

UCLA STAT 251
Statistical Methods for the Life and Health Sciences

● **Instructor: Ivo Dinov,**
 Asst. Prof. In Statistics and Neurology

University of California, Los Angeles, Winter 2002
<http://www.stat.ucla.edu/~dinov/>

STAT 251, UCLA, Ivo Dinov Slide 1

Principle Component Analysis (PCA)

● Introduction to PCA

STAT 251, UCLA, Ivo Dinov Slide 2

PCA - purpose

- The main applications of PCA analytic techniques are:
 - to *reduce* the number of variables and
 - to *detect structure* in the relationships between variables, that is to *classify variables*.
- Therefore, PCA analysis is applied as a data reduction or structure detection method (the term *factor analysis* was first introduced by Thurstone, 1931). The topics listed below will describe the principles of factor analysis, and how it can be applied for addressing these two goals.

Slide 3 STAT 251, UCLA, Ivo Dinov

PCA - Basics

- Suppose we conducted a (rather "silly") study in which we measure 100 people's height in inches and centimeters. Thus, we would have two variables that measure height. If in future studies, we want to research, for example, the effect of different nutritional food supplements on height, would we continue to use both measures? Probably not; height is one characteristic of a person, regardless of how it is measured, since there are measure-conversion rules!
- Let us now extrapolate from this trivial study to something that one might actually do as a researcher. Suppose we want to measure people's satisfaction with their lives. We design a satisfaction questionnaire with various items; among other things we ask our subjects how satisfied they are with their hobbies (item 1) and how intensely they are pursuing a hobby (item 2). Most likely, the responses to the two items are highly correlated with each other. Given a high correlation between the two items, we can conclude that they may be quite **redundant**.

Slide 4 STAT 251, UCLA, Ivo Dinov

PCA - Basics

- **Combining Two Variables into a Single Factor.**
 Correlations between two variables may be summarized in a scatterplot. A regression line can then be fitted that represents the "best" summary of the linear relationship between the variables. If we could define a variable that would approximate the regression line in such a plot, then that variable would capture most of the "essence" of the two items. Subjects' single scores on that new factor, represented by the regression line, could then be used in future data analyses to represent that essence of the two items. In a sense we have reduced the two variables to one factor. Note that the new factor is actually a linear combination of the two variables.



Slide 5 STAT 251, UCLA, Ivo Dinov

PCA - Basics

- **Principal Components Analysis.** The example described above, combining two correlated variables into one factor, illustrates the basic idea of factor analysis, or of principal components analysis to be precise. If we extend the two-variable example to multiple variables, then the computations become more involved, but the basic principle of expressing two or more variables by a single factor remains the same.

Slide 6 STAT 251, UCLA, Ivo Dinov

PCA - Basics

- Computational aspects of principal components analysis: Study of the extraction of principal components amounts to a *variance maximizing (varimax) rotation* of the original variable space. For example, in a scatterplot we can think of the regression line as the **original X axis, rotated so that it approximates the regression line**. This type of rotation is called *variance maximizing* because the criterion for (goal of) the rotation is to:
 - maximize the variance (variability) of the "new" variable (factor), while
 - minimizing the variance around the new variable.

Slide 7 STAT 251, UCLA, Joe Dimeo

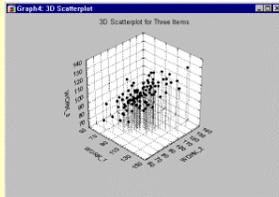
PCA - Basics

- **Generalizing to the Case of Multiple Variables.** When there are more than two variables, we can think of them as defining a "space," just as two variables defined a plane. Thus, when we have three variables, we could plot a three-dimensional scatterplot, and, again we could fit a plane through the data.

Slide 8 STAT 251, UCLA, Joe Dimeo

PCA – scatterplot example

- With more than 3 variables it becomes impossible to illustrate the points in a scatterplot, however, the logic of rotating the axes so as to maximize the variance of the new factor remains the same. But up to 3-variables we can use a scatterplot:



Slide 9 STAT 251, UCLA, Joe Dimeo

PCA - Basics

- **Multiple orthogonal factors.** After we have found the line on which the variance is maximal, there remains some variability around this line. In PCA, after the first factor has been extracted, that is, after the first line has been drawn through the data, we **iteratively** continue to define others line that **maximize the remaining variability**. In this manner, consecutive factors are extracted. Because each **consecutive factor is defined to maximize the variability that is not captured by the preceding factor**, consecutive factors are **independent** of each other. Put another way, consecutive factors are uncorrelated or **orthogonal** to each other.

Slide 10 STAT 251, UCLA, Joe Dimeo

PCA - Basics

- **How many Factors to Extract?** Remember that, so far, we are considering PCA as a data reduction method, that is, as a method for reducing the number of variables. The question then is, how many factors do we want to extract? Note that as we extract consecutive factors, they account for less and less variability. The decision of when to stop extracting factors basically depends on when there is only very little **random** variability left. The nature of this **decision is arbitrary**; however, various guidelines have been developed.

Slide 11 STAT 251, UCLA, Joe Dimeo

PCA - Basics

- **Standard results from a PCA analysis:** We are extracting factors that account for less and less variance. To simplify matters, one usually starts with the correlation matrix, where the variances of all variables are equal to 1.0. Therefore, the total variance in that matrix is equal to the number of variables. For example, if we have 10 variables each with a variance of 1 then the total variability that can potentially be extracted is equal to 10 times 1. Suppose that in the **life-satisfaction study** introduced earlier we included 10 items to measure different aspects of satisfaction at home and at work. The variance accounted for by successive factors would be summarized as follows:

Slide 12 STAT 251, UCLA, Joe Dimeo

PCA - Example

Extraction: Principal components

Value	Eigenvalue %	Total Variance	Cumulative Eigenvalue	Cumulative %
1	6.118369	61.18369	6.11837	61.1837
2	1.800682	18.00682	7.91905	79.1905
3	.472888	4.72888	8.39194	83.9194
4	.407996	4.07996	8.79993	87.9993
5	.317222	3.17222	9.11716	91.1716
6	.293300	2.93300	9.41046	94.1046
7	.195808	1.95808	9.60626	96.0626
8	.170431	1.70431	9.77670	97.7670
9	.137970	1.37970	9.91467	99.1467
10	.085334	0.85334	10.00000	100.0000

Slide 13 STAT 251, UCLA, Joe Dimeo

PCA - Eigenvalues

- **Eigenvalues** - In the second column (*Eigenvalue*) above, we find the variance on the new factors that were successively extracted. In column 3, these values are expressed as a percent of the total variance (in this example, 10). As we can see, factor 1 accounts for 61 percent of the variance, factor 2 for 18 percent, and so on. As expected, the sum of the eigenvalues is equal to the number of variables. The third column contains the cumulative variance extracted. The variances extracted by the factors are called the *eigenvalues*. This name derives from the computational issues involved.

Slide 14 STAT 251, UCLA, Joe Dimeo

PCA - Eigenvalues

- **Eigenvalues and the Number-of-Factors Problem:** Now that we have a measure of how much variance each successive factor extracts, we can return to the question of how many factors to retain. By its nature this is an arbitrary decision. However, there are some guidelines that are commonly used, and that, in practice, seem to yield the best results.

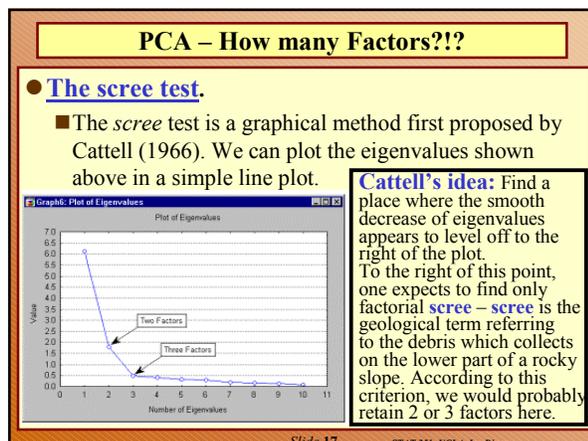
Slide 15 STAT 251, UCLA, Joe Dimeo

PCA - How many Factors?!?

- **The Kaiser criterion.**
 - First, we can retain only factors with eigenvalues greater than 1. In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, we drop it. This criterion was proposed by Kaiser (1960), and is probably the one most widely used. In our example above, using this criterion, we would retain 2 factors (principal components).

Value	Eigenvalue %	Total Variance	Cumulative Eigenvalue	Cumulative %
1	6.118369	61.18369	6.11837	61.1837
2	1.800682	18.00682	7.91905	79.1905
3	.472888	4.72888	8.39194	83.9194

Slide 16 STAT 251, UCLA, Joe Dimeo



PCA

- **Which criterion to use?**
 - Both criteria have been studied in detail (Browne, 1968; Cattell & Jaspers, 1967; Hakstian, Rogers, & Cattell, 1982; Linn, 1968; Tucker, Koopman & Linn, 1969).
 - It appears as if the first method (*Kaiser criterion*) sometimes retains too many factors, while the second technique (*scree test*) sometimes retains too few;
 - Both do quite well under normal conditions, that is, when there are relatively few factors and many cases.
 - In practice, an additional important aspect is the extent to which a solution is interpretable. Therefore, one usually examines several solutions with more or fewer factors, and chooses the one that makes the best "sense."

Slide 18 STAT 251, UCLA, Joe Dimeo

PFA – Principal Factors Analysis

● Principal Factors Analysis

- Recall our satisfaction questionnaire example. Let's slightly perturb this model for factor analysis. We can think of subjects' responses as being dependent on two components.
 - First, there are some underlying common factors, such as the "satisfaction-with-hobbies" factor we looked at before. Each item measures some part of this common aspect of satisfaction.
 - Second, each item also captures a unique aspect of satisfaction that is not addressed by any other item.

Slide 19 STAT 251, UCLA, Joe Dimeo

PCA vs. PFA

● Principal factors vs. principal components

- To distinguish between the two factor analytic models is that:
 - **PCA** assumes that *all* variability in an item should be used in the analysis,
 - **PFA** (principal factors analysis) we only use the variability in an item that it has in common with the other items.
- In most cases, these two methods usually yield very similar results. However, principal components analysis is often preferred as a method for data reduction, while principal factors analysis is often preferred when the goal of the analysis is to detect structure.

Slide 20 STAT 251, UCLA, Joe Dimeo

PCA

● Factor Analysis as a Classification Method

- We'll use the term *factor analysis* generically to encompass both PCA & PFA.
- Assume now that we are at the point in our analysis where we basically know how many factors to extract. We may now want to know the **meaning of the (composite) factors**. How can we interpret them in a meaningful manner?
- Let's try to **work backwards**, that is, begin with a meaningful structure and then see how it is reflected in the results of a factor analysis.
- In our satisfaction example, here is the **correlation matrix for items** pertaining to satisfaction at work and items pertaining to satisfaction at home.

Slide 21 STAT 251, UCLA, Joe Dimeo

PCA

● Factor Analysis as a Classification Method

- Symmetry of the Covariance matrix? Why?

Variable	Work1	Work2	Work3	Home1	Home2	Home3
Work1	1.00	0.65	0.65	0.14	0.15	0.14
Work2	0.65	1.00	0.73	0.14	0.18	0.24
Work3	0.65	0.73	1.00	0.16	0.24	0.25
Home1	0.14	0.14	0.16	1.00	0.66	0.59
Home2	0.15	0.18	0.24	0.66	1.00	0.73
Home3	0.14	0.24	0.25	0.59	0.73	1.00

Slide 22 STAT 251, UCLA, Joe Dimeo

PCA

● Factor Analysis as a Classification Method

- The **work satisfaction items** are highly correlated amongst themselves, and
- the **home satisfaction items** are highly inter-correlated amongst themselves, too.
- The **correlations across** these two types of items (work vs. home satisfaction items) is **comparatively small**. Are there only **two relatively independent factors** reflected in the correlation matrix?
 - one related to satisfaction at work,
 - the other related to satisfaction at home.

Slide 23 STAT 251, UCLA, Joe Dimeo

PCA

- PCA analysis of the two-factors. Specifically, look at the correlations between the variables and the two factors (or "derived" variables), as they are extracted.

Variable	Factor 1	Factor 2
Work1	0.654384	0.564143
Work2	0.715256	0.541444
Work3	0.741688	0.508212
Home1	0.634120	-0.563123
Home2	0.706267	-0.572658
Home3	0.707446	-0.525602
ExplainedVar	2.891313	1.791000
ProportTotal	0.481885	0.298500

Slide 24 STAT 251, UCLA, Joe Dimeo

PCA

- Apparently, the **factor 1** is generally more highly correlated with the original variables than the **factor 2**. This is to be expected because, as previously described, these factors are extracted successively and will account for less and less variance overall.

Slide 25 STAT 251, UCLA, Joe Dimeg

PCA

- **Interpreting the Factor Structure.**
 - As expected, the **first factor** is marked by high loadings on the **work satisfaction** items.
 - **Factor 2** is marked by high loadings on the **home satisfaction** items. We would thus conclude that satisfaction, as measured by our questionnaire, is composed of those two aspects;
 - And there is our **classification** of the variables.
- Consider another example, this time with **4 additional** Hobby/Misc variables added to our earlier example.
- The initial 10 variables were reduced to 3 specific factors, a **work factor**, a **home factor** and a **hobby/misc. factor**. Note that factor loadings for each factor are spread out over the values of the other two factors but are high for its own values. For example, the factor loadings for the hobby/misc variables (in green) have both high and low "work" and "home" values, but all four of these variables have high factor loadings on the "hobby/misc" factor.

Slide 26 STAT 251, UCLA, Joe Dimeg

PCA

- **If time permits do another example.**
 - SYSTAT.
 - C:\vo.dir\Research\Data.dir\WM_GM_CSF_tissueMaps.dir\
 - ATLAS_IVO_all.xls
 - PCA: Data Reduction → Factor Analysis
 - Var's: Hemi. Tisse, Method, Value

Slide 27 STAT 251, UCLA, Joe Dimeg