Arbeidsforskningsinstituttene

Arbeidsfysiologisk institutt - Arbeidspsykologisk institutt - Muskelfysiologisk institutt
Yrkeshygienisk institutt

Kontoradresse: Gydas vei 8, tlf. 02/46 68 50
Postadresse: P.b. 8149 Dep Oslo 1

Tittel: Kjemometri. Rapport fra et kurs ved Universitetet i Umeå, 13. - 17. oktober 1986

Forfatter(e): Erik Bye

Prosjektansvarlig:

Prosjektmedarbeidere:

Utgiver (institutt): Yrkeshygienisk institutt

Dato: 10/11-86 Antall sider: 27 ISSN: 0800-3777

Serie: HD 935/86 FOU

Sammendrag:

Rapporten gir en sammenfatning av kursinnhold m/eksempler på anvendelse av SIMCA-metoden for Pattern Recognition.

Stikkord: Kjemometri
Multivariat dataanalyse
SIMCA
Kurs

Key words: Chemometrics
Multivariate Data Analysis
SIMCA
Course

K J E M O M E T R I

Rapport fra et kurs ved Universitetet i Umeå

13. - 17. oktober 1986

av

Erik Bye

HD 935/86 FOU

Yrkeshygienisk institutt

1986

FORORD

I perioden 13.-17. oktober 1986 deltok undertegnede på et kurs i Kjemometri, ved Universitetet i Umeå. Kurset ble gitt av Docent Svante Wold, Kemometrigruppen, Organisk kjemi, under titelen : SIMCA - MACUP.

Denne kursrapporten er laget for å gi en kortfattet beskrivelse av hva metoden går ut på, hvilke forutsetninger som ligger til grunn og hvordan metoden kan anvendes.

Kurset har gitt meget stort faglig utbytte, og kan trygt anbefales andre som arbeider med problemstillinger av kompleks natur. I et forsøk på å motivere medarbeidere til å tenke "multivariat", er det i rapporten lagt vekt på å belyse anvendelse av kjemometri innenfor problemstillinger i nær tilknytning til arbeidsmiljøforskning.

En spesiell takk rettes til Smelteverksindustriens Helseutvalg, som har bidratt økonomisk til kursdeltagelsen.

Ytterligere informasjon om kurset kan gis av :

Førsteamanuensis Erik Bye
Yrkeshygienisk institutt
tlf. (02) 46 68 50

eller ved henvendelse til :

Docent Svante Wold
Universitetet i Umeå
Tlf. (090) 16 50 00

Oslo, 10. november 1986

Erik Bye

INNHOLD

1. KURSOPPLEGG

Kurset var lagt opp med forelesninger (8.30 - 11.30), praktiske øvelser med datamaskin (13.00 - 16.00), gjennomgang av dagens tema og oppgaver (16.00 - 17.00) og mulighet for selvstendig arbeid på kveldstid.

For de som ønsket videre arbeid med egne data var det anledning til å arbeide påfølgende uke, med veiledning.

Forelesningenes innhold fremgår av Appendiks 1, se s. 15.

For forberedelse til kurset ble det sendt ut en artikkel til deltagerne (1).

Det var ialt 28 deltagere, hvorav to fra Norge og 26 fra Sverige, og kurset ble holdt på svensk. Med deltagere fra arkeologi, skogforskning, psykologi, metallindustri, odontologi, organisk , analytisk, og medisinsk kjemi, samt yrkeshygiene antydes noe om det brede anvendelsesområdet for multivariat dataanalyse etter Pattern Recognition modellen med SIMCA-metoden.

Kursavgift var kr. 4000 og 2000 for deltagere fra henholdsvis industri og forskningsinstitutter. Hver deltager fikk utlevert manual og dokumentasjon for bruk av en BASIC-versjon av SIMCA.

2. HVA ER SIMCA-METODEN ?

I det følgende vil det bli gitt en kortfattet og summarisk beskrivelse av metoden. Det vil ikke bli gitt noen beskrivelse av det matematiske verktøy som benyttes ved beregningene. Interesserte henvises til ref. 1-2 evt. andre ref. i Appendiks 2, s. 16.

Multivariat dataanalyse benyttes ved studier av komplekse systemer, med mange objekter og mange variable pr. objekt. SIMCA-metoden benyttes for å klassifisere objekter, dernest for å studere hvilke parametre som bidrar til å karakterisere klassene, og sist til å studere relasjoner mellom uavhengige og avhengige variable. Eksempler på det siste er sammenheng mellom struktur (fysiske/kjemiske parametre) og biologisk aktivitet.

Generelt brukes betegnelsen Pattern Recognition (PARC), og metoden som blir beskrevet her bygger på at multivariate data for et visst antall "like" objekter kan tilnærmes med prinsipalkomponent-modellen.Man snakker om fire forskjellige nivåer :

NIVÅ 1

Hvert objekt beskrives som et punkt i et p-dimensjonalt rom (p variable). Ved hjelp av prinsipalkomponenter (PC) forsøker en å forklare (modellere) mest mulig av variansen i data. Hovedtanken bak en slik PC-beregning og dannelse av PC-modeller er at et stort antall variable (målte) kan beskrives ved et lite antall, nemlig prinsipalkomponenter. En PC-vektor (PC1) blir beregnet ved minste kvadraters metode tilpasset variansen i data, som en lineærkombinasjon med bidrag fra alle variable. Herved beskrives mange variable ved hjelp av en ny variabel.

Neste PC (PC2) blir beregnet på samme måte , normalt på PC1, med bidrag fra alle variable slik at det gjøres rede for ytterligere varians i data. Dersom disse to komponentene f.eks. beskriver en vesentlig del av variansen, har en nå fått to variable istedetfor p variable. I det to-dimensjonale rom (plan), dannet av vektorene PC1 og PC2 (PC12), med dataene projisert ned på planet, vil en kunne se om det er noe struktur i datamassen, f.eks. om det er informasjon om klasser av objekter. På denne måten har en fått dannet et "to-dimensjonalt vindu" inn i et mangedimensjonalt rom, et såkalt score-plot. Eventuelt må flere PC beregnes.

NIVÅ 2

Ved hjelp av PC-tilpasning til de enkelte klasser av objekter og informasjon om hvilke variable som bidrar til klassemodellene kan nye ukjente objekter klassifiseres. Outliers analyseres og eventuelt utelates på dette nivå.

NIVÅ 3

Dersom relasjoner mellom datatyper skal undersøkes, f.eks. struktur og aktivitet, bygger man også opp klassemodeller for strukturdata med PC beregninger, men nå på en slik måte at også mønsteret i aktivitetsdata forklares (modelleres). Dette kalles for Partial Least Squares Modelling with Latent Variables (PLS).

NIVÅ 4

Klassemodellene kan benyttes til å klassifisere ukjente objekter som tidligere, men den totale PLS-modellen kan benyttes til å forutsi noe om biologisk aktivitet (prediksjon). Ut fra data om hvilke parametre som bidrar til relasjonen, kan også PLS-modellen benyttes til å modellere objekter til ønsket effekt (avhengig variabel).

Til studier av klassemodeller og PLS-modeller benyttes grafiske plot, noe som i vesentlig grad letter tolkningen av resultatene.

På dette tidspunkt inntreffer gjerne depresjonen,
spesielt ved første gangs møte med metoden.
Ikke gi opp, du er på rett vei.

"Lyset i den andre enden av tunnelen er ikke et motgående tog."

Spesielt ref. 1 kan være til hjelp !

3. HVA FORUTSETTER SIMCA-METODEN ?

Den viktigste forutsetningen for bruk av metoden må sies å være at samme variabel uttrykker det samme for hvert objekt, og at det er homogenitet i data. Uavhengige variable må monotont beskrive likheten innen en klasse. Idet dette kan sjekkes i PC- og PLS-beregninger er faren for feil liten.

Sterk undergruppering i en klasse bør unngås. Outliers må utelates ved klasseberegningen. Disse kan lett sees i PC-plot i NIVÅ 1, eventuelt testes med vanlige statistiske metoder.

Her kan det kanskje også være på sin plass å si noe om hva metoden ikke forutsetter, sett med den tradisjonelle eksperimentators øyne :

a. Den forutsetter ikke mange objekter, minimum 5 pr. klasse.

b. Dessto flere variable dessto bedre.

c. Det er bedre å karakterisere hvert objekt med mange "mindre presise" målinger enn med få meget nøyaktige målinger.

(Tro er fortsatt frivillig !)

Overordnet det som forutsettes og det som ikke forutsettes om data, er det to hovedprinsipper som ble understreket under kurset :

HVA VIL JEG ? (Problemformulering - valg av variable).

ALDRI EN VARIABEL AD GANGEN (EVITA - En variabel i taget)

Det kan ikke understrekes nok hvor viktig forsøksplanlegging er, og ref. 3 ble anbefalt som veiviser frem til relevante forsøk.

Et forenklet eksempel på problemet knyttet til EVITA er utbytte ved en prosess som funksjon av trykk og temperatur. I figur 1 er det skissert en flate som beskriver dette utbytte, med optimal betingelse angitt med "$". Dersom FORSØK PÅ OPTIMERING GJØRES MED T=100 ^{0}C OG VARIERENDE trykk, fås "max" ved merket "?". Dersom nå forsøket gjøres ved konstant trykk = 1 atm. får en det submaksimale utbytte merket "%" og toppen vil aldri nås.

(På kurset ble det gitt skremmende eksempler på dette !)

4. HVA KAN SIMCA-METODEN BRUKES TIL ?

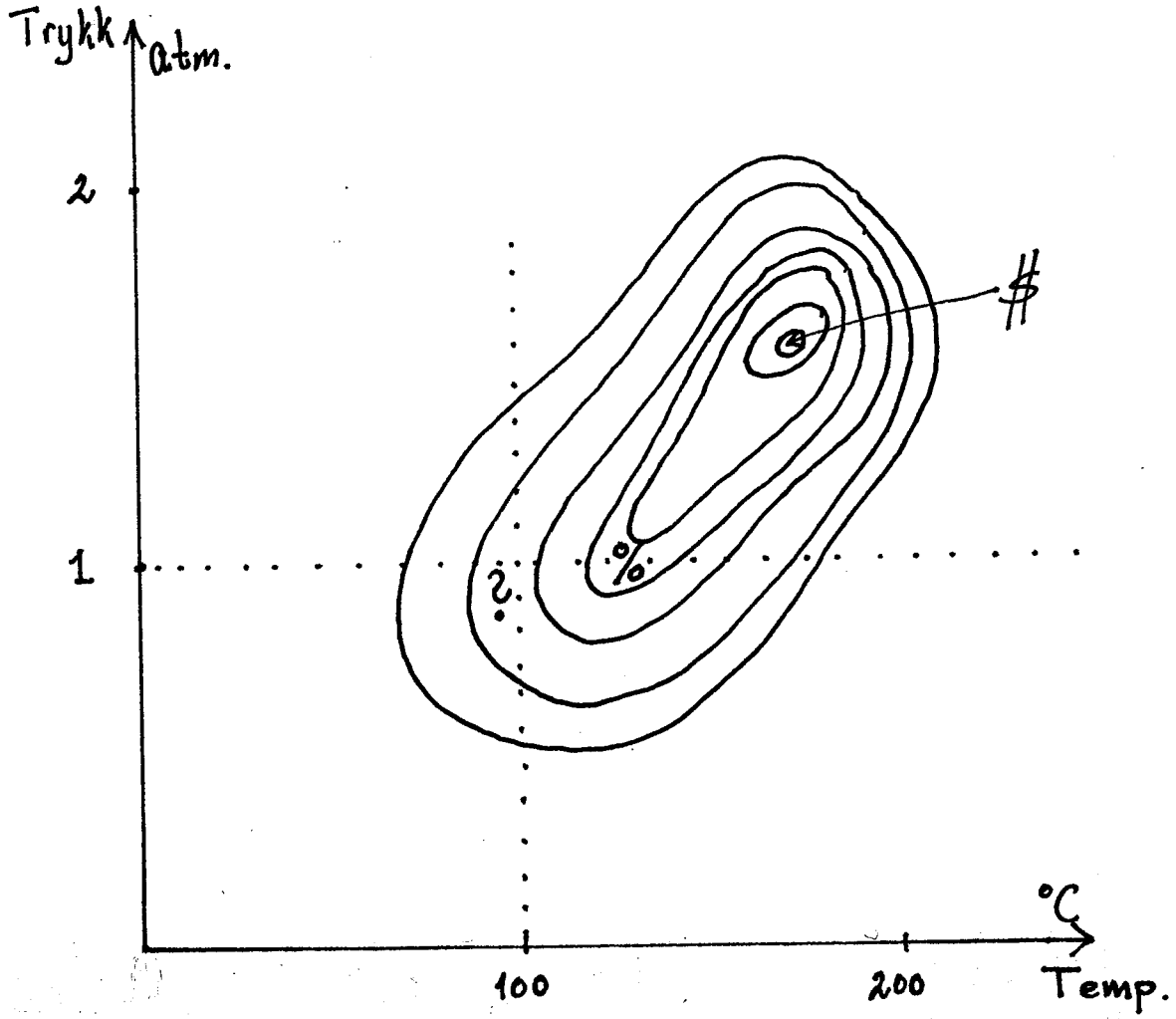
Generelt kan SIMCA-metoden benyttes til :

I. Oversikt over data (PARC 1)
II. Kalibrering (PARC 2)
III. Korrellering (PARC 3 & 4)

I. kan omfatte kontroll av data (måle-, skrive, punchefeil, "juks"), studier av datamengdens struktur, om det er flere klasser av objekter, om data inneholder outliers, planlegge videre forsøk for å dekke hele målerommet, studere hvilke parametre som bidrar til datastrukturen.

Til dette benyttes de såkalte PC-plot, som det blir gitt ett eksempel på i avsnitt 5, s.8.

II. omfatter konstruksjon av klassemodeller og kan anvendes til å kalibrere analysemetoder, til optimering av prosesser, til å klassifisere ukjente objekter til definerte klasser.



Figur 1. Utbytte ved en prosess skissert som en kolle i terrenget. Utbytte måles i "høyde over havet". Optimalisering med en parameter ad gangen er skissert med stiplede linjer.

Som konkrete eksempler kan nevnes :

- identifikasjon av kilder ved oljesøl
- klassifisering av gjenstander ved arkeologiske utgravninger
- kalibrere kvantitativ bestemmelse av protein i korn
- studere sammenheng mellom biologiske prøver og f.eks. arbeidsmiljø

Igjen benyttes grafiske plot som hjelpemidler til tolkning av resultater. Her henvises til ref. 1-2 for detaljerte studier.

III. Som tidligere nevnt vil en i siste del av analysen (PLS) forsøke å studere relasjoner mellom ulike typer variable (uavhengige og avhengige). Sistnevnte er av typen helseeffekter, biologiske effekter, egenskaper hos matvarer, metaller og maling. Forutsetningen er at en har kvantitative mål for slike egenskaper. Her kan både kontinuerlige og diskrete data anvendes. Ut fra klassemodeller vil en så prediktere slike variable ut fra målinger av uavhengige variable. Her kan (for å gjøre det ytterligere forvirrende ?) også uavhengige variable behandles som avhengige, og det kan gjøres forsøk på prediksjon, dersom det er noen relasjon mellom objektenes uavhengige variable.

Konkrete eksempler på slike arbeider er :

- sammenheng mellom struktur og Ca-antagonister (6).
- sammenheng mellom gasskromatografidata og lagrede/ulagrede matvarer (1,4).
- sammenheng mellom struktur og cancerogen effekt for PAH-forbindelser (7).
- sammenheng mellom struktur og mutagen effekt for halogenerte hydrokarboner (8).
- optimalisering av prosesser med tilhørende prediksjon av kvalitet ut fra målte parametre. Prediksjon av nødvendige parameterverdier for å få ønsket kvalitet ut fra prediksjon.

Det siste eksemplet er tatt med i generell form for å antyde at ved gode datasett er det mulig å prediktere fra uavhengige til avhengig parametre og omvendt. I referanselisten gitt i Appendiks 2 (s. 16) vil en finne eksempler på ovennevnte undersøkelser, samt en rekke andre interessante anvendelsesområder.

5. *EKSEMPLER*

Med to eksempler vil jeg anskueliggjøre PC-beregninger og klassemodeller,og outliers. Dette gjøres med data og grafiske plot fra arbeider som ble omtalt ved kurset.

Eksempel 1

FERSKE OG LAGREDE KÅLRØTTER (ref. 4)

Gasskromatografidata fra prøver av 7 ferske og lagrede kålrøtter er analysert ved hjelp av SIMCA-metoden for å undersøke :

a. Er det forskjell på de to typer kålrøtter ?

b. Kan kålrøtter klassifiseres ved hjelp av GC-målinger ?

c. Kan egenskaper predikteres ?

Tabell 1 gir oversikt over analysedata, der parametrene er log-transformert.

Databehandling :

i) Data skaleres slik at hver variabel har samme vekt : multipliser med 1/st.avvik (analogt med grafisk fremstilling og akseenheter).

ii) Middelverdien (x) for hver variabel bestemmes og $x' = x - \bar{x}$ beregnes for hver variabel.

To prinsipalkomponenter beregnes som lineærkombinasjoner av alle objektenes bidrag.

Figur 2a viser et plot av PC1 mot PC2 (PC12-plot) , der objektene er gitt ved nr., score-plot.

Et objekt (nr.7) faller langt utenfor de andre, tolkes som outlier, og holdes utenfor en ny PC12 beregning, med plot vist i Figur 2b.
De ferske og lagrede kålrøtter faller i to klasser, med kun to objekter som det kan være litt tvil om (nr. 2 og 12).

I programmet finnes statistiske metoder for test om signifikans og klassetilhørighet, uten at jeg vil gå inn på det her.

For PLS-beregninger for de separate klasser henvises til ref. 1.

Figur 3 (s.13) viser et såkalt Cooman's plot, der hvert objekt er inntegnet med avstand til begge klasser.

Table 1. The mean peak height for seven swede cultivars, fresh and stored, according to Cole and Phelps (1979). The last two rows are the average of all fourteen cultivars, fresh and stored, respectively. The first character of the sample name indicates the class, F=fresh or S=stored. The second character is the cultivar type according to Cole and Phelps (1979).

No	Name	1	2	3	4	5	6	7	8
1	FH	.37	.99	1.17	6.23	2.31	3.78	.22	.24
2	FA	.84	.78	2.02	5.47	5.41	2.8	.45	.46
3	FB	.41	.74	1.64	5.15	2.82	1.83	.37	.37
4	FI	.26	.45	1.5	4.35	3.08	2.01	.52	.49
5	FK	.99	.19	2.76	3.55	3.02	.65	.48	.48
6	FN	.7	.46	2.51	2.79	2.83	1.68	.24	.25
7	FM	1.27	.54	.9	1.24	.02	.02	1.18	1.22
8	SI	1.53	.83	3.49	2.76	10.3	1.92	.89	.86
9	SH	1.5	.53	3.72	3.2	9.02	1.85	1.01	.96
10	SA	1.55	.82	3.25	3.23	7.69	1.99	.85	.87
11	SK	1.87	.25	4.59	1.4	6.01	.67	1.12	1.06
12	SB	.8	.46	3.58	3.95	4.7	2.05	.75	.75
13	SM	1.63	1.09	2.93	6.04	4.01	2.93	1.05	1.05
14	SN	3.45	1.09	5.56	3.3	3.47	1.52	1.74	1.71
15	F-AVG	.62	.72	1.48	4.14	2.69	2.08	.45	.45
16	S-AVG	1.55	.78	3.32	3.2	5.75	1.77	1.04	1.02

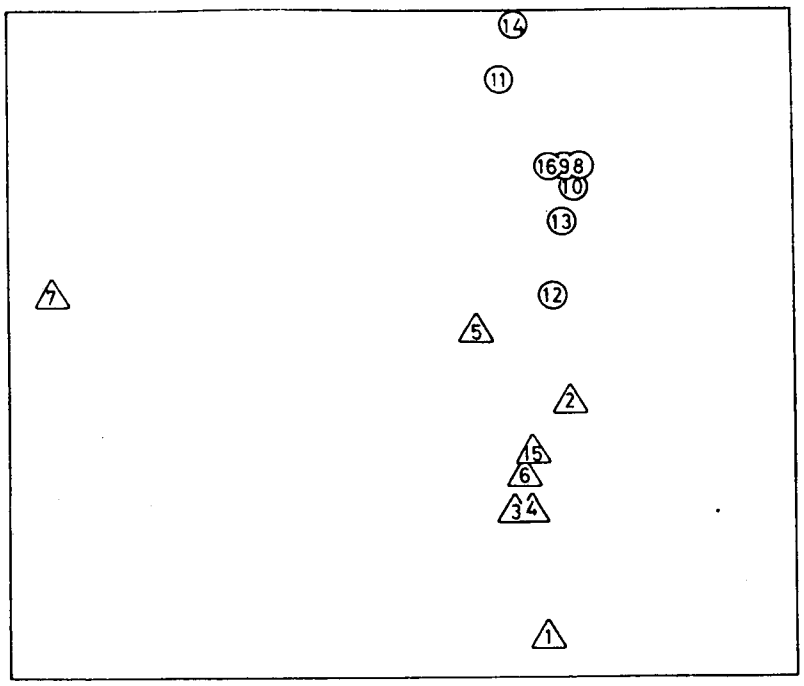


Figure 2a (above) and 2b (below). PC12 plots of the 16 swede samples (above) and of the 15 swede samples with no. 7 removed (below).

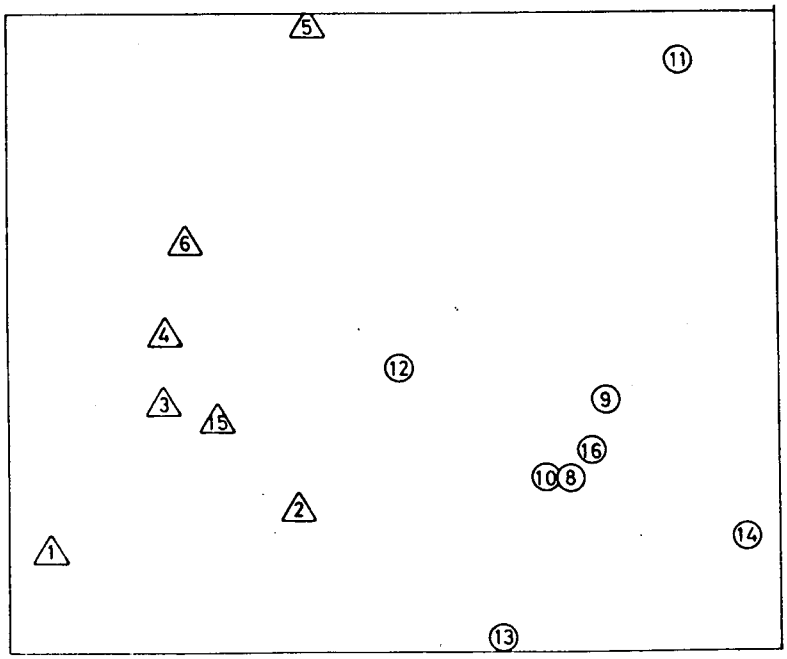
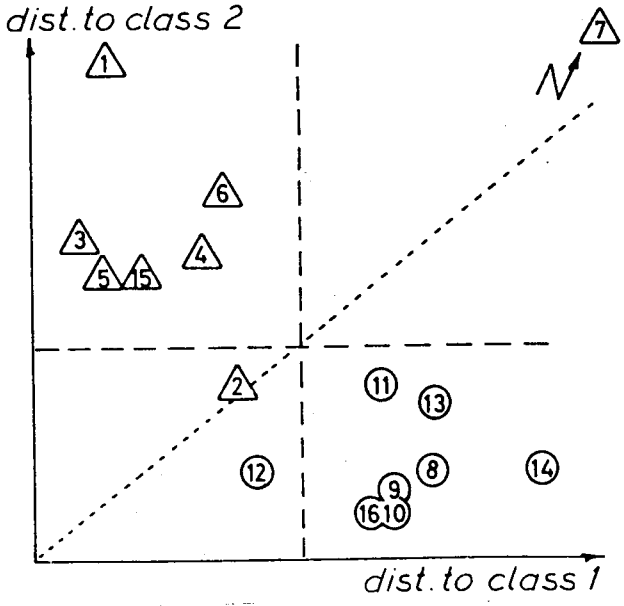


Figure 3. A class distance plot (Coomans plot) for the swedes when both classes have been modelled by one dimensional PC models. The dashed horisonal and vertical lines indicate 95 % tolerance intervals and the 45° line indicates equal class distance. For all objects the class distances differ significantly but for no. 2 and 12, which are not clearly classified by the given data. Compare fig. 4b.



Wold et al. 1983 (1)

EKSEMPEL 2

BIOLOGISKE PRØVER OG ARBEIDSMILJØ HOS SVEISERE (ref. 5)

I et forsøk på å studere arbeidsmiljø for sveisere ved hjelp av elementanalyse av blodprøver, er det samlet inn fullblodprøver fra sveisere som arbeider med forskjellige sveisemetoder, forskjellig metall samt fra kontroller. Opplysninger om elementsammensetningen ved de ulike sveisemetoder og sveiseoppgaver viste betydelig variasjon. Ialt ble fem grupper (gruppe 1 -5) studert, inkludert kontroller, og følgende kan sies kort om elementkonsentrasjoner og sammensetning :

Gruppe 1 : kontroller (68 pers.)
Gruppe 2 : Rustfritt stål - dekkelektroder; Ni og Cr 10 - 20 %; betydelig mindre for andre elementer (23 pers.)
Gruppe 3 : Rustfritt stål - TIG; elementer som over (7 pers.)
Gruppe 4 : Aluminium - MIG; Mg 1 - 5 %; Si/Zn tilstede (28 pers.)
Gruppe 5 : Aluminium - MIG; elementer som over (23 pers.)

Eksponering : Rustfritt stål : 5 mg/m3 tilsvarende Ni:0.03 og Cr:0.3 mg/m3.

Aluminium : 12 mg/m3

(Alle målinger tatt under sveisemaske).

Problemstilling :

Er det forskjell i elementsammensetningen i de analyserte blodprøver ?

Table 1 s. 13 viser data for blodanalysen, og ved hjelp av vanlig prosedyre for SIMCA-metoden fikk man et PC12-plot som vist i Fig. 3 s. 14. Selv om det kan være vanskelig å skjelne de plottede punkter, angitt med gruppe nr 1 til 5, viste dette arbeidet at blodprøvene ikke kunne klassifiseres ut fra elementsammensetningen. Dette medfører at blodprøvene heller ikke kunne si noe om arbeidsmiljøet, at data ikke innholdt nok informasjon om de ulike klasser. Om dette skyldtes utvalg av elementer eller feil organ ble ikke utredet i artikkelen.

For ytterligere informasjon om arbeidene i disse eksempler henvises til ref. 1, 4 og 5.

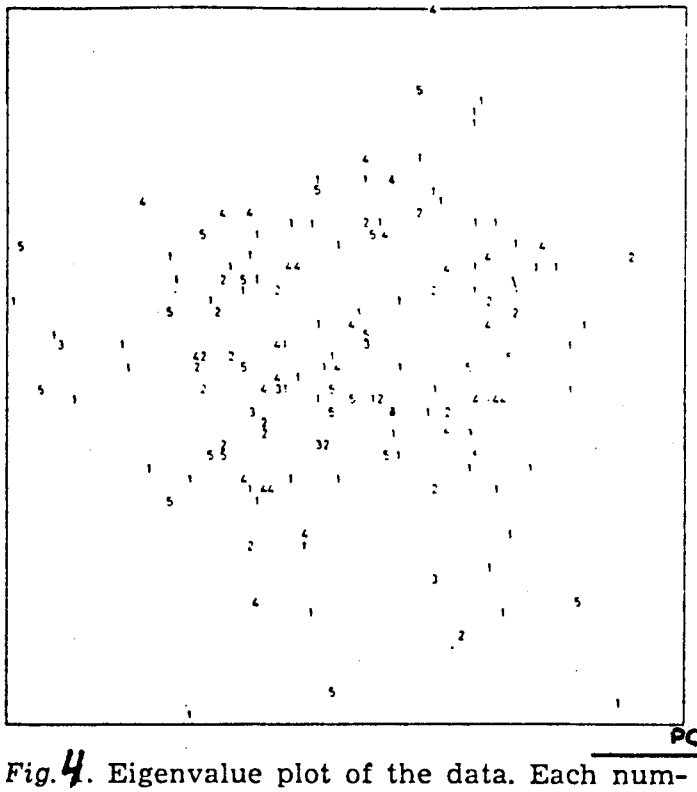


Fig. 4. Eigenvalue plot of the data. Each number in the plot corresponds to one individual in that class. The eigenvalue plot preserves as much of the variance in the original 17-dimensional space in the projection down to the 2-dimensional space of the plot.

Table 1. Properties of the data. All variables were transformed by $y \rightarrow \log_e (1 + ay)$. Normal range of each element based on the 95 % confidence interval of log (1 + ay).

Variable no.		Name	Typical value[a] μg/g wet tissue	Min	Max	Skewness of log (1 + ay)	a	Mean[b] log (1 + ay)	SD[b] log (1 + ay)
1	Pb	Lead	0.04	0	1.870	—1.1	1,000	3.664	1.935
2	Sr	Strontium	0.04	0.007	0.27	—1.8	1,000	3.825	0.896
3	Rb	Rubidium	10	3.5	25	0.6	1	2.375	0.439
4	Br	Bromine	0.2	0.06	0.75	—0.3	1,000	5.349	0.636
5	Ga	Gallium	0.025	0.0004	0.51	—1.2	1,000	3.270	1.482
6	Zn	Zinc	1.7	0.63	4.2	0.6	10	2.882	0.444
7	Cu	Copper	0.7	0.35	1.4	—0.4	10	2.113	0.301
8	Co	Cobalt	0.09	0.016	0.49	—1.9	1,000	4.520	0.835
9	Fe	Iron	520	270	980	—0.4	1	6.247	0.320
10	Mn	Manganese	0.06	0.009	0.37	—2.5	1,000	4.118	0.901
11	Cr	Chromium	0.03	0.002	0.34	—1	1,000	3.443	1.197
12	Ca	Calcium	47	27	83	0.0	1	3.873	0.275
13	K	Potassium	2,500	1,500	4,000	—1	1	7.811	0.247
14	S	Sulfur	390	200	760	0.1	1	5.965	0.333
15	P	Phosphorus	540	290	1,010	0.2	1	6.295	0.314
16	Si	Silicon	6	1.4	18	0.2	1	1.913	0.524
17	Mg	Magnesium	22	8	60	—2	1	3.135	0.483

[a] Typical values [from mean of log (1 + ay)].

[b] Data scaled by subtraction of mean and subsequent division by the standard deviation. Hence, the original data are obtained from the transformed data as

$$Y_{orig} = \{\exp[y_{tr} \cdot SD + Mean] - 1\}/a.$$

Henlet fra Ulfvarsson & Wold (1977)

6. OPPSUMMERING

Kurset i SIMCA-metoden ga en meget god og instruktiv innføring i multivariat dataanalyse, spesielt med sikte på klassifisering av objekter. Det ble lagt vekt på de mulighetene denne metoden gir til å studere mange parametre simultant, ved forsøksplanlegging , ved prosessoptimalisering og dataanalyse.

Det gode utbytte av kurset skyldes først og fremst den pedagogiske fremleggelse av stoffet, men også oppdeling i forelesninger og praktisk bruk av metoden gjennom hele kursuken.

7. REFERANSER

1. S. Wold et. al : Pattern recognition : Finding and using regularities in multivariate data. In : Food Research and Data Analysis , Ed. : H. Martens & H. Russwurm Jr., Applied Science Publishers, London, 1983.

2. S. Wold et al. : Multivariate Data Analysis in Chemistry. Proceedings NATO Adv. Study Inst. on Chemometrics, Cosenza, Italy, Sept. 1983 Ed. : B. R. Kowalski. Reidel Publ.Co., Dordrecht, Holland, 1984.

3. G.E.P. Box et al. : Statistics for experimentators. Wiley, New York 1978.

4. R.A.Cole et al : Use of canonical Variate Analysis in Differentiation of Swede Cultivars by GasLiquid Chromtography of Volatile Hydrolysis Products. J. Sci. Food Agric. 30 (1979) 669 - 676.

5. U. Ulfvarsson et al. : Trace-elemental Concentrations in Blood Samples from Welders of Stainless Steel or Aluminium and a Reference group. Scand. J. Work Environ. Health 3 (1977) 183 - 191.

6. P. Berntsson et al. : Comparison Between X-ray Crystallographic data and Physicochemical Parameters with respect to Their Information about the Calcium Channel Antagonist Activity of 4-Phenyl-1,4-dihydropyridines. Quant. Struct. Act. Relat. 5 (1986) 45 - 50.

7. B. Norden et al. : Prediction of the Carcinogenic Potency of Eleven Polycyclic Hydocarbons (PAH) having a Bay Region.
Quant. Struct. Act. Relat. 2 (1983) 73 -76.

8. U. Edlund et al. : Prediction of Cancerogenic and Mutagenic Potencies Using the PLS Method. In : Extrahepatic Drug Metabolism and Chemical Carcinogenesis (J. Rydstrøm et al. Eds.). Elsevier Science Publishers (1983).

KEMISKA INSTITUTIONEN
AVD. FÖR ORGANISK KEMI
UMEÅ UNIVERSITET
UMEÅ

PLAN för kursen i multivariat kemisk data-analys (VECKA 1):

Varje förmiddag föreläsning kl 8.30-11.30 (måndag 10-12.30)
Fikapaus 10.00-10.30
Tiderna är exakta, ingen akademisk kvart eller dyl.

Varje eftermiddag: Övningar och diskussioner 13.00-16.00. Genomgång 16.00
Fika kl 14.30.

~~Fest~~ Onsdag kväll.

Föreläsningarnas innehåll:

Måndag: Grunderna för multivariat data-analys (MVDA)
Data-tabeller, objekt, variabler, notation.
Observationsrymden (m-space), projektioner, bilder.
Tvådimensionella fönster i m-space.
Principal-komponent-analys (PCA), faktor-analys (FA)
Tolkning: projektioner, likhetsmodeller
Kors-validering
Representation, A-variabler, kontinuitet, homogenitet
Score-plot (tt), laddningsplot (bb)
En var i taget, jfr design och optimering
Exempel: Tumor, Ketone

Tisdag: Pattern recognition (PARC)
Tränings set, test set, klasser
Likhet eller olikhet, olika metoder för PARC
SIMCA, inkl. asymmetriska fallet
Antalet variabler
Transformationer, skalning, representation, A-variabler
Avvikare, homogenitet
Variablers relevans, selektion?
Design, optimering, försöksplaner
Typiska tillämpningar i olika områden
Exempel: Ketone, Arch, Swedes.

Onsdag: Problem → data. GC, HPLC, ..., NMR, MS, kurvformer, halter, geokemi, kinetik, tidsserier, ...
PARC, level 3 och 4.
PLS
Design
Typiska tillämpningar: Kalibrering, smak, struktur-egenskaps-samband, optimering
Exempel: Multivariat kalibrering, QSAR (betab2)

Torsdag: Problem-formulering (ex. oil-spills)
Representation, A-variabler, skalning, homogenitet
Degrees of freedom, beroende mellan objekt, design
Detaljerad genomgång av ett exempel (olje-spill)

Fredag: Repetition: MVDA (bilder, proj.), VAD VILL JAG?, PCA, FA, representation, design, skalning, SIMCA, KNN, asymmetriska fallet, antal variabler, PLS.

Avancerade tillämpningar, smooth, 3 dim arrayer, var.cluster,...

Vecka 2 Fria övningar med handledning.
En timme gemensam diskussion varje dag enl. önskemål.

****
*
* SIMCA 3B Copyright: Umeå University and SEPANOVA AB *
*
****

Manual, 2.nd edition, Oct.15, 1983.
Research Group for Chemometrics, Umeå Univ., 901 87 UMEÅ, Sweden.
SEPANOVA AB, Östrandsv. 14, S-122 43 ENSKEDE, Sweden.

Contents

1. General advice

Whenever something goes wrong (due to errors in the program, of course, never due to your own mistakes), press CTRL C on CP/M machines (END, CR on ABC-80) before doing anything else. This closes all files. Otherwise you can easily cause errors in the library of your B: disk (the data disk) and then you are really in trouble unless you have a backup disk. On ABC-80 this is the DR1: disk.

ALWAYS have backup disks for both programs and data (on separate disks). Create backup files on another disk immediately after data input (FINP or FINWS).

1.1 Hardware.

To run SIMCA-3B (in BASIC) you need at least a 64 K 8-bit CP/M microcomputer with two floppy disk drives and a printer. SIMCA-3B also runs on the Swedish ABC-80 micro extended to 32 K memory with a double floppy. Versions running on 8088 machines, e.g. IBM PC, are soon released as well as Fortran versions for PDP-11, VAX, PRIME and other larger machines. A M68000 version is under development in C.

Harddisks and Winchester drives are, of course, accepted by the SIMCA system. In CP/M the hard disk is usually divided into pseudo drives (A:, B:, etc.). With other operative systems such as UNIX and the like, there is no distinction between different types of secondary memory.

Depending on the computer and operative system you are running SIMCA on, file handling and printer activation is slightly different. Thus on CP/M machines the printer is activated by CRTL P. Any output to the screen is then "echoed" on the printer. FLIST and FPLOT have their printer activation made from the program, so for those modules CTRL P should not be given before the modules are entered.

On the ABC-80 micro, the printer is either hooked up on the PR: interphase or on the V24: interphase. In the former case, nothing needs to be done except selecting the PR: option in DEFINE (menu -3). With the V24: interphase a PRINTER driving program must be run before the printer becomes operative. See your ABC-80 manual.

With UNIX like operative systems (UNIX, CROMIX, etc.), the screen output can be "teed" to the printer by a suitable command. See your operative system manual.

1.2 Introduction to SIMCA

The SIMCA-MACUP methods analyse multivariate data, i.e. data tables, data matrices (see figure 1) in various ways. Such data are obtained by measuring or observing or otherwise getting the values of p or M "variables" on n or N "objects". The articles enclosed as appendix give a general introduction to multivariate data analysis in chemistry and neighboring fields.

OBJECTS: These often are analytical "samples", or chemical compounds or chemical reactions or processes or biological individuals.

VARIABLES: The variables often are derived either from spectroscopy or from separation methods such as gas and liquid chromatography (GC and LC).

SIMCA (Soft Independent Modelling of Class Analogy) is based on fitting principal components (PC) models separately to separate classes of objects (cases, individuals, ...). New objects can then be compared to one, several or all of the class models. In this way they can be assigned to a class if they are sufficiently "close" to one of the class models.

MACUP is an acronym for Modelling And Classification Using PLS. With the PLS method (partial least squares modelling with latent variables), models can be developed for the quantitative prediction of one block of variables in the data matrix (then denoted Y) from another block of variables, denoted X. These models can simultaneously be used for ordinary SIMCA classification on the basis of the X-variables.

The package also contains facilities for multivariate projections and plots and for K-Nearest-Neighbor analysis.

The reader is recommended to study the appendix (ref.s 3 c and 3d) and the other references below; here only a very brief idea of the methods and related concepts is given.

DATA TABLE: The data analysis concerns a table (matrix X) of data xki obtained by observing the values of p variables (index i) on n objects (index k). The objects may be divided into a training set containing G classes and a test set of unassigned objects. In PLS analysis, the data matrix is divided also variable wise into one block of p "predictor" variables, X, and one block of q "predicted" or "dependent" variables, Y. See figure 1.

In the programs and earlier SIMCA texts, M is often used instead of p, and N instead of n. To add to the confusion, after the Cosenza 1983 Chemometrics meeting, i and k are reversively used as object and variable indices, respectively. This standard notation will be gradually introduced in all SIMCA programs.

P-SPACE (M-SPACE): To understand the data analysis, we represent each object as a point in an p- (or M-) dimensional space obtained by letting each variable define one orthogonal coordinate axis. We refer to this space as p-space or M-space. In general, when $p > 3$ ($M > 3$), this space is difficult to visualize. We therefore use 3-spaces as models (see figures 2 and 3), remembering that p-spaces have analogous properties to the 2 and 3-spaces we can actually see and touch.

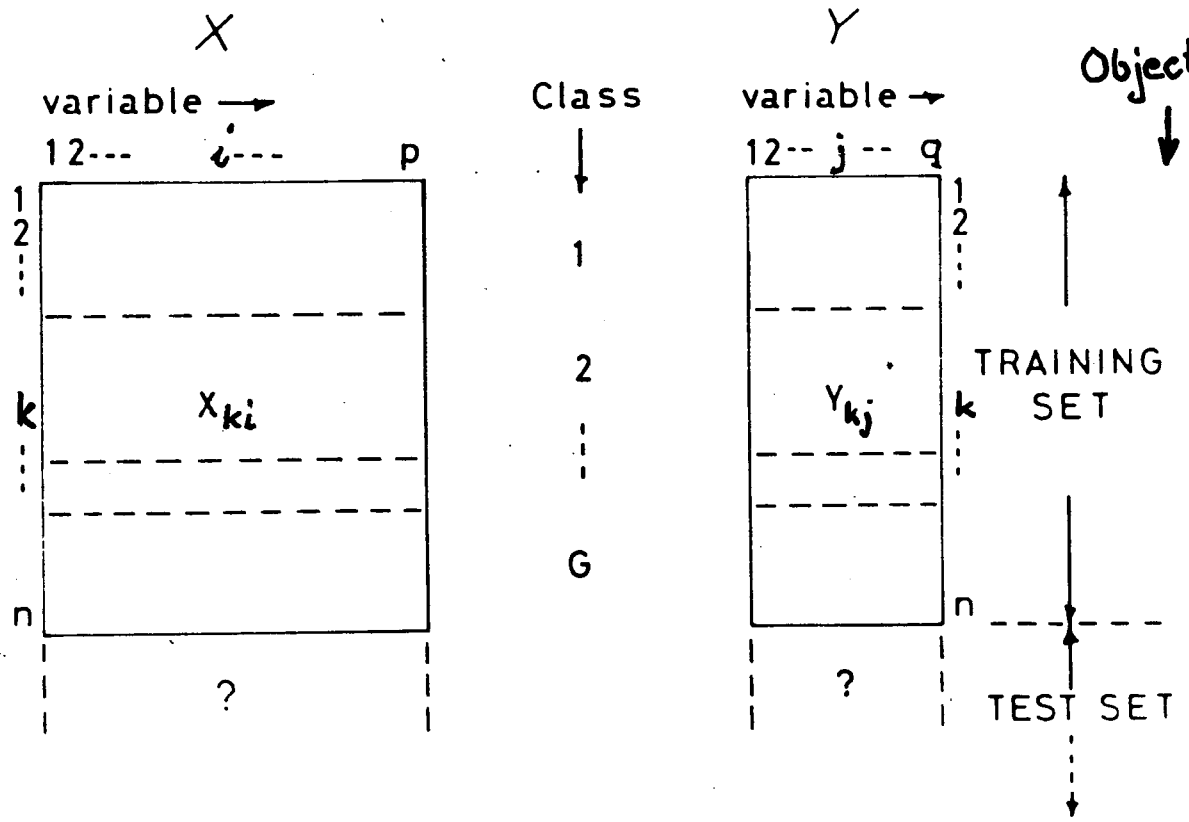


Figure 1. The data in a generalized pattern recognition problem are organized in one or two tables, matrices. The matrix X contains characterizing "independent" variables, e.g. the measured concentrations of various constituents according to GC and HPLC, the light absorbtion at a number of wavelengths, pH, conductivity, etc.

In PARC levels 3 and 4, we have also a matrix of "dependent" variables with one (PARC 3) or several columns (PARC 4). These variables may be quantitative measures of taste (e.g. panel data) or, in the multivariate calibration problem, concentrations of "protein", "fat", etc., measured by slow wet-chemistry methods.

The data X and Y are divided into two sets: (1) the training set (calibration set) used to develop the typical "data patterns" of the represented classes (1,2,...,G; can be one, two or several) and (2) the test set containing objects of unknown class assignment and, on levels 3 and 4, with unknown Y-values.

Figure 2. The data vector of one object (the x part) is represented by a point in the p-dimensional space spanned by the p x-variables. Though spaces with p > 3 are difficult to visualize, three dimensional spaces can be used as illustrations. Points, lines, planes, angles and distances have the same properties in spaces with many dimensions as in those with two or three.

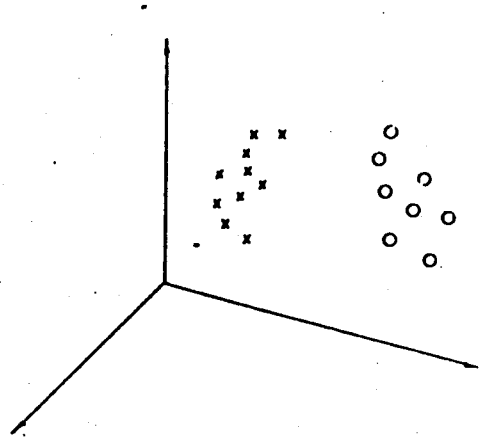
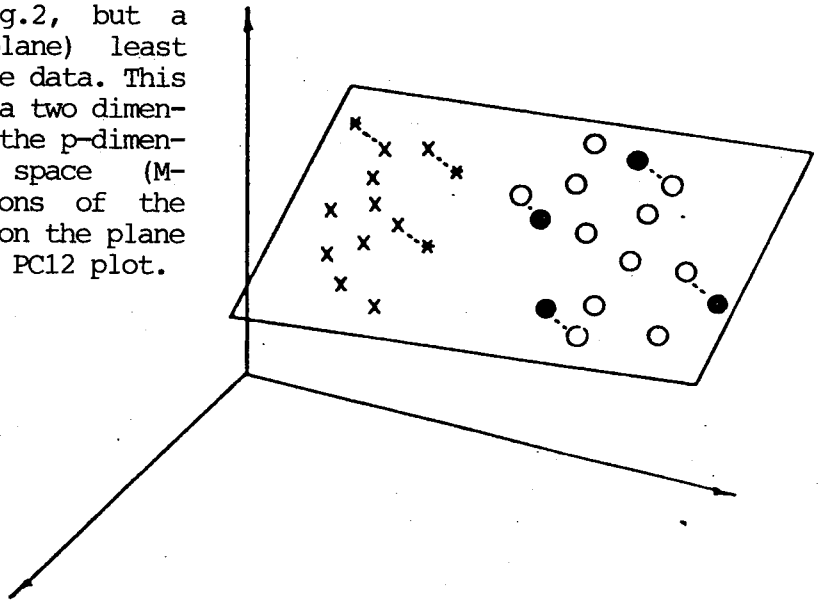


Figure 3. Same as fig.2, but a plane (the PC12 plane) least squares fitted to the data. This PC plane constitutes a two dimensional window into the p-dimensional measurement space (M-space). The projections of the object points down on the plane are visualized in the PC12 plot.



PARC and TRAINING SET: The idea with pattern recognition (abbreviated PARC) is to describe the position and extension of each class as inferred from object points "known" to belong to the classes, which together form the training set or the reference sets.

TEST SET: When the PARC models have been "calibrated" on the training set, new objects are then assigned to the classes on the basis of their similarity to the various classes. These new objects are called the test set.

WINDOWS IN P-SPACE, PROJECTIONS: The object points can be linearly projected down on a plane (figure 3). The plane can then be lifted out and plotted or shown on the computer display. Thus one can construct two-dimensional windows into p-space. Different planes give different projections, different windows. The commonly used singular vector projection (often called eigenvector projection), which is equivalent to a principal components projection, produces the "window" which fits the data set best in the least squares sense; the window shows as much of the variation in p-space as possible.

SIMCA CLASS MODEL DERIVATION: The scope of the SIMCA method is to derive a description of each class to allow the classification of "new" objects in the test set. This is done by fitting principal components (PC) models to each class reference set. The PC models range in complexity from a point (A=0), through a line (A=1), to an A-dimensional hyperplane. The adequate complexity of a class model (A) is determined by crossvalidation (ref.4).

In p-space, this corresponds to the approximation of each class by a linear geometrical structure. Using simple statistical arguments, one can then construct a tolerance interval around each class model. This corresponds to setting a maximal allowed residual standard deviation (RSD) for each object. In p-space this gives a cylinder around each class (figure 4), in general a hyper-cylinder around the class hyper-plane.

The cylinder is cut off at the top and the bottom on the basis of the range of the class object points as projected down on the model.

SIMCA CLASSIFICATION: The objects in the test set can now be classified by relating them to each class model. Numerically, the models are fitted to the data vectors by multiple linear regression. This gives for each object one RSD for each class model.

The object is likely to belong to class g if the RSD is smaller than, say, twice the class RSD. Approximate F-tests can be used for a more precise estimation of the probability of class belonging. In p-space this corresponds to the object being inside the class cylinder. Hence an object can be assigned to a single class (unique classification), several classes (ambigous classification) or none of the classes (outlier).

The classification can be made on the basis of RSDs calculated with residuals weighted according to their relevance in the class modelling (ref. 3 c). This makes the "polishing" of class models by deleting irrelevant variables less important.

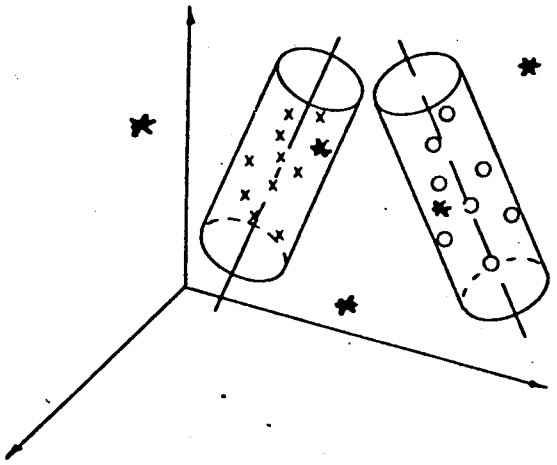


Figure 4. Same as figures 2 and 3, but with one-dimensional PC models (lines) fitted separately to each class. Tolerance regions have been constructed around each class (cylinders) on the basis of the scatter of the training set points around the models. Test set points (asterisks) are classified according to their positions inside or outside class cylinders.

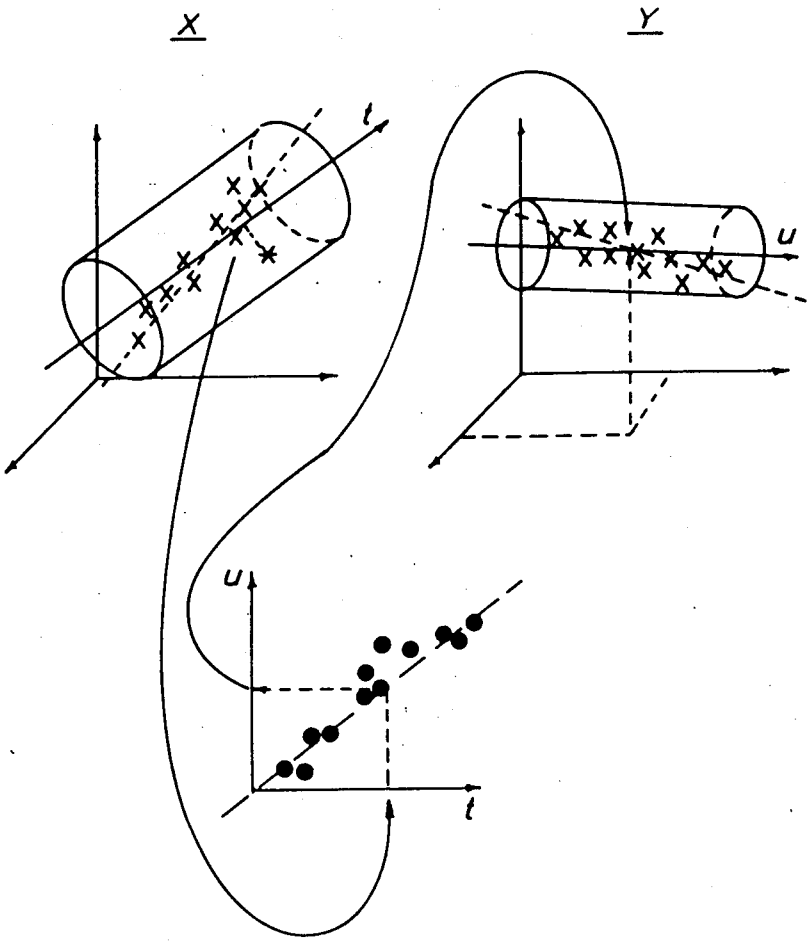


Figure 5. In PARC levels 3 and 4, relations between the X-block and the Y-block of variables are constructed separately for each separate class. With the PLS method, this can be illustrated as relations between constructs in one space for the x-data and one space for the y-data. On level 3 with only one y-variable, the latter space is one-dimensional.

In the training phase, models similar to PC models are constructed that approximate the X and Y data of a class. These models are tilted slightly to improve the correlation between the classes in terms of the correlation between the projection t and the projection u (lower part of the figure). Tolerance intervals (cylinders) are constructed just as in SIMCA level 2 (fig. 5).

In the "test phase", the prediction phase, a new object (asterix) is classified as similar to the class or not according to its position inside or outside the class tolerance interval in X-space. The position of the object point along the x-part of the model, i.e. the coordinate t, is introduced in the u-t-plot (lower part of the picture) which gives a predicted u-value for the object. This is taken into the model in Y-space to give predicted values of each of the y-variables.

RELEVANCE OF VARIABLES: The residuals in the modelling phase and the classification phase can be used to calculate measures of relevance for the variables k, their modelling power (see ref.s 1-3 and CPRIN below). OBSERVE that in PLS one must not polish the X-matrix by variable deletion. This seriously increases the risk for spurious results.

PLS: In case one has for one or several classes also "dependent" variables, Y, the data analysis has two objectives. (1): As before, to use the data X to develop class models allowing the classification of new objects as similar or dissimilar to each class. (2): To develop, for each class with Y-data, models allowing the prediction of Y from X.

These two objectives are accomplished simultaneously by the PLS method as described in ref. 3 c. The analysis can be seen as very similar to the PC modelling in SIMCA, but the PC models are "tilted" slightly to improve the prediction of the regularities in Y-space from those in X-space (figure 5).

SCALING: The PC and PLS analyses give different results with different weightings of the variables. Usually one of two standard weightings is used (the second is recommended). (i) Autoscaling where each variable is scaled to have unit variance over the training set (done with FSCAL). (ii) Class scaling where each variable is scaled to have unit variance over one class (done in CLOAD). See also ref.31.

When prior information is available about the relative importance of the variables for the given problem, the variables should be given weights proportional to this relative importance.

K nearest neighbors analysis (KNN): This method is included in the package since it forms a useful complement to the SIMCA method of PARC. Here each object in the training set is classified according to the classes of its K nearest neighbors (usually K = 1 or 3). The results are also dependent on the data scaling and the same scalings are recommended as above.

1.3 References

The statistics and mathematics of the SIMCA method are described in the following articles:

1. S. Wold. Pattern Recognition by means of Disjoint Principal Components Models. Pattern Recognition 8 (1976), 127-139.

2. S. Wold and M. Sjöström. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In Chemometrics, Theory and application (B.R. Kowalski, Ed.) ACS symposium series no. 52, 1977.

3.a. C. Albano et al. Four levels of Pattern Recognition. Anal. Chim. Acta. 103 (1978) 429-43.

3.b. C. Albano et al. Characterization and classification based on multivariate data analysis. In Proceedings 27.th IUPAC (A. Varmavuori, Ed.), Pergamon Press, Oxford, 1980.

APPENDIX 1
3.c. S.Wold et.al. Pattern recognition: Finding and using regularities in multivariate data. In Food research and data analysis (H.Martens and H.Russwurm Jr., Ed.s), Applied Science, London 1983.

APPENDIX 2
3.d. S.Wold et.al. Pattern recognition by means of disjoint principal components models (SIMCA). Philosophy and methods. In Proc. Symp. Applied Statistics, Copenhagen, Jan.1981 (A.Höskuldsson, Kim Esbensen et.al., Ed.s), NEUCC, Copenhagen 1981.

4. S. Wold. Cross validatory estimation of the number of components in factor and principal components analysis. Technometrics 20 (1978) 397-406.

Other pertinent references describing multivariate projections are

11. R. Gnanadesikan. Methods for statistical data analysis of multivariate observations. Wiley, N.Y., 1977.

12. B.R. Kowalski and C.F. Bender. Pattern recognition. A useful and powerful approach to interpreting chemical data. J. Amer. Chem. Soc. 94 (1972) 5632 and 95 (1973) 686.

13. K.V. Mardia, J.T. Kent and J.M. Bibby. Multivariate analysis. Academic Press, New York 1979.

The K-Nearest-Neighbor method is described in, for instance, the following references. These also review pattern recognition in general. Ref. 24 is recommended for chemists.

21. K. Fukunaga. Introduction to statistical pattern recognition. Academic Press, N.Y., 1972.

22. H.C. Andrews. Introduction to mathematical techniques in pattern recognition. Wiley, N.Y., 1972.

23. P.H.A. Sneath and R.R. Sokal. Numerical Taxonomy. Freeman and Co., San Fransisco, 1973.

24. K. Varmuza. Pattern recognition in chemistry. Springer Verlag, Berlin, 1980.