Index of Lecture 10: Dimension-reduction techniques (slightly revised on p. 4 and 19)

Practical Information

Today's lecture — classical statistical material on representing a set of variables by a subset containing the majority of the information,

* principal components analysis (PCA; Manly 3/4, Chapter 6; ED, Chapter 3),
* exploratory factor analysis (FA), viewed as an extension of PCA (Manly 3/4, Chapter 7; ED, Chapter 12),

— the purpose of analysis is still mostly descriptive/explorative, but statistical models/assumptions are needed for some methods (in particular for FA).

Links with other ideas/material:

* when linked with regression, PCA and FA may target the same aims as partial least squares (to be discussed briefly later in the course),
* we have already mentioned the equivalence between classical multidimensional scaling and PCA.[1]

Other news:

- home assignment #5 (on multivariate analysis) has been posted, due on Monday,
- project proposals will be returned on Monday, with comments and discussion.

[1] This is explained in a section in Manly 3/4, Chapter 12 on the relation between MDS and PCA. The precise result is that classical MDS with Euclidian distance is equivalent to PCA for the covariance matrix. The eigenvalue decomposition is the same, even if MDS eigenvalues are multiplied by $(n-1)$.

Introduction to Dimension-Reduction Techniques

Synthesis: dimension reduction may be attractive for different purposes,

- ∗ reduce information from multiple variables to a single, or a few, variable(s) in order to facilitate direct use of data, e.g. for decision-making,
- ∗ reduced dimension variable(s) may concentrate certain features of interest in the data,
- ∗ reduced dimension variables may be inputs to further analysis, e.g. predictors for regression-type analyses, where high-dimensional input is infeasible.[2]

Principal components analysis (PCA):

- ∘ one of the oldest (well before computers. . .) and most fundamental multivariate techniques (may be viewed as a building block/template for many other multivariate procedures),
- ∘ based on linear operations of the data (columns $\sim$ variables) $\Rightarrow$ draws heavily on linear algebra theory and formulae.
- ∘ questions remain about the number of PCA components and non-interval variable types.

Exploratory factor analysis[3] (FA) viewed as an addition to PCA to allow easier interpretation of components (my interpretation[4]):

- ∘ contrary to PCA, factor analysis is based on a specific statistical model and its analysis,
- ∘ however, model assumptions are largely untestable, and estimation does not lead to a unique solution (one of the main critiques of FA).

[2] Typically due to a large number of predictors relative to the number of subjects, or to strongly collinear predictors.
[3] We do not discuss confirmatory factor analysis, a methodology related to structural equation modelling and aimed at estimating hypothesized latent structures and their relationships.
[4] Factor analysis is a large and somewhat controversial subject, with the detailed discussions of its interpretation and analytical validity well beyond our scope.

PCA: the Idea

PCA is a mathematical[5] procedure to represent the "information" contained in a number (p) of variables $X_1, \ldots, X_p$, with its covariance (or correlation), by a new set $Z_1, \ldots, Z_p$ of variables such that

- each Z_j is a linear combination of the X_k's,
- the variables $Z_1, \ldots, Z_p$ are independent (uncorrelated/orthogonal),
- the total "variance" explained by the Z_j's equals that explained by the X_k's, and the Z_j's can be ordered by decreasing "variance" explained.

First interpretations:

- the principal components $(Z_j) \sim$ independent "directions" among the (X_k),
- no loss of information by switching: $(X_k) \mapsto (Z_j)$,
- the first (few) components $Z_1, Z_2, \ldots$ that explain most of the "variance",
  - may have useful subject-matter interpretations,
  - may represent a useful data reduction.
- by standardizing the (X_k) prior to PCA, all X_k are brought on same scale, and the decomposition effectively becomes of the original correlation matrix R.

[5] PCA is not based on a statistical model, and does not in itself involve statistical inference.

PCA: the Solution

Given a covariance matrix[6] S ($p \times p$) for the multivariate vector $(X_1, \ldots, X_p)^t$,

Z_j is constructed from an eigenvector for the j^{th} eigenvalue λ_j of S, where

– eigenvalues are ordered as: $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$, and $\sum_k \lambda_k = \text{tr(S)}$,

– $\text{Var}(Z_j) = \lambda_j$, (i.e., Z_j has variance λ_j)

– Z_j is expressed from the (X_k) and the eigenvector $(a_1^{(j)}, \ldots, a_p^{(j)})$ as:

$$Z_j = \sum_k a_k^{(j)} X_k = a_1^{(j)} X_1 + a_2^{(j)} X_2 + \ldots + a_p^{(j)} X_p \,, \quad \text{with } \sum_k (a_k^{(j)})^2 = 1\,, \qquad (1)$$

where the coefficients (also loadings) $(a_1^{(j)}, \ldots, a_p^{(j)})$ are unique up to a sign change.

Notes and interpretations:

- "small" loadings (e.g. $|a_k^{(j)} \leq 0.3|$) are often disregarded for interpretations of the principal components,
- when standardizing the (X_k) prior to PCA, S = correlation matrix R, with tr(R) = p,
- with data on $X_1, \ldots, X_p$, the scores for the principal components are computed from Equation (1) for each observation; these scores:
  - are useful for interpretation of the principal components,
  - may be used as predictors in subsequent regression.

[6] Typically S would be the empirical covariance matrix based on (say) n observations of $(X_1, \ldots, X_p)^t$, but in principle the matrix could have another origin, e.g. a residual covariance matrix.

PCA First Example: Sparrows

Summary: 5 measures on 49 sparrows; survival (0/1) not directly part of PCA.

- standard deviations: range $0.56-5.1 \Rightarrow$ should work with correlation matrix R,
- correlations: range $0.53-0.77 \Rightarrow$ good potential for reducing data dimension,
- table of eigenvalues and loadings for R:

Parameter/ Variable		Component number j 1	2	3	4	5
eigenvalue	λ_j	3.62	0.53	0.39	0.30	0.16
cumulative	$\sum_{k\le j} \lambda_k/5$	0.72	0.83	0.91	0.97	1.00
`len_total`	X_1	0.45	0.05	0.69	-0.42	-0.37
`ext_alar`	X_2	0.46	-0.30	0.34	0.55	0.53
`len_beakhead`	X_3	0.45	-0.33	-0.45	-0.61	0.34
`len_hum`	X_4	0.47	-0.19	-0.41	0.39	-0.65
`len_keelst`	X_5	0.40	0.88	-0.18	0.07	0.19

  - 1^{st} eigenvalue ($\approx$ average of measures $\Rightarrow$ bird size) by far largest, others "ignorable",
  - 2^{nd} component $\sim$ shape feature (X_5 vs. X_2, X_3, X_4), mostly determined by X_5,
- scores for components (e.g. 1^{st} vs. 2^{nd}) plotted against each other (score plot), with survival indicator:
  - no obvious mean differences between survivor groups,
  - however indication of larger spread among non-survivors,
- use of scores as uncorrelated predictors in logistic regression for survival: no significance for any of the scores.

PCA Second Example: Plants in Steneryd

Summary: abundances measured for $p=25$ plant species measured on $n=17$ plots.[7]

- purpose: determine whether plots can be meaningfully characterized by their abundance of plants, in particular in relation to four environmental indicators,
- descriptives for abundances (across plots):
  - measurement range $0-45 \sim$ absence $-$ complete cover in 9 subplots,
  - standard deviations: range $2.9-16.2 \sim$ different abundances,
  - correlations: range $(-0.79, 0.89) \Rightarrow$ potential for reducing data dimension,
- table of eigenvalues for the correlation matrix R, and their proportions out of the total ($p=25$):

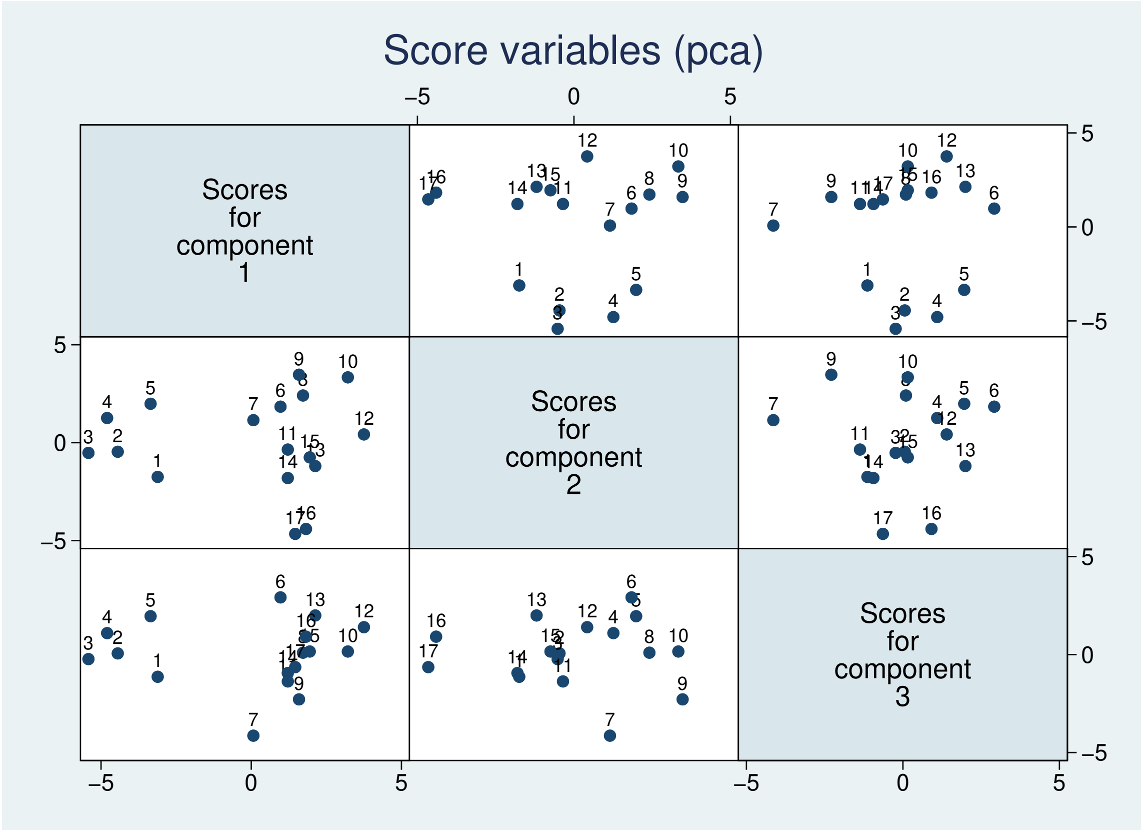
component j	eigenvalue λ_j	proportion $\lambda_j/25$	cumulative prop.
1	8.792	0.352	0.352
2	5.585	0.223	0.575
3	2.955	0.118	0.693
4	1.929	0.077	0.770
5	1.581	0.063	0.834
6	1.131	0.045	0.879
7	0.993	0.040	0.919
...	...	...	...
16	0.051	0.002	1.000

note: $\text{rank}(R) \leq \min(n-1, p)$ = upper bound for components.

[7] Study area: Steneryd Nature Reserve in the province of Blekinge, South Sweden, at one time a park-meadow but left essentially as a wilderness during 65 years before sampling, see also Additional Multivariate Exercise 6.

PCA Scores for Plants in Stenerydd

Scores give the new uncorrelated directions of the data ($\sim$ vectors of size n) — matrix score plot for first 3 components (for simplicity):

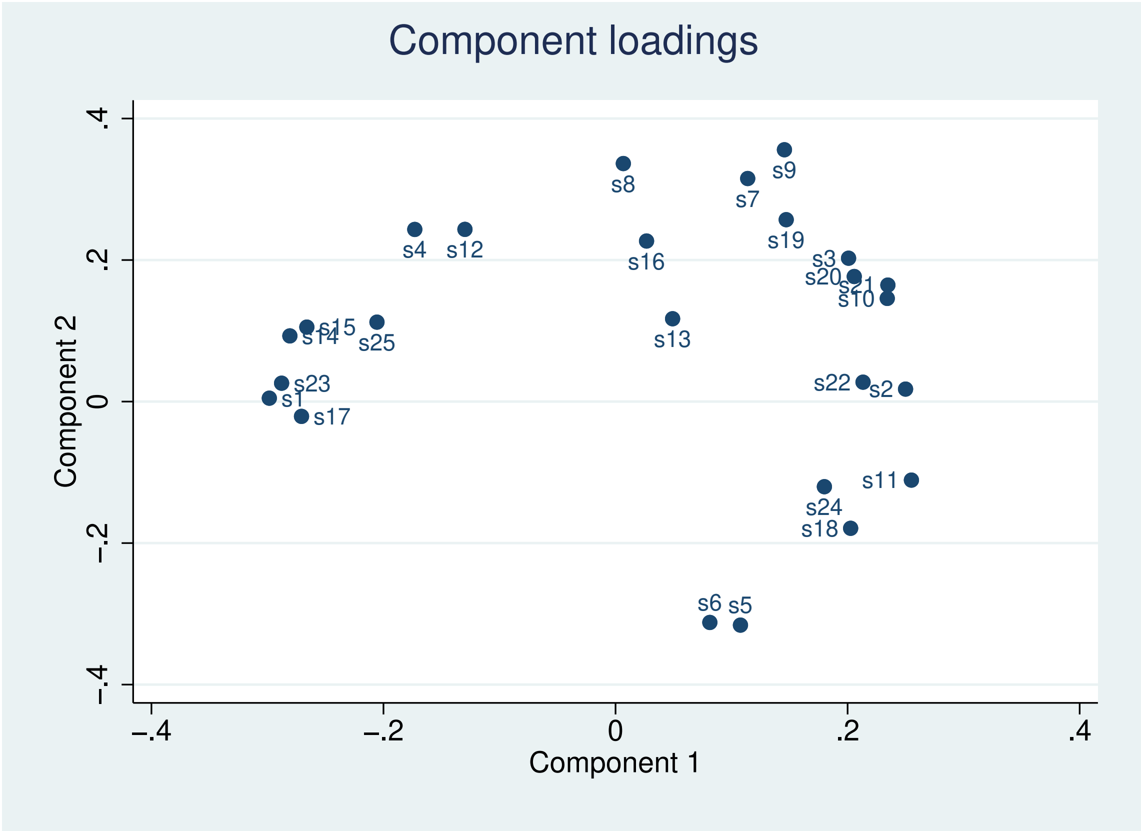


- first component appears directly related with plot order ($1-15$),
- associations may be explored with observation characteristics, e.g. here the environmental variables for the plots (see Stata do-file).

PCA Loadings for Plants in Stenerydd

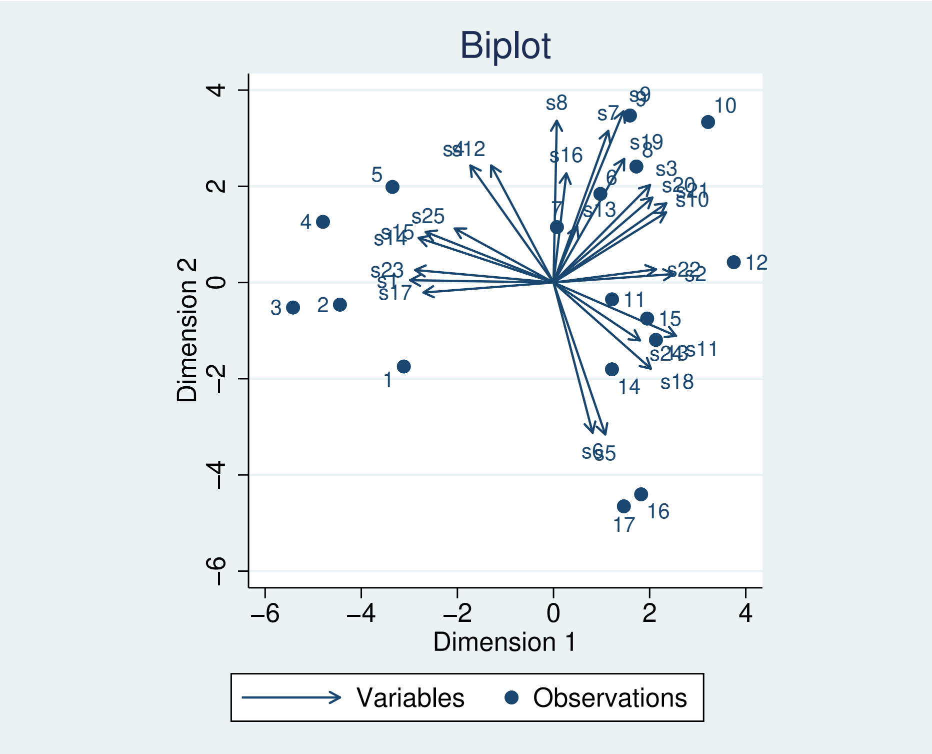
Loadings for the principal components (in Eq. (1), $a_k^{(j)}$ for component j and variable k) give the weights of the original variables in the components ($\sim$ vectors of size p) — loadings plot for first 2 components (for simplicity):

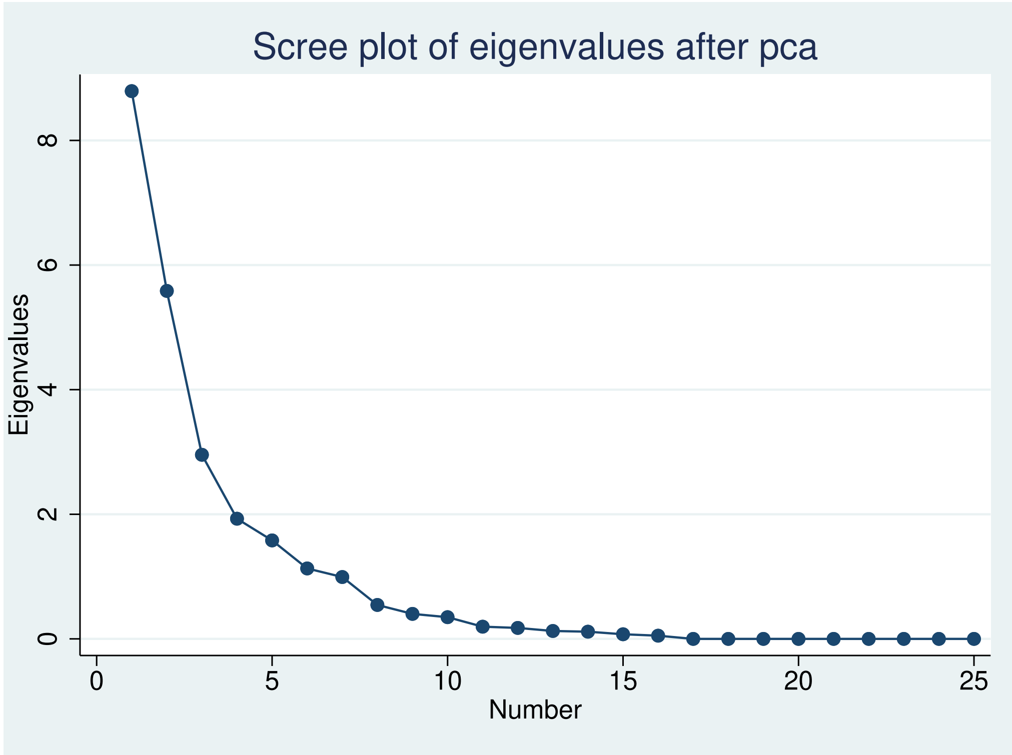
- first component gives negative weights for a few species (e.g., $\mathbf{1, 17, 23}$) with abundances concentrated on first plots (dry, light),
- strongest weights on second component $\sim$ contrast of species $\mathbf{5, 6}$ with species $\mathbf{7, 8, 9}$,
- loadings not numerically very strong (i.e., few beyond $\pm\mathbf{0.3}$).



PCA Biplot and Scree plot for Plants in Stenéryd

The biplot for a PCA displays the scores and loadings in the same plot.[8]





The scree plot depicts eigenvalues against components, as an aid to select a suitable number of components to retain (next slide).

[8] Other uses/options for the biplot exist which we will not cover here.

PCA Choices

Main choice to make: number of principal components to retain; some considerations:

- balance simplicity of analysis/exploration with loss of information,
- essentially always a subjective choice,
- some possible "rules" for deciding on # components:
  - explained correlation (for standardized variables) ≤ 1,[9]
  - cumulative % of explained variance/correlation $\geq 80\%$ (or 90%),
  - kink/bend/elbow in the scree plot (when it starts to flatten).

Standardizing the variables? — essentially a choice between the covariance matrix S or the correlation matrix R:

- default is for correlation R (also the typical software default), unless variables are on the same scale and differences in their magnitude is of actual interest.[10]

Variables to include:

- avoid "noise" variables (with no relation to scientific question),
- variables with low correlations with all other variables may be ignored in PCA.[11]

[9] The rationale being, that components that explain less than any of the original variables should be of less interest.
[10] The Stata manual has this statement about the choice of scale: "In applied research, PCA of a covariance matrix is useful only if the variables are expressed in commensurable units."
[11] In the extreme case of a variable being uncorrelated with all others, it will retain its own component in PCA.

PCA Limitations and Practical Issues[12]

- sensitivity to problems with correlations:
  - outlying observations,
  - non-linear associations,
  - unnatural ranges arising from sampling (too narrow or too wide),
- sample size recommendations: "good" starting at 300 observations, under certain circumstances "acceptable" down to 50,
- normality: no specific assumption (beyond quantitative scale) involved for PCA, but correlations will be most meaningful for continuous variables with roughly symmetrical distributions,
- outliers among cases: assessment may be based on multivariate (Mahanalobis[13]) distance,
- known structure may need to be removed first[14]; this may also apply to hierarchical structure, if the main interest lies in correlations at the lowest level.

[12] Based largely on TF, Section 13.3.2.

[13] The Mahanalobis distance (further discussed in Session 11) between a distribution (based on sample mean μ and sample covariance Σ) and a point x is:

$$d_M(\mathrm{x}, \mu, \Sigma) = (\mathrm{x} - \mu)^t \Sigma^{-1} (\mathrm{x} - \mu).$$

[14] "PCA will reveal the gross features of the data, which may already be known, and is often best applied to residuals after the known structure has been removed."; Venables & Ripley (2002), p. 304.

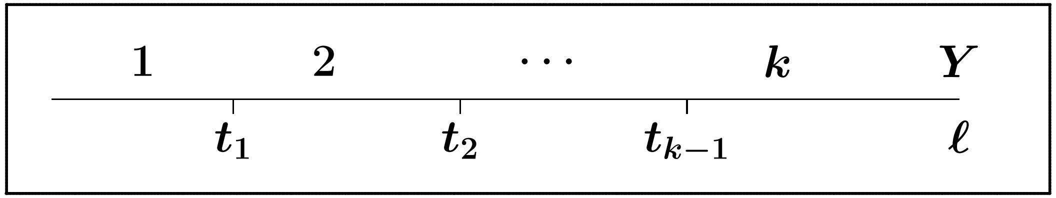
Choric Correlations

Illustration of issue: the association between two binary variables Y_1 and Y_2 is not measured well by the Pearson correlation,

$$\rho(Y_1, Y_2) = \frac{\mathrm{E}(Y_1 Y_2) - \mathrm{E}(Y_1)\mathrm{E}(Y_2)}{\mathrm{sd}(Y_1)\mathrm{sd}(Y_2)} = \frac{\Pr(Y_1 = Y_2 = 1) - \Pr(Y_1 = 1)\Pr(Y_2 = 1)}{\sqrt{p_1(1-p_1)p_2(1-p_2)}},$$

where $p_k = \Pr(Y_k = 1)$, $k = 1, 2$, because $\rho(Y_1, Y_2)$ has a much more restricted range than the usual $-1 \leq \rho \leq 1$.[15]

Latent variables underlying ordinal observations:
a k-category ordinal outcome Y can be thought of as corresponding to thresholding of an underlying ℓ at cut-off values $t_1, \ldots, t_{k-1}$:[16]



Assuming normal distributions for such underlying variables, we may now estimate (Pearson) correlations between:

- latent variables (ℓ_1, ℓ_2) for two binary variables (Y_1, Y_2) (tetrachoric correlation),
- latent variables (ℓ_1, ℓ_2) for ordinal categorical variables (Y_1, Y_2) (polychoric correlation),
- a latent variable ℓ and an observed quantitative variable (polyserial correlation),

— in all cases using iterative methods for ML estimation of the correlation.

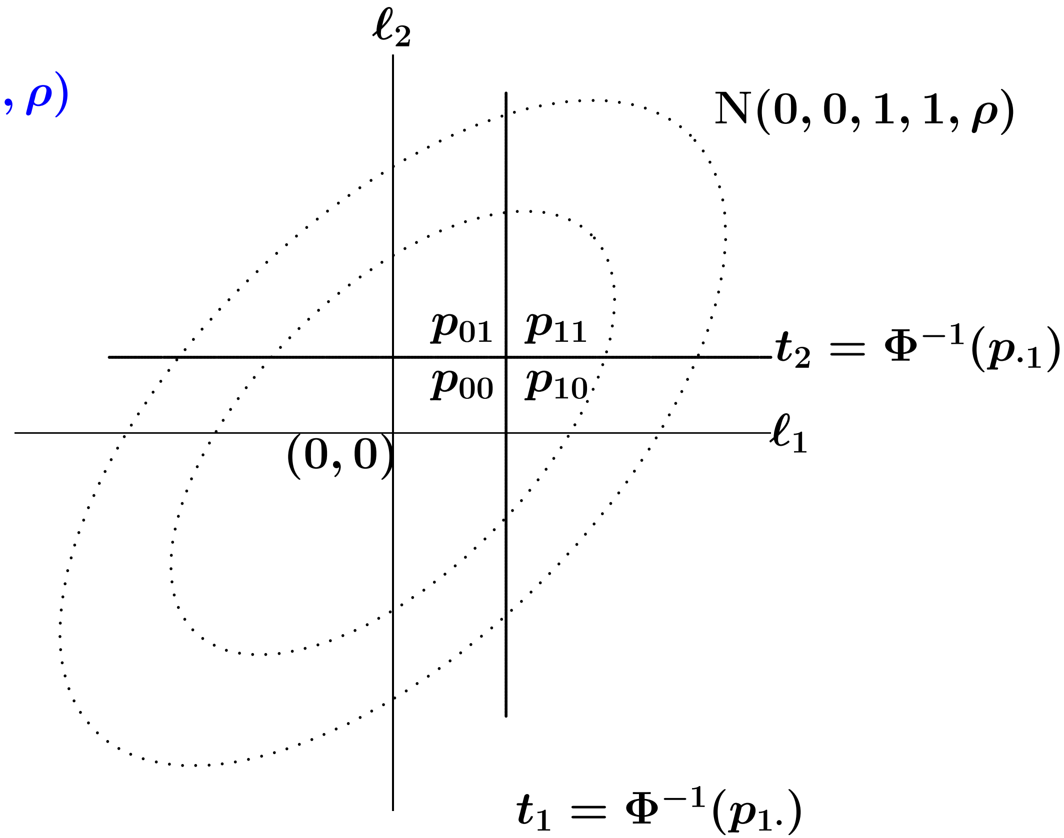
[15] For example, if $p_1 = 0.8$ and $p_2 = 0.6$ then the possible range for ρ is $(-0.41, 0.61)$.
[16] Same idea as in (ordinal) logistic regression models.

Illustration for Tetrachoric Correlation

If two binary variables Y_1 and Y_2 are represented by the probabilities,

	$Y_2=1$	$Y_2=0$	margin
$Y_1=1$	p_{11}	p_{10}	$p_{1\cdot}$
$Y_1=0$	p_{01}	p_{00}	$p_{0\cdot}$
margin	$p_{\cdot 1}$	$p_{\cdot 0}$	1

estimation of the tetrachoric ρ $\sim$ matching the (standard) bivariate normal $N(0,0,1,1,\rho)$ elliptic contours to the probabilities of the table:[17]



[17] Following Brown (1977), Olsson (1979); the MLE is of (t_1, t_2, ρ) or ρ only.

PCA with Polychoric Correlations

Proposed approach to deal with binary/ordinalvariables for PCA (and similar methods):[18]

(1) compute correlation matrix for full set of variables, where all correlations involving ordinal categorical variables will be estimated based on latent variables (i.e., will be either tetrachoric, polychoric or polyserial),

   * requires software implementation (e.g. Stata, R; not in Minitab),
   * resulting correlation matrix may not be positive semidefinite, but usually not a problem if some small eigenvalues are negative (smoothing of the correlation matrix has also been suggested),

(2) perform PCA (or other analysis) from correlation matrix (instead of the variables),

(3) manually compute scores from the eigenvector loadings obtained.

Assessment:

+ puts correlations involving ordinal variables at better scale for quantification and interpretation,
+ avoids assumptions involving labeling of ordinal categories,
– latent variables may not have a meaningful interpretation as a degree of measurement (e.g., gender, conception),
– scores will be less "correct", because the latent variables are not available.

[18] Following Everitt & Dunn (2001), or Uebersax (2015): `http://john-uebersax.com/stat/tetra.htm`.

Polychoric PCA Example: CALF Data

Summary: 11 measurements (actually a subset) obtained at admission for 254 calves admitted to the veterinary hospital at AVC (during 1989-93).

- objective: to predict whether (or not) the calf was septic at the time admission,[19]
- quantitative variables: `age`, `dehy`, `pulse`, `resp`, `temp`,
- binary variables: `sex`, `eye`, `umb`,
- ordinal variables: `attd`(3), `jnts`(0−4), `post`(3).

Pearson (below diagonal) and tetrachoric/polychoric/polyserial correlations (above diagonal) for selected variables:

	age	dehy	sex	eye	attd	jnts
age	1	-.126	.010	.044	-.141	.001
dehy	-.126	1	-.146	-.099	.382	-.030
sex	.008	-.116	1	.331	-.045	.048
eye	.018	-.034	.089	1	.150	.596
attd	-.118	.320	-.030	.041	1	.084
jnts	-.010	-.020	-.007	.221	.023	1

- the non-Pearson correlations are mostly numerically larger, in particular with binary variables.

[19] The final sepsis diagnosis in the data was obtained later based on all data available at the time of death or discharge.

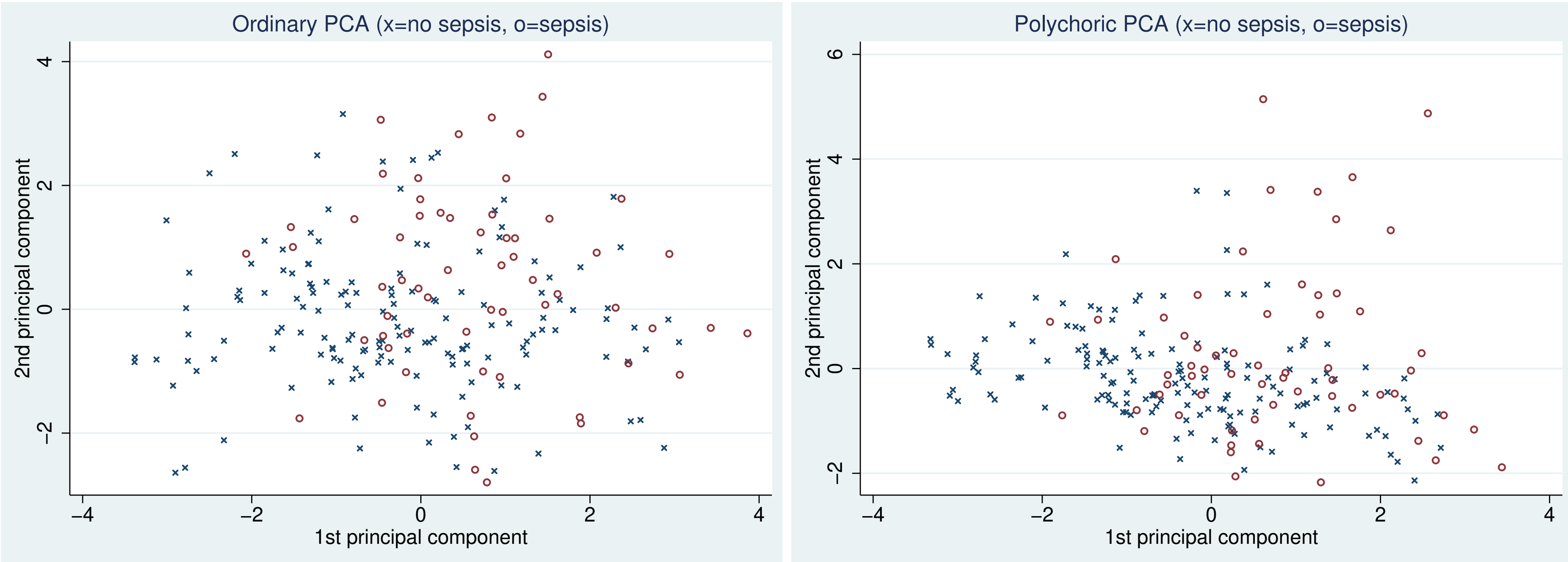
Ordinary and Polychoric PCA: calf Data

Comparison of first 4 eigenvalues and loading sets for ordinary and polychoric PCA:

	Ordinary PCA				Polychoric PCA			
j	1	2	3	4	1	2	3	4
λ_j	2.21	1.65	1.32	1.03	2.48	2.14	1.50	1.02
$\sum_{k \le j} \lambda_k / 11$	.200	.351	.470	.564	.226	.420	.557	.650
age	-.10	-.44	.31	-.05	-.11	-.09	.53	-.12
sex	-.10	.02	.33	.71	-.06	.25	.31	.54
attd	.54	.14	.05	.18	.58	.01	-.03	.13
dehy	.43	-.03	-.13	-.21	.37	-.19	-.09	-.32
eye	.02	.19	.54	-.14	.15	.51	.28	.11
jnts	.01	.24	.51	-.48	.10	.50	.17	-.44
post	.56	.14	.08	.09	.59	.03	.01	.13
pulse	-.16	.44	-.08	-.24	-.09	.30	-.35	-.44
resp	.06	.51	-.30	.13	.11	.16	-.56	.33
temp	-.41	.35	-.09	.07	-.32	.32	-.27	.20
umb	-.01	.29	.33	.28	.07	.40	.03	-.09

- slightly larger eigenvalues with polychoric correlations; (not shown): only 6 eigenvalues needed to reach 80% of total (7 components for ordinary PCA),
- first component very similar between the two PCAs,
- also some similarity among fourth components,
- other components (2,3) not the same.

Plots for Ordinary and Polychoric PCA



Separation of septic/non-septic calves by components 1−3 (mean (SE) of score diff.):

- **polychoric PCA:** $\Delta s_1 = 1.09$ (.19), $\Delta s_2 = 0.44$ (.22), $\Delta s_3 = 0.15$ (.19),
- **ordinary PCA:** $\Delta s_1 = 0.98$ (.20), $\Delta s_2 = 0.64$ (.22), $\Delta s_3 = 0.40$ (.22),

— ordinary PCA has smaller separation for first component, but larger separations for second and third components; also, logistic regression fits with 3 score predictors are similar for the two PCAs.

Factor Analysis Model and Notation

With n observations on each of p variables $X_1, \ldots, X_p$, the factor analysis model assumes the existence of $k < p$ (latent) variables (factors) $F_1, \ldots, F_k$ and "errors" $U_1, \ldots, U_p$ such that

$$\begin{aligned} X_1 &= A_{11}F_1 + A_{12}F_2 + \ldots + A_{1k}F_k + U_1, \\ X_2 &= A_{21}F_1 + A_{22}F_2 + \ldots + A_{2k}F_k + U_2, \\ &\ldots \\ X_p &= A_{p1}F_1 + A_{p2}F_2 + \ldots + A_{pk}F_k + U_p, \quad \text{or} \quad \mathrm{X} = \mathrm{AF} + \mathrm{U}, \end{aligned}$$

where F, U have n columns (just as X), and all factors and errors are uncorrelated,

- A_{jl} is the loading of factor l on variable j,
- the factors and errors can be taken with mean zero,
- the factors can be taken with variance 1, whereas $\mathrm{Var}(U_j) = \psi_j$ is the unique variance, or uniqueness, for X_j,
- the communality for X_j is the variance explained by the factors (i.e., excluding ψ_j): $\mathrm{Var}(X_j) - \psi_j = \sum_{l=1}^{k} A_{jl}^2$,[20]
- several methods exist for computing factor scores, the most common uses a formula akin to regression.[21]

[20] When X's are standardized, obviously $\mathrm{Var}(X_j) = 1$.
[21] Under normality, the conditional mean of F given X is estimated by $\mathrm{A}^t\mathrm{S}^{-1}\mathrm{X}$.

Estimation for Factor Analysis

Multiple estimation approaches exist; a brief summary of some of the more commonly used approaches:

- principal components: select the k first components (Z_l, $l = 1, \ldots, k$) from a PCA, and let the factors be those components scaled to unit variance: $F_l = Z_l/\sqrt{\lambda_l}$, and therefore inversely scaled loadings: $A_{jl} = \sqrt{\lambda_l}\, a_j^{(l)}$ (from Equation (1));
  note: this essentially enforces high communality and low uniqueness (depending on the number of components included, k),
- principal factors: the communality (ψ_j) for X_j is initially set at the multiple R^2 for a regression on the other variables, then the factors are estimated, and ψ_j is reestimated with the selected ($k < p$) factors,[22],
- maximum likelihood estimation based on assumed normality of both F and U (hence also X); has the advantage of offering statistics/inference about model fit (and hence choice of k),

— Manly (2005) suggests the principal components as initial factors, with a varimax rotation (next slide), as "a reasonable start with any set of data".

[22] In a variation of the method (in Stata termed: iterative principal factors), communalities are reestimated iteratively (in some way).

Factor Rotation

Fact: factors and loadings are not unique — any rotation (linear, invertible transformation) of **F** and **A** will produce the same fit (errors **U**):

- purpose of rotation: improve interpretability of solution,
- some orthogonal[23] rotations:
  - varimax: maximize variance of loadings on each factor ∼ simplify columns of loading matrix (**A**); most common form,
  - quartimax: maximize variance of loadings on each variable ∼ simplify rows of loading matrix,
  - equimax: a compromise between the varimax and quartimax rotations,
- oblique (non-orthogonal) rotations are also possible, in which case the resulting factors will no longer be uncorrelated and variance will hence be shared between factors; most common type is oblimin,
- rotation of PCA components is also being used, though not unproblematic and perhaps most sensible for components with close eigenvalues (and hence ill-determined), Jolliffe (1989).

[23] Orthogonal transformations preserve the factors as uncorrelated.

FA First Example: Sparrows

Factor analysis with $k = 2$ principal components, unrotated and varimax rotation:

- table of eigenvalues and loadings for R:

Factor loadings A_{jl} / Variable j	Component l 1	Component l 2	Rotated comp. l 1	Rotated comp. l 2	Communality	Uniqueness
`len_total`	0.86	0.04	0.71	0.48	0.74	0.26
`ext_alar`	0.88	-0.22	0.86	0.27	0.82	0.18
`len_beakhead`	0.86	-0.24	0.85	0.25	0.79	0.21
`len_hum`	0.90	-0.14	0.83	0.35	0.82	0.18
`len_keelst`	0.76	0.64	0.31	0.94	0.98	0.02
variance	3.62	0.53	2.77	1.38		
proportion variance	0.72	0.11	0.55	0.28		
cumulative prop.	0.72	0.83	0.55	0.83		

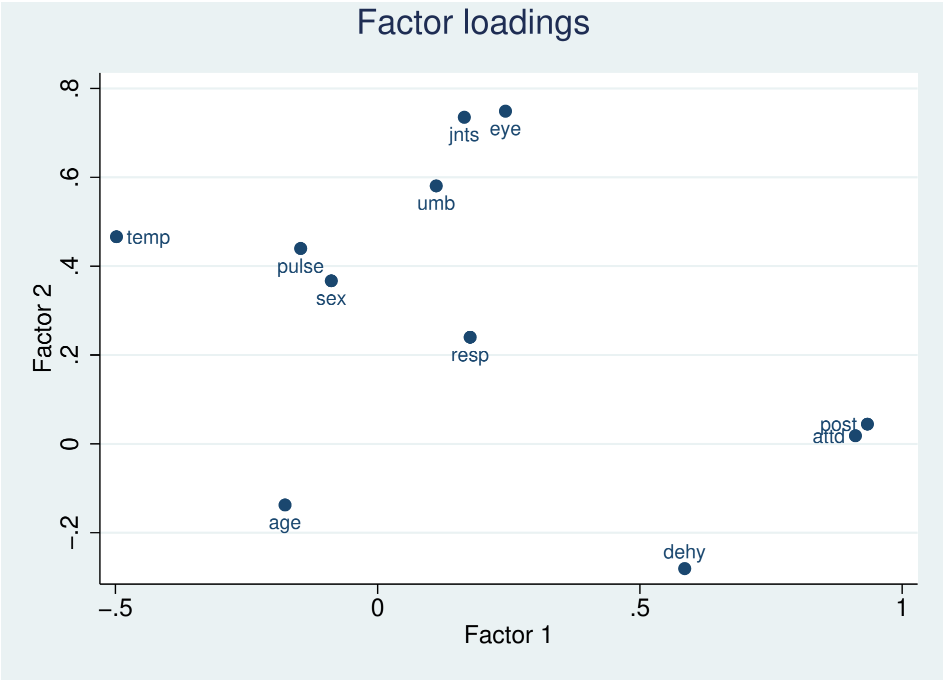
  - the first two components explain 83% of the variance (as before),
  - the factor loadings are the PCA loadings multiplied by $\sqrt{\lambda_l}$,
  - the communalities (ψ_j) are unaffected by rotations,
  - the uniqueness equals (for correlation matrix): $1 -$ communality,

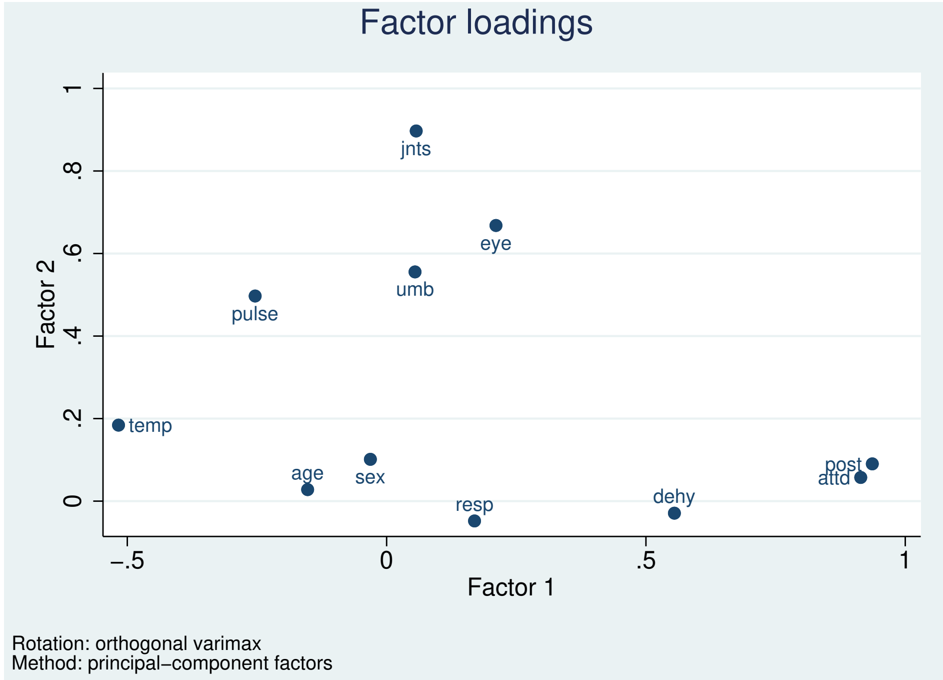
- rotation effects are most easily seen in a loading plot:

  the loading of `len_keelst` has been reduced in factor 1 and increased in factor 2, in order to separate the factors better from each other.

FA ROTATIONS FOR CALF DATA

Loading plots for first two (polychoric principal component) factors in a 4-factor model, before (left) and after varimax rotation (right):





- the first factor is pretty much unchanged,
- the second factor has its non-zero loadings concentrated on less variables after the rotation.

Factor Analysis versus PCA

Despite the similarity between methods, some further points on which they can be distinguished:[24]

- the focus of PCA is to decompose variance; for FA it is on explaining covariances/correlations (by underlying factors),
- in PCA, all variance is distributed on components, whereas FA allows for unique (non-shared) variance for each variable,
- in PCA, adding a component does not change the original components; after FA rotation of an augmented set of factors there may be little resemblance with the original factors,
- with ML estimation, the FA results from the covariance and correlation matrix are essentially equivalent, contrasting PCA and other FA estimation procedures.

Summary of comparison (TF):

> If you are interested in a theoretical solution uncontaminated by unique and error variability and have designed your study on the basis of underlying constructs that are expected to produce scores on your observed variables, FA is your choice. If, on the other hand, you simply want an empirical summary of the data set, PCA is the better choice.

[24] Following ED, Section 12.8; TF, Section 13.5.1.1.