

Korrespondanceanalyse som statistisk redskab på de gymnasiale uddannelser

Hovedopgave i matematik ved EVU. Aalborg Universitet

Af Leif Gammelgård .



Summary

In many scientific disciplines the collection and processing of empirical data plays a central role. How this treatment of data takes place, obviously depends on the type of data and which academic field in question. It is often possible to set up tables of both qualitative and quantitative data. But even with a relatively limited number of variables it can be difficult to obtain a comprehensive overview of the information collected, which again makes it difficult to draw conclusions based on the data.

An important method in getting an overview of large amounts of data is by illustrating them in charts and graphs. Common graphical presentations of data have the limitation that it is only possible to display two different categories (dimensions) in the same diagram in a clear way. Not all types of data can be illustrated in simple graphs and charts. Especially if the data is composed of several dimensions, a graphical representation is difficult. It is therefore interesting to examine which alternatives there are to graphical a presentation and analysis of data in multiple dimensions.

In this report the main focus will be an analysis of correspondence analysis, as a tool to illustrate data. Using statistical tools, it's easy to make a correspondence analysis plot, which directly reduces an otherwise unmanageable data material down to a single chart that is open to interpretation. By illustrating the multi-dimensional data in two dimensions, there is a data reduction, which means the loss of information; a loss that is not immediately visible to the reader of the plot. The interpretation of a correspondence analysis plot thus requires, as for other types of data representations, the specialist knowledge in the area of which the data belongs. But the correspondence plots also require a mathematical understanding of what the plot illustrates and how the plot is obtained.

The purpose of this project is to provide an introduction to correspondence analysis which can form the basis for teaching on a secondary school level. Such courses will give students the requirements needed for the management and interpretation of correspondence analysis plots in the context of different school subjects.

The report will start out with an introduction, followed by a section on graphical presentation of data that aims to describe some of the challenges of presenting data in more than two dimensions and introduce some of the basic concepts and terms that will be used later in the report. The section is targeted at the academic level that will be expected of students at upper secondary schools.

The section on the X^2 -value is based on the definition of the X^2 test parameter, which is a well known parameter in upper secondary school. The X^2 -value is used to define the distances in the correspondence analysis.

The chapter on dimensions of a dataset first discusses the projection of a point onto a line, and continues to explain the projection onto an arbitrary subspace. The chapter contains simple examples with figures illustrating projection in only a few dimensions. Additionally, eigenvectors of symmetric matrices is treated and the principle behind the spectral decomposition is reviewed.

In the section on singular values the principles of symmetric matrices will be extended to arbitrary non-square matrices via the principle of singular value decomposition.

The understanding of the spectral decomposition via singular value decomposition is central to the understanding of the mathematics behind correspondence analysis. The chapter also contains simple numerical examples.

The chapter about generating a correspondence plot shows how a correspondence analysis plot is built. First through examples in one and two dimensions, then in several dimensions in which there are a projection into two dimensions. The calculations are shown step by step in tabular form. In the following chapter an algorithm for calculating the matrix operators, which can be written in the statistical program “R”, will be inspected.

Finally, three examples of how the correspondence analysis can be used in a practical context, is shown.

The report ends with a conclusion that summarizes how correspondence analysis could be used in a secondary context, to illustrate statistical data as well as a starting point for a discussion on the use of statistical analysis in an interdisciplinary context.

If correspondence analysis should be implemented in upper secondary level it would be natural to create an interdisciplinary course in which the non-mathematical sciences could contribute with the academic requirements for the analysis and interpretation of the data.

In this report the focus is on the mathematical aspect. The data is only intended as illustrative examples, and in some cases are designed to clarify specific aspects. When designing a concrete course it would be natural to extend with more examples and exercises than those contained in this report, and to supplement with mathematical knowledge where the explanations given in this report is incomplete or too complex compared to the academic level of upper secondary students.

Students in Danish upper secondary schools are often asked to interpret statistical data or reflect on the interpretation of others. Therefore there is a need among upper secondary students for a general insight in the use of statistical models such as correspondence analysis. Also, there is a need for providing the students with the necessary skills to perform their own statistical analyses and make qualified interpretations based on those analyses.

Forord

Denne rapport er udarbejdet som en hovedopgave i matematik i forbindelse med uddannelsen Master i matematik under EVU på Aalborg Universitet. Rapporten er skrevet i perioden marts 2014 til december 2014.

I rapporten er matricer angivet med store bogstaver i fed, f.eks.: **D** . Elementerne i matricen angives med samme bogstav i lille skrift, med indeks for række og søjleangivelse, f.eks.: d_{ij}

Vektorer er angivet med lille bogstav i fed, f.eks.: **v** . De enkelte elementer angives med ikke-fed skrift, med indeks, f.eks.: v_i

Der benyttes det græske bogstav χ "Chi" i forbindelse med χ^2 "Chi-i-anden" parameteren. Ikke at forveksle med det latinske X .

Vejleder på projektet er Steen Anderson

Indholdsfortegnelse

Summary.....	3
Forord	5
Indholdsfortegnelse.....	7
Indledning.....	9
Problemanalyse	10
Problemformulering og problemafgrænsning	11
Rapportens opbygning	12
Grafisk præsentation af data.....	14
Relativ fordeling	15
Række- og søjleprofiler.....	16
Vægtning af fordelingerne	19
Data i matriceform	19
X^2-værdi	20
Test for uafhængighed	20
X^2 -værdien i korrespondanceanalyse	21
X^2 -afstanden	22
Et tal-eksempel.....	23
Søjleprofiler	25
Maksimal inerti.....	28
Dimensioner i et datasæt	30
Projektion på et underum.....	30
Projektion fra to dimensioner til én dimension	30
Projektion fra n dimensioner ned på en linje.....	31
Projektion fra n dimensioner ned på 2 dimensioner.....	32
Egenverdier og egenvektorer for en symmetrisk matrice	34
Ortogonale egenvektorer	35
Et tal-eksempel.....	36
Spektral dekomposition.....	37
Et tal-eksempel.....	37
Definition af de singulære vektorer	39
Singular Value Decomposition	41

Bestemmelse af SVD for en vilkårlig matrice	42
Et tal-eksempel	42
Brug af SVD til reduktion af dimensioner	45
Generering af korrespondanceanalyseplot	47
Illustration af data i to og tre dimensioner	47
Eksempel med data i to dimensioner	47
Bestemmelse af kategorier ud fra maksimal inerti.....	49
Søjleprofiler.....	50
Eksempel med data i 3 dimensioner.....	52
Algoritme for KA-plot.....	56
Definitioner af matricer	56
Procedure for beregning af koordinater i et KA-plot.....	57
Koordinater for rækkeprofilerne	58
Koordinater for søjlekategorier	59
Søjleplot	60
Et tal-eksempel	61
Eksempel med søjleprofiler	64
Symmetrisk plot	65
Praktiske eksempler på anvendelse af korrespondanceanalyseplot.....	67
Eksempel 1 - Beregning af KA plot over stemmefordeling ved folketingsvalg	67
Supplerende punkter	73
Eksempel 2 - Analyse af udvalgte forfatteres værker	75
Eksempel 3 - Individuel spørgeskemaundersøgelse	76
Konklusion.....	81
Forslag til konkret anvendelse af korrespondanceanalyse i en gymnasial sammenhæng....	81
Litteraturliste	83

Indledning

Fælles for mange videnskabelige discipliner er, at indsamling og behandling af empiri spiller en central rolle. Ofte fremkommer denne empiri i form af datamateriale i et talformat, der efterfølgende skal behandles og analyseres. Hvordan denne behandling af data foregår, afhænger selvsagt af hvilken type data og hvilket fagområde der er tale om. Nogle datatyper fremkommer direkte i et kvantificerbart format, som eksempelvis højde, vægt, indkomst. Andre typer data er i kvalitativt format, hvor det ikke umiddelbart er muligt at sige hvad der er størst eller mindst. Det kunne f.eks. være indenfor socialvidenskab hvor man har brug for at skelne mellem køn, uddannelsestyper, etnisk tilhørsforhold eller politisk ståsted. Selv om man relativt let kan kategorisere dataene, kan det være vanskeligt at opstille en rangordnet skala. En politisk højre-/ venstreskala gør det nok muligt at rangordne de forskellige politiske partier, men det vil næppe give mening at påstå af en SF'er og en DF'er tilsammen er ækvivalent med to radikale. Der er med andre ord ikke en meningsfuld afstand mellem disse variable, hvorfor de ikke kan behandles som intervallskalerede, men kun rangordnes og behandles som ordinalskalerede variable.

Det er ofte muligt at opstille tabeller over såvel, kvalitative som kvantitative data. Men selv ved et relativt begrænset antal variable kan det være vanskeligt at få et fyldestgørende overblik over de indsamlede informationer, hvilket igen gør det vanskeligt at drage konklusioner baseret på dataene. En vigtig metode til at få et overblik over større datamængder er ved at illustrere dem i diagrammer og grafer. Sædvanlige grafiske præsentationer af data har dog den begrænsning, at det kun er muligt at vise to forskellige kategorier (dimensioner) i samme diagram på en overskuelig måde.

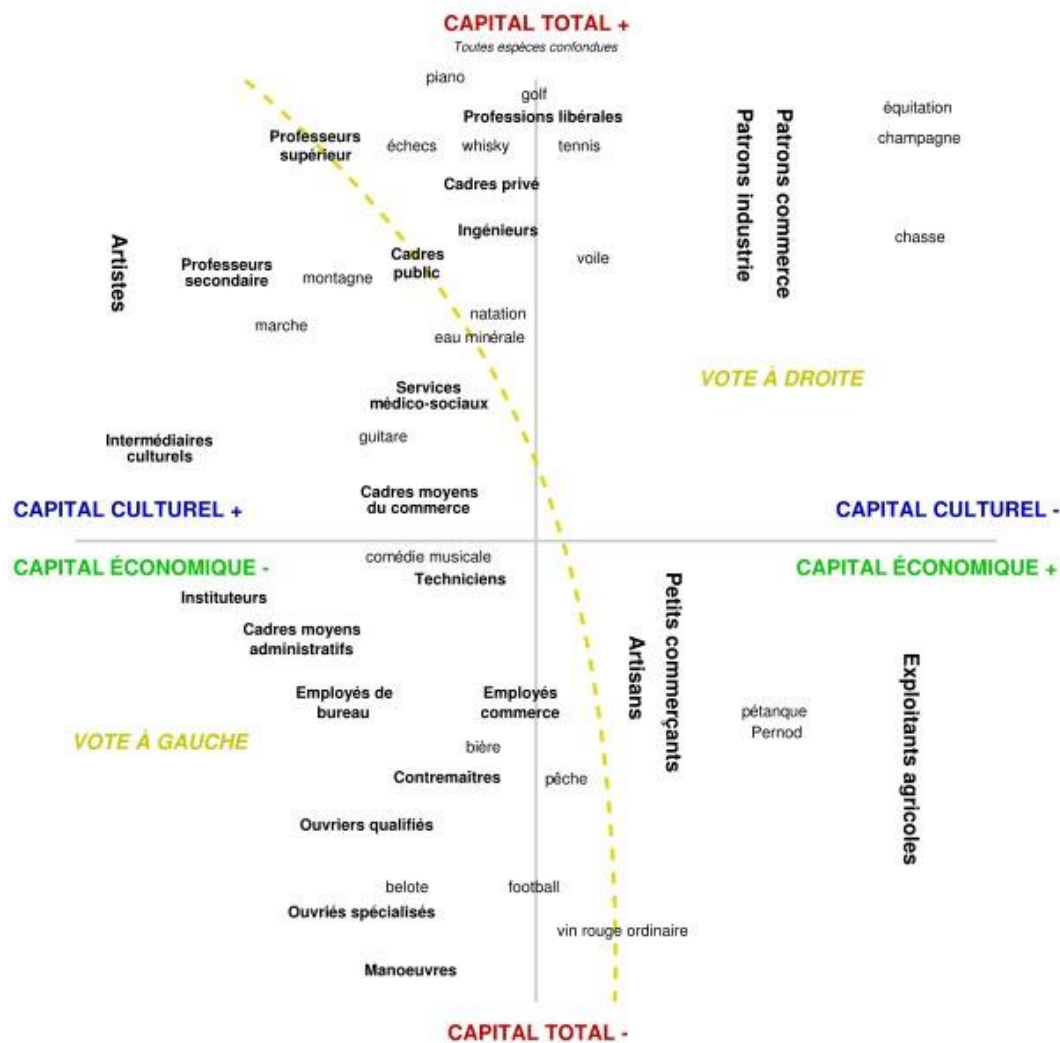
Ikke alle typer data lader sig illustrere i simple grafer og diagrammer. Specielt hvis dataene er opdelt i flere dimensioner, er en grafisk præsentation vanskelig. Hvis man eksempelvis gerne vil illustrere indkomst som funktion af alder, køn og uddannelse, er den traditionelle løsning at opdele tabeller og grafer på de forskellige uafhængige variable og f.eks. lave tre grafer for hhv. køn/løn, alder/løn og uddannelse/løn. Alternativt konstrueres ofte tre-vejstabeller, hvor den statistiske sammenhæng mellem to variable undersøges, og kontrolleres for en tredje variabel (eksempelvis sammenhæng mellem køn og indkomst, kontrolleret for uddannelsesniveau). Problemet med disse fremstillinger er, at de hurtigt bliver uoverskuelige og kun kan fremstille sammenhænge mellem få variable ad gangen. Det kan derfor være interessant at undersøge hvilke alternative muligheder der er for at lave grafisk præsentation og analyse af data i flere dimensioner.

Som baggrund for dette projektarbejde tages der udgangspunkt i følgende initierende problem:

Hvilke redskaber kan anvendes til at tolke data i flere dimensioner, som f.eks. fremkommer af spørgeskemaundersøgelser, på en hensigtsmæssig måde? Hvordan sikres det at redskaberne bliver brugt på en hensigtsmæssig måde, også for personer med begrænset indsigt i matematikken bag?

Problemanalyse

Indenfor socialvidenskaberne arbejder man ofte med data af kvalitativ karakter, og har bl.a. udviklet metoder til at illustrere disse data grafisk via korrespondanceanalyse. Specielt indenfor sociologien hvor bl.a. Pierre Bourdieu har benyttet korrespondanceanalyse til at illustrere og dokumentere sammenhænge mellem begreber som f.eks. kulturel kapital og økonomisk kapital. Den grafiske illustration af kapitalbegreberne har været et ikon for korrespondanceanalysen, se Figur 1, der illustrerer positionen af forskellige vaner og dispositioner i et rum, der angiver mængden af økonomisk og kulturel kapital.



Figur 1 Det sociale rum. Forenklet udgave fra Bourdieus "Af praktiske grunde"¹ En række økonomiske og sociale faktorer er plottet ind i et diagram der er udspændt af to akser, angivet ved hhv. økonomisk og social kapital.

Korrespondanceanalyse er en udbygning af den geometriske dataanalyse, der gør det mulig at analysere data i flere dimensioner. Den geometriske dataanalyse er bl.a. udviklet af Benzécri i midten af 60'erne, og blev en udbredt statistisk metode i Frankrig, specielt indenfor sociologien.

¹ Kilde: Bourdieus, simplifié extrait de *Raisons pratiques*, Seuil, coll. Points, 1996, p. 21

Geometrisk dataanalyse i sig selv indeholder ikke en grafisk analyse. Korrespondanceanalyse er en videreudvikling udført af bl.a. Brigitte Le Roux og Michael Greenacre.²

Først i 1980'erne fik metoden international udbredelse, da der kom tilgængelige værker om emnet på engelsk. Derudover var det matematiske grundlag for korrespondanceanalysen krævende, uden brug af den computerregnekraft, man senere har fået til rådighed. Matematisk bygger korrespondanceanalyse på abstrakt lineær algebra, men i praksis anvendes sædvanligvis matricformalismen indenfor lineær algebra. Til dels fordi matricformalismen er umiddelbart lettere at implementere i computeralgoritmer. Indenfor forskerverdenen er der dog ikke helt konsensus om, hvad den rette matematiske tilgang til korrespondanceanalysen bør være.³

I dag er det relativt simpelt at udføre selve beregningerne i korrespondanceanalysen ved hjælp af standard statistikprogrammer. Samtidig er brugen af korrespondanceanalyse til at tolke data ved at blive udbredt indenfor en lang række fagdiscipliner. Det er ikke usædvanligt at man i en præsentation af talmateriale ud over de gængse søjlediagrammer og grafer også finder korrespondanceanalyseplots til illustration af den overordnede sammenhæng.

Ved hjælp af statistikværktøjer er det enkelt at lave et korrespondanceanalyseplot, der umiddelbart reducerer et ellers uoverskueligt datamateriale ned til et enkelt diagram, der er åbent for fortolkning. Det er naturligvis forjættende at bruge dette værktøj. Og det kan uden tvivl være en effektiv metode til at få et samlet overblik eller se sammenhænge som ville være skjult i traditionelle grafer og diagrammer. Men, naturligvis har det en pris når man laver et enkelt diagram ud af data der oprindeligt indeholder mange dimensioner og variable. Ved at illustrere flerdimensionelle data i to dimensioner, sker der en datareduktion, hvilket er ensbetydende med tab af information; et tab som ikke er umiddelbart synligt for læseren af plottet. Dette er ikke i sig selv usædvanligt, men er naturligt ved en grafisk repræsentation af data. For at anvende korrespondanceanalyse korrekt skal man være opmærksom på denne reduktion af data, ligesom man naturligvis skal have kendskab til hvordan selve plottene kan og ikke kan tolkes. Tolkningen af et korrespondanceanalyseplot kræver således, ligesom for andre typer datafremstillinger, en faglig indsigt i det område dataene falder ind under, men samtidig er det med korrespondanceanalyseplots nødvendigt med en matematisk forståelse for, hvordan plottet er fremkommet. Ligesom begreber som normalfordeling, middelværdi og spredning efterhånden er udbredt som almen viden, kræver en korrekt tolkning af en normalfordelt parameter også et vist kendskab til matematikken bag. Tilsvarende vil en vis matematisk basisviden indenfor korrespondanceanalyse være en forudsætning for at kunne tolke disse plots korrekt.

Problemformulering og problemafgrænsning

Eftersom selve korrespondanceanalyseplottene vinder stigende udbredelse som illustration af statistiske data, er der behov for at også de matematiske forudsætninger for at anvende og tolke disse plots bliver udbredt. Formålet med dette projekt er at give en indføring i korrespondanceanalyse, der kan danne grundlag for et undervisningsforløb på gymnasieniveau. Et sådant undervisningsforløb vil kunne give eleverne forudsætningerne for at kunne anvende og tolke korrespondanceanalyseplots i en faglig sammenhæng. Korrespondanceanalyse er af natur

² Kilde: [1]La Roux, p 12

³ Kilde: [1]La Roux, s. 9

en tværfaglig disciplin. Den forudsætter en faglig viden indenfor det fagområde hvorfra datamaterialet stammer. Samtidig kræves også en matematisk viden om hvordan en korrespondanceanalyse udføres, hvad den illustrerer og hvordan den kan tolkes. Det vil derfor være naturligt at lave et tværfagligt forløb over korrespondanceanalyse, hvor de ikke-matematiske fag ville kunne bidrage med de faglige forudsætninger for en analyse og tolkning af dataene. Dette projekt vil fokusere på det matematiske aspekt. Datamaterialet er blot tænkt som illustrative eksempler, og er i visse tilfælde konstruerede for at tydeliggøre særlige aspekter. Ved udformning af et konkret undervisningsforløb ville det være naturligt at udbygge med yderligere eksempler og øvelser, end de der er indeholdt i denne rapport, samt at supplere med matematisk viden, der hvor forklaringerne i denne rapport ikke er fyldestgørende eller er for komplekse set i forhold til elevernes forudsætninger.

En stor del af matematikken der vil blive benyttet i rapporten vil bygge på en matrice-formalisme, da den er umiddelbart anvendelig og er relativt velkendt for elever på det gymnasiale niveau.

Rapportens opbygning

Ud over denne indledning vil rapporten indeholde et afsnit om grafisk præsentation af data, der har til formål at beskrive nogle af udfordringerne ved præsentation af data i mere end to dimensioner, samt indføre nogle af de grundlæggende ord og begreber, der vil blive anvendt senere i rapporten. Afsnittet er målrettet det niveau og de forudsætninger der vil være forventet blandt elever ved de gymnasiale uddannelser.

I afsnittet om X^2 -værdi (Chi-i-anden) tages der udgangspunkt i definitionen af X^2 testparameteren, der er en velkendt parameter i gymnasiet. X^2 -værdien benyttes til at definere afstandene i korrespondanceanalysen.

Afsnittet "Dimensioner i et datasæt" behandler først projektionen af et punkt ind på en ret linje, og fortsætter med projektion ind på et vilkårligt under rum. Afsnittet indeholder simple eksempler med tal, der illustrerer projektion i få dimensioner. Derudover vil egenvektorer for symmetriske matricer blive behandlet og princippet bag spektral dekomposition vil blive gennemgået.

I afsnittet om singulære værdier udvides principperne fra symmetriske matricer til vilkårlige, ikke kvadratiske matricer via princippet for singular value dekomposition. Forståelsen af spektral dekomposition via singular value dekomposition er central i forståelsen af matematikken bag korrespondanceanalyse. Kapitlet indeholder også simple taleksempler.

Afsnittet "Generering af KA-plot" viser hvorledes et korrespondanceanalyseplot er bygget op. Først via eksempler i en og to dimensioner, dernæst i flere dimensioner, hvor der sker en projektion ind på to dimensioner. Beregningerne er vist trin for trin i tabelform. I det efterfølgende afsnit bliver gennemgået en algoritme til beregningen af matriceoperationerne, der kan skrives i statistikprogrammet "R".

Til sidst vises tre forskellige eksempler på hvordan korrespondanceanalysen kan anvendes i en praktisk sammenhæng.

Rapporten afsluttes med en konklusion, der sammenfatter hvordan korrespondanceanalyse kunne anvendes i en gymnasial sammenhæng, til illustration af statistiske data såvel som udgangspunkt for en diskussion om anvendelse af statistiske analyser i en tværfaglig sammenhæng.

Grafisk præsentation af data

Når informationer i form af talmateriale skal formidles, er det vigtigt, at man gør sig nogle overvejelser over, hvordan man sikrer, at det væsentligste indhold i dataene kommer til at fremstå i en overskuelig form for modtageren. En naturlig måde at præsentere data kunne være i form af tabeller. Tabeller giver et overblik over data, samtidig med at alle informationerne bliver bevaret. Men, hvis datamaterialet er for stort og variablene for mange, bliver det uoverskueligt at uddrage den information, der er interessant.

I Tabel 1 nedenfor er angivet nogle data fra folketingsvalget i 2011. Af hensyn til overskueligheden er der kun vist 3 partier og 3 geografiske områder. Hver enkelt række angiver hvor mange stemmer hvert af de tre partier fik i det givne område. Det er nok muligt at uddrage, hvordan stemmefordelingen er i hvert område, men det bliver straks vanskeligere at sammenligne to forskellige områder, da der ikke er det samme antal stemmer i hvert område. Hvis tabellen havde indeholdt alle partier og valgkredse ville den blive ganske omfattende og uoverskuelig.

	Socialdemokraterne (S)	Venstre (V)	Enhedslisten (Ø)	I alt
Nørrebro	7249	3748	12382	23379
Rudersdal	6508	17895	2161	26564
Aalborg Ø	15895	9924	3220	29039
I alt	29652	31567	17763	78982

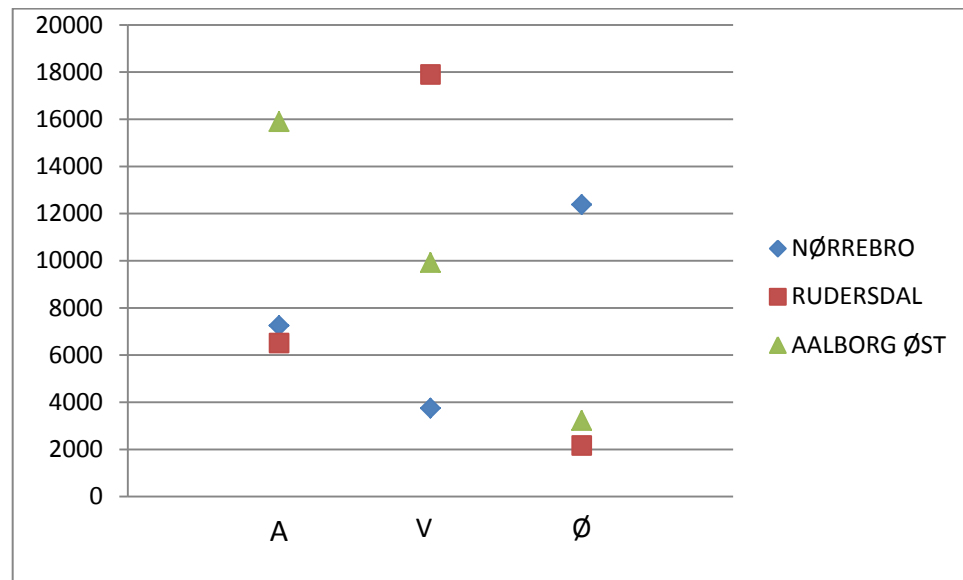
Tabel 1 Udpluk af valgresultat fra folketingsvalg 2011. Kilde Danmarks Statistik

I dette tilfælde er rækkefølgen på søjlerne uden betydning. De kunne være arrangeret alfabetisk, efter en samfundsfaglig højre-/venstreskala, eller som i dette tilfælde efter partibogstav. Tilsvarende gælder for rækkefølgen af områderne. Både række- og søjleparametrene er såkaldte kategoriske variable. Kendetegnen for kategoriske variable er, at afstanden mellem dem ikke har nogen mening. Ved kategoriske variable skelnes der mellem nominalskalerede og ordinalskalerede data. Hvis det er muligt at rangordne skalaen, er der tale om ordinalskalerede variable. I dette tilfælde ville den politiske højre-/venstreskala kunne bruges som argument for en rangordning af de forskellige partier. Men selv med en politisk højre-/venstreskala, ville man ikke kunne afgøre, om afstanden mellem Enhedslisten og Venstre er dobbelt så stor som afstanden mellem Socialdemokraterne og Venstre. En rangordning af de geografiske områder vil ikke give nogen umiddelbar mening, her er der tale om nominalskalerede data.

Hvis rækkerne ikke var arrangeret efter valgkredse men efter indkomstgrundlag ville rækkefølgen få betydning. Det ville da være naturligt at arrangere dataene efter stigende eller faldende indkomstgrundlag og man ville kunne sætte et tal på afstanden mellem rækkerne. Parametre af denne type kaldes kontinuerte variable.

Hvis dataene i tabellen skulle præsenteres grafisk, ville det ikke være muligt blot at tegne et enkelt søjlediagram. Da antallet af stemmer er afhængigt af to forskellige variable, kræver en korrekt fremstilling i princippet tre dimensioner. En mulig fremstilling af dataene kunne være i et punktdiagram, som vist på Figur 2. Her giver den lodrette afstand mellem punkterne umiddelbar mening som forskel i stemmetal, mens den vandrette afstand er uden mening.

Fordelen ved en grafisk præsentation af dataene er, at det nu er lettere at få et overblik over fordelingen af stemmerne, men med den konsekvens, at noget af præcisionen af dataene går tabt. Det er ikke muligt at genskabe de præcise stemmetal ud fra grafen alene (ved almindeligt øjemål). Men, bortset fra når det skal afgøres hvem der er valgt ind, er de præcise tal ofte mindre væsentlige end at få formidlet et let tilgængeligt overblik.



Figur 2 Diagram over stemmefordeling på tre partier i tre områder.

Selv om plottet giver et visuelt overblik over stemmetallene fra tabellen, er det måske ikke velegnet til at illustrere de interessante informationer. Man kunne for eksempel være interesseret i at få afklaret, hvordan den relative stemmefordeling er i de enkelte områder.

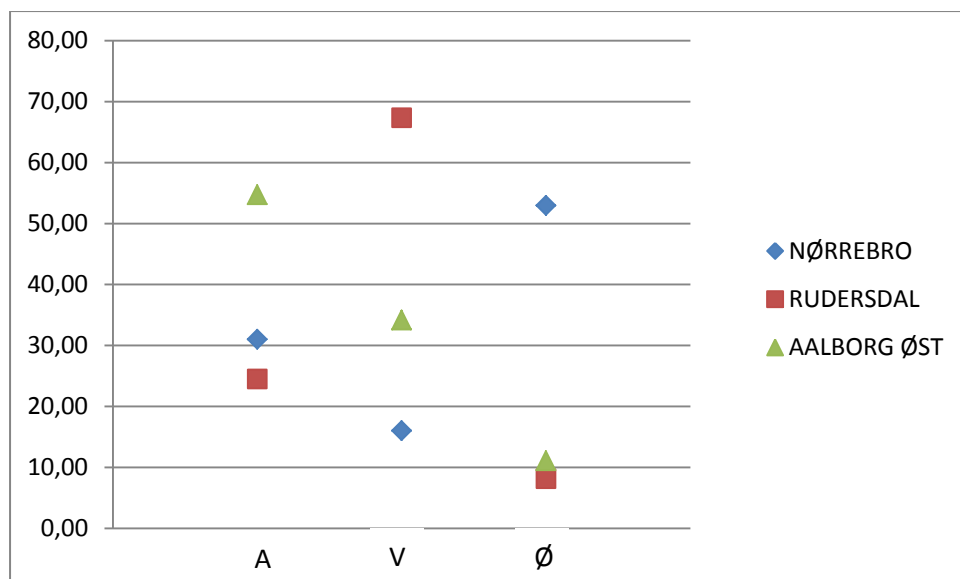
Relativ fordeling

For at gøre det muligt at sammenligne partiernes størrelse på tværs af valgkredse er det nødvendigt at tage højde for valgkredsens størrelser. Dette gøres ved at beregne den relative stemmefordeling for hver valgkreds for sig, som vist i Tabel 2.

	Socialdemokratiet	Venstre	Enhedslisten	I alt
Nørrebro	31,0 %	16,0 %	53,0 %	100 %
Rudersdal	24,5 %	67,4 %	8,1 %	100 %
Aalborg Ø	54,7 %	34,2 %	11,1 %	100 %
Gennemsnit	37,5 %	40,0 %	22,5 %	100 %

Tabel 2 Den relative stemmefordeling i forhold til antallet af stemmer i hvert område.

Disse data kan igen indtegnes i et punktdiagram som vist nedenfor, hvor summen af stemmer i hvert område giver 100 %.



Figur 3 Diagram over relativ stemmefordeling på tre partier i tre områder

Nu er det muligt, at sammenligne hvor stor en andel af stemmerne, partierne har fået i de respektive områder. Det bør bemærkes, at stemmeandelene kun er andele i forhold til de tre partier der er medtaget, og dermed ikke tager højder for stemmerne på de øvrige partier.

Række- og søjleprofiler

Hver række i Tabel 2 udgør en rækkeprofil over de relative fordelinger. Den nederste række indeholder de marginale fordelinger af det samlede stemmetal i de tre områder. F.eks. har Socialdemokraterne fået 37,5 % af de i alt 78982 stemmer. Hvis stemmefordelingerne var uafhængig af området, ville man kunne forvente at socialdemokraterne ville få 37,5 % af stemmerne i alle områderne.

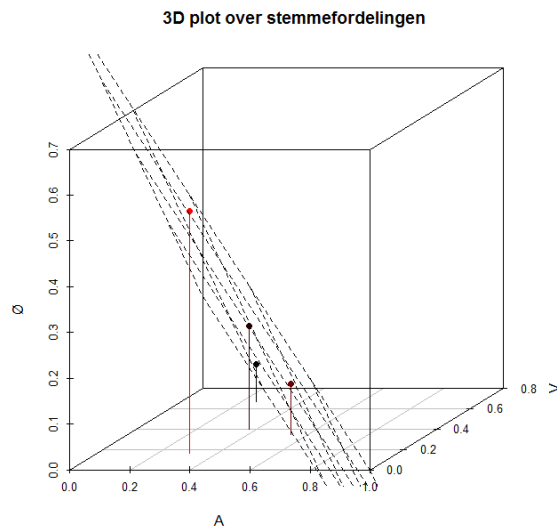
Hvis man i stedet beregnede den relative andel af stemmerne i forhold til det samlede antal stemmer i hvert område, ville man i sidste kolonne få den marginale fordeling af stemmerne på de tre områder. Det vil sige at 29,6 % af alle stemmerne kommer fra Nørrebro. De øvrige søjler viser hvordan de enkelte partiers stemmer er fordelt på de tre områder. F.eks. vil 24,4 % af de stemmer Socialdemokraterne har modtaget i de tre områder komme fra Nørrebro.

	Socialdemokratiet	Venstre	Enhedslisten	Gennemsnit
Nørrebro	24,4 %	11,9 %	69,7 %	29,6 %
Rudersdal	21,9 %	56,7 %	12,2 %	33,6 %
Aalborg Ø	53,6 %	31,4 %	18,1 %	36,8 %
I alt	100 %	100 %	100 %	100 %

Tabel 3 Søjleprofiler. Den relative stemmefordeling i forhold til antallet af stemmer på hvert parti.

Hvor dataene i Tabel 2 angiver rækkeprofiler, angiver dataene i Tabel 3 søjleprofiler. Hvilken type profil man ønsker at anvende afhænger af hvad dataene skal illustrerer. Matematisk er der ingen principiel forskel på om man arrangerer dataene efter rækkeprofiler eller søjleprofiler, men man bør gøre sig klart hvilket valg man træffer, da selve tolkningen naturligvis vil være forskellig.

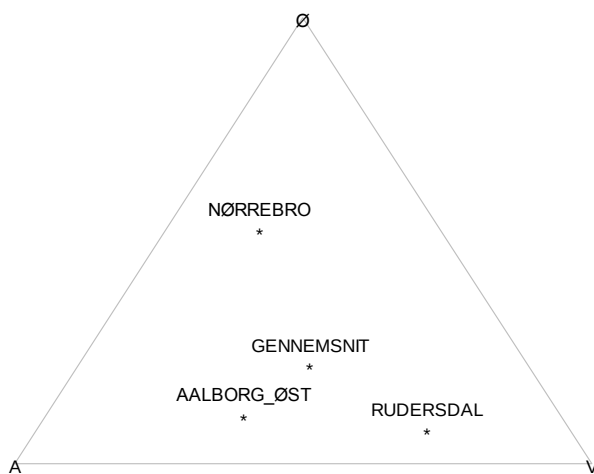
Ved at ændre dataene fra absolutte til relative værdier sker der et tab af information. Det samlede stemmetal kan ikke længere udledes ud fra tabellen. En anden effekt er at værdierne i de enkelte profiler ikke længere er uafhængige af hinanden. F.eks. kan de relative fordelinger i ét område udledes ud fra kendskabet til de øvrige, da summen skal give 100 %. Dette kan illustreres ved at lave et tredimensionelt plot af dataene i Tabel 2.



Figur 4 3D plot over relativ stemmefordeling.

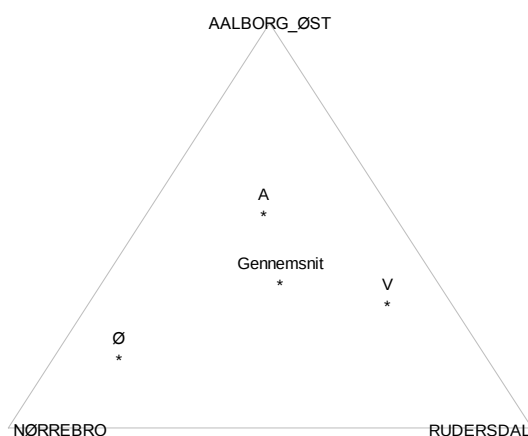
De tre akser udgør de relative stemmetal på de tre partier, mens punkterne er de tre områder. Da summen af de tre koordinater for punkterne skal give 100 % vil alle punkterne ligge på det samme plan i det tredimensionelle rum. Nemlig det plan der er udspændt mellem punkterne for 100 % på de tre akser. I stedet for at tegne punkterne ind i et pseudo-3D plot er det muligt at indtegne dem i et såkaldt triangulært plot, uden tab af information. Dette er en konsekvens af, at der reelt kun er to uafhængige variable. Se Figur 5.

Man kan betragte illustrationen af stemmefordelingerne i Figur 5 som en projektion af den tredimensionelle fordeling, der er vist i Figur 4, ind på det todimensionelle under rum, der giver den største spredning af dataene. I dette tilfælde er projektionen trivielt, da dataene i forvejen ligger i et todimensionelt under rum, men selve princippet er fundamentalt for korrespondanceanalysen.



Figur 5 Fordelingen af områder i forhold til stemmefordeling på tre partier, samt placeringen af gennemsnittet.

Gennemsnittet er fordelingen af samtlige stemmer fra alle tre områder. Positionen af gennemsnittet er ikke det geometriske centrum for de tre partier, da de tre partier bidrager med forskellig vægt til gennemsnittet. Hvis man i stedet plottede søjleprofilerne ville man få de tre partier fordelt i forhold til stemmefordelingen i de tre områder. Se Figur 6.



Figur 6 Fordelingen af partier i forhold til stemmefordeling i de tre områder.

Vægtning af fordelingerne

Ud over de enkelte rækkeprofiler i Figur 5 er rækkeprofilen for gennemsnittet også indtegnet. Rækkeprofilen for gennemsnittet er et vægtet gennemsnit af de enkelte rækkeprofiler i tabellen. De rækkeprofiler, der repræsenterer et stort stemmetal, bidrager med højere vægt til gennemsnittet end de rækkeprofiler med lavt stemmetal. I korrespondanceanalysen er det sædvanen at betegne vægten som rækkens masse. Masserne af de enkelte rækker er givet i den gennemsnitlige relative søjleprofil, mens søjlernes masse er givet ved den gennemsnitlige relative rækkeprofil.

Data i matriceform

Ved behandling af data i tabelform er det oplagt at indsætte dataene i en matrice, for at lette den matematiske håndtering. Rækkerne og søjlerne i tabellen over de rå data kan umiddelbart overføres til rækker og søjler i en matrice. Dataene fra Tabel 1 vil i matriceform blive:

$$D = \begin{bmatrix} 7249 & 3748 & 12382 \\ 6508 & 17895 & 2161 \\ 15895 & 9924 & 3220 \end{bmatrix}$$

Hvor de enkelte elementer i matricen angives ved indeks for række- og søjlenummer, f. eks. $d_{21} = 6508$. Summation af alle elementer i en række eller søjle kan angives ved et plus som indeks, f.eks. $d_{+2} = 26564$ angiver summen af tallene i række 2, og $d_{++} = 78982$ er det samlede antal stemmer på det tre partier i de tre områder.

En matrice B over rækkeprofilerne kan beregnes som:

$$B = \begin{bmatrix} \frac{d_{11}}{d_{1+}} & \frac{d_{12}}{d_{1+}} & \frac{d_{13}}{d_{1+}} \\ \frac{d_{21}}{d_{2+}} & \frac{d_{22}}{d_{2+}} & \frac{d_{23}}{d_{2+}} \\ \frac{d_{31}}{d_{3+}} & \frac{d_{32}}{d_{3+}} & \frac{d_{33}}{d_{3+}} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

Tilsvarende kan elementerne i en matrice A over søjleprofilerne beregnes som $a_{ij} = \frac{d_{ij}}{d_{+j}}$

Massen af rækkerne er givet ved $r_i = \frac{d_{i+}}{d_{++}}$, mens massen af søjlerne er givet ved $c_j = \frac{d_{+j}}{d_{++}}$.

Massen af rækkerne er samtidig den gennemsnitlige søjleprofil og massen af søjlerne er den gennemsnitlige rækkeprofil.

Hvis alle rækkeprofilerne er identiske med de gennemsnitlige profiler ville det betyde at de enkelte partier ville have den samme stemmefordeling, uafhængigt af området. Der ville dermed være tale om uafhængighed mellem række- og søjlevariable. Jo større spredning der er mellem de enkelte profiler desto større afhængighed er der mellem variablene i tabellens rækker og søjler. Hvis søjle- og rækkeparametrene er uafhængige af hinanden, ville der kun optræde variation på grund af den statistiske spredning, og punkterne ville samle sig omkring punkterne for den gennemsnitlige profil. Specielt ved store datasæt vil den statistiske spredning være relativt lille.

X²-værdi

De gennemsnitlige profiler angiver de marginale relative fordelinger af variablene. I Tabel 2 ses det, at 37,5 % af samtlige stemmer i de tre områder gik til Socialdemokraterne. Hvis der er uafhængighed mellem stemmefordelingen og område, ville Socialdemokraterne kunne forvente at få 37,5 % af stemmerne i alle områderne. Tilsvarende ses det i Tabel 3, at 29,6 % af alle stemmer er afgivet på Nørrebro. Hvis man igen antog uafhængighed mellem parti og område ville det være forventet, at 29,6 % af alle partiernes stemmer ville komme fra Nørrebro. En tabel over stemmefordelingerne ved uafhængighed ville kunne beregnes som det samlede stemmetal gange produktet af de marginale fordelinger i række- og søjleprofilerne. For Enhedslisten på Nørrebro ville det være $78982 \cdot 29,6 \% \cdot 37,5 \% = 8777$ stemmer.

	Socialdemokratiet	Venstre	Enhedslisten	I alt
Nørrebro	8777	9344	5258	23379
Rudersdal	9973	10617	5974	26564
Aalborg Ø	10902	11606	6531	29039
I alt	29652	31567	17763	78982

Tabel 4 Forventet stemmefordeling ved uafhængighed mellem område og parti.

Det bemærkes at de forventede stemmefordelinger i Tabel 4 har de samme marginale fordelinger som Tabel 1, og dermed de samme gennemsnitlige række- og søjlefordelinger. Men også de enkelte række- og søjleprofiler er identiske med de gennemsnitlige profiler.

Test for uafhængighed

Tabel 4 indeholder de forventede stemmetal ved uafhængighed, mens Tabel 1 indeholder de faktiske stemmetal. Et oplagt spørgsmål vil være, om det er muligt, at de faktiske stemmetal blot er et udtryk for den statistiske usikkerhed. Dette kan undersøges ved en såkaldt X^2 -test (Chi-i-anden test), der angiver sandsynligheden for at få et udfald der er mindst lige så ekstremt som udfaldet i Tabel 1, under antagelse af, at der ikke er en afhængighed mellem område og parti i forhold til stemmetal.

X^2 -test for uafhængighed er en velkendt disciplin i matematik i gymnasiet, og vil derfor ikke blive gennemgået i detaljer her, men der henvises til lærebøger på området. Der vil dog blive givet et kort resume her, da nogle af principperne også er fundamentale for korrespondanceanalyse.

Et vigtigt element for såvel X^2 -test som for korrespondanceanalyse er selve X^2 -afstanden. X^2 -afstanden er et mål for afstanden mellem de to fordelinger. X^2 -værdien er defineret som:

(1)

$$X^2 = \sum \frac{(\text{observeret antal} - \text{forventet antal})^2}{\text{forventet antal}}$$

Hvor det forventede antal er antallet ved uafhængighed mellem variablene. I korrespondanceanalyse vil man benytte den gennemsnitlige fordeling frem for den forventede værdi, men der er reelt tale om samme størrelse. Ud over X^2 -værdien angiver man antallet af frihedsgrader for observationerne. Antallet af frihedsgrader f beregnes som:

$$f = (m - 1)(n - 1), \text{ hvor } m, n \text{ angiver antallet af hhv. rækker og søjler.}$$

Ud fra χ^2 -værdien og antallet af frihedsgrader kan p-værdien beregnes i et statistikprogram. Jo større χ^2 -værdien er, desto mere ekstrem er observationen i forhold til den forventede fordeling. p-værdien angiver sandsynligheden for at få en χ^2 -værdi lig med eller større end den angivne værdi.

χ^2 -værdien er både afhængig af de relative fordelinger og det samlede antal observationer. Dette kan eksemplificeres ved, at det ikke er usædvanligt at slå 2 seksere ved kast med en terning 4 gange. Hvorimod det vil være langt mere usædvanligt at slå 50 seksere ved kast med samme terning 100 gange. I begge tilfælde er den gennemsnitlige succesrate 50 %.

I det konkrete tilfælde med valgresultatet er χ^2 -værdien så stor for en fordeling med fire frihedsgrader ($\chi^2 = 26106$) at der ikke kan beregnes en p-værdi større end 0. Man må derfor konkludere, at fordelingen af stemmerne på partierne bestemt ikke er uafhængig af området, hvilket nok også ville være en umiddelbar vurdering for det givne datasæt.

I gymnasiet vil en χ^2 -tests sædvanligvis alene blive brugt til at afgøre, hvilken sandsynlighed der er for at få en given observation, under antagelse af, at variablene er uafhængige, samt be- eller afkræfte en hypotese om uafhængighed, ud fra et givent signifikansniveau. Ved mange virkelige datasæt, og i særdeleshed ved de interessante, vil der ikke være tale om uafhængighed mellem observationsparametrene. Her kan korrespondanceanalysen bruges til at komme et spadestik dybere ned i analysen af dataene, da den kan bruges til at afdække, hvor afhængigheden mellem observationsparametrene ligger.

χ^2 -værdien i korrespondanceanalyse

Selve χ^2 -værdien er afhængig af det samlede antal observationer. Men i korrespondanceanalyse tager man udgangspunkt i de relative fordelinger i række- og søjleprofilerne. En omskrivning af χ^2 parameteren giver:

(2)

$$\chi^2 = \sum \frac{(d_{ij} - d_{ij}^{\#})^2}{d_{ij}^{\#}} = d_{++} \sum \frac{\left(\frac{d_{ij}}{d_{++}} - \frac{d_{ij}^{\#}}{d_{++}} \right)^2}{\frac{d_{ij}^{\#}}{d_{++}}}$$

Hvor # angiver de forventede værdier. d_{++} er det samlede antal observationer, og er derfor det samme for de observerede og forventede data. χ^2 delt med antallet af observationer kaldes fordelingsinerti og betegnes ϕ^2 . Udtrykket kan omskrives:

(3)

$$\phi^2 = \frac{\chi^2}{d_{++}} = \sum \frac{\left(\frac{d_{ij}}{d_{i+}} \frac{d_{i+}}{d_{++}} - \frac{d_{ij}^{\#}}{d_{i+}} \frac{d_{i+}}{d_{++}} \right)^2}{\frac{d_{ij}^{\#}}{d_{i+}} \frac{d_{i+}}{d_{++}}} = \sum \frac{d_{i+}}{d_{++}} \frac{\left(\frac{d_{ij}}{d_{i+}} - \frac{d_{ij}^{\#}}{d_{i+}} \right)^2}{\frac{d_{ij}^{\#}}{d_{i+}}} = \sum r_i \frac{(b_{ij} - c_j)^2}{c_j}$$

Inertien er alene beregnet ud fra de relative fordelinger. Man ville få den samme værdier for inertien ved en fordeling der er dobbelt så stor. Men selve X^2 -værdien ville naturligvis også blive dobbelt så stor.

Den samlede inerti for fordelingen kan skrives som:

(4)

$$\phi^2 = \frac{X^2}{d_{++}} = \sum_i r_i \sum_j \frac{(b_{ij} - c_j)^2}{c_j}$$

Det ses, at inertien er en sum af bidrag fra de enkelte rækker, vægtet i forhold til rækkernes masse. Det vil sige, at inertien kan beregnes som en sum af bidrag for hver række for sig. c_j angiver den gennemsnitlige rækkeprofil, også kaldet centroiden, for rækkeprofilerne.

Tilsvarende ville inertien også kunne beregnes ud fra en vægtning af inertien for søjleprofilerne:

(5)

$$\phi^2 = \frac{X^2}{d_{++}} = \sum_j c_j \sum_i \frac{(b_{ij} - r_i)^2}{r_i}$$

X²-afstanden

Hvis vi blot indtegnede rækkeprofilerne i et koordinatsystem ville afstanden mellem to profiler kunne bestemmes vha. Pythagoras' læresætning:

(6)

$$\text{Euklidisk afstand} = \sqrt{\sum_j (b_{ij} - b_{i'j})^2}$$

Hvor b_{ij} angiver koordinaterne til profil i , og $b_{i'j}$ angiver koordinaterne til profil i'

I korrespondanceanalyse er der dog en vis pointe i at benytte den vægtede afstand frem for den euklidiske afstand. I udtrykket for inertien indgår kvadratet på afstanden vægtet med en over massen af profilen. Vi kan derfor med fordel definere en X^2 -afstand som:

(7)

$$X^2 - \text{afstand} = \sqrt{\frac{(b_{ij} - b_{i'j})^2}{c_j}}$$

Koordinaterne for rækkeprofilerne kan dermed omskrives til X^2 -afstande, ved at skalere med en over kvadratroden af massen: $\frac{b_{ij}}{\sqrt{c_j}}$

Den samlede inerti for fordelingen kan nu udtrykkes vha. X^2 -afstandene:

$$\text{Inerti: } \phi^2 = \sum r_i \cdot (X^2 - \text{afstand})^2 = \sum_{i,j} r_i \left(\frac{b_{ij}}{\sqrt{c_j}} - \frac{c_j}{\sqrt{c_j}} \right)^2$$

Da vi her benytter at det er afstanden fra centroiden til profilen der skal benyttes ved beregning af inertien.

De enkelte rækker kan betragtes som punkter i et rum, hvor søjlerne repræsenterer dimensionerne. Det betyder, at dimensioner med stor masse vil komme til at have en kort afstand til centroiden, mens dimensioner med lille masse vil have en stor afstand til centroiden.

Hvis punkterne ligger tæt samlet omkring centroiden vil den samlede inerti for fordelingen være lille, da X^2 -afstandene vil være lille. Enkelte punkter kan godt ligge langt fra centroiden, uden at inertien af den grund er stor, såfremt massen af den profil er lille. Inertien siger noget om, hvor stor spredning der er mellem profilerne i fordelingen.

Den samlede inerti i fordelingen er summen af inerti bidrag for de enkelte profiler. Det giver derfor mening at tale om, hvor stort et bidrag hver enkelt profil har for den samlede inerti.

Et tal-eksempel

Hvis vi har en fordeling af en række observationer, som angivet i Tabel 5, hvor i alt 269 observationer er fordelt på 5 søjleparametre og 6 rækkeparametre, ville man kunne betragte fordelingen som en fordeling af de 6 rækkeprofiler i et 5-dimensionalt rum, angivet ved søjleparametrene.

	y ₁	y ₂	y ₃	y ₄	y ₅	I alt
x ₁	13	12	5	8	4	42
x ₂	11	16	8	6	2	43
x ₃	13	11	12	7	8	51
x ₄	5	7	16	12	4	44
x ₅	6	8	9	15	5	43
x ₆	2	10	9	17	8	46
I alt	50	64	59	65	31	269

Tabel 5 Tal-eksempel. Observeret fordeling

Tabellen kan omsættes til en relativ fordeling af observationerne i de enkelte rækker.

rækkeprofiler	y ₁	y ₂	y ₃	y ₄	y ₅	I alt	$r_i = \frac{d_{i+}}{d_{++}}$
---------------	----------------	----------------	----------------	----------------	----------------	-------	-------------------------------

x_1	0,309524	0,285714	0,119048	0,190476	0,095238	1	0,156134
x_2	0,255814	0,372093	0,186047	0,139535	0,046512	1	0,159851
x_3	0,254902	0,215686	0,235294	0,137255	0,156863	1	0,189591
x_4	0,113636	0,159091	0,363636	0,272727	0,090909	1	0,163569
x_5	0,139535	0,186047	0,209302	0,348837	0,116279	1	0,159851
x_6	0,043478	0,217391	0,195652	0,369565	0,173913	1	0,171004
$c_j = \frac{d_{+j}}{d_{++}}$	0,185874	0,237918	0,219331	0,241636	0,115242	1	1

Tabel 6 Relative rækkeprofiler for et tal-eksempel.

Hvor værdierne er beregnet ved: $b_{ij} = \frac{d_{ij}}{d_{i+}}$

Den yderste række og søjle angiver den marginale fordeling. Der er sket en reduktion af information i forhold til Tabel 5, da det samlede observerede antal ikke længere kan aflæses. Til gengæld kan informationerne i søjle y_5 aflæses ud fra informationerne i de øvrige søjler, da summen i hver række skal give 1. Så selv om punkterne har 5 koordinater vil de kun optage et 4-dimensionelt rum indenfor det 5-dimensionelle rum.

Den forventede fordeling af observationerne, hvis fordelingen mellem søjlerne var uafhængig af fordelingen mellem rækkerne, ville kunne beregnes som produktet af de gennemsnitlige række- og søjleprofiler gange det samlede antal:

(9)

$$d_{ij}^{\#} = r_i \cdot c_j^T \cdot d_{++}$$

Herefter kunne man beregne χ^2 -værdien, og teste for uafhængighed. Dette vil dog ikke blive gjort her.

Hvis man vil bestemme inertibidraget for en rækkeprofil, eller ønsker at indtegne punkterne i et koordinatsystem, vil man benytte χ^2 -afstandene som koordinater til punkterne. χ^2 -koordinaterne kan beregnes for tabellen:

χ^2 -afstand	y_1	y_2	y_3	y_4	y_5	Inerti $\phi_{række}^2$
x_1	0,717935	0,585758	0,254197	0,38749	0,280547	0,023735
x_2	0,593356	0,762847	0,397257	0,283859	0,137011	0,030559
x_3	0,591241	0,44219	0,502413	0,279221	0,462078	0,016873
x_4	0,263578	0,326161	0,776457	0,554815	0,267795	0,025889
x_5	0,323649	0,381424	0,446914	0,709647	0,342529	0,011332
x_6	0,100847	0,445685	0,417768	0,751814	0,512304	0,036084
Gennemsnit	0,431131	0,487769	0,468328	0,491565	0,339473	0

Tabel 7 χ^2 koordinater for rækkeprofilerne.

Værdierne er beregnet ud fra: $x_{ij} = b_{ij} / \sqrt{c_j}$

I sidste søjle er angivet inertien for rækkeprofilen. Denne er beregnet ud fra:

$\phi_i^2 = r_i \cdot \sum (X^2 - \text{afstand})^2$. Den samlede inerti for hele fordelingen er summen af inertien for de enkelte rækker $\phi_{total}^2 = \sum \phi_{række}^2 = 0,144471$. X^2 -værdien for fordelingen bliver dermed: $X^2 = d_{++} \cdot \phi_{total}^2 = 38,9$. Tabellen indeholder i alt 20 frihedsgrader. Den tilsvarende p-værdi er 0,69 %, hvilket indikerer, at fordelingen ikke er tilfældig, da der kun er en sandsynlighed på 0,69 % for at få et udfald der er mindst lige så ekstremt, under antagelse af uafhængighed.

Det ses ud fra Tabel 7, at de enkelte profiler ikke bidrager lige meget til den samlede inerti.

Profilen x_5 bidrager med $\frac{0,011332}{0,14447} \cdot 100 \% = 7,8 \%$ af den samlede inerti mens profilen x_6

bidrager med 25%. De to punkter har næsten samme masse, så forskellen skyldes primært, at x_5 har en rækkeprofil, der ligger relativt tæt på centroiden, mens x_6 ligger relativt langt fra. Med andre ord ligger fordelingen i x_5 tæt op ad den gennemsnitlige fordeling, mens punktet x_6 ligger langt fra den gennemsnitlige fordeling. En yderligere inspektion af X^2 afstandene afslører, at det primært er fordelingerne på parametrene y_1 og y_5 for x_6 , der har en stor afstand til centroiden.

Ud fra X^2 -afstandene er det også muligt at beregne den indbyrdes afstand mellem de forskellige punkter. X^2 -afstandene mellem punkterne er vist i Tabel 8. Dette kan benyttes til at identificere, hvilke kategorier der ligger tæt sammen i det 5-dimensionale rum.

	x_1	x_2	x_3	x_4	x_5
x_1	0				
x_2	0,314148	0			
x_3	0,378087	0,468586509	0		
x_4	0,758121	0,730591099	0,556438	0	
x_5	0,584791	0,666501842	0,527197	0,380551	0
x_6	0,783337	0,838691941	0,688136	0,517705	0,291927

Tabel 8 X^2 -afstanden mellem punkterne.

Eksemplet ovenfor er blot et tal-eksempel. Men i et reelt datasæt ville x og y parametrene repræsentere forskellige variable. F.eks. kunne x -parameteren repræsentere hårfarven og y -parameteren repræsentere øjenfarven hos en række personer. Ud fra Tabel 8 kunne man argumentere for, at kategorierne x_5 og x_6 eller x_1 og x_6 kunne slås sammen til én kategori, da de ligger relativt tæt på hinanden i rummet.

Søjleprofiler

Hidtil har vi behandlet række-kategorierne som forskellige punkter i et rum der er udspændt af søjle-kategorierne. Da der i eksemplet ovenfor er 5 forskellige søjle-kategorier, ligger rækkeprofilerne i et 5-dimensionelt rum. Alle operationerne vi hidtil har udført på rækkeprofilerne, kan man også udføre på søjleprofilerne.

Søjleprofiler $a_{ij} = \frac{d_{ij}}{d_{+j}}$						$r_i = \frac{d_{i+}}{d_{++}}$
	y_1	y_2	y_3	y_4	y_5	
x_1	0,26	0,1875	0,084746	0,123077	0,129032	0,156134

x ₂	0,22	0,25	0,135593	0,092308	0,064516	0,159851
x ₃	0,26	0,171875	0,20339	0,107692	0,258065	0,189591
x ₄	0,1	0,109375	0,271186	0,184615	0,129032	0,163569
x ₅	0,12	0,125	0,152542	0,230769	0,16129	0,159851
x ₆	0,04	0,15625	0,152542	0,261538	0,258065	0,171004
I alt	1	1	1	1	1	1
$c_j = \frac{d_{+j}}{d_{++}}$	0,185873606	0,237918	0,219331	0,241636	0,115242	

Tabel 9 Tabel over søjleprofilerne

Det bemærkes, at de marginale fordelinger r_i og c_j er uændrede. Fordelingen kan nu betragtes som en fordeling af de 5 søjleprofiler i et 6-dimensionelt rum, angivet ved rækkekategoriene.

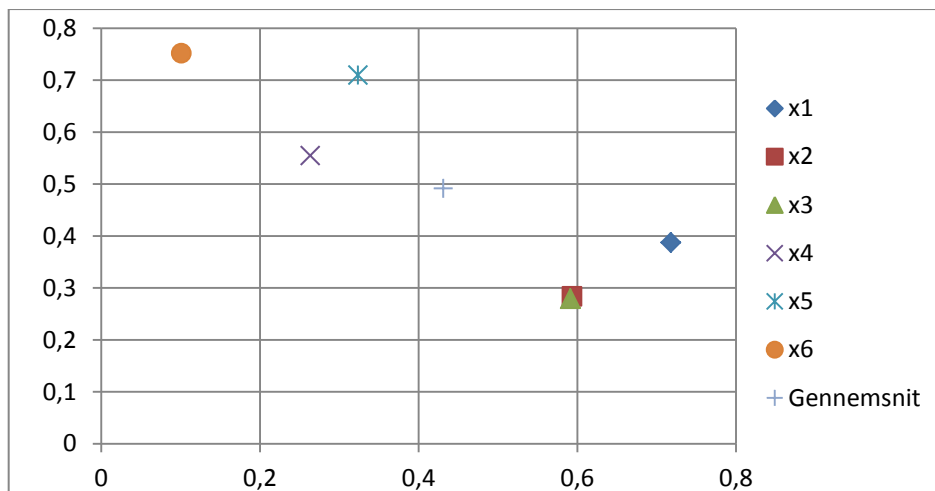
Ligesom ved rækkeprofilerne, kan der ved søjleprofilerne laves en tabel over X^2 -afstandene og hvor meget hver søjle bidrager til den samlede inerti.

X^2 -afstand	y ₁	y ₂	y ₃	y ₄	y ₅	Gennemsnit
x ₁	0,658	0,475	0,214	0,311	0,327	0,395
x ₂	0,550	0,625	0,339	0,231	0,161	0,400
x ₃	0,597	0,395	0,467	0,247	0,593	0,435
x ₄	0,247	0,270	0,671	0,456	0,319	0,404
x ₅	0,300	0,313	0,382	0,577	0,403	0,400
x ₆	0,097	0,378	0,369	0,632	0,624	0,414
Profilafstand	0,503	0,293	0,332	0,391	0,371	0,000
Inerti	0,047	0,020	0,024	0,037	0,016	0,144
Relativt inertibidrag	32,53 %	14,10 %	16,77 %	25,59 %	11,00 %	

Tabel 10 X^2 -afstand og inertien for søjleprofiler

Ud fra Tabel 10 ses det, at søjlerne y₁ og y₄ er de profiler, der bidrager mest til den samlede inerti. Tilsammen dækker de over 58,12 % af den samlede spredning i datasættet.

Det er ikke muligt at indtegne hverken rækkeprofilerne eller søjleprofilerne retvisende i et koordinatsystem, da det i praksis kun er muligt at angive spredningen i to dimensioner grafisk. Men ud fra fordelingen af inerti i Tabel 10 kan man identificere hvilke dimensioner, der vægter højest, når spredningen af datasættet skal beskrives. I dette datasæt vil dimensionerne y₁ og y₄ give den største forklaring på spredningen. Hvis man skulle illustrere dataene grafisk, alene ud fra de to dimensioner, ville fordelingen se ud som vist på Figur 7.



Figur 7 Rækkeprofiler, hvor y₁ angiver 1. koordinaten og y₄ angiver 2. koordinaten

Figur 7 viser X^2 -afstandene for dimensionerne y₁ og y₄. Men, da disse to parametre kun repræsenterer 58,1% af den samlede inertie, er der 41,9%, der ikke er illustreret. Grafen skal derfor læses med et vist forbehold. Det ses, at punkterne x₅ og x₆ har stor indbyrdes afstand, hvilket bekræftes af Tabel 8. Punkterne x₂ og x₃ er næsten sammenfaldende på figuren, hvilket de ikke er ifølge tabellen. Forklaringen ligger i, at separationen mellem x₂ og x₃ primært ligger i de dimensioner, der ikke er illustreret på figuren.

Generelt vil det være sådan, at profiler der ligger langt fra hinanden i det todimensionelle plan, også gør det i det flerdimensionelle rum, mens profiler der ligger tæt, ikke nødvendigvis ligger tæt sammen i det flerdimensionelle rum.

I eksemplet ovenfor er profilerne fra det 5-dimensionelle rum projiceret ned på et todimensionelt rum, der er udspændt af y₁ og y₄. Dette gør det muligt at præsentere dataene grafisk. Prisen er, at der sker en reduktion af information i forhold til de oprindelige data. I dette tilfælde er det kun 58,1% af spredningen, der er bevaret.

Princippet i et korrespondanceanalyseplot er meget lig det foregående eksempel. Men, i stedet for blot at udvælge to af de eksisterende parametre som basis for det todimensionelle underrum, vælges et underrum der giver minimalt tab af forklaring, og dermed bevarer mest mulig spredning af de oprindelige data. Dette vil blive behandlet i det næste afsnit.

Maksimal inert

Den maksimale inert i en fordeling vil fremkomme, når alle profilerne kun har udfald i én af kategorierne, dvs. når rækkeprofilerne er sammenfaldende med yderpunkterne (kategorierne). En sådan fordeling kan betragtes som en diagonal fordeling, hvor de diagonale værdier er 1.

	Kategori 1	Kategori 2	...	Kategori k	c
b_1	1	0	...	0	c_1
b_2	0	1	...	0	c_2
...	...	...	...	0	...
b_k	0	0	0	1	c_k
r	c_1	c_2	...	c_k	

Tabel 11 Relativ fordeling hvis alle profiler lå udelukkende i én kategori.

Det bemærkes, at ved denne fordeling vil massen af rækkerne og massen af søjlerne være identiske. For at bestemme den samlede inert vil man kunne opstille en tabel over X^2 -afstandene fra den gennemsnitlige fordeling:

X^2 -afstande	Kategori 1	Kategori 2	...	Kategori k	c
b_1	$\sqrt{\frac{(1-c_1)^2}{c_1}}$	$\sqrt{c_2}$	...	$\sqrt{c_k}$	c_1
b_2	$\sqrt{c_1}$	$\sqrt{\frac{(1-c_2)^2}{c_2}}$	...	$\sqrt{c_k}$	c_2
...	...	...	...	0	...
b_k	$\sqrt{c_1}$	$\sqrt{c_2}$	...	$\sqrt{\frac{(1-c_k)^2}{c_k}}$	c_k
r	c_1	c_2	...	c_k	

Tabel 12 X^2 -afstanden beregnet ud fra Tabel 11

Inertbidragene kan bestemmes for hvert element for sig, som kvadratet på X^2 -afstanden, vægtet med massen af profilen:

ϕ^2	Kategori 1	Kategori 2	...	Kategori k	c
b_1	$(1-c_1)^2$	$c_1 c_2$	...	$c_1 c_k$	c_1
b_2	$c_2 c_1$	$(1-c_2)^2$	...	$c_2 c_k$	c_2
...	...	...	...	0	...
b_k	$c_k c_1$	$c_k c_2$	...	$(1-c_k)^2$	c_k
r	c_1	c_2	...	c_k	

Tabel 13 Inertbidrag fra de enkelte elementer.

Den samlede inert fra første række kan omskrives til:

(10)

$$\phi_1^2 = (1-c_1)^2 + c_1 \sum_{i=2}^k c_i = (1-c_1)^2 + c_1(1-c_1) = 1-c_1$$

Tilsvarende er inertibidraget fra de øvrige profiler $\phi_i^2 = 1 - c_i$. Den samlede inerti bliver dermed:

(11)

$$\phi^2 = \sum_1^k \phi_i^2 = \sum_1^k 1 - c_1 = k - 1$$

Den maksimale inerti i en given fordeling vil være antallet af kategorier minus en.

Dimensioner i et datasæt

I det foregående afsnit så vi på datasæt i flere dimensioner. Mens man uden problemer kan beregne X^2 afstandene for data i flere dimensioner, er det ikke umiddelbart muligt at præsentere dataene grafisk, for mere end to uafhængige dimensioner.

Ofte er data opgivet i tabeller med flere dimensioner i hver variabel. I dette afsnit vil vi se nærmere på hvordan data i flere end to dimensioner kan illustreres ved at projicere dem ned på et todimensionelt underrum.

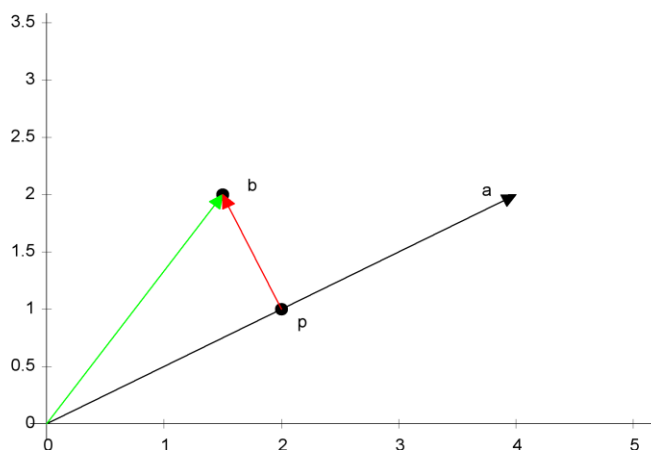
Projektion på et underrum

For at kunne reducere antallet af dimensioner, er det nødvendigt at kunne projicere dataene fra et rum med flere dimensioner ned på et underrum på f.eks. to dimensioner. I tal-eksemplet i det foregående afsnit valgte vi at se bort fra de tre dimensioner med mindst bidrag til den samlede inerti i fordelingen. Dette er måske en enkel løsning, men ikke særligt hensigtsmæssig. I eksemplet fik vi kun bevaret 58% af spredningen i datasættet, ved anvendelse af X^2 -værdien som mål for spredningen.

Da det ikke er muligt at tegne punkter i rum på 4 eller 5 dimensioner, illustreres princippet først ved en projektion af et punkt i 2 dimensioner ned på et underrum i én dimension.

Projektion fra to dimensioner til én dimension

Ved en projektion af et punkt ned på en linje vælges det punkt på linjen, der ligger tættest på punktet. På Figur 8 er illustreret projektionen af punktet b ned på linjen angivet ved vektoren a . Vektoren til punktet p kan ses som en skalering af vektoren a : $p = xa$. Vektoren $e = b - p = b - xa$ er et udtryk for den fejl, vi får ved projektionen. Målet ved projektionen er at minimere denne fejl, hvilket opnås ved at lave en vinkelret projektion ned på linjen.



Figur 8 Skitse af projektion af punktet b ned på linjen udspændt af vektor a .

Ortogonaliteten mellem vektorerne e og a kan matematisk formuleres ved, at deres indre produkt (prik-produktet) er nul:

(12)

$$a^T e = a^T (b - xa) = 0$$

Som igen kan omskrives til:

(13)

$$xa^T a = a^T b$$

Da $a^T a$ blot er en skalar, kan x -værdien bestemmes som:

(14)

$$x = \frac{a^T b}{a^T a}$$

Vektoren til punktet p kan bestemmes ved:

(15)

$$p = ax = a \frac{a^T b}{a^T a} = Pb$$

Hvor P er projektionsmatricen:

(16)

$$P = \frac{aa^T}{a^T a}$$

Det bemærkes at P altid vil være en symmetrisk matrice da $P = \frac{1}{a^T a} \begin{bmatrix} a_1 a_1 & a_1 a_2 \\ a_1 a_2 & a_2 a_2 \end{bmatrix}$.

I eksemplet fra Figur 8 er $a = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$ og projektionsmatricen:

(17)

$$P = \frac{aa^T}{a^T a} = \frac{\begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 4 & 2 \end{bmatrix}}{\begin{bmatrix} 4 & 2 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix}} = \frac{1}{20} \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix}$$

Projektionen af punktet $b=(1,5;2)$ bliver:

(18)

$$p = Pb = \frac{1}{20} \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix} \begin{bmatrix} 1,5 \\ 2 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 40 \\ 20 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Alle punkter i de todimensionelle rum vil kunne projiceres vinkelret ned på linjen udspændt ved vektoren a , ved at gange vektorerne med projektionsmatricen.

Hvis punktet allerede ligger på linjen vil det ikke ændre koordinater, og hvis projektionen udføres to gange, vil det ikke flytte sig ved den anden projektion. Det ses da også at hvis projektionsmatricen bliver ganget på vektoren $(2,1)$ får man igen vektoren $(2,1)$.

Projektion fra n dimensioner ned på en linje.

Proceduren for projektion fra to dimensioner ned på én dimension (en ret linje) kan direkte overføres til projektion fra n dimensioner ned på én dimension. Forskellen ville være at den

vektor der udspænder den rette linje nu vil have n koordinater, og projektionsmatricen vil blive en $n \times n$ matrice.

Som eksempel kan man betragte projektionen af punktet $b=(1; 2; 3)$ ned på et underrum, udspændt af vektoren $a = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$. Det ses at b ikke ligger i det underrum, der er givet ved a da man ikke kan nå b ved en skalering af a . Projektionsmatricen bliver:

(19)

$$P = \frac{aa^T}{a^T a} = \frac{\begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \begin{bmatrix} 2 & 1 & 3 \end{bmatrix}}{\begin{bmatrix} 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}} = \frac{1}{14} \begin{bmatrix} 4 & 2 & 6 \\ 2 & 1 & 3 \\ 6 & 3 & 9 \end{bmatrix}$$

Projektionen af punktet $b = (1; 2; 3)$ bliver:

(20)

$$p = Pb = \frac{1}{14} \begin{bmatrix} 4 & 2 & 6 \\ 2 & 1 & 3 \\ 6 & 3 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \frac{1}{14} \begin{bmatrix} 26 \\ 13 \\ 39 \end{bmatrix} = \begin{bmatrix} 26/14 \\ 13/14 \\ 39/14 \end{bmatrix}$$

Det kan konstateres at projektionen er vinkelret ved at vise at prikproduktet mellem fejlen e og vektoren p er 0:

$$e = b - p = \begin{bmatrix} -12/14 \\ 15/14 \\ 3/14 \end{bmatrix}$$

$$p \cdot e = \frac{1}{14^2} (26 \cdot (-12) + 13 \cdot 15 + 39 \cdot 3) = 0$$

Projektion fra n dimensioner ned på 2 dimensioner

Hvor et underrum i én dimension vil være givet ved en ret linje, der kan skrives som en skalering af en enkelt vektor, vil et underrum i to dimensioner være en flade der kan skrives som en linearkombination af to uafhængige vektorer. Dvs. at så længe de to vektorer ikke er parallelle, vil alle punkter på en flade kunne nås ved en skalering af den ene vektor plus en skalering af den anden vektor. Alternativt kan man sige at underrummet er udspændt af de to vektorer.

Hvis et punkt b ligger i et tredimensionelt rum, kan det projiceres ned på en flade der er udspændt af to uafhængige vektorer. Punktet p for projektionen vil være det punkt på fladen, der ligger tættest på punktet b . Som i det tidligere eksempel vil afstanden mellem b og p være et udtryk for fejlen ved projektionen. Hvis b og p angiver vektorerne til punkterne vil fejlen e kunne skrives som $e = b - p$. Da e angiver den korteste afstand fra planet til punktet b vil e være vinkelret på p , men også vinkelret på alle øvrige vektorer i planet. Da planet var udspændt af to vektorer (a_1 og a_2) vil e også være vinkelret på disse.

De to vektorer kan skrives som søjlevektorer i matricen $\mathbf{A}$:

(21)

$$\mathbf{A} = \begin{bmatrix} \vdots & \vdots \\ \mathbf{a}_1 & \mathbf{a}_2 \\ \vdots & \vdots \end{bmatrix}$$

Det antages at punktet $\mathbf{b}$ ikke ligger i det underrum der bliver udspændt af de to vektorer, og derfor ikke kan skrives som en linearkombination af dem. Men det kan punktet $\mathbf{p}$, da det ligger i underrummet.

(22)

$$\mathbf{p} = x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 = \mathbf{A}\mathbf{x}$$

Da vektoren $\mathbf{e}$ er vinkelret på både $\mathbf{a}_1$ og $\mathbf{a}_2$ vil følgende ligninger gælde:

(23)

$$\mathbf{a}_1^T \mathbf{e} = \mathbf{a}_1^T (\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}$$

$$\mathbf{a}_2^T \mathbf{e} = \mathbf{a}_2^T (\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}$$

Hvilket kan omskrives til:

(24)

$$\mathbf{A}^T (\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0} \Leftrightarrow$$

$$\mathbf{A}^T \mathbf{A}\mathbf{x} = \mathbf{A}^T \mathbf{b}$$

Hvor $\mathbf{A}^T \mathbf{A}$ er en $n \times n$ matrice.

$\mathbf{x}$ kan nu findes ved at gange med den inverse matrice til $\mathbf{A}^T \mathbf{A}$:

(25)

$$\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$$

Og vektoren fra projektionen af $\mathbf{b}$ bliver:

(26)

$$\mathbf{p} = \mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} = \mathbf{P}\mathbf{b}$$

Hvor projektionsmatricen bliver $\mathbf{P} = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$

Det ses at forskellen ifht. projektion på en ret linje er at $\mathbf{A}^T \mathbf{A}$ ikke er en skalar men en matrice, da $\mathbf{A}$ indeholder to søjlevektorer. Det er endvidere et krav at $\mathbf{A}^T \mathbf{A}$ har en invers, hvilket vil være tilfældet hvis vektorerne i $\mathbf{A}$ er uafhængige.⁴

Projektionsmatricen vil overføre ethvert punkt i det oprindelige n -dimensionale rum til det nærmeste punkt i underrummet udspændt af søjlevektorerne i $\mathbf{A}$. Ved at gentage projektionen

⁴ Spencer, side 395

vil man ramme det samme punkt igen. Matematisk kan en gentagelse af projektionen formuleres som projekktionsmatricen ganget med sig selv, hvilke vil resultere i den samme projekktionsmatrice. Dette kan matematisk formuleres som:

(27)

$$P^2 = P$$

Hvilket kan vises ved at indsætte udtrykket for P :

(28)

$$P^2 = A(A^T A)^{-1} A^T A(A^T A)^{-1} A^T = A I (A^T A)^{-1} A = A(A^T A)^{-1} A^T = P$$

I eksemplet ovenfor er det antaget at underrummet er todimensionelt. Principielt kan der projiceres ned på et underrum af vilkårlig dimension. A skal blot indeholde de uafhængige vektorer, der udspænder underrummet.

Vi har nu det nødvendige værktøj til at kunne projicere profilerne fra vores undersøgelse ned på et underrum af to dimensioner, men vi mangler stadig at få defineret det optimale underrum at projicere vores profiler ned på.

Egenverdier og egenvektorer for en symmetrisk matrice

For at kunne bestemme det bedste underrum starter vi med at definere egenverdier og egenvektorer for en kvadratisk matrice A . Dette er standard tekstbogsstof omkring matricer, og vil kun blive behandlet ganske kort her. Definitionen af egenverdi og egenvektor fremkommer som løsning til egenverdiligningen:

(29)

$$Ax = \lambda x$$

Hvor egenverdien λ er en skalar og x er den tilhørende egenvektor.

Når en kvadratisk matrice ganges på en egenvektor vil vektoren ikke ændre retning, men blot blive skaleret med egenverdien. Alle vektorer der er parallelle med en egenvektor er selv en egenvektor.

At bestemme egenverdier og egenvektorer for en matrice A kaldes at løse egenverdioproblemet. Ligningen kan omskrives til:

(30)

$$(A - \lambda I)x = 0$$

Da vektoren x ikke må være 0 -vektoren skal determinanten til vektoren $(A - \lambda I)$ være 0:

(31)

$$\det(A - \lambda I) = 0$$

Ved at løse denne ligning findes de mulige egenverdier, og ved at indsætte egenverdierne i den foregående ligning kan egenvektorerne bestemmes.

Specielt når A er symmetrisk dvs. $A = A^T$ gælder at egenverdierne alle er reelle.

Ortogonal egenvektorer

For en symmetrisk matrice vil det gælde at egenvektorerne for forskellige egenverdier vil være ortogonale.⁵ Dette vises i det følgende.

Egenverdierne og egenvektorerne er løsninger til egenverdiligningen. Hvis vi betragter to forskellige løsninger har vi:

(32)

$$Ax_i = \lambda_i x_i$$

$$Ax_j = \lambda_j x_j$$

Hvor $\lambda_i \neq \lambda_j$

Ved at gange med den transponerede for den anden egenvektor fremkommer:

(33)

$$x_j^T Ax_i = \lambda_i x_j^T x_i$$

$$x_i^T Ax_j = \lambda_j x_i^T x_j$$

Ved at transponere den første ligning fås:

(34)

$$x_i^T A^T x_j = \lambda_i x_i^T x_j$$

Da A er symmetrisk er $A^T = A$:

(35)

$$x_i^T A^T x_j = x_i^T Ax_j = \lambda_j x_i^T x_j = \lambda_i x_i^T x_j$$

Da de to egenverdier er forskellige, $\lambda_j \neq \lambda_i$ må:

(36)

$$x_i^T x_j = 0$$

Hvilket netop er kravet for ortogonalitet, dvs. $x_i \perp x_j$.⁶

Hvis en egenverdi er en multipel rod til egenverdiproblemet, med multiplicitet k , vil de tilhørende egenvektorer udgøre et k -dimensionelt underrum. Egenvektorerne kan da vælges som ortogonale enhedsvektorer i dette underrum.⁷ Dette vil ikke blive behandlet yderligere her.

⁵ Spence side 425

⁶ Conradsen s29

⁷ (3) s114, Spence s426

Et tal-eksempel

Vi ønsker at bestemme egenværdier og egenvektorer for følgende matrice:

(37)

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Det ses at $\mathbf{A}$ er symmetrisk.

(38)

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \det \left(\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} \right) = (2-\lambda)^2 - 1 = 0$$

$$\lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) = 0$$

Det ses at egenværdierne er $\lambda = 1$ og $\lambda = 3$

For $\lambda = 3$ har vi:

(39)

$$(\mathbf{A} - 3\mathbf{I})\mathbf{x} = 0$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

En mulig løsning er:

(40)

$$x_1 = x_2$$

$$\text{Egenvektoren bliver } \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For $\lambda = 1$ har vi:

(41)

$$(\mathbf{A} - \mathbf{I})\mathbf{x} = 0$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Her er en mulig løsning:

(42)

$$x_1 = -x_2$$

$$\text{Egenvektoren bliver } \mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Selve længden af egenvektorerne er arbitrært valgt. Man kunne for entydighedens skyld vælge

$$\text{at normere den til hhv. } \mathbf{e}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \text{ og } \mathbf{e}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

Som forventet er egenvektorerne ortogonale.

Det maksimale antal egenverdier er bestemt af graden af det polynomium der fremkommer ved løsning af ligning (31), hvis $\mathbf{A}$ er en $k \times k$ matrice er der maksimalt k mulige egenverdier. Antallet af egenverdier forskelligt fra nul er givet ved rangen af matricen $\mathbf{A}$.

Spektral dekomposition

Hvis $\mathbf{\Lambda}$ angiver $k \times k$ en diagonalmatrice med egenverdierne som elementer, arrangeret efter størrelse, og $\mathbf{X}$ angiver en matrice hvor søjlerne er de tilsvarende normerede egenvektorer, kan egenverdiligningen opskrives som:

(43)

$$\mathbf{AX} = \mathbf{X}\mathbf{\Lambda}$$

Da søjlevektorerne i $\mathbf{X}$ er ortogonale er $\mathbf{XX}^T = \mathbf{I}$, hvilket igen betyder at enhver symmetrisk matrice kan skrives som produktet af en ortogonal matrice, en diagonal matrice og en transponeret ortogonal matrice:

(44)

$$\mathbf{A} = \mathbf{AXX}^T = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T$$

Matricen kan også omskrives som en sum af bidrag fra de enkelte egenverdi- /egenvektorpar:

(45)

$$\mathbf{A} = \sum_{i=1}^k \lambda_{ii} \mathbf{e}_i \mathbf{e}_i^T$$

Idet det bemærkes at produktet af de ortogonale vektorer $\mathbf{e}_i \mathbf{e}_j^T$ giver 0. Dette kaldes også for en spektral dekomposition af $\mathbf{A}$ ⁸. Der er et bidrag fra hver egenvektor, vægtet med vektorens egenverdi. Da elementerne i $\mathbf{\Lambda}$ er sorteret efter værdi, med λ_{11} som den største, vil bidragene blive mindre for stigende i . Dette vil få betydning når vi i korrespondanceanalysen ønsker at reducere antallet af dimensioner på den mest hensigtsmæssige måde.

Et tal-eksempel

Ved tal-eksemplet fra tidligere havde vi fundet egenverdier og egenvektorer for matricen

$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Matricen vil kunne udtrykkes ved matricerne $\mathbf{X}$ og $\mathbf{\Lambda}$:

(46)

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-T} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Egenvektorerne blev bestemt til $\mathbf{e}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ og $\mathbf{e}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$

$\mathbf{A}$ kan udtrykkes ved spektral dekomposition:

⁸ lentilucci 2

(47)

$$\mathbf{A} = \sum_{i=1}^k \lambda_i \mathbf{e}_i \mathbf{e}_i^T = 3 \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} + 1 \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Definition af de singulære vektorer

De datamatricer vi ønsker at behandle i korrespondanceanalysen er sjældent hverken kvadratiske eller symmetriske. Men vil være rektangulære matricer hvor værdierne i matricen vil være angivet ved ϕ^2 -afstande.

Rækkeprofilerne i vores observationssæt udgør punkter i et n -dimensionelt rum. I korrespondanceanalysen bestemmes det todimensionelle underrum, hvor profilerne kan projiceres ind på, således at den størst mulige spredning af profilerne er bevaret. Dette vil være ensbetydende med at man projicerer ind på et plan der giver den mindst mulige fejl ved projektionen. Ved projektion af nul-vektoren skal man igen få nul-vektoren. Det vil sige at projektiionsplanet skal gå gennem origo.

Hver rækkeprofil $\mathbf{b}_i$ kan betragtes som en n -dimensional vektor, hvis den projicerede vektor kaldes $\mathbf{p}_i$ og fejlen ved projektionen $\mathbf{e}_i$ kan sammenhængen skrives som:

(48)

$$\mathbf{b}_i = \mathbf{p}_i + \mathbf{e}_i$$

Det optimale underrum i denne sammenhæng, er defineret som det underrum der giver den mindste sum af kvadrater på længden af fejlene for alle vektorerne i matricen. Det vil sige at $\sqrt{\sum |\mathbf{e}_i|^2}$ skal minimeres. Kvadratet på fejlen for hver vektor kan skrives som:

(49)

$$|\mathbf{e}_i|^2 = |\mathbf{b}_i|^2 - |\mathbf{p}_i|^2$$

Idet det bemærkes at $\mathbf{e}_i$ og $\mathbf{p}_i$ er ortogonale og længden af de tre vektorer derfor udgør en retvinklet trekant. Da længden af vektorerne $\mathbf{b}_i$ er uafhængig af hvilket underrum der vælges, kan optimeringen opnås ved at vælge et underrum, der maksimerer summen af kvadratet længden af projektiionerne i stedet.

Rækkeprofilerne kan opstilles som rækkevektorer i matricen $\mathbf{A}$. Vi antager, at $\mathbf{v}$ er en enhedsvektor langs den "bedste rette linje", dvs. den linje der giver det mindste kvadrat på fejlen, ved en projektion af rækkevektorerne ned på linjen. Længden af hver projektiion kan skrives som $|\mathbf{a}_i \cdot \mathbf{v}|$. Dermed bliver summen af kvadratet af længderne af projektiionerne $|\mathbf{A}\mathbf{v}|^2$. Den bedste rette linje følger altså den vektor $\mathbf{v}$, der maksimerer $|\mathbf{A}\mathbf{v}|^2$.

Den første singulære vektor $\mathbf{v}_1$ kan defineres som den vektor der maksimerer $|\mathbf{A}\mathbf{v}|$:

(50)

$$\mathbf{v}_1 = \arg \max_{|\mathbf{v}|=1} |\mathbf{A}\mathbf{v}|$$

Længden af vektoren kaldes den første singulære værdi: $\sigma_1(\mathbf{A}) = |\mathbf{A}\mathbf{v}_1|$. Kvadratet på denne værdi er samtidig summen af kvadratet på længden af projektiionerne ned på linjen.

Vektoren $\mathbf{v}_1$ angiver det bedste éndimensionelle underrum. Det bedste 2-dimensionelle underrum må indeholde $\mathbf{v}_1$ som den ene basisvektor, samt en vektor $\mathbf{v}_2$, der er ortogonal med $\mathbf{v}_1$. $\mathbf{v}_2$ kan bestemmes som den vektor, vinkelret på $\mathbf{v}_1$, der maksimerer $|\mathbf{A}\mathbf{v}|$:

(51)

$$\mathbf{v}_2 = \arg \max_{\mathbf{v} \perp \mathbf{v}_1, |\mathbf{v}|=1} |\mathbf{A}\mathbf{v}|$$

$\mathbf{v}_2$ er den anden singulære vektor og $\sigma_2(\mathbf{A}) = |\mathbf{A}\mathbf{v}_2|$ er den anden singulære værdi. Tilsvarende kan de efterfølgende singulære vektorer og værdier bestemmes ved at finde nye ortogonale basisvektorer for underrummet, der maksimerer længden af projektionen.

Når alle vektorer, der udspænder det underrum $\mathbf{A}$ ligger i, er fundet, vil $|\mathbf{A}\mathbf{v}| = 0$ for alle $\mathbf{v}$, der er ortogonale med de fundne singulære vektorer. Antallet af singulære vektorer, med singulære værdier forskellig fra 0, vil være givet ved rangen af matricen $\mathbf{A}$.⁹

De singulære værdier σ_i kan betragtes som størrelsen af $\mathbf{A}$, i retning af $\mathbf{v}_i$. Dermed skal det samlede kvadrat af indholdet i $\mathbf{A}$ være det samme som summen af kvadratet af de singulære værdier:

(52)

$$\sum (a_{ij})^2 = \sum \sigma_i^2$$

Hvilket er matrix-analogien til at opdele en vektor i dens komponenter langs de ortogonale enhedsvektorer.

For hver række $\mathbf{a}_j$ er længden projiceret ind på de ortonormale vektorer:

(53)

$$|\mathbf{a}_j|^2 = \sum_i (\mathbf{a}_j \cdot \mathbf{v}_i)^2$$

Summen af alle rækker giver:

(54)

$$\sum_j |\mathbf{a}_j|^2 = \sum_j \sum_i (\mathbf{a}_j \cdot \mathbf{v}_i)^2 = \sum_i \sum_j (\mathbf{a}_j \cdot \mathbf{v}_i)^2 = \sum_i |\mathbf{A}\mathbf{v}_i|^2 = \sum_i \sigma_i^2(\mathbf{A})$$

Hvor første led netop er summen af kvadratet på alle led i matricen, $\sum (a_{ij})^2$.

Kvadratroden af denne værdi kaldes også Frobenius normen for matricen:

(55)

$$\|\mathbf{A}\|_F = \sqrt{\sum (a_{ij})^2}$$

$\mathbf{v}_i$ kaldes de "højre singulære vektorer". Ved at projicere matricen ind på disse vektorer fremkommer et nyt sæt ortogonale vektorer, kaldet "venstre singulære vektorer":¹⁰

⁹ Hopcroft s112

¹⁰ Hopcroft s114

(56)

$$\mathbf{u}_i = \frac{1}{\sigma_i(\mathbf{A})} \mathbf{A} \mathbf{v}_i$$

Der normeres ved at dividere med $\sigma_i(\mathbf{A})$. Denne ligning kan omskrives til:

(57)

$$\mathbf{A} \mathbf{v}_i = \sigma_i(\mathbf{A}) \mathbf{u}_i$$

Hvis $\mathbf{v}_i$ og $\mathbf{u}_i$ indføres som søjlevektorer i matricerne $\mathbf{V}$ og $\mathbf{U}$ hhv. og de singulære værdier skrives som elementer i en diagonal matrice $\mathbf{\Sigma}$, kan sammenhængen for alle vektorerne skrives som:

(58)

$$\mathbf{A} \mathbf{V} = \mathbf{\Sigma} \mathbf{U}$$

$$\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$$

Der genkendes som en singular value decomposition af matricen $\mathbf{A}$.

Singular Value Decomposition

Ved hjælp af Singular Value Decomposition(SVD) er det muligt at udtrykke enhver $m \times n$ matrice ved et produkt af en $m \times m$ ortonormal matrice $\mathbf{U}$, en $m \times n$ diagonal matrice $\mathbf{\Sigma}$ og en $n \times n$ transponeret ortonormal matrice $\mathbf{V}^T$.

(59)

$$\mathbf{B} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$$

$\mathbf{B}$ er en $m \times n$ matrice. Det er altid muligt at lave en symmetrisk, kvadratisk matrice ved at gange med den transponerede matrice - $\mathbf{B}^T \mathbf{B}$. Dermed får man:

(60)

$$\mathbf{B}^T \mathbf{B} = \mathbf{V} \mathbf{\Sigma}^T \mathbf{U}^T \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \mathbf{\Sigma} \mathbf{V}^T = \mathbf{V} \mathbf{\Sigma}^T \mathbf{\Sigma} \mathbf{V}^T$$

Da $\mathbf{U}$ er en ortonormal matrice er $\mathbf{U}^T \mathbf{U}$ enhedsmatricen. Vi ser nu at $\mathbf{V}$ er egenvektorerne til $\mathbf{B}^T \mathbf{B}$ og $\mathbf{\Sigma}^T \mathbf{\Sigma}$ er en diagonalmatrice med de tilsvarende egenværdier:

(61)

$$\mathbf{B}^T \mathbf{B} \mathbf{V} = \mathbf{V} \mathbf{\Sigma}^T \mathbf{\Sigma}$$

Tilsvarende får vi, hvis vi ganger med $\mathbf{B}^T$ fra højre side:

(62)

$$\mathbf{B} \mathbf{B}^T = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \mathbf{V} \mathbf{\Sigma}^T \mathbf{U}^T = \mathbf{U} \mathbf{\Sigma} \mathbf{\Sigma}^T \mathbf{U}^T$$

Og vi ser at $\mathbf{U}$ er en ortonormal matrice med egenvektorerne til $\mathbf{B} \mathbf{B}^T$ som søjlevektorer og $\mathbf{\Sigma} \mathbf{\Sigma}^T$ er diagonalmatricen med egenværdier som elementer:

(63)

$$\mathbf{B}\mathbf{B}^T\mathbf{U} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T$$

De diagonale elementer (egenværdierne) er identiske med de diagonale elementer i $\mathbf{\Sigma}^T\mathbf{\Sigma}$. Hvis rangen af matricen $\mathbf{B}$ er r , er de første r diagonale elementer i $\mathbf{\Sigma}\mathbf{\Sigma}^T$ forskellige fra 0, mens alle øvrige elementer er 0. De diagonale elementer i matricen $\mathbf{\Sigma}$ kan skrives som $\sigma_i = \sqrt{\lambda_i}$, hvor λ_i er egenværdierne til de symmetriske matricer.

Ved en singular value decomposition af en vilkårlig matrice kan man udtrykke matricen ved et sæt af ortonormale vektorer. Egenværdierne (og dermed de singulære værdier) er et mål for hvor meget matricen "fylder" i retning af den tilsvarende egenvektor.

Hvis et datasæt er repræsenteret i 3 dimensioner, men alle punkter reelt ligger på en flade (dvs. i et todimensionelt under rum) vil en SVD af matricen kun give to singulære værdier der er forskellige fra 0, samt to egenvektorer der udspænder det todimensionelle under rum.

Bestemmelse af SVD for en vilkårlig matrice

For at kunne finde SVD for en matrice har vi først brug for at definere nogle størrelser ud fra den symmetriske matrice $\mathbf{A}$

Den inverse matrice til matricen $\mathbf{A}$ er pr. definition den matrice der giver enhedsmatricen ved multiplikation med $\mathbf{A}$:

(64)

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

Ved at benytte identiteten i ligning (45) får vi følgende udtryk for den inverse matrice:

(65)

$$\mathbf{A}^{-1} = \sum_{i=1}^k \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i^T$$

Vi kan kun finde en invers til en matrice der har egenværdier forskellige fra nul. Tilsvarende kan roden og den inverse rod af $\mathbf{A}$ defineres ud fra at $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$ og $\mathbf{A}^{-1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1}$

(66)

$$\mathbf{A}^{1/2} = \sum_{i=1}^k \sqrt{\lambda_i} \mathbf{e}_i \mathbf{e}_i^T$$

(67)

$$\mathbf{A}^{-1/2} = \sum_{i=1}^k \frac{1}{\sqrt{\lambda_i}} \mathbf{e}_i \mathbf{e}_i^T$$

Et tal-eksempel

Lad $\mathbf{B}$ være en ikke-symmetrisk matrice $\mathbf{B} = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 3 & -1 \end{bmatrix}$

$$\text{Det transponerede er } \mathbf{B}^T = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 3 & -1 \end{bmatrix}$$

(68)

$$\mathbf{B}^T \mathbf{B} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 3 \\ 1 & 3 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 2 \\ 4 & 10 & 0 \\ 2 & 0 & 10 \end{bmatrix}$$

(69)

$$\mathbf{B} \mathbf{B}^T = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$$

Egenverdierne findes ved at løse $\det(\mathbf{B} \mathbf{B}^T - \lambda \mathbf{I}) = 0$

Dermed finder man egenverdierne til at være $\lambda_1 = 12$ og $\lambda_2 = 10$

Egenvektorerne for $\mathbf{B} \mathbf{B}^T$ findes ved at løse:

(70)

$$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

For $\lambda_1 = 12$ fremkommer:

(71)

$$11x_1 + x_2 = 12x_1$$

$$x_1 + 11x_2 = 12x_2$$

Hvilket har løsningen:

(72)

$$e_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

For $\lambda_2 = 10$ fremkommer:

(73)

$$11x_1 + x_2 = 10x_1$$

$$x_1 + 11x_2 = 10x_2$$

Hvilket har løsningen:

(74)

$$e_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

Tilsvarende findes egenvektorerne for $\mathbf{B}^T \mathbf{B}$, for $\lambda_1 = 12$:

(75)

$$2x_1 + 4x_2 + 2x_3 = 12x_1$$

$$4x_1 + 10x_2 = 12x_2$$

$$2x_1 + 10x_3 = 12x_3$$

$x_1 = 1 \Rightarrow x_2 = 2$ og $x_3 = 1$. Den normerede vektor bliver:

(76)

$$e_1 = \begin{bmatrix} 1/\sqrt{6} \\ 2/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$$

For $\lambda_2 = 10$

$$2x_1 + 4x_2 + 2x_3 = 10x_1$$

$$4x_1 + 10x_2 = 10x_2$$

$$2x_1 + 10x_3 = 10x_3$$

$x_1 = 0$, $x_2 = -1 \Rightarrow x_3 = 2$ Den normerede vektor bliver:

(77)

$$e_2 = \begin{bmatrix} 0 \\ -1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$$

Den tredje egenværdi er nul, da matricen kun er af rank 2, $\lambda_3 = 0$

(78)

$$2x_1 + 4x_2 + 2x_3 = 0$$

$$4x_1 + 10x_2 = 0$$

$$2x_1 + 10x_3 = 0$$

$x_1 = 1 \Rightarrow x_2 = -\frac{4}{10}$ og $x_3 = -\frac{2}{10}$. Enhedsvektoren bliver:

(79)

$$e_3 = \begin{bmatrix} 5/\sqrt{30} \\ -2/\sqrt{30} \\ -1/\sqrt{30} \end{bmatrix}$$

Det bemærkes, at enhedsvektorerne som forventet er ortogonale.

Matricen for de singulære værdier bliver:

$$\mathbf{\Sigma} = \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

SVD faktoriseringen af $\mathbf{B}$ bliver:

$$\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\ 0 & -1/\sqrt{5} & 2/\sqrt{5} \\ 5/\sqrt{30} & -2/\sqrt{30} & 1/\sqrt{30} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 3 & -1 \end{bmatrix}$$

SVD faktorisering bliver hurtigt en uoverkommelig opgave at udføre manuelt ved større matricer. Men SVD faktorisering er en standardfunktion i mange statistikprogrammer og CAS-værktøjer, og kan med fordel løses med disse værktøjer.

Brug af SVD til reduktion af dimensioner

Ligesom ved den symmetriske matrice kan en vilkårlig matrice skrives som en sum, givet ud fra SVD faktoriseringen:

(80)

$$\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

Hvor k angiver rangen af matricen og dermed antallet af egenverdier forskellige fra 0. σ_i vil være aftagende, da de singulære værdier i $\mathbf{\Sigma}$ er arrangeret efter størrelse. Hver summation tilføjer en dimension til dataenes placering, givet ved de ortogonale egenvektorer.

Hvis man undlader at tage alle elementerne med i summationen i (80) vil det antal dimensioner datasættet udbreder sig over blive reduceret. For $k = 1$ kommer dataene til at ligge på en ret linje i retningen af den første principale vektor og ved $k = 2$ kommer de til at ligge på et plan, udspændt af første og anden principale vektor.

Den reducerede matrice kan skrives som:

(81)

$$\mathbf{B}_l = \sum_{i=1}^l \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

Hvor $l < k$. Her vil $\mathbf{B}_l$ blive den bedste rank l approksimation af $\mathbf{B}$. Rækkerne i $\mathbf{B}_l$ er rækkerne i $\mathbf{B}$, projiceret ned på det underrum, der er udspændt af de første l højre singulære vektorer.

Hvis $\mathbf{b}$ er en rækkevektor i $\mathbf{B}$ kan projektionen ind på rummet udspændt af $\mathbf{v}_i$ skrives som:

(82)

$$\sum_{i=1}^k (\mathbf{b} \cdot \mathbf{v}_i) \mathbf{v}_i^T$$

I tal-eksemplet ovenfor giver det første led:

(83)

$$\mathbf{B}^{\#}_1 = \sqrt{12} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix}$$

Hvor $\mathbf{B}^{\#}_i = (\mathbf{b} \cdot \mathbf{v}_i) \mathbf{v}_i^T$. Og andet led:

(84)

$$\mathbf{B}^{\#}_2 = \sqrt{10} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 0 & -1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 0 & -1 & 2 \\ 0 & 1 & -2 \end{bmatrix}$$

Summen af de to bidrag giver $\mathbf{B}$

Ved at gange $\mathbf{B}^{\#}_1$ med $\mathbf{U}$ får vi:

(85)

$$\mathbf{U}^T \mathbf{B}^{\#}_1 = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 2\sqrt{2} & \sqrt{2} \\ 0 & 0 & 0 \end{bmatrix}$$

Vi ser at der kun er et bidrag i $\mathbf{V}$'s første egenvektors retning, mens

(86)

$$\mathbf{U}^T \mathbf{B}^{\#}_2 = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 0 & -1 & 2 \\ 0 & 1 & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\sqrt{2} & 2\sqrt{2} \end{bmatrix}$$

Kun har et bidrag i retning af den anden egenvektor.

Næste kapitel handler om hvordan et korrespondanceanalyseplot konstrueres ud fra de principper der er beskrevet i dette afsnit.

Generering af korrespondanceanalyseplot

I dette afsnit vil vi se hvordan man kan generere et korrespondanceanalyseplot (KA-plot) vha. matricoperationer. Hver række eller søjle i en tabel vil blive repræsenteret ved et punkt i et KA-plot. KA-plottet er en repræsentation af dataene i to dimensioner, men selve profilerne vil i udgangspunktet sædvanligvis have flere end 2 dimensioner. KA-plottet vil indeholde en projektion af profilerne ind på det plan der bevarer den største variation af profilernes placering.

For at kunne indtegne profilerne i et todimensionelt rum (en flade), skal vi have defineret det rum der bevarer den største udbredelse af punkterne ved projektion ind på fladen. Alternativt kan man betragte det som, at man vil minimere den fejl der sker ved at foretage projektionen, dvs. minimere kvadratet på afstanden fra projektionen af profilen, til profilen selv.

Illustration af data i to og tre dimensioner

Inden vi starter på den matematiske formalisme kan det være illustrativt at se hvordan data i få dimensioner kan illustreres, for at vise princippet i et KA-plot.

Eksempel med data i to dimensioner

Lad os betragte en simpel tabel, der viser sammenhængen mellem variablene køn og rygning. Lad os antage, at vi har spurgt 100 personer om deres køn og om hvorvidt de er rygere. Resultatet fremgår af tabellen nedenfor. Vi vil i dette simple tilfælde forvente, at profilerne kan illustreres i en enkel dimension, dvs. som punkter på en ret linje.

	Ryger	Ikke ryger
Mand	17	13
Kvinde	30	40

Tabel 14 Sammenhæng mellem køn og rygning for 100 udvalgte personer. (Fiktivt eksempel)

Ud fra Tabel 14 er det simpelt at opstille en tabel over de relative rækkefordelinger, samt massen af række- og søjleprofiler:

	Ryger	Ikke ryger	r
Mand (b_1)	0,566667	0,433333	0,3
Kvinde (b_2)	0,428571	0,571429	0,7
c	0,47	0,53	

Tabel 15 Relative rækkeprofiler, samt massen af række- og søjleprofiler

Rækken c angiver massen af de enkelte søjler, og angiver samtidig den gennemsnitlige fordeling af rygere og ikke rygere i undersøgelsen. r angiver den gennemsnitlige fordeling af mænd og kvinder.

For at illustrere dataene i et KA-plot, vil vi benytte X^2 -afstandene som koordinater. Derudover ønsker vi at den gennemsnitlige fordeling skal ligge i origo.

X^2 -afstandene er afstandene vægtet med $\frac{1}{\sqrt{c_i}}$. Dvs. at koordinaterne kan beregnes som: $\frac{(b_{ij}-c_i)}{\sqrt{c_i}}$

X^2 -afstand	Ryger	Ikke-ryger	Afstand fra gennemsnit	Inerti
Mand	0,141003	-0,13278	0,193682	0,011254
Kvinde	-0,06043	0,056907	0,083007	0,004823

Tabel 16 X^2 koordinaterne til profilerne

Profilernes afstande fra gennemsnittet beregnes ved Pythagoras:

(87)

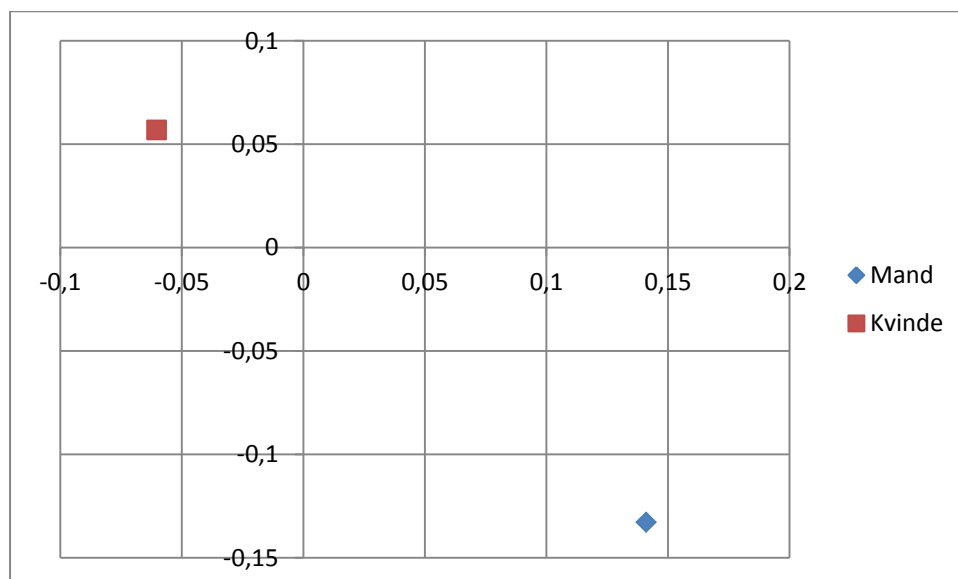
$$X^2_{afstand} = \sqrt{\sum_j \frac{(b_{ij} - c_i)^2}{c_i}}$$

Inertien er kvadratet på X^2 -afstanden vægtet med massen af profilen:

(88)

$$\phi_i^2 = r_i \sum_j \frac{(b_{ij} - c_i)^2}{c_i}$$

De to rækkeprofiler kan nu indtegnes i et koordinatsystem, med de angivne koordinater, se Figur 9. Den samlede inerti for de to profiler er 0,016077

Figur 9 Rækkeprofiler indtegnet med X^2 -afstande som koordinater

Ud over at indtegne rækkeprofilerne kan vi indtegne positionen af søjlekategorierne. Vi kan forestille os, at vi tilføjer yderligere to profiler. I den ene profil er der en enkelt ryger og i den anden profil er der en enkelt ikke-ryger. Disse to nye profiler indgår i diagrammet uden vægt, og bidrager derfor ikke til den samlede inerti.

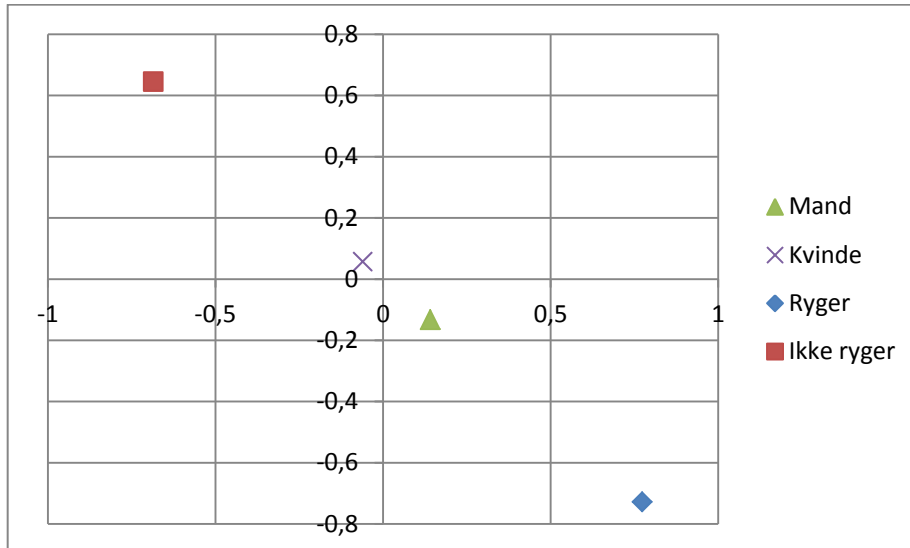
X^2 -afstanden for kategorierne for ryger og ikke-ryger er beregnet i Tabel 17.

	Ryger	Ikke-ryger	Afstand fra gennemsnit
--	-------	------------	------------------------

Ryger (profil)	$\frac{1-0,47}{\sqrt{0,47}}=0,773084$	$\frac{0-0,47}{\sqrt{0,47}} = -0,72801$	1,061913
Ikke-ryger (profil)	$\frac{0-0,53}{\sqrt{0,53}} = -0,68557$	$\frac{1-0,53}{\sqrt{0,53}} = 0,645595$	0,941697

Tabel 17 Koordinater for søjlekategorier.

Hvis profilerne for ryger og ikke-ryger tilføjes plottet får vi Figur 10.



Figur 10 Plot over rækkeprofiler inkl. kategorier.

Bestemmelse af kategorier ud fra maksimal inerti

En anden måde at bestemme positionen af punkterne for kategorierne, er ud fra den maksimale inerti for fordelingen. Den maksimale inerti i en fordeling i to kategorier vil være 1. Den maksimale inerti vil fremkomme hvis alle profiler kun havde bidrag i én af kategorierne. I så fald ville positionen af rækkeprofiler og søjlekategorier være sammenfaldende. Inertibidraget fra hver søjlekategori vil være lig 1 minus vægten af søjlekategorien. I dette eksempel vil:

(89)

$$\phi_{ryger}^2 = 1 - c_1 = c_2$$

$$\phi_{ikke-ryger}^2 = 1 - c_2 = c_1$$

Da vi nu kender inertien og massen af hver kategori kan vi bestemme X^2 -afstanden:

(90)

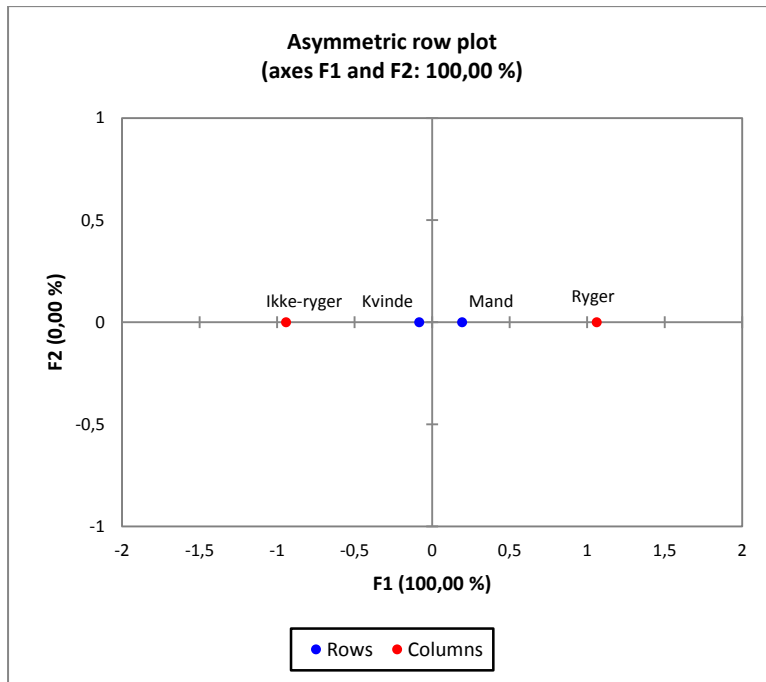
$$X_{afstand}^2(\text{Ryger}) = \sqrt{\frac{c_2}{c_1}} = 1,061913$$

$$X_{afstand}^2(\text{Ikke-ryger}) = \sqrt{\frac{c_1}{c_2}} = 0,941697$$

Hvilket giver de samme afstande som vi fandt tidligere.

I tilfældet på Figur 10, er profilernes udbredelse begrænset til én dimension. I et normalt KA-plot vil man vælge at benytte den første principale akse som 1. akse. Da der ikke er nogen

udbredelse langs den anden principale akse, vil plottet kunne præsenteres på en enkelt akse. Afstandene på KA-plottet er de samme som afstandene i Figur 10.



Figur 11 KA-plot af rækkeprofilerne i Tabel 14. Genereret via XLSTAT.

Søjleprofiler

Ud over at indtegne rækkeprofilerne fra Tabel 14 er det muligt at indtegne søjleprofilerne. De relative søjlefordelinger er da:

	Ryger	Ikke-ryger	r
Mand	0,36170	0,24528	0,3
Kvinde	0,63829	0,75471	0,7
c	0,47	0,53	

Tabel 18 Relativ fordeling af søjleprofiler

Det bemærkes at det nu er summen af søjlerne der skal give én. Den gennemsnitlige fordeling er nu angivet i massen af rækkerne r . X^2 -afstandene skal nu beregnes i forhold til massen af rækkerne.

	Ryger	Ikke-ryger
Mand	0,112652157	-0,09989908
Kvinde	-0,07374815	0,065399301
Afstand fra gennemsnit	0,134645081	0,119402242

Tabel 19 X^2 -koordinater for søjleprofilerne

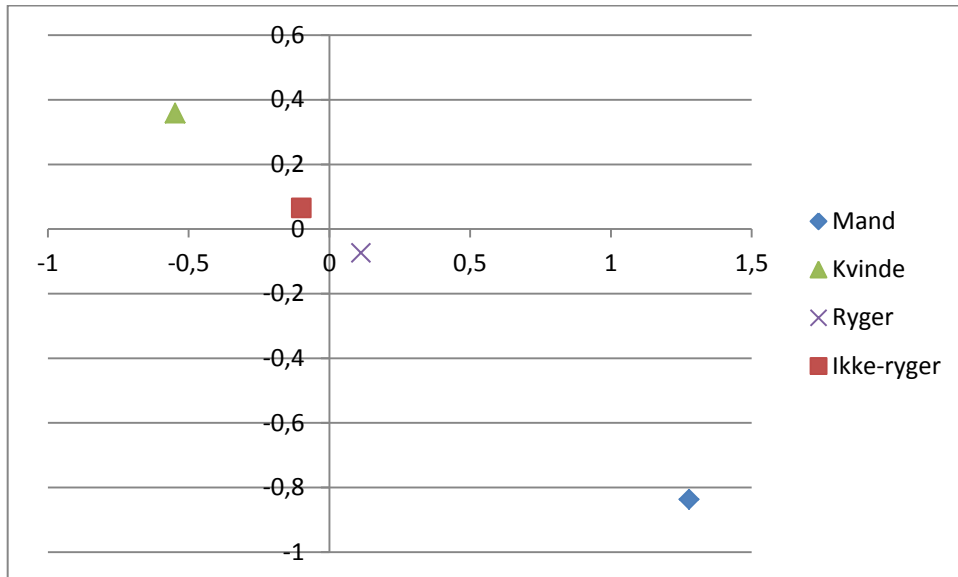
Koordinaterne for rækkekategoriene kan igen bestemmes ved at tilføje to søjleprofiler med hhv. 100 % kvinder og 100 % mænd. Disse profiler skal ikke have nogen masse.

X^2 koordinater	Mænd(profil)	Kvinder(profil)
Mand	1,278019	-0,54772

Kvinde	-0,83666	0,358569
Afstand fra gennemsnit	1,527525	0,654654

Tabel 20 X^2 -afstande til række-kategorierne

Ved at indtegne profilerne i et koordinatsystem får man et plot meget lig Figur 10. Men denne gang med søjleprofilerne placeret indenfor det én-dimensionale rum der er udspændt af række-kategorierne for mand/kvinde.



Figur 12 Plot over søjleprofiler inkl. række-kategorier

Igen er det muligt at bestemme afstanden til række-kategorierne ved forholdet mellem inertierne for de to kategorier. Den samlede maksimale inerti er 1. Inertibidraget for kategorien "Mænd" er 1 minus vægten af kategorien:

$$\phi_{mænd}^2 = 1 - r_1 = r_2$$

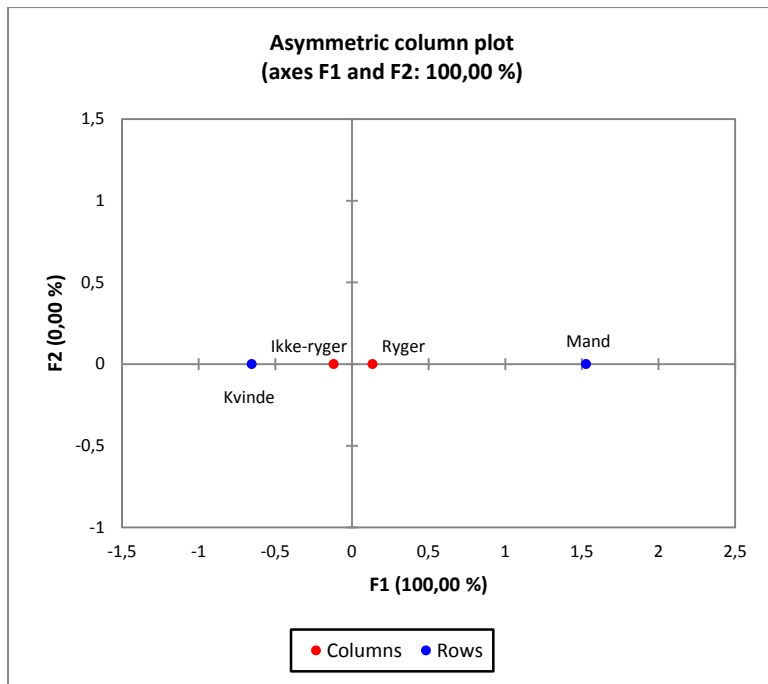
Afstandene fra gennemsnittet bliver:

$$X^2 - afstand (Mænd) = \sqrt{\frac{r_2}{r_1}} = 1,527$$

Tilsvarende bliver afstanden fra gennemsnittet til kategorien "Kvinder":

$$X^2 - afstand (Kvinder) = \sqrt{\frac{r_1}{r_2}} = 0,654$$

Indtegnet som et traditionelt KA-plot bliver plottet som vist på Figur 13.



Figur 13 KA-plot over søjleprofiler

Forholdet mellem afstanden til et punkt angivet som en profil og til et punkt angivet som en kategori er netop givet ved en faktor på kvadratroden af inertien for den pågældende akse.

Eksempel med data i 3 dimensioner

I dette eksempel vil vi betragte et datasæt i 3 dimensioner. Lad os antage, at vi har spurgt 477 gymnasieelever om, hvor lang tid de i gennemsnit bruger på hjemmearbejde om dagen. Svarene er angivet i Tabel 21.

D (d_{ij})	0-½ t	½-1 t	> 1t	i alt
1.g	70	68	32	170
2.g	55	57	46	158
3.g	58	49	42	149
i alt	183	174	120	477

Tabel 21 Undersøgelse af gymnasieelevers tid brugt på hjemmearbejde.

De relative fordelinger er vist i Tabel 22

$b_{ij} = \frac{d_{ij}}{d_{i+}}$	0-½ t	½-1 t	> 1t	r
1.g	0,411765	0,4	0,188235	0,356394
2.g	0,348101	0,360759	0,291139	0,331237
3.g	0,389262	0,328859	0,281879	0,312369
c	0,383648	0,36478	0,251572	1

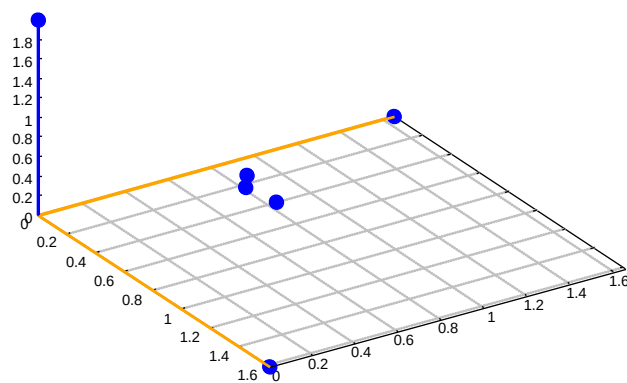
Tabel 22 Relativ fordeling af udfaldene i Tabel 21

Den gennemsnitlige fordeling er givet ved c . X^2 -afstandene kan igen beregnes ud fra $\frac{b_{ij}-c_j}{\sqrt{c_j}}$.

X^2 -koordinaterne	$0-\frac{1}{2}t$	$\frac{1}{2}-1t$	$>1t$	Inerti
1.g	0,045394	0,058314	-0,12628	0,007629
2.g	-0,05739	-0,00666	0,078886	0,003167
3.g	0,009064	-0,05947	0,060424	0,002271

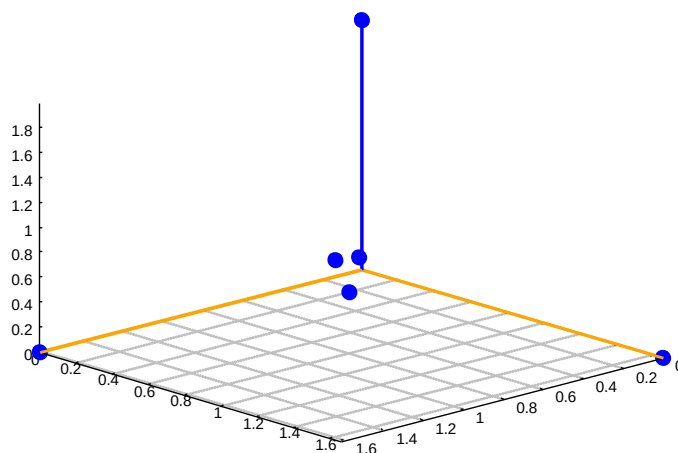
Tabel 23 X^2 afstand fra gennemsnit af rækkeprofiler

Det vil være muligt at indtegne profilerne i et tredimensionelt plot, hvor de tre akser angiver X^2 -afstandene langs de tre kategorier. Punkterne vil ligge på et plan. Dette er indlysende ved tre profiler. Hvis der havde været flere profiler ville de ligge på det samme plan. Se Figur 14.



Figur 14 3D plot af rækkeprofilerne

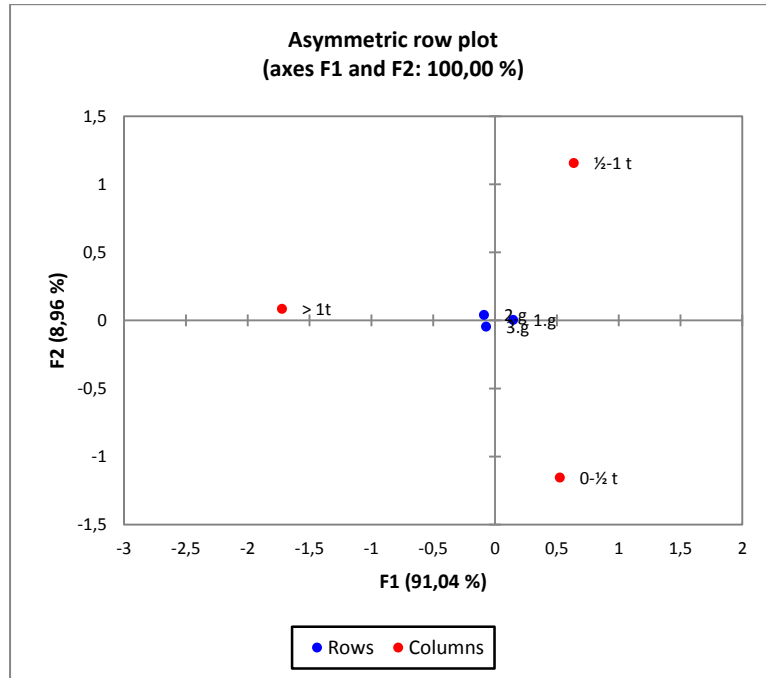
Da det kun er muligt at illustrere et plot i to dimensioner, skal plottet i tre dimensioner drejes, så planet som punkterne ligger i, er parallelt med projektiionsplanet. Dette er forsøgt illustreret på Figur 15.



Figur 15 3D plot roteret således, at spredningen af profilerne er maksimeret.

Det er muligt at beregne den samlede inerti for fordelingen til 0,013067. Denne inerti kan opdeles i to bidrag; et for hver af de principale akser der udspænder fordelingen. Inertien langs den 1. principale akse kan beregnes til 0,0119 og langs den 2. principale akse til 0,00117.

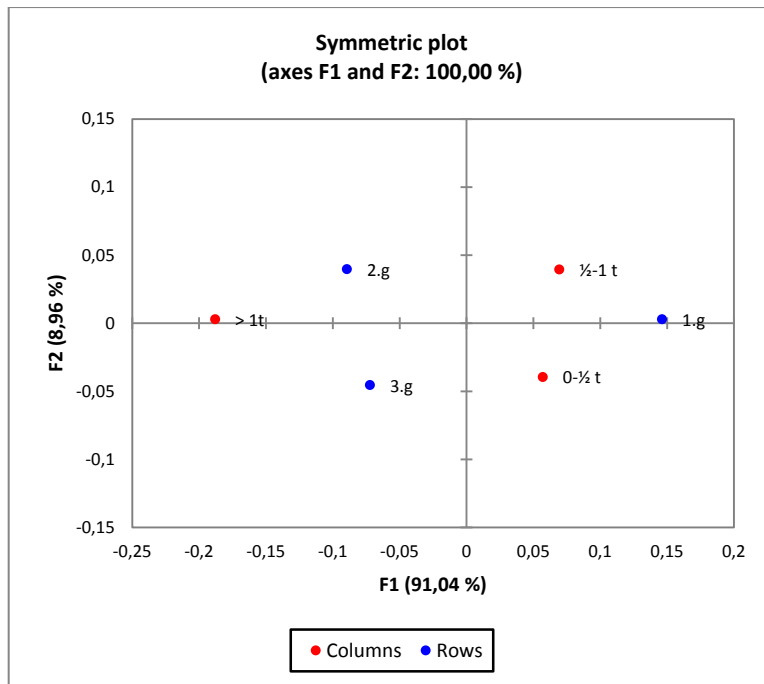
Et KA-plot over fordelingen er vist på Figur 16



Figur 16 Asymmetrisk KA-plot over rækkeprofilerne.

I plottet på Figur 16 er også søjlekategorierne indtegnet. Det ses, at plottet er identisk med plottet på Figur 15, blot roteret 90 grader, således at den 1. principale akse ligger vandret.

Da inertien i fordelingen (0,013067) er relativt lille, sammenlignet med den maksimale inerti på 2, er rækkeprofilerne klumpet sammen tæt ved den gennemsnitlige profil. For at gøre det lettere at tolke plottet kan både række- og søjleprofilerne indtegnes i samme diagram. Som vist i det foregående eksempel er forholdet mellem rækkekoordinater og kategorikoordinater angivet ved kvadratroden af inertien for den pågældende akse. På Figur 17 er vist et symmetrisk KA-plot, hvor både række- og søjleprofiler er indtegnet.



Figur 17 Symmetrisk KA-plot, hvor både række- og søjleprofiler er indtegnet.

På et symmetrisk plot er det lettere at se de indbyrdes placeringer af rækkeprofilerne og af søjleprofilerne. Det vil ikke give mening at tolke på afstanden mellem profiler for række kategorier og profiler for søjle kategorier. Men retningerne er meningsgivende. F.eks. ses det, at der er relativt flere elever fra 2. og 3.g der bruger over 1 time dagligt på lektielæsning.

Algoritme for KA-plot

Definitioner af matricer

Udgangspunktet for en korrespondanceanalyse er en matrice med data, hvor rækkerne og søjlerne angiver forskellige kategorier. Elementerne i matricen er positive værdier. Det kunne f.eks. være antallet af respondenter i en spørgeskemaundersøgelse.

Matricen med de observerede data kaldes $\mathbf{D}$ og er en $I \times J$ matrice.

Ved at dele med det samlede antal observationer får man en matrice over de relative fordelinger:

(91)

$$\mathbf{P} = \frac{1}{d_{++}} \mathbf{D}$$

der igen er en $I \times J$ matrice.

De marginale relative fordelinger, der er angivet ved massen af rækkerne og søjlerne, kan bestemmes som summen af de relative fordelinger for hhv. søjlerne og rækkerne:

(92)

$$r_i = \sum_{j=1}^J p_{ij}$$

(93)

$$c_j = \sum_{i=1}^I p_{ij}$$

Som vektorer kan de beregnes som: $\mathbf{r} = \mathbf{P}\mathbf{1}$ og $\mathbf{c} = \mathbf{P}^T\mathbf{1}$, hvor $\mathbf{1}$ benyttes som betegnelse for en vektor af passende længde, bestående af 1-taller. $\mathbf{c}$ vil være den gennemsnitlige rækkeprofil og $\mathbf{r}$ vil være den gennemsnitlige søjleprofil. Disse vektorer vil indeholde de marginale fordelinger og betegnes her som masserne af hhv. søjlerne og rækkerne. Med andre ord er fordelingen af massen af rækkeprofilerne givet ved $\mathbf{c}$ og fordelingen af massen af søjleprofilerne er givet ved $\mathbf{r}$.

Massen for rækkerne og søjlerne kan omskrives til diagonalmatricer, hvor $\mathbf{D}_r$ er en $I \times I$ matrice og $\mathbf{D}_c$ er en $J \times J$ matrice:

(94)

$$d_{r,ij} = \begin{cases} r_i & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

(95)

$$d_{c,ij} = \begin{cases} c_j & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

Det bemærkes, at summen af elementer i $\mathbf{P}$, $\mathbf{D}_r$ og $\mathbf{D}_c$ giver 1. Ved at gange med det samlede antal observationer d_{++} får man igen de oprindelige absolutte fordelinger.

Procedure for beregning af koordinater i et KA-plot

Først defineres en matrice $\mathbf{S}$ for de standardiserede residualer¹¹:

(96)

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-1/2}$$

$\mathbf{r}\mathbf{c}^T$ vil være en $I \times J$ matrice over produktet af de gennemsnitlige række- og søjlefordelinger. Dette svarer til fordelingen ved uafhængighed mellem række- og søjlekategorierne. $(\mathbf{P} - \mathbf{r}\mathbf{c}^T)$ angiver dermed forskellen mellem de observerede og de gennemsnitlige fordelinger.

$\mathbf{D}_r^{-1/2}$ er en $I \times I$ diagonalmatrice med elementerne $\frac{1}{\sqrt{r_i}}$ i diagonalen og $\mathbf{D}_c^{-1/2}$ er en $J \times J$ diagonalmatrice med elementerne $\frac{1}{\sqrt{c_j}}$ i diagonalen.

De enkelte elementer i $\mathbf{S}$ er:

(97)

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i} \sqrt{c_j}}$$

s_{ij}^2 angiver inertibidraget for dette element i tabellen. Summen af kvadratet på elementerne i $\mathbf{S}$ angiver dermed den samlede inerti ϕ^2 :

(98)

$$\phi^2 = \sum s_{ij}^2 = \sum \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

$\mathbf{S}$ kan omskrives ved at lave en SVD faktorisering:

(99)

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Hvor søjlerne i $\mathbf{U}$ indeholder de normerede egenvektorer til $\mathbf{S}\mathbf{S}^T$, dvs. de principale vektorer for rækkeprofilerne. Søjlerne i $\mathbf{V}$ indeholder de normerede egenvektorer til $\mathbf{S}^T\mathbf{S}$, dvs. de principale vektorer for søjleprofilerne.

(100)

$$\mathbf{S}\mathbf{S}^T = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T$$

(101)

$$\mathbf{S}^T\mathbf{S} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$$

¹¹ Greenacre s202

Σ indeholder en diagonalmatrice med de singulære værdier σ_{ii} , dvs. kvadratroden af egenverdierne i aftagende orden, som de diagonale elementer. Alle øvrige elementer i Σ er 0. Egenverdierne angiver også inertibidraget i den givne dimension. Dvs. inertien $\phi^2 = \sum \sigma_{ii}^2$.

Inertibidraget for hver af de principale akser er de samme for såvel række- som søjleprofilerne.

S kan skrives som summen af bidrag fra hver dimension:

(102)

$$S = U \Sigma V^T = \sum_{i=1}^k \sigma_{ii} \mathbf{u}_i \mathbf{v}_i^T$$

Første led i summationen ovenfor angiver bidraget til den samlede inert i den første principale retning, og andet led vil angive bidraget i den anden principale retning, osv. For de led der er højere end rangen k af matricen S vil der ikke være noget bidrag, da den singulære værdi herfor vil være 0.

Koordinater for rækkeprofilerne

For at afstandene mellem profilerne skal være meningsfuld, skal det være X^2 -afstanden der skal benyttes som koordinater. Fra tidligere har vi at sammenhængen mellem den samlede inert og X^2 -afstanden er givet ved:

(103)

$$\phi^2 = \sum r_i \cdot (X_i^2 - afstand)^2$$

For inertibidraget langs den 1. principale akse gælder:

(104)

$$\sigma_{11}^2 = \sum r_i \cdot (X_{i1, koordinat}^2)^2$$

Tilsvarende for de øvrige principale akser. Skrevet på matriceform bliver koordinaterne for rækkeprofilerne givet som søjlevektorer i matricen F :

(105)

$$F = D_r^{-1/2} U \Sigma$$

Hvor rækkeindeks refererer til profilen og søjleindeks refererer til koordinaterne. De enkelte elementer i matricen kan skrives som:

(106)

$$f_{ij} = \frac{1}{\sqrt{r_i}} u_{ji} \sigma_{jj}$$

Koordinaterne for den første rækkeprofil bliver:

(107)

$$\mathbf{f}_{1j} = \left(\frac{\sigma_{11}}{\sqrt{r_1}} u_{11}, \frac{\sigma_{22}}{\sqrt{r_2}} u_{21} \right)$$

Et KA-plot illustrerer projektionen af rækkeprofilerne ind på det plan der er udspændt af de ortogonale vektorer $\mathbf{u}_1$ og $\mathbf{u}_2$, der er hhv. den første og den anden principale vektor.

Koordinater for søjlekategorier

Ud over koordinaterne for rækkeprofilerne skal også koordinaterne for søjlekategorierne bestemmes. Det er nødvendigt at indtegne søjlekategorierne for at kunne tolke KA-plottet. Positionen af søjlekategorierne angiver de ekstreme udfald. Hvis man forestiller sig, at man har en rækkeprofil med, hvor 100 % af observationerne ligger i én bestemt kategori, ville positionen af denne profil netop være positionen af denne søjlekategori i KA-plottet. Den maksimale inerti for hver principale akse vil være 1. Dvs:

(108)

$$\phi_{j,max}^2 = 1 = \sum c_j (X^2 - afstand_j)^2$$

X^2 -afstanden til den j 'te kategori bliver $\frac{1}{\sqrt{c_j}}$

I det multidimensionale rum vil søjlekategorierne ligge som punkter langs akserne i rummet, med X^2 -afstanden $\frac{1}{\sqrt{c_j}}$ fra origo. Opstillet på matriceform bliver vektorerne til søjlekategorierne:

(109)

$$\begin{bmatrix} \frac{1}{\sqrt{c_1}} & 0 & 0 & \dots \\ 0 & \frac{1}{\sqrt{c_2}} & 0 & \dots \\ 0 & 0 & \frac{1}{\sqrt{c_3}} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \mathbf{D}_c^{-1/2}$$

Det bemærkes, at denne matrice er identisk med $\mathbf{D}_c^{-1/2}$, der blev defineret tidligere. For at finde positionen af søjlekategorierne i det todimensionelle KA-plot skal vektorerne til søjlekategorierne projiceres ind på de to vektorer ($\mathbf{v}_1; \mathbf{v}_2$) der er de to første principale vektorer for søjleprofilerne.

For at bestemme koordinaterne for søjlekategorierne skal de projiceres ind på de principale akser. Koordinaterne er givet som:

(110)

$$(\mathbf{V}^T \mathbf{V})^{-1} \mathbf{V}^T \mathbf{D}_c^{-1/2} = \mathbf{V}^T \mathbf{D}_c^{-1/2}$$

Det bemærkes, at $\mathbf{V}^T \mathbf{V} = \mathbf{I}$, da $\mathbf{V}$ består af ortonormale enhedsvektorer. Ved at transponere udtrykket fås koordinaterne som rækkevektorer:

(111)

$$\mathbf{\Gamma} = \mathbf{D}_c^{-1/2} \mathbf{V}$$

Herved kommer første indeks til at angive kategorien og andet indeks til at angive koordinaten. De enkelte elementer i matricen kan skrives som:

(112)

$$\gamma_{ij} = \frac{1}{\sqrt{c_i}} v_{ij}$$

Den første kategori vil få koordinaterne $(\frac{v_{11}}{\sqrt{c_1}}; \frac{v_{12}}{\sqrt{c_1}})$ i KA-plottet.

Positionen for søjlekategorierne er angivet ved standardkoordinaterne. Retningen af vektorerne for søjlekategorierne angiver retningen for denne dimensions projektion ind på KA-plottet.

Søjleplot

Ud over at indtegne rækkeprofilerne i et KA-plot, med søjlekategorierne indtegnet som yderpunkter, er det muligt at lave den omvendte illustration. Hvor man indtegner søjleprofilerne og har række kategorierne som yderpunkter. Dette svarer principielt til at bytte om på rækker og søjler i den oprindelige tabel over observationerne.

Beregningen af koordinaterne for søjleprofilerne kan findes som:

(113)

$$\mathbf{G} = \mathbf{D}_c^{-1/2} \mathbf{V} \mathbf{\Sigma}$$

Mens koordinaterne for række kategorierne bestemmes ud fra:

(114)

$$\mathbf{\Phi} = \mathbf{D}_r^{-1/2} \mathbf{U}$$

Ved at sammenligne med udtrykkene fra rækkeprofilerne ses det at:

(115)

$$\mathbf{G} = \mathbf{D}_c^{-1/2} \mathbf{V} \mathbf{\Sigma} = \mathbf{\Gamma} \mathbf{\Sigma}$$

$$\mathbf{F} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Sigma} = \mathbf{\Phi} \mathbf{\Sigma}$$

Det vil sige, at omregningen fra standardkoordinater til principale koordinater er givet ved en skalering med kvadratroden af inertien for den givne akse.

Det ses, at den vægtede sum af kvadraterne af de principale koordinater er inertien for den givne dimension:

(116)

$$\mathbf{F} \mathbf{D}_r \mathbf{F}^T = \mathbf{\Phi} \mathbf{\Sigma} \mathbf{D}_r \mathbf{\Sigma}^T \mathbf{\Phi}^T = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Sigma} \mathbf{D}_r \mathbf{\Sigma}^T \mathbf{U}^T \mathbf{D}_r^{-1/2} = \mathbf{\Sigma} \mathbf{\Sigma}^T = \mathbf{\Lambda}$$

Hvor Λ indeholder en diagonalmatrice, med egenværdierne som elementer. Tilsvarende får vi for søjleprofilerne:

(117)

$$\mathbf{G}\mathbf{D}_c\mathbf{G}^T = \Lambda$$

Den vægtede sum for standardkoordinaterne giver 1.

(118)

$$\Phi\mathbf{D}_r\Phi^T = \mathbf{I}$$

$$\Gamma\mathbf{D}_c\Gamma = \mathbf{I}$$

Det er muligt at genskabe den oprindelige relative fordeling ud fra rekonstruktionsformlen:

$$p_{ij} = r_i c_j \left(1 + \sum_{k=1}^K \sigma_{kk} \varphi_{ik} \gamma_{jk} \right)$$

Hvor φ_{ik} er standardkoordinaterne for rækkerne og γ_{jk} er standardkoordinaterne for søjlerne.

Et tal-eksempel

I sidste kapitel kiggede vi på et eksempel med data i 6 række kategorier og 5 søjle kategorier. De relative fordelinger $\mathbf{P}$ kan skrives som:

$\mathbf{P}$	y1	y2	y3	y4	y5	$\mathbf{r}$
x1	0,048	0,045	0,019	0,030	0,015	0,156
x2	0,041	0,059	0,030	0,022	0,007	0,160
x3	0,048	0,041	0,045	0,026	0,030	0,190
x4	0,019	0,026	0,059	0,045	0,015	0,164
x5	0,022	0,030	0,033	0,056	0,019	0,160
x6	0,007	0,037	0,033	0,063	0,030	0,171
$\mathbf{c}$	0,186	0,238	0,219	0,242	0,115	1,000

Tabel 24 Relativ fordeling P

Hvor den sidste række og søjle angiver masserne.

Matricen over de standardiserede residualer $\mathbf{S}$ beregnes ud fra formlen (97):

$\mathbf{S}$	y1	y2	y3	y4	y5
x1	0,113	0,039	-0,085	-0,041	-0,023

x2	0,065	0,110	-0,028	-0,083	-0,081
x3	0,070	-0,020	0,015	-0,092	0,053
x4	-0,068	-0,065	0,125	0,026	-0,029
x5	-0,043	-0,043	-0,009	0,087	0,001
x6	-0,137	-0,017	-0,021	0,108	0,071

Tabel 25 S matrice over dataene i Tabel 24

Kvadratet på elementerne i Tabel 25 angiver inertien for hvert element i datamatricen. Den samlede inertie er $\phi^2 = \sum s_{ij}^2 = 0,144$

Der kan laves en SVD faktorisering af matricen $\mathbf{S}$. Dette er dog ikke en simpel operation at udføre manuelt, men kan f.eks. udføres i statistikprogrammet R.

De singulære værdier giver:

$\sigma_1 = 0,311$	$\sigma_2 = 0,161$	$\sigma_3 = 0,127$	$\sigma_4 = 0,0784$	$\sigma_5 = 0$
--------------------	--------------------	--------------------	---------------------	----------------

Tabel 26 Singulære værdier

Tilsvarende giver inertierne, der er kvadratet på de singulære værdier, for de 5 dimensioner:

$\phi_1^2 = 0,097$	$\phi_2^2 = 0,026$	$\phi_3^2 = 0,016$	$\phi_4^2 = 0,006$	$\phi_5^2 = 0$
--------------------	--------------------	--------------------	--------------------	----------------

Tabel 27 Inertierne for hver dimension

Det bemærkes at inertien i den 5. dimension er nul, hvilket er forventet da søjlerne i $\mathbf{S}$ ikke er uafhængige, og profilerne derfor ligger i et 4-dimensionelt rum.

Ud fra inertierne ses det at 67,19 % af den samlede inertie er forklaret af den første akse, mens 17,96 % er forklaret af den anden akse. I alt er det muligt at afbilde 85,15 % af den samlede inertie i et todimensionelt plot. Dette kan sammenlignes med de blot 58,12 % det var muligt at afbilde ved blot at benytte udbredelsen i retningerne af y1 og t4, der var de to oprindelige akser, med størst variation, som vist i Figur 7.

De ortonormale vektorer i $\mathbf{U}$ er:

-0,4463	-0,3285	-0,1281	0,4680	0,4073
-0,4930	-0,0326	0,6296	-0,3772	-0,3471
-0,2458	0,3302	-0,7103	-0,2739	-0,3657
0,3469	0,7051	0,2755	0,1552	0,2097
0,2921	-0,2067	0,0553	0,5544	-0,7237
0,5401	-0,4920	-0,0612	-0,4818	0,1101

Og de ortonormale vektorer i $\mathbf{V}$ er:

-0.6724	0.0743	-0.2900	0.5218	-0.4311
-0.3566	-0.3203	0.4777	-0.5514	-0.4878
0.2489	0.8291	0.1419	-0.1054	-0.4683
0.5600	-0.4175	0.1934	0.4827	-0.4916

| 0.2121 | -0.1735 | -0.7938 | -0.4237 | -0.3395 |

Koordinaterne for rækkeprofilerne er givet ved de principale koordinater:

F	F1	F2	F3	F4
x1	-0,352	0,134	0,040	-0,093
x2	-0,384	0,013	-0,195	0,074
x3	-0,176	-0,122	0,202	0,049
x4	0,267	-0,281	-0,084	-0,030
x5	0,228	0,083	-0,017	-0,109
x6	0,407	0,192	0,018	0,091

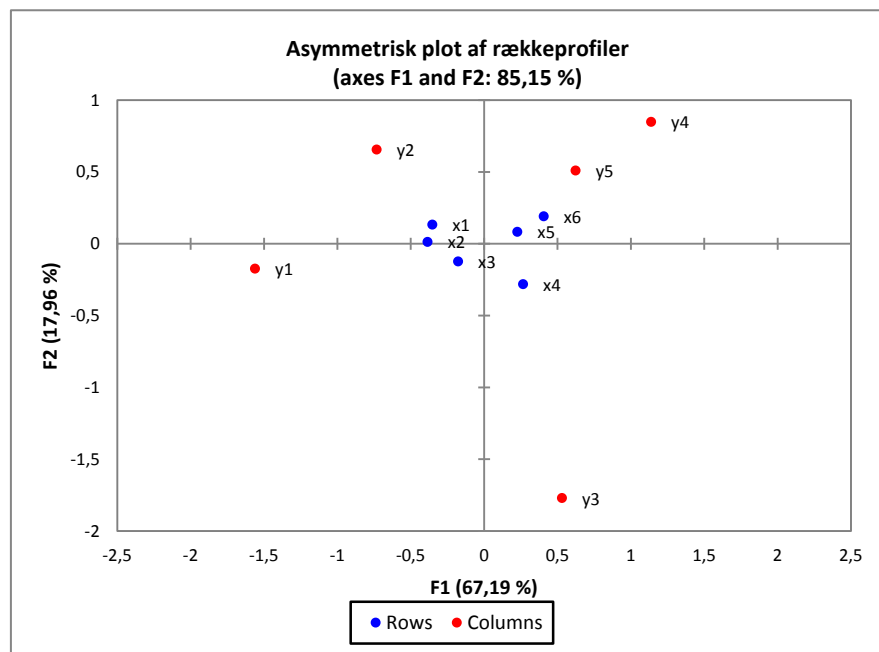
Beregnet ud fra: $F = D_r^{-1/2} U \Sigma$

Mens koordinaterne for søjlekategorierne er givet ved standardkoordinaterne:

Γ	F1	F2	F3	F4
y1	-1,560	-0,172	0,673	-1,210
y2	-0,731	0,657	-0,979	1,131
y3	0,532	-1,770	-0,303	0,225
y4	1,139	0,849	-0,394	-0,982
y5	0,625	0,511	2,338	1,248

beregnet ud fra: $\Gamma = D_c^{-1/2} V$.

Ud over at indtegne punkterne for rækkeprofilerne indtegner man også punkterne for



Figur 18 Asymmetriske plot af rækkeprofiler

søjlekategorierne. Jo mindre den samlede inerti er i en fordeling, des mere vil rækkeprofilerne klumpe sig sammen i midten af plottet. Det vil sige at de enkelte profiler vil ligge tæt op af den gennemsnitlige fordeling.

I det asymmetriske KA-plot i Figur 18 er den totale inerti 0,144. Hver rækkeprofil bidrager med massen af profilen gange kvadratet på afstanden fra origo. Søjleprofilerne er indtegnet med deres standardkoordinater. Det vil sige yderpunktet for hver søjle. Hvis der f.eks. var yderligere en række, hvor alle observationer lå i kategori y4, ville positionen af denne rækkeprofil være sammenfaldende med y4.

Eksempel med søjleprofiler

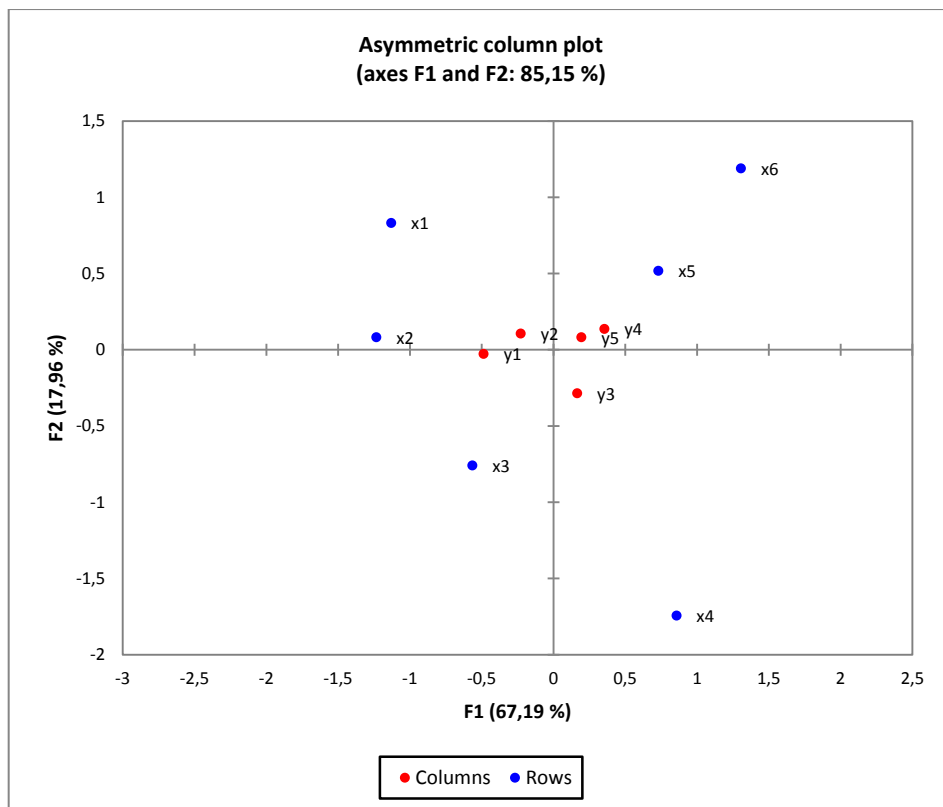
For at tegne et KA-plot over søjleprofilerne skal vi indtegne de principale koordinater for søjleprofilerne og standardkoordinaterne for rækkeprofilerne.

G	F1	F2	F3	F4
y1	-0,486	-0,028	0,083	-0,095
y2	-0,228	0,106	-0,121	0,089
y3	0,166	-0,285	-0,037	0,018
y4	0,355	0,137	-0,049	-0,077
y5	0,195	0,082	0,289	0,098

Φ	F1	F2	F3	F4
x1	-1,129	0,831	0,324	-1,184
x2	-1,233	0,082	-1,575	0,944
x3	-0,564	-0,758	1,631	0,629
x4	0,858	-1,743	-0,681	-0,384
x5	0,731	0,517	-0,138	-1,387
x6	1,306	1,190	0,148	1,165

KA-plottet er vist på Figur 19. Ved at sammenligne med KA-plottet i Figur 18 ses det, at de indbyrdes placeringer af hhv. række- og søjleprofilerne er uændret. Men, der er sket en skalering af akserne for positionen af kategorierne for at give positionen af profilerne. Denne skalering er en følge af sammenhængen i (115).

Afstanden mellem de enkelte profiler er χ^2 -afstanden. Det betyder, at to punkter med stor indbyrdes afstand også har stor forskel i fordelingen. Omvendt vil to profiler, der er sammenfaldende i plottet, ikke nødvendigvis have identisk fordeling. Her er det vigtigt at huske, at plottet netop er en projektion af en fordeling i flere dimensioner. To sammenfaldende profiler kan derfor have en vis indbyrdes afstand i en af de dimensioner der ikke er vist på plottet. For at få sikkerhed for om to profiler er sammenfaldende, også i den oprindelige tabel, er det derfor nødvendigt at bestemme den indbyrdes afstand ud fra samtlige koordinater.



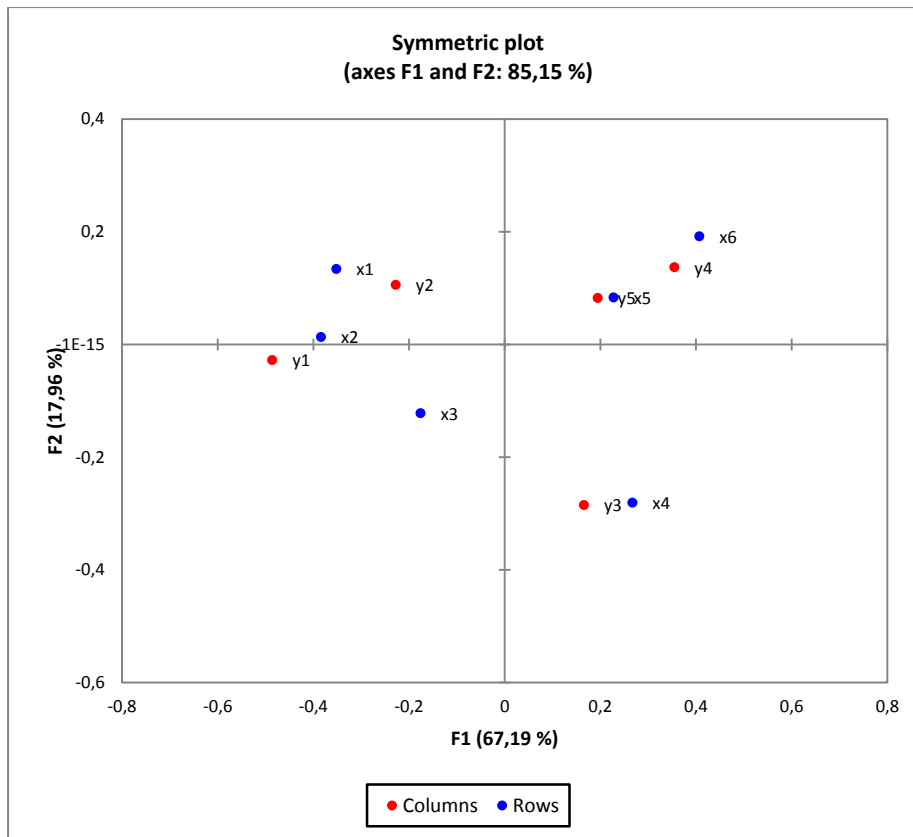
Figur 19 Asymmetrisk KA-plot over søjleprofilerne

Ud for de enkelte akser er der angivet hvor stor en del af den samlede inerti der er forklaret ved denne akse. I eksemplet i Figur 19 er 85,15 % af den samlede inerti forklaret ved de to akser. Dvs. at der er 14,85 % af inertien, der ligger i dimensioner, der ikke er illustreret. Dette er et udtryk for, hvor langt profilerne ligger fra det plan, der er illustreret.

Hvis inertien i plottet er relativt lille, dvs. hvis de observerede observationer ligger tæt på origo, der angiver den gennemsnitlige profil, vil profilerne klumpe sig sammen i midten af diagrammet. Derved kan det blive vanskelig at lave en tolkning af diagrammet ud fra et asymmetrisk plot.

Symmetrisk plot

Den sædvanlige måde at illustrere et KA-plot er at lave et symmetrisk plot. Det betyder, at både række- og søjleprofilerne indtegnes i det samme diagram med deres principale koordinater. Et sådant plot er vist på Figur 20. Her er den indbyrdes afstand rækkeprofilerne imellem χ^2 -afstandene, og ligeledes er den indbyrdes afstand søjleprofilerne imellem χ^2 -afstandene. Men afstandene mellem en række- og en søjleprofil er ikke meningsgivende.



Figur 20 Symmetrisk KA-plot.

Ud fra det symmetriske KA-plot er det muligt at tolke rækkeprofilerne for sig og søjleprofilerne for sig. Derudover er det muligt at tolke på retningerne for hhv. række- og søjleprofilerne, da disse er uændret ved skaleringen. Ved at lægge en akse ind i retning af en af søjleprofilerne, er det muligt at projicere rækkeprofilerne ind på denne akse. F.eks. kan der lægges en akse ind i retning af y4. Det ses da, at profilen x6 har det største relative indhold i den dimension og x2 det laveste, såfremt man kan se bort fra udbredelsen i de øvrige dimensioner.

I næste afsnit vil der blive vist eksempler på, hvordan et KA-plot kan tolkes i konkrete sammenhænge.

Praktiske eksempler på anvendelse af korrespondanceanalyseplot

I dette kapitel vil der blive givet nogle enkelte eksempler på hvordan korrespondanceanalyse ville kunne anvendes i en gymnasial sammenhæng.

I det første eksempel vises hvordan beregningerne kan udføres trin for trin i statikprogrammet R. Ved hvert trin i beregningerne vises resultatet i tabelform.

Eksempel 1 - Beregning af KA plot over stemmefordeling ved folketingsvalg

I et tidligere eksempel blev det illustreret, hvordan stemmefordelingen på tre partier i tre geografiske områder kunne vises i et todimensionelt plot. I dette eksempel vil vi udvide den grafiske præsentation til at dække hele valgresultatet.

Hvis der skulle vises en dækkende illustration af stemmefordelingen i forhold til geografi fra sidste folketingsvalg skulle alle opstillede partier og samtlige geografiske områder medtages. Stemmefordelingen fra folketingsvalget i 2011 fordelt på storkredse er angivet i Tabel 28.

	A	B	C	F	I	K	O	V	Ø
København	80705	71264	23262	52870	24882	1185	35887	64914	70831
Københavns omegn	81557	33992	21763	27767	16703	1086	40444	69948	23221
Nordsjælland	54182	33537	19765	21012	20613	1171	30279	91870	15062
Bornholm	9801	1509	580	2011	509	681	2972	7309	1992
Sjælland	133571	39840	24863	51952	23857	2022	85571	139888	29959
Fyn	89577	26658	16033	32990	13164	1330	38978	76845	19914
Syddjylland	108937	29742	17905	34717	21912	4714	66666	149148	17266
Østjylland	131338	50350	18028	43364	23682	3090	50274	131674	30471
Vestjylland	76208	24090	12677	26819	16061	9664	40378	113443	11505
Nordjylland	113739	25716	20171	32690	15202	3127	45277	102686	16639

Tabel 28 Stemmefordeling folketingsvalget 2011. Kilde: Danmarks Statistik

For at kunne udføre beregningerne til et korrespondanceanalyseplot, er det nødvendigt at benytte et egnet matematikværktøj. Dette kunne være statistikværktøjet R, der er et gratis open source program, og dermed umiddelbart tilgængeligt. I det følgende eksempel er de anvendte programlinjer for beregningerne i R angivet.

Første trin i proceduren er at indlæse data i R. Dette kan gøres på en række forskellige måder. I dette eksempel hentes data fra en .txt fil:

```
my.path <- "C:\\Users\\lga\\Documents\\Master\\Speciale\\R_datafiler\\" #setup
path
```

```
data<-read.table(file=paste(my.path,"storkredse.txt",sep="")) # indlæs data
```

```
data # udskriv data i terminalvinduet
```

Indholdet i variabelen *data* svarer til matricen **D** fra der foregående kapitel.

De relative fordelinger beregnes ud fra formlen $\mathbf{P} = \frac{1}{d_{++}} \mathbf{D}$, mens de marginale fordelinger beregnes som hhv. summen af rækkerne og summen af søjlerne i $\mathbf{P}$:

```
data.P <- data/sum(data) # beregn de relative fordelinger
```

```
data.r <- apply(data.P, 1, sum) # beregn summen over alle rækkerne
```

```
data.c <- apply(data.P, 2, sum) # beregner summen over alle søjlerne
```

De marginale fordelinger $\mathbf{r}$ og $\mathbf{c}$ bliver:

København	Københavns omegn	Nordsjælland	Bornholm	Sjælland	Fyn	Syddjylland	Østjylland	Vestjylland	Nordjylland
0.12016	0.08931	0.08113	0.00772	0.14999	0.08903	0.12727	0.13609	0.09336	0.10589

A	B	C	F	I	K	O	V	Ø
0.24823	0.09501	0.04939	0.09205	0.04983	0.00792	0.123246	0.26745	0.06684

Disse fordelinger angiver stemmefordelingen på hhv. storkreds og parti.

Ud fra disse vektorer dannes to diagonale matricer:

```
data.Dr <- diag(data.r) # generere en diagonalmatrice ud fra r
```

```
data.Dc <- diag(data.c) # generere en diagonalmatrice ud fra c
```

Matrice $\mathbf{S}$ for de standardiserede residualer beregnes ud fra formen:

(119)

$$\mathbf{S} = \mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r} \mathbf{c}^T) \mathbf{D}_c^{-1/2}$$

omregne elementerne i diagonalmatricerne til 1 over kvadratroden

```
data.Drmh <- diag(1/sqrt(data.r))
```

```
data.Dcmh <- diag(1/sqrt(data.c))
```

```
data.P<-as.matrix(data.P)
```

```
data.S <- data.Drmh %*% (data.P - data.r %o% data.c) %*% data.Dcmh
```

Matricen $\mathbf{S}$ kan omskrives ved en SVD faktorisering, der benytter den indbyggede SVD funktion i "ca" pakken til R:

```
data.svd <- svd(data.S)
```

```
data.svd$d # singulære værdier
```

```
data.svd$u # principale rækkevektorer
```

```
data.svd$v # principale søjlevektorer
```

Ved SVD faktoriseringen findes de singulære værdier, samt de principale vektorer for række- og søjlekategorierne. Disse variable er indeholdt i parameteren data.svd.

De singulære værdier er angivet i Tabel 29, sammen med inertien for de enkelte dimensioner:

	1.akse	2.akse	3.akse	4.akse	5.akse	6.akse	7.akse	8.akse	9.akse
Singulær værdi	0,2196	0,0942	0,0750	0,0549	0,0376	0,0167	0,0149	0,0008	0,0000
Inerti	0,048208	0,008879	0,00562	0,003012	0,001412	0,00028	0,000222	6,43E-07	1,71E-33
Inerti-fordeling	71,27875	13,12743	8,309749	4,453963	2,088067	0,413521	0,32757	0,00095	2,53E-30

Tabel 29 Singulære værdier og inertier

Det ses at 71,3 % af spredningen i fordelinger ligger langs den første principale akse. De øvrige akser har faldende inertibidrag. Inertien for den 9. akse er reelt 0, hvilket indikerer, at der ikke er nogen udbredelse i denne dimension. Dette skyldes at matricen har maksimal rang 9, og derfor højst kan ligge i 8 dimensioner.

De principale vektorer for række-kategorierne (storkredsene) og søjlekategorierne (partierne) er angivet som kolonner i Tabel 30 og Tabel 31. De enkelte vektorer i tabellerne er normerede og ortogonale.

u	X1	X2	X3	X4	X5	X6	X7	X8	X9
København	-0,851	0,183	-0,198	-0,178	-0,023	0,042	-0,193	-0,112	-0,336
Københavns omegn	-0,116	-0,198	0,232	0,141	0,564	0,351	0,512	-0,003	-0,226
Nordsjælland	0,011	0,460	0,683	0,258	0,022	-0,063	-0,204	0,250	-0,336
Bornholm	0,057	-0,070	-0,372	0,111	0,001	0,566	-0,126	0,617	-0,163
Sjælland	0,114	-0,364	0,161	-0,613	-0,038	-0,176	0,186	0,032	-0,488
Fyn	0,003	-0,339	-0,083	0,123	0,039	-0,514	-0,232	0,512	-0,190
Syddjælland	0,327	0,115	0,101	-0,342	-0,319	0,435	-0,266	-0,122	-0,233
Østjylland	0,002	-0,024	-0,134	0,450	-0,633	-0,052	0,476	-0,102	-0,372
Vestjylland	0,325	0,595	-0,474	-0,106	0,340	-0,251	0,178	-0,018	-0,302
Nordjylland	0,184	-0,309	-0,129	0,385	0,244	0,032	-0,476	-0,505	-0,372

Tabel 30 Principale vektorer for række-kategorierne

v	X1	X2	X3	X4	X5	X6	X7	X8	X9
A	0,167	-0,556	-0,331	0,496	-0,015	0,173	0,104	-0,136	0,498
B	-0,443	0,282	0,141	0,244	-0,128	-0,060	0,616	0,389	0,308
C	-0,090	-0,030	0,429	0,197	0,778	-0,015	-0,288	0,174	0,222
F	-0,185	-0,148	-0,183	-0,205	-0,032	-0,864	-0,164	-0,032	0,303
I	-0,077	0,267	0,320	0,055	-0,003	-0,029	0,157	-0,862	0,223
K	0,204	0,505	-0,654	-0,070	0,476	0,006	0,183	-0,055	0,089
O	0,200	-0,276	0,168	-0,739	0,130	0,192	0,357	0,060	0,351
V	0,447	0,429	0,153	0,023	-0,359	0,052	-0,384	0,215	0,517
Ø	-0,668	0,047	-0,267	-0,242	-0,070	0,423	-0,409	-0,062	0,259

Tabel 31 Principale vektorer for søjlekategorierne

For at indtegne rækkeprofilerne skal standardkoordinaterne bestemmes ud fra sammenhængen

$$\mathbf{F} = \mathbf{D}_r^{-1/2} :$$

```
data.rsc <- data.Drmh %*% data.svd$u # standardkoordinater for række(  $\mathbf{G}$  )
```

```
data.rpc <- data.rsc %*% diag(data.svd$d) # principal koordinat for rækker (  $\mathbf{F}$  )
```

Ud fra sammenhængen $\mathbf{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}$ bestemmes de principale koordinater for kategorierne

```
data.csc <- data.Dcmh %*% data.svd$v # standardkoordinat for søjler (  $\mathbf{\Gamma}$  )
```

```
data.cpc <- data.csc %*% diag(data.svd$d) # principal koordinat for rækker (  $\mathbf{\Phi}$  )
```

Det er kun koordinaterne for de første to akser, der skal anvendes for at indtegne punkterne i det todimensionelle plot. Koordinaterne for hhv. kategorierne ($\mathbf{\Gamma}$) og profilerne ($\mathbf{F}$) er angivet i tabellerne nedenfor:

Storkreds	1.akse	2.akse
København	-2,454	0,529
Københavns omegn	-0,390	-0,664
Nordsjælland	0,038	1,615
Bornholm	0,649	-0,797
Sjælland	0,294	-0,941
Fyn	0,010	-1,138
Sydjylland	0,917	0,322
Østjylland	0,004	-0,066
Vestjylland	1,065	1,946
Nordjylland	0,564	-0,948

Tabel 32 Standardkoordinater for kategorierne

Partibogstav	1.akse	2.akse
--------------	--------	--------

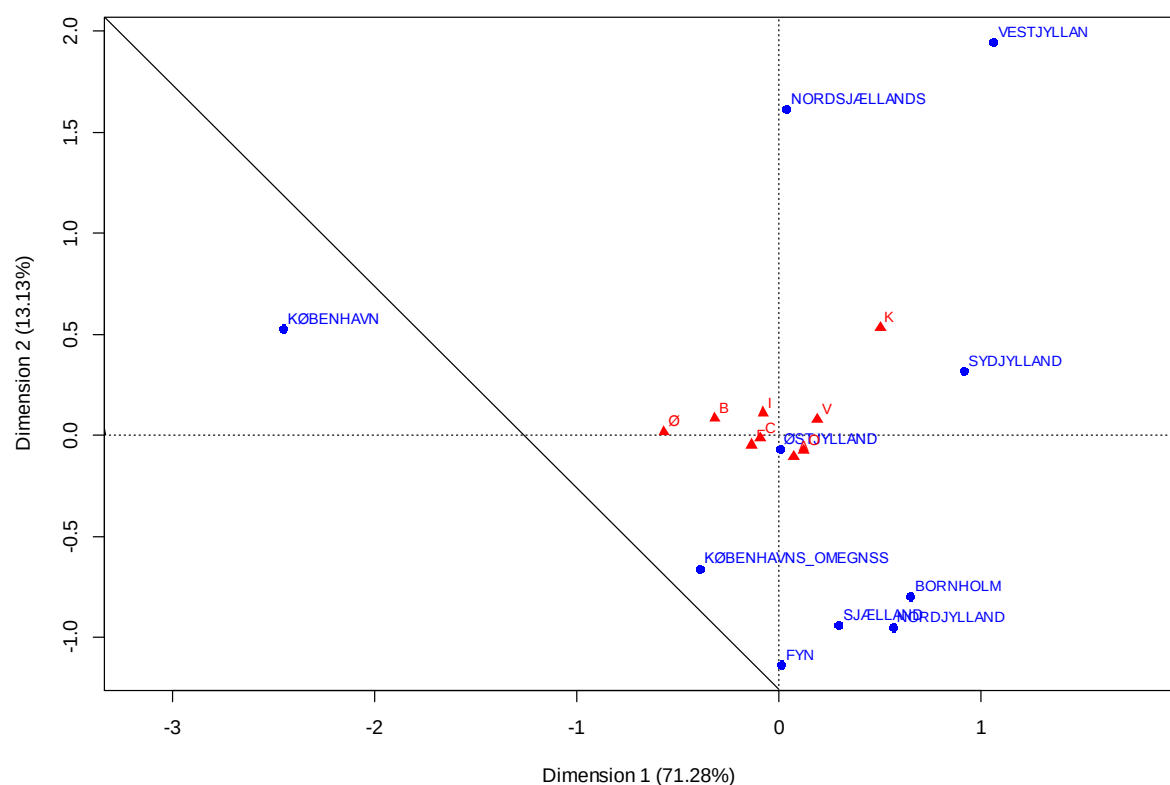
A	0,073	-0,105
B	-0,315	0,086
C	-0,089	-0,013
F	-0,134	-0,046
I	-0,076	0,112
K	0,504	0,534
O	0,125	-0,074
V	0,190	0,078
Ø	-0,567	0,017

Tabel 33 De principale koordinater for profilerne

Disse koordinater kan anvendes til at tegne korrespondanceanalyseplottet. I R findes der også en enkelt kommando der kan anvendes til at generere plottet direkte, ud fra de relative fordelinger:

```
plot(ca(data.P),map="colprincipal")
```

Denne funktion indtegner det asymmetriske korrespondanceplot, med kolonnerne som kategorier. Plottet er vist på Figur 21



Figur 21 Asymmetrisk korrespondanceplot over stemmefordelingen ved folketingsvalget 2011.

Plottet viser positionen af partier, såvel som storkredse. Første dimension indeholder en stor del af den samlede spredning, med godt 71 % af inertien, mens anden dimension indeholder 13 %. De resterende knap 16 % ligger i dimensioner der ikke er illustreret.

Ud fra plottet kan man lave en samfundsfaglig tolkning af aksernes betydning, hvilket eventuelt kunne danne grundlag for en hypotesedannelse. I dette tilfælde ville man eventuelt kunne danne en politisk højre-/venstreskala ved at projicere partierne ind på førsteaksen. Tilsvarende kunne man projicere storkredsene ind på førsteaksen hvilket kunne være en afspejling af urbaniseringsgraden i de enkelte storkredse. Det er også muligt at tolke på placeringen af de enkelte profiler. En hypotese kunne være at der er en sammenhæng mellem urbanisering og politisk orientering. At partier som Enhedslisten og Kristendemokraterne begge har en geografisk slagside ved Enhedslistens relativt store repræsentation i København og Kristendemokraternes tilsvarende i Vestjylland, er også synlig i plottet. Plottet illustrerer samtidig at vælgerfordelingen i Østjyllands storkreds er repræsentativ for landsgennemsnittet, da positionen af denne ligger tæt op ad origo.

I plottet ovenfor er partierne indtegnet som profiler med deres standardkoordinater, mens områderne er indtegnet som kategorier, med deres principale koordinater. Det er muligt at lave det omvendte plot ved at beregne standardkoordinaterne for kolonnerne ved sammenhængen $\mathbf{G} = \mathbf{D}_c^{-1/2} \mathbf{V} \mathbf{\Sigma}$ samt de principale koordinater for rækkerne ud fra sammenhængen $\mathbf{\Phi} = \mathbf{D}_r^{-1/2} \mathbf{U}$. Beregningerne af disse er vist tidligere.

Det asymmetriske korrespondanceplot kan tegnes ud fra disse koordinater, eller direkte ved funktionen:

```
plot(ca(data.P),map="rowprincipal")
```

For at tegne det symmetriske korrespondanceplot benyttes standardkoordinaterne for både række- og søjleprofilerne. Det symmetriske plot kan også tegnes direkte ved funktionen:

```
plot(ca(data.P))
```


Supplerende punkter

Ud over de indtegnede storkredse er det muligt at indtegne supplerende punkter. Man kunne f.eks. være interesseret i hvorledes enkelte opstillingskredse vil placere sig i det generelle billede.

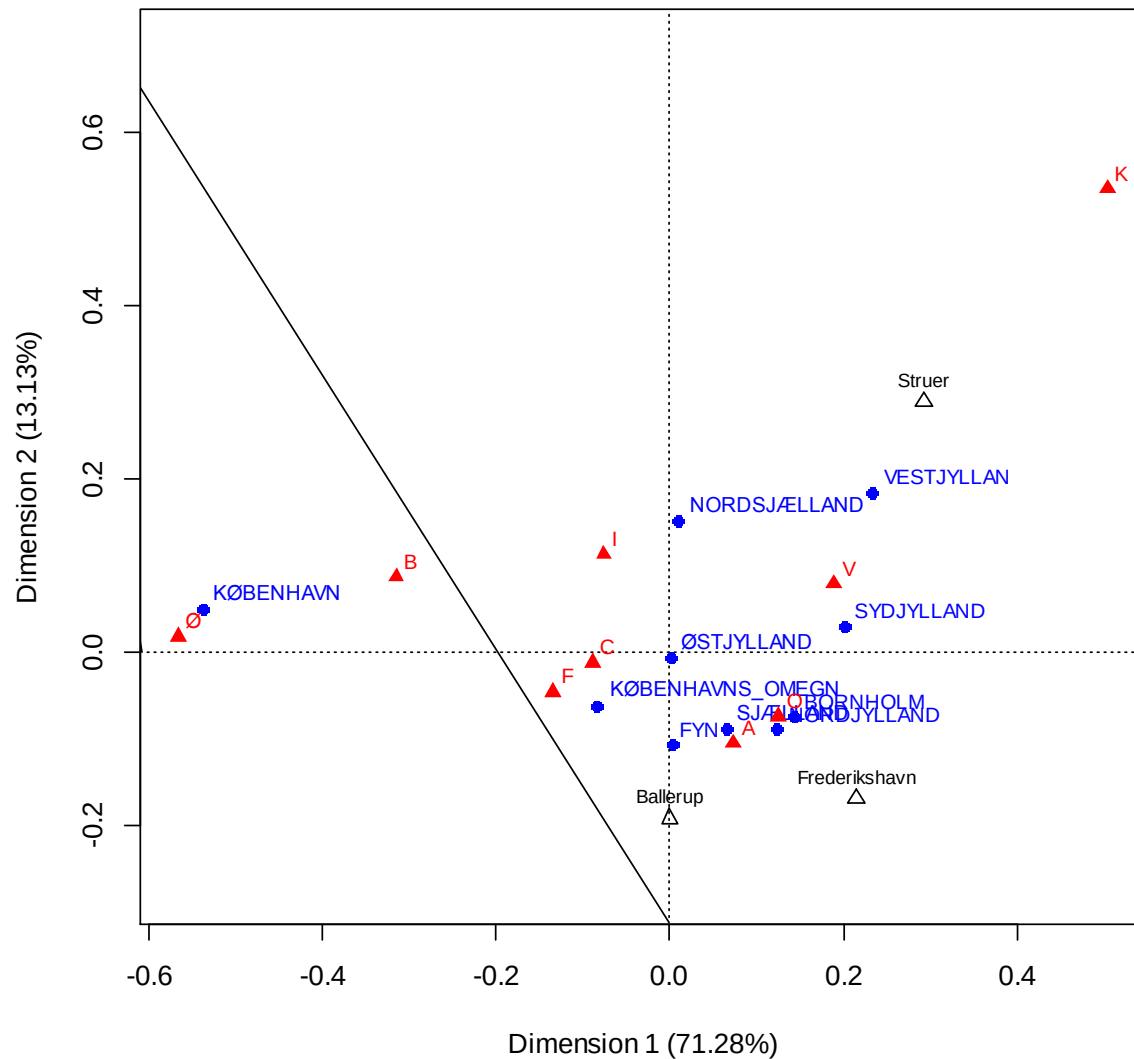
	BALLERUP	STRUER	FREDERIKSHAVN
A. Socialdemokratiet	13919	6958	14069
B. Radikale Venstre	3606	1770	1970
C. Det Konservative Folkeparti	2024	934	1876
F. SF - Socialistisk Folkeparti	3661	2975	3227
I. Liberal Alliance	1858	1306	1403
K. Kristendemokraterne	156	1765	310
O. Dansk Folkeparti	6599	3536	5869
V. Venstre, Danmarks Liberale Parti	8959	8470	11560
Ø. Enhedslisten - De Rød-Grønne	2962	833	1456

Tabel 34 Stemmedfordelingen for udvalgte opstillingskredse ved folketingsvalget 2011

I Tabel 34 er angivet stemmedfordelingen i tre udvalgte opstillingskredse. Da disse data allerede er omfattet af de storkredse der er benyttet til at tegne korrespondanceplottet skal de indtegnes i plottet uden vægt, dvs. uden at bidrage til inertien eller placeringen af profilerne. Koordinaterne for de supplerende punkter bestemmes ved først at bestemme den relative fordeling i opstillingskredsen. Dernæst bestemmes koordinaterne ved at projicere vektoren ind på standardvektorerne for rækkerne:

```
Ballerup<-c(13919,3606,2024,3661,1858,156,6599,8959,2962)
sup.Ballerup<-Ballerup/sum(Ballerup)
point.Ballerup<-t(sup.Ballerup)%*%data.csc[,1:2]
points(point.Ballerup,pch=2)
```

I Figur 22 er vist det symmetriske KA-plot med opstillingskredsene Ballerup, Struer og Frederikshavn indtegnet som supplerende punkter.



Figur 22 Symmetrisk KA-plot med opstillingskredse indtegnet som supplerende profiler.

Hvis man indtegnede samtlige opstillingskredse for en given storkreds ville positionen af disse opstillingskredse fordele sig omkring storkredsen, med storkredsens placering som det vægtede gennemsnit.

I eksemplet her er anvendt den geografiske stemmefordeling. De bagvedliggende data er frit tilgængelige fra Danmarks Statistik. Hvis man havde adgang til mere detaljerede information om vælernes baggrund kan korrespondanceanalysen benyttes til at illustrere andre sammenhænge i forhold til politisk tilhørsforhold. I en samfundsfaglig sammenhæng ville det være interessant at belyse sammenhængen mellem politisk ståsted og f.eks. økonomisk baggrund, uddannelse eller alder. For at undersøge disse sammenhænge vil det være nødvendigt med adgang til mere detaljerede survey data.

Eksempel 2 - Analyse af udvalgte forfatters værker

I "ca" pakken til R findes et indbygget datasæt "author", hvor man har optalt hyppigheden af sandtlige bogstaver i uddrag fra nogle kendte forfatters værker. De værker der er undersøgt fremgår af listen nedenunder:

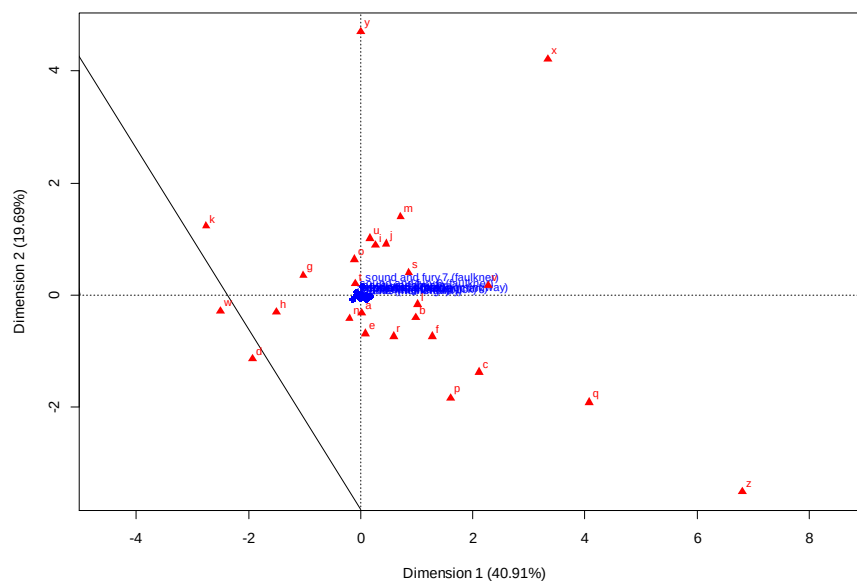
Three daughters (Pearl S. Buck)	Sound and fury 6 (William Faulkner)
Drifters (James A. Michener)	Profiles of future (Arthur C. Clarke)
Lost world (Arthur C. Clarke)	Islands (Ernest Hemingway)
East wind (Pearl S. Buck)	Pendorric 3 (Victoria Holt)
Farewell to arms (Ernest Hemingway)	Asia (James A. Michener)
Sound and fury 7 (William Faulkner)	Pendorric 2 (Victoria Holt)

Datasættet hentes og plottes med R kommandoerne:

```
data(author)
```

```
plot(ca(author),map="rowprincipal")
```

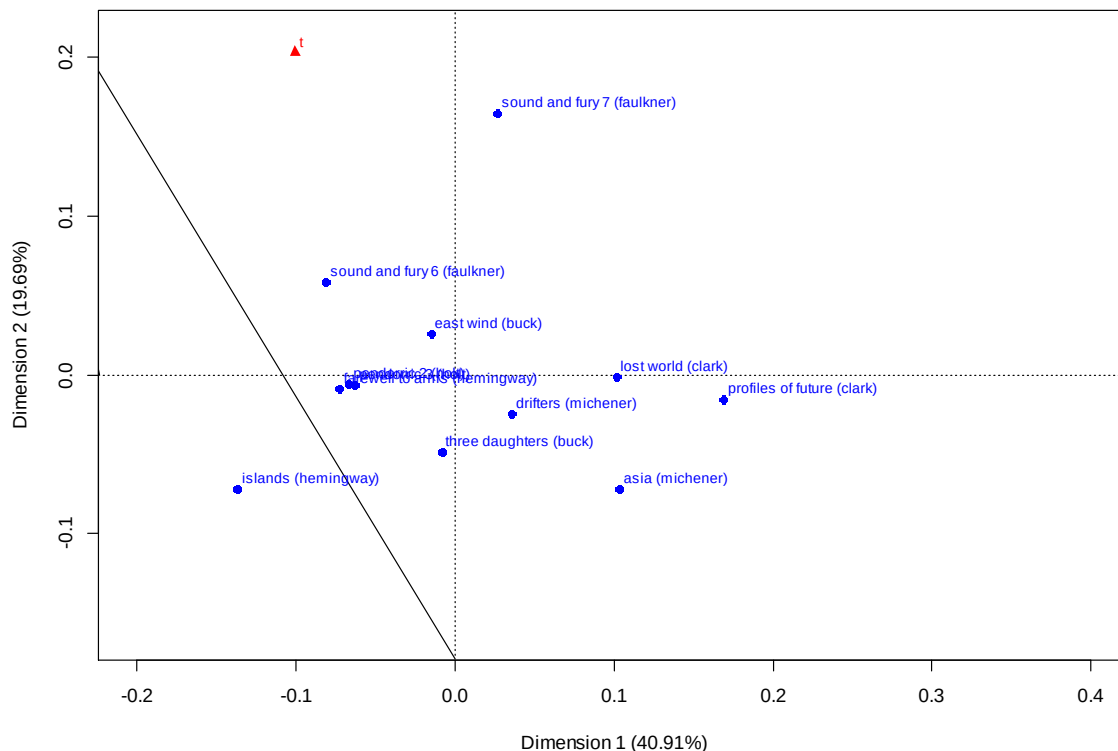
Inertien i fordelingen er meget lille. Hvilket resulterer i at punkterne for de enkelte værker klumper sig sammen nær origo af plottet.



Figur 23 KA-plot over udvalgte værker.

Ved at zoome ind på fordelingen nær origo bliver det muligt at se at der er en korrelation mellem værker skrevet af de samme forfattere.

```
plot(ca(author),map="rowprincipal", xlim=c(-0.2, 0.4), ylim=c(-0.15, 0.2))
```



Figur 24 KA-plot der viser hvorledes forskellige værker af samme forfatter viser en korrelation mht. frekvensen af bogstaver i værkerne.

Eksemplet her viser at det kan være muligt at uddrage interessante konklusioner, selv om en fordeling i udgangspunktet har en meget lav inert. Profiler for værker af den samme forfatter placerer sig i nærheden af hinanden i KP-plottet. Modellen kunne f.eks. benyttes til at undersøge om det er muligt at identificere en ukendt forfatter til et tekstuddrag via et KA-plot.

Eksempel 3 - Individuel spørgeskemaundersøgelse

Det er oplagt at benytte korrespondanceanalyse til at illustrere resultater af surveys - f.eks. spørgeskemaundersøgelser. Her er man undertiden interesseret i at benytte de enkelte respondenter som den ene dimension i et skema, og de mulige svarkategorier som den anden dimension.

Uddrag af et sådant spørgeskema er vist i Tabel 35. I det originale spørgeskema indgår svar fra 20 personer på 25 forskellige spørgsmål. Datasættet bygger i vid udstrækning på elevernes selv vurdering på en række spørgsmål, de angivne svar er derfor subjektive af natur, hvilket kan have indflydelse på undersøgelsens anvendelighed.

	Person 1	Person 2	Person 3	Person 4	Person 5
Køn (Mand=0, Kvinde=1)	0	0	0	0	1
Fritidsarbejde (skala 1 til 10)	1	7	9	2	3
Fritidsaktiviteter (skala 1 til 10)	7	3	2	8	8
Lektier (skala 1 til 10)	6	4	4	4	5
Forberedelse (Skala 1 til 10)	7	5	7	5	7
Egen vurdering af skriftligt niveau (12-skala)	7	4	10	10	7
Egen vurdering af mundtligt niveau (12-skala)	7	4	10	7	4

Tabel 35 Uddrag af resultat af spørgeskemaundersøgelse

Værdierne i tabellen kan ikke direkte tolkes som frekvensfordeling af respondenternes besvarelse. For at kunne benytte resultaterne fra spørgeskemaet i en korrespondanceanalyse skal resultaterne først omsættes til en tabel der indeholder frekvenserne. Dette gøres ved at lave én svarkategori til hver svarmulighed i hvert delspørgsmål. F.eks. vil kategorien Køn erstattes af to kategorier - Mand og Kvinde. I princippet skulle der for spørgsmålene med mulighed for at svare på en skala fra 1 til 10, indføres 10 nye kategorier, hvor ni svarmuligheder vil have et 0 og den kategori respondenter faktisk har svaret, vil have et 1. I det anvendte spørgeskema vil det betyde at hele datasættet kan oversættes til et nyt datasæt, med ca 200 svarkategorier i stedet for de oprindelige 25 svarkategorier. Da hver respondent kun har mulighed for at angive én svarmulighed for hvert spørgsmål, vil det omskrevne datasæt udelukkende bestå af 1 (sand) og 0 (falsk). En del af kategorierne vil udelukkende bestå af respondenter der har angivet 0 til spørgsmålet. Disse kategorier kan uden tab af information fjernes fra datasættet.

Et korrespondanceanalyseplot med op mod 200 kategorier vil dels blive uoverskuelig, dels vil inertien i de to dimensioner blive relativt lille. Det kan derfor være en fordel at samle nogle af kategorierne i fælleskategorier inden analysen. I eksemplet fra Tabel 35 kan resultaterne eksempelvis omskrives til Tabel 36.

Ved kategorierne "Egen vurdering" er de oprindelige kategorier for mundtlig og skriftlig egenvurdering slået sammen, da en adskillelse mellem skriftligt og mundtligt niveau ikke er et mål for undersøgelsen. Dermed bliver den maksimale værdi i disse kategorier 2.

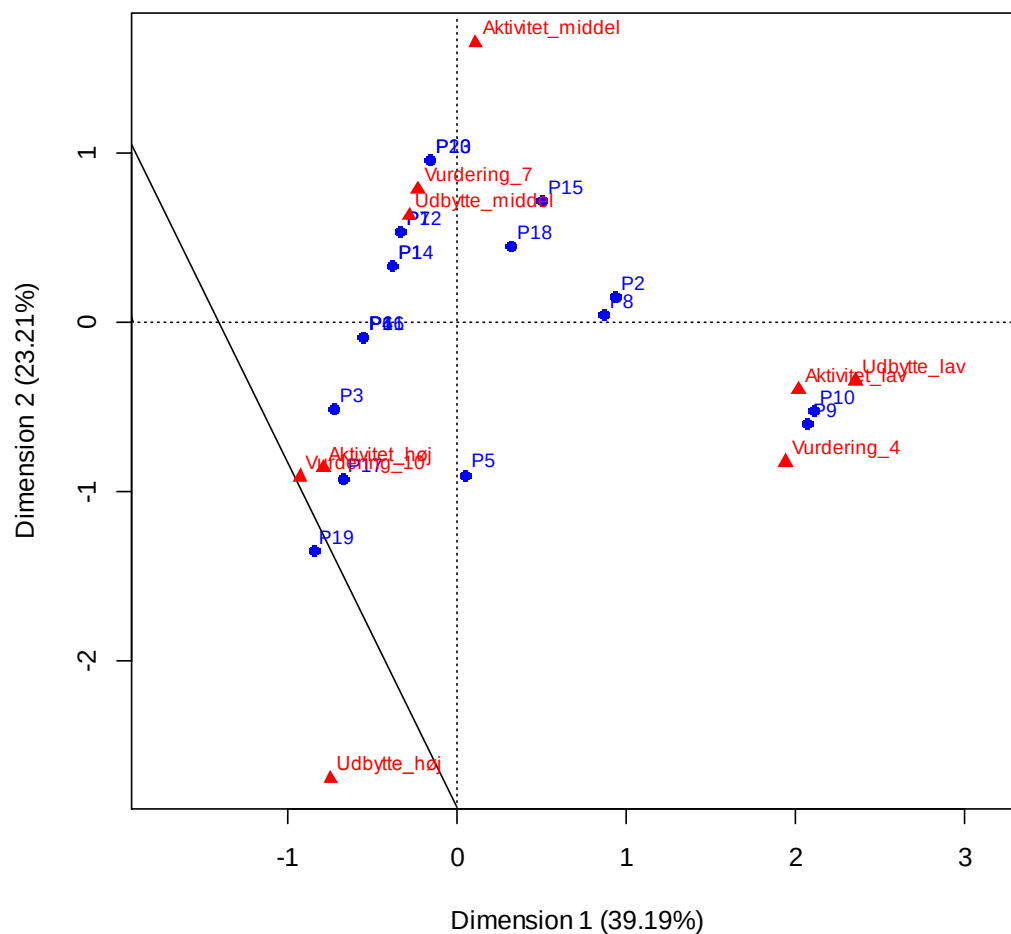
	Person 1	Person 2	Person 3	Person 4	Person 5
Mand	1	1	1	1	0
Kvinde	0	0	0	0	1
Fritidsarbejde_lav (under 3 timer/uge)	1	0	0	1	1
Fritidsarbejde_middel (3 til 6 timer/uge)	0	0	0	0	0
Fritidsarbejde_høj (over 6 timer/uge)	0	1	1	0	0
Fritidsaktiviteter_lav (under 3 timer/uge)	0	1	1	0	0
Fritidsaktiviteter_middel (3 til 6 timer/uge)	0	0	0	0	0
Fritidsaktiviteter_høj (over 6 timer/uge)	1	0	0	1	1
Lektier_lav (under 4 timer/uge)	0	0	0	0	0
Lektier_middel (4 til 10 timer/uge)	1	1	1	1	1
Lektier_høj (over 10 timer/uge)	0	0	0	0	0
Forberedelse (Skala under 3)	0	0	0	0	0
Forberedelse_middel (Skala 4 til 7)	0	1	0	1	0
Forberedelse_høj (Skala over 6)	1	0	1	0	1
Egen vurdering 4	0	2	0	0	1
Egen vurdering 7	2	0	0	1	1
Egen vurdering 10	0	0	2	1	0

Tabel 36 Uddrag af resultat af spørgeskemaundersøgelse

Før korrespondanceanalyseplottet genereres skal man beslutte hvilke kategorier der er relevante for at udspænde rummet på en meningsfyldt måde. I dette eksempel er et af formålene med undersøgelsen at afdække de enkelte elevers faglige niveau, set i relation til resten af klassen. Kategorier som køn, fritidsarbejde og fritidsaktiviteter ignoreres derfor ved generering af plottet.

Det har stor indflydelse på det endelige plot, hvilke kategorier der medtages ved genereringen af rummet for korrespondanceplottet. På den ene side kan der være et ønske om at få illustreret så mange facetter af spørgeskemaets resultater som muligt. På den anden side kan flere kategorier også resultere i flere dimensioner i datasættet, med deraf lavere andel af inertien vist i de første to dimensioner. På den måde adskiller korrespondanceanalyse sig ikke fra andre statistiske metoder, hvor man også må udvælge variable, kollapse kategorier, foretage faktoranalyser mv. inden man foretager de endelige statistiske tests.

I eksemplet i Figur 18 er vist et korrespondanceplot, hvor rummet er udspændt af kategorierne for selv vurderet aktivitetsniveau, karakterniveau samt udbytte af undervisningen.



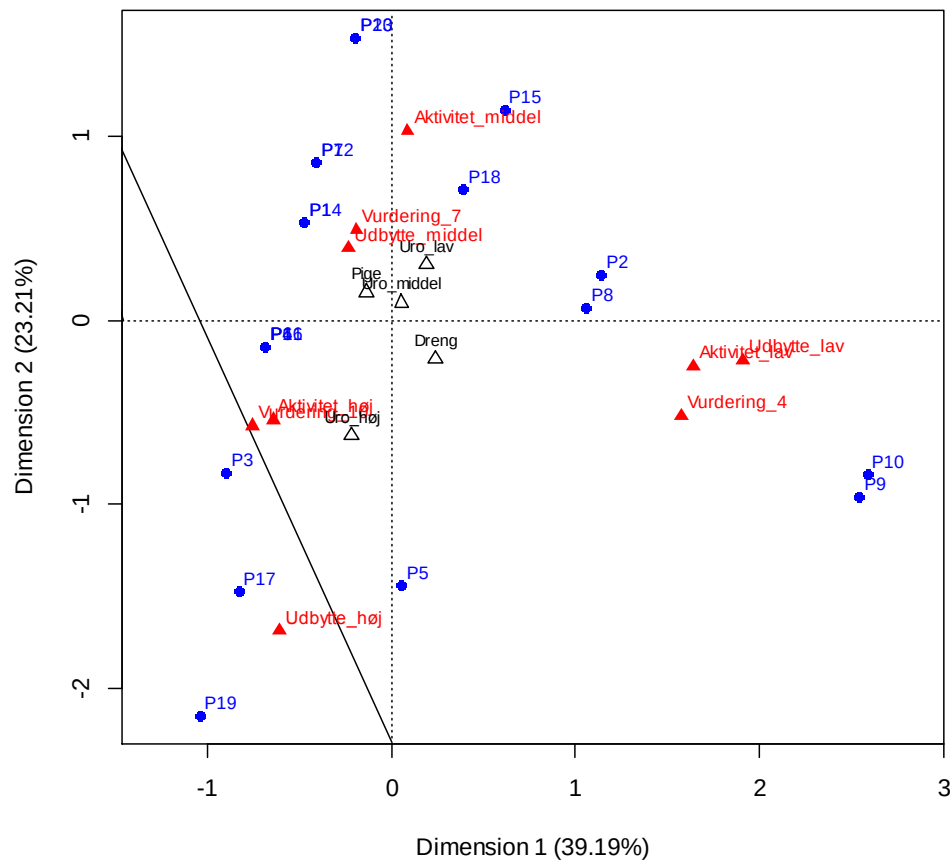
Figur 25 Asymmetrisk korrespondanceplot med kategorierne som principale elementer. "P" punkterne repræsenterer de individuelle respondenter

Hver kategori er repræsenteret ved tre punkter, der ikke er indbyrdes uafhængige, da det vægtede gennemsnit for hver kategori vil ligge i origo.

I dette rum ses det hvordan de enkelte elever placerer sig i forhold til deres selv vurdering. I det viste plot er der, måske ikke overraskende, en sammenhæng mellem selv vurderet karakterniveau, udbytte af undervisningen og aktivitetsniveau. Et sådant plot kunne give et udgangspunkt for en samtale med eleven i forbindelse med karaktergivningen.

Det kunne måske være interessant at undersøge om der er en tydelig kønsforskel i det faglige niveau. Ved at indtegne køn som en supplerende kategori ses det at Dreng/Pige kategorierne ligger relativt tæt på origo, dvs. primært i en dimension der ikke er velrepræsenteret i plottet. Derudover går Dreng/Pige akser næsten vinkelret på akser mellem højt og lavt niveau. Det kan dermed ikke konkluderes at forskellen mellem højt og lavt fagligt niveau har en sammenhæng med køn. Se Figur 26.

Tilsvarende kan kategorien for hvor problematisk man oplever uro i klassen indtegnes som et supplerende punkt. Her er der noget det indikerer at eleverne med et højt fagligt niveau også er dem der er mest generet af uro i klassen.



Figur 26 Korrespondanceplot med køn og gener fra uro i klassen indtegnet som supplerende punkter.

Konklusion

Der er en velkendt talemåde, der siger at der findes tre slags løgn: Løgn, forbandet løgn og statistik. Sandhedsværdien af denne talemåde vil ikke blive bedømt her. Men den korrekte tolkning af statistiske data er en disciplin der ofte giver anledning til mange diskussioner. Specielt ved datasæt med mange variable kan det være svært at få et klart overblik over de sammenhænge der ligger bag observationerne.

For at kunne uddrage de relevante konklusioner er det væsentligt at kunne stille de rigtige spørgsmål. Dette er ofte en mindst lige så vanskelig disciplin som at udføre selve de statistiske undersøgelser.

I gymnasieverdenen bliver eleverne ofte stillet overfor at skulle tolke statistiske data, eller forholde sig kritisk til andres tolkninger af data. Der er derfor behov for en overordnet indsigt i hvordan statistiske modeller som f.eks. korrespondanceanalyse kan anvendes. Ligesom der er et behov for selv at kunne udføre statistiske analyser og uddrage konklusioner baseret på disse analyser.

I denne rapport er den basale matematiske baggrund for korrespondanceanalysen gennemgået. De enkelte trin i korrespondanceanalysen er beskrevet på en let tilgængelig måde, med brug af simple eksempler til illustration. Endelig er der beskrevet eksempler på hvordan korrespondanceanalysen kan udføres i statistikprogrammet R.

De væsentligste matematiske elementer i korrespondanceanalysen er anvendelsen af X^2 -værdien som et mål for afstandene mellem de enkelte elementer i datasættet, samt anvendelsen af SVD faktorisering til at reducere antallet af dimensioner i et datasæt på en optimal måde. Begge elementer er beskrevet i detaljer i rapporten.

Forslag til konkret anvendelse af korrespondanceanalyse i en gymnasial sammenhæng

Et oplagt sted at inddrage korrespondanceanalyse i gymnasiet ville være i et tværfagligt forløb mellem matematik og et andet fagområde som f.eks. samfundsfag. Her kunne man tage udgangspunkt i et konkret datasæt, der kan danne grundlag for en analyse.

Ved at udfører en korrespondanceanalyse, skal eleverne tage stilling til hvilke parametre der skal udvælges for at udspænde rummet for korrespondanceplottet, og hvilke parametre der evt. skal lægges ind som supplerende punkter. I denne proces vil der skulle foretages en afvejning af hvor stor en andel af inertien der kan illustreres i de to dimensioner, i forhold til hvor mange dimensioner der medtages. Derudover skal det vurderes hvorvidt det er muligt at give en fornuftig tolkning af akserne i korrespondanceplottet.

I arbejdet med at udvælge data og teste forskellige kombinationsmuligheder vil eleverne opleve en kreativ proces hvor forskellige sammenhænge vil vise sig i plottet. Disse sammenhænge kan danne grundlag for fremsættelse af hypoteser. Hypoteserne kan efterfølgende undersøges ved yderligere analyse af datamaterialet, eller ved indsamling af ny data.

En anden mulighed for at inddrage korrespondanceanalysen i gymnasiet kunne være i forbindelse med de større skriftlige opgaver / studieretningsopgaver, hvor eleverne har

mulighed for individuelt at gå mere i dybden med et emne, end der er mulighed for i den normale klasseundervisning. Her ville der være mulighed for at arbejde mere detaljeret med emner som f.eks. projektion på et underrum eller spektral dekomposition.

Korrespondanceanalyse som selvstændig fagdisciplin indeholder mange flere facetter end der er beskrevet i denne rapport. Der vil naturligvis også være mulighed for at gå mere i dybden med nogle af de mere avancerede teknikker indenfor korrespondanceanalyse eller beslægtede statistiske metoder. Men dette vil dog sandsynligvis være for krævende for den standardiserede gymnasieelev.

Litteraturliste

- (1) Geometric Data Analysis.
Brigitte Le Roux and Henry Rouanet
Kluwer Academic Publisher, 2004
- (2) Correspondence Analysis in Practice
Michael Greenacre
Chapman & Hall/CRC, 2007
- (3) Elementary Linear Algebra
L. E. Spencer, A. J. Insel and S. H. Freidberg
Pearson Education Inc., 2008
- (4) Using the Singular Value Decomposition
E. J. Ientilucci
Rochester Institute of Technology, 2003
- (5) Singular Value Decomposition
John Hopcroft and Ravi Kannan
<http://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage>, 2012
- (6) Singular Value Decomposition Tutorial
Kirk Baker, 2013
- (7) An Introduction to R.
W. N. Venables and D. M. Smith