

# 1 Bakgrund - korrespondensanalysens historia

Mitt intresse för korrespondensanalys uppstod när jag läste Donald Broadys doktorsavhandling om Bourdieus författarskap: *Sociologi och Epistemologi*,<sup>1</sup> och Donald Broady och forskningsgruppen för utbildnings- och kultursociologi i Uppsala är också de som prövat metoden i sociologiska analyser i Sverige (De höll t.ex. en workshop om korrespondensanalys i Uppsala hösten 2001, och anordnar även kurser i korrespondensanalys sedan hösten 2001). I Frankrike har korrespondensanalysen blivit alltmer vanlig sedan slutet av 1960-talet, och det hänger säkert samman med Bourdieus användning av den. Under första halvan av 1980-talet kom en del, relativt tekniskt orienterade, texter på engelska om korrespondensanalys, men det är först under 1990-talet som mer lättillgängliga introduktioner publicerats: t.ex. [Cla98] och [WR90]. En sökning i GUNDA visar att av de 15 verk om korrespondensanalys som omnämns i [Cla98] fanns enbart ett verk att låna, två andra verk hade varit inlånade via fjärrlån, men UB hade inga egna exemplar. Huvudverket i litteraturlistan till den kurs om korrespondensanalys som hålls i Uppsala finns inte heller med i GUNDA. Under 2003 har det tillkommit två verk i libris, båda författade av en statistiker i Uppsala.

Bourdieu har hämtat korrespondensanalysen från den franske statistikern Jean-Paul Benzécri som utvecklade metoden under 1960-talet. Som vanligt finns det mindre kända föregångare i andra delar av världen, t.ex. i U.S.A. men enligt Broady, har Benzécri hävdade att han i stort sett var obekant med dessa fram till mitten av 1970-talet. Mer om metodens historia och släktskapet med andra metoder t.ex. faktoranalys finns i [WR90, p. 14f] [Bro90, p. 489f], [Cla98, p. 5f] och [Beh02]. Förvirringen har stundom varit mycket stor, det var först 1985 som man visade att flera tekniker (t.ex. korrespondensanalys och principal component analysis) som tidigare ansetts vara skilda, delade samma matematiska grund. Det första exemplet på en genomförd analys med denna matematiska grund hittade Weller och Romney i en text av Fischer från 1940, och hans beräkningar stämde med exakt med vad dagens datorprogram ger.<sup>2</sup> Fischers text "The precision of discriminant functions" publicerades i en tidskrift med namnet: *Annals of eugenics*.

Korrespondensanalysen är alltså inte specifik för sociologin. Benzécri's syfte med att utveckla den var för analysera lingvistiska data med en induktiv metod i kontrast till Chomskys deduktionistiska arbetssätt.<sup>3</sup> Psykologen

---

<sup>1</sup>[Bro90]

<sup>2</sup>[WR90, p. 15]

<sup>3</sup>[Bro90, p. 486]

Guttman publicerade 1941 obekant med Fischers arbete, artikeln ”The quantification of a class of attributes: a theory and method of scale construction” om metoden och han har givit upphov till en av dess benämningar ”Guttman weighting”. I Frankrike blev en avhandling i psykologi 1965 först med att använda den.<sup>4</sup>

## 2 Användningsområde - skillnader gentemot logistisk regressions analys

Korrespondens analysens styrka ligger i att presentera en tabell, dvs en matris, på ett sätt så att relationer mellan rader och kolumner i tabellen blir möjliga att se direkt utan att behöva göra en mängd beräkningar i huvudet.

Korrespondens analysen skiljer sig gentemot logistisk regressions analys på flera punkter:

- Det finns inga formaliserade metoder för hypotesprövning
- Alla variabler kan användas, ingen särskild skal-nivå krävs utan allt behandlas som kategori data. För tabeller med annat än frekvensdata måste man ibland göra vissa transformationer, t.ex. för att inga negativa värden får ingå.
- Distinktionen beroende-oberoende variabel upphävs.
- Man kan med använda variabler med många kategorier.

### 2.1 Hypotes- och signifikansprövningar

Benzécri är kritisk mot  $\chi^2$  test och använder istället tolkningsbarheten som ett kriterium, dvs signifikanta samband som inte är tolkningsbara lägger man inget intresse vid, och icke-signifikanta samband som är tolkningsbara fortsätter man att analysera.<sup>5</sup> Clausen respektive Weller och Romney å andra sidan rekommenderar att man gör ett  $\chi^2$  test på tabellen för att fastställa att det finns något samband att analysera.<sup>6</sup> Det finns mått som beskriver hur bra varje punkt är representerad i de grafer som man vanligen presenterar korrespondensanalysen med, och ett annat viktigt sätt att validera sina resultat är att jämföra med sådan information som inte ingår i analysen, t.ex. genom att lägga in så kallade supplementära punkter. Metodens huvudsyfte är dock explorativt.

---

<sup>4</sup>[Bro90, p. 489]

<sup>5</sup>[Bro90, p. 506f]

<sup>6</sup>[Cla98, p. 43f], [WR90, p. 64]

## 2.2 Distinktionen beroende - oberoende variabler

Korrespondensanalysen kräver inte att man gör en distinktion mellan beroende och oberoende variabler.<sup>7</sup> Ett problem som jag uppfattat i logisk regression är att man måste göra en distinktion mellan variabler som är oberoende och variabeln som är beroende. Om man, vilket jag tenderar att göra rätt ofta, föreställer sig modeller där variabler ömsesidigt förstärker varandras effekter, snarare än att uppfatta någon radikal skillnad mellan den beroende variabeln och de andra, är min erfarenhet att det kan vara svårt att finna signifikans för sådana effekter. Och om man inte har signifikans för alla kombinationseffekter man föreställer sig kanske man tenderar att ta bort dem och nöja sig med en förenklad modell.

För logistisk regression gäller att de oberoende variablerna inte enbart måste bara oberoende av den beroende variabeln, de måste också vara oberoende av varandra, vilket jag nästan aldrig kan tro att de är. Det kan hända att felet inte blir så stort när de s.k. oberoende variablerna har effekt på varandra, men det är ändå ett fundamentalt problem om förutsättningar för analysen att analysen ska vara giltig sällan är uppfyllda. Det handlar förstas om vilken verklighetsuppfattning man har, och vilken sorts giltighet man vill kräva av analyserna.

## 2.3 Många variabler - få fall

Syftet med korrespondensanalysen är att få en uppfattning om *relationer mellan variabler* snarare än mellan relationer mellan *fall*. I så motto är det en metod för en relationistisk strukturalism. Förhållandena mellan variablerna A, B och C kan förändras - dvs preciseras genom att konkretiseras - genom att variabeln D läggs till analysen. Å andra sidan innebär ytterligare variabler att de approximationer vi kan visualisera i ett begränsat antal dimensioner får ett större fel.

## 2.4 Kontextberoende

Jag är på jakt efter analysmetoder som kan hjälpa mej att finna komplexa korrelationer mellan variabler. Ett problem med logisk regression är att den behandlar varje oberoende variabls effekt på den beroende variabeln *för sig*, utan hänsyn till vilka värden varje case har på andra variabler. Antag att en variabel A har positiv effekt på den beroende variabeln X, om A kombineras

---

<sup>7</sup>[WR90, p. 7]. Vid multivariata analyser *kan* man använda den sådan distinktion för att avgöra vilka variabler man vill slå ihop, men det finns andra användbara kriterier för att avgöra det, se [Cla98, p. 46f]

med höga värden på B och C, men negativ effekt på X om den kombineras med låga värden på B och C. Om spridningen på B och C är stor kommer As effekt på (eller komplexa korrelationer med) X inte att kunna upptäckas i en logistisk regression, eftersom den *genomsnittliga* effekten av A på X är nära noll.

Beroende på kontexten kan ”samma” fenomen uppfattas olika, och ha olika konsekvenser. Metoder som hjälper oss att hitta sådana förhållanden kan vara ett sätt att överbrygga klyftan mellan kvantitativ och kvalitativ forskning.

Såvitt jag förstår erbjuder korrespondensanalysen en del möjligheter att finna sådana samband, men de verkar inte vara triviala att utforma.<sup>8</sup> Jag återkommer till detta.

### 3 Hur går analysen till?

Korrespondensanalysen kan delas in i följande faser: normera tabellen och dela upp den i en radtabell och kolumntabell, beräkna normerade avstånd, (finn en optimal approximativ lösning för ett reducerat antal dimensioner) rita en graf.

#### 3.1 Normera tabellen och dela upp tabellen i en rad- och en kolumnprofiltabell

Skapa två nya tabeller: en radprofiltabell och en kolumnprofiltabell. I radtabellen har varje värde i den ursprungliga tabellen dividerats med radsumman i den ursprungliga tabellen. I kolumntabellen divideras varje värde i den ursprungliga tabellen med kolumnsumman i den ursprungliga tabellen. Som exempel kan vi ta en tabell över brottsplats och brottstyp:

Landsända	Brottstyp			Total
	Inbrott	Bedrägeri	Vandalism	
Oslo	395	2.456	1.758	4.609
Mellersta Norge	147	153	916	1.216
Nordnorge	694	327	1.347	2.368
Total	1.236	2.936	4.021	8.193

Tabell 1: Ursprunglig tabell

<sup>8</sup>Den enklare framställningen i [Cla98, p. 30] är delvis otydlig på denna punkt, se istället [WR90, p. 85f]

Landsända	Brottstyp			Total	Radmassa
	Inbrott	Bedrägeri	Vandalism		
Oslo	0,086	0,533	0,381	1,000	0,563
Mellersta Norge	0,121	0,126	0,753	1,000	0,148
Nordnorge	0,293	0,138	0,569	1,000	0,289
Genomsnittlig Radprofil	0,151	0,358	0,491	1,000	

Tabell 2: Radprofiltabell

Landsända	Brottstyp			Genomsnittlig Kol.profil
	Inbrott	Bedrägeri	Vandalism	
Oslo	0,320	0,837	0,437	0,563
Mellersta Norge	0,119	0,052	0,228	0,148
Nordnorge	0,561	0,111	0,335	0,289
Total	1,000	1,000	1,000	
Kolumnmassor	0,151	0,358	0,491	

Tabell 3: Kolumnprofiltabell

Den första raden i radprofiltabellen ges alltså genom att dividera värdena i den ursprungliga tabellen 395, 2.456 och 1.758 med radsumman för den första raden (Oslo) 4.609. Den andra raden i radprofiltabellen ges på liknande sätt genom att värdena i den ursprungliga tabellen 147, 153 och 916 divideras med radsumman 1.216.

I Kolumnprofiltabellen ges värdena genom att varje värde i utgångstabellen divideras med den aktuella kolumnsumman: värdena i inbrottskolumnen 395, 147 och 694 delas med kolumnsumman för inbrott 1.236, värdena i bedrägerikolumnen (2.456, 153 och 327) delas med kolumnsumman för bedrägeri 2.936, och värdena i vandalismkolumnen (1.758, 916 och 1.347) delas med kolumnsumman för vandalism 4.021. De genomsnittliga rad- respektive kolumnprofilerna innehåller information om antalet "observationer" i raderna respektive kolumnerna, så att informationen att Oslo hade 4.609 brott vilket i förhållande till t.ex. Nordnorges 2.368 var nästan dubbelt så många finns kvar i förhållandet mellan Osloradens "massa" 0,563 och Nordnorges radmassa 0,289. På samma sätt bevaras magnitudförhållandet mellan kolumnerna i kolumnmassorna, t.ex. är vandalisms kolumnmassa ca tre gånger större än kolumnmassan för inbrott.

Vi har nu normerat värdena och delat upp i två tabeller. Vi skulle utifrån

dessa tabeller kunna återskapa den ursprungliga tabellen om bara hade något av de absoluta värdena i den.

Det första tricket i korrespondensanalys består i att behandla radvärdena som koordinater i ett rum. Gör man det kan man rita ut städernas positioner och få en grafisk presentation av *avståndet* mellan dem (med avseende på brott). Ju närmare städerna hamnar varandra, desto mer liknar de varandra.

Om vi betraktar radvärdena som koordinater, kan vi använda olika formler för att bestämma avståndet mellan punkterna. I korrespondensanalys använder man begreppet  $\chi^2$ -avstånd, vilket bestäms av en formel som lättast förstås med följande exempel:  $\chi^2$  avståndet mellan Oslo (1) och Mellersta Norge (2) definieras enligt följande:

$$\chi_{(1,2)}^2 = \sqrt{\frac{(0,086-0,121)^2}{0,151} + \frac{(0,533-0,126)^2}{0,358} + \frac{(0,381-0,735)^2}{0,491}}$$

$$\chi_{(1,2)}^2 = 0,868$$

För att förstå formeln, jämför med tabell 2 på föregående sida. De två raderna (radprofilerna) jämförs kolumn för kolumn och skillnaden divideras med den aktuella kolumnens massa. Divisionen med kolumnmassan för brottstypen gör  $\chi^2$  avståndet mellan områdena i större utsträckning än vanligt avstånd (ej division med kolumnmassan) *känsligt för de variabler som har liten frekvens*. Det är hela tiden den relativa skillnaden i frekvenser som spelar roll. En effekt av divisionen med kolumnmassan innebär att de ovanliga fallen hamnar långt bort från nollpunkten (genomsnittsfallet).

Område	Origo	Oslo	Mellersta Norge	Nordnorge
Origo	0	0,371	0,545	0,531
Oslo	0,371	0	0,868	0,890
Mellersta Norge	0,545	0,868	0	0,515
Nordnorge	0,531	0,890	0,515	0

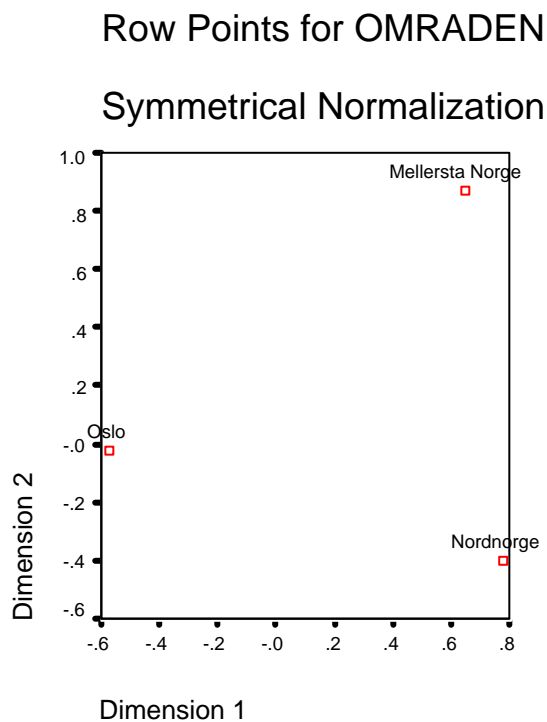
Tabell 4:  $\chi^2$  avstånden mellan områden (raderna)

Origo, grafens mittpunkt, är ett tänkt genomsnittsområde, vars radprofil är identisk med den genomsnittliga radprofilen.

I det här exemplet där vi bara har tre punkter (plus en mittpunkt) att definiera, är det inga problem att hitta en lösning, dvs. att rita en graf där samtliga punkters avstånd till varandra och mittpunkten stämmer exakt med värdena i tabellen. Om vi hade haft fyra punkter då hade det behövts det (eller kunde behövts) tre dimensioner för att alla avstånd ska stämma. Fem

punkter kräver (maximalt) fyra dimensioner och då kan vi tvingas nöja oss med en approximativ graf.

Exakt hur man går från tabellen med avstånd till en grafisk lösning kommer jag inte att gå in på här. Jag nöjer mej med att konstatera att datorn provar sig fram tills den har en tillräckligt exakt lösning.



Figur 1: Relationerna mellan områdena (raderna)

Vi gör nu samma beräkningar fast för brotten (kolumnerna) istället, och använder därvid kolumnprofiltabellen istället för som tidigare radprofiltabellen. Avståndet mellan till exempel bedrägeri och vandalism definieras enligt följande:

$$\chi_{(2,3)}^2 = \sqrt{\frac{(0,837-0,437)^2}{0,563} + \frac{(0,052-0,228)^2}{0,148} + \frac{(0,111-0,335)^2}{0,289}}$$

$$\chi_{(2,3)}^2 = 0,815$$

Det andra tricket i korrespondensanalysen, och som också är en unik fördel med metoden, är att vi nu kan rita in punkterna för brotten (kolumnerna)

Brottstyp	Origo	Inbrott	Bedrägeri	Vandalism
Origo	0	0,606	0,552	0,279
Inbrott	0,606	0	1,098	0,531
Bedrägeri	0,552	1,098	0	0,815
Vandalism	0,279	0,531	0,815	0

Tabell 5:  $\chi^2$  avstånden mellan brottstyperna (kolumnerna)

och områdena (raderna) *i samma graf*. Vi kan nu *se* vilka brott som ”hör ihop” med vilka områden. (Se figur 3 på sidan 10).

The placing of both row variables and column variables explicitly in the same space is one of the important advantages of CA and is a great aid in interpretation of the data.<sup>9</sup>

När datorn ritar ut både radpunkter och kolumnpunkter i samma diagram har jag fattat det så att den roterar de två punktsvärmarna så att avstånden mellan ett område och samtliga brott stämmer så bra som möjligt med områdets rad i kolumnprofiltabellen, och vilket är samma sak, att avstånden mellan ett brott och samtliga områden stämmer så bra som möjligt med brottets kolumn i radprofiltabellen.

Avståndet mellan två punkter av olika slag, t.ex. mellan ett område och ett brott har ingen mening *i sig*, men avstånden mellan ett område och två brott - dvs *skillnaden* mellan avståndet mellan ett område och ett visst brott och avståndet mellan samma område och ett annat brott - är meningsfull.<sup>10</sup> Med den utgångspunkten avgör man hur man ska kombinera punktsvärmarna från två olika kategorier. Vi har ingen *definition* på avståndet mellan Oslo och inbrott, men vi vet, genom att studera Oslos rad i kolumnprofiltabellen, att avståndet inbrott $\leftrightarrow$ Oslo ska vara *större än* avståndet vandalism $\leftrightarrow$ Oslo som i sin tur ska vara *större än* avståndet bedrägeri $\leftrightarrow$ Oslo. Och vi vet, genom att studera inbrottskolumnen i radprofiltabellen, att avståndet inbrott $\leftrightarrow$ Nord Norge ska vara *mindre än* avståndet inbrott $\leftrightarrow$ Mellersta Norge som i sin tur ska vara *mindre än* avståndet inbrott $\leftrightarrow$ Oslo.

## 4 Tolkning av grafer

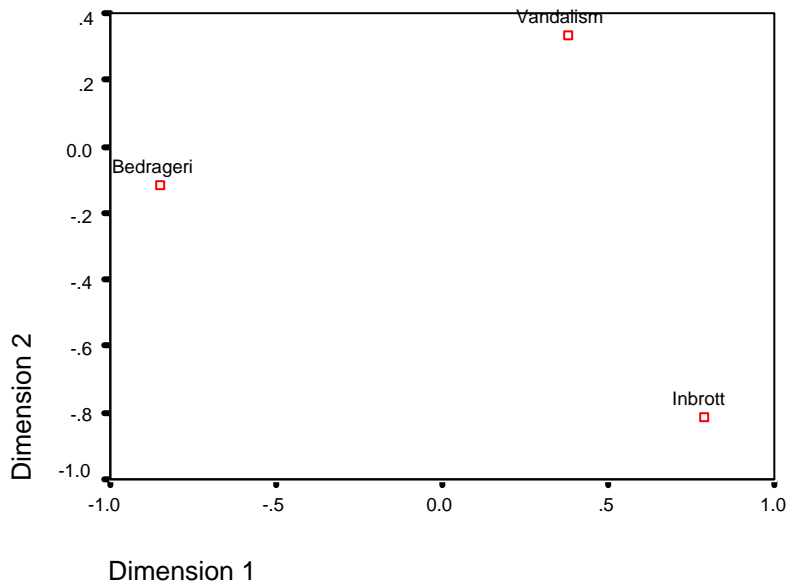
Vad betyder det då att två punkter ligger nära varandra i grafen?

<sup>9</sup>[WR90, p. 73]

<sup>10</sup>[Cla98, p. 21]

## Column Points for BROTT

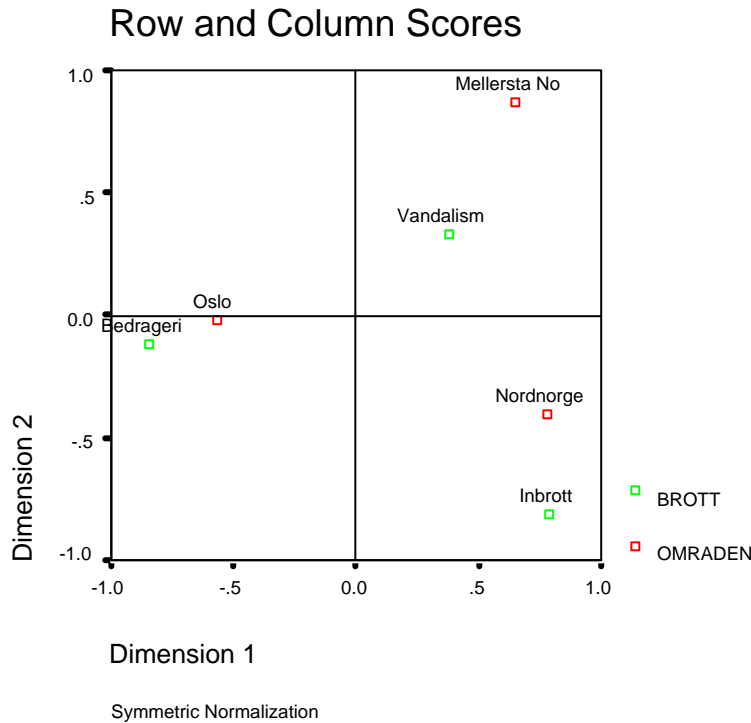
### Symmetrical Normalization



Figur 2: Relationerna mellan brottstyperna (kolumnerna)

Avståndet mellan två punkter av olika slag, t.ex. mellan ett område och ett brott har ingen mening *i sig*, men avstånden mellan ett område och två brott - dvs *skillnaden* mellan avståndet mellan ett område och ett visst brott och avståndet mellan samma område och ett annat brott - är meningsfull, men för att förstå den skillnaden måste den jämföras med skillnaden mellan ett annat område och dessa två brottstyper.

När jag satte mig in metoden dröjde det länge innan jag insåg detta. Först trodde jag att skillnaden i avstånd mellan ett område och två brottstyper korresponderar mot sannolikheten att ett brott rapporterat i detta område är av den ena eller den andra typen. Men så enkelt är det inte. Ta till exempel området Nordnorge, det ligger närmast inbrott och något längre bort ligger vandalism och längst bort av brottstyperna ligger bedrägeri. Inbrott är emellertid *inte det vanligaste brottet i Nordnorge* (694 inbrott men hela 1347 fall av vandalism). Att inbrott och Nordnorge ligger nära varandra i grafen beror *inte heller på att de flesta inbrotten sker i Nordnorge*. Det råkar vara så just i detta fallet (694 inbrott i Nordnorge, 395 i Oslo), men inbrott skulle ligga närmast Nordnorge även om antalet var 390 istället för 694). Betrakta



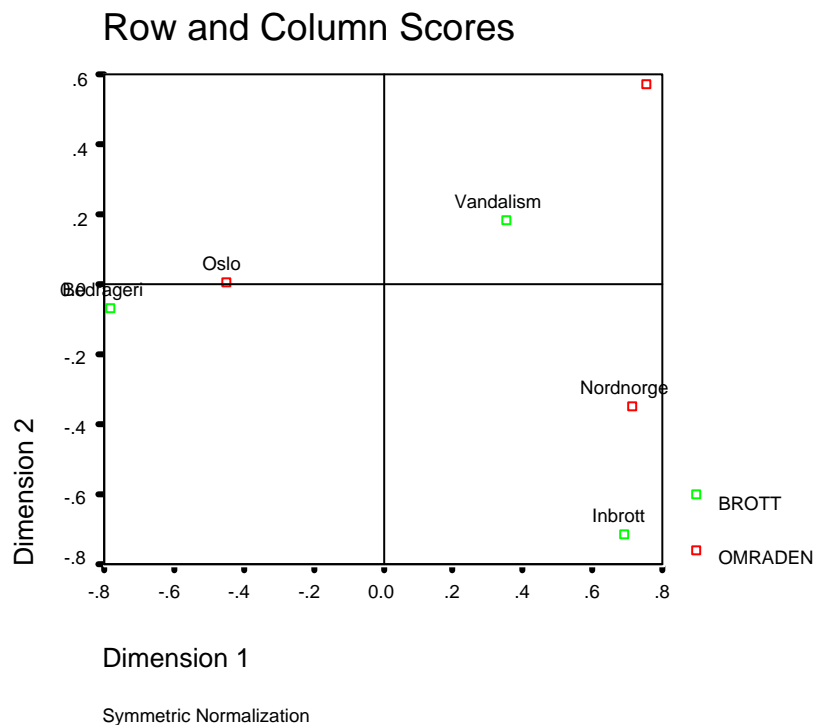
Figur 3: Relationerna mellan områden och brottstyper (raderna och kolumnerna)

brottstypen vandalism, den ligger närmast Mellersta Norge och relativt nära Nordnorge medan Oslo ligger längst bort från den. Dock begicks de flesta vandalismbrott i Oslo (1.758 mot 1347 i Nordnorge). För att visa hur oberoende avstånden i korrespondensanalysen är av de absoluta frekvenserna, har jag ändrat värdena på inbrott i Nordnorge från 694 till 390 (så att flest inbrott begås i Oslo som har 395 inbrott) och Vandalismbrott i Oslo från 1.758 till 2.460 (så att de flesta brott som begås i Oslo är vandalismbrott och de flesta vandalismbrott begås i Oslo). Därmed ändras Oslos radprofil och kolumnprofilerna för Inbrott och Vandalism.<sup>11</sup> Men förändrar det relationerna mellan avstånden i grafen? Figur 4 på följande sida är en graf utifrån modifierade data.

Oslo har visserligen kommit något närmare vandalism, men Oslo är fortfarande det område som ligger längst bort från vandalism trots att:

- De flesta vandalismbrott begås i Oslo

<sup>11</sup>Nordnorges radprofil ändras också, men ändringen innebär inte att storleksförhållandena i radprofilen ändras.



Figur 4: Relationerna mellan områden och brottstyper (raderna och kolumnerna), *modifierade data*

- De flesta brott som begås i Oslo är vandalismbrott

Storleksordningen på avstånden är desamma i de båda graferna, vilket måste förstås i relation till radprofilerna i tabell 6 på nästa sida.

I kolumnen för inbrott har fortfarande Nordnorge det största värdet, och i kolumnen för vandalism har Oslo fortfarande det minsta värdet. I Mellersta Norge och Nordnorge är vandalism överrepresenterat eftersom de har högre värden för vandalism än det genomsnittliga området (den genomsnittliga radprofilen). I Oslo är vandalism underrepresenterat eftersom Oslo har ett lägre värde för vandalism än genomsnittsområdet.

Vad gäller inbrott är Oslo återigen underrepresenterat: Oslos värde 0,074 är mindre än den genomsnittliga radprofilen 0,108. Både Mellersta Norge och Nordnorge har värden över 0,108 och där är alltså inbrott överrepresenterat.

Så, vad betyder det då att avståndet mellan inbrott och Nordnorge är mindre än avståndet mellan inbrott och Oslo och avståndet mellan inbrott och Mellersta Norge? Det betyder att, i jämförelse med Oslo och Mellersta Norge, är *andelen* inbrott, jämfört med alla brott i Nordnorge, större i

Landsända	Brottstyp			Total	Radmassa
	Inbrott	Bedrägeri	Vandalism		
Oslo	0,074	0,462	0,463	1,000	0,618
Mellersta Norge	0,121	0,126	0,753	1,000	0,148
Nordnorge	0,189	0,158	0,653	1,000	0,240
Genomsnittlig Radprofil	0,108	0,342	0,550	1,000	

Tabell 6: Radprofiltabell utifrån modifierade data

Nordnorge. Flest inbrott i absoluta tal kan ha begåtts någon annanstans.

Vad betyder det att avståndet mellan inbrott och Nordnorge är mindre än avståndet mellan bedrägeri och Nordnorge och avståndet mellan vandalism och Nordnorge? Det betyder att, i jämförelse med bedrägeri och vandalism, är *andelen* inbrott begåvna i Nordnorge, jämfört med alla inbrott, större. Inbrott behöver inte vara den vanligaste brottstypen i Nordnorge (mätt i absoluta tal).

Jörg Blausius kommenterar tolkningen av grafer så här:

[...] when interpreting CA solutions one has to avoid expressions such as "most of ..." and rather use terms such as "above the average ..."<sup>12</sup>

Jag tycker begreppet överrepresenterad och underrepresenterad är träffande.

Vårt exempel är lite väl enkelt efter som vi har just tre punkter och därför klarar oss med två dimensioner (för förhållandena inom brott och inom landändorna, men inte mellan dem). I ett verkligt exempel skulle en graf i två dimensioner behöva vara en approximativ lösning och det påpekas i litteraturen att för stora tabeller kan man inte vara säker på att avståndet mellan två punkter (av samma variabel) svarar proportionellt mot deras "likhet".

## 5 Några begrepp

Variansen definieras i korrespondensanalysen som summan av varje punkts avstånd till mittpunkten multiplicerat med radmassan för den punkten.

Varje dimension, eller axel, förklarar en del av den totala variationen, och man kan räkna ut hur stor del av variationen varje dimension förklarar.

<sup>12</sup>[Bla94, p. 52]

Grafiskt kan man se detta omedelbart genom att se hur stor längd av de olika axlarna som "används" av punkter. Varje axels bidrag till variansen kallas axelns, eller dimensionens, *eigenvalue*.

## 6 Multivariata analyser

Varje kolumn och varje rad i ursprungstabellen, dvs varje kategori, kommer att representeras av en punkt i grafen. Om vi vill att en kategori ska kunna dyka upp på mer än ett ställe i grafen måste vi *kombinera* denna kategori med en annan. Detta är ofta fallet om man vill se hur variabler som kön, klass, etnicitet och ålder *tillsammans* spelar roll för de övriga variablerna i materialet. Ta t.ex. Bourdieus graf över rummet av sociala positioner.<sup>13</sup> Här är yrket den variabel som de övriga relateras till.<sup>14</sup> Vi får här *genomsnittsvärden* för hur dessa relationer gäller för män respektive kvinnor, äldre respektive yngre etc. Är vi intresserade av t.ex. könsskillnader måste vi skapa kombinationsvariabler (kvinna+lantarbetare, man+lantarbetare, kvinna+bonde, man+bonde, kvinna+ej facklörd arbetare, man+ej facklörd arbetare, kvinna+facklörd arbetare, man+facklörd arbetare o.s.v.). Ett exempel på hur en sådan graf kan se ut finns i [BBP02], reproducerat i figur 5 på följande sida.

Ett annat exempel på multivariat analys har vi här:

Först en univariat analys med enbart ålder: Tabell 3.1 och Diagram 3.1 från [Cla98, p. 27f]

Sedan en multivariat analys med ålder och kön: Tabell 3.4 och Diagram 3.2 och 3.3 från [Cla98, p. 31ff]

### 6.1 Tolkning av grafer - flera dimensioner

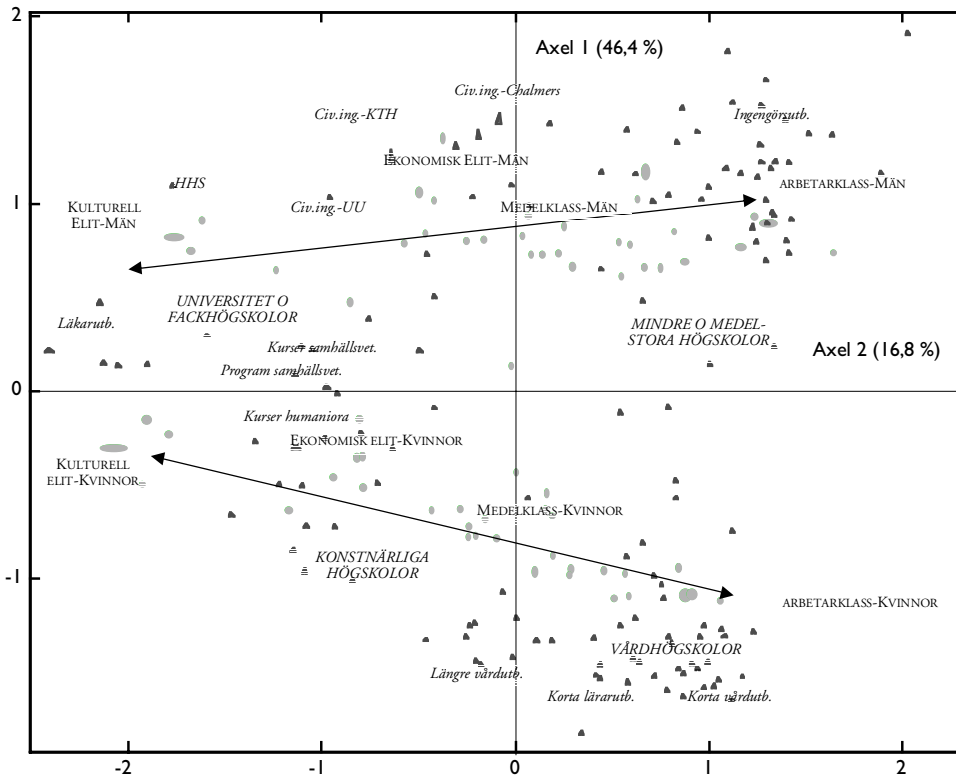
I exemplet ovan visas dimension 1-2 i det vänstra diagrammet och dimension 2-3 i det högra. Man kan föreställa sig det som att det vänstra diagrammet är ett foto av en tredimensionell lösning taget vinkelrätt från den axel som förklarar mest, dvs den axel som punkterna sprids mest längst (dvs den axel som har högst *eigenvalue*). Det andra diagrammet kan betraktas som ett foto på samma tredimensionella lösning, men där kameran placerats på den axeln som i det förra diagrammet var motivet. Datorer kan användas för att presentera tredimensionella lösningar. Genom att studera *eigenvalue* statistiken kan man avgöra hur många dimensioner som är intressanta. Se t.ex. tabell 3.5, där det klart framgår att tre dimensioner är fullt tillräckligt.

---

<sup>13</sup>[Bou84, p. 128f]

<sup>14</sup>Se tabellerna A2-A11, ej A7, [Bou84, p. 526-545]

Graf 1. Högskolefältet 1998, 300 utbildningar, 32 sociala grupper, män och kvinnor separerade, dimension 1 och 2.\*



\* Cirkarna (grå) avser sociala grupper och triangelarna (svarta) utbildningar. En fulltextversion av grafen återfinns i appendix. Storleken på cirkarna och triangelarna visar på bidragsvärdet för modaliteten till axel 1 och axel 2.

Figur 5: Kön, klassbakgrund och valt lärosäte

Ta en titt på Bourdieus diagram igen. Antag att Bourdieus data var felaktiga och behövde korrigeras därför att en större andel universitetslärare hade smak för potatis än vad Bourdieus korrespondensanalys byggde på. Vilken effekt skulle det få för grafen? Antag vidare att du föreställde dej att dessa universitetslärare, dvs just de som ofta åt potatis, vore intressanta att analysera för sig, hur skulle du gå till väga?

[återstår: outliers, tilläggsinformation, mått på hur bra varje punkt beskrivs av dimensionen, mått på hur bra varje dimension beskrivs av punkterna]

## 7 Ett exempel till

Ytterligare ett exempel, denna gång med data som presenterades på Europeiska Sociologförbundets kongress i Murcia hösten 2003: I ett antal europeiska länder har man genomfört enkätundersökningar till kvinnor och män och bland annat frågat om antal arbetade timmar per vecka i lönearbete, inkomst, arbetade timmar per vecka i oavlönat vårdande arbete. Majella Kilkey har tittat på de par där båda är lönearbetande (minns inte riktigt definitionen), som jag tror är särskilt intressanta av den anledningen att andelen sådana par ökar i Europa,<sup>15</sup> och delar in paren efter om om fördelningen av antal arbetade timmar per vecka, fördelningen av inkomst, fördelningen av antalet arbete oavlönade timmar i vårdande arbete är jämställd, om kvinnan gör mer eller om mannen gör mer.

För varje aspekt (lönearbete, inkomst, vårdande arbete) finns alltså tre möjliga utfall:

- mannen gör mer än kvinnan
- kvinnan gör mer än mannen
- de gör ungefär lika mycket (skillnaden är mindre än 20%)

Sammantaget ger detta 27 olika möjliga utfall eller sätt att vara par på. För varje land så har Kilkey räknat antal par i varje av de 27 möjliga positionerna och sammanställt uppgifterna (i procent) i tabell 6.1.

I Appendix 1 har Kilkey markerat de fyra dominerande typerna av par: 10 (Jämställda par), 14 (Män lönearbetar och tjänar mer, kvinnor vårdar mer), 15 (De lönearbetar lika mycket, män tjänar mer och kvinnor vårdar mer) och 18 (De lönearbetar och tjänar lika mycket, kvinnor vårdar mer).

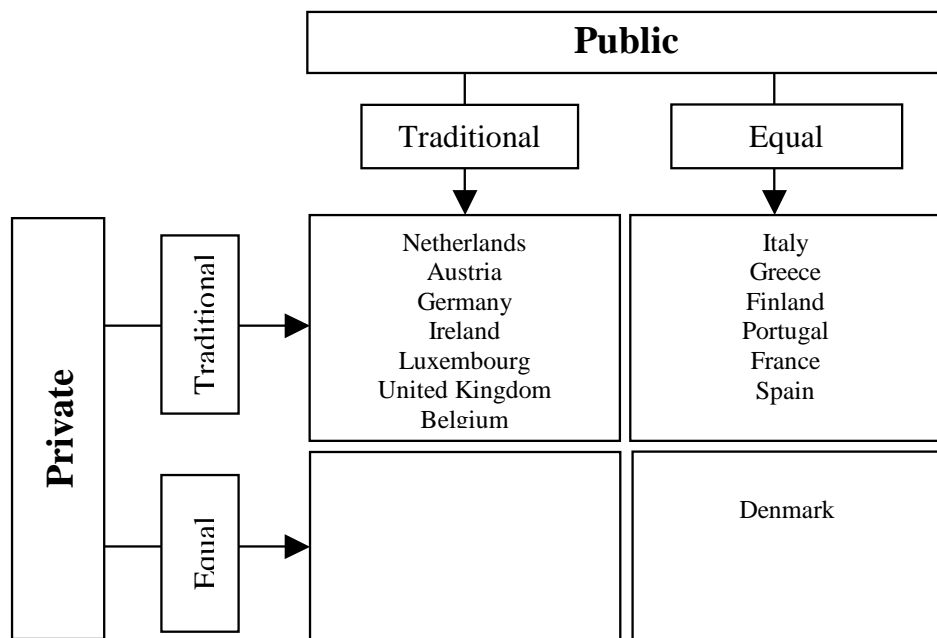
I den presentation som Kilkey gjorde i Murcia gjorde hon en fyrfältstabell och placerade in länderna utifrån vilken partyp som dominerade i varje land: se figur

Korrespondensanalys skulle kunna ge en rikare bild av förhållandena i tabellen. Nedan följer i schematisk form de steg jag behövde ta för att kunna köra korrespondensanalysen:

1. Open kilkey.xls
2. Transform/Automatic recode: v1->country

---

<sup>15</sup>Eftersom man enbart har med tvåförsörjar familjer säger dessa siffror lite eller inget om hur jämställd arbetsdelningen är på det hela taget i de olika länderna, relevansen ligger som sagt istället i att tvåförsörjar familjer ökar i antal och att de i policy dokument framställs som bra.



Figur 6: Arbetsfördelning och inkomst i tvåförsörjarfamiljer i olika länder, analys

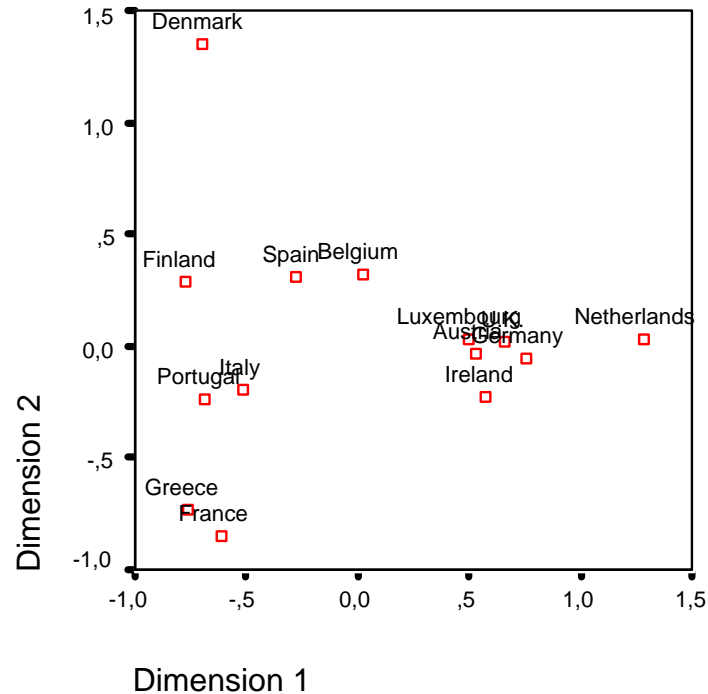
3. Variable view: ta bort v1
4. Data/Restructure: No id
5. Transform/Recode into same: trans1 if 0 -> 0.25
6. Open syntax: gender\_CA.sps
7. Data/Select cases: index = 10 | index = 14 | index = 15 | index = 18

Det sista steget innebär att enbart kolumn 10, 14, 15 och 18 används i analysen. Nedan följer både grafer med hänsyn tagen till samtliga och enbart dessa fyra kolumner:

Den ursprungliga klassificeringen stämmer bäst för Netherlands, Germany, Ireland, Austria, Luxembourg, UK i traditional public/traditional private. Belgium är det land som ligger närmast medelfallet och alltså beskrivs sämst i modellen. Speciellt Spain, men även Finland avviker från den övriga i gruppen traditional private/equal public, vars mest typiska exempel är Grekland. När man tar avgränsar analysen till att gälla de fyra vanligaste partyperna sker, vad gäller länderna, en tydlig förändring: France hamnar närmare medelfallet, vilket beror på att France hade relativt höga värden i kolumnerna

## Row Points for Country/Type

### Symmetrical Normalization



Figur 7: förhållanden mellan länder, korrespondenanalys med samtliga partyper

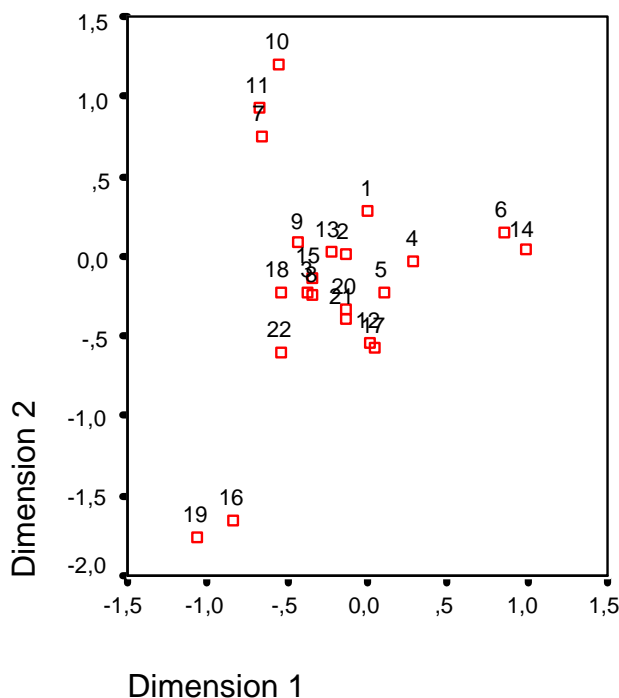
19 och 16. Vad gäller länderna innebär reduceringen till 4 partyper att en av de tre polerna byts ut? När alla partyper togs med, utgjorde 7, 11 och 10 en pol, 6 och 14 en pol och 16 och 19 den tredje polen. När vi bara tar hänsyn till 10, 14, 15 och 18 försvinner - förstås - polen med 16, 19 och ersätts med 18 och 15 istället.

## Referenser

[BBP02] Donald Broady, Mikael Börjesson, and Mikael Palme. Det svenska högskolefältet under 1990-talet: Den sociala snedrekryteringen och konkurrensen mellan lärosätena. In Thomas Furusten, editor,

## Column Points for INDEX1

### Symmetrical Normalization



Figur 8: förhållanden mellan partyper, korrespondenanalys med samtliga partyper

*Perspektiv på högskolan i ett förändrat Sverige*, pages 13–47. Högskoleverket, 2002. ISBN-91-88874-91-5.

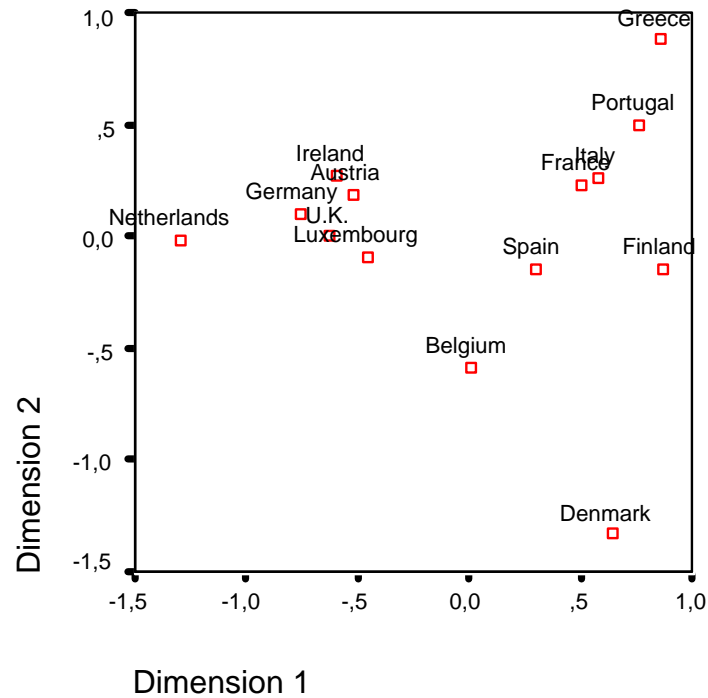
[Beh02] Eric J Beh. Simple correspondence analysis: a bibliographic review. <http://www.uws.edu.au/qmms/research/reports/20029.pdf>, 2002. QM & MS Research Reports, University of Sydney, 2002:9.

[Bla94] Jörg Blausius. Correspondence analysis in social science research. In Michael Greenacre and Jörg Blausius, editors, *Correspondence Analysis in the Social Sciences*, pages 23–52. Academic Press, 1994.

[Bou84] Pierre Bourdieu. *Distinction: a social critique of the judgement of taste*. Routledge, 1984.

## Row Points for Country/Type

### Symmetrical Normalization

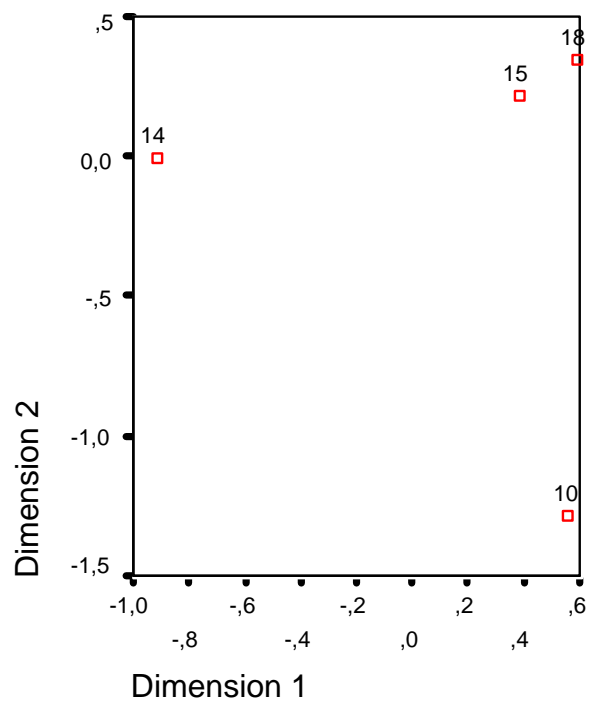


Figur 9: förhållanden mellan länder, korresponderanalys med de fyra vanligaste partyperna

- [Bro90] Donald Broady. *Sociologi och Epistemologi: Om Pierre Bourdieus författarskap och den historiska epistemologin*. HLS Förlag, 1990.
- [Cla98] Sten-Erik Clausen. *Applied correspondence analysis*. Quantitative Applications in the Social Sciences. Sage, 1998.
- [WR90] Susan C Weller and A Kimball Romney. *Metric Scaling*. Quantitative Applications in the Social Sciences. Sage, 1990.

## Column Points for INDEX1

### Symmetrical Normalization



Figur 10: förhållanden mellan partyper, korresponderanalys med de fyra vanligaste partyperna