Faktoranalyse - en metode til datakomprimering

Af Niels J. Blunch*)

Resumé

Et ofte tilbagevendende problem ved behandling af data fra markedsanalyser, er den store datamængde, der kan gøre det vanskeligt at overskue sammenhænge mellem de mange variable, der indgår i en sådan analyse.
Man løber en risiko for at »drukne i data«, og der opstår behov for fremgangsmåder, der kan hjælpe til at besvare spørgsmål som: (1) hvilke variable skal jeg vælge til videre analyse? og (2) hvad er det egentlig jeg måler med alle de mange variable?
I denne artikel skal gives en kortfattet introduktion til den såkaldte faktoranalyseteknik, der i de senere år har fundet stadig større anvendelse ved løsning af disse problemer. Artiklen er disponeret som følger:
1. Om univariable og multivariable statistiske modeller. – 2. Problemet. – 3. Faktoranalyse og komponentanalyse. – 4 En geometrisk fortolkning af de principale komponenter. – 5. Eksempler på anvendelse af faktoranalyse inden for marketingområdet. – Matematisk appendix.

1. Om univariable og multivariable statistiske modeller

De statistiske metoder, der gennem årene har vundet indpas i markedsanalysearbejdet og som i dag udgør en væsentlig del af enhver lærebog inden for området, er karakteriseret ved stort set at være begrænset til univeriable metoder, d.v.s. til metoder, hvor der kun knyttes én statistisk variabel til hvert af de elementer (personer, virksomheder el. lign.), der er genstand for undersøgelse.

*) Cand. merc., amanuensis ved Institut for Markedsøkonomi, Handelshøjskolen i Århus.

Faktoranalyse - en metode til datakomprimering

Af Niels J. Blunch*)

Resumé

Et ofte tilbagevendende problem ved behandling af data fra markedsanalyser, er den store datamængde, der kan gøre det vanskeligt at overskue sammenhænge mellem de mange variable, der indgår i en sådan analyse.
Man løber en risiko for at »drukne i data«, og der opstår behov for fremgangsmåder, der kan hjælpe til at besvare spørgsmål som: (1) hvilke variable skal jeg vælge til videre analyse? og (2) hvad er det egentlig jeg måler med alle de mange variable?
I denne artikel skal gives en kortfattet introduktion til den såkaldte faktoranalyseteknik, der i de senere år har fundet stadig større anvendelse ved løsning af disse problemer. Artiklen er disponeret som følger:
1. Om univariable og multivariable statistiske modeller. – 2. Problemet. – 3. Faktoranalyse og komponentanalyse. – 4 En geometrisk fortolkning af de principale komponenter. – 5. Eksempler på anvendelse af faktoranalyse inden for marketingområdet. – Matematisk appendix.

1. Om univariable og multivariable statistiske modeller

De statistiske metoder, der gennem årene har vundet indpas i markedsanalysearbejdet og som i dag udgør en væsentlig del af enhver lærebog inden for området, er karakteriseret ved stort set at være begrænset til univeriable metoder, d.v.s. til metoder, hvor der kun knyttes én statistisk variabel til hvert af de elementer (personer, virksomheder el. lign.), der er genstand for undersøgelse.

*) Cand. merc., amanuensis ved Institut for Markedsøkonomi, Handelshøjskolen i Århus.

Det er velkendt, at dette gælder for hele dataindsamlingsområdet, hvor teorierne omkring de repræsentative undersøgelser (stratifiering, klyngeudvælgelse etc.) implicit bygger på den forudsætning, at kun en enkelt variabel er genstand for undersøgelse; medens teorien stort set er ude af stand til at give velfunderede operationelle anvisninger på, hvorledes man skal løse de problemer, der opstår, når dette ikke er tilfældet.
Inden for det område, der omfatter analysen af de indsamlede data, gør samme tendens sig imidlertid gældende: De statistiske modeller, man anvender, er næsten alle univariable. Dette gælder endog modeller, hvori indgår måling af flere egenskaber. F. eks. er den almindelige multiple regressionsmodel

$$y = a + \Sigma_j b_j x_j + \varepsilon, \text{ hvor } E(\varepsilon) = 0$$

ikke multivariabel i statistisk forstand, da den eneste statistiske variabel er ε.
Da man imidlertid inden for markedsanalysen næsten altid er interesseret i at måle *flere* egenskaber ved de elementer(f. eks. personer, husstande eller virksomheder), der er genstand for undersøgelse, opstår der behov for en teoribygning for *multivariable statistiske modeller*. En sådan har da også været under stadig udvikling siden tyverne. Det er dog først inden for den sidste halve snes år, at de udviklede metoder har fundet udbredelse inden for markedsanalysen, men så er der til gengæld også tale om en eksplosionsagtig udvikling, således at man i dag næppe kan åbne et markedsanalysetidsskrift uden at støde på betegnelser som diskriminantanalyse og faktoranalyse.
Hvorfor er der da gået så lang tid inden metoderne har fundet anvendelse inden for markedsanalysearbejdet, når der nu er et så enormt behov? Hertil kan gives to svar:

For det første støder opbygningen af en statistisk teori for multivariable fordelinger på større *teoretiske* vanskeligheder, end tilfældet har været med den univariable teori: De matematiske formuleringer bliver mere komplicerede, og ikke mindst opstår der problemer, når de opnåede resultater skal fortolkes. Udviklingen er derfor gået langsomt, og problemer, der for længst er klaret for de univariable tilfælde, er stadig uløste.

For det andet medfører anvendelsen af de udviklede metoder så enorme beregningsarbejder, at alene dette rent *praktiske* forhold indtil fremkomsten af EDB har kunnet sætte en stopper for en mere udbredt anvendelse af de teoretiske resultater. Det er i den forbindelse værd at gøre sig klart, at selv om nogle af de pågældende metoder har været anvendt i antropologi, biologi og botanik i snart et halvt hundrede år

og først har nået markedsanalysen via psykologers og sociologers anvendelse, er det først i de sidste år, at metoderne har fundet virkelig udbredt anvendelse inden for såvel disse som andre videnskaber, og der er ingen tvivl om, at dette må sættes i forbindelse med EDB-anlæggenes voksende udbredelse.

Jeg har i en tidligere artikel (En metode til klassificering af kundeemner. Erhvervsøkonomisk Tidsskrift nr. 2, 1968, p. 81) beskæftiget mig med den multivariable model, der går under navnet *multipel diskriminantanalyse* og vil i det følgende give en introduktion til den vel nok mest anvendte af alle multiple modeller, nemlig *faktoranalysen* eller rettere *komponentanalysen* (en nærmere præcisering af disse begreber foretages i afsnit 3). Dog først lige et par ord om multivariable sandsynlighedsfordelinger.

Foretages der måling af p egenskaber ved de elementer, der er genstand for undersøgelse og opfattes disse p målinger som statistiske variable, bliver deres simultane sandsynlighedsfordeling p-dimensional, medens det sæt målinger, der er knyttet til det enkelte element, kan opfattes som en søjlevektor i p dimensioner:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_i \\ \cdot \\ \cdot \\ x_p \end{pmatrix} \tag{1.1}$$

Vektorer betegnes her som overalt i det følgende med tykke små bogstaver, medens matricer betegnes med tykke store bogstaver. Vektoren **x** opfattes da som en statistisk variabel, og dens forventede værdi er den vektor, hvis elementer er de forventede værdier for elementerne i (1.1), altså:

$$E(\mathbf{x}) = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \cdot \\ \cdot \\ E(x_i) \\ \cdot \\ \cdot \\ E(x_p) \end{pmatrix} \tag{1.2}$$

Som bkendt defineres variansen på en statistisk variabel x som:

$$\mathrm{Var}(x) = E(x - E(x))^2 \tag{1.3}$$

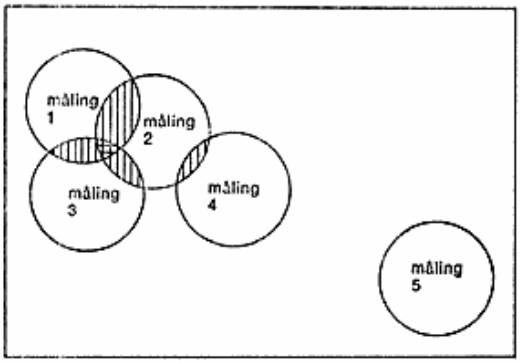
En analog anvendelse af denne definition på det multivariable tilfælde fører til nedenstående definition på *kovariansmatricen,* der er den multivariable analog til variansen i det univariable tilfælde:

$$\mathbf{V} = \mathrm{E}\left[(\mathbf{x}-\mathrm{E}(\mathbf{x}))\ (\mathbf{x}-\mathrm{E}(\mathbf{x}))^{\mathrm{T}}\right] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1j} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2j} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{i1} & \sigma_{i2} & \cdots & \sigma_{ij} & \cdots & \sigma_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pj} & \cdots & \sigma_{pp} \end{pmatrix} \qquad (1.4)$$

hvor T angiver, at vektoren er transponeret. σ_{ij} betegner kovariansen mellem den i'te og den j'te variabel. Hvor $i = j$ angiver σ_{ij} variansen i den i'te marginalfordeling. I overensstemmelse med den almindelige konvention skrives som regel σ_i^2 i dette tilfælde.

2. Problemet

Ved markedsanalyse såvel som ved anden dataindsamling, hvor der indsamles data om flere egenskaber ved de elementer, der er genstand for undersøgelse, vil man ofte have en mere eller mindre velbegrundet mistanke om, at i hvert fald nogle af de foretagne målinger ved det enkelte element er korrelerede. Skulle dette være tilfældet, indeholder de indsamlede data redundant information forstået på den måde, at flere af *målingerne* i større eller mindre omfang dækker over samme grundlæggende *egenskaber* ved det enkelte element. Dette kan afbildes i et Venn-diagram, hvor hver enkelt cirkel illustrerer den mængde af egenskaber, der registreres ved den pågældende måling.



Figur 1.

Faktoranalysens problemstilling.

Vi ser, at i det afbildede tilfælde registrerer målingerne 1 og 2, 1 og 3, 2 og 3 samt 2 og 4 parvis nogle fælles egenskaber (angivet ved skravering), ligesom en mindre mængde egenskaber registreres af både måling 1, 2 og 3 (angivet ved dobbelt skravering). Derimod registrerer måling 5 udelukkende egenskaber, der ikke registreres af andre målinger. *Faktoranalyse* er netop den fælles betegnelse for en række teknikker til at kortlægge denne underliggende struktur i elementernes egenskaber. En sådan kortlægning har stor teoretisk og praktisk interesse.

For det første kan det have interesse at fastslå, hvor mange »egenskaber«, vi egentlig måler ved hvert element. Er der overlapninger således som vist i figur 1, er antallet af egenskaber mindre end antallet af målinger. Det vil derfor eventuelt være muligt – med kendskab til måleteknikken – at identificere (»navngive«) disse egenskaber.

For det andet må tilstedeværelsen af redundant information betyde, at det er muligt at komprimere de indsamlede data uden at miste for meget information. En sådan komprimering må kunne ske enten ved at nogle af målingerne bortkastes, eller ved at der skabes en enkelt eller nogle få nye ukorrelerede variable, der er funktioner af de oprindelige. Faktoranalysen anviser netop, hvorledes denne komprimering kan foretages med mindst mulig tab af information.
I denne sidste anvendelse er faktoranalysen altså ikke nogen erstatning for eller konkurrent til de gammelkendte analysemetoder (som f. eks. regressionsanalyse), men derimod et glimrende supplement til disse på de tidligere og mere explorative stadier i analysearbejdet.

Disse er jo ofte karakteriseret ved, at man endnu ikke har klart formulerede hypoteser med hensyn til hvilke variable, der er relevante for problemstillingen. Under den indledende søgen efter brugbare arbejdshypoteser, vil man derfor ofte starte med at kaste det store men ret grove net ud: En lang række forhold gøres til genstand for en ret overfladisk undersøgelse, og på grundlag heraf vælges så de variable, der findes relevante for en nøjere undersøgelse.

3. Faktoranalyse og komponentanalyse

Der må på dette sted skelnes mellem *faktoranalyse* og *komponentanalyse*. Som regel benyttes begreberne i flæng, eller faktoranalyse anvendes som en fællesbetegnelse. I virkeligheden dækker de to betegnelser to forskellige problemstilinger, som det nok er værd at holde sig for øje, selv om de som regel sammenblandes.

Ved faktoranalyse har vi a priori opstillet den teori, at de p målinger, vi foretager, i virkeligheden kun dækker over m forskellige egenskaber ved hvert element ($m < p$), således at i hvert fald nogle af de p må-

linger måler fælles egenskaber. Spørgsmålet er nu: Hvorledes er sammenhængene mellem et elements m egenskaber og de foretagne p målinger?
Dette kan lidt mere formelt udtrykkes i følgende model:

$$x_i = \sum_{k=1}^{m} l_{ik} f_k + \varepsilon_i \qquad i = 1, 2, \ldots\ldots, p \qquad (3.1)$$

Denne model fortolkes som følger: Resultatet af den i'te måling opfattes som summen af

(1) m *ukorrelerede* egenskaber, f_k, der er fælles for flere målinger idet hver af egenskaberne tillægges vægten l_{ik}.

(2) En risidualværdi, ε_i, der repræsenterer enten en egenskab, der *kun* registreres af måling i eller en variation som følge af måleusikkerhed eller eventuelt en kombination af disse to virkninger.

De m egenskaber kaldes *fælles faktorer* (common factors) og de mp vægte kaldes *faktorvægte* (factor loadings).
Bemærk at de p målinger, x_i, udtrykker m variable, der ikke er direkte konstaterbare. Derimod udtrykker (3.1) en *model,* og problemet er at estimere denne models parametre, l_{ik}, samt at teste modellen. For at kunne gøre dette, er det nødvendigt at opstille forudsætninger vedrørende fordelingen for **x.** Denne forudsættes som regel at være multinormal, da dette foruden at være det eneste gennemarbejdede tilfælde, ofte vil være den rimeligste forudsætning.
Af ovenstående følger, at der er uendelig mange *faktormatricer* $\mathbf{L} = (l_{ik})$, der opfylder (3.1). Faktoranalysearbejdet falder derfor i to faser. Først bestemmes en vilkårlig faktormatrix, **L.** De ved **L** bestemte faktorer vil kun sjældent kunne fortolkes meningsfuldt. Derfor vil man i anden fase søge at transformere **L** til en ny matrix, der med bibeholdelse af samme »usikkerhed«, ε, er lettere at fortolke. Idealet er, at den ny matrix – den såkaldte *roterede faktormatrix* – har rækker (eller søjler) hvori der kun forekommer én værdi nær 1, og at alle andre elementer i rækken (søjlen) har værdier nær 0. D.v.s. at hver måling kun er stærkt korreleret med én faktor (eller at hver faktor kun er stærkt korreleret med én måling). Det vil da være lettere, med kendskab til måleteknikken og til den foreliggende teoribygning inden for det område, der er genstand for undersøgelse, at »navngive« faktorerne. Der kommer ved denne *rotation* – der foregår ved fortsatte multiplikationer af **L** med ortogonale matricer – et subjektivt element ind i faktoranalysen, som man ingenlunde må være blind for.
Ved komponentanalyse har vi derimod ikke a priori opstillet nogen teori angående antallet af fælles faktorer, men kan tværtimod have til hensigt at benytte de indsamlede data som udgangspunkt for opstilling af en sådan model.

Da vi ikke a priori har fastlagt m, vil vi i stedet søge at udtrykke målingerne, x_i , ved p *ukorrelerede* variable, y_k , altså:

$$x_i = \sum_{k=1}^{p} l_{ik} y_k \qquad i = 1, 2, \ldots\ldots, p \qquad (.32)$$

y_k kaldes i komponentanalysen *principale komponenter* (principal components). Det er muligt, at vi kan udtrykke de p målinger, x_i, i færre end p variable, y_k, og i så fald har vi opnået en reduktion i vort problems dimension (ligesom ved faktoranalyse). Dette vil imidlertid være undtagelsen, og i stedet vil vi søge en reduktion »med tilnærmelse« ved følgende teknik:

Vi har i (3.2) udtrykt x'erne som funktioner af y'erne, men det er naturligvis også muligt at udtrykke y'erne som funktioner af x'erne:

$$y_k = \sum_{i=1}^{p} a_{ik} x_i \qquad k = 1, 2, \ldots\ldots, p \qquad (3.3)$$

Vi kan betragte (3.3) som en transformation af de p x'er til p nye variable y_k. Grunden til at vi foretager målingerne x_i er, at de varierer fra element til element, således at vi kan studere sammenhængene i variationerne i x_i. Hvis det var således, at en enkelt måling gav samme resultat ved alle elementer, var der ingen grund til at foretage den.
Hvis vi nu vælger koefficienterne a_{ik} på en sådan måde, at den første variabel, y_1, får størst mulig varians, den anden, y_2, størst mulig varians og samtidig ikke er korreleret med y_1. den tredie, y_3, størst mulig varians og samtidig ikke er korreleret med y_1 og y_2, o.s.fr. er det muligt, at de første få (f. eks. 2–3) variable får tillagt »næsten« hele den variation, der er i materialet, medens de sidste p–2 eller p–3 variable kun tegner sig for nogle få procent af variationen, således at det er muligt at se bort fra dem og koncentrere sig om nogle få variable, y_k, idet det kan bevises, at summen af varianserne for y'erne er lig summen af varianserne for x'erne.
På den måde får vi altså ved at se bort fra nogle af y'erne summen i (3.2) til at indeholde færre end p led, mod at lighedstegnet nu kun gælder »med tilnærmelse«. Selv om man ikke kan (eller vil) »ofre« nogle af y'erne, men i stedet vælger at arbejde videre med dem alle p, vil det forhold, at de – i modsætning til x'erne – er ukorrelerede være en fordel i sig selv ved mange former for videre statistisk analyse af materialet.
I modsætning til, hvad der er tilfældet med faktoranalysen, er bestemmelsen af matricen $\mathbf{L} = (l_{jk})$ eller den dermed forbundne matrix $\mathbf{A} = (a_{jk})$ entydig bortset fra fastlæggelsen af en skala. Dette betyder, at fortsat rotation for at finde mere »meningsfulde« dimensioner, vil medføre, at de principale komponenters egenskab, at de er ordnet efter

faldende (og maksimal) varians, falder væk. Dette er et forhold, som tilsyneladende ofte overses, når komponentanalyse bruges som datakomprimerende metode.

Læg mærke til, at ved faktoranalyse arbejder vi os *fra en model mod nogle data,* medens vi ved komponentanalysen arbejder *fra data mod model.* Det er altså ikke ved komponentanalyse nødvendigt med nogen form for forudsætninger angående variablerne x_i. Kun deres kovariansmatrix er af interesse, medens deres statistiske fordeling i øvrigt er uinteressant.

Jeg har her prøvet at drage et skel mellem faktor- og komponentanalyse. Som tidligere nævnt anvendes disse udtryk oftest i flæng, eller man benytter betegnelsen faktoranalyse om dem begge. Dette skyldes to forhold:

(1) I praksis vil man oftest arbejde »i ring«: Nogle data giver anledning til opstilling af en model (komponentanalyse), og senere søger man at verificere modellen på andre data (faktoranalyse). Dette fører måske til modifikation af modellen, der så igen søges verificeret o.s.fr.

(2) Jeg har ovenfor skitseret formålene og principperne i de to former for analyse, medens de til rådighed stående beregningsteknikker ikke er behandlet. Af sådanne findes der et utal, og det gælder for en del af disse, at de kan anvendes i nogenlunde samme form både til komponent- og faktoranalyse, hvilket medvirker til sammenblandingen af begreberne. Den teknik, der efter fremkomsten af EDB sædvanligvis regnes for den bedste, har navnet »principale komponenters metode«. Den er i sit udgangspunkt en komponentanalyse, og vil i appendix blive gennemgået i denne form, men den finder efterhånden i tillempet form også anvendelse inden for det, der ovenfor er kaldt faktoranalyse.

4. En geometrisk fortolkning af de principale komponenter

Begrænser vi os til det to-dimensionale tilfælde – hvor der foretages to målinger, x_1 og x_2, på hvert element, kan resultatet af en stikprøveundersøgelse baseret på n observationer afbildes i et sædvanligt retvinklet koordinatsystem, idet vi afbilder den j'te observation som det af stedvektoren

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{4.1}$$

bestemte punkt. En sådan afbildning er foretaget i figur 2, hvor det er forudsat, at $E(\mathbf{x}) = \mathbf{0}$, hvilket altid vil kunne opnås ved hensigtsmæssigt valg af nulpunkt.

Som nævnt i appendix svarer den ved komponentanalysen foretagne ortogonale transformation til en drejning af koordinatakserne omkring origo, således at observationerne i det ny koordinatsystem er givet ved ligning (3.3). I det to-dimensionale tilfælde fås:

$$y_1 = a_{11}x_1 + a_{21}x_2$$
$$y_2 = a_{12}x_1 + a_{22}x_2 \qquad (4.2)$$

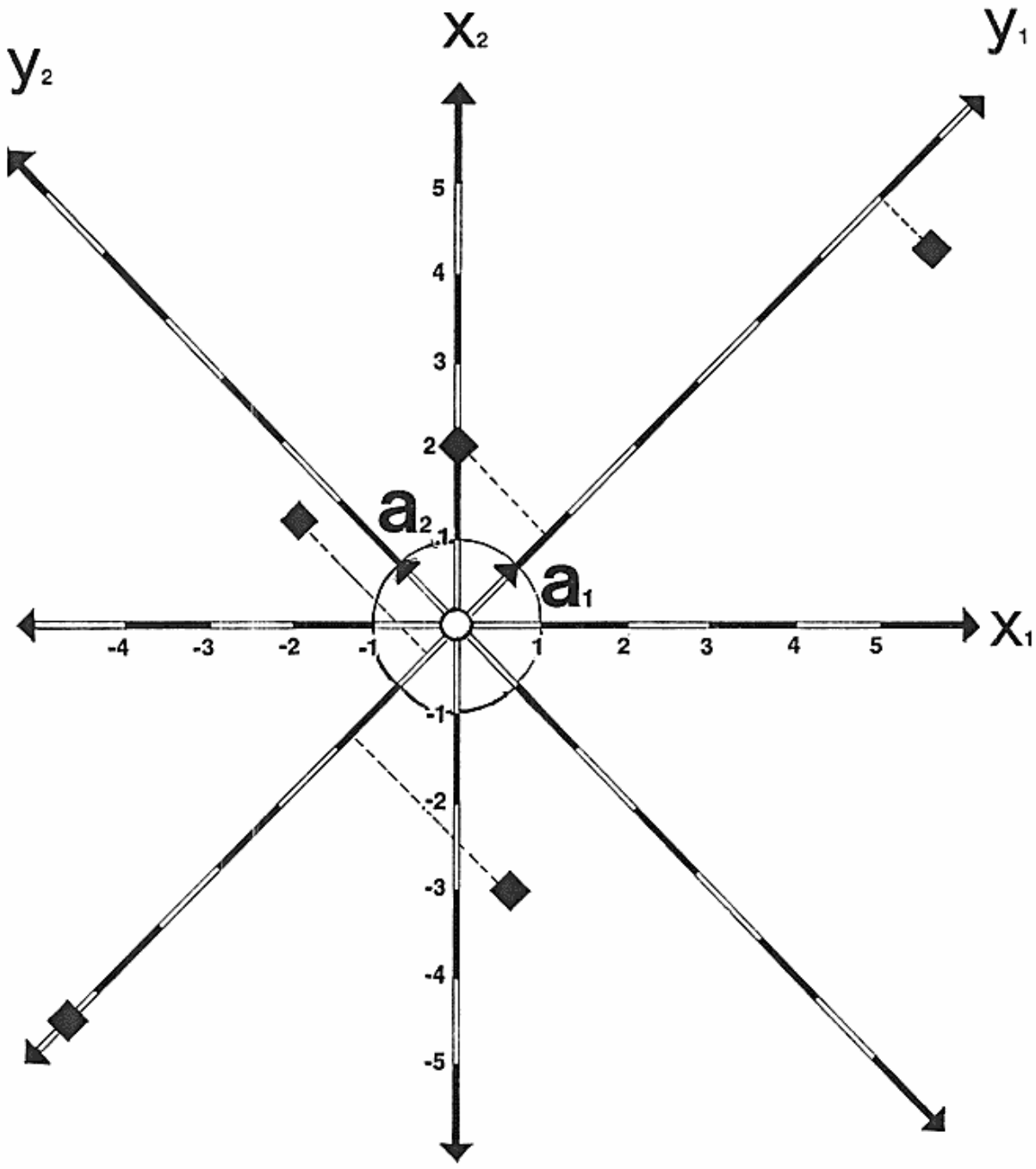
hvor koefficienterne er bestemt ved, at y_1 skal have størst mulig varians, medens y_2 skal have størst mulig varians under bibetingelsen, at y_2 ikke er korreleret med y_1.

Lad os først se, hvad der sker, hvis vi drejer x_1-aksen om origo. Det fremgår, at hvis en sådan drejning foregår mod uret, vil punkternes x_1-koordinater i begyndelsen blive »spredt mere ud«: Variansen vil vokse. Fortsætter drejningen vil x_1-koordinaterne igen blive »trykket sammen«: Deres varians vil blive mindre.

Observationsmaterialets maksimale varians målt på den således drejede x_1-akse opnås åbenbart når aksen går »midt gennem« punktklyngen. Benytter vi nu betegnelsen y_1-aksen for den således drejede x_1-akse, har vi åbenbart opnået, at kunne udtrykke vore to målinger x_1 og x_2 ved en ny variabel y_1, der er den lineære funktion af x_1 og x_2, der har størst varians. y_2-aksen må da stå vinkelret på y_1-aksen i origo, da y_1 og y_2 skal være ukorrelerede. I figur 2 er de nye koordinatakser indtegnet ligesom de karakteristiske vektorer er angivet. Vi ser, at disse éntydigt bestemmer beliggenheden af det nye koordinatsystem.

Det skal til sidst bemærkes, at y_1-aksen ikke må forveksles med den sædvanlige regressionslinie, der bestemmes ved minimering af summen af punkternes lodrette afstand fra linien. Derimod kan det bevises, at summen af punkternes vinkelrette afstand fra y_1-aksen er mindre end til nogen anden ret linie. Denne egenskab, der er en konsekvens af Phytagoras læresætning, betyder, at bestemmelsen af y_1-aksens beliggenhed i anden forbindelse ofte omtales som ortogonal regression.

Denne mere håndfaste geometriske fortolkning af de principale komponenter lader sig uden videre generalisere til flere end to dimensioner – selv om man ved flere end tre dimensioner må tage fantasien til hjælp. At transformere de oprindelige variable, **x**, til nye ukorrelerede variable, **y**, med maksimal varians er altså ensbetydende med at foretage en drejning af koordinatakserne, således at summen af punkternes kvadratafvigelser fra den første nye koordinatakse minimeres. Dernæst bestemmes den anden koordinatakse således, at den er vinkelret på den første i origo og således, at summen af punkternes kvadratafvigelse fra den minimeres o.s.fr. Vi ser, at det ej heller i denne formulering af problemstillingen er nødvendigt at gøre forudsætninger om den statistiske fordeling for **x**.



Figur 2. Grafisk fremstilling af komponentanalyse.

5. Eksempler på anvendelse af faktoranalyse inden for marketingområdet.

Massy grupperer de formål, man kan forfølge med faktoranalysen som følger:

(1) Bestemmelse af antallet af dimensioner, der er nødvendigt til beskrivelse af et elements egenskaber.

(2) Gruppering af de undersøgte elementer i mindre, homogene grupper.

(3) Bestemmelse af hvilke målinger, der skal udvælges til videre analyse.

(4) Skabelse af helt nye variable, der – som funktioner af de oprindelige – kan gøres til genstand for videre analyse.

Man vil bemærke, at i de to første anvendelser er faktoranalysen den endelige analysemetode og altså et mål i sig selv, medens den i de to sidste anvendelser ikke er en konkurrent til andre analysemetoder, men et supplement til disse. Det er min opfattelse, at det er i denne – måske mere ydmyge rolle – at metoden vil komme til at få sin største betydning inden for markedsanalysearbejdet.
Der er i den sidste halve snes år fremkommet en lang række referater af praktiske anvendelser inden for marketingområdet, således at en blot nogenlunde dækkende gennemgang af disse ikke er mulig inden for en tidsskriftsartikels rammer. Jeg vil derfor i stedet knytte nogle kommentarer til de af Massy opstillede fire problemformuleringer og illustrere dem med enkelte henvisninger til offentliggjorte analyserapporter.

Ad (1). Bestemmelse af antallet af dimensioner, der er nødvendigt til beskrivelse af et elements egenskaber.

Dette er den oprindelige anvendelse af faktoranalyseteknikken. Problemet opstod ved psykologers forsøg på ved hjælp af en lang række tests at beskrive et menneskes personlighed. Spørgsmålet var nu: Måler disse mange tests nu virkelig forskellige egenskaber, og hvis ikke, hvor mange dimensioner er da nødvendig til en tilfredsstillende beskrivelse af personligheden? Dette fører imidlertid frem til en gruppering af tests, der måler samme dimension. En sådan gruppering er af stor praktisk interesse, da den giver mulighed for inden for hver gruppe, at vælge den bedste test, d.v.s. den test, der er stærkeste korreleret med den pågældende faktor (= dimension), smlgn. formål (3) i Massy's opstilling. Af større teoretisk interesse er det imidlertid, at man med udgangspunkt i en sådan gruppering eventuelt kan opnå

at identificere (»navngive«) de enkelte faktorer og dermed give disse matematiske abstraktioner en faglig meningsfuld fortolkning.

Fra psykologerne er faktoranalysen kommet til marketingområdet sammen med skaleringsteknikkerne. De første – og fleste – markedsanalyseanvendelser er da også sket i forbindelse med fortolkningen af indstillingsundersøgelser af forskellig art ved hjælp af skaleringsteknik. Et eksempel herpå er *Eastlack's* anvendelse af faktoranalyse af præferancedata ved benyttelse af Osgood's semantiske differentialskala.

Formålet med undersøgelsen var at bestemme hvilke egenskaber en ny kaffeblanding, der skulle sælges som mærkevare, skulle være i besiddelse af for bedst muligt at appellere til en større andel af forbrugere på et lokalt marked.

Som et eksempel på en anden problemstilling, der kan angribes med faktoranalyseværktøjet kan nævnes *Dudek's* analyse af en række af de af amerikanske markedsanalyseinstitutter publiserede index for TV-kigning. Formålet var her at undersøge, i hvilket omfang de forskellige index måler samme dimensioner og hvilke dimensioner, der da er tale om.

Der er et aspekt ved anvendelsen af faktoranalyse som middel til at fortolke måleresultater, som det nok er værd at ofre et par ord på: I denne anvendelse, hvor man »navngiver« de matematiske abstraktioner, der optræder under den neutrale betegnelse faktorer, benytter man sig i realiteten af faktoranalysen som et udgangspunkt for en generering af teorier.

Såfremt man benytter faktoranalysens resultater forudsætningsløst, d.v.s. lader dem tale for sig selv uden at referere til den almindelige viden og teoribygning, der findes inden for det fagområde analysen anvendes på, løber man en meget alvorlig risiko for, at der kommer meningsløse eller direkte vildledende resultater frem.

Det er jo velkendt, at korrelation ikke er ensbetydende med årsagssammenhæng, men blot er et udtryk for samvariation. Utallige er de advarsler, vi har fået mod at drage for vidtgående slutninger på grundlag af den såkaldte »nonsenskorrelation« mellem f. eks. antallet af skudte ræve i Finland og antallet af indgåede ægteskaber i Paris, for nu at nævne et klassisk eksempel.

Ved faktoranalyse er risikoen for at gå vild på denne måde i virkeligheden mange gange større end ved korrelationsanalyse, fordi faktoranalysen er så langt mindre gennemskuelig. Man skal altså tænke sig godt om, før man drister sig til at navngive de forskellige faktorer.

Et eksempel herpå er *Stoetzel's* analyse af franske forbrugeres præferancer for spiritus, hvortil *Vincent* har bragt en alternativ »løsning« baseret på de samme data.

Ad (2). Gruppering af de undersøgte elementer i mindre, homogene grupper.

I den første anvendelse af faktoranalyse var udgangspunktet en gruppering af målinger, således at målingerne inden for hver gruppe var stærkt korrelerede indbyrdes, og derfor forudsattes at måle samme egenskabsdimension(er), kaldet faktorer.
Foretager man fuldstændig samme beregninger, men nu således at man lader målinger og elementer bytte plads overalt i beregningerne, vil man åbenbart opnå, at få opdelt elementerne i mindre grupper, der er homogene med hensyn til de foretagne målinger.
Med et blot nogenlunde rimeligt antal elementer vil kovariansmatricen, **V**, få så enorme dimensioner, at der opstår problemer med lagerkapacitet ved EDB-anvendelse, og det er da heller ikke lykkedes mig at finde offentliggjorte eksempler på anvendelse af denne såkaldte *omvendte faktoranalyse* på realistiske problemer. De steder, den har været anvendt har nærmest haft form af pilot-undersøgelser med det formål at studere metoden og ikke problemet.
En nærliggende anvendelse af omvendt faktoranalyse er gruppering af forbrugerne i grupper, der hver især er mere homogene i deres følsomhed over for de forskellige handlingsparametre end alle forbrugere taget under ét. En sådan viden kan danne grundlag for at differentiere en virksomheds politik over for enkelte segmenter af markedet.
En lignende fremgangsmåde kan anvendes til gruppering af geografiske områder (f. eks. handelsområder) i grupper, der er homogene med hensyn til socio-økonomiske og demografiske egenskaber. En sådan gruppering kan da være et hjælpemiddel ved udvælgelse af test- og kontrolmarkeder til anvendelse i forbindelse med markedseksperimenter.

Ad (3). Bestemmelse af hvilke målinger, der skal udvælges til nærmere analyse.

Som tidligere nævnt, kan den ved faktoranalyse fremkomne gruppering af målinger, der for det enkelte element er stærkt indbyrdes korrelerede, benyttes til inden for hver gruppe at udvælge én variabel til videregående analyse. Blandt de utallige offentliggjorte rapporter om denne anvendelse skal kun to nævnes. *Frank og Boyd* har benyttet metoden til at bestemme hvilke af en lang række socio-økonomiske variable, der mest hensigtsmæssigt kunne indgå som forklarende variable i en regressionsanalyse med det formål at forudsige husholdningers forbrug af private mærkevarer.
Twedt var interesseret i at bestemme hvilke egenskaber ved en fagbladsannonce (f. eks. størrelse, antal ord, anvendelse af farve etc.), der

var bestemmende for hvor mange læsere den fik. Foruden at identificere de underliggende dimensioner, blev faktoranalysen benyttet til blandt de oprindelige 20 variable at udvælge 7 – én for hver af de identificerede faktorer –, der skulle indgå i en regressionsanalyse, som skulle gøre det muligt at forudsige størrelsen af en given annonces læserkreds.

Ad (4). Skabelse af helt nye variable, der – som funktioner af de oprindelige – kan gøres til genstand for videre analyse.

I stedet for at udvælge nogle af de oprindelige variable – x-værdierne – til videre analyse, kan faktoranalysen benyttes til skabelse af nye variable – y-værdierne – der så kan indgå i den videre analyse. Hver y-værdi vil da repræsentere en række indbyrdes korrelerede x-værdier, medens y-værdierne i sig selv vil være indbyrdes ukorrelerede, hvilket gør dem velegnede til at indgå som forklarende variable i en regressionsanalyse eller diskriminantanalyse. *Farley* har benyttet denne teknik i en undersøgelse, hvis formål det var at konstatere årsagerne til, at mærkeloyaliteten varierer fra produktgruppe til produktgruppe inden for fødevareområdet.

6. Matematisk appendix

Principale komponenters metode.

Vi tager udgangspunkt i formuleringen (3.3), der skrevet på matrixform er

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad (A.1)$$

Vi kan betragte en sådan transformation til nye variable som en ændring i det koordinatsystem, hvori vi udtrykker vore målinger. Af sådanne transformationer er der naturligvis uendelig mange, hver bestemt ved udformningen af *transformationsmatricen* **A**. Vi vil nu stille følgende krav til transformationsmatricen **A**:

(1) At transformationen er afstandsbevarende.

(2) At variablerne y_k er ukorrelerede.

(3) At $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \ldots\ldots\ldots \geq \mathrm{Var}(y_p)$.

(4) At $\mathrm{Var}(y_k)$ er størst mulig for samtlige k.

Lad vektorerne $\mathbf{x}\alpha$ og $\mathbf{x}\beta$ repræsentere to punkter i det p-dimensiale rum. Kvadratet på afstanden mellem disse to punkter er da $D(\mathbf{x}\alpha, \mathbf{x}\beta) = (\mathbf{x}\alpha - \mathbf{x}\beta)^T (\mathbf{x}\alpha - \mathbf{x}\beta)$. Ved transformationen $\mathbf{y} = \mathbf{A}\mathbf{x}$ føres de to punkter over i punkterne repræsenteret ved vektorerne $\mathbf{y}\alpha = \mathbf{A}\mathbf{x}\alpha$ og $\mathbf{y}\beta = \mathbf{A}\mathbf{x}\beta$. Kvadratet på afstanden mellem disse to punkter er da

$$\begin{aligned} D(\mathbf{A}\mathbf{x}_\alpha - \mathbf{A}\mathbf{x}_\beta) &= (\mathbf{A}\mathbf{x}_\alpha - \mathbf{A}\mathbf{x}_\beta)^T(\mathbf{A}\mathbf{x}_\alpha - \mathbf{A}\mathbf{x}_\beta) \\ &= (\mathbf{A}(\mathbf{x}_\alpha - \mathbf{x}_\beta))^T(\mathbf{A}\mathbf{x}_\alpha - \mathbf{x}_\beta) \\ &= (\mathbf{x}_\alpha - \mathbf{x}_\beta)^T\mathbf{A}^T\mathbf{A}(\mathbf{x}_\alpha - \mathbf{x}_\beta) \end{aligned}$$

Vi ser altså, at transformationen kun er afstandsbevarende såfremt $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, hvor $\mathbf{I}$ er enhedsmatricen. En matrix, der opfylder denne betingelse kaldes *ortogonal*. Er transformationsmatrixen, $\mathbf{A}$, i (A.1) ortogonal kaldes transformationen ortogonal. En ortogonal transformation svarer til en drejning af koordinatsystemet omkring origo, hvilket er illustreret i afsnit 4.

Da vi kun er interesseret i *variationerne* i $\mathbf{x}$, vil vi i det følgende forudsætte, at $E(\mathbf{x}) = \mathbf{0}$. Kovariansmatricen for $\mathbf{x}$ er da

$$\mathbf{V} = E(\mathbf{x}\mathbf{x}^T)$$

Er $\mathbf{a}$ en søjlevektor i p dimensioner i $\mathbf{A}$, har vi

$$y = \mathbf{a}^T\mathbf{x} \tag{A.2}$$

hvor

$$\mathbf{a}^T\mathbf{a} = 1 \tag{A.3}$$

da $\mathbf{A}$ *er* ortogonal. Vi har nu

$$\mathrm{Var}(y) = E(\mathbf{a}^T\mathbf{x})^2 = E(\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{a}) = \mathbf{a}^T\mathbf{V}\mathbf{a} \tag{A.4}$$

At finde den linearkombination, y, af $\mathbf{x}$, der har maksimal varians, er ensbetydende med at maksimere (A.4) under bibetingelsen (A.3), hvilket ved intruduktion af la Grange-multiplikatoren, λ, er ensbetydende med at maksimere

$$L_1 = \mathbf{a}^T\mathbf{V}\mathbf{a} - \lambda(\mathbf{a}^T\mathbf{a} - 1) \tag{A.5}$$

Vektoren af partielle afledede m.h.t. $\mathbf{a}$ er

$$\frac{\delta L_1}{\delta \mathbf{a}} = 2\mathbf{V}\mathbf{a} - 2\lambda\mathbf{a} \tag{A.6}$$

Vi må kræve, at (A.6) er lig **0**. Med andre ord, der må, for at (A.5) skal have maksimum, findes en skalar, λ, og en vektor **a**, således at

$$\mathbf{Va} = \lambda\mathbf{a} \tag{A.7}$$

λ og **a** kaldes henholdsvis en karakteristisk værdi (eller egenværdi) og dennes tilhørende karakteristiske vektor for matricen **V**. (A.7) kan skrives på formen

$$(\mathbf{V} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0} \tag{A.8}$$

En løsning på (A.8), der samtidig sikrer, at (A.3) er opfyldt, må medføre, at determinanten

$$|\mathbf{V} - \lambda\mathbf{I}| = 0 \tag{A.9}$$

(A.9) er et polynomium i λ af p'te grad og har derfor p rødder – – af hvilke nogle dog kan være identiske. Lad os kalde disse rødder $\lambda_1 \geq \lambda_2 \geq \ldots\ldots\ldots\ldots \geq \lambda p$.

Ved multiplikation med $\mathbf{a}^T$ på begge sider af lighedstegnet i (A.8) fås

$$\begin{aligned} &\mathbf{a}^T(\mathbf{V} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0} \\ &\mathbf{a}^T\mathbf{V} - \mathbf{a}^T\lambda\mathbf{I}\mathbf{a} = \mathbf{0} \\ &\mathbf{a}^T\mathbf{V}\mathbf{a} = \lambda\mathbf{a}^T\mathbf{a} \end{aligned} \tag{A.10}$$

hvoraf ved indsættelse af (A.4) og (A.3) på henholdsvis venstre og højre side af lighedstegnet fås

$$\mathrm{Var}(y) = \lambda \tag{A.11}$$

Da vi ønsker at maksimere Var(y), vælges den største rod i (A.9).

Denne rod har vi tidligere betegnet λ_1. Hvis $\mathbf{a}_1$ betegner en normeret løsning til ligningen

$$(\mathbf{V} - \lambda_1\mathbf{I})\mathbf{a} = \mathbf{0}$$

er

$$y_1 = \underset{1}{\mathbf{a}^T}\mathbf{x}$$

den normerede linearkombination af **x**, der har maksimal varians.

Det kan bevises (Anderson 1958, kap. II), at den normerede linearkombination, y_2, der har maksimal varians under bibetingelsen, at y_2

$$y_2 = \underset{2}{\mathbf{a}^T}\mathbf{x}$$

hvor $\mathbf{a}_2$ er den til den næsthøjeste karakteristiske værdi, λ_2, svarende karakteristiske vektor. Det kan herefter bevises, at den linearkombination, y_3, der har maksimal varians under bibetingelsen, at y_3 hverken er korreleret med y_1 eller y_2, er

$$y_3 = \mathbf{a}_3^T x$$

hvor $\mathbf{a}_3$ er den til den trediestørste karakteristiske værdi, λ_3 svarende karakteristiske vektor.o.s.fr.

Vi kan altså beregne $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots\ldots, \mathbf{a}_k, \ldots\ldots, \mathbf{a}_p$ og dermed bestemme transformationsmatricen, $\mathbf{A}$, (hvori $\mathbf{a}_k$ udgør den k'te søjle) ved følgende fremgangsmåde:

(1) Samtlige rødder i ligningen (A.9) bestemmes.

(2) Rødderne opstilles efter faldende størrelse.

(3) Den første (største) rod, λ_1, indsættes i (A.8), der løses m.h.t. vektoren $\mathbf{a}$. Den fundne vektor er den første søjle, $\mathbf{a}_1$ i $\mathbf{A}$ Dernæst indsættes den anden (næststørste) rod, λ_2, i (A.8), der igen løses m.h.t. $\mathbf{a}$. Den fundne løsning er nu $\mathbf{a}_2$, den anden søjle i $\mathbf{A}$ o.s.fr. Ved løsningen af (A.8) benyttes endvidere, at $\mathbf{a}^T\mathbf{a} = 1$ og at kovariansen mellem y_k og y_h skal være 0 for alle $h \neq k$.

Da $\mathbf{A}$ er ortogonal, medfører $\mathbf{y} = \mathbf{Ax}$ at $\mathbf{x} = \mathbf{A}^T\mathbf{y}$. D.v.s. at vi i én arbejdsgang får bestemt såvel koefficienterne i (3.3) som søjlerne i $\mathbf{A}$, som koefficienterne i (3.2) som rækkerne i $\mathbf{A}$. Med andre ord er $l_{ik} = a_{ki}$.

Det bemærkes endelig, at da λ er en kvadreret størrelse (varians), er $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots\ldots\ldots\ldots \geq \lambda_p \geq 0$

hvor samtlige lighedstegn kun gælder i det helt uinteressante tilfælde, hvor $\mathbf{V} = \mathbf{0}$. Derimod kan der opstå to mindre ekstreme tilfælde, der kan have en vis interesse:

(1) Nogle af λ'erne kan være 0. så fald er kovariansmatricen, $\mathbf{V}$ af lavere rang end p. Dette medfører, at datamatrixen $\mathbf{X} = (x_{ig})$, hvor x_{ig} betegner den i'te måling på det g'te element, er af lavere rang end p. En hver måling x_{ig} kan da udtrykkes i færre end p y-værdier.

(2) Nogle af λ'eren kan være lige store. I så fald findes der uendelig mange ortogonale transformationer, der opfylder kravet om ukorrelerede y'er.

Endelig skal det bemærkes, at det ved anvendelse af faktoranalyse er koutume at standardisere y'erne, således at de alle får variansen 1. Dette gøres ved at foretage transformationen $z_k = (1/\sqrt{\lambda_k})y_k$ for $k = 1, 2, \ldots, p$.

Ligeledes ser man ofte, at x'erne standardisere før beregningerne påbegyndes, således at udgangspunktet ikke bliver en kovariansmatrix, men en korrelationsmatrix d.v.s. en matrix af korrelationskoefficienter.

Referencer

Standardværket er

1. Harry H. Harman: *Modern Factor Analysis.* University of Chicago Press, 1960.

En statistisk orienteret fremstilling findes i

2. T. W. Anderson: *Introduction to Multivariate Statistical Analysis.* Wiley, 1958.

Computerprogrammer kan søges i

3. W. W. Cooley and P. R. Lohnes: *Multivariate Procedures for the Behavioral Sciences.* Wiley, 1962.
4. W. J. Dickson (Ed.): *BMD Biomedical Computer Programs.* University of California Press, 1968.
5. *IBM Scientific Subroutine Package.*

Eksempler på anvendelse af faktoranalyse inden for marketingområdet er talrige, og mange offentliggøres i tidsskrifterne Journal of Marketing Research og Journal of Advertising Research. De i artiklens afsnit 5 refererede kilder er:

6. Frank J. Dudek: *Relations among Television Rating Indices.* JAR Vol. 4, 1964, No. 3, pp. 24–28.
7. J. L. Eastlack: *Consumer Flavor Preference Factors in Food Product Design.* JMR Vol. 1, 1964, No. 1, pp. 38–42.
8. J. V. Farley: *Why Does Brand Loyality Vary Over Products?* JMR Vol. 1, 1964, No. 4, pp. 9–14.
9. R. E. Frank and H. W. Boyd Jr.: *Are Private Brand-Prone Grocery Customers Really Different?* JAR Vol. 5, 1965, No. 4, pp. 27–35.
10. W. F. Massy: *Applying Factor Analysis to a Specific Marketing Problem.* Proceedings of the American Marketing Association, December 1963, pp. 291–307.
11. Jan Stoetzel: *A Factor Analysis of the Liquor Preferences of French Consumers.* JAR Vol. 1, 1961, No. 2, pp. 7–11.
12. D. Twedt: *A Multiple Factor Analysis of Advertising Readership.* Journal of Applied Psychology, Vol. 1, November 1964, pp. 9–14.
13. Norman L. Vincent: *A Note on Stoetzel's Factor Analysis of Liquor Preferences.* JAR Vol. 2, 1962, No. 1, pp. 24–27.

Endelig skal det bemærkes, at det ved anvendelse af faktoranalyse er koutume at standardisere y'erne, således at de alle får variansen 1. Dette gøres ved at foretage transformationen $z_k = (1/\sqrt{\lambda_k})y_k$ for $k = 1, 2, \ldots, p$.

Ligeledes ser man ofte, at x'erne standardisere før beregningerne påbegyndes, således at udgangspunktet ikke bliver en kovariansmatrix, men en korrelationsmatrix d.v.s. en matrix af korrelationskoefficienter.

Referencer

Standardværket er

1. Harry H. Harman: *Modern Factor Analysis.* University of Chicago Press, 1960.

En statistisk orienteret fremstilling findes i

2. T. W. Anderson: *Introduction to Multivariate Statistical Analysis.* Wiley, 1958.

Computerprogrammer kan søges i

3. W. W. Cooley and P. R. Lohnes: *Multivariate Procedures for the Behavioral Sciences.* Wiley, 1962.
4. W. J. Dickson (Ed.): *BMD Biomedical Computer Programs.* University of California Press, 1968.
5. *IBM Scientific Subroutine Package.*

Eksempler på anvendelse af faktoranalyse inden for marketingområdet er talrige, og mange offentliggøres i tidsskrifterne Journal of Marketing Research og Journal of Advertising Research. De i artiklens afsnit 5 refererede kilder er:

6. Frank J. Dudek: *Relations among Television Rating Indices.* JAR Vol. 4, 1964, No. 3, pp. 24–28.
7. J. L. Eastlack: *Consumer Flavor Preference Factors in Food Product Design.* JMR Vol. 1, 1964, No. 1, pp. 38–42.
8. J. V. Farley: *Why Does Brand Loyality Vary Over Products?* JMR Vol. 1, 1964, No. 4, pp. 9–14.
9. R. E. Frank and H. W. Boyd Jr.: *Are Private Brand-Prone Grocery Customers Really Different?* JAR Vol. 5, 1965, No. 4, pp. 27–35.
10. W. F. Massy: *Applying Factor Analysis to a Specific Marketing Problem.* Proceedings of the American Marketing Association, December 1963, pp. 291–307.
11. Jan Stoetzel: *A Factor Analysis of the Liquor Preferences of French Consumers.* JAR Vol. 1, 1961, No. 2, pp. 7–11.
12. D. Twedt: *A Multiple Factor Analysis of Advertising Readership.* Journal of Applied Psychology, Vol. 1, November 1964, pp. 9–14.
13. Norman L. Vincent: *A Note on Stoetzel's Factor Analysis of Liquor Preferences.* JAR Vol. 2, 1962, No. 1, pp. 24–27.