dataset 'b', hvor de variable renses for ubesvarede spørgsmål i linierne 6 til 28, mens de 6 variable, der skal indgå i stianalysen, defineres i linierne 29 til 34. I linierne 35 til 41 sættes der labels på de 6 manifeste variable V1 til V6 af hensyn til læsningen af SAS-output'et.

```
1. libname eba 's:\eba-data';
2. title 'Path analyse with manifest variables. ISSP 1989';
3. data a;
4. set eba.d89um;
5. keep v24-v26 v28-v31 v69 v72-v74 v90 v110 v100 ;
6. data b;
7. set a;
8. if v24>5 then delete;
9. if v25>5 then delete;
10. if v26>5 then delete;
11. if v28>5 then delete;
12. if v29>5 then delete;
13. if v30>5 then delete;
14. if v31>5 then delete;
15. if v69>5 then delete;
16. if v72>5 then delete;
17. if v73>5 then delete;
18. if v74>5 then delete;
19. if v69=0 then delete;
20. if v72=0 then delete;
21. if v73=0 then delete;
22. if v74=0 then delete;
23. if v90>70 then delete;
24. if v90=0 then delete;
25. if v100>14 then delete;
26. if v100=0 then delete;
27. if v110>6 then delete;
28. if v110=0 then delete;
29. v1=v24+v25+v26;
30. v2=v28+v29+v30+v31;
31. v3=v69+v72+v73+v74;
32. v4=v90;
33. v5=v110;
34. v6=v100;
35. label
36.  v1='Good job'
37.  v2='Interesting job'
38.  v3='Dangerous job'
39.  v4='Hours per week'
40.  v5='Social class'
41.  v6='Years in school';
```

```
1. proc calis corr residual;
2.   lineqs
3.    v2 = pv2v1 v1 + pv2v3 v3 + pv2v4 v4 + e2,
4.    v1 = pv1v3 v3 + pv1v4 v4 + pv1v5 v5 + pv1v6 v6 + e1;
5.   std
6.    e1 = vare1,
7.    e2 = vare2,
8.    v3 = varv3,
9.    v4 = varv4,
10.   v5 = varv5,
11.   v6 = varv6;
12.  cov
13.   v3 v4 = cv3v4,
14.   v3 v5 = cv3v5,
15.   v3 v6 = cv3v6,
16.   v4 v5 = cv4v5,
17.   v4 v6 = cv4v6,
18.   v5 v6 = cv5v6;
19.  var v1-v6;
20. run;
21. quit;
```

I denne boks er kaldet af PROC CALIS vist. I linie 1 er anført SAS-optionerne: 'corr', der betyder, at alle estimerede korrelationskoefficienter udskrives, og 'residual', der medfører, at differensen mellem de observerede korrelationer og de af stimodellen estimerede korrelationer udskrives. Disse differenser kaldes **residualer**, og der er altså én residual for hver observeret korrelation.

Linie 3 og 4 er definitionsligningerne for V2 og V1, som vist på figuren. Linie 5-11 er alle de varianser, der indgår som parametre i modellen. Bemærk, at varianserne på V2 og V1 kaldes E2 og E1, idet de har en lidt anden status end de andre varianser på grund af definitionsligningerne for V2 og V1 i linie 3 og 4. Endelig står i linie 12-18 de korrellationer, der indgår i modellen som parametre. Bemærk her, at der ingen korrelationer er mellem V1, V2 og de øvrige 4 variable.

SAS output'et fra dette program er vist i de følgende bokse.

```
Manifest Variable Equations

 V1      =     .    *V3    +  .    *V4    +  .    *V5    +  .    *V6    + 1.0000 E1
                     PV1V3          PV1V4          PV1V5          PV1V6

 V2      =     .    *V1    +  .    *V3    +  .    *V4    + 1.0000 E2
                     PV2V1          PV2V3          PV2V4

    3834 Observations       Model Terms          1
       6 Variables          Model Matrices       4
      21 Informations       Parameters          19
```

I denne første boks vises de to definitionsligninger, linie 3 og 4, for de 'afledte' variable V1 og V2 på symbolsk form, så man kan checke, at ligningerne er korrekt skrevet op i SAS-programmet. Desuden vises 6 nøgletal fra datasættet og modellen, bl.a. at der er 3834 observationer, som indgår i beregningen af korrelationerne, at der er 6 variable og 21 elementer i korrelationsmatricen, nemlig $7 \cdot 6/2$. Endelig ser vi, at der 19 parametre, der skal estimeres i modellen.

I den næste boks vises en lang række teststørrelser og tilpasningsmål. De to tal, vi har brug for er fremhævet.

```
Fit criterion . . . . . .     0.0077
Goodness of Fit Index (GFI) . . . . . .     0.9975
GFI Adjusted for Degrees of Freedom (AGFI). . . .     0.9733
Root Mean Square Residual (RMR) . . . . . .     0.0176
Parsimonious GFI (Mulaik, 1989) . . . . . .     0.1330

Chi-square = 29.4095       df = 2       Prob>chi**2 = 0.0001
Null Model Chi-square:     df = 15                  721.1757

RMSEA Estimate  . . . . . .  0.0598  90%C.I.[0.0419, 0.0798]
Probability of Close Fit  . . . . . .     0.1731
ECVI Estimate . . . . . .  0.0176  90%C.I.[0.0140, 0.0232]
Bentler's Comparative Fit Index . . . . . .     0.9612
Normal Theory Reweighted LS Chi-square  . . . . .    29.2969
Akaike's Information Criterion. . . . . .    25.4095
Bozdogan's (1987) CAIC. . . . . .    10.9061
Schwarz's Bayesian Criterion. . . . . .    12.9061
McDonald's (1989) Centrality. . . . . .     0.9964
Bentler & Bonett's (1980) Non-normed Index. . . .     0.7089
Bentler & Bonett's (1980) NFI . . . . . .     0.9592
James, Mulaik, & Brett (1982) Parsimonious NFI. .     0.1279
Z-Test of Wilson & Hilferty (1931). . . . . .     4.6831
Bollen (1986) Normed Index Rho1 . . . . . .     0.6942
Bollen (1988) Non-normed Index Delta2 . . . . . .     0.9619
Hoelter's (1983) Critical N . . . . . .        782
```

Kvotienttteststørrelsen for modeltilpasningen, som vi i det følgende vil kalde q-teststørrelsen, er på q = 29.41 med 2 frihedsgrader, som er klart signifikant. Det betyder, at modellens tilpasning til data kan forbedres, men udelukker ikke, at nogle af korrelationerne er insignifikante, svarende til, at de tilsvarende stier på grafen kan udelades. De næste bokse belyser dette.

Bemærk, at SAS output'et altid viser q-teststørrelsen for den såkaldte 'Null Model', hvor der overhovedet ingen korrelationer er mellem de variable. Det vil sige, at modellen kun har 6 parametre, svarende til de 6 restledsvarianser. Frihedsgraderne for dette test er derfor 21 - 6 = 15. Hvorfor dette test, totalt uden informationsværdi, vises af SAS, er mig en gåde.

I den næste boks vises de standardiserede residualer. Da de asymptotiske standardfejl på residualerne beregnes af SAS-programmet, kan der også beregnes en 'Asymptotic Standardized Residual Matrix', der viser residualerne divideret med deres standard fejl. Elementerne i 'Asymptotic Standardized Residual Matrix' kan sammenholdes med tallet 2 for at se, om de er signifikante. Denne tabel skal, som sagt, inspiceres for værdier, der er klart større end 2. Alle 0'erne i tabellen svarer til korrelationer, der estimeres af modellen.

```
Asymptotically Standardized Residual Matrix

             V1           V2           V3           V4           V5           V6

V1       0.0000       0.0000       0.0000       0.0000       0.0000       0.0000     Good job
V2       0.0000       0.0000       0.0000       0.0000      -5.1623       1.2626     Interest. job
V3       0.0000       0.0000       0.0000       0.0000       0.0000       0.0000     Dangerous job
V4       0.0000       0.0000       0.0000       0.0000       0.0000       0.0000     Hours a week
V5       0.0000      -5.1623       0.0000       0.0000       0.0000       0.0000     Social class
V6       0.0000       1.2626       0.0000       0.0000       0.0000       0.0000     School Years
Average Standardized Residual = 0.3059
Average Off-diagonal Standardized Residual = 0.4283

Rank Order of 2 Largest Asymptotically Standardized Residuals

     V5,V2      V6,V2
   -5.1623     1.2626
```

Som det fremgår er der én signifikant modelafvigelse, svarende til korrelationen mellem V2 og V5. Vi venter lidt med at forfølge denne oplysning.

Først betragter vi definitionsligningerne for V1 og V2, hvor regressionskoefficienterne såvel som deres standardfejl og 't Values' er vist. En 't-Value' er den estimerede regressions-koefficient divideret med sin standardfejl. Disse størrelser er signifikante, hvis deres værdier er større end ca. 2. Benyttes denne 'regel' fremgår det, at hele 4 koefficienter er insignifikante, nemlig den mellem V1 og V3, den mellem V1 og V5, den mellem V2 og V3 og den mellem V2 og V4.

```
Manifest Variable Equations

 V1      =     0.0148*V3    -  0.0543*V4    +  0.0158*V5    +  0.0731*V6    +  1.0000 E1
 Std Err       0.0168 PV1V3    0.0163 PV1V4    0.0165 PV1V5    0.0162 PV1V6
 t Value       0.8797         -3.3238          0.9536          4.5051

 V2      =     0.2658*V1    -  0.0048*V3    +  0.0098*V4    +  1.0000 E2
 Std Err       0.0156 PV2V1    0.0158 PV2V3    0.0158 PV2V4
 t Value      17.0301         -0.3048          0.6230
```

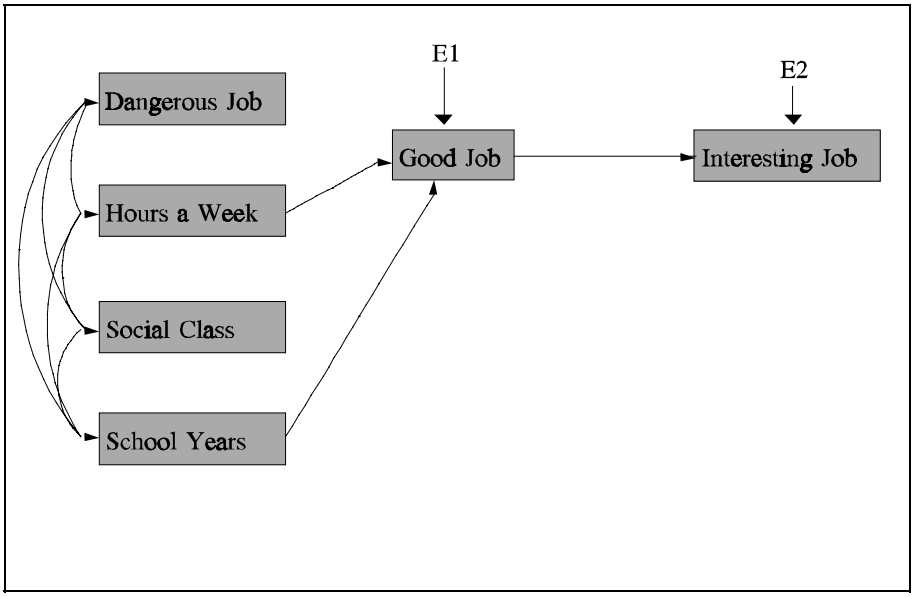
Disse informationer fra SAS output'et fortæller os, som vi skal se om et øjeblik, at grafen for stimodellen kan forenkles betydeligt uden at væsentligt forringe modeltilpasningen.

Endelig viser den næste boks korrelationerne mellem 'baggrundsvariablene' V3, V4 og V5.

```
              Covariances among Exogenous Variables
-----------------------------------------------------------------
                                       Standard
    Parameter          Estimate          Error          t Value
-----------------------------------------------------------------
V4    V3    CV3V4       -0.159551        0.016356          -9.755
V5    V3    CV3V5        0.210387        0.016506          12.746
V5    V4    CV4V5        0.044621        0.016168           2.760
V6    V3    CV3V6        0.121825        0.016272           7.487
V6    V4    CV4V6       -0.026603        0.016158          -1.646
V6    V5    CV5V6        0.093889        0.016223           5.787
```

Det fremgår heraf, at der er en enkelt insignifikant korrelation, nemlig mellem V4 og V6. I modsætning til de andre signifikante størrelser, hvor modelmodifikationer klart må overvejes, er det knap så væsentligt at modificere modellen ved at fjerne sammenhængen mellem V4 og V6. Dette skyldes, at den væsentlige del af modelfortolkningen består i at fortolke de signifikante sammenhænge mellem 'baggrundvariablene' V3 til V6 og de 'afledte' variable V1 og V2.

Hvis vi starter modelmodifikatonerne med at fjerne de sammenhænge, der svarer til insignifikante regressionskoefficienter i definitionsligningerne for V1 og V2, får vi den forenklede graf:



Modellen svarende til denne graf analyseres ved følgende kald af PROC CALIS, hvor leddene

```
pv2v3 v3 , pv2v4 v4
```

i linie 5 og

```
pv1v3 v3 , pv1v5 v5
```

i linie 6 er udeladt, mens resten af SAS programmet er uændret.

```
1. proc calis
2.  corr
3.   residual;
4.   lineqs
5.   v2 = pv2v1 v1 + e2,
6.   v1 = pv1v4 v4 + pv1v6 v6 + e1;
7.   std
8.    e1 = vare1,
9.    e2 = vare2,
10.   v3 = varv3,
11.   v4 = varv4,
12.   v5 = varv5,
13.   v6 = varv6;
14.  cov
15.   v3 v4 = cv3v4,
16.   v3 v5 = cv3v5,
17.   v3 v6 = cv3v6,
18.   v4 v5 = cv4v5,
19.   v4 v6 = cv4v6,
20.   v5 v6 = cv5v6;
21.  var v1-v6;
22. run;
23. quit;
```

I det næste - korte - uddrag af SAS output'et, ses, at data stadig ikke beskriver modellen godt nok, idet q = 32.10 med 6 frihedsgrader er klart signifikant. Til gengæld er modellens tilpasningsevne ikke blevet forringet. Differensen mellem q-værdien for denne model og den forrige er nemlig

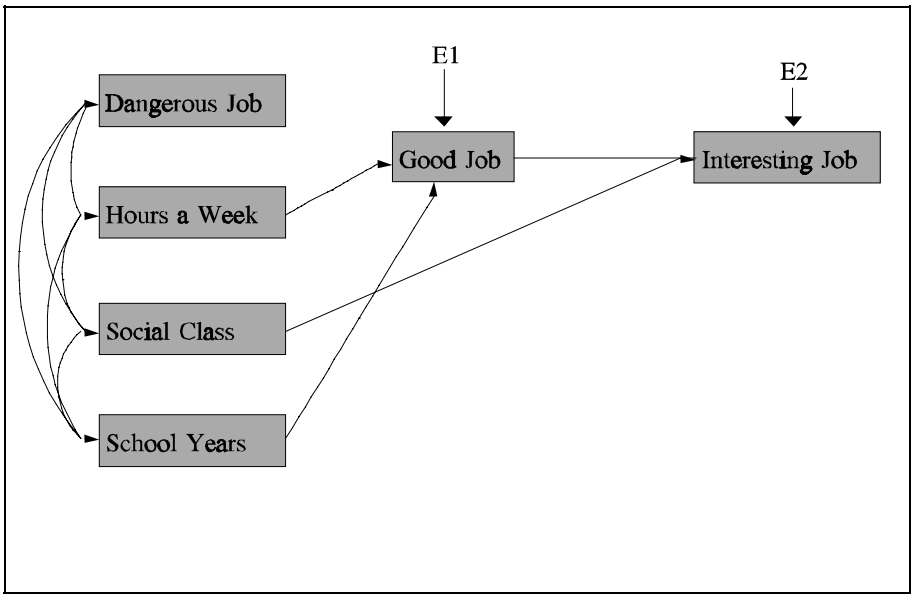
$$Q \text{ (differens)} = 32.10 - 29.41 = 2.69 ,$$

som med df = 6 - 2 = 4 har en signifikanssandsynlighed på 0.611, jvf. Kapitel 17.

Samtidig ser vi, at alle de tilbageværende korrelationer i definitionsligningerne for V1 og V2 har signifikante koefficienter.

```
Chi-square = 32.0976       df = 6       Prob>chi**2 = 0.0001

Manifest Variable Equations

 V1      =  -  0.0559*V4    +  0.0764*V6    +  1.0000 E1
 Std Err       0.0161 PV1V4    0.0161 PV1V6
 t Value      -3.4735          4.7490

 V2      =     0.2650*V1    +  1.0000 E2
 Std Err       0.0156 PV2V1
 t Value      17.0160
```

Der er endnu en måde, vi kan forbedre modellen på, nemlig ved at inddrage den signifikante sammenhæng mellem V2 og V5, som vi 'opdagede' i forbindelse med inspektionen af de standardiserede residualer. Denne nye modificerede model har grafen



Kaldet af PROC CALIS, svarende til denne graf, ser sådan ud, idet vi har tilføjet

pv2v5 v5

i ligningen for V2.

```
proc calis
  corr
  residual;
  lineqs
   v2 = pv2v1 v1 + pv2v5 v5 + e2,
   v1 = pv1v4 v4 + pv1v6 v6 + e1;
  std
   e1 = vare1,
   e2 = vare2,
   v3 = varv3,
   v4 = varv4,
   v5 = varv5,
   v6 = varv6;
  cov
   v3 v4 = cv3v4,
   v3 v5 = cv3v5,
   v3 v6 = cv3v6,
   v4 v5 = cv4v5,
   v4 v6 = cv4v6,
   v5 v6 = cv5v6;
 var v1-v6;
run;
```

Nøgletallene i SAS-output'et for denne model ser sådan ud:

```
Chi-square = 6.3494        df = 5       Prob>chi**2 = 0.2737

Manifest Variable Equations

 V1      =  -  0.0559*V4    +  0.0764*V6    +  1.0000 E1
 Std Err       0.0161 PV1V4    0.0161 PV1V6
 t Value      -3.4735          4.7490

 V2      =     0.2669*V1    -  0.0789*V5    +  1.0000 E2
 Std Err       0.0155 PV2V1    0.0155 PV2V5
 t Value      17.1916         -5.0842
```

Sammenligner man q-teststørrelserne mellem de to sidste modeller, ser man, at

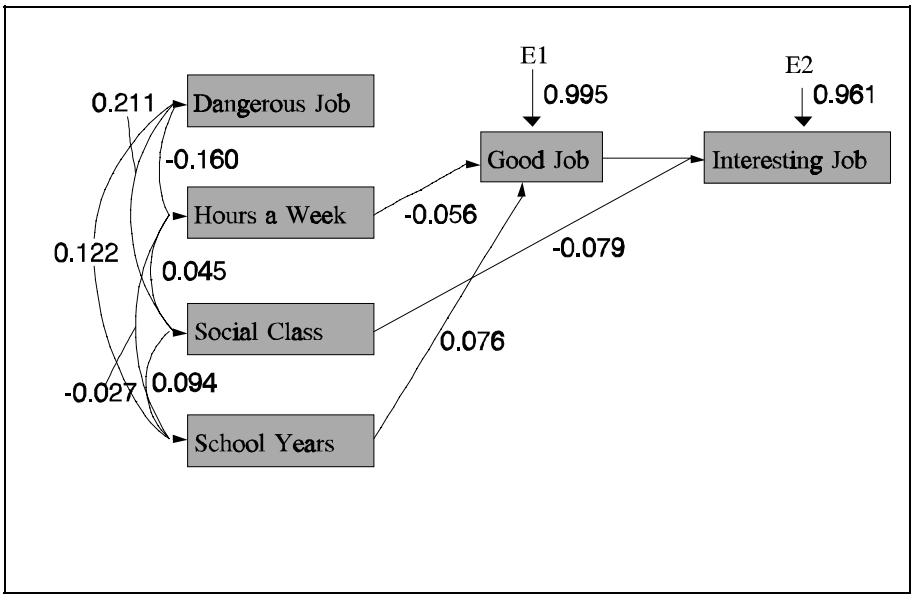
$$q \text{ (differens)} = 32.10 - 6.35 = 25.75 .$$

Med 6 - 5 = 1 frihedsgrad er dette resultat klart signifikant, så at inddrage sammenhængen mellem V2 og V5 er en klar forbedring af modellens tilpasning til data. Ja, endda så meget, at signifikanssandsynligheden er 0.274, svarende til et meget smukt fit til data. Modellen kan derfor accepteres som en tilfredsstillende beskrivelse af variationen i den observerede korrelationsmatrix.

Vi kan således lade dette være vores endelige model. I næste boks er vist de standardiserede koefficienter til stierne fra ’baggrundsvariable’ til ’afledte’ variable og korrelationer mellem ’baggrundsvariablene’, dvs. de tal vi skal bruge til den graf, der belyser styrken at sammenhængene mellem de variable.

```
Equations with Standardized Coefficients

 V1      =  - 0.0559*V4    + 0.0764*V6    + 0.9954 E1
                  PV1V4          PV1V6

 V2      =    0.2668*V1    - 0.0789*V5    + 0.9606 E2
                  PV2V1          PV2V5


Correlations among Exogenous Variables
---------------------------------
    Parameter          Estimate
---------------------------------
V4    V3    CV3V4       -0.159551
V5    V3    CV3V5        0.210387
V5    V4    CV4V5        0.044621
V6    V3    CV3V6        0.121825
V6    V4    CV4V6       -0.026603
V6    V5    CV5V6        0.093889
```

De standardiserede regressionskoefficienter er fremkommet ved at normere så alle indgående variable har varians 1, dvs. der er tale om korrelationskoefficienter. På denne måde er alle de viste koefficienter på den næste figur sammenlignelige.



17. Kvotienttestteori.

17.1 Test af modellen M mod modellen M_1.

Betragt en model M med m parametre ($\theta_1, \ldots, \theta_m$) og likelihoodfunktion

$$M : L(\theta_1, \ldots, \theta_m) .$$

Betragt dernæst modellen M under hypotesen

$$H_1 : \theta_{k+1} = \ldots = \theta_m = 0 ,$$

dvs. modellen M_1 med parametre ($\theta_1, \ldots, \theta_k, 0, \ldots, 0$) og likelihoodfunktion

$$M_1 : L(\theta_1, \ldots, \theta_k, 0, \ldots, 0) .$$

Spørgsmålet er, om modellen M_1 beskriver data lige så godt som modellen M.

Tænker man i hypoteseprøvningsterminologi, svarer det til, om H_1 kan accepteres. Vurderingen bygger derfor på **kvotienttesstørrelsen** Q_1, defineret som

$$Q_1 = -2\ln\frac{L(\tilde{\theta}_1, \ldots, \tilde{\theta}_k, 0, \ldots, 0)}{L(\hat{\theta}_1, \ldots, \hat{\theta}_m)} ,$$

hvor

$$(\hat{\theta}_1, \ldots, \hat{\theta}_m)$$

er ML-estimatorerne under M, og

$$(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$$

er ML-estimatorerne under M_1, dvs. de værdier, der maksimaliserer

$$M_1 : L(\theta_1, \ldots, \theta_k, 0 \ldots, 0)$$

Testet Q_1 har ifølge den generelle teori for kvotienttesstørrelser en approksimativ χ^2-fordeling med m-k frihedsgrader, jvf. Kapitel 9.

Dette skriver vi som

$$Q_1 \sim \chi^2(m-k) .$$

Vi forkaster derfor hypotesen, dvs. at modellen M_1 beskriver data lige så godt som M, på niveau α, hvis det for den observerede værdi q_1 af Q_1 gælder, at
eller man bestemmer signifikanssandsynligheden

$$q_1 > \chi^2_{1-\alpha}(m-k) ,$$

$$p = P(Q_1 \geq q_1) ,$$

hvor $Q_1 \sim \chi^2(m-k)$. Hvis p er lille, beskriver M_1 ikke data lige så godt som M, mens en stor værdi (f.eks. større end 0.05) betyder, at M_1 beskriver data lige så godt som M.

17.2. Test af modellen M_1 mod en simplere model M_2.

Hvis M_1 beskriver data lige så godt som M, kan man gå videre og teste om en endnu simplere model M_2 kan beskrive data lige så godt som M_1. En simplere model betyder, at der er flere parametre - end i M_1 -, der sættes lig med 0. Det vil sige en model, hvor p flere end k parametre er sat til 0. Det svarer til hypotesen

$$H_2 : \theta_{k-p+1} = ... = \theta_k = \theta_{k+1} = ... = \theta_m = 0 .$$

Her bliver kvotienttesteststørrelsen

$$Q_2^* = -2\ln\frac{L(\tilde{\theta}_1,...,\tilde{\theta}_{k-p},0,...,0)}{L(\hat{\theta}_1,...,\hat{\theta}_k,0,...0)} ,$$

hvor nu

$$(\hat{\theta}_1,...,\hat{\theta}_k)$$

er ML-estimatorerne for $(\theta_1, ... ,\theta_k)$ under M_1, og

$$(\tilde{\theta}_1,...,\tilde{\theta}_{k-p})$$

er ML-estimatorerne for $(\theta_1, ... ,\theta_{k-p})$ under M_2. (Hvorfor jeg skriver Q_2^*, vil fremgå af det følgende.)

Da vi ønsker at vurdere om modellen M_2, der svarer til hypotesen

$$H_2 : \theta_{k-p+1} = ... = \theta_m = 0 ,$$

beskriver data lige så godt som modellen M_1, dvs. **givet at** hypotesen

$$H_1 : \theta_{k+1} = ... = \theta_m = 0$$

allerede er opfyldt, er der, som sagt, p flere θ'er, der sættes lig med 0 under H_2 end under H_1.

Det betyder, at

$$Q_2^* \sim \chi^2(p) .$$

17.3. Test af modellen M_2 mod den oprindelige model M.

Man kan selvfølgelig også teste modellen M_2 mod den oprindelige model M. Da modellen M_2 er en model af samme type som M_1, blot med færre parametre, vil kvotienttesstørrelsen for M_2 mod M være

$$Q_2 = -2\ln\frac{L\left(\tilde{\theta}_1,\ldots,\tilde{\theta}_{k-p},0,\ldots,0\right)}{L\left(\hat{\theta}_1,\ldots,\hat{\theta}_m\right)} .$$

I forhold til M er antallet af θ'er, der er 0 under M_2, lig med m - (k - p) = m - k + p. Derfor gælder

$$Q_2 \sim \chi^2(m-k+p) .$$

17.4. Sekventiel testning.

Sekventiel testning bygger på følgende resultat

$$Q_2 = Q_2^* + Q_1 .$$

Bevis: Vi får

$$Q_2 = -2\ln L\left(\tilde{\theta}_1,\ldots\tilde{\theta}_{k-p},0,\ldots,0\right) + 2\ln L\left(\hat{\theta}_1,\ldots,\hat{\theta}_m\right) =$$

$$-2\ln L\left(\tilde{\theta}_1,\ldots,\tilde{\theta}_{k-p},0,\ldots,0\right) + 2\ln L\left(\hat{\theta}_1,\ldots\hat{\theta}_k,0,\ldots,0\right) + \left(-2\ln L\left(\tilde{\theta}_1,\ldots\tilde{\theta}_k,0\ldots0\right) + 2\ln L\left(\hat{\theta}_1,\ldots,\theta_m\right)\right) ,$$

eller

$$Q_2 = Q_2^* + Q_1 .$$

Læg mærke til, at det passer med frihedsgraderne. Hvis vi kalder frihedsgraderne for M_2 mod M for df_2, frihedsgraderne for M_2 mod M_1 df_2^* og frihedsgraderne for M_1 mod M for df_1, gælder

$$df_2 = m-k+p = p+(m-k) = df_2^* + df_1 .$$

Resultatet skrives ofte som:

$$Q_2^* = Q_2 - Q_1 = \chi_2 (df_2 - df_1) ,$$

Og testet med Q_2^*, som q-teststørrelse kaldes et sekventielt q-test.

Man kalder modellerne M, M_1, M_2 **hierakiske**, fordi de succesivt er indeholdt i hinanden.

Dette skrives ofte som

$$M_2 \subset M_1 \subset M .$$

For hierakiske modeller kan en model, f.eks. M_2, der indeholdt i en anden model, f.eks. M_1, dvs. at i M_2 er èn eller flere af parametrene i M_1 sat lig med 0, testes mod M_1 ved et kvotienttest. I eksemplet her ved kvotienteststørrelsen $Q_2^* = Q_2 - Q_1$. Brugen af differenser mellem q-teststørrelser kaldes **sekventiel testning**.

For at opstille resultaterne, når der testes sekventielt, i en overskuelig tabel, er det nyttigt at indføre betegnelsen **den mættede model** (Eng. 'saturated model') for den model, hvor ingen af parametrene er sat lig med 0. (Dén, der hidtil har heddet M). Det betyder, at kvotienttest-størrelserne for M_1 og M_2 mod den mættede model er Q_2 og Q_1, mens q-teststørrelsen Q_2^* for M_2 mod M_1 er differensen mellem q-teststørrelserne for de to hypotester mod den mættede model.

Det kan være praktisk at opstille skemaet:

Model	Q-størrelse mod mættet model	Antal fri-hedsgrader	Sekventiel teststørrelse	Antal fri-hedsgrader
M_1	Q_1	df_1	-	
M_2	Q_2	df_2	$Q_2 - Q_1$	$df_2 - df_1$

Skemaet kan naturligvis fortsættes. Hvis vi tester 5 modeller både direkte og sekventielt mod den mættede model, får man skemaet:

Model	Q-størrelse mod mættet model	Antal frihedsgrader	Sekventiel teststørrelse	Antal frihedsgrader
M_1	Q_1	df_1	-	
M_2	Q_2	df_2	$Q_2 - Q_1$	$df_2 - df_1$
M_3	Q_3	df_3	$Q_3 - Q_2$	$df_3 - df_2$
M_4	Q_4	df_4	$Q_4 - Q_3$	$df_4 - df_3$
M_5	Q_5	df_5	$Q_5 - Q_4$	$df_5 - df_4$

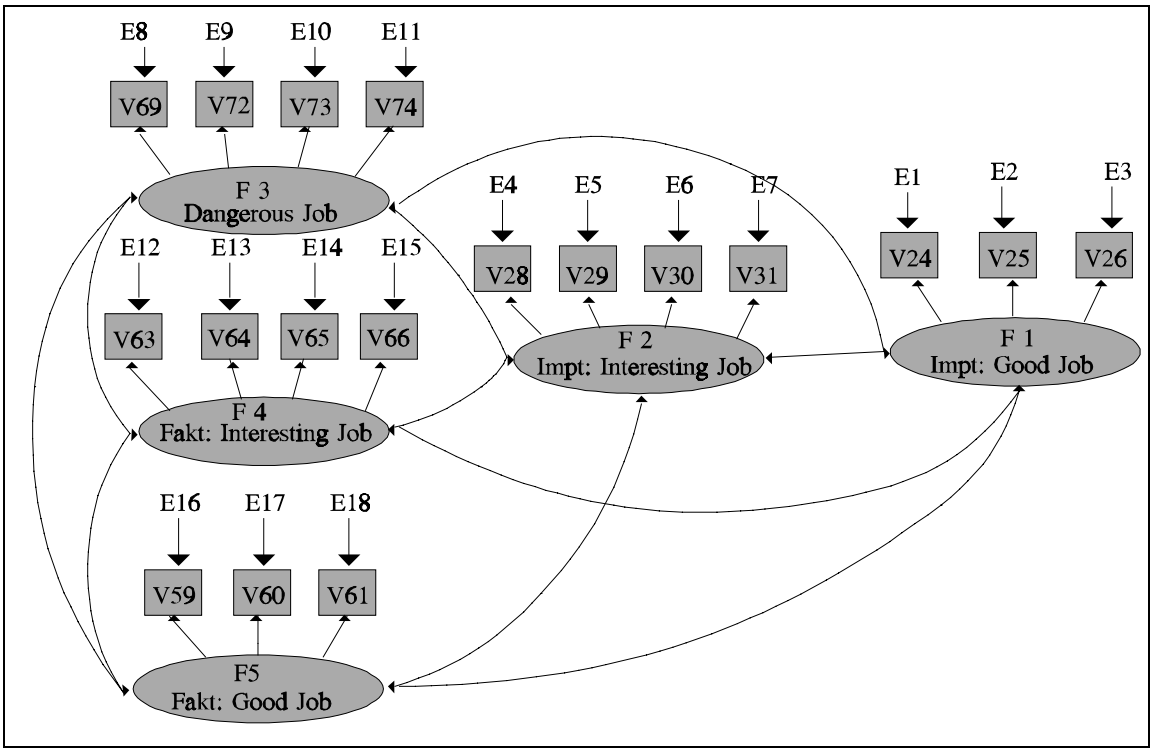
Her vurderes modellerne M_1 til M_5 mod den mættede model ved q-teststørrelserne Q_1 til Q_5, mens de forenklede modeller M_2 til M_5 vurderes mod den forrige model ved differenserne mellem de respektive q-teststørrelser.

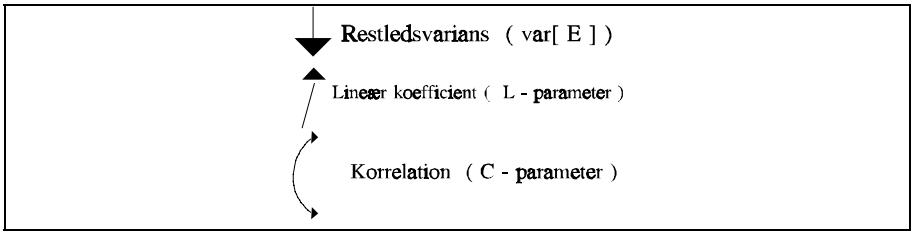
Vigtig bemærkning: *En kvotienttesstørrelses værdi forøges altid, hvis man simplificerer modellen, idet en model med færre parametre naturligvis ikke kan beskrive en model lige så godt som den oprindelige model. Det er derfor ikke nødvendigt at spekulere over, hvilke kvotienttesstørrelser, der skal trækkes fra hinanden. En differens mellem to Q-værdier må nemlig ikke blive negativ. (At kvotienttesstørrelsen ikke forøger sin værdi, når modellen simplificeres, men er lig med den forrige, sker ikke i praksis.)*

18. En målingsmodel (Confirmative Factor Analysis).

Vi har tidligere analyseret de tre variabelgrupper (1) V24 - V26, (2) V28 - V31 og (3) V69 plus V72 - V74, der i stianalysen med manifeste variable optrådte som repræsentanter for de afledede variable (1) 'Good job', (2) 'Interesting job' og (3) 'Dangerous Job'. Vi skal nu tilføje variabelgrupperne (4) V63 - V66 og (5) V59 - V61. Der er tale om de samme spørgsmål, som i grupperne (1) og (2), men hvor man i (1) og (2) spurgte om, hvorvidt de forskellige forhold ved jobbet var væsentlige ('important') efter respondentens opfattelse, spurgte man i grupperne (4) og (5) om forholdene ved arbejdet rent faktisk var, som påstanden var formuleret ('Statement'). Så vi må omdøbe de afledede variable (1) og (2) til 'Impt: Good Job' og 'Impt: Interesting Job'. Herved bliver de 2 afledede variable svarende til (1) og (2), men nu relateret til de faktiske forhold, til (4) 'Fakt: Interesting Job' og (5) 'Fakt: Good Job'. Her har jeg, af uransagelige grunde, byttet om på rækkefølgen og skrevet ordet 'fact' med k på engelsk!

En målingsmodel er en model, der grafisk kan illustreres således:





I formler ser modellen sådan ud.

$$x_1 = b_{11}\, f_1 + e_1$$

$$x_2 = b_{21}\, f_1 + e_2$$

$$x_3 = b_{31}\, f_1 + e_3$$

$$x_4 = b_{12}\, f_2 + e_4$$

$$x_5 = b_{22}\, f_2 + e_5$$

$$x_6 = b_{32}\, f_2 + e_6$$

$$x_7 = b_{42}\, f_2 + e_7$$

$$x_8 = b_{13}\, f_3 + e_8$$

$$x_9 = b_{23}\, f_3 + e_9$$

$$x_{10} = b_{33}\, f_3 + e_{10}$$

$$x_{11} = b_{43}\, f_3 + e_{11}$$

$$x_{12} = b_{14}\, f_4 + e_{12}$$

$$x_{13} = b_{24}\, f_4 + e_{13}$$

$$x_{14} = b_{34}\, f_4 + e_{14}$$

$$x_{15} = b_{44}\, f_4 + e_{15}$$

$$x_{16} = b_{15}\, f_5 + e_{16}$$

$$x_{17} = b_{25}\, f_5 + e_{17}$$

$$x_{18} = b_{35}\, f_5 + e_{18}$$

Her er x'erne de observerbare eller **manifeste** variable, mens f_1 til f_5 er de 5 ikke-observerbare eller **latente** variable. Restledene, e'erne, er uafhængige normalfordelte stokastiske variable, mens b'erne er de koefficienter, der knytter de latente variable sammen med de observede variable. Som det fremgår af figuren, er det yderligere en antagelse, at de V'er, der indgår i definitionen af en latent variabel er uafhængige givet den latente variabel.

Idet kovariansmatricen for V'erne danner udgangspunkt for estimationen af modellens parametre, kan man godt opskrive de 18 ligninger ovenfor på matrix form, som

$$\mathbf{V}_X = \mathbf{B}\, \mathbf{V}_F\, \mathbf{B}^T + \mathbf{E}\ ,$$

hvor $\mathbf{V}_X$ er en 18x18 matrix, dannet af x'ernes varianser og kovarianser. **B** er en 18x5 matrix bestående af b'koefficienterne, $\mathbf{V}_F$ er kovariansmatricen for de 5 latente variable F_1, ... ,F_5. **E** er en diagonalmatrix af varianserne for restleddene E_1, ... ,E_{18}. Men det er en del besværligt, idet der er mange 0'er i **B** som følge af, at de kun er bestemt af de x'er, der er funktioner af en bestemt latent variabel. Heldigvis behøver vi ikke at studere denne matrix relation. For at estimere modellens parametre, skal vi blot i SAS-programmet skrive de 18 ligninger op, stort set som de står oven for.

Men først må vi identificere modellens parametre. Det sker lettest ud fra figuren for modellen, idet hver forbindelseslinie svarer til en parameter. Der er således:

(1) 18 restledsvarianser, svarende til pilene fra restleddene til de observerede variable, dvs. fra et 'E' til et 'V'.

(2) 3 + 4 + 4 + 4 + 3 = 18 koefficienter til de latente variable, svarende til linierne fra V'erne til F'erne.

(3) 10 kovarianser mellem de 5 latente variable, idet der kan trækkes

$$\binom{5}{2} = \frac{5 \cdot 4}{2} = 10$$

linier mellem de 5 ovaler, svarende til F1 til F5.

Der er altså ialt 18 + 18 + 10 = 46 parametre, der skal estimeres i denne **målingsmodel**. Til gengæld er der 19x18/2 = 171 datapunkter i form af kovarianserne mellem de 18 observerede variable, der indgår i modellen. Der er derfor 171 - 46 = 125 frihedsgrader til rådighed til at teste, om modellen, med de estimerede parametre, kan beskrive variationen i de observerede data.

Vi skal til senere brug kalde denne målingsmodel M_m.

De første 32 linier af SAS-programmet til analyse af målingsmodellen, vist neden for, er data-delen, hvor de 18 variable, der indgår i analyse bliver udtrukket af databasen og bliver renset for ubesvarede spørgsmål. Denne del svarer helt til de tidligere viste programmer.

```
1. libname eba 'S:\eba-data';
2. title 'ISSP89: All countries';
3. data a;
4. set eba.d89um;
5. keep v24-v26 v28-v31 v69 v72-v74 v59-v61 v63-v66;
6. data b;
7. set a;
8. if v24>5 then delete;
9. if v25>5 then delete;
10. if v26>5 then delete;
11. if v28>5 then delete;
12. if v29>5 then delete;
13. if v30>5 then delete;
14. if v31>5 then delete;
15. if v69>5 then delete;
16. if v72>5 then delete;
17. if v73>5 then delete;
18. if v74>5 then delete;
19. if v59>5 then delete;
20. if v60>5 then delete;
21. if v61>5 then delete;
22. if v63>5 then delete;
23. if v64>5 then delete;
24. if v65>5 then delete;
25. if v66>5 then delete;
26. if v59=0 then delete;
27. if v60=0 then delete;
28. if v61=0 then delete;
29. if v63=0 then delete;
30. if v64=0 then delete;
31. if v65=0 then delete;
32. if v66=0 then delete;
```

Den næste del af programmet er kaldet af SAS proceduren PROC CALIS, som udfører analysen af de 18 variable som funktioner af 5 latente variable.

Linie 33 er selve kaldet af PROC CALIS, hvor man kan vælge mellem forskellige optioner. Her er valgt 'covariance', der betyder, at det er kovariansmatricen $\mathbf{V}_X$, der bliver analyseret, 'corr', der betyder, at alle korrelationer bliver vist, samt 'residual', der betyder, at residualer og standardiserede residualer mellem den observerede kovariansmatrix $\mathbf{V}_X$ og den - udfra de estimerede parametre - beregnede, bliver vist. Man kan godt udelade optionen 'residual', da den, som i eksemplet her, sjældent yder noget væsentligt bidrag til fortolkningen af resultaterne.

Linie 34 til 52, med overskriften 'lineqs', er de 19 linier, der definerer de 5 latente variable, svarende til ligningerne oven for, så f.eks. $x_1 = b_{11}\, f_1 + e_1$ svarer til linie 35

```
v24 = lv24f1 f1 +e1 .
```

```
33. proc calis covariance corr residual;
34.  lineqs
35.   v24 = lv24f1 f1 +e1,
36.   v25 = lv25f1 f1 +e2,
37.   v26 = lv26f1 f1 +e3,
38.   v28 = lv28f2 f2 +e4,
39.   v29 = lv29f2 f2 +e5,
40.   v30 = lv30f2 f2 +e6,
41.   v31 = lv31f2 f2 +e7,
42.   v69 = lv69f3 f3 +e8,
43.   v72 = lv72f3 f3 +e9,
44.   v73 = lv73f3 f3 +e10,
45.   v74 = lv74f3 f3 +e11,
46.   v63 = lv63f4 f4 +e12,
47.   v64 = lv64f4 f4 +e13,
48.   v65 = lv65f4 f4 +e14,
49.   v66 = lv66f4 f4 +e15,
50.   v59 = lv59f5 f5 +e16,
51.   v60 = lv60f5 f5 +e17,
52.   v61 = lv61f5 f5 +e18;
53.  std
54.   f1=1,
55.   f2=1,
56.   f3=1,
57.   f4=1,
58.   f5=1,
59.   e1-e18 = vare1-vare18;
60.  cov
61.   f1 f2 = cf1f2,
62.   f1 f3 = cf1f3,
63.   f1 f4 = cf1f4,
64.   f1 f5 = cf1f5,
64.   f2 f3 = cf2f3,
65.   f2 f4 = cf2f4,
66.   f2 f5 = cf2f5,
67.   f3 f4 = cf3f4,
68.   f3 f5 = cf3f5,
69.   f4 f5 = cf4f5;
70.  var v24-v26 v28-v31 v69 v72-v74 v63-v66 v59-v61;
71. run;
72. quit.
```

Bemærk, at b-koefficienterne her er skrevet som et 'l', der står for 'lineær' efterfulgt af navnet på både x-variablen og f-variablen. Størrelserne 'lv24f1' til 'lv61f5' er altså de 18 b-parametre. Dernæst følger i linie 53 til 59 definitionen af de variansparametre, der indgår i modellen. De 18 restledsvarianser kaldes 'e1' til 'e18'. Varianserne til de 5 latente variable 'f1' til 'f5' sættes i linierne 54 til 58 lig med 1. Dette skyldes det, at man i produktled, som b_{ij} f_j enten må normere b'erne eller f'erne, ellers er modellen ikke identificerbar. Her bliver f'erne normeret ved at sætte f'ernes varianser til 1. I linie 60 til 69 defineres de 10 kovarianser i modellen, nemlig alle kovarianser mellem de 5 latente variable. Her bruges navnene 'cf1f2'

til ’cf4f5’, hvor ’c’ står for kovarians og efterfølges af navnene på de indgående variable. Endelig i linie 70 står hvilke 18 variable, der indgår i analysen.

Man behøver ikke i SAS at bruge så ’lange’ koefficientnavne som f.eks. ’LV24F1’, men det har den fordel, at man kan holde rede på, hvilke to variable, der er tale om. Dog er ’l’-et obligatorisk, så SAS véd, at der er tale om en lineær regressionskoefficient. Hvis man kan holde rede på, hvad der er hvad, kan man godt blot skrive ’L1’ til ’L18’, så ’L1’ svarer til ’LV24F1’, osv.

Den første boks neden for viser den del af SAS-output’et, hvor de 18 definitionsligninger er angivet på symbolsk form, så man kan checke, at alt er med. At der skrives et 1-tal foran fejlleddet er blot et udtryk for, at PROC CALIS skal kunne håndtere mange forskellige modeller med strukturelle ligninger. I målingsmodellen, og i stimodellerne i næste kapitel, er koefficienterne til fejlledene lig med 1, men det betyder ikke at E’ernes varianser er ens. Tværtimod optræder kan man se af ’std’-afsnittet i SAS-programmet, at alle varianser på restled optræder som parametre i modellen, benævnt ’vare1’ til ’vare18’.

Endelig vises i de 3 sidste linier af output-boksen, at der 6776 observationer, 18 variable, 171 elementer i $\mathbf{V}_X$ og 46 parametre.

```
Manifest Variable Equations

 V24     =      .     *F1      + 1.0000 E1
                       LV24F1

 V25     =      .     *F1      + 1.0000 E2
                       LV25F1

 V26     =      .     *F1      + 1.0000 E3
                       LV26F1

 V28     =      .     *F2      + 1.0000 E4
                       LV28F2

 V29     =      .     *F2      + 1.0000 E5
                       LV29F2

 V30     =      .     *F2      + 1.0000 E6
                       LV30F2

 V31     =      .     *F2      + 1.0000 E7
                       LV31F2

 V69     =      .     *F3      + 1.0000 E8
                       LV69F3

 V72     =      .     *F3      + 1.0000 E9
                       LV72F3

 V73     =      .     *F3      + 1.0000 E10
                       LV73F3

 V74     =      .     *F3      + 1.0000 E11
                       LV74F3

 V63     =      .     *F4      + 1.0000 E12
                       LV63F4

 V64     =      .     *F4      + 1.0000 E13
                       LV64F4

 V65     =      .     *F4      + 1.0000 E14
                       LV65F4

 V66     =      .     *F4      + 1.0000 E15
                       LV66F4

 V59     =      .     *F5      + 1.0000 E16
                       LV59F5

 V60     =      .     *F5      + 1.0000 E17
                       LV60F5

 V61     =      .     *F5      + 1.0000 E18
                       LV61F5

    6776 Observations       Model Terms          1
      18 Variables          Model Matrices       4
     171 Informations       Parameters          46
```

```
Dual Quasi-Newton Optimization
Dual Broyden - Fletcher - Goldfarb - Shanno Update (DBFGS)
Number of Parameter Estimates 46
Number of Functions (Observations) 171

Optimization Start: Active Constraints= 0  Criterion= 1.010 Maximum Gradient Element= 0.302

 Iter rest nfun act   optcrit  difcrit maxgrad   alpha   slope
    1    0    4   0    0.9202   0.0902   0.477  0.0402  -7.167
    2    0    6   0    0.8133   0.1069   0.183   0.803  -0.239
    3    0    8   0    0.7796   0.0337   0.139   0.372  -0.138
    4    0    9   0    0.7539   0.0257  0.0715   0.745 -0.0681
    5    0   10   0    0.7461  0.00786  0.0886   1.000 -0.0160
    6    0   11   0    0.7418  0.00426  0.0646   1.000 -0.0134
    7    0   12   0    0.7402  0.00166  0.0548   1.000 -0.0066
    8    0   13   0    0.7391  0.00103  0.0257   1.000  -0.003
    9    0   14   0    0.7389 0.000217  0.0169   1.000 -0.0008
   10    0   15   0    0.7389 0.000035  0.0174   1.000 -0.0003
   11    0   16   0    0.7388 0.000042 0.00847   1.000 -0.0002
   12    0   17   0    0.7388 0.000043 0.00532   1.000 -0.0001
   13    0   18   0    0.7388 4.125E-6 0.00486   1.000 -263E-7
   14    0   19   0    0.7388 4.529E-6 0.00188   1.000 -158E-7
   15    0   20   0    0.7388 2.129E-6  0.0009   1.000 -538E-8
   16    0   21   0    0.7388 5.406E-7 0.00029   1.000 -107E-8
   17    0   22   0    0.7388 2.397E-7 0.00036   1.000 -442E-9
   18    0   23   0    0.7388 1.297E-7 0.00017   1.000 -235E-9
   19    0   24   0    0.7388 2.608E-8 0.00008   1.000  -54E-9
   20    0   26   0    0.7388 1.687E-8 0.00011   2.000  -19E-9
   21    0   27   0    0.7388 9.746E-9 0.00003   1.000  -18E-9
   22    0   28   0    0.7388 2.394E-9 0.00001   1.000 -43E-10

Optimization Results: Iterations= 22 Function Calls= 29 Gradient Calls= 26
Active Constraints= 0  Criterion= 0.73877962 Maximum Gradient Element= 0.000014872
Slope= -4.29007E-9

NOTE:  GCONV convergence criterion satisfied.
```

I denne boks fra SAS output'et er vist en række tekniske detaljer fra den iterative løsning af likelihoodligningerne. De første 4 linier viser, hvilken iterationsmetode, der er benyttet. Som man kan se, er det en tillempet Newton iterationsmetode, som der åbenbart har været flere om at tillempe. Dernæst er der en række, ligeledes tekniske, detaljer vedrørende de 22 iterationer, der er nødvendige. Den eneste størrelser, der her har en vis interesse er 'optcrit', som - på nær et par konstanter - er lig med -2 gange log-likelihoodfunktionen. Kriteriet er derfor minimum, hvis likelihoodfunktionen er maksimum og 0 hvis der er fuldstændig overensstemmelse mellem den observerede og den estimerede kovariansmatrix. Denne fortolkning af 'optcrit' er nærmere gennemgået i Kapitel 11. De øvrige oplysninger har kun interesse for specialister, og især i de tilfælde, hvor proceduren ikke konvergerer.

I dette tilfælde er der konvergens, hvilket noteres med 'NOTE: GCONV convergence criterion satisfied.'

```
Fit criterion . . . . . .     0.7388
Goodness of Fit Index (GFI) . . . . . .     0.9144
GFI Adjusted for Degrees of Freedom (AGFI). . . .     0.8829
Root Mean Square Residual (RMR) . . . . . .     0.0551
Parsimonious GFI (Mulaik, 1989) . . . . . .     0.7470

Chi-square = 5005.2319     df = 125     Prob>chi**2 = 0.0001
Null Model Chi-square:     df = 153               25973.7793

RMSEA Estimate  . . . . .  0.0759  90%C.I.[0.0741, 0.0777]
Probability of Close Fit  . . . . . .     .
ECVI Estimate . . . . . .  0.7524  90%C.I.[0.7188, 0.7871]
Bentler's Comparative Fit Index . . . . . .     0.8110
Normal Theory Reweighted LS Chi-square  . . . . .  5709.2068
Akaike's Information Criterion. . . . . .  4755.2319
Bozdogan's (1987) CAIC. . . . . .  3777.5892
Schwarz's Bayesian Criterion. . . . . .  3902.5892
McDonald's (1989) Centrality. . . . . .     0.6976
Bentler & Bonett's (1980) Non-normed Index. . . .     0.7687
Bentler & Bonett's (1980) NFI . . . . . .     0.8073
James, Mulaik, & Brett (1982) Parsimonious NFI. .     0.6596
Z-Test of Wilson & Hilferty (1931). . . . . .    57.4646
Bollen (1986) Normed Index Rho1 . . . . . .     0.7641
Bollen (1988) Non-normed Index Delta2 . . . . . .     0.8112
Hoelter's (1983) Critical N . . . . . .        207
```

I denne boks vises en lang række mål for, hvor god tilpasningen mellem model og data er, startende med 'Fit criterion', som er identisk med 'optcrit' fra den forrige boks. SAS-programmer har den tradition, at man hellere angiver for mange, end for få, statistiske størrelser. Her har det taget en hel del overhånd. Vi skal koncentrere os om de 2 linier, der er fremhævet [af mig] i output'et. Størrelsen 5005.2319 med df = 125 er det transformerede kvotienttestskøn for sammenligningen af den mættede model, hvor parametrene er alle de 171 kovarianser i $\mathbf{V}_X$ og målingsmodellen med 46 parametre. Antal frihedsgrader 'df' er altså df = 171 - 46 = 125. Den viste q-teststørrelse er klart signifikant. SAS skriver '= 0.0001' både hvis signifikanssandsynligheden er lig 0.0001 efter afrunding, og når signifikanssandsynligheden er < 0.00005.

Den model, der kaldes 'Null Model', er en model, hvor der ingen korrelationer er, hverken mellem latente, mellem manifeste eller mellem latente og manifeste variable. Denne model har 171 - 153 = 18 frihedsgrader, da det alene er de 18 restledsvarianser, der er parametre i modellen. Hvorfor SAS insisterer på at gengive q-teststørrelsen for denne model i alle output er mig en gåde. Det må vist høre til de helt store sjældenheder, at man kan få brug for at sammenligne en model med denne trivielle model. Men det er nok typisk for programmørerne i SAS, at de har en svaghed for at inkludere en række mere eller mindre trivielle nøgletal for modellerne, selv om disse nøgletal sjældent vil blive brugt i praksis.

I den næste boks vises alle de estimerede lineære koefficienter - b'erne - sammen med deres standardfejl og kvotienten mellem estimatet og standardfejlen. Denne kvotient 't Value' er approksimativt normeret normalfordelt i store stikprøver. Dvs. at estimatet for en b-parameter, eller L - parameter, er signifikant forskellig fra 0, hvis værdien af 't Value' er større end ca. 2. Det er en del misvisende, at SAS altid skriver 't Value' også hvis størrelse, som det er tilfældet her, ikke er t-fordelt. Men sådan gør man altså! Tallene i boksen viser, at alle koefficienter er signifikante, dvs. at alle de latente variable bidrager signifikant til at forklare de manifeste variables variation. Det er naturligvis betryggende, da en ikke-signifikant koefficient jo ville betyde, at den pågældende variabel ikke skulle medtages i modellen.

```
V24      =     0.2476*F1     +  1.0000 E1
Std Err       0.0110 LV24F1
t Value      22.5496

V25      =     0.3858*F1     +  1.0000 E2
Std Err       0.0130 LV25F1
t Value      29.7263

V26      =     0.5474*F1     +  1.0000 E3
Std Err       0.0166 LV26F1
t Value      32.9705

V28      =     0.2355*F2     +  1.0000 E4
Std Err       0.0085 LV28F2
t Value      27.5853

V29      =     0.3201*F2     +  1.0000 E5
Std Err       0.0110 LV29F2
t Value      29.1305

V30      =     0.7155*F2     +  1.0000 E6
Std Err       0.0113 LV30F2
t Value      63.0401

V31      =     0.6742*F2     +  1.0000 E7
Std Err       0.0112 LV31F2
t Value      60.3099

V69      =     0.6926*F3     +  1.0000 E8
Std Err       0.0159 LV69F3
t Value      43.4694

V72      =     0.8716*F3     +  1.0000 E9
Std Err       0.0142 LV72F3
t Value      61.2299

V73      =     0.8906*F3     +  1.0000 E10
Std Err       0.0138 LV73F3
t Value      64.4274

V74      =     0.7501*F3     +  1.0000 E11
Std Err       0.0149 LV74F3
t Value      50.4350

V63      =     0.5187*F4     +  1.0000 E12
Std Err       0.0122 LV63F4
t Value      42.4859

V64      =     0.4218*F4     +  1.0000 E13
Std Err       0.0135 LV64F4
t Value      31.3003

V65      =     0.7612*F4     +  1.0000 E14
Std Err       0.0136 LV65F4
t Value      55.9705

V66      =     0.6757*F4     +  1.0000 E15
Std Err       0.0127 LV66F4
t Value      53.2188

V59      =     0.4654*F5     +  1.0000 E16
Std Err       0.0159 LV59F5
t Value      29.3291

V60      =     0.6690*F5     +  1.0000 E17
Std Err       0.0166 LV60F5
t Value      40.3543

V61      =     0.7164*F5     +  1.0000 E18
Std Err       0.0179 LV61F5
t Value      40.0005
```

Covariances among Exogenous Variables

Parameter			Estimate	Standard Error	t Value
F2	F1	CF1F2	0.321473	0.017606	18.259
F3	F1	CF1F3	0.028704	0.018305	1.568
F3	F2	CF2F3	0.011339	0.015531	0.730
F4	F1	CF1F4	0.073583	0.019086	3.855
F4	F2	CF2F4	0.488811	0.013579	35.998
F4	F3	CF3F4	-0.025022	0.016056	-1.558
F5	F1	CF1F5	0.199822	0.020139	9.922
F5	F2	CF2F5	0.082211	0.017288	4.755
F5	F3	CF3F5	-0.130227	0.017001	-7.660
F5	F4	CF4F5	0.313064	0.016933	18.488

I denne boks er de estimerede kovarianser CF1F2 til CF4F5 mellem de latente variable F1 til F5 vist. Også standardfejlene og de standardiserede estimater, dvs. 't Value' er vist. Som nævnt er et estimat signifikant forskelligt fra 0, hvis 't Value' er større end ca. 2. At et estimat ikke er signifikant forskelligt fra 0 betyder, at korrelationen mellem de to pågældende variable kan sættes til 0 uden at forringe modellens forklaringsevne. Her gælder det alle de tre korrelationer, hvor F3 indgår.

I den næste boks er vist de standardiserede koefficienter, som man vanligvis indskriver på grafen for at vise størrelsesordenen af den pågældende sammenhæng. For estimaterne af de lineære parametre (L - parametrene) er estimaterne standardiserede så variansen på alle de variable både de manifeste (V'erne) og de latente (F'erne) er lig med 1. Det betyder, ifølge Kapitel 12 at de er lig med korrelationskoefficienterne mellem V'erne og F'erne. De standardiserede koefficienter er altså sammenlignelige. Bemærk, at da både V'erne og F'erne standardiseres til varians 1, vil varianserne på E'erne naturligvis også ændre sig.

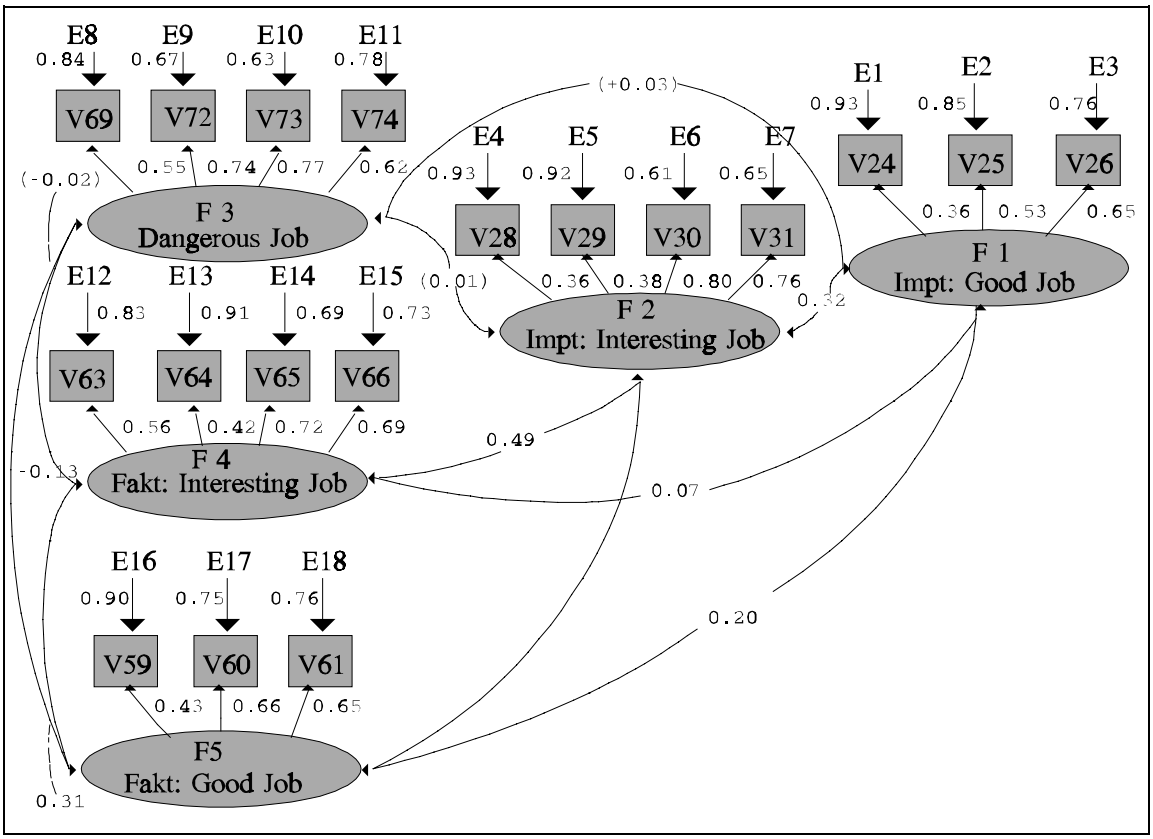
```
Equations with Standardized Coefficients

 V24      =     0.3629*F1      + 0.9318 E1
                        LV24F1

 V25      =     0.5269*F1      + 0.8499 E2
                        LV25F1

 V26      =     0.6473*F1      + 0.7622 E3
                        LV26F1

 V28      =     0.3650*F2      + 0.9310 E4
                        LV28F2

 V29      =     0.3841*F2      + 0.9233 E5
                        LV29F2

 V30      =     0.7958*F2      + 0.6055 E6
                        LV30F2

 V31      =     0.7608*F2      + 0.6489 E7
                        LV31F2

 V69      =     0.5489*F3      + 0.8359 E8
                        LV69F3

 V72      =     0.7395*F3      + 0.6732 E9
                        LV72F3

 V73      =     0.7741*F3      + 0.6331 E10
                        LV73F3

 V74      =     0.6233*F3      + 0.7820 E11
                        LV74F3

 V63      =     0.5566*F4      + 0.8308 E12
                        LV63F4

 V64      =     0.4220*F4      + 0.9066 E13
                        LV64F4

 V65      =     0.7192*F4      + 0.6948 E14
                        LV65F4

 V66      =     0.6852*F4      + 0.7284 E15
                        LV66F4

 V59      =     0.4332*F5      + 0.9013 E16
                        LV59F5

 V60      =     0.6581*F5      + 0.7529 E17
                        LV60F5

 V61      =     0.6488*F5      + 0.7609 E18
```

I den sidste boks er korrelationskoefficienterne mellem de 5 latente variable vist. Også her af hensyn til en grafisk fremstilling, hvor man kan skrive disse korrelationer ind.

Correlations among Exogenous Variables

Parameter			Estimate
F2	F1	CF1F2	0.311122
F3	F1	CF1F3	0.054118
F3	F2	CF2F3	0.008131
F4	F1	CF1F4	0.047377
F4	F2	CF2F4	0.492078
F4	F3	CF3F4	-0.018468
F5	F1	CF1F5	0.192202
F5	F2	CF2F5	0.001949
F5	F3	CF3F5	0.070247
F5	F4	CF3F5	0.070247

I figuren neden for er alle de standardiserede lineære koefficienter og alle korrelationerne indtegnet på stifiguren, baseret på tallene i de to sidste output-bokse.

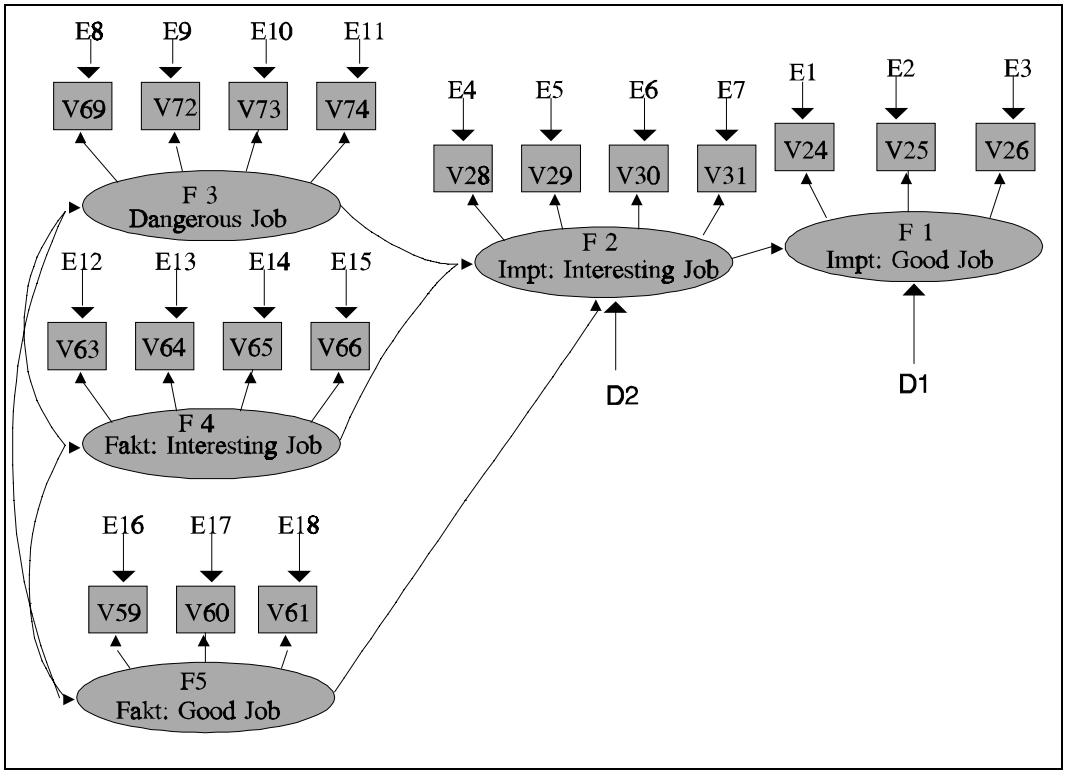


Figuren skal læses på den måde, at høje korrelationer tyder på en stærk sammenhæng, mens små korrelationer tyder på små (men dog signifikante) sammenhænge.

19. Stianalyse med latente variable.

Ved en stianalyse med latente variable specificeres både de lineære ligninger, som binder de latente og de manifeste variable sammen, men også ligninger, der lader nogle af de latente variable være funktioner af nogle af de øvrige. Navnet 'stianalyse' skal illustere, at nogle af sammenhængene 'peger fremad' fra nogle variable til andre 'afledte' variable. Det er ikke ualmindeligt, at sådanne relationer opfattes som kausale sammenhænge, men det behøver ikke at være tilfældet. Man skal under alle omstændigheder være omhyggelig med at fortolke sammenhænge som kausale - alene udfra statistiske analyser af relationer mellem variable.

Grafen neden for viser en stimodel, der bygger på de samme variable, som målingsmodellen i Kapitel 18. Vi skal somme tider kalde en sådan graf et **stidiagram**.



Her er det de to spørgsmål vedrørende (F1), hvor væsentligt man synes ens job er med hensyn til løn, avancementsmuligheder mv. ('Good Job'), og (F2), hvor væsentligt man synes ens job er, med hensyn til at være til gavn for samfundet, hjælpe andre mennesker, mv. ('Interesting Job'), der opfattes som 'afledte' variable. Vi skal forsøge at forklare disse variable som afledte af de 3 andre latente variable (F3), hvor farligt jobbet er (Dangeous Job), F4 hvordan ens job rent faktisk er med hensyn til at være til gavn for samfundet, hjælpe andre mennesker, mv. ('Fakt. Interesting Job') og (F5) med hensyn til løn, avancementsmuligheder, mv. ('Fakt. Good Job').

Læg mærke til, at linierne fra de 3 'baggrundsvariable' F3, F4 og F5 til de 'afledte' variable F2 og F1 nu kun har en pil i en den ene ende, nemlig den, der viser fremad fra 'baggrunds-variablen' til den 'afledte variabel'.

SAS-programmet til analyse af denne model ser således ud:

```
1. libname eba 's:\eba-data';
2. title 'ISSP89: All countries';
3. data a;
4. set eba.d89um;
5. keep v24-v26 v28-v31 v69 v72-v74 v59-v61 v63-v66;
6. data b;
7. set a;
8. if v24>5 then delete;
9. if v25>5 then delete;
.....

10. if v66=0 then delete;
11. proc calis covariance corr residual;
12.   lineqs
13.    v24 = lv24f1 f1 + e1,
14.    v25 = lv25f1 f1 + e2,
15.    v26 = f1 + e3,
16.    v28 = lv28f2 f2 + e4,
17.    v29 = lv29f2 f2 + e5,
18.    v30 = lv30f2 f2 + e6,
19.    v31 = f2 + e7,
20.    v69 = lv69f3 f3 + e8,
21.    v72 = lv72f3 f3 + e9,
22.    v73 = lv73f3 f3 + e10,
23.    v74 = f3 + e11,
24.    v63 = lv63f4 f4 + e12,
25.    v64 = lv64f4 f4 + e13,
26.    v65 = lv65f4 f4 + e14,
27.    v66 = f4 + e15,
28.    v59 = lv59f5 f5 + e16,
29.    v60 = lv60f5 f5 + e17,
30.    v61 = f5 + e18,
31.    f1 = pf1f2 f2 + d1,
32.    f2 = pf2f3 f3 + pf2f4 f4 + pf2f5 f5 + d2;
33.   std
34.    f3 - f5 = varf3 - varf5,
35.    d1-d2 = vard1 - vard2,
36.    e1-e18 = vare1-vare18;
37.   cov
38.    f3 f4 = cf3f4,
39.    f3 f5 = cf3f5,
40.    f4 f5 = cf4f5;
41.   var v24-v26 v28-v31 v69 v72-v74 v63-v66 v59-v61;
42.  run;
43.  quit;
```

Linie 1-10 svarer til de tidligere programmer, hvor data indlæses fra databasen og renses for ubesvarede spørgsmål. I linie 11-41 følger så selve PROC CALIS-kaldet. Det adskiller sig fra det tilsvarende program for målingsmodellen på flere måder.

For det første er der kommet to nye ligninger til, nemlig linie 31 og linie 32, hvor de to lineære relationer mellem 'baggrundsvariablene' F3 til F5 og de afledte variable F1 og F2 står. De lineære koefficienter betegnes her ved forbogstavet 'p' for 'path' og de to restled kaldes D1 og D2 for at adskille dem fra restleddene i ligningerne linie 13 - 30.

Det er nødvendigt at lave en ny normering af de manifeste variable. Dette skyldes at to af de latente variable F1 og F2 optræder både i ligningerne linie 13 til 30 og i de to ligninger, der definerer F1 og F2. Man vælger derfor at normere F'erne, så L-koefficienten for den sidste af V-variablene i hver gruppe, omfattende de samme latente variable, er lig med 1. Det betyder, at ligningerne for f.eks. V26 og V31 bliver

$$v26 = f1 + e3,$$

$$v31 = f2 + e7,$$

Til gengæld må vi indføre varianserne for de 3 latente 'baggrundsvariable' som parametre i modellen. Det sker i 'std'-afsnittet som linie 34, hvor de navngives 'varf3' til 'varf5'.

Varianserne af de to 'afledte' latente variable bestemmes af de nye restled D1 og D2 i linie 31 og 32, så vi må også i 'std'-afsnittet linie 35 definere D1 og D2's varianser som parametre med navnene 'vard1' og 'vard2'.

Vi skal kalde denne **'basis' stimodel** for M_{p0}.

Den første vigtige del af output viser de indgående ligninger på symbolsk form plus nogle grundlæggende tal, f.eks. at der er 18 variable, 171 datapunkter og 43 parametre. Ligesom ved målingsmodellen skrives der ’1.000’ foran alle restled.

```
Manifest Variable Equations

 V24      =      .     *F1      + 1.0000 E1
                         LV24F1

 V25      =      .     *F1      + 1.0000 E2
                         LV25F1

 V26      =    1.0000 F1      + 1.0000 E3

 V28      =      .     *F2      + 1.0000 E4
                         LV28F2

 V29      =      .     *F2      + 1.0000 E5
                         LV29F2

 V30      =      .     *F2      + 1.0000 E6
                         LV30F2

 V31      =    1.0000 F2      + 1.0000 E7

 V69      =      .     *F3      + 1.0000 E8
                         LV69F3

 V72      =      .     *F3      + 1.0000 E9
                         LV72F3

 V73      =      .     *F3      + 1.0000 E10
                         LV73F3

 V74      =    1.0000 F3      + 1.0000 E11

 V63      =      .     *F4      + 1.0000 E12
                         LV63F4

 V64      =      .     *F4      + 1.0000 E13
                         LV64F4

 V65      =      .     *F4      + 1.0000 E14
                         LV65F4

 V66      =    1.0000 F4      + 1.0000 E15

 V59      =      .     *F5      + 1.0000 E16
                         LV59F5

 V60      =      .     *F5      + 1.0000 E17
                         LV60F5

 V61      =    1.0000 F5      + 1.0000 E18

 F1       =      .     *F2      + 1.0000 D1
                         PF1F2

 F2       =      .     *F3      +  .     *F4      +  .     *F5      + 1.0000 D2
                         PF2F3             PF2F4              PF2F5

    6776 Observations       Model Terms          1
      18 Variables          Model Matrices       4
     171 Informations       Parameters          43
```

```
Fit criterion . . . . . .     0.7559
Goodness of Fit Index (GFI) . . . . . .     0.9131
GFI Adjusted for Degrees of Freedom (AGFI). . . .     0.8840
Root Mean Square Residual (RMR) . . . . . .     0.0567
Parsimonious GFI (Mulaik, 1989) . . . . . .     0.7639

Chi-square = 5121.0244     df = 128     Prob>chi**2 = 0.0001
Null Model Chi-square:     df = 153               25973.7793

RMSEA Estimate  . . . . . .  0.0759  90%C.I.[0.0741, 0.0777]
Probability of Close Fit  . . . . . .      .
ECVI Estimate . . . . . .  0.7686  90%C.I.[0.7346, 0.8037]
Bentler's Comparative Fit Index . . . . . .     0.8066
Normal Theory Reweighted LS Chi-square  . . . . .  5799.5037
Akaike's Information Criterion. . . . . .  4865.0244
Bozdogan's (1987) CAIC. . . . . .  3863.9182
Schwarz's Bayesian Criterion. . . . . .  3991.9182
McDonald's (1989) Centrality. . . . . .     0.6918
Bentler & Bonett's (1980) Non-normed Index. . . .     0.7689
Bentler & Bonett's (1980) NFI . . . . . .     0.8028
James, Mulaik, & Brett (1982) Parsimonious NFI. .     0.6717
Z-Test of Wilson & Hilferty (1931). . . . . .    58.1260
Bollen (1986) Normed Index Rho1 . . . . . .     0.7643
Bollen (1988) Non-normed Index Delta2 . . . . . .     0.8068
Hoelter's (1983) Critical N . . . . . .        207
```

Her er vist alle de mange teststørrelser og indeks for modeltilpasning. Igen har jeg fremhævet de to linier, vi skal bruge. Vi ser at modeltilpasningen ved et Q-test er beskrevet af q = 5121.02 med 128 frihedsgrader, som er endog meget signifikant. Bemærk, at SAS skriver signifikanssandsynligheden som p = 0.0001, selv om den faktiske værdi er langt mindre end 0.00005.

I næste boks er vist alle de estimerede ligninger med 't Value' for alle lineære koefficienter. Alle L-koefficienter for de manifeste variable er signifikante, som vi skulle forvente. Men bemærk, at fordi vi har omnormeret de latente variable, får vi intet signifikanstest for koefficienterne til V26, V31, V74, V61 og V66. Til gengæld ser vi, at P-koefficienten til F3 i ligningen for F2 ikke er signifikant. Dette betyder, at én af de modelmodificeringer, der bør overvejes, er at fjerne F3 fra ligningen, der definerer F2.

```
Manifest Variable Equations

 V24     =     0.5296*F1     +  1.0000 E1
 Std Err       0.0289 LV24F1
 t Value      18.3273
```

Den næste boks viser de estimerede korrelationer mellem de 3 latente 'baggrundsvariable' F3 til F5, sammen med værdierne af 't Values'. Tallene viser en ikke signifikant korrelation mellem F3 og F4.

Den næste modelmodifikation, der bør overvejes, er derfor at fjerne denne korrelation, svarende til den dobbelte pil mellem F3 og F4 på stidiagrammet.

Covariances among Exogenous Variables

Parameter			Estimate	Standard Error	t Value
F4	F3	CF3F4	-0.012483	0.008161	-1.530
F5	F3	CF3F5	-0.068911	0.009271	-7.433
F5	F4	CF4F5	0.145906	0.009519	15.328

Den næste boks viser de standardiserede koefficienter.

```
Equations with Standardized Coefficients

 V24      =     0.3938*F1      + 0.9192 E1
                       LV24F1

 V25      =     0.5565*F1      + 0.8309 E2
                       LV25F1

 V26      =     0.5999 F1      + 0.8000 E3

 V28      =     0.3615*F2      + 0.9324 E4
                       LV28F2

 V29      =     0.3828*F2      + 0.9238 E5
                       LV29F2

 V30      =     0.7993*F2      + 0.6009 E6
                       LV30F2

 V31      =     0.7611 F2      + 0.6487 E7

 V69      =     0.5485*F3      + 0.8362 E8
                       LV69F3

 V72      =     0.7393*F3      + 0.6734 E9
                       LV72F3

 V73      =     0.7742*F3      + 0.6330 E10
                       LV73F3

 V74      =     0.6237 F3      + 0.7816 E11

 V63      =     0.5550*F4      + 0.8319 E12
                       LV63F4

 V64      =     0.4199*F4      + 0.9076 E13
                       LV64F4

 V65      =     0.7207*F4      + 0.6932 E14
                       LV65F4

 V66      =     0.6861 F4      + 0.7275 E15

 V59      =     0.4317*F5      + 0.9020 E16
                       LV59F5

 V60      =     0.6744*F5      + 0.7384 E17
                       LV60F5

 V61      =     0.6351 F5      + 0.7724 E18

 F1       =     0.3033*F2      + 0.9529 D1
                       PF1F2

 F2       =     0.0173*F3      + 0.5042*F4      - 0.0652*F5      + 0.8727 D2
                       PF2F3            PF2F4            PF2F5
```

I Kapitel 12 diskuterede vi definitionen og betydningen af denne omskalering af L- og P- koefficienterne. De viste koefficienter skal således nu opfattes som korrelationer og er derfor sammenlignelige i størrelsesforhold. Sammenhængen mellem f.eks. V30 og F2 er altså - alt andet lige - meget stærkere end sammenhængen mellem V24 og F1.

```
Correlations among Exogenous Variables
-----------------------------------
     Parameter            Estimate
-----------------------------------
F4      F3      CF3F4        -0.024579
F5      F3      CF3F5        -0.130897
F5      F4      CF4F5         0.307511
```

I denne sidste boks vises korrelationerne mellem de latente 'baggrundsvariable'. Som for målingsmodellen bruges disse først og fremmest til at skrive ind i en graf, hvis man ønsker, at den skal vise 'styrken' af de forskellige sammenhænge målt ved korrelationskoefficienter. Den insignifikante sammenhæng mellem F3 og F4 afspejles dog klart ved den meget lille korrelation på -0.025 mellem de to variable.

Hele PROC CALIS - programdelen får derfor udseendet:

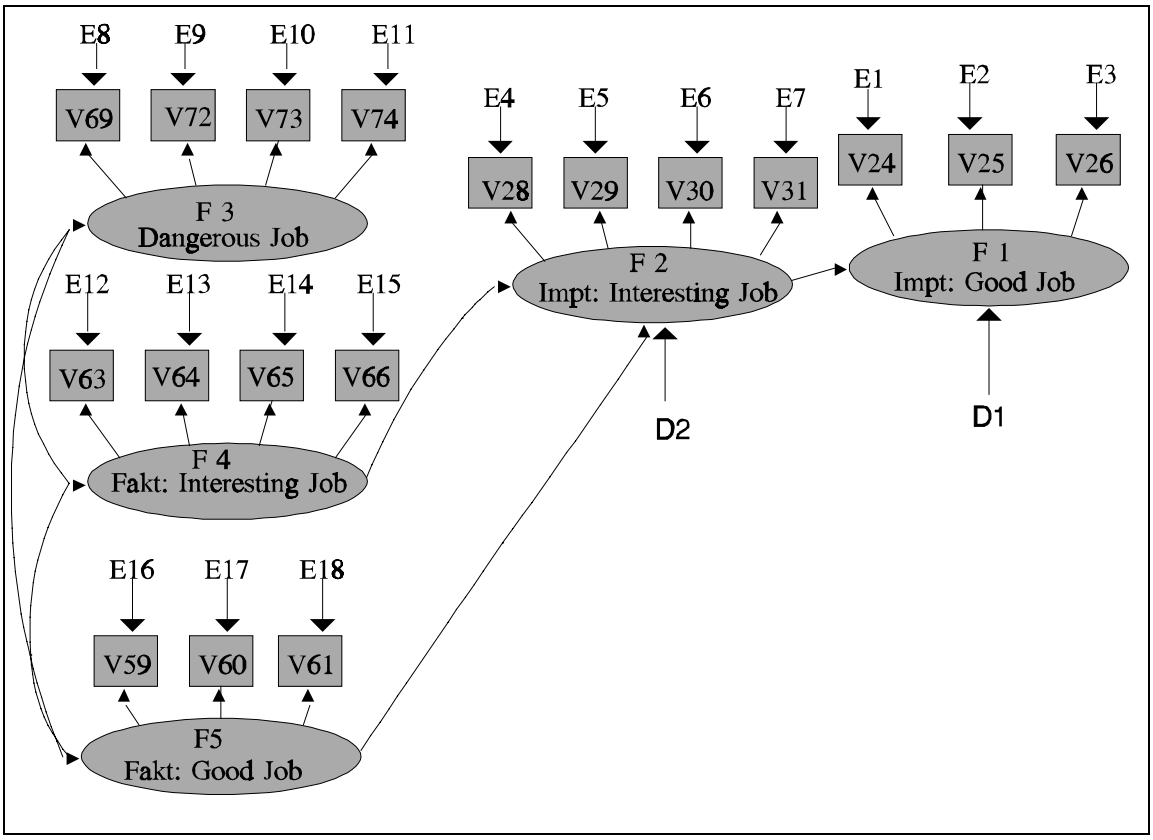
```
proc calis covariance corr residual;
  lineqs
   v24 = lv24f1 f1 + e1,
   .....................

   v61 = f5 + e18,
   f1 = pf1f2 f2 + d1,
   f2 = pf2f4 f4 + pf2f5 f5 + d2;
  std
   f3-f5 = varf3 - varf5,
   d1-d2 = vard1 - vard2,
   e1-e18 = vare1 - vare18;
  cov
   f3 f4 = cf3f4,
   f3 f5 = cf3f5,
   f4 f5 = cf4f5;
 var v24-v26 v28-v31 v69 v72-v74 v63-v66 v59-v61;
run;
quit;
```

Den første modelmodifikation, den valgte stimodel pegede på, var at fjerne den lineære relation mellem F2 og F3. Det betyder, at ligningen, der beskriver sammenhængen mellem F2 og F3 til F5, skal omformes, så de to ligninger, der definerer de latente variable F1 og F2, får udseendet

```
f1 = pf1f2 f2 + d1,
f2 = pf2f4 f4 + pf2f5 f5 + d2;
```

Vi har her udeladt de fleste af definitionsligningerne for de manifeste variable.

Derved får den modificerede stimodel, som vi skal kalde M_{p1}, det grafiske udseende:



De dele af SAS-output'et, der ændres, vises i de følgende bokse.

På symbolsk form viser SAS de nye ligninger for F1 og F2 som:

```
Latent Variable Equations

 F1      =      .     *F2      + 1.0000 D1
                       PF1F2

 F2      =      .     *F4      +   .     *F5      + 1.0000 D2
                       PF2F4              PF2F5

    6776 Observations        Model Terms           1
      18 Variables           Model Matrices        4
     171 Informations        Parameters           42

Chi-square = 5122.3772     df = 129     Prob>chi**2 = 0.0001
Null Model Chi-square:     df = 153                  25973.7793
```

Det fremgår samtidigt, at der nu kun er 42 parametre, sammenlignet med de 43 parametre i M_{p0}. Og at q-teststørrelsen for modellen er q = 5122.38 med 129 frihedsgrader.

I den næste boks er de nye estimater for de lineære parametre vist, sammen med deres standardfejl og 't Value'. Det fremgår, som forventet, at der ikke længere optræder insignifikante koefficienter.

```
  Latent Variable Equations

F1      =     0.2282*F2     +  1.0000 D1
Std Err       0.0152 PF1F2
t Value      15.0211

F2      =     0.5028*F4     -  0.0659*F5     +  1.0000 D2
Std Err       0.0195 PF2F4     0.0171 PF2F5
t Value      25.7255          -3.8513
```

Som det fremgår af den næste output-boks, er der imidlertid stadig en insignifikant korrelation mellem de 2 latente 'baggrundsvariable' F3 og F4.

```
              Covariances among Exogenous Variables
-----------------------------------------------------------------------
                                          Standard
     Parameter            Estimate          Error          t Value
-----------------------------------------------------------------------
F4      F3      CF3F4        -0.010950        0.008071          -1.357
F5      F3      CF3F5        -0.069200        0.009268          -7.466
F5      F4      CF4F5         0.145948        0.009523          15.325
```

Så det betaler sig måske at fjerne sammenhængen mellem de latente 'baggrundsvariable' F3 og F4.

Dette gør man ved at sætte korrelationen mellem F3 og F4 lig med 0, eller simpelthen at skrive

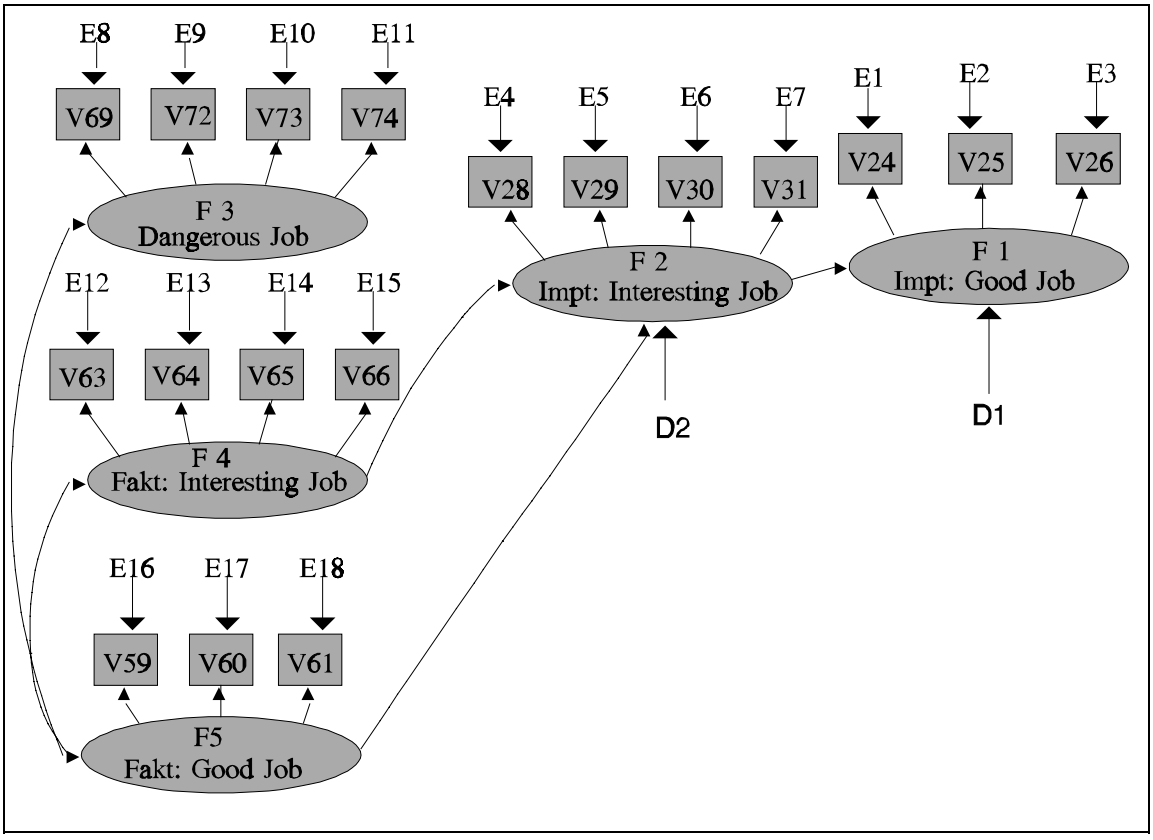
```
f3 f4 = 0
```

i 'cov'-afsnittet af PROC CALIS programmet.

De væsentlige del af SAS-programmet skal derfor ændres som følger.

Grafen eller stidiagrammet for denne model, som vi skal kalde M_{p2}, ser sådan ud, idet dobbeltpilen mellem F3 og F4 er fjernet:

```
proc calis covariance corr;
  lineqs
    ......................

    f1 = pf1f2 f2 + d1,
    f2 = pf2f4 f4 + pf2f5 f5 + d2;
  std
    f3-f5 = varf3 - varf5,
    d1-d2 = vard1 - vard2,
    e1-e18 = vare1 - vare18;
  cov
    f3 f4 = 0,
    f3 f5 = cf3f5,
    f4 f5 = cf4f5;
 var v24-v26 v28-v31 v69 v72-v74 v63-v66 v59-v61;
run;
quit;
```



Det første SAS-output viser, at der nu kun er 41 parametre i modellen, og at den relevante q-teststørrelse nu er 5124.17 med 130 frihedsgrader. Dette resultat er stadig kraftigt signifikant.

```
Latent Variable Equations

 F1      =     .    *F2     + 1.0000 D1
                     PF1F2

 F2      =     .    *F4     +  .    *F5     + 1.0000 D2
                     PF2F4           PF2F5

    6776 Observations       Model Terms          1
      18 Variables          Model Matrices       4
     171 Informations       Parameters          41

Chi-square = 5124.1684     df = 130     Prob>chi**2 = 0.0001
Null Model Chi-square:     df = 153                25973.7793
```

Vi venter imidlertid til Kapitel 20 med at sammenligne de forskellige modeller. Her skal vi se, at det ikke så meget er den absolutte værdi af q-teststørrelsen, der er afgørende, som differenserne mellem q-teststørrelserne. Disse er nemlig et mål for, hvor godt én model beskriver data i forhold til en anden model med flere estimerede parametre.

Den næste boks viser de to ligninger, der definerer de latente variable, sammen med de estimerede L-koefficienter, deres standardafvigelser og 't Value'. Det fremgår, at alle koefficienter er signifikant forskellige fra 0. Fjernelsen af korrelationen mellem F3 og F4 har altså ikke ændret på dette forhold.

```
  Latent Variable Equations

 F1      =     0.2282*F2     +  1.0000 D1
 Std Err       0.0152 PF1F2
 t Value      15.0225

 F2      =     0.5016*F4     -  0.0649*F5     +  1.0000 D2
 Std Err       0.0195 PF2F4     0.0171 PF2F5
 t Value      25.7803          -3.7995
```

Derimod viser den næste boks, at der kun er 2 estimerede korrelationer mellem de latente 'baggrundsvariable ' tilbage, nu hvor korrelationen mellem F3 og F4 er udeladt af modellen. Til gengæld er begge disse C-koefficienter signifikant forskellige fra 0.

```
              Covariances among Exogenous Variables
-------------------------------------------------------------------
                                          Standard
     Parameter            Estimate          Error          t Value
-------------------------------------------------------------------
F5     F3     CF3F5        -0.066589        0.009040          -7.366
F5     F4     CF4F5         0.144706        0.009486          15.254
```

I den sidste boks er de standardiserede L- og C-koefficienter vist, hvis man har lyst til at sætte estimerede korrelationer på grafen for denne model.

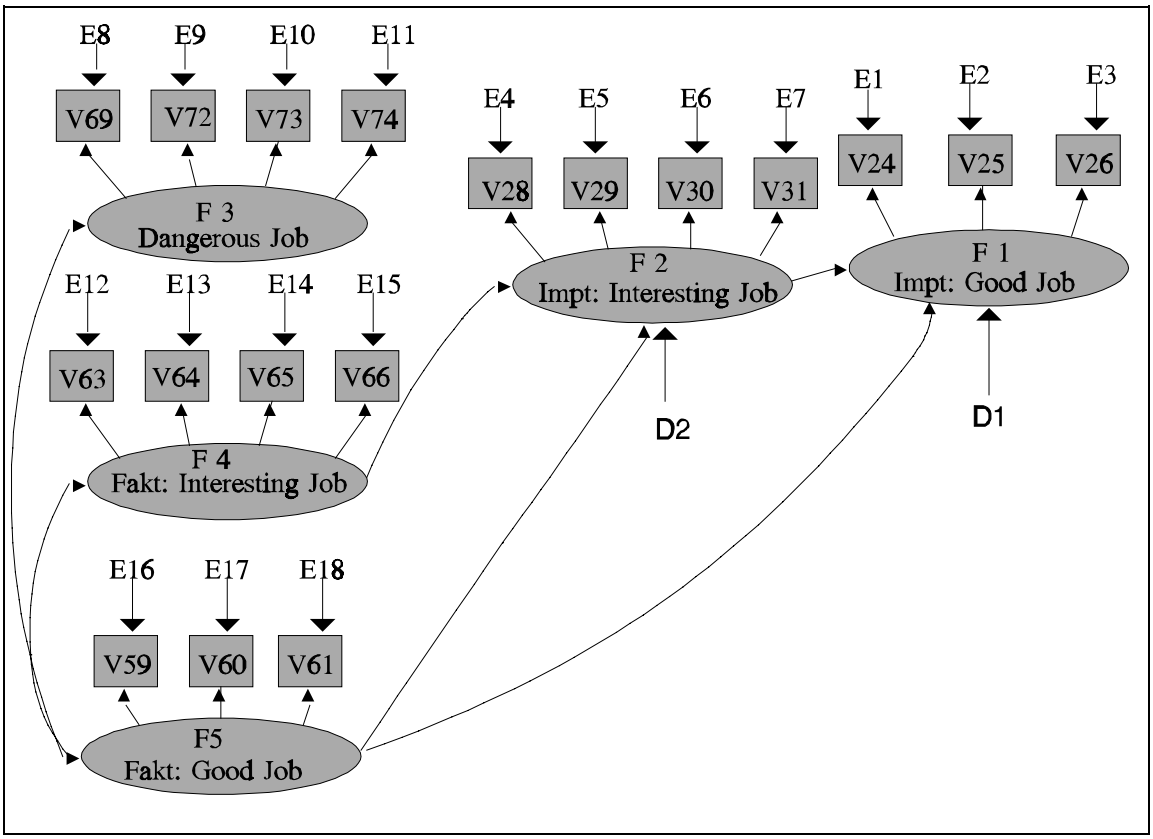
```
 F1       =     0.3032*F2      + 0.9529 D1
                        PF1F2

 F2       =     0.5047*F4      - 0.0675*F5      + 0.8726 D2
                        PF2F4            PF2F5


Correlations among Exogenous Variables
-------------------------------------
     Parameter             Estimate
-------------------------------------
F5      F3      CF3F5         -0.123663
F5      F4      CF4F5          0.299410
```

Vi mangler en enkelt model, som bør overvejes. Hvis man ser tilbage på målingsmodellen i Kapitel 18, fremgik det, at der var signifikante korrelationer mellem F1 og F4 og mellem F1 og F3. Dog var 't-value' for korrelationen mellem F1 og F3 kun lidt over 3. Denne sammenhæng var mellem om jobbet rent faktisk var et 'Good Job' og om respondenten mente at det var væsentligt, at jobbet var et 'Interesting Job'. Man kunne måske håbe på, at denne sammenhæng kunne 'opsluges' af sammenhængen mellem F5 og F2, nemlig at jobbet er et 'Good Job' både rent faktisk, og bliver anset som sådan af respondenten.

Derimod kan vi ikke ignorere sammenhængen mellem F5 og F1, som udtrykker at jobbet er et 'Interesting Job' både rent faktisk, og bliver anset som sådan af respondenten. Så vi bør prøve at indføje en sammenhæng mellem F5 og F1 i modellen. De relevante linier i PROC CALIS programmet ændres derfor til:

```
proc calis covariance corr residual;
  lineqs
   v24 = lv24f1 f1 + e1,
   .....

   f1 = pf1f2 f2 + pf1f5 f5 + d1,
   f2 = pf2f4 f4 + pf2f5 f5 + d2;
  std
   f3-f5 = varf3 - varf5,
   d1-d2 = vard1 - vard2,
   e1-e18 = vare1 - vare18;
  cov
   f3 f4 = 0,
   f3 f5 = cf3f5,
   f4 f5 = cf4f5;
 var v24-v26 v28-v31 v69 v72-v74 v63-v66 v59-v61;
run;
```

Denne model, som vi skal kalde M_{p3}, har grafen:



Den første output-boks vises de symbolske ligninger, så vi kan checke, at modellen er korrekt skrevet op i SAS-programmet. Desuden fremgår det, at der er 42 parametre i modellen, så antal frihedsgrader til rådighed for at teste modellen er 171 - 42 = 129.

```
Latent Variable Equations

 F1      =     .     *F2      +  .     *F5      + 1.0000 D1
                      PF1F2             PF1F5

 F2      =     .     *F4      +  .     *F5      + 1.0000 D2
                      PF2F4             PF2F5

    6776 Observations        Model Terms           1
      18 Variables           Model Matrices        4
     171 Informations        Parameters           42
```

Lidt længere nede i output finder vi listen over teststørrelser og mål for modeltilpasningen.

I boksen neden for er kun vist de 2 linier med q-teststørrelser. Det fremgår, at modellen er signifikant med en observeret q-værdi på 5071.96 med - som beregnet ovenfor - 129 frihedsgrader. Men igen her henviser jeg til Kapitel 20, hvor de forskellige modeller sammenlignes udfra **differenser** mellem observerede q-teststørrelser.

```
Chi-square = 5071.9582     df = 129     Prob>chi**2 = 0.0001
Null Model Chi-square:     df = 153               25973.7793
```

Boksen neden for viser definitionsligningerne for de latente 'afledte' variable, med standard fejl og tilhørende 't-Value'. Værdierne af 't-value' for sammenhængene mellem F1 og F2, F2 og F4, samt F2 og F5, har stort set ikke ændret sig, men P-koefficienten svarende til sammenhængen mellem F1 og F5 er klart signifikant med en 't-value' på 7.74.

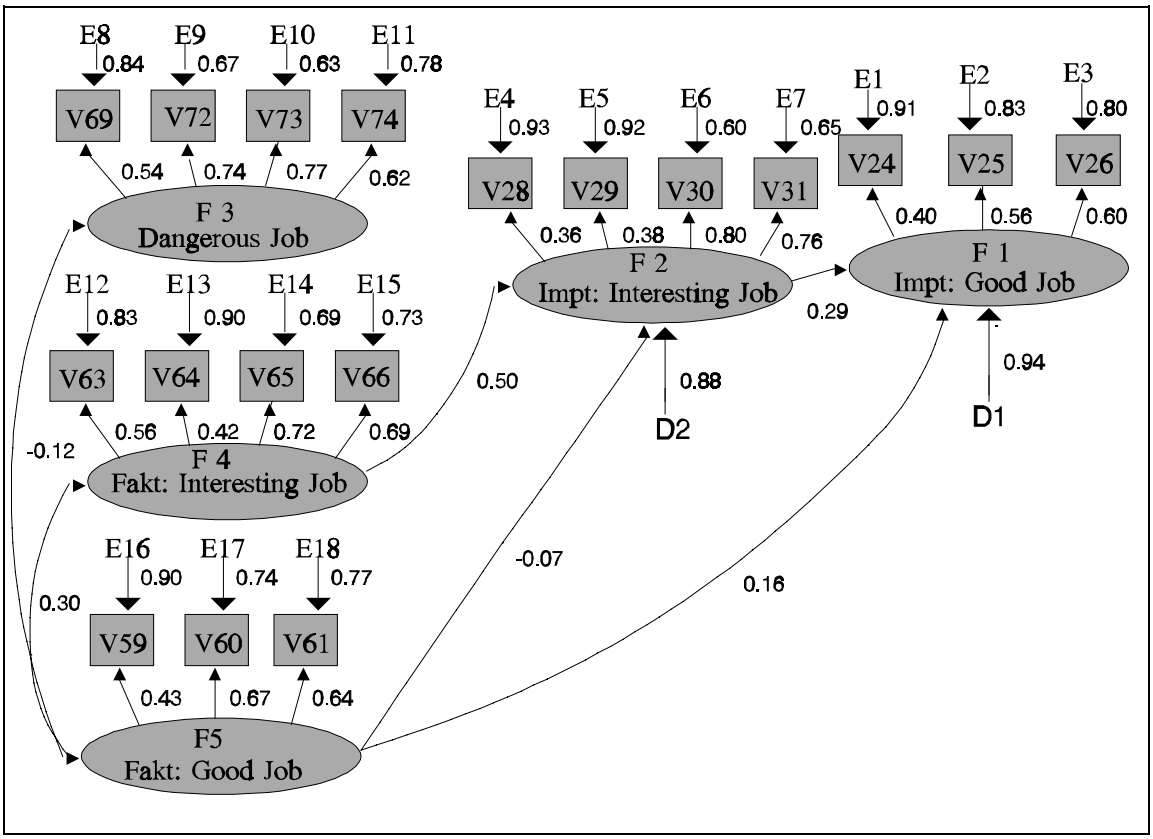
```
Latent Variable Equations

 F1      =     0.2422*F2     +  0.1233*F5     +  1.0000 D1
 Std Err       0.0159 PF1F2     0.0159 PF1F5
 t Value      15.2222           7.7407

 F2      =     0.4994*F4     -  0.0660*F5     +  1.0000 D2
 Std Err       0.0194 PF2F4     0.0168 PF2F5
 t Value      25.7801          -3.9345
```

Hvis man ønsker at tegne modellens graf med de estimerede korrelationer, viser de 2 næste bokse de dele af SAS-output'et, som skal benyttes.

```
Equations with Standardized Coefficients

 F1      =    0.2942*F2     + 0.1584*F5     + 0.9385 D1
                    PF1F2           PF1F5

 F2      =    0.5029*F4     - 0.0697*F5     + 0.8736 D2
                    PF2F4           PF2F5
```

```
Correlations among Exogenous Variables
-------------------------------------
     Parameter            Estimate
-------------------------------------
F5     F3     CF3F5        -0.123663
F5     F4     CF4F5         0.299410
```

Grafen eller stidiagrammet for den 'endelige' model med styrken af de enkelte sammenhænge udtrykt ved korrelationer ser sådan ud:


E8
E9
E10
E11
0.84
0.67
0.63
0.78
V69
V72
V73
V74
0.54
0.74
0.77
0.62
F 3
Dangerous Job
E12
E13
E14
E15
0.83
0.90
0.69
0.73
V63
V64
V65
V66
0.56
0.42
0.72
0.69
-0.12
F 4
Fakt: Interesting Job
E16
E17
E18
0.90
0.74
0.77
0.30
V59
V60
V61
0.43
0.67
0.64
F5
Fakt: Good Job
E4
E5
E6
E7
0.93
0.92
0.60
0.65
V28
V29
V30
V31
0.36
0.38
0.80
0.76
F 2
Impt: Interesting Job
0.50
0.88
D2
-0.07
0.16
E1
E2
E3
0.91
0.83
0.80
V24
V25
V26
0.40
0.56
0.60
F 1
Impt: Good Job
0.29
0.94
D1

20. Sammenligning af målingsmodellen og stimodellerne.

Vi skal bruge resultaterne i Kapitel 17 til at sammenligne de forskellige modeller, der blev analyseret i Kapitel 18 og 19. De modeller, der er tale om, er M_m, M_{p0}, M_{p1}, M_{p2} og M_{p3}.

De af SAS beregnede q-teststørrelserne og de tilhørende frihedsgrader (df) er vist i tabellen nedenfor. Jeg har ikke vist signifikanssandsynlighederne for disse q-værdier, idet de alle er vildt signifikante.

Derimod er der, for hver af de 4 sidste linier, vist differensen mellem q-størrelserne, antal frihedsgrader for q-differensen og de tilhørende signifikanssandsynligheder.

Model	q	df	q-differens	df	Sign.ssh.
M_m	5005.23	125	-	-	-
M_{p0}	5121.02	128	115.79	3	< 0.0005
M_{p1}	5121.38	129	1.36	1	0.244
M_{p2}	5124.17	130	1.79	1	0.181
M_{p3}	5071.96	129	52.21	1	< 0.0005

Her får man overblik over situationen, som kan samles i følgende resultater:

(a) Målingmodellen (M_m) kan ikke beskrive materialet. Det kan derfor heller ikke nogle af de 4 stimodeller (M_{p0} til M_{p3}), der jo er enklere modeller end målingsmodellen. Så egentlig burde man stoppe her. Men de følgende kommentarer viser, at det behøver man ikke.

(b) Da q-differensen mellem målingsmodellen (M_m) og den første stimodel (M_{p0}) er signifikant, kan vi konstatere, at vi i basis stimodellen har udeladt signifikante sammenhænge mellem de latente variable.

(c) q-differensen mellem M_{p1} og M_{p0} er ikke signifikant. Det betyder, at de to modeller har en lige god (omend generelt meget dårlig) tilpasning til data. Så korrelationen mellem F3 og F2 kan altså udelades uden at forringe modellens tilpasningsevne.

(d) På samme måde er q-differensen mellem M_{p2} og M_{p1} ikke signifikant. Det betyder, at de to modeller har en lige god tilpasning til data. Korrelationen mellem F3 og F4 kan altså også udelades uden at forringe modellens tilpasningsevne.

(e) Her skal vi være opmærksom på at M_{p3} er en **forbedring** af modellen i forhold til M_{p2}, dvs. M_{p3} har **flere** parametre end M_{p2}. Det betyder, at vi nu skal trække det nederste tal fra **dét lige over**, hvor vi i de andre linier har gjort det modsatte. Her er resultatet klart signifikant, så vi får en markant forbedring af modellens tilpasning til data ved at tilføje en korrelation mellem F5 og F1.

21. Simpel correspondance analyse

Den observerede værdi af Pearsons Q-teststørrelse

$$q = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(x_{ij} - \frac{x_{i\cdot}x_{\cdot j}}{n}\right)^2}{\frac{x_{i\cdot}x_{\cdot j}}{n}}$$

er et mål for hvor meget de observerede antal i en kontingenstabel $\{x_{ij}\}$ afviger fra de forventede antal

$$\frac{x_{i\cdot}x_{\cdot j}}{n}$$

under en hypotese om uafhængighed. Idéen i correspondance analyse er at beskrive forskellene mellem de observerede og forventede værdier ved hjælp af passende størrelser, der kan afbildes i en graf. Det er traditionen i correspondance analyse, at man arbejder med hyppigheder og ikke antal. Man omskriver derfor Pearsons Q-teststørrelse til

$$q = n\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(f_{ij} - f_{i\cdot}f_{\cdot j}\right)^2}{f_{i\cdot}f_{\cdot j}} ,$$

hvor $f_{ij} = x_{ij}/n$, $f_{i\cdot} = x_{i\cdot}/n$ og $f_{\cdot j} = x_{\cdot j}/n$, så

$$q = n\sum_{i=1}^{I}\sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\cdot}f_{\cdot j}} - 1\right)^2 f_{i\cdot}f_{\cdot j} ,$$

Jo større leddene

$$r_{ij} = \frac{f_{ij}}{f_{i\cdot}f_{\cdot j}} - 1$$

derfor er, jo mere afviger de observerede data fra uafhængighedsantagelsen.

I correspondance analyse dekomponerer man matricen **R** af residualled r_{ij} i egenvektorer og egenværdier, dvs.

$$\mathbf{R} = \Phi\,\Lambda\,\Psi^T$$

Hvor Λ, som sædvanligt, er en diagonal matrix med egenværdierne i diagonalen. Hvad φ'erne og ψ'erne, dvs elementerne i Φ og Ψ, betyder, vender vi tilbage til.

Der er en mest teknisk detalje af stor betydning for anvendelserne, man skal kende til. I modsætning til en almindelig egenværdidekomponering af en matrix, hvor søjlerne i Φ og Ψ er ortogonale, anvendes i correspondance analyse matricer, der er ortogonale med en lidt anden definition. Man kræver nemlig, at kvadratsummerne af søjlerne i Φ skal være 1, **vægtet med rækkefrekvenserne**, og at produktsummerne af søjlerne i Φ skal være 0, **vægtet med rækkefrekvenserne.** Det tilsvarende gælder for matricen Ψ. Normeringerne af φ'rne og ψ'erne får derfor formen:

$$\sum_{i=1}^{I} \varphi_{im}^2 f_{i\cdot} = \sum_{j=1}^{J} \psi_{jm}^2 f_{\cdot j} = 1 \quad , \quad \text{for alle } m \ ,$$

mens kravet om, at produktsummerne skal være 0, erstattes af

$$\sum_{i=1}^{I} \varphi_{im} \varphi_{i\mu} f_{i\cdot} = \sum_{j=1}^{J} \psi_{jm} \psi_{j\mu} f_{\cdot j} = 0 \quad , \quad \text{for alle } m \neq \mu \ .$$

Ligningen

$$\mathbf{R} = \Phi \, \Lambda \, \Psi^T$$

giver en 'fuldstændig løsning' i den forstand, at alle afvigelser fra uafhængighed er beskrevet.

Af hensyn til det følgende skriver vi denne matrix ligning fuldt ud:

$$r_{ij} = \sum_{m=1}^{M} \lambda_m \varphi_{im} \psi_{jm} \quad , \quad i = 1, \ldots, I, \ j = 1, \ldots J \ .$$

Matricen **R** har rang min{ I-1 , J-1 }, idet

$$\sum_{i} f_{i\cdot} r_{ij} = 0 \quad \text{og} \quad \sum_{j} f_{\cdot j} r_{ij} = 0 \ .$$

Dvs. rangen af M er lig med min{ I-1 , J-1 }, så selv for kontingenstabeller af moderat orden, f.eks. 3x4 eller 5x5, kan der blive mange parametre at holde rede på.

Man vælger derfor i correspondance analyse at beskrive residualmatricen, dvs. afvigelserne mellem det observerede og det forventede under uafhængighed, ved en egenværdi/egenvektor dekomponering af mindre dimension end den maksimale. I den franske tradition, som SAS følger, vælges altid to dimensioner.

Det betyder ganske vist, at man kun får en approksimation til residualerne, idet vi nu har

$$\mathbf{R} \simeq \Phi_0 \, \Lambda_0 \, \Psi^T{}_0$$

eller
hvor Λ_0 betegner Λ med kun de M_0 største λ'er, Φ_0 består af de første M_0 søjler i Ψ og Ψ_0 består af de første M_0 søjler i Ψ. Denne approksimation bliver naturligvis bedre, større dimension M_0 vi vælger, dvs. hvor tæt M_0 kommer på M.

$$r_{ij} = \sum_{m=1}^{M} \lambda_m \varphi_{im} \psi_{jm} \simeq \sum_{m=1}^{M_0} \lambda_m \varphi_{im} \psi_{jm} ,$$

Lige som for principalkomponentmetoden kan correspondance analyse med M_0 dimensioner begrundes ved et mindste kvadraters resultat.

SÆTNING: Hvis kvadratsummen $Q(M_0)$ er defineret som

$$Q(M_0) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(r_{ij} - \sum_{m=1}^{M_0} \lambda_m \varphi_{im} \psi_{jm} \right)^2 f_{i\cdot} f_{\cdot j} ,$$

da antager $Q(M_0)$ sit minimum, hvis

$$\lambda_1, \lambda_2, \ldots, \lambda_{M_0}$$

er de M_0 største egenværdier, mens φ_{im}, i=1,...,I, m=1,...,M_0 er de søjler i Φ, der svarer til de M_0 største egenværdier og ψ_{jm}, j=1,...,J, m=1,...,M_0 er de søjler i Ψ, der svarer til de M_0 største egenværdier.

Approksimationen

$$\mathbf{R} \simeq \Phi_0 \Lambda_0 \Psi_0^T ,$$

som vi også udledte oven for, kan således tillige begrundes som en mindste kvadraters løsning.

I correspondance analyse tegner man altid det såkaldte 'correspondance analyse diagram' med $M_0 = 2$. I et sådant diagram er der ét punkt for hver rækkekategori og ét punkt for hver søjlekategori. x-koordinater for de punkter, der svarer til rækkekategorierne, er

$$(\lambda_1 \varphi_{11}, \ldots, \lambda_1 \varphi_{I1}) ,$$

mens y-koordinaterne er

$$(\lambda_2 \varphi_{12}, \ldots, \lambda_2 \varphi_{I2}) .$$

På tilsvarende måde bliver x-koordinaterne for de punkter, der svarer til søjlekategorierne

$$(\lambda_1 \psi_{11}, \ldots, \lambda_1 \psi_{J1}) ,$$

og y-koordinaterne

For at vise, hvordan et correspondance analyse diagram ser ud, betragter vi følgende eksempel.

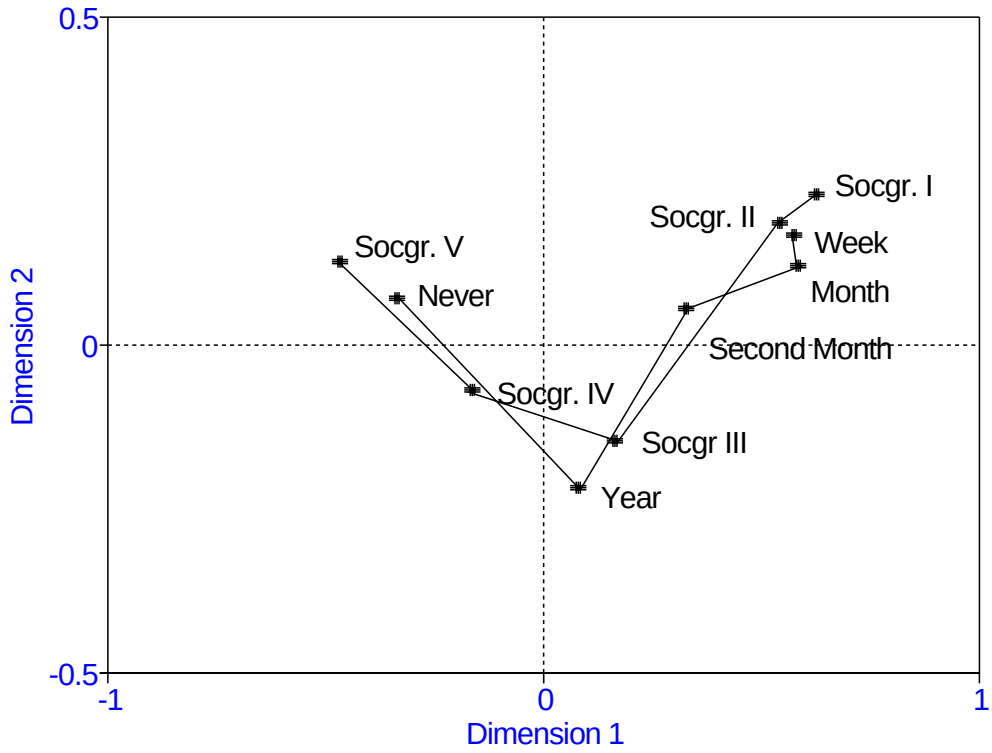
$$\left(\lambda_2\psi_{12}, \ldots, \lambda_2\psi_{J2}\right) .$$

Tabellen neden for viser hvordan et tilfældigt udvalg af danskere mellem 40 og 59 år gamle i 1974 fordelte sig på Social Gruppe og hvor ofte de deltog i møder uden for arbejdstiden.

	Hyppighed af møder uden for arbejdstiden				
Social Gruppe	Én eller to gange om ugen	Èn eller to gange om måneden	Ca. én gang hver anden måned	Nogle få gange om året	Aldrig
I	17	27	13	24	25
II	25	57	17	49	55
III	38	91	41	217	213
IV	22	33	21	133	222
V	9	21	17	87	305

Det er også denne tabel, der bliver analyseret i forbindelse med correspondance analyse med SAS i næste kapitel.

Foretager vi en egenværdi/egenvektor dekomponering af residual matricen **R**, som gennemgået oven for og taster de viste x- og y-koordinater for både rækker og søjler ind i et regneark, får vi grafen på næste side. Det fremgår, at de to højeste socialgrupper ligner hinanden, og at det er dem, der oftest går til møder. Socialgruppe III ligger i midten og går kun af og til til møder. Endelig går personer i de to laveste socialgruppe næsten aldrig til møder uden for arbejdstiden. For den laveste socialgruppe, ser det endda ud til, at de aldrig går til møder.



Den observerede værdi af Pearsons Q-teststørrelse

$$q = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(x_{ij} - \frac{x_{i\cdot}x_{\cdot j}}{n}\right)^2}{\frac{x_{i\cdot}x_{\cdot j}}{n}}$$

kan også skrives

$$q = \sum_{i=1}^{I}\sum_{j=1}^{J} \left(\frac{x_{ij}}{x_{i\cdot}} - \frac{x_{\cdot j}}{n}\right)^2 \left(\frac{n\,x_{i\cdot}}{x_{\cdot j}}\right),$$

eller

$$q = n\sum_{i=1}^{I} f_{i\cdot}\sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}\right)^2 \left(\frac{1}{f_{\cdot j}}\right).$$

Ved at bytte om på i og j, får man det analoge resultat

$$q = n \sum_{j=1}^{J} f_{\cdot j} \sum_{i=1}^{I} \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot} \right)^2 \left(\frac{1}{f_{i\cdot}} \right).$$

Man kalder i correspondance analyse vektoren

$$\frac{f_{ij}}{f_{i\cdot}} \quad , \quad j = 1, \ldots , J$$

for rækkeprofilerne, idet disse størrelser er de relative hyppigheder for tallene i række i. Under uafhængighedshypotesen skal disse være ens bortset fra tilfældige afvigelser. På samme måde defineres søjleprofilerne som

$$\frac{f_{ij}}{f_{\cdot j}} \quad , \quad i = 1, \ldots ,I \ .$$

Også disse skal under uafhængighedshypotesen være ens bortset fra tilfældige afvigelser.

Pearsons Q-teststatistic kan derfor fortolkes enten som en vejet sum af afstandene mellem søjleprofilerne og de estimerede søjleprofiler under hypotesen om uafhængighed, eller som en vejet sum af afstandene mellem rækkeprofilerne og de estimerede rækkeprofiler under hypotesen om uafhængighed.

Der er en meget vigtig omskrivning af Pearsons Q-teststørrelse, som spiller en stor rolle ved fortolkningen af correspondance analysens resultater.

Da q kan skrives

$$q = n \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\cdot} f_{\cdot j}} - 1 \right)^2 f_{i\cdot} f_{\cdot j}$$

og

$$\frac{f_{ij}}{f_{i\cdot} f_{\cdot j}} - 1 = \sum_{m=1}^{M} \lambda_m \varphi_{im} \psi_{jm} \ ,$$

får vi

$$q = n \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\sum_{m=1}^{M} \lambda_m \varphi_{im} \psi_{jm} \right)^2 f_{i\cdot} f_{\cdot j}$$

$$= n \sum_{m=1}^{M} \sum_{\mu=1}^{M} \lambda_m \lambda_\mu \sum_{i=1}^{I} \varphi_{im} \varphi_{i\mu} f_{i\cdot} \sum_{j=1}^{J} \psi_{jm} \psi_{j\mu} f_{\cdot j},$$

Men da de to sidste summer (over i og j) er lig 0, undtagen for m = μ, får vi udtrykket

$$q = n\sum_{m=1}^{M} \lambda_m^2.$$

På den anden side er q_0, defineret som

$$q_0 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{f_{ij}}{f_{i\cdot}f_{\cdot j}} - 1 - \sum_{m=1}^{M_0}\lambda_m\varphi_{im}\psi_{jm}\right)^2 f_{i\cdot}f_{\cdot j} \ ,$$

et mål for hvor godt data er beskrevet, efter vi har indført en correspondance analyse model med $M = M_0$.

Benytter vi

$$r_{ij} = \left(\frac{f_{ij}}{f_{i\cdot}f_{\cdot j}} - 1\right) = \sum_{m=1}^{M}\lambda_m\varphi_{im}\psi_{jm}$$

kan ovenstående ligning omskrives til

$$q_0 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{m=1}^{M}\lambda_m\varphi_{im}\psi_{jm} - \sum_{m=1}^{M_0}\lambda_m\varphi_{im}\psi_{jm}\right)^2 f_{i\cdot}f_{\cdot j}$$

$$= n\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{m=M_0+1}^{M}\lambda_m\varphi_{im}\psi_{jm}\right)^2 f_{i\cdot}f_{\cdot j} \ ,$$

der let omskrives til

$$q_0 = n\sum_{m=M_0+1}^{M}\lambda_m^2$$

jvf. reduktionen af q til n gange en sum af alle λ'erne oven for.

Værdien af q_0 er lille, hvis næsten al variation i data er forklaret af correspondance analyse modellen, og stor, hvis kun en lille del af variationen er forklaret.

Man kan derfor benytte
som et mål for hvor meget af variationen i data, der er forklaret af en correspondance analyse model med M_0 dimensioner.

$$q - q_0 = n\sum_{m=1}^{M_0} \lambda_m^2$$

Et relativt mål for den forklarede del bliver herefter

$$r^2(M_0) = \frac{n\sum_{m=1}^{M_0} \lambda_m^2}{n\sum_{m=1}^{M} \lambda_m^2} = \frac{q - q_0}{q} .$$

For vores eksempel får vi tallene

	Dimension			
	$m / M_0 = 1$	2	3	4
λ_m	0.354	0.139	0.048	0.021
$r^2(M_0)$	0.850	0.981	0.997	1.000

Ud fra udtrykket

$$q = n \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\sum_{m=1}^{M} \lambda_m \varphi_{im} \psi_{jm} \right)^2 f_{i\cdot} f_{\cdot j} \ .$$

for Pearsons Q-teststatistic kan man danne sig et indtryk af, hvor meget rækkevariablen i bedrager til beskrivelsen af afvigelserne mellem data og model. Vi behøver blot at fastholde indeks i, dvs.

$$q_i = n \sum_{j=1}^{J} \left(\sum_{m=1}^{M} \lambda_m \varphi_{im} \psi_{jm} \right)^2 f_{i\cdot} f_{\cdot j} \ .$$

Lidt omformninger af dette udtryk giver:

$$q_i = n f_{i\cdot} \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{\mu=1}^{M} \lambda_m \lambda_\mu \varphi_{im} \psi_{jm} \varphi_{i\mu} \psi_{j\mu} f_{\cdot j} \ .$$

eller

$$q_i = n f_{i\cdot} \sum_{m=1}^{M} \sum_{\mu=1}^{M} \lambda_m \lambda_\mu \varphi_{im} \varphi_{i\mu} \sum_{j=1}^{J} \psi_{jm} \psi_{j\mu} f_{\cdot j} \ .$$

Men da summen over j i dette udtryk på grund af ψ-matricens ortogonalitet, vejet med $f_{\cdot j}$, er 1 for $m = \mu$ og 0 for $m \neq \mu$, får vi

$$q_i = n f_{i\cdot} \sum_{m=1}^{M} \lambda_m^2 \varphi_{im}^2 \ .$$

Det **relative bidrag fra række i til dimension m**, bliver altså

$$D_{im}^{A} = \frac{n f_{i\cdot} \lambda_m^2 \varphi_{im}^2 \ .}{\sum_{i=1}^{I} n f_{i\cdot} \lambda_m^2 \varphi_{im}^2}$$

Men igen på grund af Φ-maricens vejede ortogonalitet, dvs.

$$\sum_{i=1}^{I} f_{i\cdot}\varphi_{im}^2 = 1 \ ,$$

har $D_{i\ m}^{A}$ den simple form

$$D_{im}^{A} = f_{i\cdot}\varphi_{im}^2 \ .$$

På samme måde defineres det **relative bidrag fra søjle j til dimension m** som

$$D_{jm}^{B} = f_{\cdot j}\psi_{jm}^2 \ .$$

Hvis man i stedet sammenligner det (i,m)'te led i q_i med summen over m ved at dividere med summen over m, får man det **relative bidrag til række i fra dimension m** som

$$d_{im}^{A} = \frac{n f_{i\cdot}\lambda_m^2\varphi_{im}^2}{\sum_{m=1}^{M} n f_{i\cdot}\lambda_m^2\varphi_{im}^2} = \frac{\lambda_m^2\varphi_{im}^2}{\sum_{m=1}^{M}\lambda_m^2\varphi_{im}^2}$$

Tilsvarende bliver det **relative bidrag til søjle j fra dimension m**

$$d_{jm}^{B} = \frac{\lambda_m^2\psi_{jm}^2}{\sum_{m=1}^{M}\lambda_m^2\psi_{jm}^2}$$

22. Simpel correspondance analyse i SAS

For eksemplet i Kapitel 21, hvor 'Social Gruppe' er krydset med 'Hyppigheden af at deltage i møder uden for arbejdstiden', udføres en correspondance analyse ved hjælp af SAS proceduren PROC CORRESP. SAS-programmet kan udformes forskelligt.

Dette SAS program er en version, der et sat op, så man får et simpelt plot.

```
1 title 'Meeting data';
2 data meet;
3 input socgr $ w m s y n;
4   label w = 'Week'
5         m = 'Month'
6         s = 'Second Month'
7         y = 'Year'
8         n = 'Never';
9 cards;
10    1    17     27     13     24     25
11    2    25     57     17     49     55
12    3    38     91     41    217    213
13    4    22     33     21    133    222
14    5     9     21     17     87    305
15 proc corresp all data = meet out = results;
16   var w m s y n;
17   id socgr;
18 proc plot data = results;
19 plot dim2*dim1 = socgr;
20 run;
21 quit;
```

Linie 1 angiver en overskrift til hver side.

Linie 2-14 viser hvordan datasættet 'meet' sættes op.

```
2 data meet;
3 input socgr $ w m s y n;
4   label w = 'Week'
5         m = 'Month'
6         s = 'Second Month'
7         y = 'Year'
8         n = 'Never';
9 cards;
10   1   17    27    13    24    25
11   2   25    57    17    49    55
12   3   38    91    41   217   213
13   4   22    33    21   133   222
14   5    9    21    17    87   305
```

I linie 3 kaldes rækkevariablen 'socgr', der står for Social Gruppe, mens søjlevariablene kaldes w, m, s, y og n. Det er af hensyn til grafen, jeg har valgt disse fem bogstaver. I linie 4 til 8 tildeles de labels. Jeg har her, igen af hensyn til grafen, forkortet sådan:

'Week' = 'Once a week'
'Month' = 'Once a month'
'Second month' = 'Once every second month'
'Year' = 'Once a year'
'Never' = 'Never'

Pointen er, at alle labels starter med et forskelligt første bogstav.

Fra linie 9 til 14 står selve data. Her er socialgrupperne I, II, III, IV og V navngivet '1', '2', '3', '4' og '5' - igen af hensyn til grafen. Bemærk, at man stadig skal skrive 'cards' (i linie 9) som en nostalgisk påmindelse om hulkortenes tidsalder fra 1960 til 1975.

Kaldet af PROC CORRESP står i linie 15 - 17.

```
15 proc corresp all data = meet out = results;
16  var w m s y n;
17  id socgr;
```

Her er valgt optionen 'all', der giver en række grundtal, mens 'data = meet' betyder, at der analyseres på datasættet 'meet'. Endelig angiver 'out = results', at der laves et output datasæt, hvor bl.a. koordinaterne til correspondance analyse diagrammet står. De fem variable, der skal analyseres står i linie 16. Som det ses, er det søjlevariablene. At rækkevariablene skal være de 5 socialgrupper markeres i linie 17 med 'id socgr'.

Selve grafen laves i linie 18 og 19:

```
18 proc plot data = results;
19 plot dim2*dim1 = socgr;
```

Læg mærke til, at i output datasættet 'results' fra PROC CORRESP er plotkoordinaterne gemt som 'dim1' og 'dim2', derfor skal Proc Plot anvendes på SAS-datasættet 'results'. Man skal skrive 'plot dim2*dim1 = socgr', idet 'dim1' og 'dim2', som sagt er plotkoordinaterne, og vi ønsker et punkt for hver værdi af 'socgr'.

Endelig skal der, som altid stå 'run' og 'quit' i de to sidste linier, linie 20 og 21.

SAS leverer en del output, hvor noget er relateret til den korte teoretiske gennemgang i Kapitel 21, andet ikke. Correspondance analyse, som den blev lanceret af franskmænd i midten af 60'erne var baseret på en geometrisk beskrivelse af sammenhænge mellem kategoriserede data, hvorfor mange af betegnelserne - som er overtaget af SAS-programmørerne -, stadig refererer til de franske geometriske betegnelser. Imidlertid viser SAS-output'et alle de væsentlige størrelser.

Så i det følgende, skal vi koncentrere os om de dele af SAS output'et, der har interesse i forbindelse med gennemgangen af correspondance analyse i Kapitel 21. Output'et i uddrag ser sådan ud:

De første 3 SAS-output bokse viser de resultater, man får ved at bruge optionen 'all' i kaldet af PROC CORRESP. I boksen neden for står de observerede antal med række- og søjlesummer, de forventede antal man får under uafhængighedshypotesen

$$n\hat{\pi}_{ij} = \frac{x_{i.}\, x_{.j}}{x_{..}} ,$$

samt forskellene mellem observerede og de forventede antal under uafhængighedshypotesen.

Contingency Table

	Week	Month	Second Month	Year	Never	Sum
1	17	27	13	24	25	106
2	25	57	17	49	55	203
3	38	91	41	217	213	600
4	22	33	21	133	222	431
5	9	21	17	87	305	439
Sum	111	229	109	510	820	1779

Chi-Square Statistic Expected Values

	Week	Month	Second Month	Year	Never
1	6.614	13.645	6.495	30.388	48.859
2	12.666	26.131	12.438	58.196	93.569
3	37.437	77.234	36.762	172.007	276.560
4	26.892	55.480	26.408	123.558	198.662
5	27.391	56.510	26.898	125.852	202.350

Observed Minus Expected Values

	Week	Month	Second Month	Year	Never
1	10.386	13.355	6.505	-6.388	-23.859
2	12.334	30.869	4.562	-9.196	-38.569
3	0.563	13.766	4.238	44.993	-63.560
4	-4.892	-22.480	-5.408	9.442	23.338
5	-18.391	-35.510	-9.898	-38.852	102.650

Den næste boks viser de individuelle bidrag til Pearsons Q-teststørrelse. Den samlede værdi af Q-teststørreslen er q = 262.66, der med 16 frihedsgrader klart forkaster uafhængigheds-hypotesen.

Contributions to the Total Chi-Square Statistic

	Week	Month	Second Month	Year	Never	Sum
1	16.310	13.072	6.516	1.343	11.651	48.892
2	12.010	36.466	1.673	1.453	15.898	67.501
3	0.008	2.453	0.489	11.769	14.608	29.327
4	0.890	9.109	1.107	0.722	2.742	14.569
5	12.348	22.314	3.642	11.994	52.074	102.372
Sum	41.567	83.414	13.427	27.280	96.972	262.661

De sidste tabeller viser, hvad man i correspondance analyse kalder ’rækkeprofiler’ og ’søjleprofiler’. Der er blot tale om, at man i den første tabel kalder de relative hyppigheder i hver række ’Row Profiles’. F.eks er 0.160 = 17/106. I tabellen med søjleprofiler kaldes omvendt de relative hyppigheder i hver søjle ’Column Profiles’. F.eks. er 0.153 = 17/111.

```
Row Profiles

                Week            Month      Second Month             Year            Never

1           0.160377         0.254717          0.122642         0.226415         0.235849
2           0.123153         0.280788          0.083744         0.241379         0.270936
3           0.063333         0.151667          0.068333         0.361667         0.355000
4           0.051044         0.076566          0.048724         0.308585         0.515081
5           0.020501         0.047836          0.038724         0.198178         0.694761

Column Profiles

                Week            Month      Second Month             Year            Never

1           0.153153         0.117904          0.119266         0.047059         0.030488
2           0.225225         0.248908          0.155963         0.096078         0.067073
3           0.342342         0.397380          0.376147         0.425490         0.259756
4           0.198198         0.144105          0.192661         0.260784         0.270732
5           0.081081         0.091703          0.155963         0.170588         0.371951
```

I det næste output står selve resultaterne fra correspondance analysen. Den første boks ser sådan ud:

```
Inertia and Chi-Square Decomposition

Singular  Principal Chi-
Values    Inertias  Squares Percents   17   34   51   68   85
                                    ----+----+----+----+----+---
0.35420   0.12546   223.188  84.97% **************************
0.13946   0.01945    34.602  13.17% ****
0.04798   0.00230     4.096   1.56%
0.02087   0.00044     0.775   0.30%
          -------   -------
          0.14765   262.661 (Degrees of Freedom = 16)
```

Se bort fra 'Singular Values'. 'Principal Inertias'(lydskrift for 'inertias': 'i-nø-tjas' med tryk på første stavelse) er egenværdierne λ_1 til λ_4.

'Chi-Squares' er fordelingen af den samlede Pearson Q-teststørrelse på de 4 dimensioner, dvs. med betegnelserne i Kapitel 21: Tallet 223.188 i linie 1 er $q - q_0$ for $M_0 = 1$, mens summen 223.188 + 34.602 = 257.790 af de to første tal er $q - q_0$ for $M_0 = 2$. Således fortsættes, så summen af de tre første tal 223.188 + 34.602 + 4.096 = 261.886 er $q - q_0$ for $M_0 = 3$. Summen af alle 'Chi-Squares' er lig med den samlede Pearson Q-teststørrelse 262.661, idet rangen af den matrix, der dekomponeres, er $M_0 = 4$.

'Percents' er 'Chi-Squares' procentvise fordeling på dimensionerne. Skal man bruge $q - q_0$ divideret med q for hver dimension, skal man danne de kumulative procenter i 'Percent'-søjlen.

Den næste del af SAS-output'et ser således ud:

```
Row Coordinates

            Dim1          Dim2

1       0.623811      0.228697
2       0.540517      0.185881
3       0.164169      -.146287
4       -.162536      -.069750
5       -.465369      0.127241


Summary Statistics for the Row Points

         Quality          Mass       Inertia

1       0.957071      0.059584      0.186140
2       0.982533      0.114109      0.256990
3       0.989210      0.337268      0.111654
4       0.925454      0.242271      0.055467
5       0.998137      0.246768      0.389749
```

'Row coordinates' i søjlen under 'Dim1' er de punkter, der skal bruges i correspondance analyse diagrammet som x-koordinater for socialgruppe I til V, her 1 til 5. Tallene 'Column coordinates' i søjlen under 'Dim2' er y-kooordinaterne for socialgruppe I til V i correspondance analyse diagrammet. Disse tal kan, hvis man ønsker en bedre graf, overføres til et regneark, f.eks. EXCEL eller QUATTRO PRO, hvor grafen så laves.

Dernæst i dette SAS-output står en række nøgletal, som man i den franske geometriske tradition lægger stor vægt på.

Søjlen med 'Quality' skal du blot se bort fra. Søjlen betegnet 'Mass' er sådan set triviel, idet den blot viser den relative fordeling af rækkemarginalerne i den observerede tabel. Tallet 0.059584 er således 106/1779 fra den observerede kontingenstabel. Dvs. søjlerne i 'Mass' viser blot $f_{i.}$ for i = 1, ... 5.

I den sidste sølje 'inertia' [i -nø - tja] vises de størrelser, vi i Kapitel 21 kaldte q_i, dvs. hvor meget af værdien af Pearsons Q-teststørrelse q, der skyldes et bidrag fra række i, dvs. med

$$q_i = \sum_{j=1}^{J} \frac{\left(x_{ij} - \frac{x_{i.}\, x_{.j}}{n} \right)^2}{\frac{x_{i.}\, x_{.j}}{n}}$$

er 'inertia' q_i/q, f.eks. 0.186 = 48.89/626.66.

I den næste del af SAS-output'et står bidragene **fra række kategori i til dimension m**, dvs. D^A_{im} med notationen i Kapitel 21, men i SAS-sprog benævnt 'Partial Contribution to Inertia

from the Row Points', samt størrelserne $d^{A}_{im,}$, i Kapitel 21 benævnt **bidragene fra dimension til kategori m**, men i SAS-sprog benævnt 'Squared Cosines for the Row Points'.

```
Partial Contributions to Inertia for the Row Points

            Dim1          Dim2

1       0.184817      0.160222
2       0.265732      0.202704
3       0.072454      0.371071
4       0.051016      0.060599
5       0.425981      0.205405

Squared Cosines for the Row Points

            Dim1          Dim2

1       0.843677      0.113394
2       0.878623      0.103910
3       0.551395      0.437815
4       0.781530      0.143924
5       0.928709      0.069428
```

Man ser at det er socialgruppe V, der bidrager mest til dimension 1, og socialgruppe III, der bidrager mest til dimension 2. På den anden side bidrager dimension 1 mest til at forklare socialgruppe I, II, IV og V, mens dimension 1 og dimension 2 bidrager lige meget til at forklare socialgruppe III.

For søjlevariablene ser SAS-output'et, svarende til output for rækkerne, således ud:

```
Column Coordinates

                       Dim1          Dim2

Week               0.574625      0.167089
Month              0.585062      0.119867
Second Month       0.329072      0.054837
Year               0.082910      -.215870
Never              -.336483      0.070878


Summary Statistics for the Column Points

                    Quality          Mass       Inertia

Week               0.956292      0.062395      0.158255
Month              0.979169      0.128724      0.317573
Second Month       0.903474      0.061270      0.051120
Year               0.999677      0.286678      0.103862
Never              0.999878      0.460933      0.369191
```

'Column coordinates' er de punkter, der skal bruges i correspondance analyse diagrammet som x-koordinater for de 5 'meeting' grupper w, m, s, y og n, i søjlen under 'Dim1', og y-koordinaterne for de 5 'meeting' grupper w, m, s, y og n, i søjlen under 'Dim2'. Også disse tal kan, hvis man ønsker en bedre graf, overføres til et regneark, f.eks. EXCEL eller QUATTRO PRO, hvor grafen så laves.

Dernæst i dette SAS-output står, som for rækkerne, en række nøgletal, som man i den franske geometriske tradition lægger stor vægt på.

Søjlen med 'Quality' skal du igen blot se bort fra. Søjlen betegnet 'Mass' viser blot den relative fordeling af søjlemarginalerne i den observerede tabel. Dvs. med betegnelserne i Kapitel 21, $f_{.j}$ for j = 1, ... 5.

I den sidste sølje, 'inertia', vises størrelserne q_j/q, hvor q_j svarer til q_j i Kapitel 21, dvs. hvor meget af værdien af Pearsons Q-teststørrelse q, der skyldes et bidrag fra søjle j, dvs.

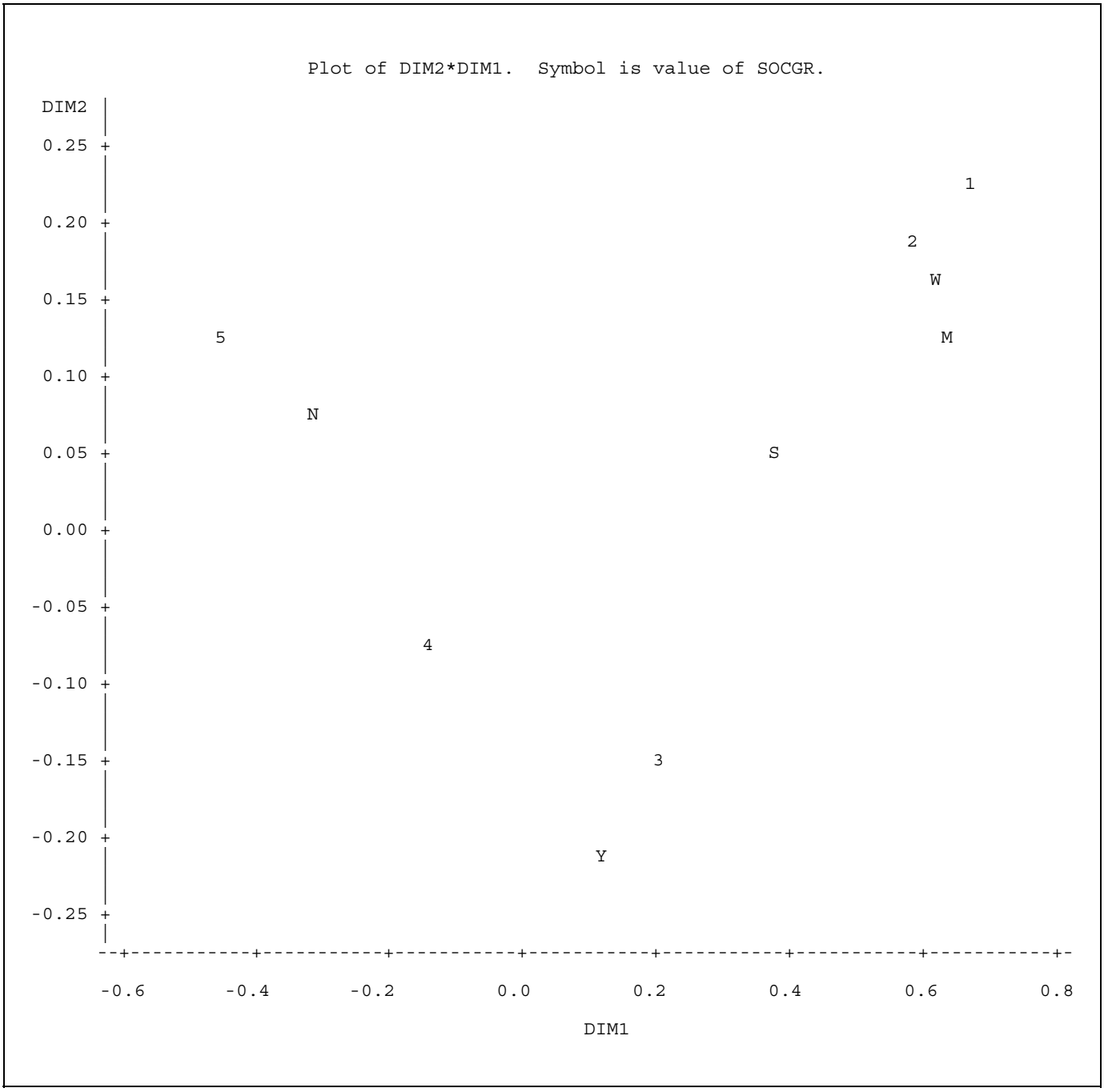
$$q_j = \sum_{i=1}^{I} \frac{\left(x_{ij} - \frac{x_{i.}\, x_{.j}}{n} \right)^2}{\frac{x_{i.}\, x_{.j}}{n}} .$$

I den næste del af SAS-output'et står bidragene **fra søjle kategori j til dimension m**, dvs. D^B_{jm} med notationen i Kapitel21, men i SAS-sprog benævnt 'Partial contribution to Inertia from the Column Points', samt størrelserne $d^B_{jm,}$, i Kapitel 21 benævnt bidragene **til søjle kategori j fra dimension m**, men i SAS-sprog benævnt 'Squared Cosines for the Column Points'.

```
Partial Contributions to Inertia for the Column Points

                        Dim1          Dim2

Week                0.164218      0.089560
Month               0.351212      0.095089
Second Month        0.052886      0.009473
Year                0.015708      0.686828
Never               0.415976      0.119050

Squared Cosines for the Column Points

                        Dim1          Dim2

Week                0.881739      0.074553
Month               0.939723      0.039445
Second Month        0.879063      0.024411
Year                0.128509      0.871168
Never               0.957397      0.042480
```

Man ser at det er 'Month' og 'Never', der bidrager mest til dimension 1, og 'Year', der bidrager mest til dimension 2. På den anden side bidrager dimension 1 mest til at forklare de

3 første og den sidste søjle, mens dimension 2 kun bidrager væsentligt til at forklare søjle 4: 'Year'. Det correspondance analyse diagram denne version af SAS-programmet frembringer, ser det sådan ud:



Det er det samme correspondance analyse diagram som figuren i Kapitel 21, men i en lidt mere primitiv udgave.

23. Multipel correspondance analyse i SAS.

Desværre er der flere måder at lave correspondance analyse for kontingenstabeller med flere end to variable. Den mest udbredte kaldes Multiple Correspondance Analyse, eller blot 'MCA'. Denne metode bygger på den såkaldte 'Burt matrix'. Vi skal her udelukkende betragte tilfældet med 3 variable. I eksemplet bliver det Indkomst, Formue og om man ejer eller lejer sin bolig. Kalder vi disse tre variable 'Indk', 'Form' og 'Ejer/Lejer' dannes Burt matricen ved at stille alle kategorierne for de tre variable op på rad og række, dvs. med 5 indkomst intervaller, 5 formueintervaller og de to kategorier 'Ejer' og 'Lejer', bliver det 12 kategorier i alt. Disse 12 kategorier lader man da være både rækkerne og søjlerne i kontingenstabellen. Dermed får man 9 dele af kontingenstabellen som anskueliggjort i følgende skema:

Datamatricer	Indk	Form	Ejer/Lejer
Indk	D_I	C_{IF}	C_{IE}
Form	$C_{FI} = C_{IF}^T$	D_F	C_{FE}
Ejer/Lejer	$C_{EI} = C_{IE}^T$	$C_{EF} = C_{FE}^T$	D_E

Her er C-matricerne de to-dimensionale kontingenstabeller, der er dannet af de to variable vist i indekset. F.eks. er C_{IF} 5X5 kontingenstabellen dannet af Indkomst og Formue, og C_{IE} 5x2 kontingenstabellen dannet af Indkomst og Ejer/Lejer.

D-matricerne er diagonal matricer, dvs. med nuller uden for diagonalen. I diagonalen står marginalerne for den pågældende variabel. F.eks. er diagonalleddene i D_I de marginale antal for hver af variabel 'Indk's 5 kategorier.

Selve Burt matricen for dette eksempel står i det første SAS-output efter programmet for en MCA-analyse.

SAS-programmet for en MCA-analyse er således ud:

```
 1 title 'Indkomst Formue data. Burt-matrix';
 2 data a;
 3 input v1 $ g h i j k a b c d f e l;
 4   label g = '0 kr'
 5         h = 'lav formue'
 6         i = 'mellem formue'
 7         j = 'høj formue'
 8         k = 'meget høj formue'
 9         a = 'lav indkmst'
10         b = 'ret lav indkomst'
11         c = 'mellem indkomst'
12         d = 'ret høj indkomst'
13         f = 'høj indkomst'
14         e = 'ejer'
15         l = 'lejer';
16 cards;
17 1    1391   0   0   0   0 360 283 269 286 193 347 348
18 2       0 961   0   0   0 196 196 197 220 151 400 366
19 3       0   0 831   0   0 120 134 193 209 174 489 358
20 4       0   0   0 578   0  80  94 127 122 155 533 361
21 5       0   0   0   0 414  39  59  63  77 176 571 278
22 a     360 196 120  80  39 795   0   0   0   0 448 843
23 b     283 196 134  94  59   0 766   0   0   0 374 586
24 c     269 197 193 127  63   0   0 850   0   0 641 189
25 d     286 220 209 122  77   0   0   0 915   0 528  50
26 f     193 151 174 155 176   0   0   0   0 849 369  43
27 e     347 400 489 553 571 448 374 641 528 369 2360  0
28 l     348 366 358 361 278 843 586 189  50  43 0  1711
29 ;
30 proc corresp all data=a out= results;
31   var g h i j k a b c d f e l;
32   id v1;
33 proc plot data= results;
34 plot dim2*dim1 = v1;
35 run;
36 quit;
```

Den eneste måde, dette program adskiller sig fra de andre SAS-programmmer til correspondance analyse, er ved (1) variabelnavnene og (2) den indlæste datamatrix.

I programlinierne, vist i næste SAS output boks, skal der tages hensyn til at rækkevariablene er lig med søjlevariablene

```
 1 input v1 $ g h i j k a b c d f e l;
 2   label g = '0 kr'
 3         h = 'lav formue'
 4         i = 'mellem formue'
 5         j = 'høj formue'
 6         k = 'meget høj formue'
 7         a = 'lav indkmst'
 8         b = 'ret lav indkomst'
 9         c = 'mellem indkomst'
10         d = 'ret høj indkomst'
11         f = 'høj indkomst'
12         e = 'ejer'
13         l = 'lejer';
14 cards;
15 1    1391    0    0    0    0 360 283 269 286 193 347 348
16 2       0 961    0    0    0 196 196 197 220 151 400 366
17 3       0    0 831    0    0 120 134 193 209 174 489 358
18 4       0    0    0 578    0  80  94 127 122 155 533 361
19 5       0    0    0    0 414  39  59  63  77 176 571 278
20 a     360 196 120  80  39 795    0    0    0    0 448 843
21 b     283 196 134  94  59    0 766    0    0    0 374 586
22 c     269 197 193 127  63    0    0 850    0    0 641 189
23 d     286 220 209 122  77    0    0    0 915    0 528  50
24 f     193 151 174 155 176    0    0    0    0 849 369  43
25 e     347 400 489 553 571 448 374 641 528 369 2360   0
26 l     348 366 358 361 278 843 586 189  50  43 0   1711
```

Som det fremgår, sker det ved at variabel betegnelserne i linie 1 er de samme som første tegn i de 12 datalinier, linie 15 - 26. Når g, h, i, j og k er ændret til tallene 1 - 5 i linie 15 - 19, skyldes det, at variabelnavnene i 'input'-linien ikke kan være tal. Men på den simple graf senere, benytter SAS de de 12 værdier af variabel 'v1' i linie 15 - 26 som betegnelser på punkterne. Hvis man ikke ønsker en simpel graf, kan man benytter de samme bogstaver - eller andre - i 'input-linien, linie 1, og i starten af de 12 data linier efter 'cards', dvs. linie 15 - 26.

Den 12x12 matrix, der står i linierne 15 - 26, bortset fra linie nummereringen og værdien af 'v1', er Burt matricen.

De 12 linier med labels i linierne 2 til 13 sørger for informative overskrifter på output.

Selve programmet for PROC CORRESP ser således ud:

```
proc corresp data = a out = results;
  var g h i j k a b c d f e l;
  id v1;
proc plot data = results;
plot dim2*dim1 = v1;
run;
quit;
```

Det svarer til de tidligere kald af PROC CORRESP, men naturligvis med alle 12 involverede variable deklareret i 'var'-linien.

Den første del af output ser lidt anderledes ud for en MCA analyse end for en simpel correspondance analyse.

```
Indkomst Formue data. Burt-matrix

Singular  Principal Chi-
Values    Inertias  Squares Percents    5   10   15   20   25
                                    ----+----+----+----+----+---
0.51706   0.26735   9900.24  23.65% ************************
0.41712   0.17399   6442.95  15.39% ***************
0.38998   0.15208   5631.74  13.45% *************
0.35910   0.12895   4775.27  11.41% ***********
0.34132   0.11650   4314.12  10.31% **********
0.31312   0.09804   3630.57   8.67% *********
0.29658   0.08796   3257.18   7.78% ********
0.26495   0.07020   2599.55   6.21% ******
0.16883   0.02850   1055.53   2.52% ***
0.08261   0.00682    252.72   0.60% *
0.00016   0.00000      0.00   0.00%
          -------   -------
          1.13040   41859.9 (Degrees of Freedom = 121)
```

Når man analyserer en Burt matrix, skal egenværdierne, 'Principal Inertias', og bidragene til Pearsons Q-teststørrelse 'Chi Squares' fortolkes noget anderledes end ved en simpel correspondance analyse. Det er lidt kompliceret, og jeg undlader detaljer. Det er dog værd at bemærke, at selv om 2 dimensioner - skulle man tro udfra 'Percents'-søjlen - kun beskriver ca. 39% af den samlede variation, er det ikke tilfældet. SAS-programmet fortsætter da også ufortøvet med kun at analysere de to første dimensioner.

De koordinater, der skal bruges i correspondance analyse diagrammet bliver nu ens for søjler og rækker. Det fremgår også af SAS output'et - bortset fra nogle mindre beregningsunøjagtigheder - i den næste output boks.

Row Coordinates

	Dim1	Dim2
1	-.309701	-.588644
2	-.096359	-.047709
3	0.228350	0.052766
4	0.192229	0.314509
5	0.422224	0.492901
a	-.721117	0.066072
b	-.581778	0.095500
c	0.295464	-.244177
d	0.576445	-.976775
f	0.829469	0.931470
e	0.441247	-.013471
l	-.744341	0.241864

Column Coordinates

	Dim1	Dim2
0 kr	-.309558	-.588673
lav formue	-.096179	-.047634
mellem formue	0.228559	0.052936
høj formue	0.196627	0.312835
meget høj formue	0.422513	0.493437
lav indkmst	-.721125	0.066435
ret lav indkomst	-.581875	0.095844
mellem indkomst	0.295261	-.243663
ret høj indkomst	0.576206	-.976355
høj indkomst	0.829106	0.931823
ejer	0.440969	-.014764
lejer	-.744737	0.241795

Også bidragene fra de 12 række kategorier til dimension 1 og dimension 2 og bidragene fra dimension 1 og dimension 2 til de 12 række kategorier bliver de samme som bidragene fra de 12 søjle kategorier til dimension 1 og dimension 2 og bidragene fra dimension 1 og dimension 2 til de 12 søjle kategorier. Dette fremgår af den næste output boks.

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
1	0.033686	0.186993
2	0.002520	0.000949
3	0.013209	0.001084
4	0.007651	0.031473
5	0.030198	0.063237
a	0.151325	0.001952
b	0.085196	0.003528
c	0.022300	0.023403
d	0.080788	0.356435
f	0.146635	0.284143
e	0.139236	0.000199
l	0.287256	0.046604

Squared Cosines for the Row Points

	Dim1	Dim2
1	0.073157	0.264287
2	0.006643	0.001629
3	0.039585	0.002114
4	0.029274	0.078363
5	0.125205	0.170631
a	0.521683	0.004380
b	0.276470	0.007450
c	0.063248	0.043197
d	0.172700	0.495868
f	0.273556	0.344973
e	0.605099	0.000564
l	0.781898	0.082556

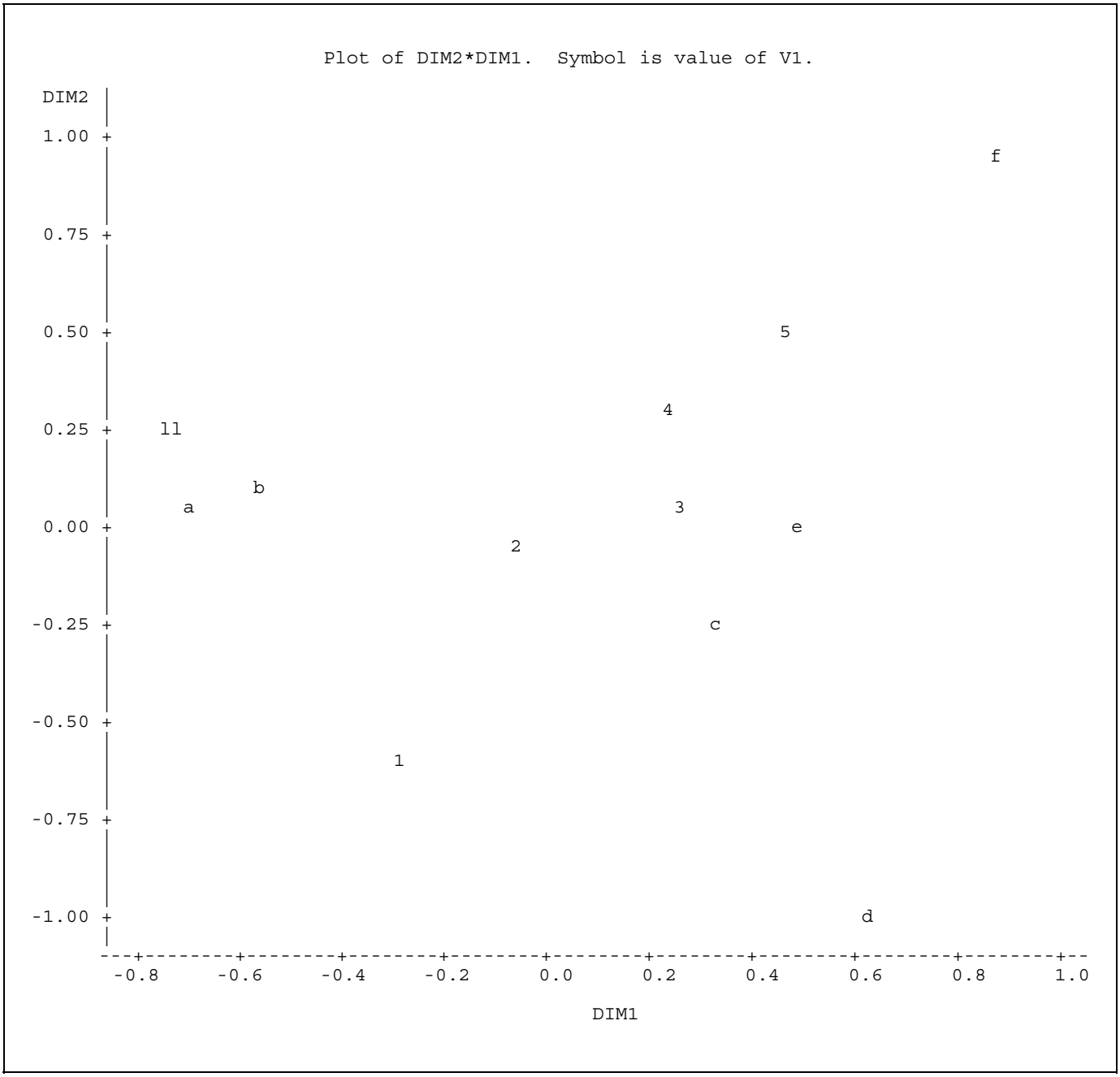
Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
0 kr	0.033654	0.187011
lav formue	0.002511	0.000946
mellem formue	0.013234	0.001091
høj formue	0.008084	0.031443
meget høj formue	0.030239	0.063374
lav indkomst	0.151328	0.001974
ret lav indkomst	0.085224	0.003553
mellem indkomst	0.022270	0.023305
ret høj indkomst	0.080721	0.356129
høj indkomst	0.146506	0.284358
ejer	0.138667	0.000239
lejer	0.287562	0.046578

Squared Cosines for the Column Points

	Dim1	Dim2
0 kr	0.073098	0.264343
lav formue	0.006620	0.001624
mellem formue	0.039674	0.002128
høj formue	0.030735	0.077801
meget høj formue	0.125528	0.171208
lav indkomst	0.521810	0.004429
ret lav indkomst	0.276581	0.007504
mellem indkomst	0.063185	0.043031
ret høj indkomst	0.172581	0.495510
høj indkomst	0.273264	0.345167
ejer	0.606139	0.000679
lejer	0.781778	0.082408

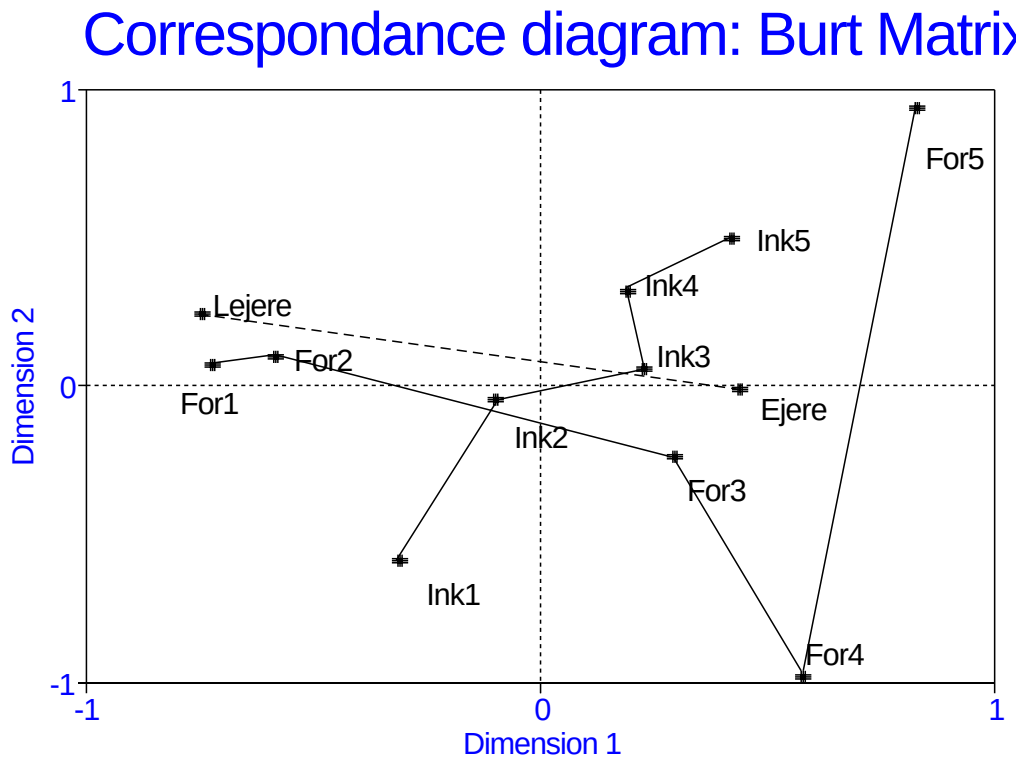
Det simple correspondance analyse diagram for en MCA analyse af de tre variable 'Indk', 'Form' og 'Ejer/Lejer' ser sådan ud.



Det dobbelte 'l', skyldes de regneunøjagtigheder, som vi bemærkede oven for, vedr. plot-værdierne for rækker og søjler. Bortset fra dette viser figuren, at de 5 punkter for Indkomst, benænt 1 til 5, følger de 5 punkter for Formue, benævnt a til e, meget godt, men dog med en snævrere variation for Indkomst, især på dimension 1. Samtidig fremgår det, at punktet 'e' for 'Ejere' så klart ligger tæt ved de højeste formuer, men ikke rigtigt placerer sig i forhold til indkomster. Vi vender tilbage til dette forhold. Punktet for 'Lejere' (dobbeltpunktet 'll')

befinder sig tydeligvis nær de laveste formuer. Dette er der intet overraskende i. For langt de fleste danskere er formuen jo bundet i ejerboligen, hvis man har en sådan.

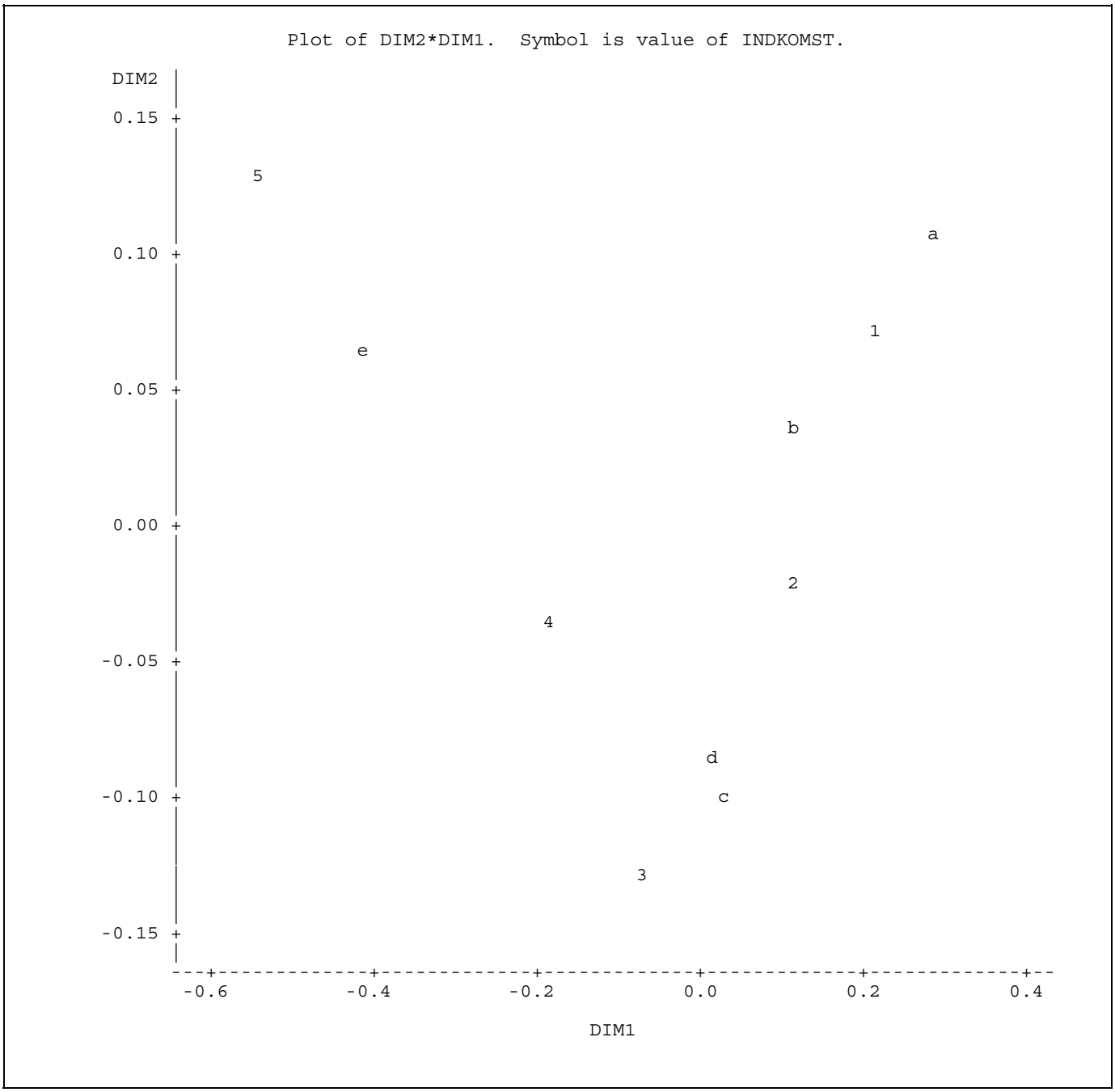
Disse kommentarer bliver tydeligere ved at se på følgende figur tegnet i regnearket Quattro Pro.



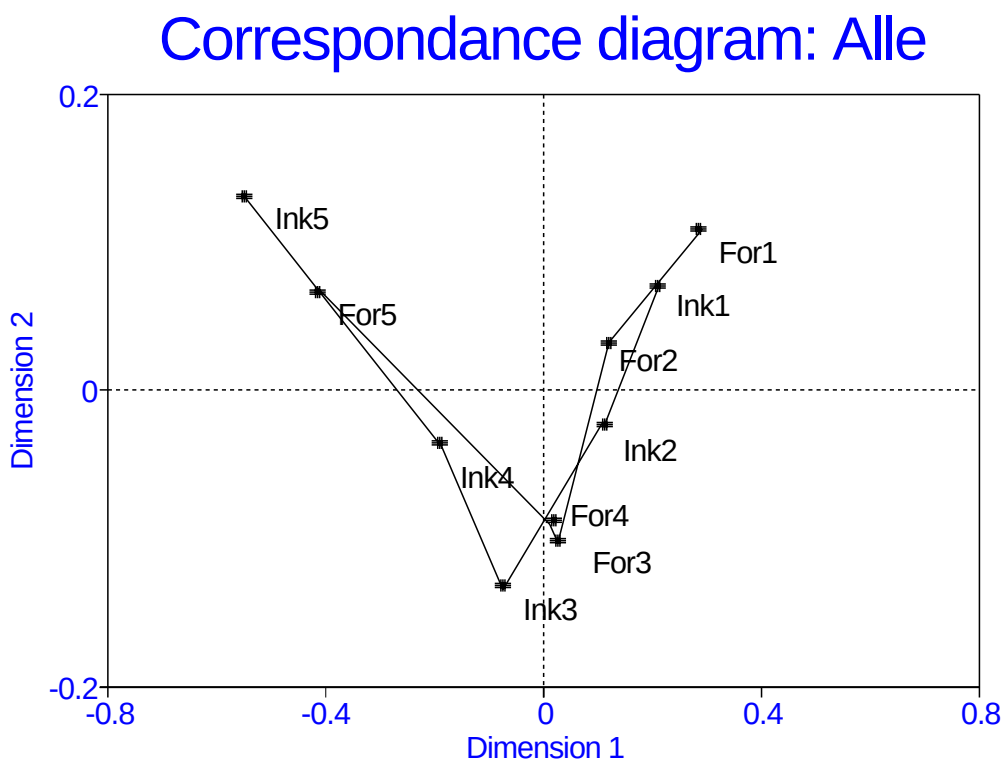
Laver man en simpel correspondance analyse på ’Indk’ mod ’Form’, skal man benytte følgende SAS program:

```
title 'Indkomst Formue data. Alle';
data a;
input indkomst $ a b c d e;
  label a = 'a=0 kr'
        b = 'b=lav formue'
        c = 'c=mellem formue'
        d = 'd=høj formue'
        e = 'e=meget høj formue' ;

cards;
1      360 283 269 286 193
2      196 196 197 220 151
3      120 134 193 209 174
4       80  94 127 122 155
5       39  59  63  77 176
;
proc corresp all data = a out = results;
  var a b c d e;
  id indkomst;
proc plot data= results;
plot dim2*dim1 = indkomst;
run;
```

Grunden til, at 'labels' sektionen er skrevet på denne måde, er, at den simple SAS-graf benytter det første bogstav i 'labels'-linierne, og det første tegn i 'cards'-linierne til at skrive ind på grafen. Man får derfor - med dette program - grafen vist neden for.



Tegnet i Quattro Pro bliver tegningen en del pænere:



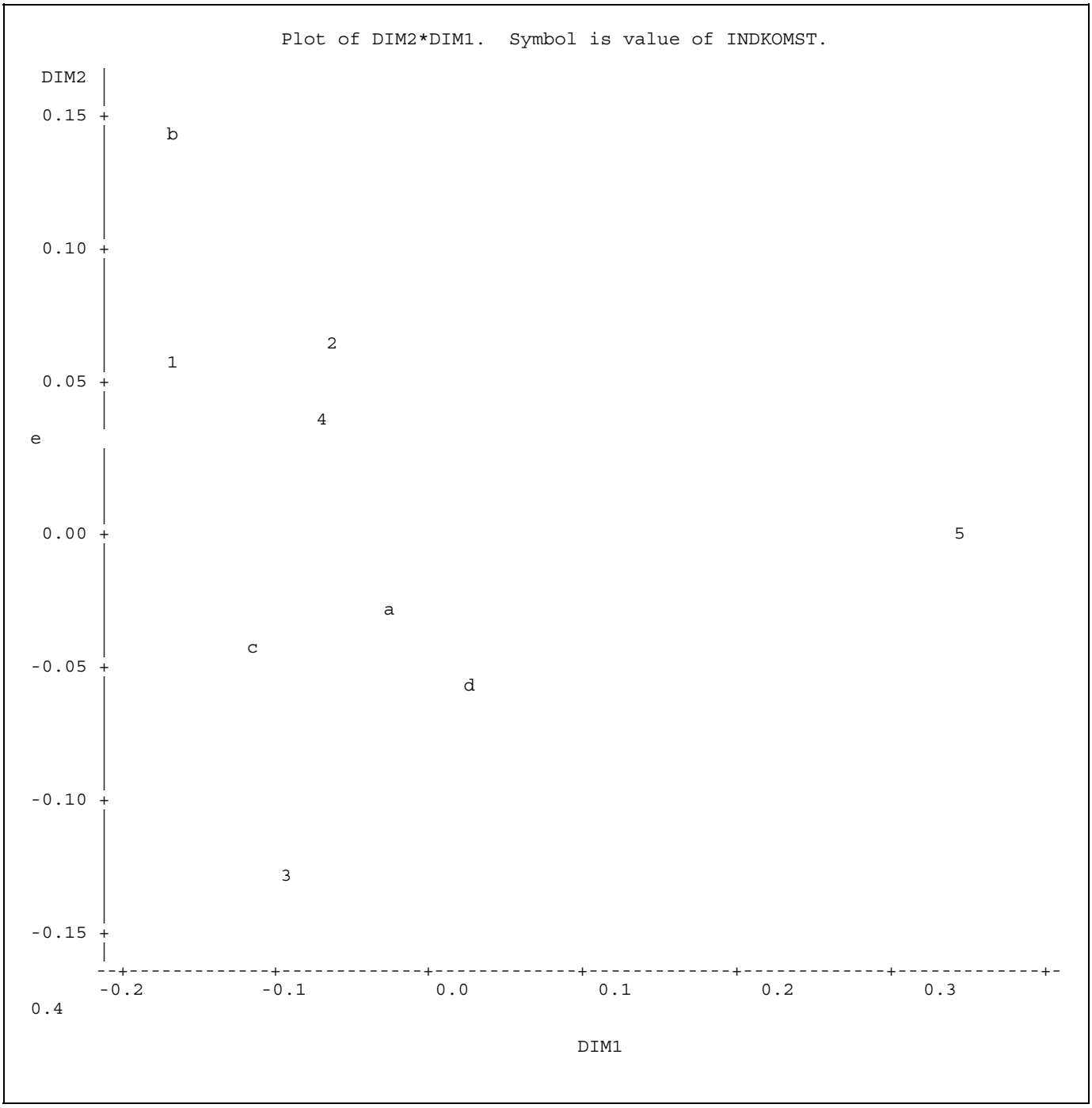
Her ser vi, at Indkomst, punkterne 1 til 5, følges meget pænt med Formue, punkterne a til e.

Ét af kritikpunkterne ved at benytte MCA, dvs. en analyse af Burt matricen, er, at man egentlig slet ikke laver en multidimensional analyse. I Burt matricen indgår nemlig kun de marginale kontingenstabeller mellem de variable, og ikke den fler-dimensionale kontingenstabel mellem alle de variable. For at illustrere dette, skal vi se på to correspondance analyser, der bygger på den 3-dimensionale kontingenstabel mellem 'Indk', 'Form' og 'Ejer/Lejer'. Man kan således dele den 3-dimensionale tabel op i 2 dele. Den første er kontingenstabellen mellem 'Indk' og 'Form' alene for Ejere. Den anden er kontingenstabellen mellem 'Indk' og 'Form' alene for Lejere. Det giver et ganske overraskende resultat i forhold til MCA-analysen. De to programmer ser sådan ud:

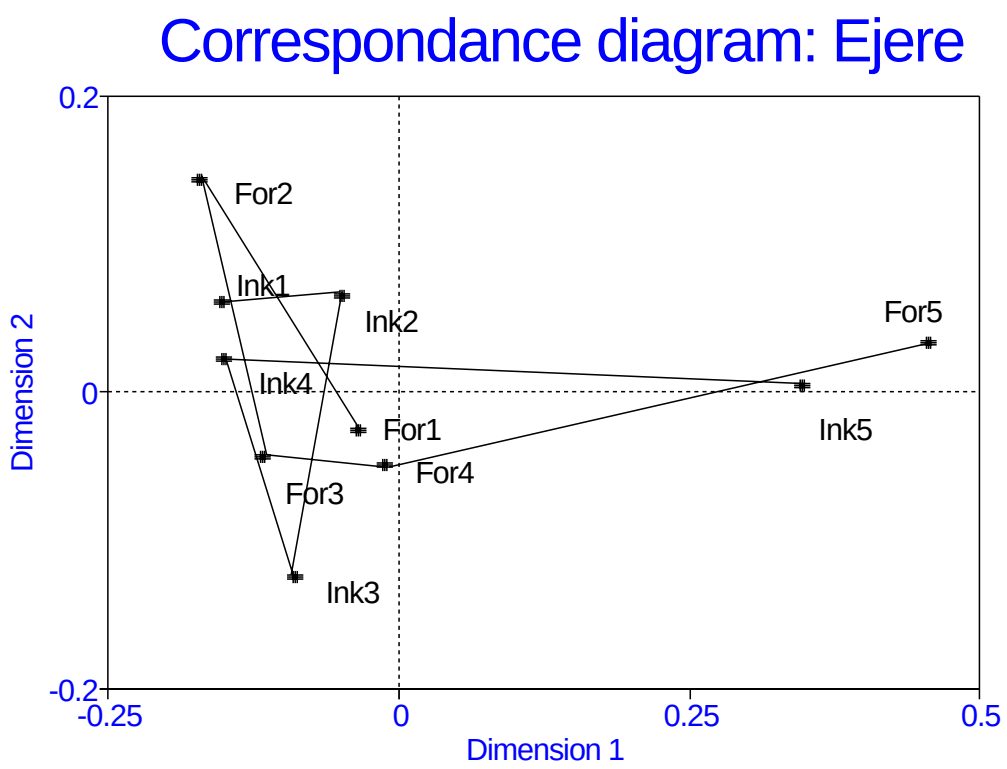
Kun for ejere:

```
title 'Indkomst Formue data. Kun ejere';
data a;
input indkomst $ a b c d e;
  label a = 'a=0 kr'
        b = 'b=lav formue'
        c = 'c=mellem formue'
        d = 'd=høj formue'
        e = 'e=meget høj formue' ;

cards;
1     69 70 98 75 35
2     68 76 113 87 56
3     98 64 153 120 54
4     110 100 155 115 73
5     103 64 122 131 151
;
proc corresp all data=a out= results;
  var a b c d e;
  id indkomst;
proc plot data= results;
plot dim2*dim1 = indkomst;
run;
```

Med correspondance analyse diagram:



Eller tegnet i Quattro Pro.

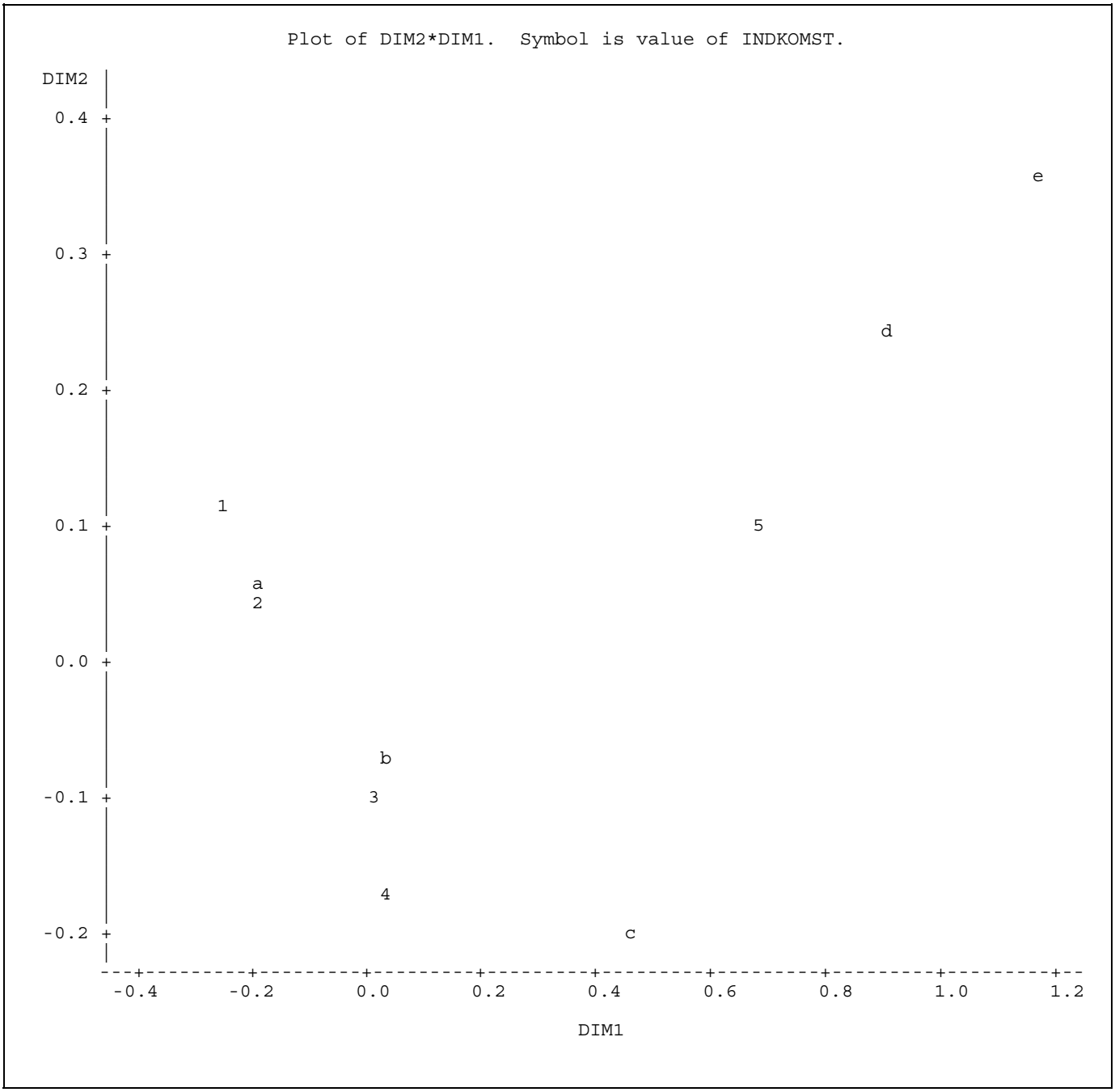


Som man ser, følges Indkomst og Formue slet ikke ad. Grunden er den oplagte, at de mange ældre boligejere, for hvilke formuen især står i deres bolig, ved deres tilbagetrækning fra arbejdsmarkedet generelt går kraftigt ned i indtægt, mens formuen stadig er relativt høj, idet den er bundet i ejendommen. Kun det laveste indkomst punkt, punktet '5', ligger tæt ved punktet 'e' for den laveste formue.

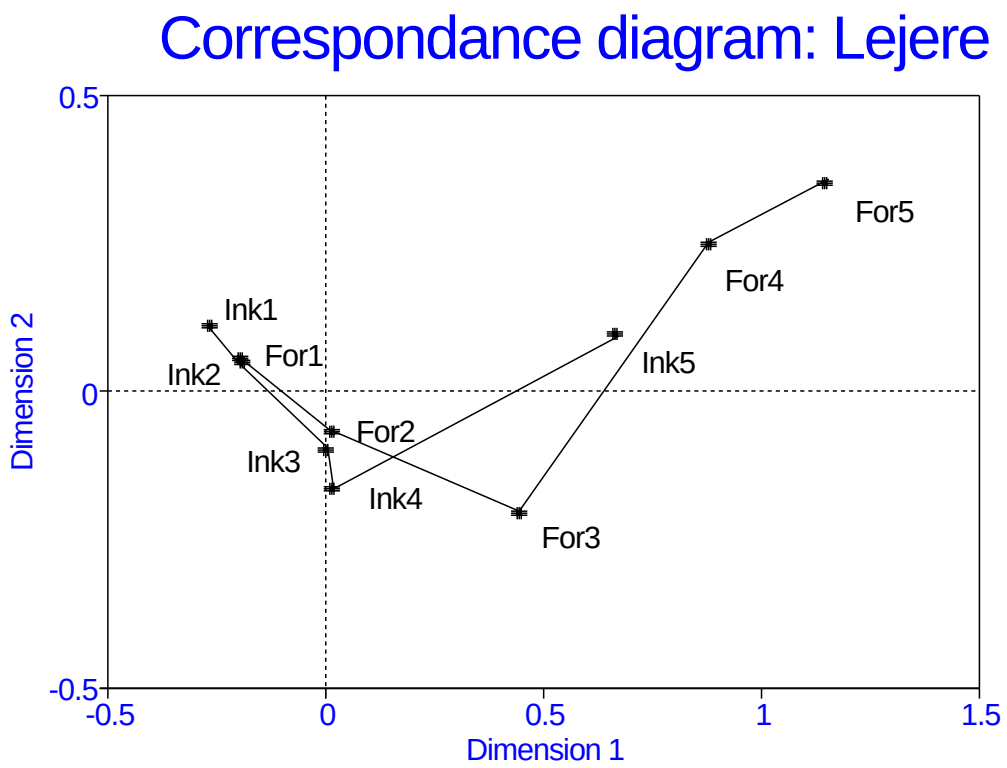
For kun lejere bliver SAS programmet for en correspondance analyse:

```
title 'Indkomst Formue data. Kun lejere';
data a;
input indkomst $ a b c d e;
  label a = 'a=0 kr'
        b = 'b=lav formue'
        c = 'c=mellem formue'
        d = 'd=høj formue'
        e = 'e=meget høj formue' ;

cards;
1     291 126  22   5   4
2     215 120  21   7   3
3     171 133  40   7   7
4     176 120  54   7   4
5      90  87  52  24  25
;
proc corresp all data=a out= results;
  var a b c d e;
  id indkomst;
proc plot data= results;
plot dim2*dim1 = indkomst;
run;
```

For denne gruppe bliver correspondance analyse diagrammet:


Plot of DIM2*DIM1. Symbol is value of INDKOMST.
DIM2
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
e
d
1
5
a
2
b
3
4
c
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
DIM1

Eller tegnet i Quattro Pro.



Nu er billedet til gengæld nogenlunde svarende til det generelle billede. Her gælder fænomenet med at formuen - efter pensionering - forbliver stor, mens indkomsten går ned, jo ikke.

Stikordsregister.