Lecture 16: Conjoint analysis

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

March 18, 2024

Intro

Plan

1) Conjoint analysis

2) Preparation for exam

- ANOVA, ANCOVA, Multiple comparison corrections

- Contingency tables, Chi-squared tests, Partial correlation

- PCA, Factor analysis

- Discriminant analysis, Logit

- Cluster analysis, Dendrogramms

Conjoint analysis

WHAT IS CONJOINT ANALYSIS?

Conjoint Analysis is a technique used to understand preference or relative importance given to various attributes of a product by the customer while making purchase decisions.

- A popular research method for predicting how people make complex choices.

- Comes out of the psychology, economics, and market research academic areas (from the 1970s).

- Today, conjoint analysis studies are conducted annually by research firms, govt. agencies, in academia, and by companies of all sorts.

Conjoint analysis

Basic concepts in conjoint analysis

A technique that attempts to determine the relative importance that consumers attach to salient attributes and the utilities they attach to the levels of attributes

- This information is derived from consumers' evaluations of brands, or from brand profiles composed of these attributes and their levels.

- The participants are presented with stimuli that consist of combinations of attribute levels.

- They are asked to evaluate the se stimuli in terms of their desirability.

- Conjoint procedures attempt to assign values to the levels of each attribute so that the resulting values or utilities attached to the stimuli match, as closely as possible, the input evaluations provided by the participants.

- The underlying assumption is that any set of stimuli - such as products, brands or companies - is evaluated as a bundle of attributes.augment.

Conjoint analysis

The basic conjoint analysis model may be represented by the following formula:

$$U(X) = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \alpha_{ij} x_{ij}$$

where $U(X) =$ overall utility of an alternative

$\alpha_{ij} =$ the part-worth contribution or utility associated with the j th level $(j = 1, 2, \ldots, k_j)$ of the ith attribute $(i = 1, 2, \ldots, m)$

$k_i =$ number of levels of attribute i

$m =$ number of attributes

$x_{ij} = 1$ if the j th level of the ith attribute is present $= 0$ otherwise.

The importance of an attribute, I_i, is defined in terms of the range of the part-worths, α_{ij}, across the levels of that attribute:

$$I_j = \left\{ \max\left(\alpha_{ij}\right) - \min\left(\alpha_{ij}\right) \right\} \text{ for each } i$$

The attribute's importance is normalised to ascertain its importance relative to other attributes, W_i :

$$W_i = \frac{I_i}{\sum_{i=1}^{m} I_i}$$

so that

$$\sum_{j=1}^{m} W_i = 1$$

Conjoint analysis

Several different procedures are available for estimating the basic model.

- The simplest: dummy variable regression

The predictor variables consist of dummy variables for the attribute levels.

If an attribute has k_i levels, it is coded in terms of $k_i - 1$ dummy variables.

For non-metric data, the participants are typically required to provide rank-order evaluations (the rankings may be converted to 0 or 1 by making paired comparisons between brands)

In metric form, the participants provide ratings, rather than rankings (the ratings, assumed to be interval scaled)

- Other procedures: that are appropriate for nonmetric data include

LINMAP (The Linear Programming Technique for Multidimensional Analysis of Preference)

MONANOVA (monotone ANOVA with transformed scores)

LOGIT model

Conjoint analysis

Hybrid conjoint analysis uses hybrid models which have been developed to serve two main purposes:

- Simplify the data collection task by imposing less of a burden on each participant

- Permit the estimation of selected interactions (at the subgroup level) as well as all main (or simple) effects at the individual level.

In the hybrid approach, the participants evaluate a limited number, generally no more than nine, conjoint stimuli, such as full profiles.

These profiles are drawn from a large master design and different participants evaluate different sets of profiles so that, over a group of participants, all the profiles of interest are evaluated.

In addition, participants directly evaluate the relative importance of each attribute and desirability of the levels of each attribute.

By combining the direct evaluations with those derived from the evaluations of the conjoint stimuli, it is possible to estimate a model at the aggregate level and still retain some individual differences.

Conjoint analysis

The individual participant vs the aggregate level:

- At the individual level, the data of each participant are analysed separately.

- If an aggregate-level analysis is to be conducted, some procedure for grouping the participants must be devised.

One common approach is to estimate individual level part-worth or utility functions first.

Participants are then clustered on the basis of the similarity of their part-worth functions.

Aggregate analysis is then conducted for each cluster.

Conjoint analysis

Assumptions and limitations of conjoint analysis:

- The important attributes of a product can be identified.

- Consumers evaluate the choice alternatives in terms of these attributes and make trade-offs.

- Data collection may be complex, particularly if a large number of attributes are involved and the model must be estimated at the individual level.

Solution: interactive or adaptive conjoint analysis and hybrid conjoint analysis.

Conjoint analysis

Statistics and terms associated with conjoint analysis:

- Part-worth functions. The part-worth or utility functions describe the utility that consumers attach to the levels of each attribute.

- Relative importance weights. The relative importance weights are estimated and indicate which attributes are important in influencing consumer choice.

- Attribute levels. These levels denote the values assumed by the attributes.

- Full profiles. Full or complete profiles of brands are constructed in terms of all the attributes by using the attribute levels specified by the design.

- Pairwise tables. Participants evaluate two attributes at a time until all the required pairs of attributes have been evaluated.

- Cyclical designs. These designs are employed to reduce the number of paired comparisons.

- Fractional factorial designs. These designs are employed to reduce the number of stimulus profiles to be evaluated in the full-profile approach.

- Orthogonal arrays. These are a special class of fractional designs that enable the efficient estimation of all main effects.

- Internal validity. This involves correlations of the predicted evaluations for the holdout or validation stimuli with those obtained from the participants.

Conjoint analysis

Conjoint analysis relies on participants' subjective evaluations.

Conjoint analysis, on the other hand, seeks to develop the part-worth or utility functions describing the utility that consumers attach to the levels of each attribute.

Conjoint analysis has been used in marketing for a variety of purposes, including the following:

- **Determining the relative importance** of attributes in the consumer choice process.

The relative importance weights indicate which attributes are important in influencing consumer choice.

- **Estimating market share of brands** that differ in attribute levels.

The utilities derived from conjoint analysis can be used as input into a choice simulator to determine the share of choices, and hence the market share, of different brands.

- **Determining the composition of the most-preferred brand**.

Brand features can be varied in terms of attribute levels and the corresponding utilities determined.

The brand features that yield the highest utility indicate the composition of the most-preferred brand.

- **Segmenting the market based on similarity** of preferences for attribute levels.

The partworth functions derived for the attributes may be used as a basis for clustering participants to arrive at homogeneous preference segments.

Conjoint analysis

Example 1: CHOOSING A RESTAURANT

Type of food (3)

- (American, Chinese, Italian)

Distance from your house (3)

- (5 minutes, 15 minutes, 30 minutes)

Typical cost per person (3)

- ($20, $30, $50)

How much you think your partner will like it (2)

- (a lot, just OK)

Q: HOW IMPORTANT ARE THESE FACTORS (ATTRIBUTES) TO THE CLIENT?

Conjoint analysis

ASKING IMPORTANCES IN RESEARCH CONTEXTS

We often want to know how important things are to people that we wish to survey

The problem is that folks have a hard time giving us good data on those sorts of questions...

- Lack of consistency in how they use the scale

- Lack of discrimination among items they rate

- Straightlining

Why not ask people to make specific tradeoffs like they do in the real world?

- And then figure out what must be driving peoples' decisions by observing their choices?

Conjoint analysis

CONJOINT ASKS PEOPLE TO MAKE TRADEOFFS JUST LIKE THEY DO IN REAL WORLD

A TRADEOFF QUESTION:

Chinese	Italian
30 minutes from your house	5 minutes from your house
Typically $20 per person	Typically $50 per person
You think it your partner will like it "just OK"	You think it pleases your will like it "a lot"

Which Would You Choose?

Conjoint analysis

THEORETICALLY, LOTS OF POSSIBLE CHOICES:

(3 Themes) $\times$ (3 Travel Distances) $\times$ (3 Cost Levels) $\times$ (2 Please Your Partner)

54 possible combinations!

But 1431 ways that these combinations could be displayed in pairs of two options

What conjoint analysis software does?

Picks a small subset of the possible combinations and sets of combinations for each respondent (typically 8 to 12 sets)

Asks respondents which product profile (alternative) they prefer in each set

Uses the answers to figure out:

- How important each attribute is in driving the decisions

- Which levels within each attribute are preferred (and by how much)

Conjoint analysis

WHAT-IF (MARKET SIMULATOR)

- The Really Cool output is an interactive what-if simulator.

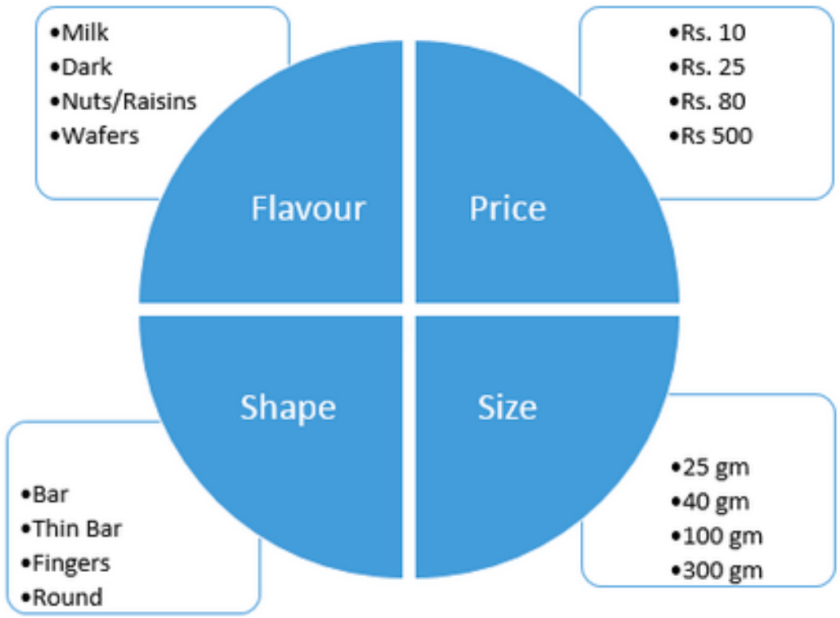
- If you had one of these, you could predict which of 20 possible restaurants in your area would please you and would please your partner.

- If you were a marketer, you could predict how modifications to your existing product line or extensions of your product line would likely perform in the market place (versus your competitor's products).

Conjoint analysis

Example 2:

While buying chocolate, let us say that there are 4 important attributes to be taken into consideration - Flavour, Shape, Size and Price. Each of these 4 attributes have 4 sub-levels each given below:



Conjoint analysis

- Conjoint Analysis helps in assigning utility values for each attribute (Flavour, Price, Shape and Size) and to each of the sub-levels.

- The attribute and the sublevel getting the highest Utility value is the most favoured by the customer.

- In this case, $4 * 4 * 4 * 4$ i.e. 256 combinations of the given attributes and their sublevels would be formed.

- Out of these combinations, let us say, we pick 16 combinations which make more practical sense.

- For e.g. a 300-gm chocolate would not be sold by any brand for only Rs. 10 .

Conjoint analysis

Steps to perform the analysis:

1. Ask the customers to rank the 16 chocolate types based on their preferences on an ordinal scale.

- The most preferred chocolate out of the given 16 varieties would be given Rank 1 and the least preferred chocolate would be given Rank 16.

2. Create two files for the conjoint analysis.

- One file should have all the 16 possible combinations of chocolates

- the other should have data of all the 100 respondents, in which 16 combinations were ranked from 1 to 16.

3. Then import the data into R/Python for Conjoint Analysis

Note:

The dependent variable is usually preference or intention to buy.

However, the conjoint methodology is flexible and can accommodate a range of other dependent variables, including actual purchase or choice.

Conjoint analysis

In evaluating boot profiles, participants were required to provide preference. Consider the hypothetical results below for a participant:

Profile number	Attribute levels			Preference rating
	Upper	Country	Price	
1	1	1	1	9
2	1	2	2	7
3	1	3	3	5
4	2	1	2	6
5	2	2	3	5
6	2	3	1	6
7	3	1	3	5
8	3	2	1	7

Conjoint analysis

The estimated model may be represented as:

$$U = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6$$

where:

- $X_1 X_2$ = dummy variables representing uppers

- $X_3 X_4$ = dummy variables representing country

- $X_5 X_6$ = dummy variables representing price.

For uppers, for example, the attribute levels were coded as follows:

	X_1	X_2
Level 1	1	0
Level 2	0	1
Level 3	0	0

Conjoint analysis

The levels of the other attributes were coded similarly.

The parameters were estimated as follows:

$$\hat{\beta}_0 = 4.22, \hat{\beta}_1 = 1.000, \hat{\beta}_2 = -0.333, \hat{\beta}_3 = 1.000, \hat{\beta}_4 = 0.667, \hat{\beta}_5 = 2.333, \hat{\beta}_6 = 1.333$$

Given the dummy variable coding, in which level 3 is the base level, the coefficients may be related to the part-worths:

$$\alpha_{11} - a_{13} = \hat{\beta}_1 \quad \text{and} \quad a_{12} - a_{13} = \hat{\beta}_2.$$

To solve for the part-worths, an additional constraint is necessary:

$$\alpha_{11} + a_{12} + a_{13} = 0$$

These equations for the first attribute, uppers, are:

$$\alpha_{11} - a_{13} = 1.000, a_{12} - a_{13} = -0.333, a_{11} + a_{12} + a_{13} = 0$$

Solving these equations, we get:

$$a_{11} = 0.778, a_{12} = -0.556, \quad a_{13} = -0.222$$

The part-worths for other attributes can be estimated similarly. For the second attribute, country, we have:

$$a_{21} - a_{23} = \hat{\beta}_3, \quad a_{22} - a_{23} = \hat{\beta}_4, \quad a_{21} + a_{22} + a_{23} = 0$$

For the third attribute, price, we have:

$$a_{31} - a_{33} = \hat{\beta}_5, \quad a_{32} - a_{33} = \hat{\beta}_6, \quad a_{31} + a_{32} + a_{33} = 0$$

Conjoint analysis

Estimates of regression model coefficients

β_0	4.222	$\alpha_{11} - \alpha_{13} =$	1	$\alpha_1 1=$	0.777667	
β_1	1	$\alpha_{12} - \alpha_{13} =$	-0.333	$\alpha_1 2=$	-0.55533	
β_2	-0.333			$\alpha_{13}=$	-0.22233	Sum = 0
β_3	1	$\alpha_{21} - \alpha_{23} =$	1			
β_4	0.667	$\alpha_{22} - \alpha_{23}=$	0.667	$\alpha_2 1=$	0.444333	
β_5	2.333			$\alpha_{22}=$	0.111333	
β_6	1.333	$\alpha_{31} - \alpha_{33} =$	2.333	$\alpha_{23}=$	-0.55567	Sum = 0
		$\alpha_{32} - \alpha_{33} =$	1.333001			
				$\alpha_{31}=$	1.111	
				$\alpha_{32}=$	0.111001	
				$\alpha_{33}=$	-1.222	Sum =0

The relative importance weights are calculated based on ranges of part-worths as follows: sum of ranges of part-worths
$= (0.778 - (-0.556)) + (0.445 - (-0.556)) + (1.111 - (-1.222)) = 4.668.$

Hence:

- Relative importance of uppers $= 1.334/4.668 = 0.286$

- Relative importance of country $= 1.001/4.668 = 0.214 =$

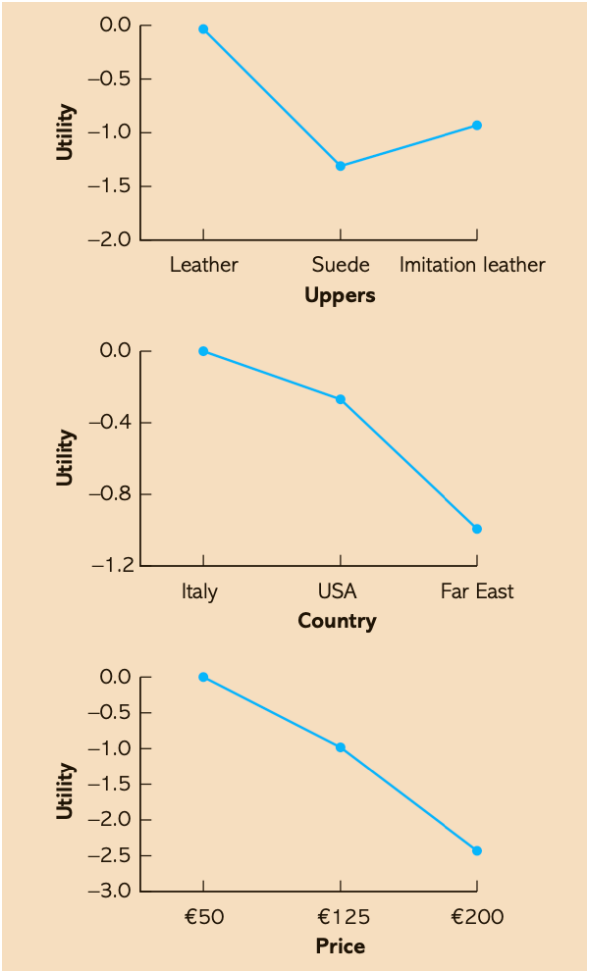
- Relative importance of price $= 2.333/4.668 = 0.500.$

Conjoint analysis

For interpreting the results, it is helpful to plot the part-worth functions.

The utility values have only interval-scale properties and their origin is arbitrary.

The relative importance of attributes should be considered.



Conjoint analysis

Assess reliability and validity

- The goodness-of-fit of the estimated model should be evaluated.

- For example, if dummy variable regression is used, the value of R^2 will indicate the extent to which the model fits the data.

- Test-retest reliability can be assessed by obtaining a few replicated judgements later in data collection.

- The evaluations for the holdout or validation stimuli can be predicted by the estimated part-worth functions.

- The predicted evaluations can then be correlated with those obtained from the participants to determine internal validity.

- If an aggregate-level analysis has been conducted, the estimation sample can be split in several ways and conjoint analysis conducted on each subsample.

The results can be compared across subsamples to assess the stability of conjoint analysis solutions.

UoL Tasks

Q1

i. What are projective techniques? Under what circumstances should projective techniques be used?

Projective techniques are unstructured and indirect data collection techniques.

- They may be defined as a form of questioning which encourages the respondent to project their underlying motivations, beliefs, attitudes or feelings regarding the issues of concern.

- Respondents are not directly asked about their own behaviour, but are asked to reflect on the behaviour of others, indirectly projecting their own motivations, beliefs, attitudes or feelings.

- Projective techniques should be employed when the required information cannot be accurately obtained by direct methods because the information is not part of conscious memory. Direct questioning in these circumstances would generate shallow and meaningless responses.

UoL Tasks

ii. Describe the word association technique. Give an example of a situation in which this technique is especially useful.

- In the word association technique, respondents are presented with a list of words, one at a time.

- The underlying assumption is that by freely associating with certain words, respondents will reveal their inner feelings about the topic of interest.

- Word association is frequently used in testing brand names, and occasionally for measuring attitudes about particular products, services, brands, packages or advertisements.

iii. Describe the story completion technique. Give an example of the type of respondent and the context in which such a technique would work.

- With the story completion technique, respondents are given part of a story, enough to direct attention to a particular topic but not to hint at the ending.

- One characteristic they would need is a willingness to participate (right introduction and motivation)

UoL Tasks

Q2. Define a 'response rate' and non-response. What strategies are available for minimising non-response and adjusting for it? In your response, consider how these are tackled across different data collection methods.

Response rate = N of completions/N in survey.

Non-response is refusal to respond. There are two types:

(i.) unit non-response (refusal to participate in survey)

(ii.) item non-response (refusal to answer a specific question in the survey).

Improve response rates by prior notification, incentives, better questionnaire design, callbacks etc. Strategies for adjusting for non-response include the following.

- Sub-sampling of non-respondents. A concerted effort is made to contact a sub-sample of the respondents, usually by means of telephone or personal interviews.

- Replacement. The non-respondents in the current survey are replaced with non-respondents from an earlier, similar survey.

- Substitution. The non-respondents are substituted with other elements from the sampling frame that are expected to respond.

UoL Tasks

- Subjective estimates. This involves making a subjective evaluation of the likely effects of non-response based on experience and available information.

- Trend analysis. The researcher tries to discern a trend between early and late respondents. This trend is projected to non-respondents to estimate their characteristic of interest.

- Simple weighting. Depending on the response rates, differential weights are assigned to the data to account for non-response.

- Imputation. Imputing the characteristic of interest to the non-respondents based on the similarity of the variables available for both non-respondents and respondents.

i. What is non-response bias and why is it problematic?

Non-response bias leads to non-representative samples due to selected respondents refusing to participate in the survey.

ii. Identify three ways to improve response rates, explaining how these could be effective.

Response rates could be improved through callbacks, prior notification, incentives, follow-ups, different facilitators etc.

UoL Tasks

Q3.

i. Define the terms 'incidence rate' and 'completion rate'.

The incidence rate refers to the rate of occurrence, or the percentage, of persons eligible to participate in a study.

In general, if there are c qualifying factors with an incidence of $Q_1, Q_2, Q_3, \ldots, Q_c$, each expressed as a proportion, then:

$$\text{incidence rate } = Q_1 \times Q_2 \times Q_3 \times \cdots \times Q_c.$$

The completion rate is the percentage of qualified respondents who complete the interview, enabling researchers to account for anticipated refusals by people who qualify.

ii. Explain how you would adjust the statistically-determined sample size, n, in light of incidence and completion rates.

In light of incidence and completion rates, the initial sample size is:

$$\text{initial sample size } = \frac{\text{final sample size}}{\text{incidence rate } \times \text{ completion rate}}.$$

UoL Tasks

Q4.

i. What are the major decisions involved in constructing an itemised rating scale?

The major decisions involved in constructing an itemised rating scale are the following.

- The number of scale categories to use – i.e. the means to finely discriminate.

- Balanced versus unbalanced scale – i.e. should the scale be skewed or not?

- Odd or even number of categories – i.e. what role does the middle scale position have?

- Forced versus non-forced nature of the scale – i.e. do respondents really think through the issues or opt for a simple middle path?

- The nature and degree of verbal description to employ – i.e. how the use of words in scale items clarifies the issues for respondents.

- The physical form of the scale – i.e. how attractive, interesting or simple the scale may appear, drawing in the respondent to want to engage with it.

ii. How many scale categories should be used in an itemised rating scale? Briefly explain your answer.

The number of scale categories that should be used in an itemised rating scale should be between three and ten. However, there is no single, optimal number of categories, which would be applicable for all scaling situations.

UoL Tasks

Q5.

i. Describe the semantic differential scale and the Likert scale. For what purposes are these scales used? Provide an example of each scale.

- A semantic differential scale is a seven-point rating scale with bipolar labels which have semantic meaning.

- In a typical application, respondents rate objects on a series of itemised, seven-point rating scales, bounded at each end by one of two bipolar adjectives, such as 'powerful' or 'weak'.

- The Likert scale typically has five response categories ranging from 'strongly disagree' to 'strongly agree'.

- Respondents are required to indicate a degree of agreement or disagreement with each of a series of statements related to the stimulus objects.

These scales are used to measure the strength of feeling about the individual constructs or components of marketing phenomena such as brand, product and company images, feelings about advertising and promotion strategies, new product development studies and in a variety of other applications.

UoL Tasks

ii. How does the nature and degree of verbal description affect the response to itemised rating scales?

- The nature and degree of verbal description associated with scale categories varies considerably and can affect the responses. Scale categories may have verbal, numerical or even pictorial descriptions.

- Furthermore, the researcher must decide whether to label every scale category, scale some categories or scale only extreme categories.

- The category descriptions should be located as close to the response categories as possible.

- The strength of the adjectives used to anchor the scale may influence the distribution of the responses.

With strong anchors (1 = completely disagree, 7 = completely agree), respondents are less likely to use the extreme scale categories.

This results in less variable and more peaked response distributions.

Weak anchors (1 = generally disagree, 7 = generally agree), in contrast, produce uniform or flat distributions.

ANOVA (ANalysis Of VAriance)

- ANOVA test used to compare the means of more than 2 groups (t-test and it's variations can be used to compare 2 groups)
- Groups mean differences inferred by analyzing variances
- ANOVA uses variance-based F test to check the group mean equality. Sometimes, ANOVA F test is also called omnibus test as it tests non-specific null hypothesis i.e. all group means are equal

Null hypothesis:

Groups means are equal (no variation in means of groups)

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

k – is the number of groups

Alternative hypothesis:

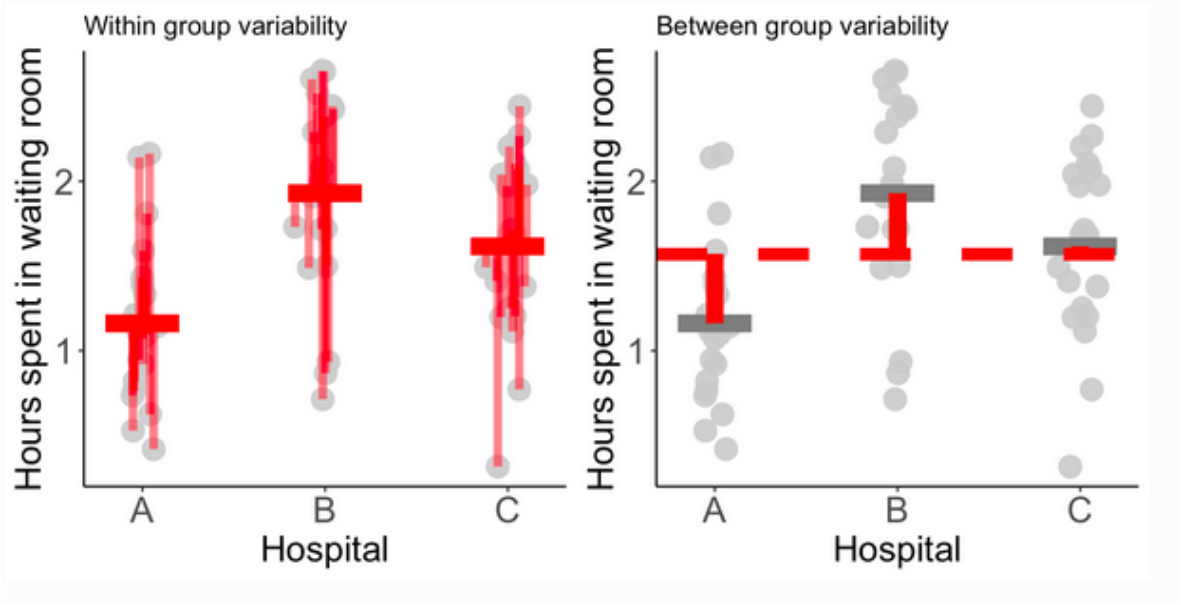
At least, one group mean is different from other groups H_1 : All μ are not equal

Idea

Compare

- within-group variability: the variance of the individual observations within a group, and

- between-group variability: the variance between the averages of the groups.

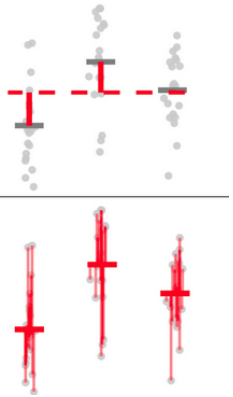


The basic idea is that if the variability between the groups is greater than the variability within the groups, then we have evidence that the differences between the groups is not simply reflecting random noise.

Idea

Related F-statistics:

$$\frac{\text{BSS / (K - 1)}}{\text{WSS / (N - K)}} =$$



$$WSS = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2 \quad \text{and} \quad BSS = \sum_{i=1}^{K} (\bar{y}_{..} - \bar{y}_{i.})^2$$

where y_{ij}, defines the waiting room time (outcome) for patient j from hospital $i, \bar{y}_{..}$ defines the global average waiting time and $\bar{y}_{i.}$ defines the average waiting time for hospital i.K is the number of hospitals, and n_i is the number of patients sampled from hospital i.

Hence:

$$F = \frac{Var_{between}}{Var_{within}} = \frac{BSS/(K-1)}{WSS/(N-K)} \sim F_{K-1,N-k}$$

Sum of Squares (SS)

Inside the One-Way ANOVA Table:

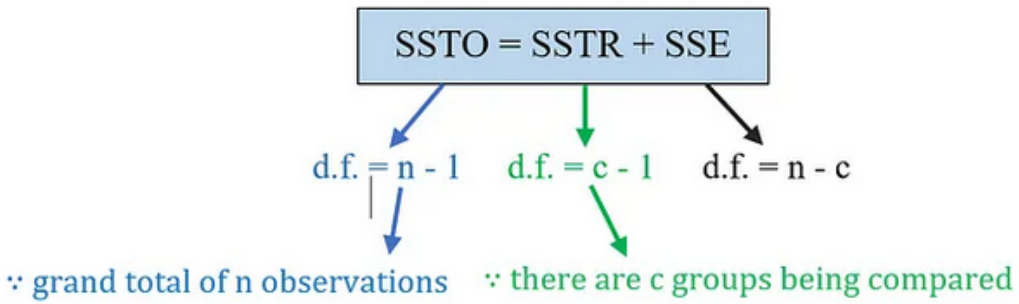
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma\, n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = SSB / (k - 1)$	$f = MSB / MSE$
Error	$SSE = \Sigma\Sigma(X - \bar{X}_j)^2$	$df_2 = N - k$	$MSE = SSE / (N - k)$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

The total amount of variability comes from two possible sources, namely:

1. Difference among the groups, called treatment (TR)

2. Difference within the groups, called error (E)

The sum of the squares due to treatment (SSTR) and the sum of squares due to error (SSE) are listed in the one-way ANOVA table. The sum of SSTR and SSE is equal to the total sum of squares (SSTO).



Regression form

$$y_{ik} = \mu + \alpha_k + \epsilon_{ik}$$

$$SS_T = SS_B + SS_E$$

$$TSS = ESS + RSS$$

Where, y_{ik} – i^{th} observation of k^{th} level of groups, μ = overall population mean (unknown), α_k = Main effect for groups (deviation from the μ), ϵ_{ik} = Error, k = levels for groups $k = 1, 2 \ldots, \mathrm{p})$, i = Observations or replicates for each group $(i = 1, 2 \ldots, r)$,

Where,
$SS_B = \sum_i p_i \left(\overline{y_{i.}} - \bar{y}..\right)^2$, $SS_E = \sum_{ik} \left(y_{ik} - \overline{y_{i.}}\right)^2$, $SS_T = SS_B + SS_E = \sum_{ik} \left(y_{ik} - \overline{y_{..}}\right)^2$

F-test for regression significance:

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} \sim F_{k-1,n-k}$$

Assumptions for the one-way ANOVA hypothesis test

- **Observations are i.i.d** – sample data are randomly selected from populations and randomly assigned to each of the treatment groups
- **Normality**
  Check: normal probability plot, Q-Q plot, Shapiro-Wilks test, etc.
- **Homoscedasticity or Homogeneity of variance** – all the k group variances are equal, that is ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2 = \ldots = \sigma_k^2$.
  Rule of thumb: the ratio of the largest to the smallest sample standard deviation is less than 2
  Check: Levene's, Bartlett's, or Brown-Forsythe test
- **Dependent variable is continuous**

  If the dependent variable is ordinal or rank (e.g. Likert item data), it is more likely to violate the assumptions of normality and homogeneity of variances.

  If these assumptions are violated, you should consider the non-parametric tests (e.g. Mann-Whitney U test, Kruskal-Wallis test).

ANOVA is a powerful method when the assumptions of normality and homogeneity of variances are valid.

ANOVA is less powerful, if the assumption of normality is violated while variances are equal.

Assumptions for the one-way ANOVA hypothesis test

ANCOVA

(i) One-way ANOVA structural model

$$\boldsymbol{X}_{ij} = \mu + \tau_j + \mathbf{e}_{ij}$$

(ii) One-way ANCOVA structural model

$$X_{ij} = \mu + \alpha_j + \beta Z_{ij} + e_{ijk}$$

Covariate is just another source of variance

- Use the term βZ_{ij} because of continuous nature;

- Implicitly, we have specified no interaction between covariate and the independent variable (α)

Uses of ANCOVA

1. To control unwanted variation that would otherwise inflate the error with which we test our models (classical usage)

2. To control for group differences, esp. in the analysis of clinical trials or other pre/post designs

Assumptions for the one-way ANOVA hypothesis test

Assumption: relationship between covariate and DV is **linear**

Non-linear relationships generally cannot be detected with ANCOVA - degrades power.

Assumption: relationship between DV and covariate is equal across treatment groups - **homogeneity of regression slopes**

Homogeneity of regression slopes is important because adjustments to treatment means are based upon an average within-cell regression coefficient

Multiple hypothesis testing

The p value obtained from ANOVA analysis is significant ($p < 0.05$), we conclude that treatment differences are statistically significant

Problem:

ANOVA does not tell which treatments are significantly different from each other.

Solution: To know the pairs of significant different treatments, we will can perform **multiple pairwise comparison** (post hoc comparison) analysis

Note: When the ANOVA is significant, post hoc tests are used to see differences between specific groups.

Post hoc tests should control the family-wise error rate (inflated type I error rate) due to multiple comparisons

Post hoc tests adjust the p-values (e.g. Bonferroni correction) or critical value (e.g. Tukey's HSD test).

Multiple hypothesis testing

Example:

There are 20 features you are interested in as independent (predictor) features to create your machine learning model.

We want to select which features are useful for our prediction model

Null Hypothesis (H0): There is no relationship between the variables

Alternative Hypothesis (H1): There is a relationship between variables

Bad idea:

To test each feature using hypothesis testing separately with some level of significance $\alpha = 0.05$.

Why? Let's calculate the probability of one significant result just due to chance?

$P(\text{ at least one significant result }) = 1 - P(\text{ no significant results })$

$P(\text{ at least one significant result }) = 1 - (1 - 0.05)^{20} \approx 0.64$

With 20 hypotheses were made, there is around a 64% chance that at least one hypothesis testing result is significant, even if all the tests are actually not significant.

With a higher number of features to consider, the chance would even higher.

That is why there are methods developed for dealing with *multiple testing error.*

Multiple hypothesis testing

This method is called the **multiple testing correction**.

Number of errors committed when testing m null hypotheses

	Declared non-significant	Declared significlicant	Total
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m -$ **R**	**R**	m

What was actually corrected?

(1) Controlling the Family-wise error rate (FWER)

FWER is the probability of making at least one false discoveries (type I errors)

$$\text{FWER} = \Pr(V \geq 1) = 1 - \Pr(V = 0)$$

Thus, by assuring FWER $\leq \alpha$, the probability of making one or more type I errors in the family is controlled at level α.

(2) Controlling the False Discovery Rate (FDR) = Type I error/False Positive Error.

Multiple hypothesis testing

1) Bonferroni correction – simplest yet the strictest method, controls FWER

α is divide it with the number of the testing/number of the hypothesis for each hypothesis.

$$\alpha_{Bon} = \alpha/m$$

Example:

Let's assume we have 10 features.

Normally, when we get the P-value < 0.05, we might see a significant result due to a chance.

In our case if we have 20 hypothesis testing.

$$\alpha_{Bon} = \alpha/m = 0.05/20 = 0.0025$$

Hence

$P($ at least one significant result $) = 1 - (1 - 0.0025)^{20} \approx 0.049$

Note:

Bonferroni Correction is proven too strict at correcting the α level where Type II error/False Negative rate is higher than what it should be.

2) Holm-Bonferroni correction method – less strict, controls FWER

The α level correction is not uniform for each hypothesis testing; instead, it was varied depending on the P-value ranking.

By ranking, it means a P-value of the hypothesis testing we had from lowest to highest.

Feature	P-Value
Feature #4 – Rank 1	0.001
Feature #3	0.003
Feature #1	0.01
Feature #8	0.0134
Feature #7	0.02
Feature #10	0.025
Feature #9	0.044
Feature #2	0.067
Feature #6	0.33
Feature #5 – Rank 10	0.5

(1)

Let's try to rank our previous hypothesis from the P-value we have before. The rank should look like this.

Multiple hypothesis testing

After we rank the P-value, we would the correct α level and test the individual hypothesis using this equation below.

$$P_k < \frac{\alpha}{m + 1 - k}$$

Where k is the ranking and m is the number of hypotheses tested.

Example, we test rank 1:

$$P_1 < \frac{0.05}{10+1-1}$$
$$P_1 < 0.005$$

Example, we test rank 10:

$$P_1 < \frac{0.05}{10+1-10}$$
$$P_1 < 0.05$$

Multiple hypothesis testing

3) Shidak correction – controls FWER

$P(V \leq 1) = 1 - P(V = 0) \leq 1 - (1 - \alpha_1)^m = \alpha$, where α is the significance level we set for the family hypotheses and α_1- the desired significance level for testing each single hypothesis.

Let's express α_1 in terms of α and get $\alpha_1 = 1 - (1 - \alpha)^{1/m}$ |

4) Shidak-Holm method – controls FWER

Iterative adjustment. Similarly, we sort our p-values in ascending order and correct them according to the Shidak correction:

$\alpha_1 = 1 - (1 - \alpha)^{\frac{\pi}{m}}$

$$\alpha_i = 1 - (1 - \alpha)^{\frac{\alpha}{w - l + 1}}$$
$$\dots$$
$$\alpha_m = \alpha$$

'=

Multiple hypothesis testing

Has several properties:

1. Controls FWER at the α significance level if the statistics are collectively independent.

2. If the statistics are collectively independent, it is impossible to construct a procedure that controls FWER at the α level and is more powerful than the Shidak-Holm method.

3. For large m it differs little from the Holm method

Note:

- Without additional assumptions it is impossible to construct a more powerful procedure than Holm's method

- Given the independence of the experiments, it is impossible to construct a more powerful procedure than the Shidak-Holm method

But you can create a powerful procedure for FDR - and, as practice shows.

Multiple hypothesis testing

5) Benjamini-Hochberg method – controls FDR

Recall, FDR (False Discovery Rate) is the average proportion of falsely rejected $\boldsymbol{H_0}$ among all rejected

$$H_0, FDR = E\left(\frac{V}{R}\middle| R > 0\right)$$

- when considering FWER, we were concerned about the probability that at least one null hypothesis would be falsely rejected

- when considering FDR, we lower the bar and assume that there will be several such hypotheses - but no more than α.

Note that $FDR \leq FWER$

Benjamini-Hochberg (BH) method or often called the BH Step-up procedure:

First, ranks the P-value from the lowest to the highest.

The hypothesis is then compared to the α level by the following equation.

$$P_k < \frac{k}{m}\alpha$$

- where k is the rank and m is the number of the hypotheses.

Cross-tabulation

While a frequency distribution describes one variable at a time, a cross-tabulation describes two or more variables simultaneously.

Cross-tabulation results in tables which reflect the **joint distribution** of two or more variables with a limited number of categories or distinct values, for example:

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

Cross-tabulation

Often the introduction of a *third variable* clarifies the initial association (or lack of it) observed between two variables.

1. It can refine the association observed between the two original variables.

2. It can indicate no association between the two variables, although an association was initially observed. In other words, the third variable indicates that the initial association between the two variables was spurious.

3. It can reveal some association between the two original variables, although no association was initially observed. In this case, the third variable reveals a suppressed association between the first two variables: a suppressor effect.

4. It can indicate no change in the initial association.

Refined initial relationship

Purchase of luxury branded clothing	Marital status	
	Married	Unmarried
High	31%	52%
Low	69%	48%
Column	100%	100%
Number of participants	700	300

52% of unmarried participants fell in the high-purchase category, as opposed to 31% of the married participants. Before concluding that unmarried participants purchase more luxury branded clothing than those who are married, a third variable, the buyer's gender, was introduced into the analysis.

Purchase of luxury branded clothing	Gender			
	Male marital status		Female marital status	
	Married	Unmarried	Married	Unmarried
High	35%	40%	25%	60%
Low	65%	60%	75%	40%
Column	100%	100%	100%	100%
Number of participants	400	120	300	180

In the case of females, 60% of the unmarried participants fall in the high-purchase category compared with 25% of those who are married. On the other hand, the percentages are much closer for males.

Hence, the introduction of gender (third variable) has **refined the relationship** between marital status and purchase of luxury branded clothing (original variables).

Unmarried participants are more likely to fall into the high-purchase category than married ones, and this effect is much more pronounced for females than for males.

Initial relationship was spurious

	Education	
Own expensive car	Degree	No degree
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of participants	250	750

The table shows that 32% of those with university degrees own an expensive (more than €80,000) car, compared with 21% of those without university degrees.

Conclusion: education influenced ownership of expensive cars.

However, income may also be an important factor for determining car ownership.

	Income			
Own expensive car	Low-income education		High-income education	
	Degree	No degree	Degree	No degree
Yes	20%	20%	40%	40%
No	80%	80%	60%	60%
Column totals	100%	100%	100%	100%
Number of participants	100	700	150	50

The percentages of those with and without university degrees who own expensive cars are the same for each income group.

When the data for the high-income and low-income groups are examined separately, the association between education and ownership of expensive cars disappears, indicating that the **initial relationship** observed between these two variables was **spurious**.

Reveal suppressed association

	Age	
Desire to travel abroad	Under **45**	**45** or older
Yes	50%	50%
No	50%	50%
Column totals	100%	100%
Number of participants	500	500

Cross-tabulation indicate no association.

Let's introduced gender as the third variable.

	Gender			
	Male age		Female age	
Desire to travel abroad	Under **45**	**45** or older	Under **45**	**45** or older
Yes	60%	40%	35%	65%
No	40%	60%	65%	35%
Column totals	100%	100%	100%	100%
Number of participants	300	300	200	200

Among men, 60% of those under 45 indicated a desire to travel abroad compared with 40% of those 45 or older. The pattern was reversed for women.

Since the association between desire to travel abroad and age runs in the *opposite direction* for males and females, the relationship between these two variables is masked when the data are aggregated across gender.

But when the effect of gender is controlled, the **suppressed association** between preference and age is **revealed** for the separate categories of males and females.

No change in initial relationship

	Family size	
Eat frequently in fast-food restaurants	Small	Large
Yes	65%	65%
No	35%	35%
Column totals	100%	100%
Number of participants	500	500

No association is observed.

The participants were further classified into high- or low-income groups based on a median split.

Eat frequently in fast-food restaurants	Income			
	Low-income family size		High-income family size	
	Small	Large	Small	Large
Yes	65%	65%	65%	65%
No	35%	35%	35%	35%
Column total	100%	100%	100%	100%
Number of participants	250	250	250	250

Again, no association was observed.

In some cases, the introduction of the third variable **does not change the initial relationship** observed, regardless of whether the original variables were associated.

This suggests that the third variable does not influence the relationship between the first two.

Chi-square test

Q: Is there a systematic association exists between the two variables?

The chi-square statistic (χ^2) is used to test the statistical significance of the observed association in a cross-tabulation.

H_0 : there is no association between the variables.

H_a : there is association between the variables.

Idea: compare the cell frequencies that would be expected if no association were present between the variables, given the existing row and column totals.

f_e – expected cell frequencies

f_o – actual observed frequencies.

The greater the discrepancies between the expected and observed frequencies, the larger the value of the statistic.

Chi-square test

The expected frequency for each cell can be calculated by using a simple formula:

$$f_e = \frac{n_r n_e}{n}$$

where

$n_r =$ total number in the row
$n_c =$ total number in the column
$n =$ total sample size.

Then the value of χ^2 is calculated as follows:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} \sim \chi^2_{(r-1)\times(c-1)}$$

The null hypothesis (H_0) of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution with the appropriate degrees of freedom.

The chi-square distribution is a skewed distribution whose shape depends solely on the number of degrees of freedom.

As the number of degrees of freedom increases, the chi-square distribution becomes more symmetrical.

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

The expected frequencies for the cells, going from left to right and from top to bottom, are:

$$15 \times 15/30 = 7.50, 1515/30 = 7.50, 15 \times 15/30 = 7.50, 15 \times 15/30 = 7.50$$

The value of χ^2 is calculated as:

$$\begin{aligned}\chi^2 &= (5-7.5)^2/7.5 + (10-7.5)^2/7.5 + (10-7.5)^2/7.5 + (5-7.5)^2/7.5 \\ &= 0.833 + 0.833 + 0.833 + 0.833 \\ &= 3.333\end{aligned}$$

Number of degree of freedom: $df = (2-1) \times (2-1) = 1$

Critical value at 5%: $\chi^2_{1,0.95} = 3.841$

The null hypothesis of no association cannot be rejected, indicating that the association is not statistically significant at the 0.05 level.

Note that this lack of significance is mainly due to the small sample size (30).

If, instead, the sample size were 300 and each data entry were multiplied by 10, test statistics would be multiplied by 10 $\chi^2_{obs} = 33.33$, which is significant at the 0.05 level.

Test assumptions

– The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers.

– The observations are drawn independently.

– Chi-square analysis should not be conducted when the expected or theoretical frequency in any of the cells is less than five.

– If the number of observations in any cell is less than 10, or if the table has two rows and two columns (a 2×2 table), a correction factor should be applied (Yates's chi-squared test).

Strength of association

1) **Phi coefficient** (ϕ)

– works for the special case of a table with two rows and two columns (a 2×2 table)

For a sample of size n, this statistic is calculated as:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

It takes the value of 0 when there is no association, which would be indicated by a chi-square value of 0 as well.

When the variables are perfectly associated, phi assumes the value of 1 and all the observations fall just on the main or minor diagonal.

Example: in our case, because the association was not significant at the 0.05 level, we would not normally compute the phi value. However, for the purpose of illustration, the value of phi is:

$$\phi = \sqrt{3.333/30} = 0.333$$

Thus, the association is not very strong.

Strength of association

2) **Contingency coefficient** (C)

– a more general case involving a table of any size

This index is also related to chi-square, as follows:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

The contingency coefficient varies between 0 and 1.

The value of 0 occurs in the case of no association (i.e. the variables are statistically independent), but the maximum value of 1 is never achieved.

The maximum value of the contingency coefficient *depends on the size of the table* (number of rows and number of columns), hence, it should be used only to compare tables of the *same size*.

Example: The value of the contingency coefficient:

$$C = \sqrt{3.333/(3.333 + 30)} = 0.31$$

This value of C indicates that the association is not very strong.

Strength of association

3) **Cramer's** V

– a modified version of the phi correlation coefficient used in tables larger than 2×2.

For a table with r rows and c columns:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1),(c-1)}}$$

When phi is calculated for a table larger than 2×2, it has no upper limit. Cramer's V is obtained by adjusting phi for either the number of rows or the number of columns in the table, based on which of the two is smaller. Hence, V will range from 0 to 1.

Example: The value of Cramer's V:

$$V = \sqrt{(3.333/30)/1 = 0.333}$$

Thus, the association is not very strong.

As can be seen, in this case $V = \phi$, which is always the case for a 2×2 table.

Strength of association

4) **Lambda coefficient**

The lambda coefficient assumes that the variables are measured on a nominal scale.

Asymmetric lambda – measures the percentage improvement in predicting the value of the dependent variable, given the value of the independent variable.

Symmetric lambda – does not make an assumption about which variable is dependent. It measures the overall improvement when prediction is done in both directions.

5) **Non-parametric coefficients of association**

Other statistics, such as tau *b*, tau *c* and gamma, are available to measure association between two *ordinal-level variables*.

All these statistics use information about the ordering of categories of variables by considering every possible pair of cases in the table.

Each pair is examined to determine whether its relative ordering on the first variable is the same as its relative ordering on the second variable (concordant), the ordering is reversed (discordant), or the pair is tied.

PCA — Principal Component Analysis

Goal: Reducing the number of dimensions of a data set

Idea:

– When a dataset contains a large number of variables, there is often a serious amount of overlap between those variables.

– The components that are found by PCA are ordered from the highest information content to the lowest information content.

$=>$ "regrouping" the variables into a smaller number of variables, called components based on variation common to multiple variables:

– the first (newly created) component contains a maximum of variation

– the second component contains the second-largest amount of variation, etc.

– the last component logically contains the smallest amount of variation.

Choose only few of the newly created components rather than the original variables, while still retaining a maximum amount of variation.

Use:

– data exploration, finding patterns in data of high dimension

– face recognition and image compression

PCA — Principal Component Analysis

Mathematical Model - Maximize variance of the new components

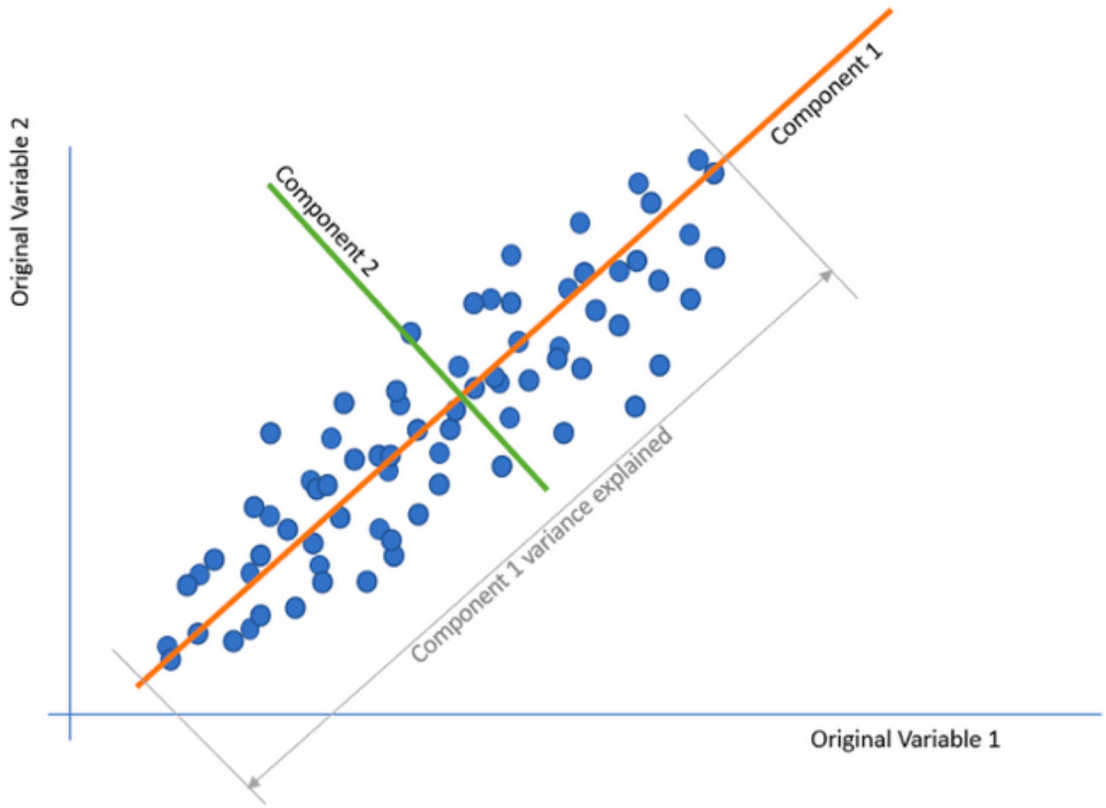


Figure: Schematic model of PCA

Mathematical definition of the PCA problem: find a linear combination of the original variables with maximum variance.

PCA — Principal Component Analysis

Mathematical Model - Maximize variance of the new components

1) Start with a (new) component z.

z is going to be computed based on our original variables $(\mathrm{X}_1, \mathrm{X}_2, \dots)$ multiplied by a weight for each of our variables $(u_1, u_2, ...)$.

This can be written as $\mathbf{z} = \mathbf{Xu}$.

2) The mathematical goal is to find the values for **u** that will maximize the variance of **z**, with a constraint of unit length on **u**.

Solution:

– This problem is mathematically called **a constrained optimization using Lagrange Multiplier**

– In practice, sequential numerical optimization.

– Can be described as applying matrix decomposition to the correlation matrix of the original variables.

PCA — Principal Component Analysis

PCA: Two Interpretations

E.g., for the first component.

I. Maximum Variance Direction: 1^{st} PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}^T\mathbf{x}_i\right)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

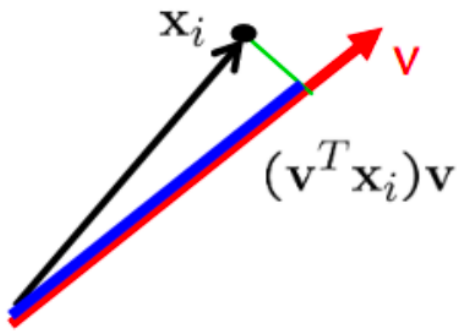
II. Minimum Reconstruction Error: 1^{st} PC a vector v such that projection on to this vector yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i - \left(\mathbf{v}^T\mathbf{x}_i\right)\mathbf{v}\right\|^2$$

blue ^{2} + green ^{2} = black ^{2}

black ^{2} is fixed (it's just the data)

So, maximizing blue ^{2} is equivalent to minimizing green ^{2}



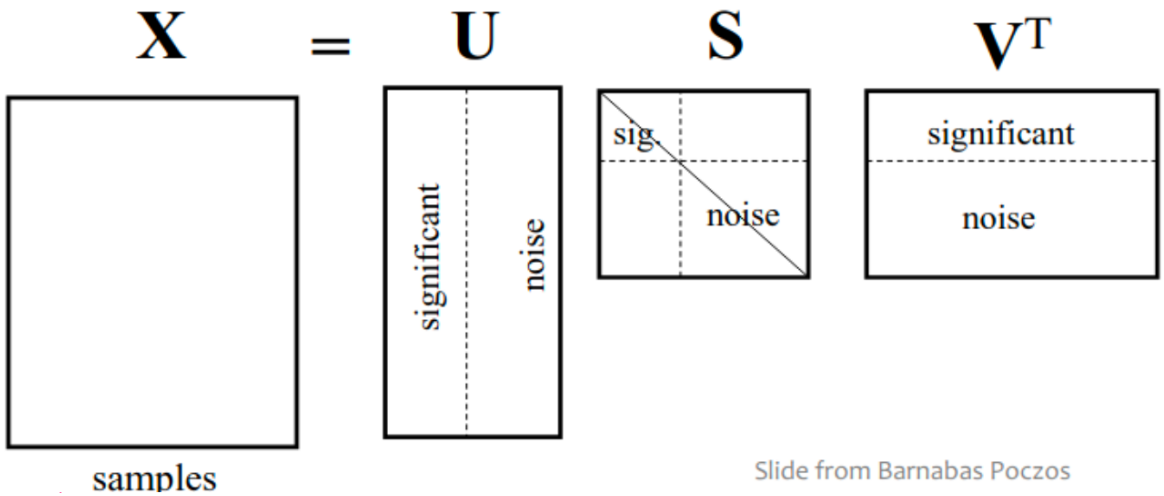
PCA — Principal Component Analysis

PCA algorithm: **SVD of the data matrix**

Singular Value Decomposition of the centered data matrix $\mathbf{X}$.

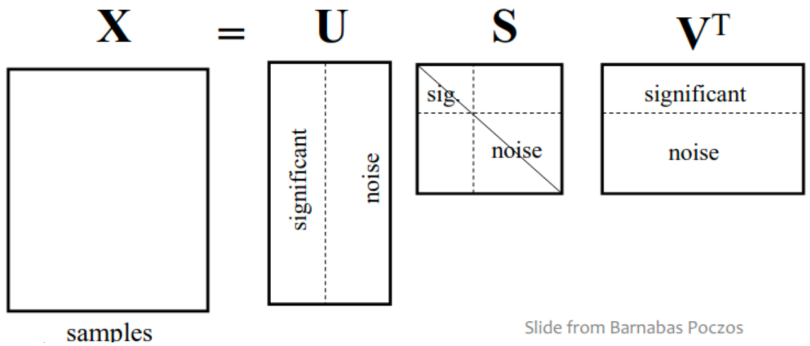
$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{N \times m}, \quad m: \text{ number of instances }, N: \text{ number of dimensions}$$

$$\mathbf{X}_{\text{features} \times \text{samples}} = \text{USV}^{\top}$$



Slide from Barnabas Poczos

PCA — Principal Component Analysis



Slide from Barnabas Poczos

Columns of U

- the principal vectors, $\{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}\}$

- orthogonal and has unit norm - so $U'U = I$

- Can reconstruct the data using linear combinations of $\{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}\}$

Matrix S

- Diagonal

- Shows importance of each eigenvector

Columns of $\mathbf{V}^\top$

- The coefficients for reconstructing the samples

CFA – Common Factor Analysis

PCA is efficient in finding the components that maximize variance.

Sometimes, however, we are not purely interested in maximizating variance, but to get the most useful **interpretations** to our newly defined dimensions.

Solution: Common Factor Analysis: an alternative to PCA that has a little bit more flexibility

Factor Analysis Vs. Principle Component Analysis

- PCA components explain the maximum amount of variance while CFA explains the covariance in data.

- PCA components are fully orthogonal to each other whereas CFA does not require factors to be orthogonal.

- PCA component is a linear combination of the observed variable while in CFA, the observed variables are linear combinations of the unobserved variable or factor.

- PCA components are uninterpretable. In CFA, underlying factors are labelable and interpretable.

- PCA is a kind of dimensionality reduction method whereas factor analysis is the latent variable method.

- PCA is a type of factor analysis. PCA is observational whereas FA is a modeling technique.

CFA – Common Factor Analysis

Goal - Finding latent variables in a data set

CFA allows reducing information in a larger number of variables into a smaller number of variables – "latent variables" – that make sense to us.

Unlike PCA, we can rotate the solution until we find latent variables that have a clear interpretation.

The principle:

– there a certain number of factors in a data set

– each of the measured variables captures a part of one or more of those factors.

CFA – Common Factor Analysis

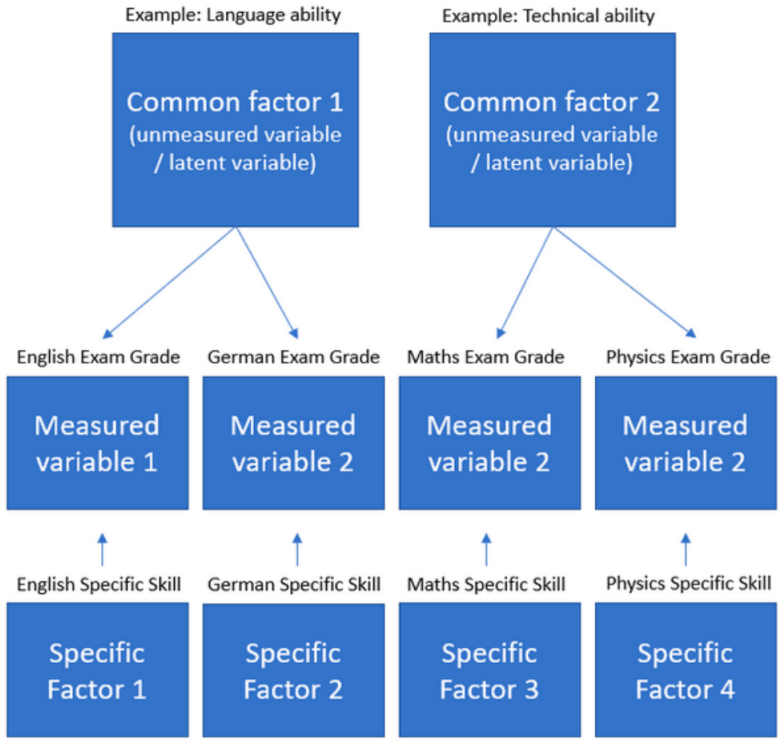
Example:

There are many students in a school with grades for many subjects.

Different grades are partly correlated: a more intellectually gifted student would have higher grades overall $\Rightarrow$ a latent variable.

But we could also imagine having students who are overall good in languages, but bad in technical subjects.

In this case, we could try to find a latent variable for language ability and a second latent variable for technical ability.



CFA – Common Factor Analysis

Mathematically, CFA is similar to multiple regression analysis in that each variable is expressed as a linear combination of underlying factors.

The covariation among the variables consists of two terms (not overtly observed)

– a small number of common factors

– plus a unique factor for each variable.

If the variables are standardised, the factor model may be represented as:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \ldots + A_{im}F_m + V_iU_i$$

where $X_i = i$ th standardised variable

A_{ij} = standardised multiple regression coefficient of variable i on common factor j

F = common factor

V_i = standardised regression coefficient of variable i on unique factor i

U_i = the unique factor for variable i

m = number of common factors.

Def.: Communality – the amount of variance a variable shares with all other variables included in the analysis

CFA – Common Factor Analysis

The factor model:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \ldots + A_{im}F_m + V_i U_i$$

where $X_i = i$ th standardised variable

The unique factors U_i are correlated with each other and with the common factors F.

The common factors themselves can be expressed as linear combinations of the observed variables:

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}W_k$$

where $F_i =$ estimate of i th factor

$W_i =$ weight or factor score coefficient

$k =$ number of variables.

It is possible to select weights or factor score coefficients so that:

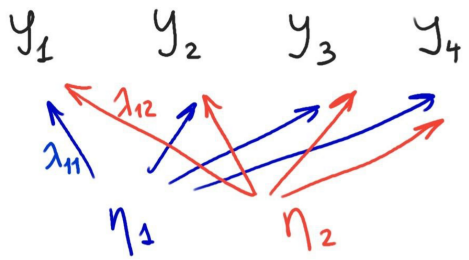
– the first factor explains the largest portion of the total variance

– the second factor accounts for most of the residual variance, subject to being uncorrelated with the first factor (the second highest variance), etc.

CFA – Common Factor Analysis

Example (cont.):

Consider latent variables: the general ability of a student for Language and Technical subjects.



Possible that some students are great at languages overall, but that they are just bad at German $=>$ specific factors which measure the impact of one variable on the measured variable.

$$y_1 = \lambda_{11}\eta_1 + \lambda_{12}\eta_2 + \varepsilon_1$$

$$\vdots$$

$$y_4 = \lambda_{41}\eta_1 + \lambda_{42}\eta_2 + \varepsilon_4$$

common unique

– λ_{ij} shows "Ability for learning German while taking into account the general ability for learning languages".

CFA – Common Factor Analysis

Common factor model:

$$y_1 = \lambda_{11}\eta_1 + \lambda_{12}\eta_2 + \varepsilon_1$$

$$\vdots$$

$$y_4 = \lambda_{41}\eta_1 + \lambda_{42}\eta_2 + \varepsilon_4$$

common unique

$$\begin{pmatrix} y_1 \\ \vdots \\ y_4 \end{pmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \vdots & \vdots \\ \lambda_{41} & \lambda_{42} \end{bmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_1 \end{pmatrix}$$

– λ_{i1} attributes one part of its variation to first common latent variable (η_1)

– λ_{i2} attributes one part of its variation to first common latent variable (η_2)

– part of variation related to a specific factor (specific to this variable; ϵ).

CFA – Common Factor Analysis

$$\begin{pmatrix} y_1 \\ \vdots \\ y_4 \end{pmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \vdots & \vdots \\ \lambda_{41} & \lambda_{42} \end{bmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_1 \end{pmatrix}$$

$$\bar{y} = \Lambda\bar{\eta} + \bar{\varepsilon}$$

where $\Lambda\bar{\eta}$ – common fixed weighting of unobserved factors

Λ are the values that we need to estimate.

To solve this, the same mathematical solution as in PCA is used, except for a small difference.

– In PCA we apply matrix decomposition to the correlation matrix.

– In Factor Analysis, we apply matrix decomposition to a correlation matrix in which the diagonal entries are replaced by $1 - \text{var}(\mathbf{d})$, one minus the variance of the specific factor of the variable.

CFA – Common Factor Analysis

Assumptions:

1. There are no outliers in data.

2. Sample size should be greater than the factor.

3. The variables used in factor analysis should be linearly related to each other. This can be checked by looking at scatterplots of pairs of variables.

4. There should not be perfect multicollinearity.

However, the variables must also be at least moderately correlated to each other, otherwise the number of factors will be almost the same as the number of original variables, which means that carrying out a factor analysis would be pointless.

5. There should not be homoscedasticity between the variables.

6. Factor analysis is designed for interval data, although it can also be used for ordinal data (e.g. scores assigned to Likert scales).

CFA – Common Factor Analysis

Terminology:

Factor – a latent (hidden, unobserved) variable which describes the association among the number of observed variables.

– The maximum number of factors are equal to a number of observed variables.

– Every factor explains a certain variance in observed variables.

–The factors with the lowest amount of variance were dropped.

Eigenvalues – represent variance explained each factor from the total variance.

Factor Rotation – re-distributed the commonalities with a clear pattern of loadings.

– Rotation is a tool for better interpretation of factor analysis.

– Rotation can be orthogonal or oblique.

CFA – Common Factor Analysis

Factor loading – a matrix which shows

– the relationship of each variable to the underlying factor,

– the correlation coefficient for observed variable and factor,

– the variance explained by the observed variables.

Communalities – the sum of the squared loadings for each variable.

– represents the common variance

– ranges from $0 - 1$ and value close to 1 represents more variance.

CFA – Common Factor Analysis

There are three main steps in a factor analysis:

1. Calculate initial factor loadings.

– **Principal component method:** looks for a set of factors which can account for the total variability in the original variables.

– **Principal axis factoring:** tries to find the lowest number of factors which can account for the variability in the original variables that is associated with these factors

Two methods will tend to give similar results if

- the variables are quite highly correlated

- the number of original variables is quite high.

Whichever method is used, the resulting factors at this stage will be uncorrelated.

CFA – Common Factor Analysis

2. Factor Rotation:

Goal: to improve the overall interpretability.

If there are 'clusters' (groups) of variables that are strongly inter-related

– the rotation is done to try to make variables **within a subgroup** score as highly (positively or negatively) as possible on one particular factor while

– ensuring that the loadings for these variables on the remaining factors are as low as possible.

In other words, the object of the rotation is to try to ensure that all variables have **high loadings only on one factor**.

Types of rotation method:

– orthogonal rotation the rotated factors will remain uncorrelated (varimax rotation)

– in oblique rotation the resulting factors will be correlated.

3. Calculation of factor scores:

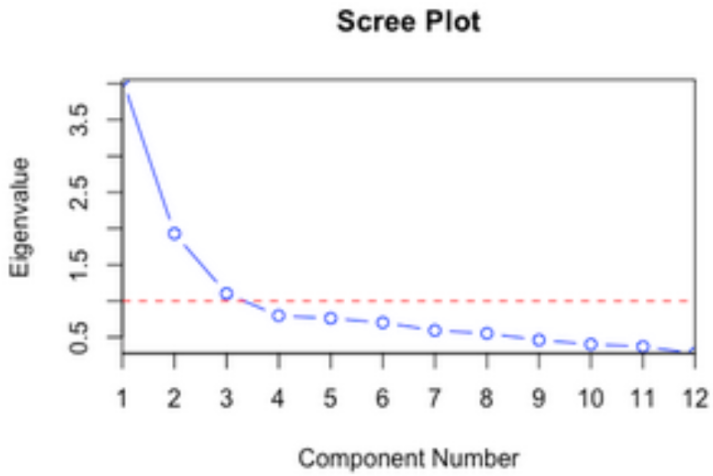
When calculating the final factor scores (the values of the m factors, $F_1, F_2, \ldots, F_m$, for each observation), a decision needs to be made as to how many factors to include.

This is usually done using one of the following methods:

- Choose m such that the factors account for a particular percentage (e.g. 75%) of the total variability in the original variables.

- Choose m to be equal to the number of eigenvalues over 1 (if using the correlation matrix). [A different criteria must be used if using the covariance matrix.]

- Use the scree plot of the eigenvalues. This will indicate whether there is an obvious cut-off between large and small eigenvalues.



Logistic regression

Suppose we work with binary outputs, i.e., $y_i \in \{0, 1\}$.

Linear regression may not be the best model.

- $x^T\beta \in \mathbb{R}$ not in $\{0, 1\}$.

- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y = 1$? (e.g. probability of going to the beach vs outdoors temperature).

Logistic regression: Different perspective. Instead of modelling the $\{0, 1\}$ output, we model the probability that $Y = 0, 1$.

Idea: We model $P(Y = 1 \mid X = x)$.

- Now: $P(Y = 1 \mid X = x) \in [0, 1]$ instead of $\{0, 1\}$.

- We want to relate that probability to $x^T\beta$.

Logistic regression

We assume the log-odds

$$\begin{aligned}\text{logit}(P(Y=1 \mid X=x)) &= \log \frac{P(Y=1 \mid X=x)}{1-P(Y=1 \mid X=x)} \\ &= \log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = x^T\beta.\end{aligned}$$

Equivalently,

$$P(Y=1 \mid X=x) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$$

$$P(Y=0 \mid X=x) = 1 - P(Y=1 \mid X=x) = \frac{1}{1+e^{x^T\beta}}$$

The logistic function is given by the inverse-logit:

$$\text{logit}^{-1}(x) = \text{logistic}(x) = \frac{e^x}{(1+e^x)} = \frac{1}{(1+e^{-x})}$$

Hence, the logit function is the quantile function associated with the standard logistic distribution.

Logistic regression

In summary, we are assuming:

- $Y \mid X = x \sim \text{Bernoulli}(p)$.

- $\text{logit}(p) = \text{logit}(E(Y \mid X = x)) = x^T \beta$.

More generally, one can use a generalized linear model (GLM). A GLM consists of:

- A probability distribution for $Y \mid X = x$ from the exponential family.

- A linear predictor $\eta = x^T \beta$.

- A link function g such that $g(E(Y \mid X = x)) = \eta$.

Example

If we analyze the case of a pairwise relationship (i.e. the case when the probability of an event occurring depends on a single factor $\boldsymbol{x}$), then the logit model can be written as follows:

$$P\left(y_i = 1\right) = F\left(z_i\right) = \frac{1}{1 + e^{-z_i}},$$

where $z_i = \beta_1 + \beta_2 x_i$;

Say, we have estimated the coefficients and the estimated probability of passing the test on hours of study:

$$\hat{P}\left(y_i = 1\right) = \frac{1}{1 + e^{-(-9+0.5x_i)}}.$$

How can such results be interpreted?

– Estimate the probability of the occurrence under certain conditions.

$$\hat{P}\left(y_i = 1 \mid x_i = 15\right) = 0.18.$$

That is, for a student who prepared for 15 hours, the probability of passing the test is 18%.

Logistic regression

– Interpret the results in terms of the change in the dependent variable resulting from a subtle change in the regressor, i.e. marginal effect

To do this, we calculate the derivative of the probability with respect to x:

$$\frac{dP\left(y_i = 1\right)}{dx} = \frac{e^{-(\beta_1+\beta_2 x)}}{\left(1 + e^{-(\beta_1+\beta_2 x)}\right)^2} \cdot \beta_2$$

From an applied point of view, the inconstancy of the marginal effect gives rise to some complexity: it is not very clear at what point to calculate it.

In practice,

1) marginal effect for the sample average.

E.g. marginal effect at point $\bar{x}$ – the average sample preparation time for the test

2) average marginal effect: calculate the marginal effect for each student, then calculate the average of the n marginal effects.

Multiple classes of data

Logistic regression with more than 2 classes

- Suppose now the response can take any of $\{1, \ldots, K\}$ values.

- Can still use logistic regression.

- We use the categorical distribution instead of the Bernoulli distribution.

- $P(Y = i \mid X = x) = p_i, 0 \leq p_i \leq 1, \sum_{i=1}^{K} p_i = 1.$

- Each category has its own set of coefficients:

$$P(Y = i \mid X = x) = \frac{e^{x^T \beta^{(i)}}}{\sum_{i=1}^{K} e^{x^T \beta^{(i)}}}.$$

- Estimation can be done using maximum likelihood as for the binary case.

Linear discriminant analysis (LDA)

- Categorical data Y. Predictors $X_1, \ldots, X_p$.

- We saw how logistic regression can be used to predict Y by modelling the log-odds

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = x^T \beta.$$

- More now examine other models for $P(Y = i \mid X = x)$.

Recall: Bayes' theorem. Given two events A, B :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Using Bayes' theorem

- $P(Y = i \mid X = x)$ harder to model.

- $P(X = x \mid Y = i)$ easier to model.

Going back to our prediction using Bayes' theorem:

$$P(Y = i \mid X = x) = \frac{P(X = x \mid Y = i)P(Y = i)}{P(X = x)}$$

Linear discriminant analysis (LDA)

More precisely, suppose

- $Y \in \{1, \ldots, k\}$.

- $P(Y = i) = \pi_i \quad (i = 1, \ldots, k)$.

- $P(X = x \mid Y = i) \sim f_i(x) \quad (i = 1, \ldots, k)$.

Then

$$\begin{aligned} P(Y = i \mid X = x) &= \frac{P(X = x \mid Y = i)P(Y = i)}{P(X = x)} \\ &= \frac{P(X = x \mid Y = i)P(Y = i)}{\sum_{j=1}^{k} P(X = x \mid Y = j)P(Y = j)} \\ &= \frac{f_i(x)\pi_i}{\sum_{j=1}^{k} f_j(x)\pi_j}. \end{aligned}$$

- We can easily estimate π_i using the proportion of observations in category i.

- We need a model for $f_i(x)$.

Linear discriminant analysis (LDA)

A natural model for the f_j s is the multivariate Gaussian distribution:

$$f_j(x) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma_j}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}.$$

Linear discriminant analysis (LDA): We assume $\Sigma_j = \Sigma$ for all $j = 1, \ldots, k$.

Quadratic discriminant analysis (QDA): general case, i.e., Σ_j can be distinct.

Note: When p is large, using QDA instead of LDA can dramatically increase the number of parameters to estimate.

In order to use LDA or QDA, we need:

- An estimate of the class probabilities π_j.

- An estimate of the mean vectors μ_j.

- An estimate of the covariance matrices Σ_j (or Σ for LDA).

Linear discriminant analysis (LDA)

Estimating the parameters

LDA: Suppose we have N observations, and N_j of these observations belong to the j category ($j = 1, \ldots, k$). We use

- $\hat{\pi}_j = N_j / N$.

- $\hat{\mu}_j = \frac{1}{N_j} \sum_{y_i = j} x_i$ (average of x over each category).

- $\hat{\Sigma} = \frac{1}{N-k} \sum_{j=1}^{k} \sum_{y_i = j} \left(x_i - \hat{\mu}_j\right) \left(x_i - \hat{\mu}_j\right)^T$. (Pooled variance.)

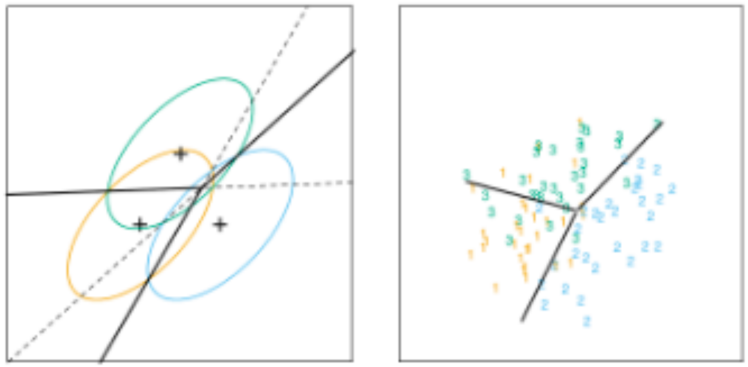


Figure: The left panel shows three Gaussian distributions, with the same covariance and different means. Contours of constant density (95% prob.).
Broken straight lines – The Bayes decision boundaries between each pair of classes Solid lines – Bayes decision boundaries.
Right panel a sample of 30 obs., and the fitted LDA decision boundaries.

Linear discriminant analysis (LDA)

LDA: linearity of the decision boundary In the previous figure, we saw that the decision boundary is linear. Indeed, examining the log-odds:

$$\begin{aligned}
&\log \frac{P(Y = l \mid X = x)}{P(Y = m \mid X = x)} = \log \frac{f_l(x)}{f_m(x)} + \log \frac{\pi_l}{\pi_m} \\
&= \log \frac{\pi_l}{\pi_m} - \frac{1}{2} (\mu_l + \mu_m)^T \Sigma^{-1} (\mu_l - \mu_m) + x^T \Sigma^{-1} (\mu_l - \mu_m) \\
&= \beta_0 + x^T \beta.
\end{aligned}$$

Note that the previous expression is linear in x. Recall that for logistic regression, we model

$$\log \frac{P(Y = l \mid X = x)}{P(Y = m \mid X = x)} = \beta_0 + x^T \beta.$$

How is this different from LDA?

- In LDA, the parameters are more constrained and are not estimated the same way.

- Can lead to smaller variance if the Gaussian model is correct.

- In practice, logistic regression is considered safer and more robust.

- LDA and logistic regression often return similar results.

QDA: quadratic decision boundary

Let us now examing the log-odds for QDA: in that case no simplification occurs as before

$$\log \frac{P(Y = l \mid X = x)}{P(Y = m \mid X = x)}$$
$$= \log \frac{\pi_l}{\pi_m} + \frac{1}{2} \log \frac{\det \Sigma_m}{\det \Sigma_l}$$
$$- \frac{1}{2} (x - \mu_l)^T \Sigma_l^{-1} (x - \mu_l) - \frac{1}{2} (x - \mu_m)^T \Sigma_l^{-1} (x - \mu_m) .$$

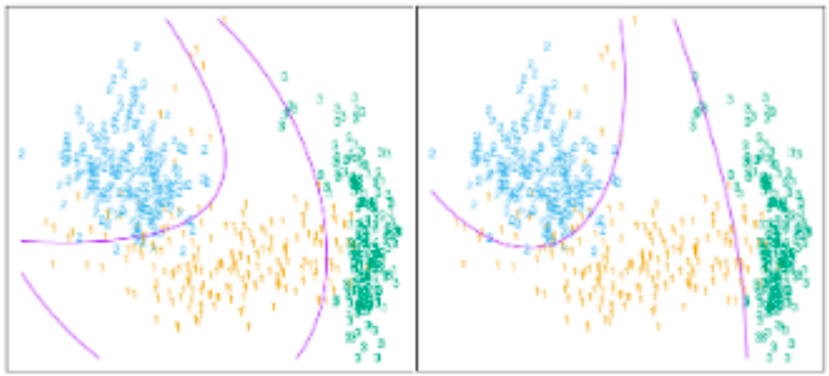


Figure: Two methods for fitting quadratic boundaries.
The left plot the quadratic decision boundaries obtained using LDA in the five-dimensional space $X_1, X_2, X_1 X_2, X_1^2, X_2^2$.
The right plot shows the quadratic decision boundaries found by QDA.
The differences are small, as is usually the case.

QDA: quadratic decision boundary

- Despite their simplicity, LDA and QDA often perform very well.

- Both techniques are widely used.

Problems when $n < p$:

- Estimating covariance matrices when n is small compared to p is challenging.

- The sample covariance (MLE for Gaussian) $S = \frac{1}{n-1} \sum_{j=1}^{n} (x_i - \hat{\mu}) (x_i - \hat{\mu})^T$ has rank at most $\min(n, p)$ so is singular when $n < p$.

- This is a problem since Σ needs to be inverted in LDA and QDA.

Many strategies exist to obtain better estimates of Σ (or Σ_j).

Among them:

- Regularization methods. E.g. $\hat{\Sigma}(\lambda) = \hat{\Sigma} + \lambda I$.

- Graphical modelling (discussed later during the course).

Class Prediction

Linear Discriminant Analysis uses Bayes's theorem to estimate the probabilities.

Assume we have three classes class 0 , class 1 , class 2 in the data set.

Step 1. LDA calculates the prior probabilities of each of the classes $P(y = 0), P(y = 1), P(y = 2)$ of the data set.

Step 2. Consider an observation x.

$P(x \mid y = 0), P(x \mid y = 1), P(x \mid y = 2)$ represent the likelihood functions.

Step 3. Calculates the Posterior probabilities to make predictions.

$$P(y = 0 \mid x) = P(x \mid y = 0) * P(y = 0)/P(x)$$
$$P(y = 1 \mid x) = P(x \mid y = 1) * P(y = 1)/P(x)$$
$$P(y = 2 \mid x) = P(x \mid y = 2) * P(y = 2)/P(x)$$

In general: $P(y = y_i \mid x) = P(x \mid y = y_i) * P(y_i)/P(x)$

The posterior probability = Likelihood*Prior/Evidence.

LDA vs PCA

4 or more genes:

Same problems as with PCA:

1. Dimensionality Reduction.

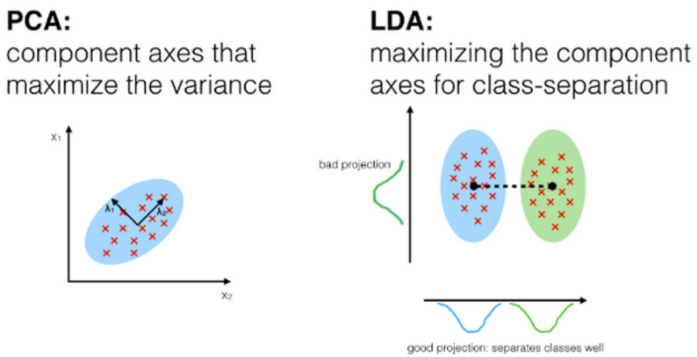
2. Visualize Classes.

- PCA reduces dimensions by focusing on the genes with the most variation.

- PCA is useful for plotting data with a lot of dimensions (or a lot of genes) onto a simple X/Y plot.

NOT interested in the genes with the most variation, BUT in maximizing the seperatibility between the two groups so we can make the best decisions.

- Linear Discriminant Analsys (LDA) is like PCA, but it focuses on maximzing the seperatibility among known categories.



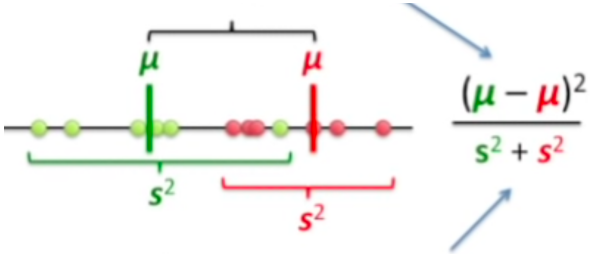
LDA vs PCA

LDA vs PCA

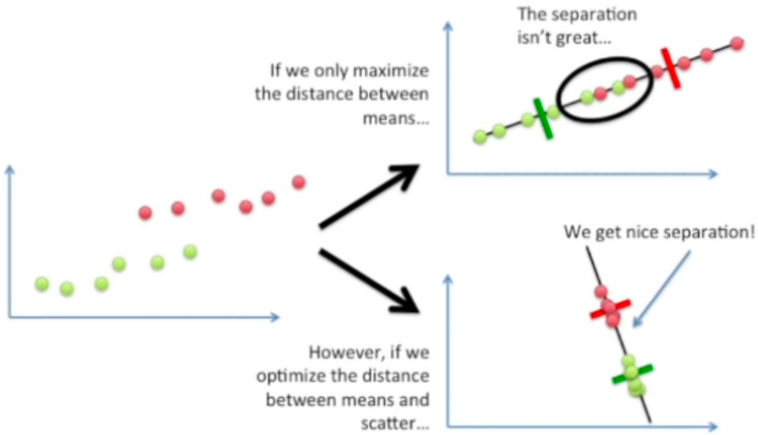
How LDA creates a new axis

The new axis is created according to two criteria (considered simultaneously):

1) Maximize the distance between means.



2) Minimize the variation (which LDA calls "scatter" and is represented by s^2) within each category.



Assumptions of Linear Discriminant Analysis

1) Multivariate normality within groups

2) Homoscedascity: the independent variables have equal variances and covariances across all the categories.

This assumption helps the Linear Discriminant Analysis to create the linear decision boundary between the categories.

When this assumption fails $=>$ Quadratic Discriminant Analysis: the mathematical function which separates the categories will now be quadratic.

Check: Bartlett's test

3) No multicollinearity

The performance of prediction can decrease with the increased correlation between the independent variables.

Note: Studies show that LDA is robust to slight violations of these assumptions.

4) Linearity among all pairs of variables.

Modeling approach

The main objectives of discriminant analysis are to:

1. develop linear combinations of the predictor variables

2. test the existence of significant differences among the groups in terms of the predictor variables

3. identify the predictor variables which contribute most to the inter-group differences

4. classify cases to one of the groups based on the values of the predictor variables

5. evaluate the accuracy of classification.

In LDA we take a different approach and impose grouping on the objects with a view to asking one or or more of three key questions:

1. Is it possible to differentiate among groups on the basis of differences in a set of shared variables

2. Which variables are most important in differentiating among groups?

3. What are the chances that the observed differences are not due to random variation?

Modeling approach

Goal: derive a variate, the linear combination of two (or more) independent variables that will discriminate best between a-priori defined groups.

Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance relative to the within-group variance.

The linear combination for a discriminant analysis, also known as the discriminant function, is derived from an equation that takes the following form:

$$Z_{ik} = b_{0i} + b_{1i}X_{1k} + \ldots + b_{ji}X_{Jk}$$

Z_{ik} ... discriminant score of discriminant function i for object k

$Z_{ik} = b_{0i} + b_{1i}X_{1k} + \ldots + b_{ji}X_{Jk}$

Z_{ik} ... discriminant score of discriminant function i for object k,$i = 1, \ldots, G-1$

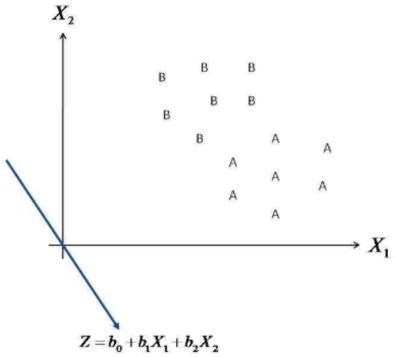
X_{jk} ... independent variable j for object k,$j = 1, 2, \ldots, J$

b_{ji} ... discriminant weight for independent variable j and discriminant function i

b_{0i} ... constant of discriminant function i

Note: Different kinds of specifications for DA functions are available.

Modeling approach



Estimation of the DA function(s)

$$\max_b \gamma = \frac{\sum_{g=1}^{G} I_g (\bar{Z}_g - \bar{Z})^2}{\sum_{g=1}^{G} \sum_{i=1}^{I_g} (\bar{Z}_{gi} - \bar{Z}_g)^2} = \frac{SS_b}{SS_w}$$

γ – discriminant criteria (eigen value)

I_g – size of group g

Z_{gi} – i-th discriminant value of group g

SS_b – sum of squared deviations between groups, explained deviation

SS_w – sum of squared deviations within groups, remaining/unexplained deviations

Note: in this sence LDA is an extension of MANOVA.

Modeling approach

Wilks' lambda evaluates the statistical significance of the discriminatory power of the discriminant function.

As γ gives the maximal value of the discriminant criteria, a high value of γ indicates high quality. However, γ has no upper limit. Therefore appropriate transformations of γ are used:

$$\frac{\gamma}{1+\gamma} = \frac{SS_b}{SS_b + SS_w} = \frac{\text{explained variation}}{\text{total variation}}$$

$$\frac{1}{1+\gamma} = \frac{SS_w}{SS_b + SS_w} = \frac{\text{unexplained variation}}{\text{total variation}}$$

$\sqrt{\frac{\gamma}{1+\gamma}}$ – canonical correlation coefficient c.

$\frac{1}{1+\gamma} \in [0; 1]$ is called Wilks' Lambda Λ.

Λ also shows whether the group means are equal.

– Large values (near 1) ndicate that the group means may be similar.

– Small values (near 0) indicate that the group means may be different.

Modeling approach

Wilks' Lambda (Λ) is an inverse quality criterium.

If K DA functions are computed the characteristics γ, c, and Λ are computed separately for each DA function.

In order to analyze the dissimilarity of the groups multivariate Wilks' Lambda is calculated:

$$\Lambda = \prod_{k=1}^{K} \frac{1}{1+\gamma_k}$$

γ_k denotes the eigen value of the k-th DA function.

Modeling approach

A suitable transformation of Λ allows a significance test regarding the DA function:

$$\chi^2_B = -\left(N - \frac{J+G}{2} - 1\right) \ln(\Lambda) \sim \chi^2_{J\cdot(G-1)}$$

N – number of observation units

J – number of variables

G – number of groups

Λ – multivariate Wilks' Lambda

Hypothesis:

H_0 : The groups are not different from each other.

H_1 : There are different groups.

Modeling approach

In order to test whether an additional DA function is necessary given k DA functions are already estimated Wilks' Lambda for the residual discriminant value can be used:

$$\Lambda_k = \prod_{q=k+1}^{K} \frac{1}{1+\gamma_q} \sim \chi^2_{(J-k)\cdot(G-k-1)}$$

Note:

– If one or more functions are deemed not statistically significant, the discriminant model should be re-estimated with the number of functions to be derived limited to the number of significant functions.

– Assessment of predictive accuracy and the interpretation of the discriminant functions will be based only on significant functions.

– It can be useful not to use all significant DA functions. Generally, 2-3 DA functions are sufficient.

Assessing Group Membership Prediction Accuracy

Hit Ratio – the correctly classified observation units divided by the number of observation units.

$$HR = \frac{c_{11} + c_{22}}{c_{11} + c_{12} + c_{21} + c_{22}}$$

The results are summarized in a classification matrix:

true class membership	predicted class membership	
	Group A	Group B
Group A	C_{11}	C_{12}
Group B	C_{21}	C_{22}

- The hit ratio must be compared with the a-priori hit ratio or a random assignment.

Modeling approach

Importance of the Independent Variables

– If the discriminant function is statistically significant

– and the classification accuracy is acceptable,

the focus lies on making substantive interpretations of the findings.

Three methods of determining the relative importance have been proposed:

(1) Standardized discriminant weights.

(2) Discriminant loadings (structure correlation).

(3) Partial F-values.

Modeling approach

(1) Standardized discriminant weights.

Goal: examine the sign and magnitude of the standardized discriminant weight (discriminant coefficient) assigned to each variable in computing the discriminant functions.

$$b_j^* = b_j \cdot s_j$$

b_j ... discriminant coefficient of variable j s_j^2 ... within group variance of variable j

– When the sign is ignored, each weight represents the relative contribution of its associated variable to that function.

– Independent variables with relatively larger weights contribute more to the discriminating power of the function than do variables with smaller weights.

For more than one DA function mean standardized discriminant weight for each variable is calculated:

$$\bar{b}_j = \sum_{k=1}^{K} \left| b_{jk}^* \right| \cdot EP_k$$

b_{jk}^* ... standardized discriminant weight for variable j and discriminant function k
EP_k ... Eigenvalue proportion of DA function k

Modeling approach

(2) Discriminant loadings (structure correlation) – measures of simple correlations between each predictor variable and the discriminant function(s).

– represents the variance which the predictor variable shares with the discriminant function(s)

– assessing the relative contribution of each independent variable to the DA function.

(3) Partial F-values – the absolute sizes of the significant F values.

Large F values indicate greater discriminatory power .

– In practice, rankings using the F-values approach are the same as the ranking derived from using discriminant weights, but the F values indicate the associated level of significance for each variable.