Topics in Survey Methodology and Survey Analysis

Part 2

Kimmo Vehkalahti

University Lecturer, University of Helsinki
Department of Social Research, Statistics
http://www.helsinki.fi/people/Kimmo.Vehkalahti

fall 2011

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Outline of Part 2

Part 2 focuses heavily on **measurement**, which is one of the sources of **uncertainty** in survey research.

The following topics will be covered:

- Reliability, validity and measurement errors
- Exploratory and confirmatory analysis
- Data reduction/compression with factor analysis
- Visualization of multidimensional data

As the topics are more or less intertwined, there will not be a strict order of things. Activity of the participants is much appreciated and it will certainly affect the way (and order) we proceed.

The material includes some notes of rather basic statistics and graphs as well. They might be familiar to many participants, depending on everyone's previous studies. **The mathematical formulas represent additional information only.**

Bibliography and other optional material for Part 2

Measurement, reliability, validity

- Alwin, Duane F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement.* Wiley.
- Fowler, Floyd J. (1995). *Improving Survey Questions: Design and Evaluation.* Applied Social Research Methods Series, Volume 38, Sage.
- Payne, Stanley L. (1951). *The Art of Asking Questions.* Princeton University Press.
- Saris, Willem E. & Gallhofer, Irmtraud N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research.* Wiley.

Factor analysis and other multivariate (survey) methods

- Cudeck, Robert & MacCallum, Robert C., eds. (2007). *Factor Analysis at 100: Historical Developments and Future.* Lawrence Erlbaum Associates.
- Everitt, Brian (2009). *Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences.* Chapman & Hall/CRC.
- Greenacre, Michael (2007). *Correspondence Analysis in Practice*, Second Edition, Chapman & Hall/CRC.
- Greenacre, Michael & Blasius, Jörg, eds. (2006). *Multiple Correspondence Analysis and Related Methods.* Chapman & Hall/CRC.
- Groves, Robert M.; Fowler Jr., Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor & Tourangeau, Roger (2004). *Survey Methodology.* Wiley.
- Mulaik, Stanley A. (2009). *Foundations of Factor Analysis,* Second Edition. Chapman & Hall/CRC.

Visualization of statistical data

- Chen, Chun-houh; Härdle, Wolfgang & Unwin, Antony, eds. (2008). *Handbook of data Visualization.* Springer.
- Cleveland, William S. (1993). *Visualizing Data.* Hobart Press, Summit, NJ.
- Greenacre, Michael (2010). *Biplots in Practice.* Fundación BBVA. `http://www.multivariatestatistics.org`
- Robbins, Naomi B. (2005). *Creating More Effective Graphs.* Wiley.
- Tufte, Edward R. (2001). *The Visual Display of Quantitative Information*, Second Edition. Graphics Press. (Fifth printing, August 2007.)

Bibliography etc. for Part 2 (continued)

Some books in Finnish

- Karjalainen, Leila & Karjalainen, Juha (2009). *Tilastojen graafinen esittäminen.* Pii-Kirjat.
- Ketokivi, Mikko (2009). *Tilastollinen päättely ja tieteellinen argumentointi.* Palmenia.
- Kuusela, Vesa (2000). *Tilastografiikan perusteet.* Edita.
- Mustonen, Seppo (1995). *Tilastolliset monimuuttujamenetelmät.* Survo Systems. `http://www.survo.fi/mustonen/monim.pdf`
- Nummenmaa, Tapio; Konttinen, Raimo; Kuusinen, Jorma & Leskinen, Esko (1996). *Tutkimusaineiston analyysi.* WSOY.
- Vehkalahti, Kimmo (2008). *Kyselytutkimuksen mittarit ja menetelmät.* Tammi.

Some studies used in examples and demos (partially in Finnish)

- European Social Survey, Round 4, subset of Finland `http://ess.nsd.uib.no/ess/round4/`
- Economic Freedom `http://www.heritage.org/Index/`
- Prices and Earnings around the Globe `http://www.ubs.com/research` `http://www.macrofocus.com/public/products/infoscope/datasets/pricesandearnings/`
- Suomalaisten arvot ja uskonnollisuus `http://www.fsd.uta.fi/aineistot/luettelo/FSD2410/meF2410.html`
- Nuorisobarometri `http://www.fsd.uta.fi/aineistot/luettelo/FSD2293/meF2293.html`

Effects of measurement to (survey) data analysis

Measurement and measures in survey research:

- measurement model: **what** to measure
- measuring instrument: **how** to measure
- instrument in survey research: **questionnaire**
- **pattern:** collection of **items** (questions, statements)

Results are affected by the **measurement quality:**

1. **validity:** are we (really) measuring the right thing?
2. **reliability:** are we measuring accurately enough?

Measurement level sets the limits for the methods:

- **classification — ordering — numeric measurement**
  (*cf.* "nominal", "ordinal", "interval"/"ratio")
- most methods require numeric measurement
- in some methods classification is enough
- the level of ordering is often most problematic

Examples of items – what are their measurement levels?

Source: ESS (European Social Survey), `http://ess.nsd.uib.no/ess/`
Here: modified and abbreviated (DK = Don't Know).

- How interested are you in politics?
  Very interested (1), quite interested (2), hardly interested (3), not at all interested? (4), DK (8)
- Did you vote in the last national election?
  Yes (1), No (2), Not eligible to vote (3), DK (8)
- Have you boycotted certain products?
  Yes (1), No (2), DK (8)
- How satisfied are you with the present state of the economy?
  Extremely dissatisfied 00 01 02 03 04 05 06 07 08 09 10 Extremely satisfied, DK (88)
- The government should take measures to reduce differences in income levels.
  Agree strongly (1), Agree (2), Neither agree nor disagree (3), Disagree (4), Disagree strongly (5), DK (8)
- What is your current situation?
  paid work (1), education (2), unemployed (3), sick or disabled (4), retired (5), housework (6), other (7)
- How many hours do you work weekly: _____
- How much do you use internet: no access (00), never (01), less than once a month (02), once a month (03), several times a month (04), once a week (05), several times a week (06), every day (07), DK (88)
- Any particular religion you have considered yourself as belonging to?
  Roman Catholic (01), Protestant (02), Eastern Orthodox (03), Other Christian denomination (04), Jewish (05), Islamic (06), Eastern religions (07), Other non-Christian religions (08)

General aim: compressing the data

A general aim of **statistical methods** is to **compress** the information in the **data** into a form of **graphs** and **statistics**.

- compressing and other analyses will absolutely require a proper knowledge of the data (and the study in question)
- central for the knowledge: graphs (and statistics) of the distributions
  - empirical distribution: all the measured (and coded) values of one variable in the data
- other ways of compressing the data:
  - combining the variables, e.g., by forming summated variables (meaningful only if summing is reasonable)
  - **multivariate methods**, such as **factor analysis**

Types of variables in the data

- **quantitative variables:**
  - **continuous** variables (such as age, length, weight etc.) (only a few identical values)
  - **discrete** variables (such as scales of opinions, counts) (many identical values, i.e., only a few different ones)

In practice, measuring something and saving it on the computer is possible only on a finite precision ("everything is discrete").

A variable may, however, be *interpreted* as continuous, if it reflects a continuous phenomenon or issue (e.g., age).

- **qualitative variables** (all discrete)**:**
  - **ordered** variables (e.g., education)
  - **classified** variables (e.g., gender)

Quantitative variables may always be transformed to qualitative ones (by classifying) but **not** the other way. Hence, it is worthwhile to **measure** as precisely and accurately as possible! It cannot be redone...

Statistics and their interpretation

When data is compressed into **statistics**:

- some of the information is always lost
- a plain statistic (typically one number!) does not say much

The most general statistics is the **mean** (average):

- the sum of the values divided by the number of observations
- meaningful only if summing is reasonable
- applicable only for quantitative variables
- plain mean is not enough (tells nothing about **variation**)

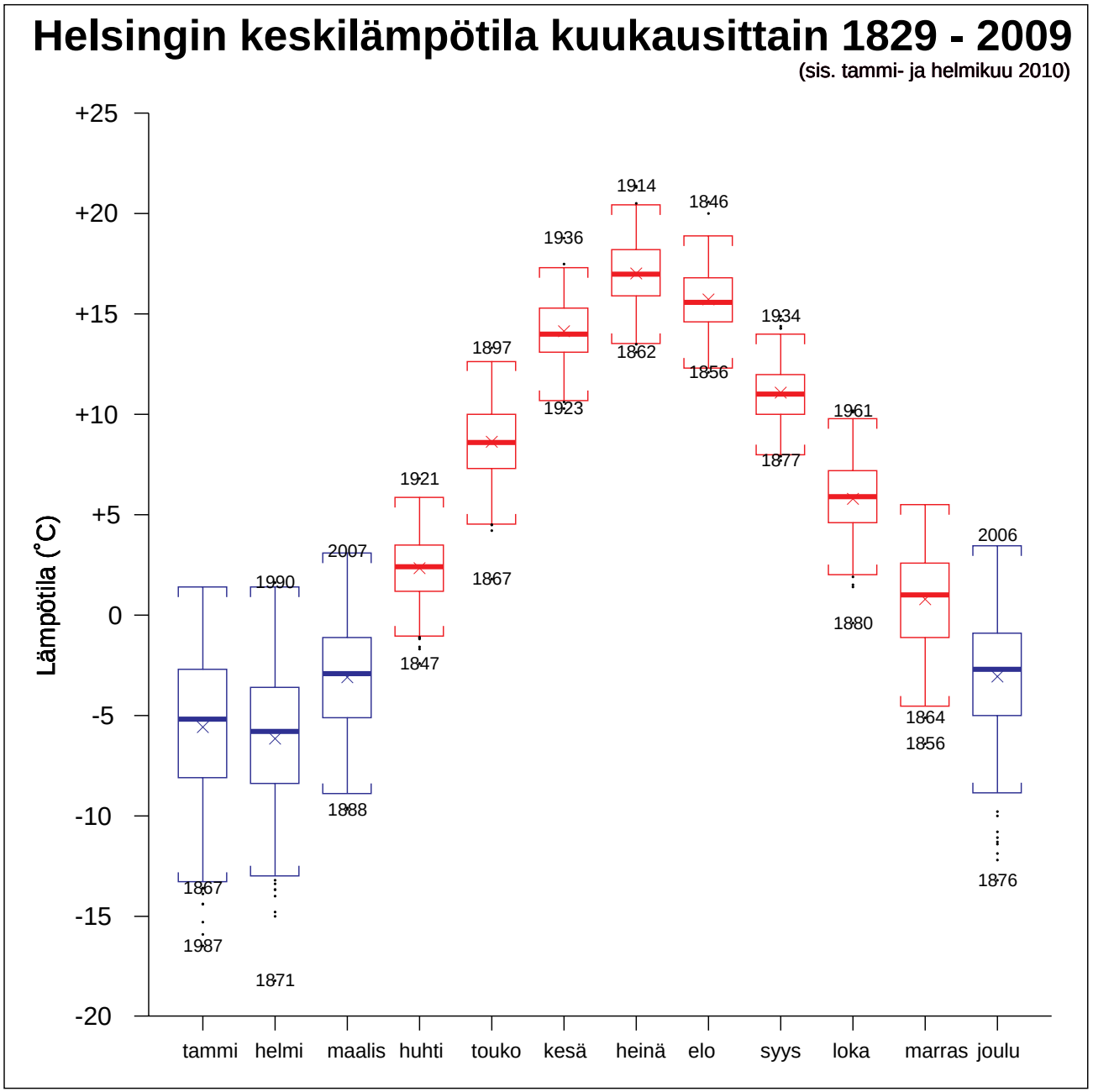
Often a better alternative is offered by the **median**:

- the central value of the ordered variable
- no calculations are employed: also valid for ordering level
- more robust with possible outliers (unlike mean)
- plain median is not enough (tells nothing about **variation**)

A useful collection of **five** (order) **statistics:**

- min, lower quartile (25 %), median (50 %), upper quartile (75 %), max
- graphical representation: box (and whiskers) plot
- note: the "box" will then include half of the observations

Example: average temperatures in Helsinki, 1829–2009



Variation and dependence

One of the key concepts of statistics is **variation:**

- the more variation, the more information
- variable with no variation is *constant* (same value for all)
  - no statistical information (of course, may be interesting)
- measures of variation (*cf.* box plot and its statistics):
  - **range:** [min(imum), max(imum)]
  - **quartile range:** [lower q, upper q]
    (also the lengths of the ranges may be used)
- most typical measure of variation is the **standard deviation:**
  - "the average deviation" of the observations from their mean
  - given in same units with the mean (easy to interpret?)
  - mean $\pm$ 1 std devs covers ca. 68 % of the values
  - mean $\pm$ 2 std devs covers ca. 95 % of the values
    (assuming the distribution is quite *symmetric* and *unimodal*)
  - the square of the std dev is *variance* (more theoretical)

Dependence and correlation

Another key concept is **dependence:** most research questions are somehow related to *dependence of different aspects or issues*.

- the character of the dependence can be evaluated by examining the scatter plot (or a cross tabulation)
- important case: **linear dependence**
- *but:* dependence may often be **nonlinear**

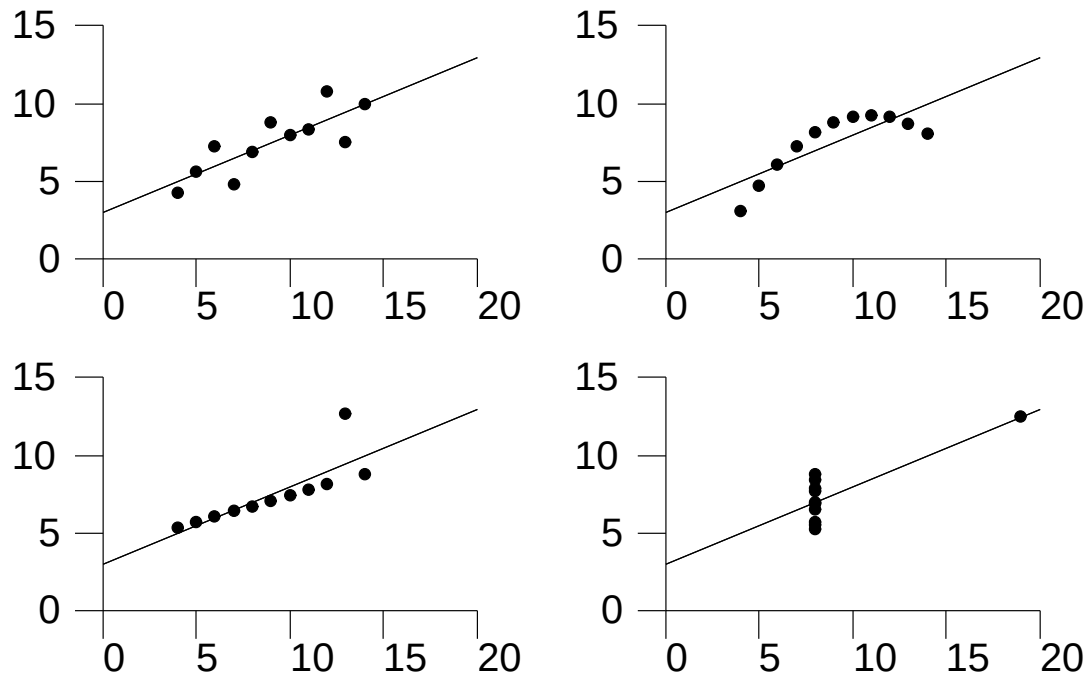
In case of linear dependence the **correlation** might be useful:

- correlation (coefficient) is a statistic of two variables (measure of variation of which is the standard deviation)
- correlation is a number on the interval $[-1, 1]$, e.g., -0.72
- most important: learning to **interpret** the correlation:
  - is the correlation positive of negative?
  - when is the correlation practically zero?
  - what if the correlation is almost $+1$ or -1?
  - *is one statistic again enough for the purpose?*
- correlation describes merely a relation, **not** a causation (causal inference is a subject matter, not a statistical matter!)

Visualization of variation and dependence: scatter diagram

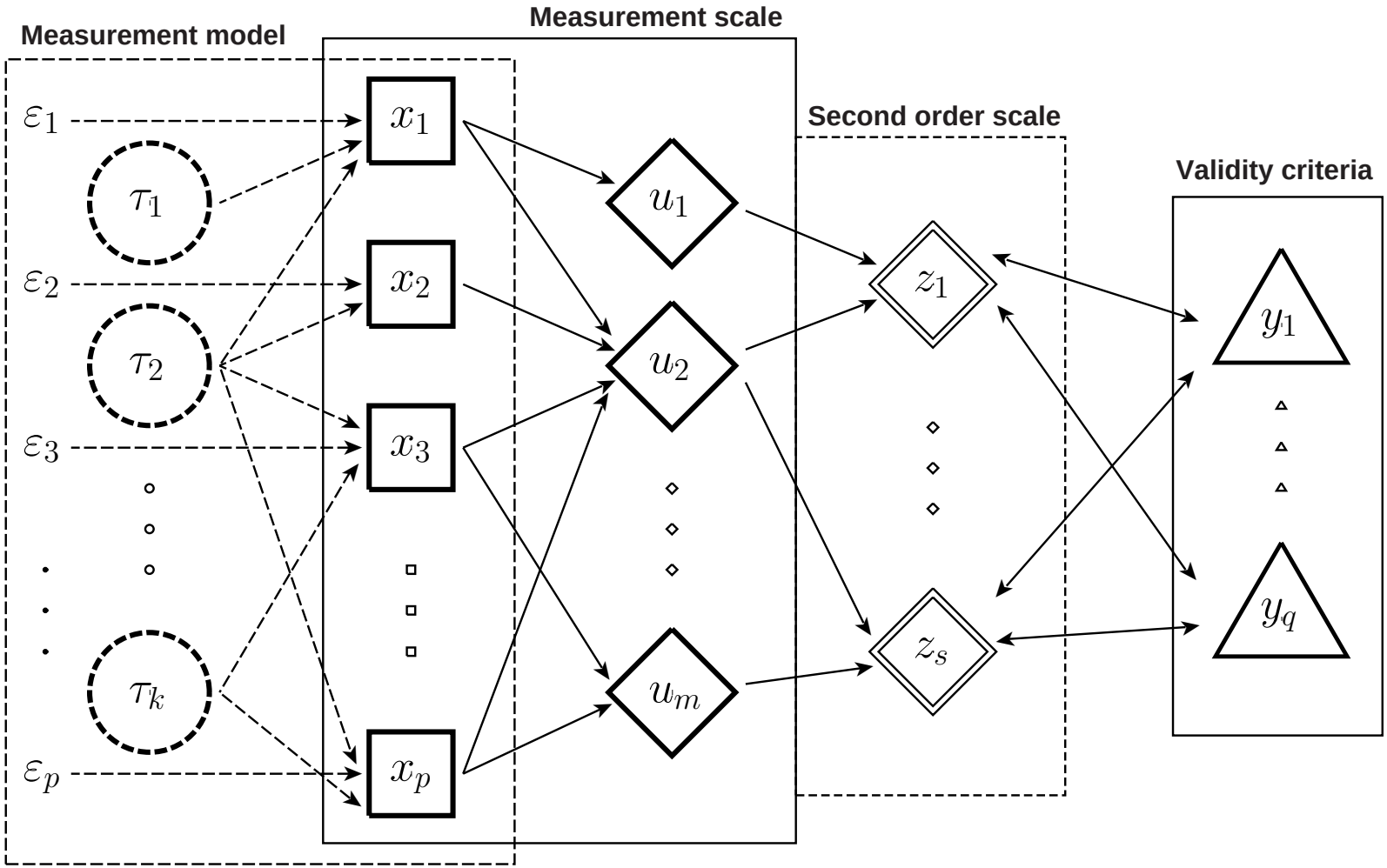
A **scatter diagram** is an excellent way to visualize and analyse **variation** and **dependence** simultaneously:

- basically a graph of two continuous variables
- numerous variations for different settings
- capable of presenting a large amount of information
- also reveals surprises such as outliers



Anscombe, F. (1973). Graphs in statistical analysis, *American Statistician*, **27**, 17-21.

Measurement framework: "guide from plans to analyses"



- a basis for assessing the measurement quality
- **essential statistical method: factor analysis**
- includes other multivariate methods as well

Four parts of the measurement framework

Measurement model

- What is the *phenomenon* under study?
- How many *dimensions* it might consist of?
- How could those dimensions be *measured*?
- **factor analysis** (exploratory—confirmatory)

Measurement scale

- combination of measures items
- examples: factor scores, summated scales, indices
- **compressing the data**

Second order scale

- result of e.g., regression or discriminant analyses
- connects measurements with other multivariate methods

Validity criteria

- a criteria defined outside of the measurement model
- for comparisons, orderings, classifications etc. of respondents

Theory behind the framework: Measurement model

In this material, the mathematical formulas represent additional information only.

Tarkkonen, L. & Vehkalahti, K. (2005). Measurement errors in multivariate measurement scales, *Journal of Multivariate Analysis*, **96**, 172–189.

Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ measure k (**important here:** $k < p$)
unobservable **true scores** $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)'$ with
unobservable **measurement errors** $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$.

Assume $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\tau}, \boldsymbol{\varepsilon}) = \mathbf{0}$. The measurement model is

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{B} \in \mathbb{R}^{p \times k}$ specifies the relationship between $\boldsymbol{x}$ and $\boldsymbol{\tau}$.

Denoting $\mathrm{cov}(\boldsymbol{\tau}) = \boldsymbol{\Phi}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$ we have

$$\mathrm{cov}(\boldsymbol{x}) = \boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}' + \boldsymbol{\Psi}, \tag{2}$$

- where it is assumed that $\boldsymbol{\Sigma} > \mathbf{0}$ and $\boldsymbol{B}$ has full column rank.

Theory behind the framework: Estimation of parameters

In this material, the mathematical formulas represent additional information only.

The **parameters** are the $pk + k(k+1)/2 + p(p+1)/2$ (unique) elements of the matrices $\boldsymbol{B}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$. In general, there are too many, since $\boldsymbol{\Sigma}$ has only $p(p+1)/2$ elements.

- Identifiability is obtained by imposing assumptions on the true scores and the measurement errors.
- **Typical:** assume that $\text{cov}(\boldsymbol{\tau}) = \boldsymbol{I}_k$, an identity matrix of order k, and $\text{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}_d = \text{diag}(\psi_1^2, \ldots, \psi_p^2)$.
- With these the model conforms with the orthogonal factor analysis model where the *common factors are directly associated with the true scores* and the *specific factors are interpreted as measurement errors*.

Assuming **multinormality** the parameters can be estimated using e.g., **the maximum likelihood** method of factor analysis.

Theory behind the framework: Structural validity

In this material, the mathematical formulas represent additional information only.

Structural validity is a property of the measurement model.

- Important, as the model forms the core of the framework and hence affects the quality of all scales created.
- Lack of structural validity can be revealed by testing
  - hypotheses on the dimension of $\boldsymbol{\tau}$
  - hypotheses on the effects of $\boldsymbol{\tau}$ on $\boldsymbol{x}$ (matrix $\boldsymbol{B}$)
- The whole approach could be called *semi-confirmatory*.
- Residuals of the model obtained by estimation of $\text{var}(\boldsymbol{\varepsilon})$.
- Dimension of $\boldsymbol{\tau}$ will make the reliabilities identified.
- Appropriate (e.g. **graphical**) factor rotation is essential.

Similarly with other questions of validity, knowledge of the theory and practice of the application is crucial.

Theory behind the framework: Measurement scale

In this material, the mathematical formulas represent additional information only.

In further analyses, the variables $\boldsymbol{x}$ are best used by creating **multivariate measurement scales** $\boldsymbol{u} = \boldsymbol{A}'\boldsymbol{x}$, where $\boldsymbol{A} \in \mathbb{R}^{p \times m}$ is a matrix of the weights. Using (2) we obtain

$$\text{cov}(\boldsymbol{u}) = \boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}'\boldsymbol{A} + \boldsymbol{A}'\boldsymbol{\Psi}\boldsymbol{A}, \tag{3}$$

the (co)variances generated by the **true scores** and the (co)variances generated by the **measurement errors**.

Examples of measurement scales include factor scores, psychological test scales, or any other linear combinations of the observed variables. The weights of the scale may also be predetermined values according to a theory.

Theory behind the framework: Predictive validity

In this material, the mathematical formulas represent additional information only.

Predictive validity is a property of the measurement scale.

- Assessed by the correlation(s) between the (second order) scale and an *external criterion*.
- In general, a second order scale is denoted by $\boldsymbol{z} = \boldsymbol{W}'\boldsymbol{u} = \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{x}$, where $\boldsymbol{W} \in \mathbb{R}^{m \times s}$ is a weight matrix and a criterion is denoted by $\boldsymbol{y} = (y_1, \ldots, y_q)'$.
- Often, these scales are produced by regression analysis, discriminant analysis, or other multivariate statistical methods.

In the most general case, the predictive validity would be assessed by the **canonical correlations** between $\boldsymbol{z}$ and $\boldsymbol{y}$.

Theory behind the framework: Predictive validity

In this material, the mathematical formulas represent additional information only.

Example: consider the regression model $y = \beta_0 + \boldsymbol{\beta}'\boldsymbol{u} + \delta$, where y is the response variable, β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ is the vector of the regression coefficients, $\boldsymbol{u}$ is the vector of the predictors (e.g., factor scores), and δ is a model error.

Now, the criterion y is a scalar, and the second order scale is given by the prediction scale $z = \hat{\boldsymbol{\beta}}'\boldsymbol{u}$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)'$.
Hence the predictive validity is equal to ρ_{zy}, the multiple correlation of the regression model.

Monte Carlo simulations indicate that the factor scores offer the most stable method for predictor selection in the regression model.
Details are found in

Vehkalahti, K., Puntanen, S. & Tarkkonen, L. (2007). Effects of measurement errors in predictor selection of linear regression model, *Computational Statistics & Data Analysis*, **52**(2), 1183–1195.

Theory behind the framework: Reliability (Tarkkonen's rho)

In this material, the mathematical formulas represent additional information only.

According to the definition of reliability, Tarkkonen's rho is obtained as a **ratio of the variances**, i.e., the diagonal elements of the matrices in (3). Hence we have

$$\boldsymbol{\rho_u} = \text{diag}\left(\frac{\boldsymbol{a}_1'\boldsymbol{B\Phi B'a}_1}{\boldsymbol{a}_1'\boldsymbol{\Sigma a}_1}, \ldots, \frac{\boldsymbol{a}_m'\boldsymbol{B\Phi B'a}_m}{\boldsymbol{a}_m'\boldsymbol{\Sigma a}_m}\right)$$

$$= (\boldsymbol{A'B\Phi B'A})_d \times [(\boldsymbol{A'\Sigma A})_d]^{-1}$$

or, in a form where the matrix $\boldsymbol{\Psi}$ is explicitly present:

$$\boldsymbol{\rho_u} = \text{diag}\left(\left[1 + \frac{\boldsymbol{a}_1'\boldsymbol{\Psi a}_1}{\boldsymbol{a}_1'\boldsymbol{B\Phi B'a}_1}\right]^{-1}, \ldots, \left[1 + \frac{\boldsymbol{a}_m'\boldsymbol{\Psi a}_m}{\boldsymbol{a}_m'\boldsymbol{B\Phi B'a}_m}\right]^{-1}\right)$$

$$= \{\boldsymbol{I}_m + (\boldsymbol{A'\Psi A})_d \times [(\boldsymbol{A'B\Phi B'A})_d]^{-1}\}^{-1}$$

Theory behind the framework: Reliability (special cases)

In this material, the mathematical formulas represent additional information only.

Many models, scales, and reliability coefficients established in the test theory of psychometrics are special cases of the framework.

Example: $\boldsymbol{x} = \boldsymbol{\mu} + \mathbf{1}\tau + \boldsymbol{\varepsilon}$ and $u = \mathbf{1}'\boldsymbol{x}$ (unweighted sum).

Now, $\boldsymbol{\Sigma} = \sigma_\tau^2 \mathbf{1}\mathbf{1}' + \boldsymbol{\Psi}_d$ and $\sigma_u^2 = \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} = p^2\sigma_\tau^2 + \mathrm{tr}(\boldsymbol{\Psi}_d)$.

$$\rho_{uu} = \frac{p^2\sigma_\tau^2}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} = \frac{p}{p-1}\left(\frac{p^2\sigma_\tau^2 - p\sigma_\tau^2}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right)$$

$$= \frac{p}{p-1}\left(\frac{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} - \mathrm{tr}(\boldsymbol{\Psi}_d) - \mathrm{tr}(\boldsymbol{\Sigma}) + \mathrm{tr}(\boldsymbol{\Psi}_d)}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right)$$

$$= \frac{p}{p-1}\left(1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right) = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^p \sigma_{x_i}^2}{\sigma_u^2}\right),$$

which is the original form of Cronbach's alpha, given in Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**, 297–334.

Demo: EU 1996 (exploration with a tiny data)

With old barriers gone, people, goods, services and money move around Europe as freely as within one country.

1 January 1993

The single market and its four freedoms are established: the free movement of goods, services, people and money is now reality. More than 200 laws have been agreed since 1986 covering tax policy, business regulations, professional qualifications and other barriers to open frontiers. The free movement of some services is delayed.

1 January 1995

Austria, Finland and Sweden join the EU. The 15 members now cover almost the whole of western Europe. In October 1990, Germany was unified and therefore former East Germany became part of the EU.

Member States: Germany, France, Italy, the Netherlands, Belgium, Luxembourg, Denmark, Ireland, United Kingdom, Greece, Spain and Portugal.

New Member States: Austria, Finland and Sweden.

See animated map of all EU enlargements.

Passport-free travel across frontiers.

26 March 1995

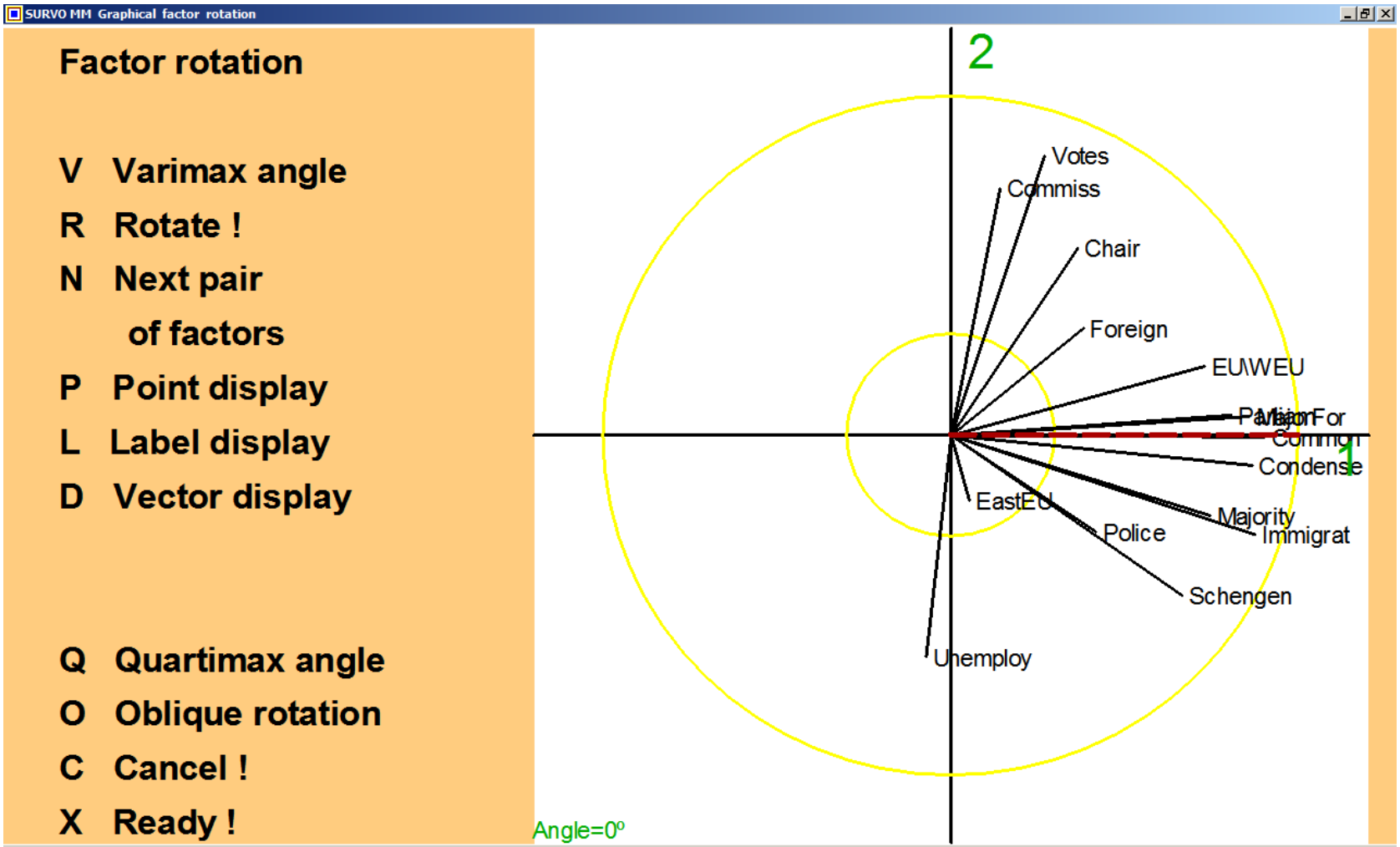
The Schengen Agreement takes effect in seven countries — Belgium, Germany, Spain, France, Luxembourg, the Netherlands and Portugal. Travellers of any nationality can travel between all these countries without any passport control at the frontiers. Other countries have since joined the passport-free Schengen area.

During negotiations in Amsterdam, EU leaders try out a typical Dutch bike.

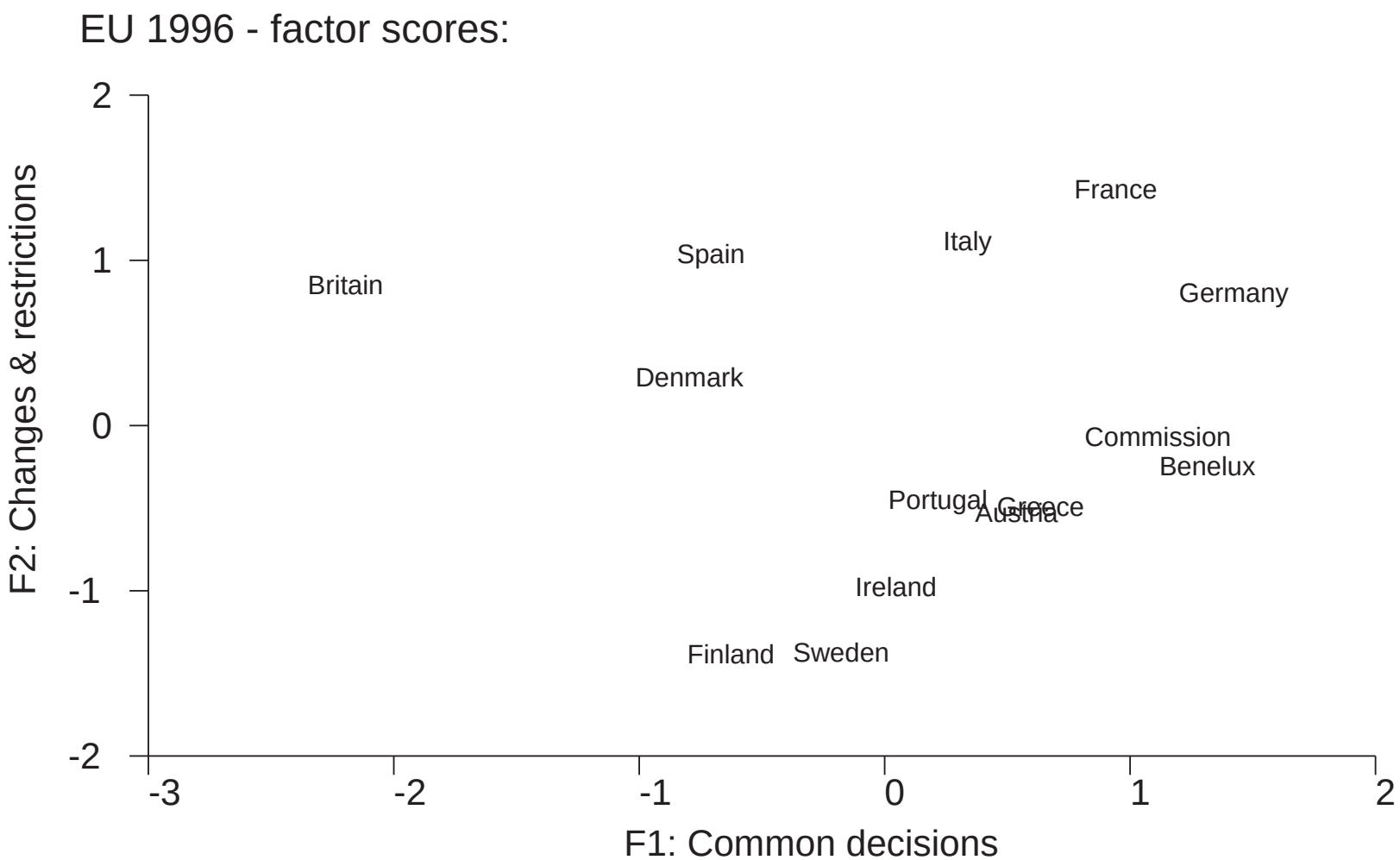
17 June 1997

Signature of the Treaty of Amsterdam. It builds on the achievements of the treaty from Maastricht, laying down plans to reform EU institutions, to give Europe a stronger voice in the world, and to concentrate more resources on employment and the rights of citizens.

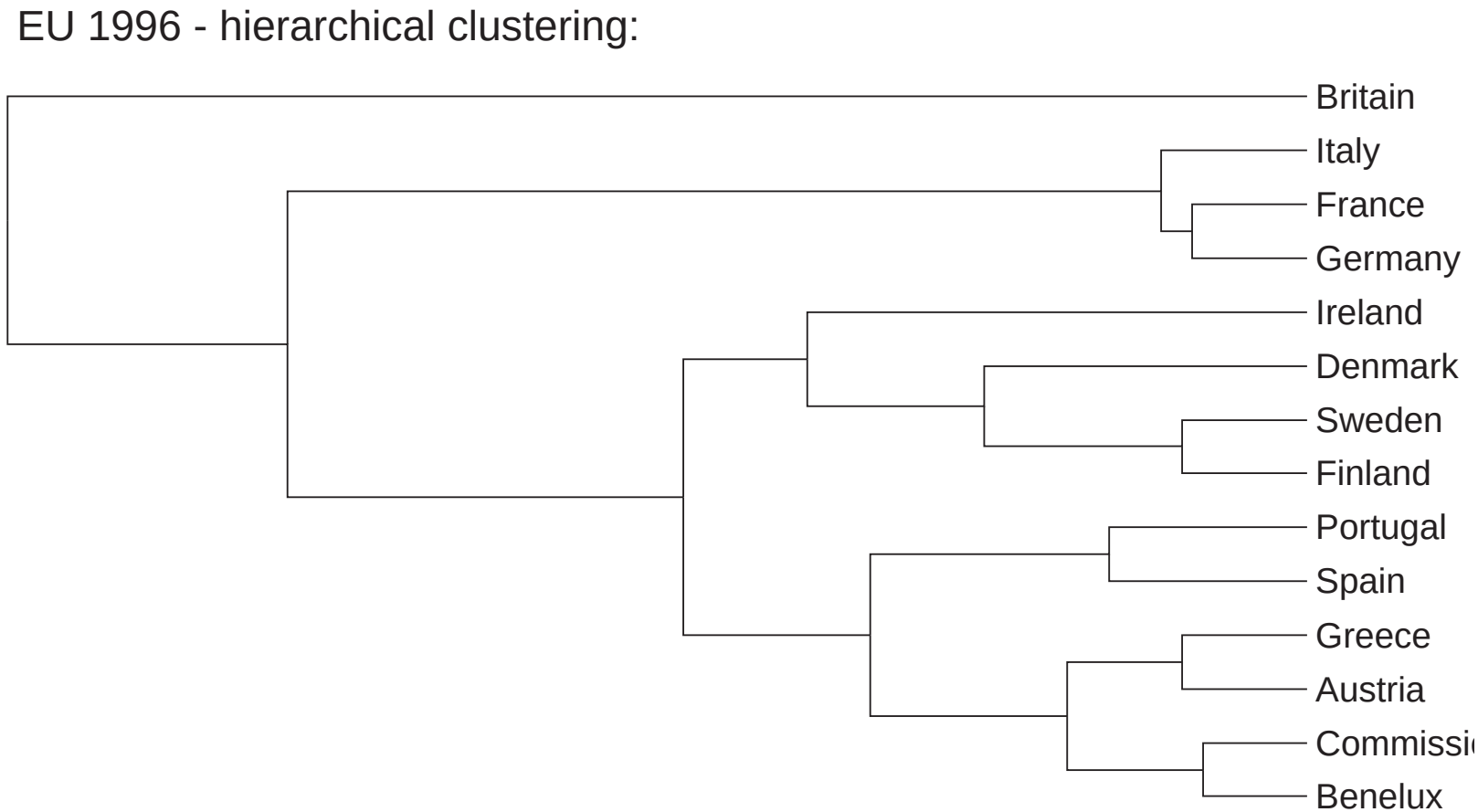
Demo: EU 1996 (graphical factor rotation)



Demo: EU 1996 (factor scores)

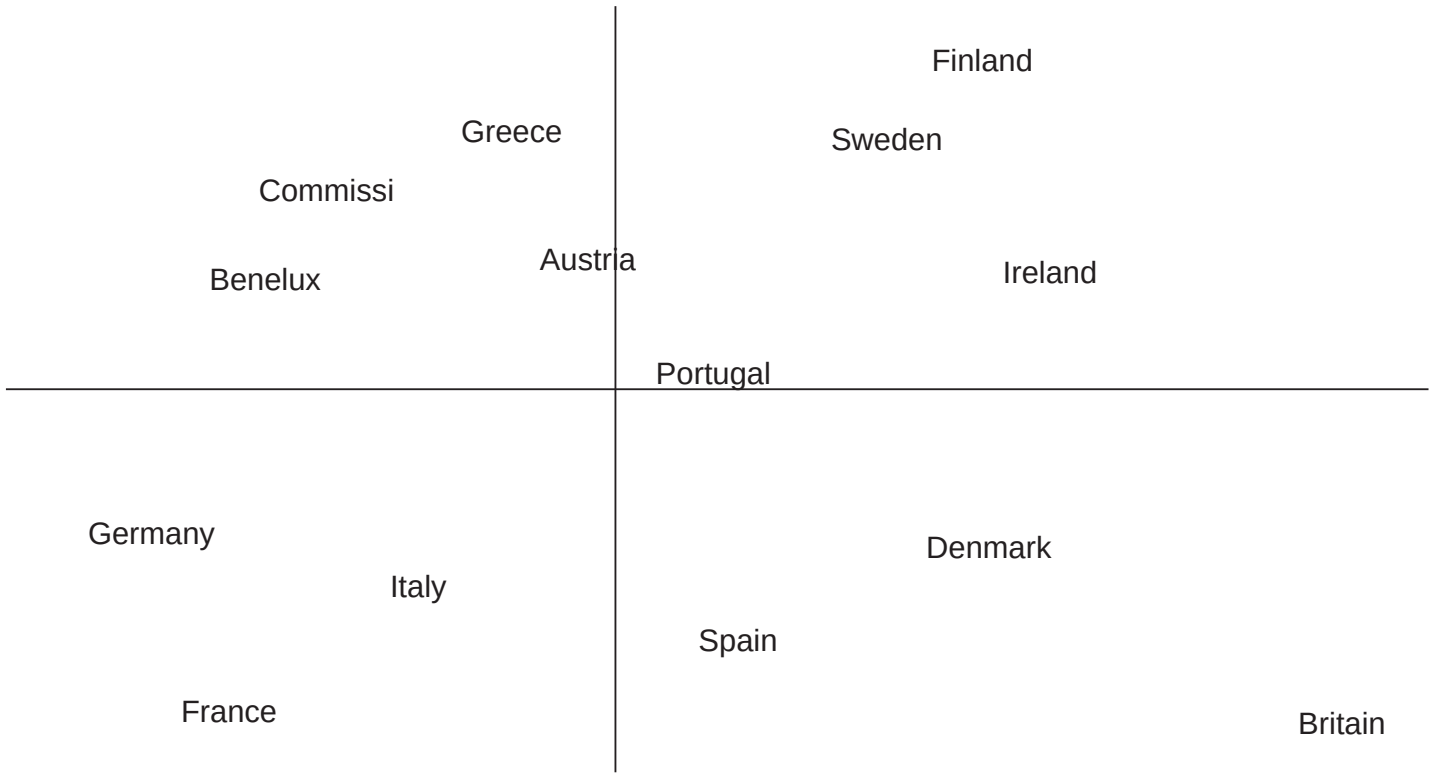


Demo: EU 1996 (hierarchical clustering)



Demo: EU 1996 (multidimensional scaling)

EU 1996 - multidimensional scaling:



Compressing data with factor analysis

Example: Attitudes among Finnish, Dutch and British Consumers

- **Source:** K. Roininen, H. Tuorila, E.H. Zandstra, C. de Graaf, K. Vehkalahti, K. Stubenitsky, and D.J. Mela (2001). Differences in Health and Taste Attitudes and Reported Behaviour among Finnish, Dutch and British Consumers: a Cross-National Validation of the Health and Taste Attitude Scales (HTAS). *Appetite*, **37**, 33–45.

In the following, the focus is on the Finnish data ($N = 467$).

The **Health** subset of HTAS pattern measures 3 dimensions:

1. General Health Interest
2. Light Product Interest
3. Natural Product Interest

The **Taste** subset is rather similar:

1. Craving for sweet foods
2. Using food as a reward
3. Pleasure

Presenting and interpreting the results of factor analysis

Factor structure of Health sub-scales (Finland)	F1	F2	F3	h^2
General Health Interest				
I am very particular about the healthiness of food.	**0.75**	0.16	0.13	0.61
I always follow a healthy and balanced diet.	**0.73**	0.13	-0.02	0.56
It is important for me that my diet is low in fat.	**0.65**	0.12	0.26	0.50
It is important for me that my daily diet contains a lot of vitamins and minerals.	**0.64**	*0.31*	0.10	0.52
(R) I eat what I like and I do not worry about healthiness of food.	**0.47**	*0.34*	*0.31*	0.43
(R) I do not avoid any foods, even if they may raise my cholesterol.	**0.46**	*0.40*	0.27	0.44
(R) The healthiness of food has little impact on my food choices.	**0.42**	*0.31*	*0.50*	0.53
(R) The healthiness of snacks makes no difference to me.	**0.32**	0.24	*0.50*	0.41
Light Product Interest				
(R) In my opinion, the use of light products does not improve ones health.	0.03	**0.77**	0.26	0.67
(R) I do not think that light products are healthier than conventional products.	-0.09	**0.74**	0.17	0.58
I believe that eating light products keeps one's cholesterol level under control.	0.29	**0.61**	-0.04	0.46
(R) In my opinion light products don't help to drop cholesterol levels.	-0.07	**0.61**	0.09	0.38
I believe that eating light products keeps one's body in good shape.	*0.37*	**0.54**	-0.10	0.43
In my opinion by eating light products one can eat more without getting too many calories.	0.23	**0.33**	-0.18	0.19
Natural Product Interest				
(R) I do not care about additives in my daily diet.	*0.49*	0.06	**0.52**	0.51
(R) In my opinion, organically grown foods are no better for my health than those grown conventionally.	0.11	0.20	**0.52**	0.32
(R) In my opinion, artificially flavored foods are not harmful for my health.	0.08	-0.09	**0.50**	0.27
I try to eat foods that do not contain additives.	*0.63*	-0.07	**0.37**	0.54
I would like to eat only organically grown vegetables.	*0.46*	-0.03	**0.34**	0.32
I do not eat processed foods, because I do not know what they contain.	*0.50*	-0.15	**0.23**	0.33
Sum of squares	4.05	2.94	2.01	9.00
Variance explained %	20.3	14.7	10.0	45.0

(R) = reversed (negative) statements, h^2 = communalities

Presenting and interpreting the results of factor analysis

Factor structure of Taste sub-scales (Finland)	F1	F2	F3	h^2
Craving for sweet foods				
(R) In my opinion it is strange that some people have cravings for chocolate.	**0.80**	0.16	-0.02	0.67
(R) In my opinion it is strange that some people have cravings for sweets.	**0.77**	0.08	0.10	0.61
(R) In my opinion it is strange that some people have cravings for ice-cream.	**0.73**	0.00	0.10	0.55
I often have cravings for sweets.	**0.62**	*0.43*	-0.06	0.58
I often have cravings for chocolate.	**0.54**	*0.43*	-0.08	0.48
I often have cravings for ice-cream.	**0.44**	*0.35*	-0.09	0.33
Using food as a reward				
I reward myself by buying something really tasty.	0.14	**0.85**	0.19	0.78
I indulge myself by buying something really delicious.	0.20	**0.83**	0.22	0.79
When I am feeling down I want to treat myself with something really delicious.	0.24	**0.56**	0.14	0.39
(R) I avoid rewarding myself with food.	0.29	**0.16**	0.21	0.15
(R) In my opinion, comforting oneself by eating is self-deception.	0.23	**0.15**	0.21	0.12
(R) I try to avoid eating delicious food when I am feeling down.	*0.33*	**0.15**	0.02	0.13
Pleasure				
(R) I do not believe that food should always be source of pleasure.	0.11	0.10	**0.60**	0.38
(R) The appearance of food makes no difference to me.	0.00	-0.02	**0.51**	0.26
It is important for me to eat delicious food on weekdays as well as weekends.	-0.01	0.15	**0.48**	0.26
When I eat, I concentrate on enjoying the taste of food.	-0.11	0.12	**0.44**	0.22
(R) I finish my meal even when I do not like the taste of a food.	0.08	0.04	**0.43**	0.19
An essential part of my weekend is eating delicious food.	0.15	*0.32*	**0.20**	0.17
Sum of squares	3.07	2.49	1.50	7.06
Variance explained %	17.1	13.8	8.3	39.2

(R) = reversed (negative) statements, h^2 = communalities

Methods for classifying and clustering the respondents

After compressing the data we can dig deeper in it, and ask:

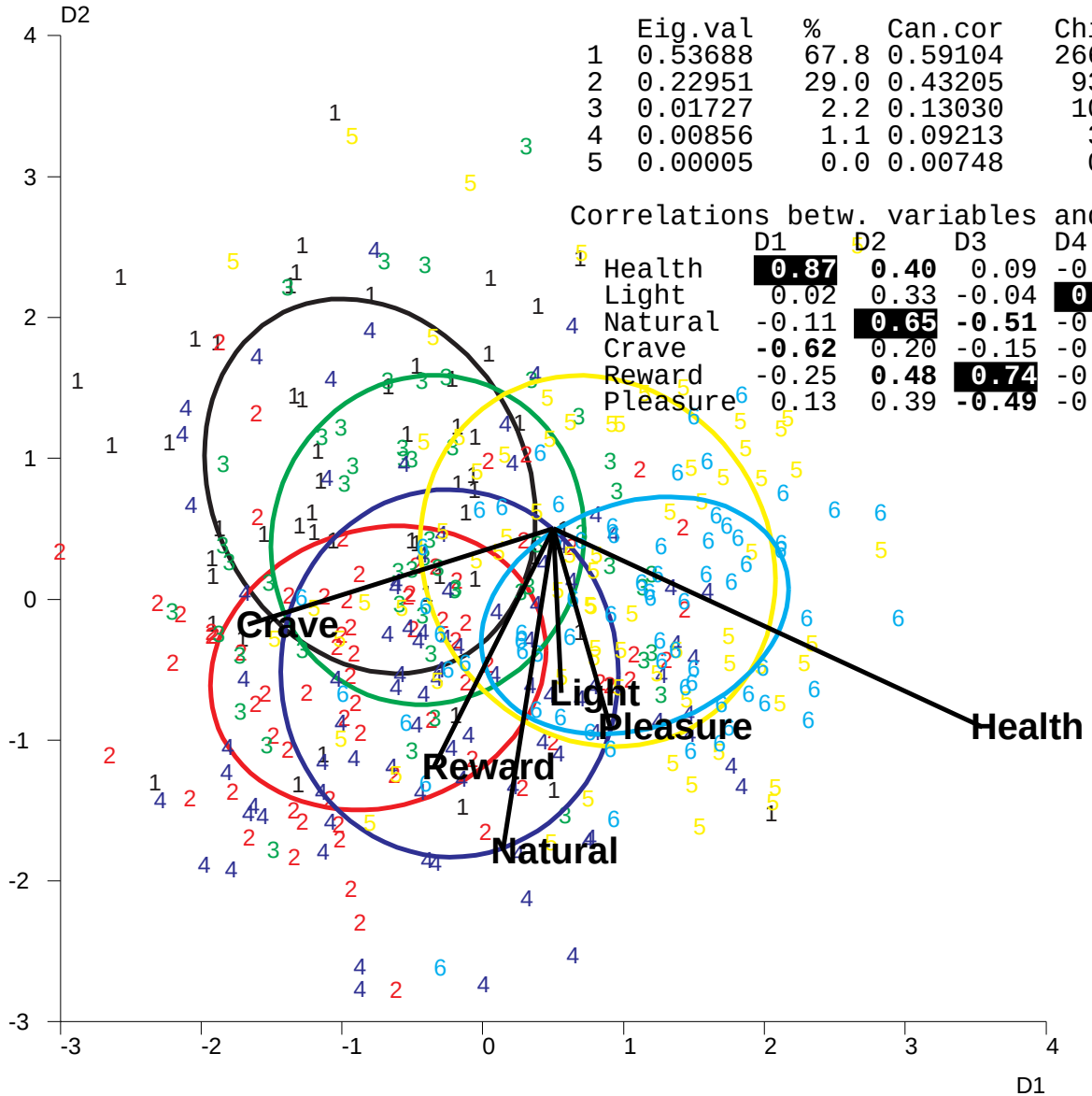
- "what type of (respondent) groups could be found in data?" and "how should we interpret and label them?"
  - **clustering methods**
- "what makes the difference between the (known) groups?" and "into which group we would classify a new respondent?"
  - **discriminant analysis**
- "how do the respondents/groups settle and relate to each other in respect of the background variables and other classifications?"
  - **correspondence analysis**

The results of multivariate methods are best explained by visualizing them — e.g. using variations of the scatter plot.

Tools for immediate visualization of multidimensional data may be used for small data, aggregated subsets, clusters etc.

Example: visualizing the results of multivariate methods

Discriminant analysis (factor dimensions, age and gender):



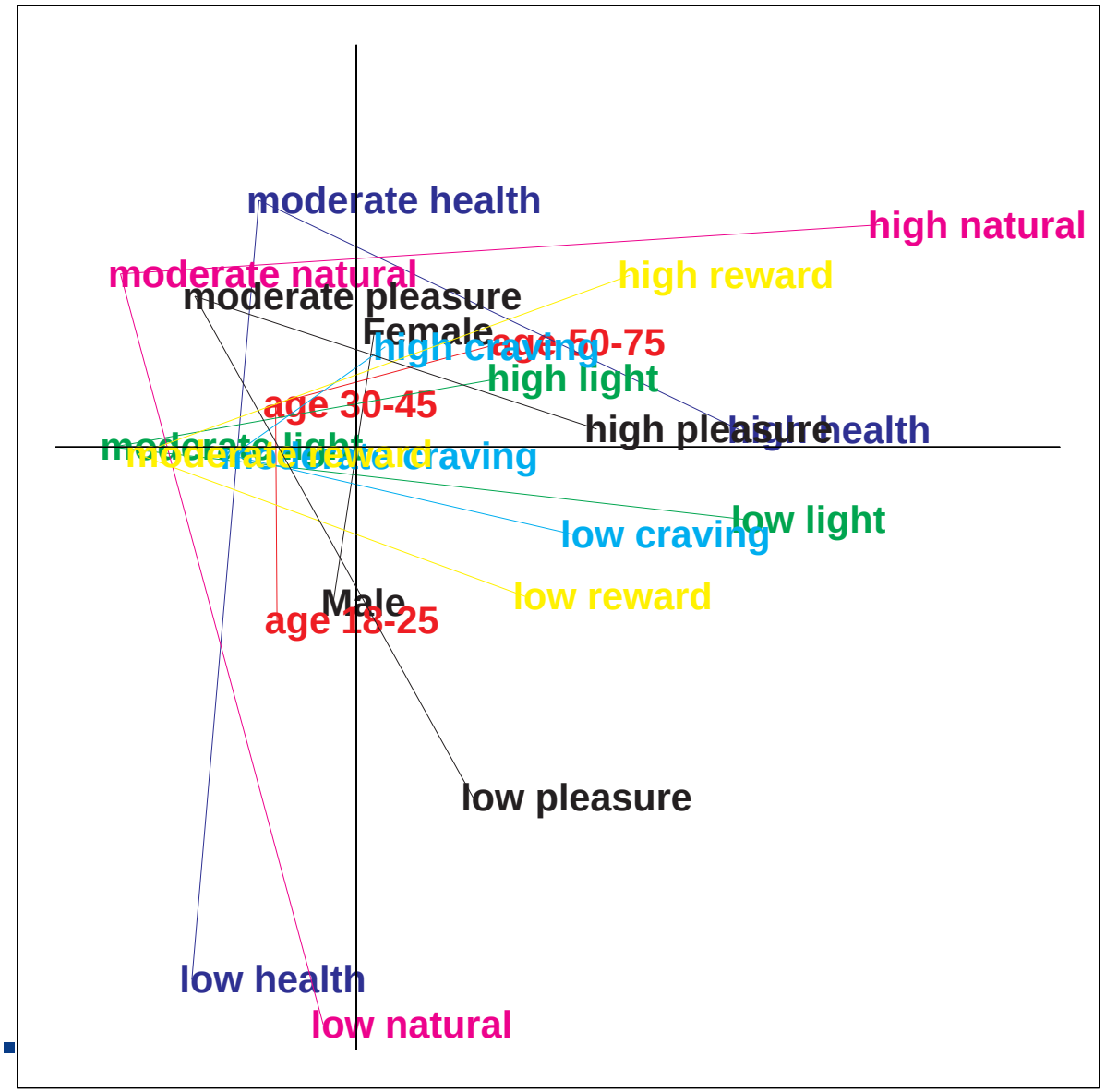
	Eig.val	%	Can.cor	Chi^2	df	P
1	0.53688	67.8	0.59104	266.15	30	0.999
2	0.22951	29.0	0.43205	93.39	20	0.999
3	0.01727	2.2	0.13030	10.33	12	0.413
4	0.00856	1.1	0.09213	3.44	6	0.249
5	0.00005	0.0	0.00748	0.02	2	0.011

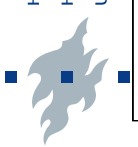
Correlations betw. variables and discriminators

	D1	D2	D3	D4	D5
Health	**0.87**	**0.40**	0.09	-0.18	0.12
Light	0.02	0.33	-0.04	**0.90**	0.13
Natural	-0.11	**0.65**	**-0.51**	-0.22	**-0.42**
Crave	**-0.62**	0.20	-0.15	-0.25	**0.68**
Reward	-0.25	**0.48**	**0.74**	-0.08	-0.01
Pleasure	0.13	0.39	**-0.49**	-0.03	0.08

Example: visualizing the results of multivariate methods

Correspondence analysis (factor dimensions, age and gender):

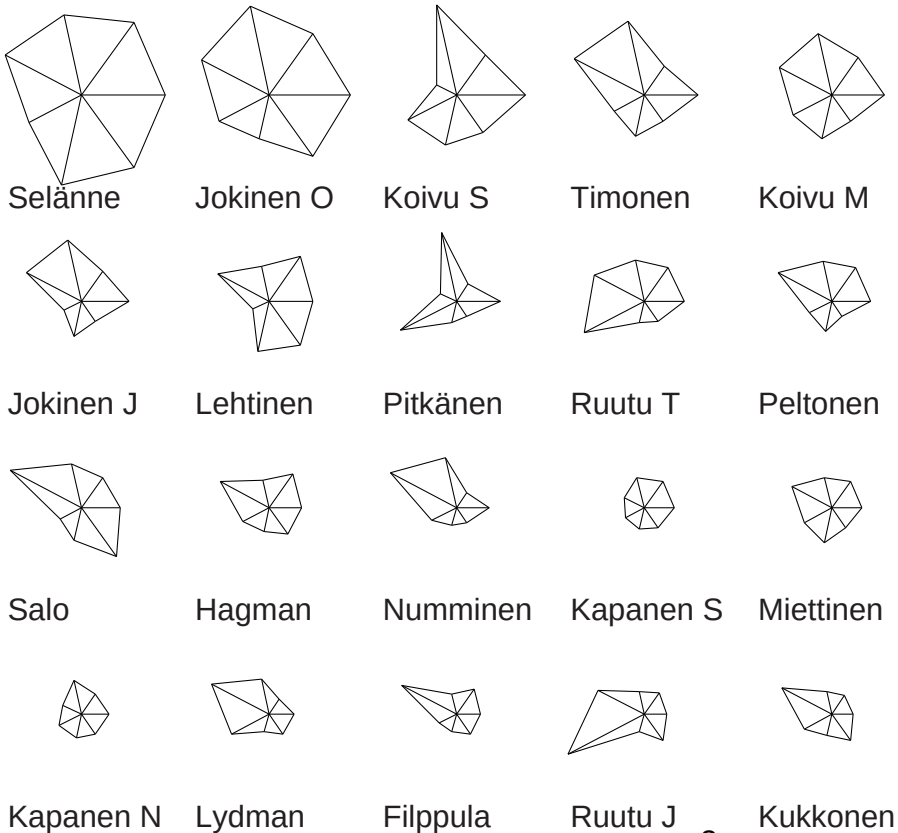




Example of visualizing player ("respondent") profiles

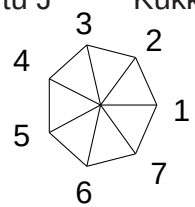
Suomalaistähtien pelaajaprofiilit

NHL:n runkosarja 2006-2007, yli 50 ottelua pelanneet



Muuttujat:

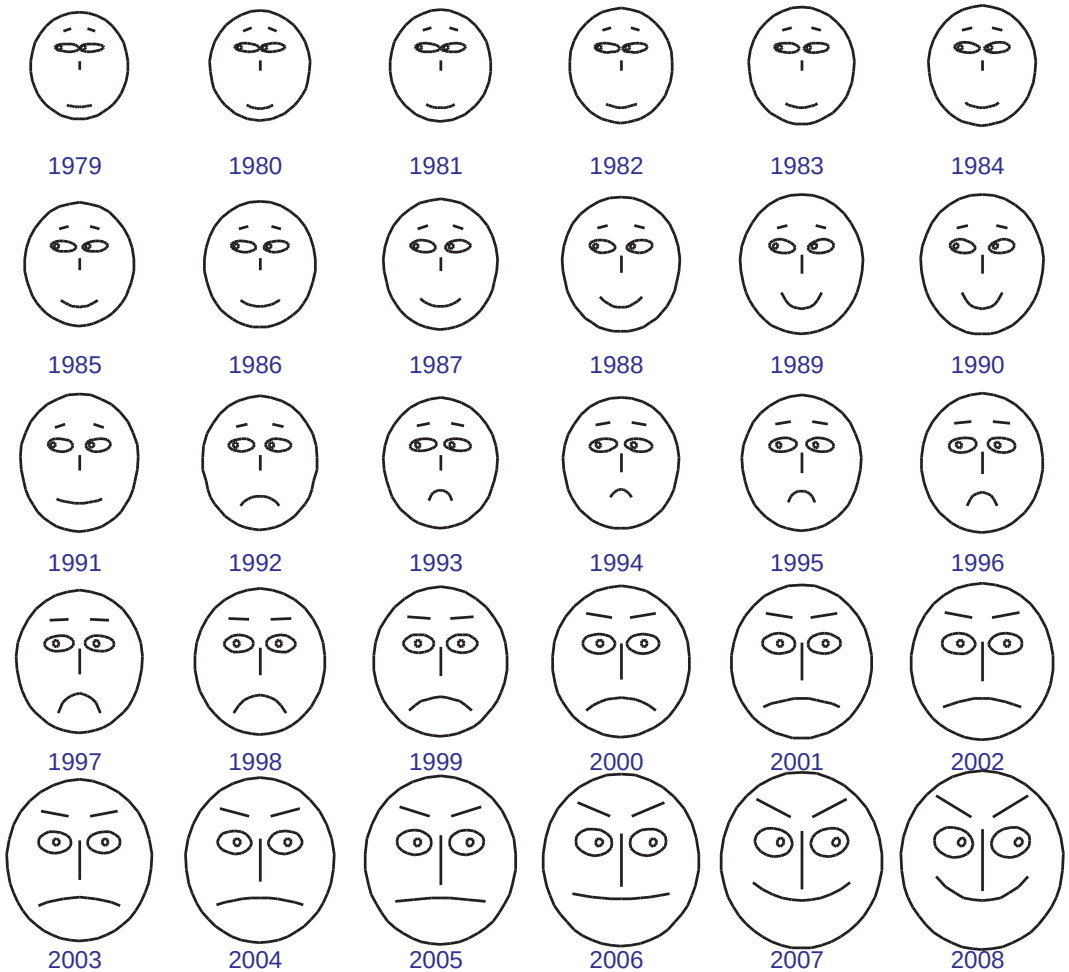
1: Pelatut pelit
2: Tehdyt maalit
3: Annetut maalisyötöt
4: +/- pisteet
5: Jäähyminuutit



6: Ylivoimamaalit
7: Voittomaalit

Example of visualizing time-series of aggregate data

Suomen vointi taloudellisin ilmein 1979 - 2008



Muuttujat: BKT | tuonti | vienti | työttömyysaste | kulutusmenot

Yhteydet kasvonpiirteisiin: *www.helsinki.fi/%7ekvehkala/naamat.html*

Aineiston lähde: Tilastokeskus *Chernoffin naamat:* **SURVO MM**

Some comments on the previous visualization

Type of the graph: **Chernoff's faces**, described in detail by `http://www.helsinki.fi/~kvehkala/naamat.html` *(translation in progress).*

Reference: Chernoff, Herman (1973). The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, **68**, 361–368.

The five variables and their connections with the selected features of the faces:

- Gross Domestic Product
  - Shape of the head and width of the mouth
- Import of products and services
  - Length of the nose
- Export of products and services
  - Size of the eyes and direction of the look
  - Eyebrows (position, slant, size)
- Unemployment rate (men and women)
  - Curvature and vertical position of the mouth
  - Slant of the eyes
- Consumption expenditures (private and public)
  - Size of the head
  - Separation and eccentricity of the eyes

Source of data: Statistics Finland (`http://www.stat.fi/`), drawn by KV using Survo (`http://www.survo.fi`)