**Customer analytics**

[**EXPLAINING CUSTOMER BEHAVIOR 3**](#_heading=)

[**Factor analysis 3**](#_heading=)

[Difference between PCA and Factor analysis 4](#_heading=h.qvckrq3q1r30)

[Eigen values and communality. 5](#_heading=h.wlfo8m1edjqx)

[EFA main rules: 5](#_heading=h.6i63hoqnjyj2)

[EFA acceptance criteria 5](#_heading=h.v4xql4c63jbs)

[How to select the number of factors in EFA. 6](#_heading=h.afddvc6aggzs)

[Guidelines for loadings in EFA 6](#_heading=h.rq99dhweywgv)

[**CB - SEM = Covariance Based Structured equation modelling 6**](#_heading=h.2vnwcwx7xohx)

[Difference from linear regression 8](#_heading=h.bu69z19db4kh)

[DISTINGUISHING FEATURES OF SEM 8](#_heading=h.gmp1z25psb3b)

[GRAPHICAL REPRESENTATION OF OBSERVABLE AND LATENT VARIABLES 9](#_heading=h.nhhjw1jh7wem)

[TYPES OF RELATIONSHIPS INVOLVED IN SEM 10](#_heading=h.109050ndsoay)

[TYPES OF ESTIMATES INVOLVED IN SEM 10](#_heading=h.gsgbfx20hy2v)

[Dependence vs causality 10](#_heading=h.vdffys7lp9ly)

[**CFA – Confirmatory factor analysis. 11**](#_heading=h.raj66hoe1q8k)

[Uni-dimensionality = no cross loadings 11](#_heading=h.4ui0wtasuqo)

[Correlated error variance. 12](#_heading=h.kg51qatop33e)

[Numbers of items pr. Construct 12](#_heading=h.6nkfbrbksyz)

[Transformation of our measurable variable in PLS-SEM. 13](#_heading=h.6ssnvd73unzd)

[Assumptions and sample size. 13](#_heading=h.gtkk80bxstur)

[Accessing model fit in CFA 13](#_heading=h.xbfcfg5v51ts)

[Construct validity 14](#_heading=h.wpit0cywejj6)

[Modification indeks 16](#_heading=h.fl3le15r8weq)

[Evaluation of the structural model. 17](#_heading=h.5zdjbtvnzn2c)

[Recursive vs nonrecursive models. 17](#_heading=h.kiaonguedu3f)

[Nested models: 17](#_heading=h.72il23zxmb1)

[**PLS-SEM (Morten) 18**](#_heading=h.dm762u7b79hz)

[**Diagrammer og notation 18**](#_heading=)

[**Assessing Reflectively measurement model. 20**](#_heading=)

[Evaluating formatively measured constructs. 21](#_heading=h.rdf0ev5sct5o)

[**Assessing structural model: 22**](#_heading=)

[**Mediating and moderating effects 23**](#_heading=)

[**CAUSAL DISCOVERY , PRODUCT RECOMMENDATION & TARGETING (I 25**](#_heading=h.ywug47caic6c)

[**Bayesian networks 25**](#_heading=h.prlfkmq83qi3)

[Terminology 25](#_heading=h.p5riei9mpf1q)

[Characteristics 26](#_heading=h.gzbmvneknnve)

[Inference 26](#_heading=h.30gg22rsuw2w)

[BN structure 27](#_heading=h.2ddsi5qdwilz)

[**1.Naïve Bayes 27**](#_heading=)

[**2.Tree Augmented Naïve Bayes (TAN) 27**](#_heading=)

[**Sames as Naive Bayes but allows relationships between features. 27**](#_heading=)

[**3.Constrained-based algorithms 27**](#_heading=)

[**4. Score-based algorithms 29**](#_heading=)

[Dag elements 29](#_heading=h.hrdbvawczw89)

[**Setting evidence to infer something 30**](#_heading=)

[**SERIAL CONNECTION 31**](#_heading=)

[**DISCOUNTING 31**](#_heading=)

[**DIVERGENT (COMMON CAUSE) 32**](#_heading=)

[D – SEPARATION CONCEPT 32](#_heading=h.9m0n2hnwaggq)

[MARKOV BLANKET 33](#_heading=h.brmcvj8iy7e1)

[MARKOV PROPERTY (conditional independent) 34](#_heading=h.1z6ebhkdcn6g)

[**Usecases for Baysiean network 35**](#_heading=)

[**a. Customer retention 35**](#_heading=)

[Foreward inference 35](#_heading=h.ijweneyc2jb6)

[Backwards inference 36](#_heading=h.xl7x4vs24rcp)

[**b. Targeted advertising 37**](#_heading=)

[**c. Product recommendation 38**](#_heading=)

[**Clickpath analysis and choice modelling 41**](#_heading=h.j6q67gpwzb55)

[Choice analysis. 41](#_heading=h.l2hd3p6486p)

[interpreting results: 42](#_heading=h.pa8shyh5lbey)

[**Segmentation: 43**](#_heading=)

[**Segmentation as Clustering and Classification 43**](#_heading=)

[Focus on clustering methods. 44](#_heading=h.odu5cv2r00t)

[**Two main distinctions: 44**](#_heading=)

[**Hierarchical Clustering: hclust() 44**](#_heading=)

[**Mean-Based Clustering: kmeans() 45**](#_heading=)

[**Model-Based Clustering: Mclust() 46**](#_heading=)

[**Latent Class Analysis: poLCA() 46**](#_heading=)

[**Comparing Models with BIC() and log likelyhood 46**](#_heading=)

[Outcome 47](#_heading=h.v4ojkngzcczy)

[Good practice 47](#_heading=h.45j51tz6anz1)

[Research design 47](#_heading=h.quwnuxsstktg)

[Assumptions 48](#_heading=h.fexk85egxez)

[Measures of similarity 48](#_heading=h.3udy4uw81bj3)

[Standardization 49](#_heading=h.wjhnmtave3ea)

[Hierarchical methods 49](#_heading=h.2r76jjy5lz39)

[Linkage 49](#_heading=h.o18znnu9jnvj)

[Example Euclidean distance algorithms + single linkage: 50](#_heading=h.my01nzfqu6wu)

[Nonhierarchical methods 51](#_heading=h.9oewduw9l0ad)

[Comparing hierarchical and nonhierarchical methods 52](#_heading=h.iva8jjind6pm)

[Deciding on the number of clusters 52](#_heading=h.4ys0igxalkga)

[Dendograms 52](#_heading=h.9rndbg9nynvr)

[Profiling 53](#_heading=h.zfe71dcukegb)

[Validation and description 53](#_heading=h.s9bl8gkpvsr5)

[Key lessons from hierarchical and non-hierarchical cluster analysis 54](#_heading=h.soenx0of9jj2)

[**Segmentation II - Model-based clustering 54**](#_heading=h.xii5ncuoym2)

[The formal model 55](#_heading=h.2arn8eqnmt51)

[Determining the number of clusters (cont’d) 55](#_heading=h.gxb0ftt7cj74)

[**Segmentation III - Finding, assessing, and predicting customer segments 56**](#_heading=h.nrlz5y4tdp4u)

[What is segmentation and why do it 56](#_heading=h.4kyk3vynnwa)

[Approaches to segmentation 56](#_heading=h.9m0nh0mpnqow)

[General quality criteria to assess segmentation solutions 57](#_heading=h.9ywrxj626yx9)

[Measures of reproducibility/stability 57](#_heading=h.bmgqv3ww9diu)

[Gorge plot 59](#_heading=h.7jistjlf4ne6)

[Segment level stability 59](#_heading=h.rtr1h4p1gzdd)

[Segment level stability across solutions (SLSA) 60](#_heading=h.sc7u6fev2m26)

[Segment level stability within solutions (SLSW) 60](#_heading=h.7t62ezhd6usr)

[**Segmentation IV - Introduction to latent class analysis 61**](#_heading=h.1cu56799d2ev)

[**Latent class analysis. 61**](#_heading=)

[Latent class vs. cluster analysis 61](#_heading=h.f22v8gd1xfjq)

[Measures of association for binary data 61](#_heading=h.f5x1ekvtk582)

[The data matrix (cont’d) 62](#_heading=h.v15hhu9wkc7a)

[Assumptions 62](#_heading=h.bmmi7qq06n3b)

[The J-class model 63](#_heading=h.g88ewbsi13j)

[Goodness-of-fit – global tests 63](#_heading=h.xh02uiuut3el)

[Goodness-of-fit – model comparisons 64](#_heading=h.olvo07n6exxk)

[**Segmentation V - Advanced latent class analysis 64**](#_heading=h.r8p9n4uxsk0n)

[LCA with concomitant variables 64](#_heading=h.re5rrc5p8rao)

[Finite mixture of regressions 66](#_heading=h.585l2v8ty12d)

[**PRODUCT RECOMMENDATION & CUSTOMER TARGETING (II) 67**](#_heading=h.w2pys48ofn0m)

[**Association rules: 67**](#_heading=h.pcoelhy1l6yp)

[**Collaborative filtering: 69**](#_heading=h.7uwb25hhq5ym)

[user based similarity: 69](#_heading=h.efzgkakkdbt6)

[Item based collaborative filtering: 70](#_heading=h.w98vhq3g26hi)

[SVD and SVD Funk 71](#_heading=h.m689han2h0op)

# EXPLAINING CUSTOMER BEHAVIOR

## Factor analysis

Within customer analysis we use both exploratory (EFA) and confirmatory factor analysis (CFA). In short:

* In exploratory FA, factors are discovered
* In confirmatory FA, factors are confirmed

In general factors analysis is much like PCA. When using FA, we assume that there is some factor or construct we believe affect the dependent variable. However, there may not be one variable which measures this construct. Therefore, we want to measure variables to get an idea as to how we define an understand our factor/latent construct. (This could be job satisfaction)

In general, the factors are linear combinations of the variables, that we are able to measure. This is often done by looking at the correlation matrix, where we want to summarize items which have a high correlation.

Objectives:

* for developing objective measures for abstract constructs e.g. personality, intellectual ability
* for summarizing data and testing a theory about underlying process
* for data reduction and multicollinearity resolution

**Matrices in the solution** (Slide 8 lecture 1)

**Orthogonal rotation** - factors are uncorrelated

* loading matrix - meaning of factors

**Oblique rotation** - factors are correlated

* factor correlation matrix
* structure matrix
* pattern matrix - meaning of factors¨

### Difference between PCA and Factor analysis

Et billede, der indeholder tekst, diagram, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

As one can tell in the factor analysis, we measure the correlation/loading between the factor and the variable. This is done using many simple linear regressions, so that we can tell how much of the variance in each variable is explained by the factor. This I kind of the reflective measures of or factors as in PLS SEM.

It is in essence the latent factors, which affects the item and not the items which affects the latent factors.

### Eigen values and communality.

When a factor has an Eigenvalue greater than one, we decide to keep the factor or principal component. The Eigenvalue is the squared sum of deviation (blue) and the communality is the squared sum of deviance across all the factors (red).Et billede, der indeholder tekst, skærmbillede, nummer/tal, Font/skrifttype

Automatisk genereret beskrivelse

**Eigen values** = Det er summen af vores loadings i anden. Bidrager variablene meget vil vi altså bruge vores Principal component.

**Comunality** - hvor meget en variable forklarer i alle loading. Er den lav, kan vi vælge ikke at bruge en variable

### EFA main rules:

* Have between 5-6 variables for each factors (as some might be dropped in the process)
  + Factors with only 1-2 variables are not stable.
    - Find other variables or drop the factor.
* Items should only be highly correlated with one factor.
* Sample should consist of minimum 300 observations. (However dependent on the number of factors)
* Multivariate normality is assumed when statistical inference is used to determine the no. of factors. (This assumption may however not always be fulfilled).
* It may be preferable to standardize the variables, to make sure that great differences in levels don’t affect our output.
* Number of observed variable \* 10

### EFA acceptance criteria

1. Matrix of correlations should be greater than +- 0,30.
2. Partial correlations should be small
3. Barlett test – should be not significant (This test is however very sensitive to the size of N)
4. Kaiser MSA test should be above 0,60.

Below one can see a one and two factor example.

Et billede, der indeholder tekst, diagram, linje/række, Font/skrifttype

Automatisk genereret beskrivelse

Altså bare en lineær analyse for at undersøge hvor meget vores faktor forklare af hver variable = vores loading. I det tilfælde, hvor der er en univariate analyse, så vil vores loadings/beta/lambda også være vores forklaringsgrad for den enkelte univariate regression.

Et billede, der indeholder diagram, linje/række, cirkel

Automatisk genereret beskrivelse

### How to select the number of factors in EFA.

1. Based on prior knowledge
2. Until the cumulative proportion of variance explained is greater that 0,60
3. Only use factors, where eigenvalues is greater than 1
4. Inspect scree plots and follow the elbow rule
5. Use other methods like cross validation

### Guidelines for loadings in EFA

1. Std. loading should as a minimum be between 0,3 and 0,4.
2. For any practical use the loading should be greater than 0,50
3. The communality should be greater than 0,50 to include factor.
4. Eigen values should be greater than 1.
5. Factor internal consistency with Cronback Alpha greater than 0,6.

## CB - SEM = Covariance Based Structured equation modelling

SEM is a form of graphical modeling, and therefore a system in which relationships can be represented in either graphical or equational form. An equation is said to be structural is there exist sufficient evidence from all available sources to support the interpretation that x1 has a causal effect on y1. In SEM we are examining how factors affect each other, rather than how individual items/variables affect each other.

Et billede, der indeholder tekst, Font/skrifttype, diagram, linje/række

Automatisk genereret beskrivelse

SEM can be defined as the use of two or more structural equations to represent complex hypotheses. In this case our x’s are the latent variables/factors we have identified in the EFA.

3. SEM is a process for testing a structural theory (Hair et al., 2010)

4. SEM is explicitly aimed at complex testing of theory, and superbly combines methods hitherto considered and used separately. It also makes possible the rigorous testing of theories that have until now been very difficult to test adequately.” (Kerlinger, 1977)

5. SEM is a framework for building and evaluating multivariate hypotheses about multipleprocesses. It is not depended on a particular estimation method. (Grace, 2014) Et billede, der indeholder tekst, diagram, linje/række, Font/skrifttype

Automatisk genereret beskrivelse

### Difference from linear regression

Et billede, der indeholder diagram, skitse, design

Automatisk genereret beskrivelse

To left is a linear regression where we only have one independent variable. To the right is SEM where we can have more than one dependent variable. Using SEM, we are also able to examine the effect of one construct to another like Y1 -> Y2 while we are still trying to determine Y3..

### DISTINGUISHING FEATURES OF SEM

In SEM we are dealing with variables, which are things we can measure in a questionnaire. We are also dealing with constructs/factors which is unobservable like job satisfaction, see below.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

As we cannot measure job satisfaction directly, we “make it up” from 5-6 questions to make sure that we capture alle the essence of variables, which affects job satisfaction. (In PLS-SEM we assume that all constructs are reflectively measured)

### GRAPHICAL REPRESENTATION OF OBSERVABLE AND LATENT VARIABLES

Et billede, der indeholder skærmbillede, tekst, diagram, linje/række

Automatisk genereret beskrivelse

As seen above, we assume that it is the latent construct that explains variance in our measurable variables. The things that we don’t explain with our latent variable is the error of measurement of each variable.

### TYPES OF RELATIONSHIPS INVOLVED IN SEM

Et billede, der indeholder tekst, diagram, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

### TYPES OF ESTIMATES INVOLVED IN SEM

Et billede, der indeholder tekst, diagram, skærmbillede

Automatisk genereret beskrivelse

### Dependence vs causality

The age old question” Is correlation = causality “. SEM is in essence measuring correlation between our latent variable. The path estimate can be interpreated as beta in linear regression, that a one unit increase in x will give an increase in y of …

To determine that a relationship reflect causal dependence, we need 4 types of evidence to be present. However even though they are present, they are no guaranty for a causal relationship.

* **Covariation**
  + The path coefficient has to be significant.
  + This shows that a factors affects another
* **Sequence**
  + In SEM we often cannot differentiate between time 0 and time 1 as seen below. Therefore, SEM can’t by itself determine the sequence and thereby the direction of the causal effect. We therefore need to use other experiments, to test what way the causality flows.

**Et billede, der indeholder diagram, skærmbillede, tekst, linje/række

Automatisk genereret beskrivelse**

* **Nonspurious correlation**
  + Think of omitted variable bias. If we don’t have/control for the right factors, there could be another factors determining that changes in our factors.

Et billede, der indeholder tekst, skærmbillede, cirkel, diagram

Automatisk genereret beskrivelse

* **Theoretical support**
  + As mentioned before, SEM do not tell us everything. Therefore we often other theoretical support to claim any type of causal relationship.

## CFA – Confirmatory factor analysis.

### Uni-dimensionality = no cross loadingsEt billede, der indeholder diagram, skitse, tegning, hvid Automatisk genereret beskrivelse

Here we based on other knowledge have an idea of the factors we need to create and what variables they are made up of. In an ideal scenario, we want to end up with something as seen below, where each variable only is associated with one construct.

However there may be situation where a variable seem to determine both teamwork and compensation as seen below. We do not allow this loading to be too high, as this indicates a problem with construct validity. If a variable can determine 2 factors, it may mean that the factors measure the same thing.

Et billede, der indeholder diagram, tegning, skitse, linje/række

Automatisk genereret beskrivelse

### Correlated error variance.

Et billede, der indeholder diagram, skitse, cirkel, linje/række

Automatisk genereret beskrivelse

Correlated error variance is bad. If the correlation is too high (above 0,5) we should consider dropping the factors as high correlation indicates that there is another construct, which explain the variance in X1 and X2.

We instead want to see, that one variable is only linked to one construct and have no covariance with other measurement errors.

Et billede, der indeholder diagram, skitse, linje/række, cirkel

Automatisk genereret beskrivelse

### Numbers of items pr. Construct

Different from the EFA, we here only need 3-4 variables pr. Construct. As these are based on prior knowledge we are less likely to drop these item. Therefore we do not need to collect 6 as in EFA.

### Transformation of our measurable variable in PLS-SEM.

In SEM we can apply different types of standardization. This may however not always be the best way to approach a problem with SEM. In SEM we often want to understand how a one unit increase in X affect Y. Therefore, scaling or normalization of our variable could make interpretation harder.

### Assumptions and sample size.

How many observations to use depends on the number of construct and so on. But is also limited as we assume normality among our variables and SEM have a hard time dealing with missing values.

As a rule of thumb, we want at least 350 observations – if there is less observations or the assumption of normality is violated we should instead use PLS SEM.

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, algebra

Automatisk genereret beskrivelse

General suggestions for sample size.

Et billede, der indeholder tekst, kvittering, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

### Accessing model fit in CFA

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, algebra

Automatisk genereret beskrivelse

When assessing the model we look at the following:

**CMIN = Chi Square Statistic**

* Here we are testing whether there is a difference between our estimated and the observed covariance matrix.
* Here we seek a non-significant effect, as we want the estimated to be a true approximation of the observed covariance matrix.

**RMSEA = Root Mean Square Error Of Approximation**

* reflects how well the model, with unknown but optimally chosen parameter estimates, would fit the population's covariance matrix. It takes into account the model complexity by penalizing the addition of parameters, hence favoring parsimonious models.
* Desired value below 0,08

**RMSR = Root Mean Square Residual**

* Is the difference between the observed correlation matrix and the predicted correlation matrix. It is calculated as the square root of the sum of the squared differences between the observed and predicted correlations, averaged over all elements in the correlation matrix.
* Desired value below 0,08

**NFI = Normed Fit Indeks**

* NFI is a measure used to assess the goodness of fit of a structural equation model (SEM). It compares the fit of a user-specified model to the fit of a null model, where the null model assumes no relationships among the variables.
* Should be 0,95 or above

**IFI = Incremental Fit Indeks**

* IFI evaluates how well a specified model fits the data compared to a baseline model, adjusting for the degrees of freedom of the models. It is designed to be less sensitive to sample size than some other fit indices
* Should be 0,95 or above

**TLI = Tucker Lewis indeks**

* PLS-SEM aims to maximize the explained variance of the endogenous constructs
* Should be above 0,90 or 0,95

**CFI = Comparative Fit Indeks**

* Should be above 0,90 or 0,95

**Standardized Root Mean Square Residual (SRMR)**

* Values below 0.08 suggest acceptable fit.

### Construct validity

Is the extent to which the measured items reflect the theoretical latent variable, that they are designed to measure.

**Convergent validity**

* Inspect factor loadings.
  + Above 0,5 but ideally above 0,7 to be useful.
  + Below 0,5 should be considered for deletion.
* Average variance extracted (AVE)
  + Above 0,5
  + Shows how much variance our construct on average determines in our indicators.
* Construct reliability (CR)
  + Whether our items measure our latent construct. How much it our items correlate with our construct.
  + Above 0,7

**Discriminant validity**

* Is the extent to which a construct is truly distinct from other constructs. AVE estimates for two factors should be greater than the square of the correlation between the two factors to provide evidence of discriminant validity.

**Nomological validity**

* The extent to which a construct fits into a theoretical network of related constructs and the expected relationships between them are confirmed by empirical data. It helps validate the theoretical relationships between constructs.
* In essence we want to see if the path coefficients are aligned with what previous research suggests.

**Face validity**

* In short do the construct make sense at face value and does other theory support it.
* If this is the case we would most likely want to keep it in our analysis.

**Standardized residuals**

* Should be below 2,5.
* Between 2,5 and 4 – nothing should be done unless other problems
* Above 4 we might delete an item as the error is to large.

### Modification indeks

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

“~~” —--> **Covariance or Residual Covariance**

* Usage: This symbol represents the covariance between two observed variables or the residual covariance between two latent variables.
* Interpretation in Modification Indices: When you see ~~ in the modification indices output, it suggests adding a covariance path between two variables that are not currently covaried in your model. This could mean allowing for correlated errors between two observed variables or specifying a covariance between two latent variables.

“=~” —-> **Latent Variable Indicator**

* Usage: This symbol represents the relationship between a latent variable and its indicators (observed variables). It specifies that the latent variable is measured by the observed variables.
* Interpretation in Modification Indices: When you see =~ in the modification indices output, it suggests adding or modifying the loading of an observed variable on a latent variable. This means including a new indicator for a latent construct or adjusting the factor loading structure.

“~” —> **Regression Path**

* Usage: This symbol represents a regression path or directional effect from one variable to another. It specifies that one variable predicts or influences another.
* Interpretation in Modification Indices: When you see ~ in the modification indices output, it suggests adding or modifying a regression path between two variables. This could mean including a new direct effect between two variables that are not currently linked in your model.

### Evaluation of the structural model.

We look at the path coefficients and discuss the effect and whether they are significant.

### Recursive vs nonrecursive models.

The primary difference between recursive and nonrecursive models in PLS-SEM lies in the directionality and complexity of the relationships among variables. Recursive models feature unidirectional, simple relationships with no feedback loops, making them easier to estimate and interpret. In contrast, nonrecursive models involve bidirectional relationships and feedback loops, adding complexity to the model estimation and interpretation process. Understanding these differences is crucial for accurately specifying and analyzing SEM models in research.

### Nested models:

A model is nested within another model if it contains the same number of variables and can be formed from the other model by altering relationship such as changing, adding or deleting relations ships, thereby imposing other constraints.

Et billede, der indeholder skitse, diagram, Stregtegning, tegning

Automatisk genereret beskrivelse

If the models are nested we can easily compare the performance by looking at the changes in the chi squared goodness of fit test and determine if this is significant. Et billede, der indeholder tekst, håndskrift

Automatisk genereret beskrivelse

If we find that model B have a lover chi squared statistic and this is significant, we may determine that model B does a better job approximating the underlying relationship. However as with most other things in SEM we would of course like theoretical support for this extra path.

If models are not nested we compare the BIC to determine what models is best.

Hvordan skal vi forstå det der modification index.

## PLS-SEM (Morten)

### Diagrammer og notation

Et billede, der indeholder diagram, cirkel, skitse, linje/række

Automatisk genereret beskrivelse

**Outer/Measurement model:** Is how the different indicators make up our latent constructs.

**Inner/structural model:** Is how the different latent constructs affect each other.

**Indicators/variables:** Are variables that we can clearly and easily measure in the real world. They are not our construct but may be part of how we define them. In the picture above the x’s are all indicators.

**Error – e :** Is the variance within a indicator not explained by a latent construct. We will therefore only have error for reflectively measured constructs.

**Error – Z :** Is the variance within a latent construct not explained the other latent construct pointing towards it.

**Single measure construct:** In the picture Y4 is a single measure construct and therefore have no error.

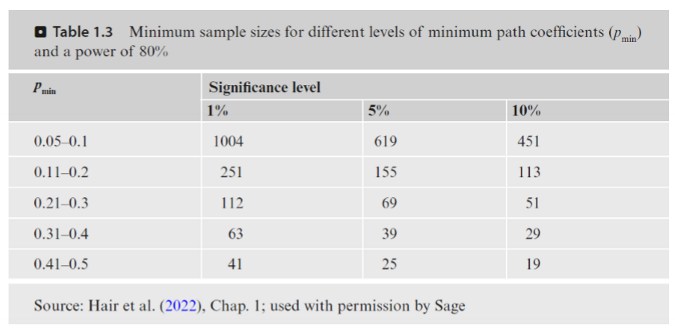
**Reflectively measured constructs (Mode A):** Have the arrow pointing from the latent construct to the indicator, they furthermore have an error term. Reflectively describe the way we think of the construct. We believe that reflectively constructs causes the indicators, which is also why we have an error term. We believe that a change in the construct, will also lead to a change in the indicators. As all indicators have a high correlation with latent construct, we don’t often weaken the construct a lot by removing an indicator, as they to some extent explain the same thing.

**Formatively measured constructs (Mode B):** Have an arrow pointing from the indicator to the latent construct. Here we assume that a linear combination of indicators make up the indicators, make up the construct. There may therefore not be a high correlation among indicators. As the different indicators explain different parts of the latent construct, we can’t as easily exclude a indicator, without messing with the explanation of our latent construct.

Et billede, der indeholder tekst, cirkel, Font/skrifttype, skærmbillede

Automatisk genereret beskrivelse

**Assumption and sample size.**

How do we interpret the plot below:

The strength of PLS-SEM is that we do not have any assumptions for the data. However as best practice we still remove outliers and standardize the numeric variables. However there is no need for the data to follow a normal distribution.

When dealing with missing data the most common thing to do is just deleting the variable or simply using mean replacement. There are other more sophisticated methods, however it has not been investigated how these methods affect the model.

**Nogle som er drøn skarpe i den indre udregning af vægte i formative og reflectively measured models i Mortens slide 8b.**

### Assessing Reflectively measurement model.

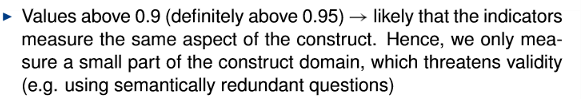
We should always discus the following to make sure that our model is valid for prediction and interpretation.

**Indicator reliability -** We inspect how much of the variance in the indicator that our construct actually explain.

* To do so we inspect the loadings.:
  + Size of correlation weights/loadings.
    - Keep if loading greater than 0,708
    - Evaluate if loading between 0,4-0,7
    - Delete if loading smaller than 0,40

**Internal consistency reliability**

* Have we chosen the right indicators to measure the construct, how do they perform as a whole? We look at the average correlation between indicators and constructs.
  + Cronbach’s alpha
    - Assume equal loading
    - Look at the covariance between the constructs and all indicators. If the covariance is high, we will get a small fraction. If we minus 1, we get something close to 1.
    - At least above 0,7
    - Ideally around 0,8-0,9
  + Composite reliability
    - Same as above but dont assume equal loading
  + Cronbach tenders to underestimate and composite to overestimate

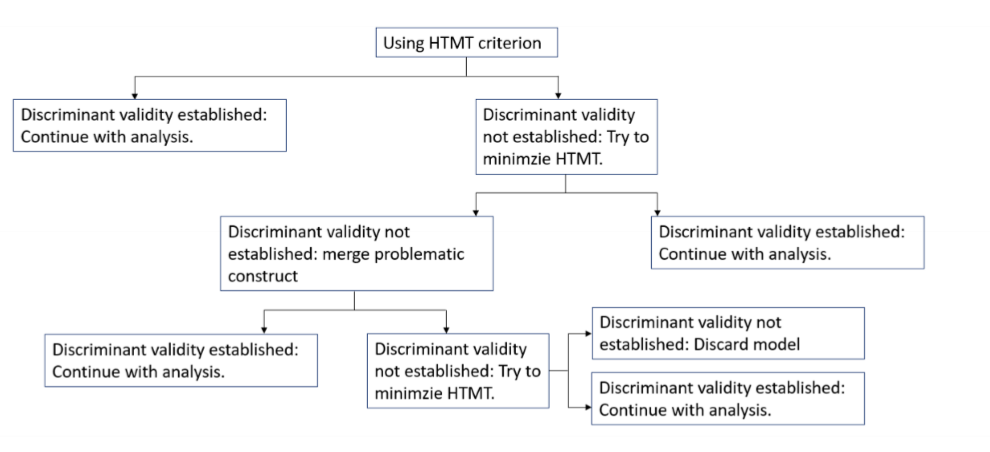


**Convergent validity**

* How much of the variance in all our indicators is explained by the construct. How well does our construct explain variance in all our indicators for the construct
  + Average variance extracted
    - We simply look at the average of variance explained in all the indicators, by our latent construct.
    - This should be greater than 0,50

**Discriminant validity**

* We want to make sure that each construct is well defined and significantly different from other constructs.
  + Fornell-Larcker criterion
    - Squarerot of AVE most be higher than the correlation with other reflectivity measured constructs
  + Heterotrait-monotrait (HTMT) ratio of correlations assuming that each construct was measured perfectly.
    - Should be smaller than 0,85
    - If higher than 0,85 - consider removing indicators with cross loadings
    - Or delete construct who appear to be to much alike.



### Evaluating formatively measured constructs.

**Formatively measuret constructs:**  Are evaluated differently, as there is not the same correlation between the different indicators - this because they measure different element of what we believe make up the construct. For the same reason, we should be careful when deleting indicators. More often than not, we “just” do a sanity check and determine whether or not it makes sense to keep the indicator.

**Convergent validity - redundancy analysis**: Convergent validity is the extent to which a formatively measured construct correlates positively with other (reflective) measures ofthe same construct. Therefore the path coefficient between Y1 and Y2 is at least above 0.7. We can also use a single measure construct to evaluate the formatively measured construct, so as to not have to many questions, which we don’t use in the actual model.

(A problem doing so is that we might get people to answear a lot of question only for it to be used for validation. We risk that people get annoyed)

Et billede, der indeholder diagram, tekst, Font/skrifttype, cirkel

Automatisk genereret beskrivelse

**Multicollinarity**: Due to the underlying assumptions of the formatively measured constructs, we do not want high correlation, as this increases the standard errors, making it harder to find significant effect and understand the effect of each indicator. (predictions may however still be good)

* **Tolerance and VIF**
  + See slides 9 – where we make a regression of one indicator using the other. The tolerance is 1 – the r^2. We can't have a tol lower than 0,2 or 0,33 as this indicates that the other indicators explain another indicator.
  + Vif is 1/tolerance. It should therefore not be greater than 5. Helst under 3.

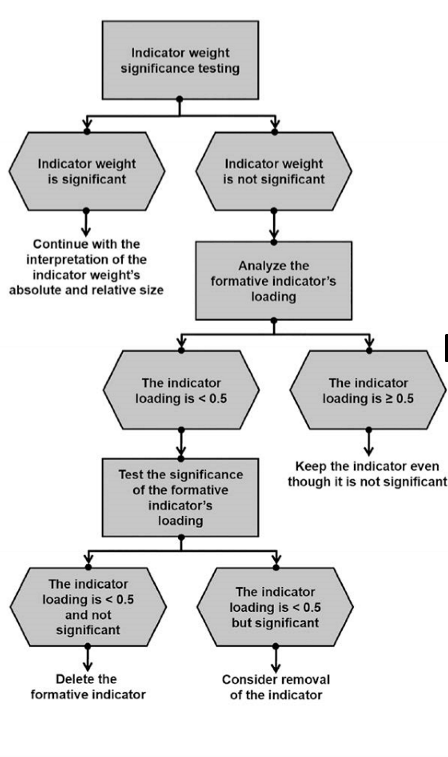
If the tolerance or VIF is too high, we should consider removing the indicator. But still be aware that the construct won't be as well defined.

If we cannot mitigate the collinearity problems, we cannot interpret the formative measurement model (but the constructs relationships with other constructs can still be analysed) – consider to dismiss the use of a formative measurement model.

In short we cant use the measurement model, but still look at path coefficients.

**Weights**. Are the contributions of each indicator to the construct.

* We simply evaluate this by determining whether or not the weight is significant
* If it is not significant, then look at slide 12 in 10. PLS.



### Assessing structural model:

Within the structural model we can't allow multicollinearity among our constructs. The VIF between constructs should not be higher than 3-5. However even if the VIF is high, we will only choose to delete a latent construct, if theory agree.

We should also inspect if the path are statistically significant. We do so by using bootstrapping.

Lastly we look at the the r-squared of each construct on others.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, dokument

Automatisk genereret beskrivelse

### Mediating and moderating effectsEt billede, der indeholder cirkel, diagram, design Automatisk genereret beskrivelse

**Mediating effects**: Is the effects from variable Y1 to Y3 that goes through the variable Y2. Se below to understand the direct, indirect, and total effects from Y1 to Y3.

A mediation occurs when a third variable, referred to as a mediator variable, intervenes between two related constructs. In the case above, Y2 is a mediator in the relationship between Y1 and Y3

* p12 \* p23 is the indirect of Y1 on Y3
* p13 is the direct effect of Y1 on Y3
* (p12 \* p23) + p13 is the total effect of Y1 on Y3

**The mediating analysis procedure**

**Et billede, der indeholder tekst, diagram, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse**

In this case **complementary** means that the effect are the same way e.g increase likelihood of being happy. **Competitive** is the other way around, that they have different effect on happines.

**Moderation :** is a variable that changes the direction and strength of relation between constructs. It is widely used in multigroup analysis e.g is their the same relationship between a set of constructs and happiness for both men or woman.

The moderating variable can also be continues, say that the effects between constructs and happiness is different on different income levels.

One can think of it as the interaction term which we also know from statistics.

**Two stage approach.**Et billede, der indeholder cirkel, skærmbillede, diagram, jackstik

Automatisk genereret beskrivelse

Here we first estimate a model without the interaction term.

Afterwards we create the interaction term by multiplying the latent variable scores of the exogenous and the moderator latent variable.  
  
Use all latent variables scores as single indicators and run the estimation again.

| Ark navn | Omhandler | Model navn i bogen + opgaver |
| --- | --- | --- |
| 8. PLS - a | Simpel reflective PLS model | Customer satisfaction model |
| 8. PLS - b | Simpel reflective PLS model med bootstrapped paths | Corporate reputation model  Page 71 |
| 9. PLS | Fuld refletive PLS model + samlet evaluering af hele modellen. | Corporate reputation model  Page 89 |
| 10. PLS a | Fuld model med formative og reflective measurements + evaluering af modellen. | Corporate reputation model  Page 111 |
| 10. PLS b | Assessing the structural model and path between construct + model comparison. | Corporate reputation model  Page 136 |
| 11. PLS | Full model formative and reflective + evaluation of model  Mediating and moderating effects | Corporate reputation model  Page 151 and 170 |

# CAUSAL DISCOVERY , PRODUCT RECOMMENDATION & TARGETING (I

## Bayesian networks

Using Bayesian networks, we want to find causal relationships. This is done by looking at variables and how setting one variable affect the probability of another one happening. In that sense we use the underlying thought of the bay’s theorem.

If we have data with respect to cancer and independent variables like exercise, food and smoking. We may want to see how setting smoking = 1 affects the overall probability of having or developing cancer later in life.

### Terminology

Et billede, der indeholder Font/skrifttype, tekst, typografi, håndskrift

Automatisk genereret beskrivelse

A – effect

* P (A) is the prior (unconditional probability), without taking into account any information about B. (overall probability of having cancer)
* P(A|B) is the posterior (conditional probability of A, given B). It takes into account information about B. ( probability of having cancer given that you are a smoker)

B – plausible cause or predictor

* P(B) is the marginal probability of the B.
* P(B|A) is the likelihood (conditional probability of B, given A).

### Characteristics

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, diagram

Automatisk genereret beskrivelse

The structure/DAG where:  
nodes = variables  
arcs = links = relationships

The strength of these relationships is defined by the Conditional Probability Tables attached to each node.

### Inference

When making inference, the analyst uses the BN to update the posterior probability in the light of new evidence. As mentioned before, we put in the information we know about some of the predictor variables.

Although not so evident at this point, this is one of the essence of BN: In the light of new evidence and conditioning on multiple variables, the use of BN facilitates the estimation of conditional probabilities in big networks

**Example**

Say we want to predict fraudulent use of credit cards. We already know that theifs are more likely to by jewelry and gas. Therefore the are predictor variable. We may however also know that middle age women are likely to buy jewelry. Our variable are therefore:

* AgeEt billede, der indeholder cirkel, skitse, tegning, diagram

  Automatisk genereret beskrivelse
* Sex
* Gas
* Jewelry
* Fraud (what we want to predict)

Define prior probabilities of the root nodes (Fraud, Age, Sex) and conditional probabilities of the children nodes (Gas, Jewlery) for every combination of the states of its parents.

Based on the probalilites, we can calculate the probability of an transaction be fraud given that we know, that we know the cardholder is a female older than 30 years.

### BN structure

Algorithms for classification

### 1.Naïve Bayes

Used to predict class membership probabilities based on the common aspects of each subject data. Meaning that based on some predictors, we want to classify people in to different groups.

Et billede, der indeholder diagram, linje/række, design

Automatisk genereret beskrivelse

## 2.Tree Augmented Naïve Bayes (TAN)

## Sames as Naive Bayes but allows relationships between features.

Et billede, der indeholder cirkel, skitse, diagram, linje/række

Automatisk genereret beskrivelse

## 3.Constrained-based algorithms

To identify the causal structures in the observational data, when causes are present in the data. To be sure there is causality we need to include at multivariate approach. As this insure that we have all the right predictor variables and model the right probabilies.

**INDEPENDENCE**

Often, it is preferred the characterization of independent variables that does not involve the density of X. That is:

Et billede, der indeholder Font/skrifttype, tekst, typografi, design

Automatisk genereret beskrivelse

Literally meaning that the marginal probability of Y does not change as a function of X.

**CONDITIONAL INDEPENDENCE**

X and Y are conditionally independent given Z, if for each value of Z, X and Y are independent. ****

**Example**

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse Et billede, der indeholder tekst, diagram, Teknisk tegning, Plan

Automatisk genereret beskrivelse

.1. BMI distribution does not change **neither by gender nor by year** => var are independent.

As seen below, they are all the same therefore, there is not relation between BMI and the year or gender.

Et billede, der indeholder tekst, Font/skrifttype, diagram, skærmbillede

Automatisk genereret beskrivelse

.2 If BMI distribution changes by gender but **does not change by year** given we control for gender.

Et billede, der indeholder tekst, diagram, skærmbillede, linje/række

Automatisk genereret beskrivelse

.3 If BMI distribution changes by year and **does not change by gender** given we control for year:

Et billede, der indeholder tekst, diagram, Font/skrifttype, linje/række

Automatisk genereret beskrivelse

## 4. Score-based algorithms

* All possible DAGs are given the same probability of occurrence
* Given the data, the DAG with the highest network score is chosen
* Several scores:
  + Bayesian Information Criterion (BIC)
  + Akaike Information Criterion (AIC)
  + Bayesian Dirichlet equivalent (BDe)

### Dag elements

Set of nodes, Directed edges or arcs, A path, a chain, a cycle.

Et billede, der indeholder tekst, diagram, skærmbillede, linje/række

Automatisk genereret beskrivelse

**Complete (saturated) graph :** if there is an arc between every pair of nodes

**Connected graph**: if there is a path between every pair of nodes

**Empty graph:** a graph with no arcs

Example, where X is the root node.  
Et billede, der indeholder diagram, skærmbillede, linje/række, tekst

Automatisk genereret beskrivelse

### Setting evidence to infer something

In other words, setting evidence changes the conditional probability of something happening. We may therefore

Et billede, der indeholder tekst, skærmbillede

Automatisk genereret beskrivelse

### SERIAL CONNECTIONEt billede, der indeholder skærmbillede, Font/skrifttype, diagram, tekst Automatisk genereret beskrivelse

Et billede, der indeholder skærmbillede, diagram, design

Automatisk genereret beskrivelse

In Figure 1a, if we get evidence for C, that changes the probability of B, which in turn changes the probability of A

Yet, if we set evidence for B (Figure 1b), then any change to C have no effect on A. Nor can any changes to A affect C. We say that the certainty of B blocks any dependence formerly shared between A and C. To be precise, there may be dependencies introduced by other relationships in the net, but not via B.

### DISCOUNTING

Et billede, der indeholder tekst, skærmbillede, diagram, linje/række

Automatisk genereret beskrivelse

The principle of "explaining away" works as follows:

- You start with evidence in the child node (C).

- This evidence updates the probabilities of the parent nodes (A) and (B).

- Additional evidence about one parent node (A) will then adjust the probability of the other parent node (B), and vice versa.

1. \*\*Evidence in \(C\):\*\* A patient with a cough (evidence for \(C\)) increases the probabilities of both the flu (\(A\)) and COVID-19 (\(B\)).

2. \*\*New Information about (B):\*\* If we get a test result showing the patient has COVID-19 (B):

- Knowing the patient has COVID-19, (P(A|C, B)\) decreases because the cough is now explained by COVID-19, reducing the need to attribute the cough to the flu.

### DIVERGENT (COMMON CAUSE)Et billede, der indeholder diagram, skærmbillede, tekst, design Automatisk genereret beskrivelse

In Figure 1a, if we get evidence for A, this

increases the chances of B which in turn

increases the chances of C

Yet, in Figure 1b, if we set evidence for B, then

any changes to A have no effect on C (and by

symmetry, no changes in C can affect A).

Again, to be precise, there may be

dependencies introduced by other

relationships in the net, but not via B.

Given the value of ( B), the child nodes ( C) and (A) are conditionally independent. This means that if you know the value of (B), learning the value of (B) does not provide any additional information about (C), and vice versa.

### D – SEPARATION CONCEPT

We have 3 criterions to determine whether or not a set X of variable are independent of another set Y. Given that we control for another variable Z.

**EXAMPLE 1**

Which pairs of variables are independent in the graphical model below, given that none of them have been observed (no evidence)?Et billede, der indeholder cirkel, skærmbillede, diagram, linje/række

Automatisk genereret beskrivelse

Answer:

If no evidence set, cf. principle 1,

A, B are independent as there are no

active trails between them.

**Example 2**

Now assume that the value of E is known (we set evidence for E). Which pairs of variables (not including E) are independent in the same graphical model, given E?  
Et billede, der indeholder cirkel, ur, design

Automatisk genereret beskrivelse

Answer: There are no pairs of variables that

are independent. Observing E, activates

the V-structures around C and E, giving rise

to active trails between every pair of

variables in the network.

### MARKOV BLANKET

* The Markov Blanket of the node Xi contains all the nodes that, if we know their states ( i.e. we have evidence for these nodes), it will isolate the node Xi from the rest of the network ( i.e. it will make Xi independent of all the other nodes given its Markov Blanket).
* It is the set of nodes that includes all the knowledge needed to do inference on the current node.
* Learning a Markov Blanket is particularly helpful when there is a large number of variables to select from in a dataset. It can also serve as a highly-efficient variable selection method in preparation for other types of modeling e.g. Regression, Neural Nets, etc.
* The Markov blanket of a node in a Bayesian network is a concept used to identify the subset of nodes that renders the node conditionally independent of the rest of the network. In other words, once you know the values of the nodes in the Markov blanket of a node \(X\), knowing the values of any other nodes in the network provides no additional information about \(X\).

As seen below, the Markov blanket is fulfilled for X9, given that we have information on X1, X2, X10 and X7. This is useful in prediction, meaning that if we have knowledge on these variables, we can predict X9 without consideration on other variable.

Et billede, der indeholder cirkel, diagram, design

Automatisk genereret beskrivelse

### MARKOV PROPERTY (conditional independent)

Et billede, der indeholder tekst, skærmbillede, linje/række, Font/skrifttype

Automatisk genereret beskrivelse

Se at vi bare siger at variablen er conditional independet af sine nodedecendents givet dens parents.

* Technically, the Markov condition is guaranteed by learning the DAG from the data. The algorithms will construct a directed graph by catching the conditional independencies and drawing edges between variables pairs that are not conditionally independent.
* However, if we had reason to believe that there is a hidden cause and this was not recorded in the dataset, and consequently not included in the graph, the Markov property will not hold.
* In that case, we cannot claim it is a causal BN network, although we can still use it as a predictor tool. See examples next.

Med andre ord det vigtigt at vi antager at vi har modelleret af vigtige variable eller i det mindre kontrolleret for dem. Har vi ikke kontrolleret for dem er C ikke uafhængig af G. Dog vil det være tilfældet hvis vi kontrollerer for A.

Et billede, der indeholder skærmbillede, tekst, cirkel, diagram

Automatisk genereret beskrivelse

**Markov Property**: Primarily used to describe the structure of conditional dependencies in a stochastic process or a Bayesian network. It simplifies the computation of the joint probability distribution by leveraging conditional independence.

**Markov Blanket**: Used in inference and learning within Bayesian networks. It helps identify the minimal set of nodes required to make a node conditionally independent from the rest of the network, which is useful for simplifying computations, making predictions, and feature selection in machine learning.

**CONCLUSION**

In a causal BN, the Markov property is satisfied given that **all common causes are represented** in the graph (i.e. there are no hidden common causes acting as confounding factors).   
Et billede, der indeholder linje/række, halskæde, kunst

Automatisk genereret beskrivelse

In our Season example, we can say there

is no way for Rain to influence Slippery

except by way of causing Wet or not.

Thus we can assume there is no hidden

variable connecting Rain and Slippery.

Every independency suggested by the

lack of an arrow between Rain and

Slippery is actually real in the system.

## Usecases for Baysiean network

### Customer retention

It can be expensive and hard to acquire new customers. Therefore many companies focus on retaining their existing customers. To do so they need to understand, what makes a customer likely to do a reaped purchase. This can be modelled using SEM (their purchase intention), both also through Bayesian networks.

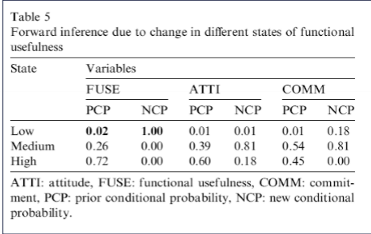
#### Foreward inference

Forward inference is a process used in Bayesian networks to compute the probability of certain variables given evidence or observations. It involves propagating information forward through the network, starting from the observed variables and moving towards the variables of interest.

Here's how it works:

1. Start with the variables that have been observed or are known.
2. Use the conditional probability distributions encoded in the network to calculate the probabilities of their immediate descendants (variables directly influenced by them).
3. Continue this process iteratively, moving through the network, updating the probabilities of each variable based on the probabilities of its parents until reaching the variables of interest.
4. Finally, combine the updated probabilities to obtain the probability distribution of the variables of interest given the observed evidence.

In essence, forward inference allows us to make probabilistic predictions or inferences about unobserved variables based on the known information in the Bayesian network.



Say we have information on a customers shopping pattern. Based on this, we can calculate the conditional probability that she will have a repead purchase.

#### 

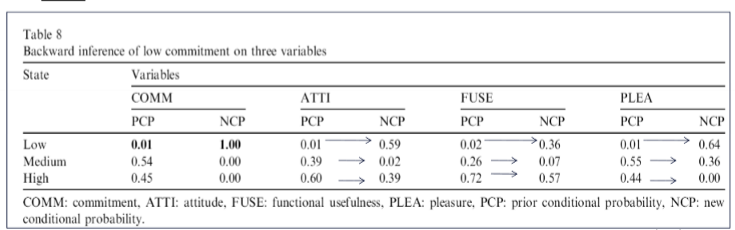
#### Backwards inference

Backward inference is the opposite of forward inference in the context of Bayesian networks. While forward inference starts from observed variables and propagates information forward through the network to infer the probabilities of unobserved variables, backward inference works the other way around.

In backward inference:

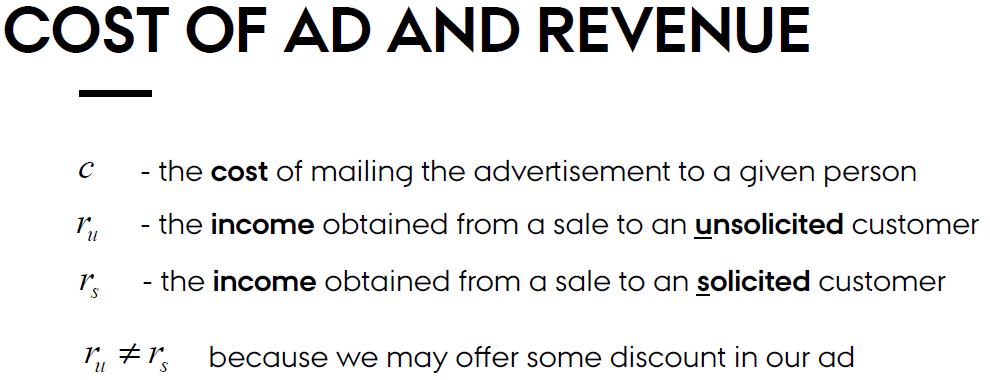
1. Start with the variables of interest, which are typically unobserved.
2. Use the conditional probability distributions encoded in the network to calculate the probabilities of their immediate ancestors (variables that directly influence them).
3. Continue this process iteratively, moving backward through the network, updating the probabilities of each variable based on the probabilities of its children until reaching the observed variables.
4. Finally, combine the updated probabilities to obtain the probability distribution of the variables of interest given the observed evidence.

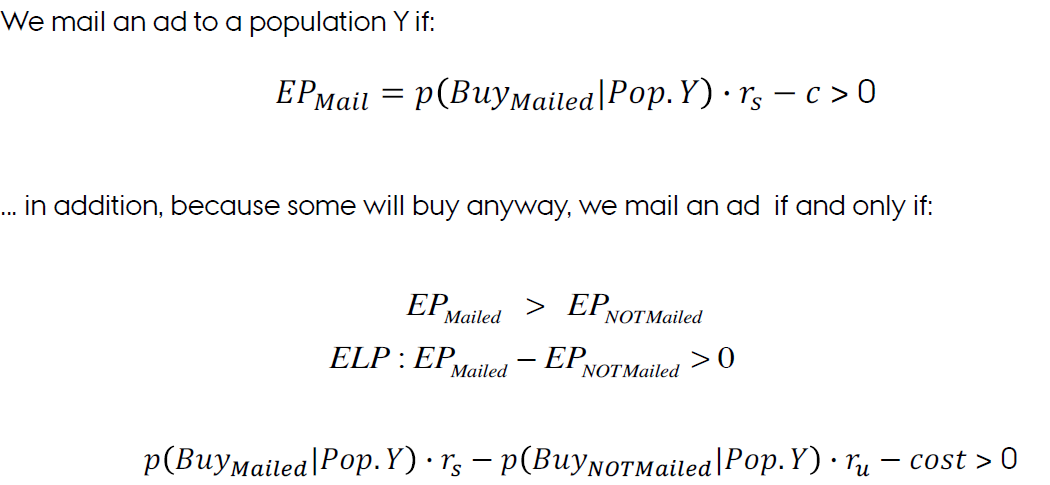
In essence, backward inference allows us to infer the probabilities of unobserved variables by working backward through the Bayesian network from the variables of interest to the observed variables.



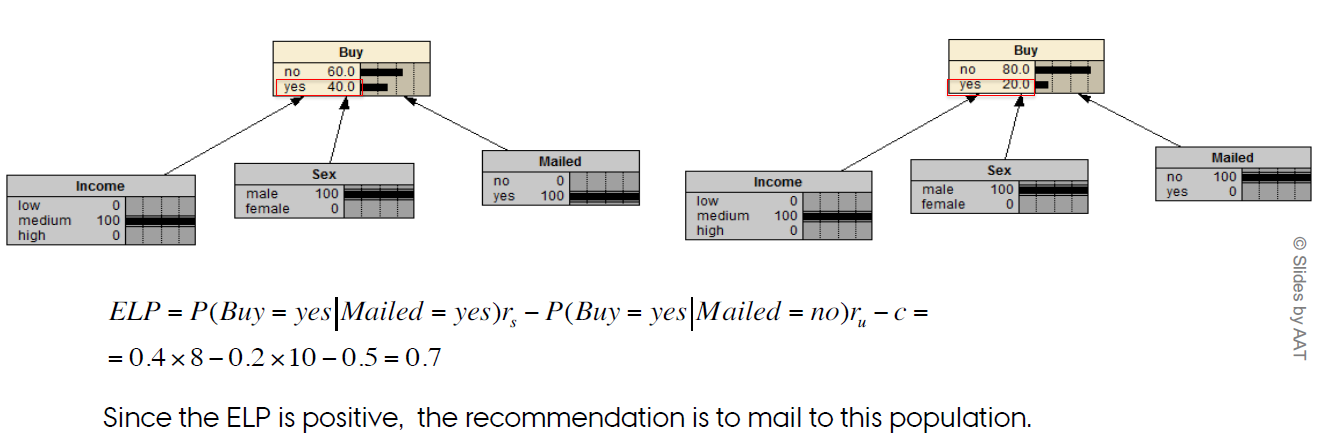
### Targeted advertising

We want to make targeted advertising, which makes it more likely that they will purchase the product. We determine whether or not to advertise if the ELP (expected lift in profit) is positive or above some threshold.





Eksempel:



### c. Product recommendation

**Collaborative Filtering =** the process of recommending items of interest to an individual, based on his/her interests or the interests of similar individuals.

**Explicit voting:**

It learns individual´s preferences from individuals’ **reported preferences** by asking the individuals explicitly to rank the items on some scaleEt billede, der indeholder tekst, skærmbillede, nummer/tal, Font/skrifttype

Automatisk genereret beskrivelse

A data set with individuals ranking 4 products (X, Y, Z, W) based on their preferences on a scale from 1 to 5.

Say that we get a new customer Joe, who have the preferences as seen to the right. A way to determine whether Joe will like item W is to compare the rest of his preferences to that of other customers.Et billede, der indeholder skærmbillede, tekst, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

As one can tell, we need a lot of data on other customers and their explicit preferences. We do also need a lot of data one Joe, this may be hard if he is not a frequent customer.

**Implicit voting:** Et billede, der indeholder tekst, skærmbillede, display/skærm/fremvisning, nummer/tal

Automatisk genereret beskrivelse

It learns individual’s preferences from individuals’ past **behaviour.** Here we will only be able to predict based on customers previous purchases.

The problem here is also that we may not have a lot of data if Joe is a new customer.

Furthermore, just because a person bought the product, does not mean that they liked it.

**Memory based algorithms (Classical machine learning)**

These methods find users that are similar to the active user (i.e. the user we want to make predictions for), and uses their preferences to predict ratings for the active user, by using the entire dataset.

**• Common similarity measures:**

**Pearson corr**.: how much two users vary together.  
We find the correlation between the user of interest and all other users

**Et billede, der indeholder tekst, skærmbillede, nummer/tal, Font/skrifttype

Automatisk genereret beskrivelse**

**Advantages**

* The quality of predictions are rather good.
* This is a relatively simple algorithm to implement for any situation.

**Disadvantages**

* It uses the entire database every time it makes a prediction, so it needs to be in memory => extremely slow.
* It can sometimes not make a prediction for certain active users/items. This can occur if the active user has no items in common with all people who have rated the target item.
* Overfitting the data: it takes all random variability in people's ratings as causation,which can be a real problem.

**Distance measures**: Manhattan distance, Euclidian distance, etc.

**Vector similarity:** we can treat two users as vectors and take the cosine of the angle btw. two vectors

**MODEL-BASED ALG.**

Based only on selecting **a portion of the existing users/items** and use that as a "model" to make recommendations without having to use the complete dataset every time.

* Adv: speed and scalability

**Three possible approaches:**

1. Enhancement of memory-based alg.: calculate similarities using only k-most similar users or items – these models were seen in the previous lectures with HJJ

2. As a linear algebra problem: use linear equations

3. As a probability model: Bayesian nets – **discussed here**

When using Bayesian networks we first calculate the prior probability of Joe ratting product W either 1, 2, 3, 4, 5. After having the prior probability we can calculated the conditional probability, given that Joe have rated the other product (1,5,5).

We then take the conditional probability and multiply by each scores. This gives us and estimate of Joe’s rating of product W. As one can see below, we believe that Joe’s score will be 4,23 implying that Joe would like the product.

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, håndskrift

Automatisk genereret beskrivelse

As seen below, the Bayesian network without plotting Joe’s preferences just shows the average score of each product. However when putting in the Joe’s preferences we get the updated probability for each scores. When multiplying the scores and probability, we get an estimate for Joe’s scores.

Et billede, der indeholder tekst, diagram, skærmbillede, Rektangel

Automatisk genereret beskrivelseEt billede, der indeholder tekst, diagram, Rektangel, skærmbillede

Automatisk genereret beskrivelse

**Clustering**

We may also assume that each customer can be divieded into a specific cluster of customers. This class/cluster of customer, will determine the score we give Joe.

However this also require a lot of data, as we create a subset of clusters. As seen below, we get the mean score in V4, for all customers in class 2, with a rating of (1,5,5).

## Clickpath analysis and choice modelling

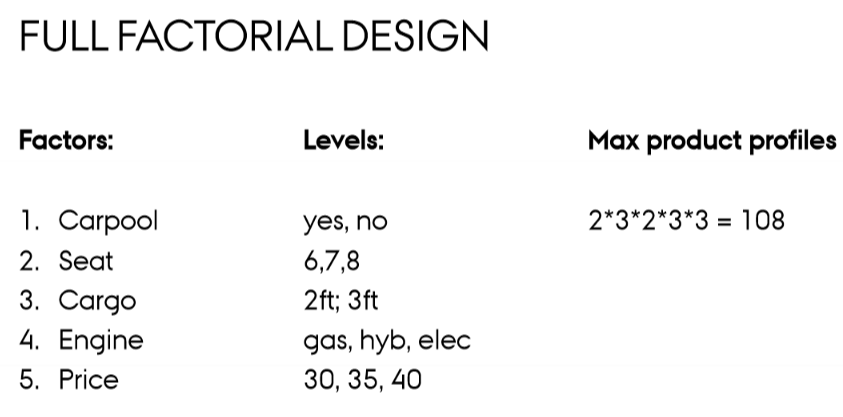
As we have not really done any clickpath analysis i dont assume that this will be part of the exam. In short we just look at how people click through our website and try to determine what will be the customers next step. For more intuition please see the slides.

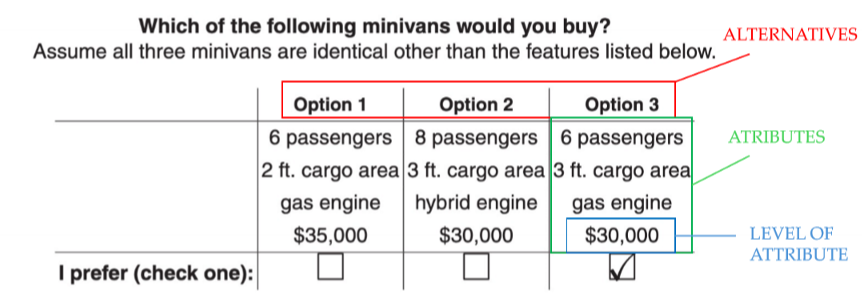
### Choice analysis.

We use choice analysis to help us better the right combination of features for our product. In the example below, we for instance look at what specifications a car should have to ensure that most customers will like it.

We do so by providing customers with a set of cars with different specifications and ask them to pick the one they prefer. We only do so for a subset of combinations - based on this we determine what features our customers like the most. This is a partial factorial design, if we used all it would be a full factorial design.

Choice based conjoint analysis is when the individual choose between different products meaning that we have a categorical dependent variable. There is also a rating based conjoint analysis, where the individual give a rating to each product, this emplying that the dependent variable i metric. We have however only worked with the first one.





### interpreting results:

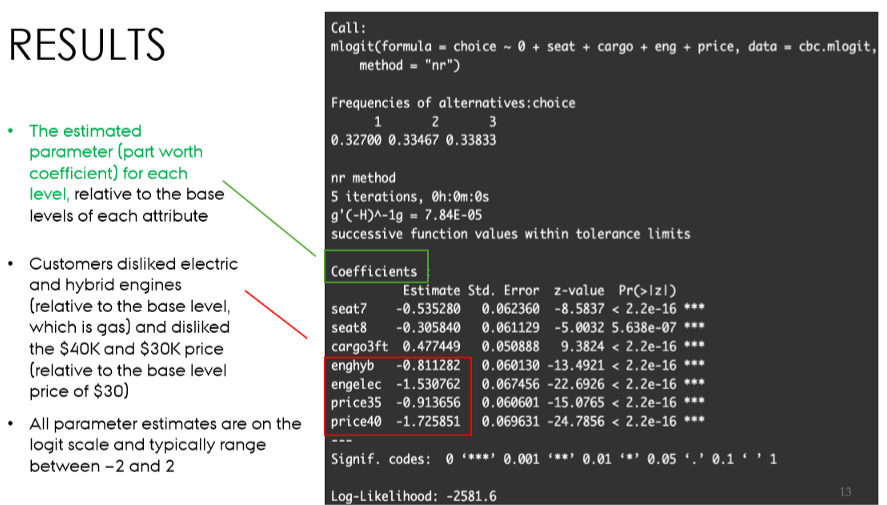
Vores baseline er den første observation i hver klasse - i dette tilfælde en bil med følgende specifikationer: sæder = 6, Engine = gas, Cargo = 2ft og Price = 30.

Er koefficienten positiv, vil folk hellere have denne option, er den omvendt negativ vil de i lavere grad have denne option. Altså vil folk mindst have en elektrisk bil til 40.000 (husk at tjekke signifikans!)

LOGIT KAN ÆNDRES TIL WILLINGNESS TO PAY.

SÅ VI KAN UDREGNE HVORNÅR FOLK ER INDIFFERENTE.

Vi kan kun lave willingness to pay når price er numerisk.



# Segmentation:

The general goal of market segmentation is to find groups of customers that differ in important ways associated with product interest, market participation, or response to marketing efforts. By understanding the differences among groups, a marketer can make better strategic choices about opportunities, product definition, and positioning, and can engage in more effective promotion. Of course, the segmentation would have to be implemented and used. Furthermore, as all business initiatives, the increase in e.g. sales should be greater than the cost of making the segmentation.

## Segmentation as Clustering and Classification

Segmentation can be clustering, which is an unsupervised machine learning method. Here we feed the data to the algorithm and ask it to find groups, which have characteristics that are different from others. In most cases clustering is about identifying groups in data.

Classification is a supervised segmentation approach, where beforehand we before hand have an understanding of groups within our data and want to predict whether a new observation belongs to a certain group.

As we are trying to find groups which are different one should eliminate variables which do not vary at all, as well as those who do not vary across groups. We should also delete variable which are not directly related to the task, as they can cause noise and spurious correlations.

### Focus on clustering methods.

### Two main distinctions:

**Distance based methods** like “ hclust “ and “k-means” try to attempt to find groups that minimize the distance between members within the group, while maximizing the distance of members from other groups.

* hclust() does this by modeling the data in a tree structure
* kmeans() uses group centroids (central points). (Numeric)

**model-based clustering** methods, Mclust() and poLCA() attempt to find groups by modeling the data such that the observed variance can be best represented by a small number of groups with specific distribution characteristics such as different means and standard deviations.

* Mclust() models the data as a mixture of Gaussian (normal) variables (numric)
* poLCA() uses a latent class model with categorical (nominal) variables. (categorial)

### Hierarchical Clustering: hclust()

Hierarchical clustering is a popular method that groups observations according to their similarity. This is a **distance-based** algorithm that operates on a dissimilarity matrix, an N-by-N matrix that reports a metric for the distance between each pair of observations.

Each observation is at first its own cluster, but is then merged together with other observations, which are alike(measured by Euclidean distance). This will slowly make larger clusters. However as we are measuring distance **we may only work with numeric data.**

**Evaluation:**

CPCC is interpreted similarly to Pearson’s r. In this case, CPCC > 0.7 indicates a relatively strong fit, meaning that the hierarchical tree represents the distances between customers well.

**Interpretation:**

To interpret the dendrogram, we specify the number of cluster we want to withdraw. Specifying 4 groups as seen below, we get that we cut the dendrogram around 0,55.Et billede, der indeholder diagram, skitse, Teknisk tegning, tegning

Automatisk genereret beskrivelse

### Mean-Based Clustering: kmeans()

K-means clustering attempts to find groups that are most compact, in terms of the mean sum-of-squares deviation of each observation from the multivariate center (centroid) of its assigned group. As we compute a distance, kmeans also rely on **Euclidean distance and is therefore only suited for numeric data.**

Here we specify the K and find distinct groups. However as seen in the plot below, all groups may not be interesting as group 3 and 4 are clearly overlapping. There is however more difference between group 1 and 2, which may help us target advertisement.Et billede, der indeholder tekst, diagram, skærmbillede

Automatisk genereret beskrivelse

### Model-Based Clustering: Mclust()

The key idea for model-based clustering is that observations come from groups with different statistical distributions (such as different means and variances). The algorithms try to find the best set of such underlying distributions to explain the observed data. **We believe that the underlying distribution is normal (gaussian).** Because mclust models data with normal distributions**, it uses only numeric data.**

### Latent Class Analysis: poLCA()

Latent class analysis (LCA) is similar to mixture modeling in the assumption that differences are attributable to unobserved groups that one wishes to uncover. **poLCA uses only categorical variables**. Using LCA we also set the number of groups, that we want the software to identify.

### Comparing Models with BIC() and log likelyhood

The key point to interpreting BIC is to remember this: the lower the value of BIC, on an infinite number line, the better. BIC of −1000 is better than BIC of −990; and BIC of 60 is better than BIC of 90.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

Alternatively, you could also check the log-likelihood values, where higher log-likelihood values are better (e.g., −1000 is better than −1100).

We may also compare model, by how they are similar. In other words, do the different clusters found using different methods agree. Is group “2” identified with kmeans also present in LCA ?  
See the script to understand this code.

### Outcome

Doing clustering we want to isolate subgroups within a population, there are similar within the groups, but significantly different from the other groups. Looking at the simple toy example below, it should be evident that there are 3 subgroups with respect to education and income. We would like the algorithm to identify 3 cluster (S1,S2), (S3,S4) and (S5,S6).

Et billede, der indeholder tekst, skærmbillede, diagram, linje/række

Automatisk genereret beskrivelse

### 

### Good practice

As we are trying to identify differences between groups, we should be careful with variable selection. It therefore wont make sense to include variable which do not differ between groups.

Furthermore we should be selective in variable selection and not include variable which do not relate to the specific objectives of the cluster analysis as it may create noise, which the algorithm have to manage through.

### Research design

**Sample size:**

* As (classical) cluster analysis isn’t a statistical inference technique sample size considerations are solely related to the possibility of representing the structure and small groups in the population
* Thus, given the objective of the analysis the researcher must make sure that the sample is large enough to represent the desired groups

**Outlier detection**

* An object which doesn’t fit into a given pattern may represent: a true outlier, a small but insignificant segment of the population or an actual and relevant group which is underrepresented in the sample
* In the first two cases deletion of the object is the appropriate action. However, in the latter case this could mean disaster
* A profile diagram/snake plot is a graphical approach to detect outliers. However, this approach isn’t suitable for large samples
* Another approach is to calculate a measure of deviance from the average observation and then rank the objects according to this measure

### Assumptions

For the classical or distance based clustering there are no assumptions, as it is mostly based on heuristics. We may however still be aware:

* That we have a somewhat representative sample of a satisfying size
  + Say we know that 5 % of our customers are high spender. We want a sample size which ensure that they make up a substantial part of the sample.
* That multicollinearity is not present. If there is a large collinearity between predictor variables, they will implicit pose a higher weighting in determining clusters.
  + E.g that we include income before and after tax to determine spending.
  + If collinearity is high, we may delete one or more variables.

### Measures of similarity

In most cases we will work with data the is interval scaled. Having data measured on an interval scale we can determine similarity using either distance measures or correlation measures.

When using the correlation, we look at the Pearson correlation. However there is a downside to using the correlation, this being that it has a to high focus on patterns and shapes, more than on forming useful clusters. From the illustration below, we would wrongly believe that observation 1 and 2 are closely related because they follow the same pattern (High correlation). When in reality it is only observation 1 and 3 that er closely related.

Et billede, der indeholder tekst, diagram, Font/skrifttype, Kurve

Automatisk genereret beskrivelse

**Different distances:**

**Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse**

### Standardization

When dealing with variable, which are not measured on the same scale. We are at the risk that variable with higher scale/variance will have a larger impact on the solution.

We therefore inspect the standard deviation of all variables, if they are not somewhat the same or we have reason as to why one should be more influential, we standardize our predictors.

### Hierarchical methods

The hierarchical methods comprise the agglomerative and divisive methods.

▶ In the **agglomerative methods** all objects start out as their own cluster and objects are the successively joined

▶ In the **divisive methods** all objects start out as a single cluster and objects are then successively divided - these methods are seldom used

The main difference between the algorithms is how the distance is measured and how the algorithm starts out.

### Linkage

**These methods only depend on the ordinal properties of the distances**

**Single-linkage/nearest-neighbor** measures the distance between clusters as the minimum of the distance between all possible pairs of objects in the two clusters – tends to produce unbalanced and “straggly” clusters

**Complete-linkage/farthest-neighbor** measures the distance between clusters as the maximum of the distance between all possible pairs of objects in the two clusters – tends to find compact clusters with equal diameters

**These methods depend on metrical properties of the data matrix and hence the distances**

**Average-linkage** measures the distance between clusters as the average distance between all possible pairs of objects in the two clusters

**Centroid measures** the distance between clusters by forming average objects for each cluster and then measuring the distance between these average objects  
Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

**Ward’s method** forms clusters by minimizing the within-cluster sums of squares – tends to find same-size clusters. This be looking at the lowest sum of squares by combining each possible combination of observartion.

It is always recommended to try different solution with different algorithms to insure that we have a robust fit. We do however for the most part use complete linkage or wards methods.

And remember, a useful cluster solution is one that provides clusters that provide insights.

### Example Euclidean distance algorithms + single linkage:

In the example below, the Euclidean distance is measured for 6 objects. In this case we can see that the lowest distance is between observation S1 and S2, as well as S3 and S4. In both cases the distance is 2. Therefore or first combination will be either (S1,S2) or (S3,S4).

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, nummer/tal

Automatisk genereret beskrivelse

Having combined (S1,S2) we recalculate the distance matrix and get the results below. As we can tell the observations with the lowest distance er are (S3,S4).

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, nummer/tal

Automatisk genereret beskrivelse

Having merged (S3,S4) together, we recalculate the similarity matrix. As we can see below, the observations with the lowest distance are now observation (S5,S6)

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, nummer/tal

Automatisk genereret beskrivelse

As seen below, we now have 3 clusters, where the difference between the others are quite large (above 100). We could choose to stop here based on prior knowledge.

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, nummer/tal

Automatisk genereret beskrivelse

### Nonhierarchical methods

The general purpose of nonhierarchical clustering is to classify objects into *k* clusters such that

1. The objects within the same cluster are as similar as possible – **high intra-class similarity**

2. The objects from different clusters are as dissimilar as possible – **low inter-class similarity**

The main difference between nonhierarchical and hierarchical clustering, is that the number of cluster = K, must be defined before starting the clustering process.

Once an object has been assigned a particular cluster it can later in the process be assigned a different cluster - thus the treelike construction process doesn’t apply

**For a k-cluster analysis the nonhierarchical method comprises the following steps**Et billede, der indeholder blomst, mønster

Automatisk genereret beskrivelse

1. Select k initial cluster seeds (centroids)

2. Assign each of the observations to the nearest cluster

3. Reassign each object to one of the k clusters according to a stopping rule

4. Stop if there is no reassignment

This description also covers situations where no reassignment is done.

As one can tell from the picture to the left. The centroid start by random, then all observation er assigned to a cluster. Afterward we calculate the new centroid and assign one more, until we don’t get a new solution any more.

**There are a number of possible ways to select the initial cluster centroids (cluster seeds)**

▶ The researcher supplies the seeds - this may be relevant if prior research has already defined segment profiles and the purpose of clustering is solely allocation

▶ Select the first k objects with nonmissing data

▶ Select the first seed as the first object with nonmissing data. The second seed is the next observation with nonmissing data that is separated from the first seed by a specified distance

▶ Randomly select k nonmissing objects

### Comparing hierarchical and nonhierarchical methods

It is recommended that a combination of the hierarchical an nonhierarchical methods is used

▶ Use the hierarchical methods to get a qualified estimate on the number of clusters

▶ Use the nonhierarchical methods to determine the cluster affiliation for each object

### Deciding on the number of clusters

▶ The decision on the number of clusters is a key decision in connection with the hierarchical methods and it is a part of the input in the nonhierarchical procedures

▶ The researcher is advised to use the hierarchical procedures to come up with a limited number of possible cluster sizes and then utilize these cluster sizes in the nonhierarchical analysis

▶ No standard objective selection procedure exists which is why the researcher is advised to consider several possible solutions

▶ A number of ad hoc rules have been developed, often these rules are tied to a particular software package

▶ The measures can either be categorized as measures of heterogeneity change or direct measures of heterogeneity

### Dendograms

▶ A key tool for determining the number of clusters in association with hierarchical clustering is the dendrogram

▶ It is a tree-like figure, where one dimension holds the distances between observations/clusters and the other dimension identifies the observations

▶ A tendency for a group of branches to come together at the same point and then not be involved in further amalgamations for a while indicates a cluster

▶ We will look at “big jumps in the dendrogram” and the percentage increase in agglomeration coefficients to inform on the number of clusters

### Profiling

In short profiling is looking at the different clusters, and the variable that define them. Eg. That customers have a high income, is a male and so on. Having a clear profile for the different clusters is needed to turn knowledge in to marketing actionpoints.

▶ The profiling is carried out by looking at the centroids of the cluster

▶ The idea is to look for characteristics on one or several of the clustering variables that identify a cluster

▶ In this way one can name each of the clusters based on the information just identified

▶ The profiling may also be helpful in choosing from different potential cluster solutions

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

### Validation and description

The purpose of the validation is to ensure that results are representative of the population and thus generalizable. Meaning that our solution it not just relevant for our single sample.

▶ One possible technique is to divide the sample into subsamples and assess the relationship between the solutions

▶ Another possibility is to use a set of different variables known to vary across clusters and then test for differences – criterion validity. In that sense, we would expect variable alike those used in the original solution, to vary in unisen with our clusters.

Description

▶ This stage involves the use variables not previously used in the analysis

▶ The idea is to describe characteristics of the clusters, typically in form of demographics, psychographic profiles, consumption etc.

Distance-based algorithms are generally assessed based on functionals

of the between- and within-cluster sum of squares – a prime example

being the Calinski-Harabas measure:

Et billede, der indeholder Font/skrifttype, tekst, linje/række, hvid

Automatisk genereret beskrivelse

where SSB and SSW are sum of squares between and within segments, respectively – a bigger number signify a better solution

### Key lessons from hierarchical and non-hierarchical cluster analysis

There is substantial ambiguity associated with

* The number of clusters
* The measurement of proximity of individuals
* The measurement of proximity of groups

There is no single index to compare different cluster solutions. Existing validation techniques depend on data and/or the cluster algorithm

Both Hierarchical and non- Hierarchical approaches are heuristic procedures. There are no assumptions. Choice of method an number of cluster may be determined by previous research or by comparing different methods.

## Segmentation II - Model-based clustering

**my words:**

The essence of model-based clustering is that we apply some assumptions to the underlying data. This being that the data is e.g multivariate normal density in higher dimensions.

Having this assumption, our approach is much more like classical statistics. This means that we also can get actual measures like BIC to evaluate this fit of different functions.

Because we make assumptions to the underlying distribution, we can get statistical measurements, which should help us evaluate the models. In that sense it becomes more of a science, when choosing a model.

This is the main difference between model-based models and distance-based models, which primarily works by applying some heuristic. In that sense the model-based approach may be better if the assumptions are meet. However, if they are not meet, chosen between distance and model-based clustering I primarily choosing on preference.

**Mortens words:**

▶ Model-based clustering assumes that the population is made up of several distinct subsets/clusters, each governed by a different multivariate probability density function

▶ The parameters associated with the model can be used to assign each observation a posterior probability of belonging to a cluster

▶ The problems of identifying the number of clusters and selecting the clustering method boil down to a model selection problem – for which we have a number of procedures

▶ Furthermore, being based on a genuine statistical model, model-based clustering readily accommodates missing data in a way similar to the state-of-the-art ML methods

▶ Are also known as latent-class cluster analysis or finite mixture modeling

▶ K-means clustering are approximate estimation methods for certain finite mixture probability models lending credibility to these

▶ They allow for an integral representation of the cluster model together with predictor variables such as demographics

### The formal model

Observed data, x are assumed to originate from a mixture of probability density functions like this:

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, algebra

Automatisk genereret beskrivelse

Once the parameters of the model have been estimated each observation can be assigned to a cluster using estimated posterior probabilities- We then calculate the probability that each observation belongs to each cluster. Then we assign it to the cluster with the highest probability, kind of like what we did in LDA.

### Determining the number of clusters (cont’d)

▶ In the literature two alternatives have been suggested **– bootstrapped LRT and Lo-Mendell-Rubin LRT**

▶ Both suggestions perform well in simulations with a slight advantage to the bootstrapped LRT

▶ Comparisons using information criteria allow for the possibility of comparing a set of models

▶ The criteria most often used is the **BIC criteria**

▶ The assumptions justifying the use of information criteria are various regularity conditions which in many situations are not fulfilled

▶ However, there is ample theoretical and empirical support for the use of these criteria in connection with model-based clustering

▶ The BIC criteria is made up of two opposite contributions – a large value of the maximized likelihood function and a penalty based on the number of parameters as well as the number of observations

▶ As such we are looking for large BIC values but since the likelihood function is negative this corresponds to **numerically small BIC values**

## Segmentation III - Finding, assessing, and predicting customer segments

This lecture will help you to understand

▶ The idea behind segmentation

▶ Distinguish between natural, reproducible and constructive segment structure

▶ Global criteria to assess segmentation solutions including reproducibility

▶ Incremental/segment level criteria to assess segmentation solutions

### 

### What is segmentation and why do it

According to Kotler and Armstrong (2006) the aim of market segmentation is to

*“divide a market into smaller groups of buyers with distinct needs, characteristics or behaviors who might require separate products or marketing mixes”*

▶ Segmentation is often done utilizing cluster analysis as the preferred tool

▶ A good segmentation strategy can lead to competitive advantages but obviously, the quality of the strategy depends on the quality of the segmentation solution

▶ The quality of a segmentation solution needs to consider uncertainty originating from the fact that the analysis is based on a single (typically random) sample as well as the fact that many clustering algorithms are stochastic

### Approaches to segmentation

It can either be **common sense segmentation:**

Here we segment based on only one variable – like age or profitability.

We can also choose a **data-driven segmentation approach:**

Here segments are made using several segmentation variables. This of course make the characterization more difficult. However they may also provide more insight as we have more variable(information) about the different segments.

**The following steps make up a data-driven segmentation**

▶ Decide which variables to use as segmentation variables

▶ Collect data

▶ Extract segments – this should involve a range of number of segments

▶ Select the best performing solution

▶ Describe the segments in this solution in terms of the segmentation variables as well as other descriptive variables

▶ Select the most optimal target segment(s)

However a key issue with this approach is the fact that the best solution is assessed via global measures and hence might miss more attractive individual segments

### General quality criteria to assess segmentation solutions

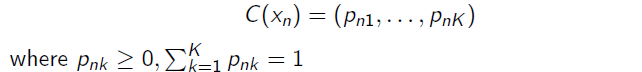
▶ Measurability – size, purchasing power and demographic profiles of the segments must be easy enough to measure  
  
▶ Accessibility – the company must be able to reach the market segmentseffectively  
  
▶ Substantiality – the segments must be large and proftable enough  
  
▶ Differentiability – the segments should be conceptually distinguishable and should respond differently to the marketing mix elements  
  
▶ Actionability – it must be possible to design effective programmes for attracting the segments

### Measures of reproducibility/stability

▶ Reproducibility can be assessed with respect to replications of the sample and of the algorithm – the focus is on the former

▶ For a given sample, χN, a partition C(·) = C(·|χN) is a random variable depending on the algorithm and the sample

▶ Specifically, for a K-segment solution each observation, xn, is assigned a vector:



▶ For classical partitioning algorithms, exactly one pnk = 1 for each n

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

▶ For two partitions,*C*1(・)*, C*2(・), we can observe one of these four outcomes for two consumers

a. Both consumers are assigned to the same segment twice

b. The two consumers are in the same segment in *C*1(・) but not in *C*2(・)

c. The two consumers are in the same segment in *C*2(・) but not in *C*1(・)

d. The two consumers are assigned to different segments twice

▶ For *n* consumers there are *n*(*n*−1)*/*2 possible pairs so we let *a, b, c, d* denote the number of pairs in each category (a+b+c+d=n(n-1)/2)

▶ The Rand index is defined as

Et billede, der indeholder Font/skrifttype, linje/række, diagram, nummer/tal

Automatisk genereret beskrivelse

In other word, we can evaluate to what extent is bootstrap sample classifies agrees on with segment to place the custuomer. This is also known as the **Rand Index.**

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, linje/række

Automatisk genereret beskrivelse

In general we would like the index to be close to one as this suggests that all different bootstrap samples agree as to which cluster we should put the observation.

### Gorge plot

Is a plot that shows the distance between consumer I and the centroid of the cluster h asEt billede, der indeholder Font/skrifttype, linje/række, tekst, hvid

Automatisk genereret beskrivelse

The Gorge plot is made for each segment to the stability. By construction similarities are between 0 (the consumer is far away from the centroid) and 1 (the consumer is close to the centroid) and add up to 1 over all segments. Therefor we expect to see a G-form in well defined clusters as we want observation either really close or far away from the centroid.

Et billede, der indeholder tekst, diagram, Plan

Automatisk genereret beskrivelse

### Segment level stability

▶ The stability assessments such as calculations of the adjusted Rand index is based on a comparison of solutions with the same number of segments

▶ Segment level stability assesses the behavior as additional segments are added to a solution

▶ The key benefit is that they allow the analyst to focus on finding one or a small number of good individual segments

#### Segment level stability across solutions (*SLSA*)

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

#### Segment level stability within solutions (SLSW)

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, algebra

Automatisk genereret beskrivelse

Segment level stability refers to the consistency or reliability of a particular segment within a dataset or system. "Across solutions" and "within solutions" denote two different perspectives for evaluating this stability:

1. **Across Solutions:**
   * This refers to comparing the stability of segments across different solutions or methods. For example, if you have multiple algorithms or models for segmenting data, you might assess how consistent the segmentation is across these different approaches. High stability across solutions suggests that the segmentation is robust and not heavily influenced by the specific method used.
2. **Within Solutions:**
   * Within a single solution or method, this refers to the stability of segments over time, iterations, or different subsets of data. It assesses whether the segmentation remains consistent when the data or conditions change. High stability within a solution implies that the segmentation is reliable and not overly sensitive to minor variations in the data or process.

In summary, across solutions compares stability between different segmentation methods, while within solutions evaluates stability within a single method across different conditions or iterations.

## Segmentation IV - Introduction to latent class analysis

## Latent class analysis.

In this section, we focus on understanding how different latent classes/segment affect how people act. This latent class models are focusing on categorical (binary) data.

We therefore assume that the latent variable are categorical (eg. Different segments) and our predictor variable er binary.

Lastly we assess whether the model can reproduce the original relationship. Measurement for evaluation is ….

### Latent class vs. cluster analysis

▶ Latent class analysis can also due to its second stage be seen as a form of cluster analysis

▶ However, the latent class model is based on a probability model

▶ Cluster analysis is rooted in the similarities between rows of the data matrix

▶ Latent class analysis is based on the probabilities of the elements in the rows

▶ Seeking to form clusters of similar probabilities, the latent class analysis focuses on rows of similar expectations

▶ This is accomplished via the assumption of local independence

In my own words:

We can use latent class analysis as a form of cluster analysis. However latent class analysis is a probabilistic methods, and thereby a bit different from the distance based cluster analysis methods discussed earlier.

### Measures of association for binary data

So, if we want to inspect the association between two binary variables, we look at the contingency table. In other words we look at the pair wise association, kind of like conditional probability. If we see a strong pair wise association, we may contribute this to a common latent factor.

In other words, if many observation are alike and have the same distribution of answers. We assume that this is do to a common latent factor, like belonging to a specific cluster.

### The data matrix (cont’d)

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

▶ In many situations we use the frequencies for the observed score patterns instead of the raw data

▶ This is an efficient way of reporting the answers from many respondents – provided the number of patterns isn’t too large

▶ The columns for the raw data hold the answers to each indicator and hence an average will tell you how large a fraction of the respondents agree on this indicator

### Assumptions

▶ The responses to the p observed binary items are independent given the latent variable, y, – conditional independence. In other words, if we control for the fact that people belong to different segments, the answers are independent. Therefore the answers, themselves do not tell us which segment people belong to.

▶ This assumption can only be tested indirectly via an assessment of the fit of the model to the data

### The J-class model

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

We look at the probability that a given observation will belong to a class, when just choosing it by random.

### Goodness-of-fit – global tests

When assesing thise models we use a GOF -test to access whether the observed frequency is different from the expected frequency:

Et billede, der indeholder tekst, Font/skrifttype, hvid, skærmbillede

Automatisk genereret beskrivelse

Or we can use the Pearson Chi square GOF test.

Et billede, der indeholder Font/skrifttype, tekst, hvid, linje/række

Automatisk genereret beskrivelse

As done with Chi squared tests we have to have at least 5 observations within each observation. If this isn’t the case, we may collapse categories together.

A rule of thumb for evaluating chi-squared is to look at the residual for each chi squared destribtuion to see where the biggest deviation is present. Thus, chi-square residuals above 4 (sometimes 3) are seen as indications of a poor fit.

### Goodness-of-fit – model comparisons

We may also want to determine whether our fit is good by comparing our model to a reference model.

The idea is that we want to access how much of the variance is explained by our model. The comparasin model, will be a independence model, assuming that there is no explaining factors.

Our model should at least bet this !!

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

## Segmentation V - Advanced latent class analysis

In the last lecture we looked at how a segment could determine what people answered. This meant that the other predictor variable was conditional independent if we controlled for segments.

### LCA with concomitant variables

▶ The possibility of introducing explanatory variables/covariates for the finite mixture of distributions/LCA is typically done via the class membership probabilities. In other words, we now look at what people have scored and use I to determine which class they belong to. Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelseEt billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

### Finite mixture of regressions

▶ The “classical” segmentation approach (hierarchical and nonhierarchical) resembles in many ways the latent class analysis/finite mixtures of distributions and result often in comparable solutions

▶ The same kind of input data can be used and the overall segmentation process is very similar

▶ A related set of methods, finite mixtures of regression models, takes a rather different approach

▶ As the name suggests, we have a dependent variable and a set of independent variables, but the functional relationship between them depends on which of a finite set of segments the object belongs to.

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, linje/række

Automatisk genereret beskrivelse

Just think of this as a normal regression, but where the beta coefficient for all of our predictors. Kind of the same when we have a dummy variable. However, in this case, each predictor changes, based on which class people belong to, as seen in the example below:

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, linje/række

Automatisk genereret beskrivelse

When looking at reflective constructs, the loading indicates the relationship between latent constructs and their indicators.

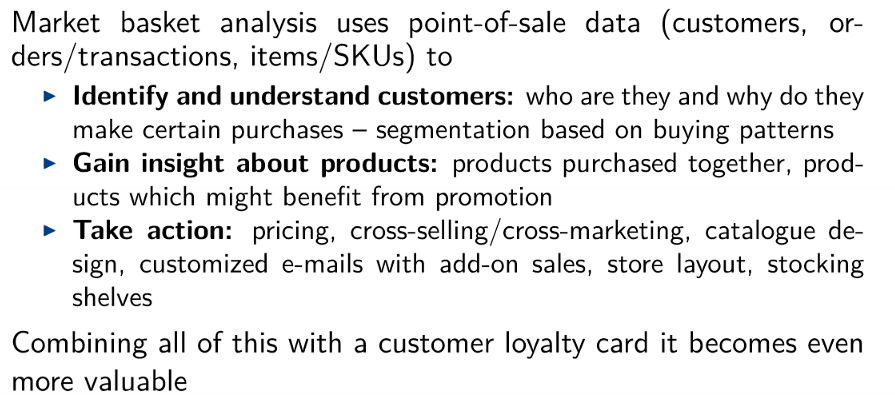
When looking at the formative constructs, the weights indicates the contribution to the latent construct.

# PRODUCT RECOMMENDATION & CUSTOMER TARGETING (II)

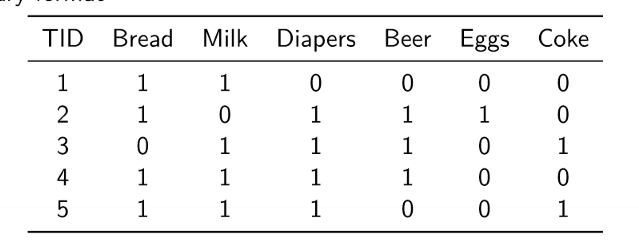
## Association rules:

Association rules are usually used when doing market basket analysis. We in short want to look at all purchases done in a store and determine whether or not some items are bought more frequently together. This is done by comparing the conditional and unconditional probability.

This might be a useful approach as we have a lot of data regarding transactions, which can be turned into knowledge for marketing promotion, inventory management and so on. (mer og krydssalg):



We do the analysis by looking at the items bought in each transaction. In we therefore treat our data like binary. Either the item is bought or not. However we miss out on information like, was this product on sale how many items where bought and so on.



**Support** = The number of times an item is present in the dataset. Often just displayed as a factor.

We often want the support count to be high, as this indicates that the pattern is repeated often. Which means that the store can earn a lot of extra money if they are able to cross sell.

However one should be aware that luxury items often have a low support, but may still be profitable for the shop to focus on.

**Confidence =**  Is an estimate of the conditional probability given any item.

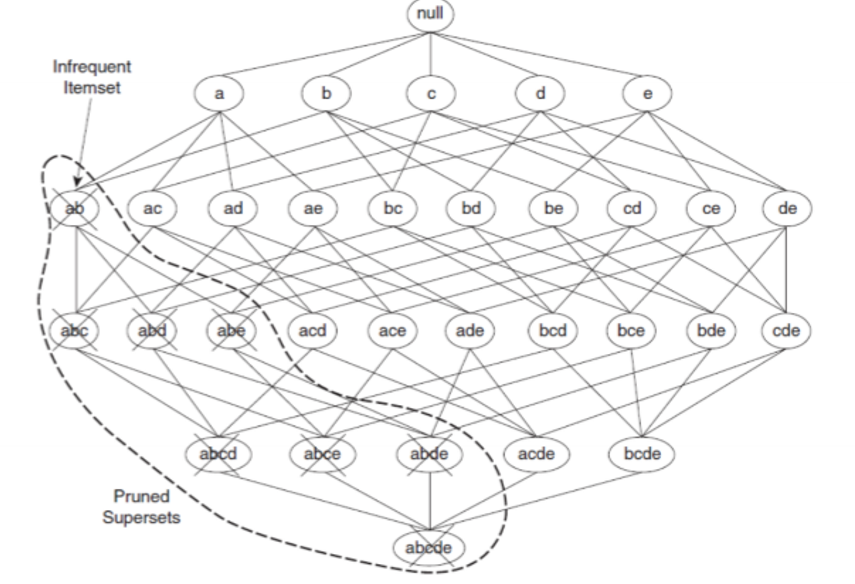
The higher this is the better, as the confidence indicates how much higher the probability is than if we just assumed no relationship. It is therefore a measure of how strong the relationship is.

**Bruteforce** = We simply make the associationrules for all items, by getting their support and confidence. However in real world this is not computational effective.

We therefore use **aprior** which assumes that if an item in itself is frequent - most of its subsets will also be frequent.

If the itemset is infrequent - most of the itemsets will also be infrequent.

In essence this means that we do not look at items which have a low support count in general. This can be seen in the picture below, where to combi AB is not very present. This means that we don't look at the subset where it is present:



**Lift =** Lift is the factor by which prediction improves when we apply the rule, compared to what we would be able to predict if we did not apply the rule.

* If this ratio is larger than 1 we have an upward lift – knowing that X has happened increases the probability that Y occurs

## Collaborative filtering:

This collaborative filtering/recommender system are all about finding interesting time to display to customers, with the end goal of increasing sales, customer loyalty, conversion rate and so on.

* Relevance – recommend items that are relevant to the user
* Novelty – recommend items that the user has not seen in the past
* Serendipity – recommend items that are unexpected
* Diversity – recommend different types of items

In general we have two types of recommendation. “ Tell me what is popular among other people who are like me” or “ tell me what is similar to something i like”.

**General problems:**

When we get a new user we can use their past rating, bought product and interest scores to recommend new unseen items. This is most likely done by learning from the k most relevant items. If we do not have a lot of rating o customers with similar rating, it may be hard to make a new prediction by users.

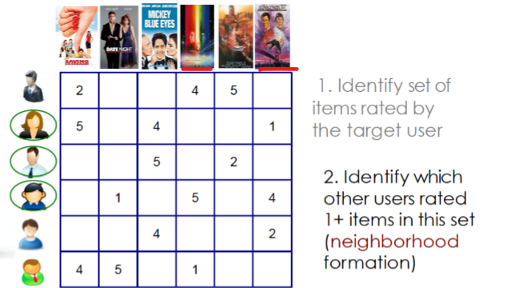
The same i know for items, if the customer have not bought anything yet, we cant really recommend anything. And if the customers have only bought one thing - we are likely to recommend items they have already bought.

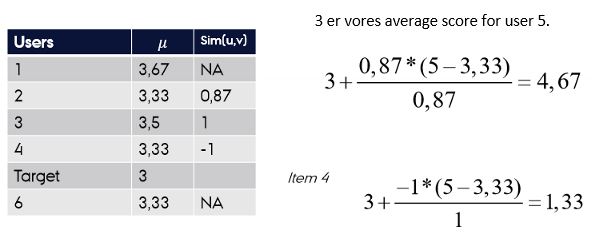
Both methods also take a lot of computation and it is hard to recommend niche products to some users.

We may also discuss how many user we want to compare to.

### user based similarity:

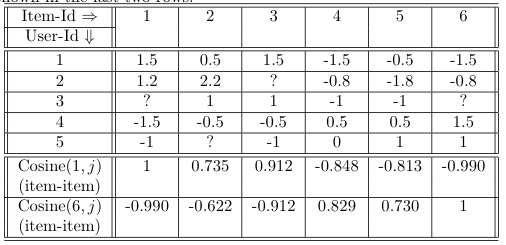
See the slides - But in short we find the standardized correlation between users who have related one of the 2 red movies. We standardise to make sure to account for the effect that some people in general rate movies higher.

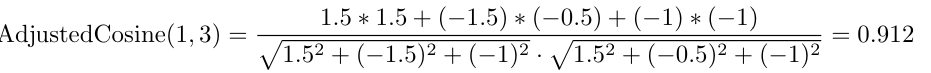


Example of calculating scores for item 5 and item 4.  


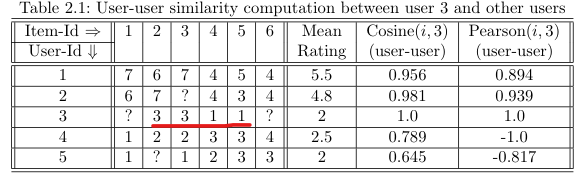
### Item based collaborative filtering:

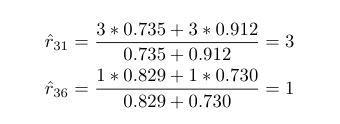
Here we look at the correlation between item rating of the customers. Once again we look at the standardized rating:





To then calculate the expected rating of an item by using the 2 highest cosine relation scores and multiply them by the unstandardised scores of the user.





In short:

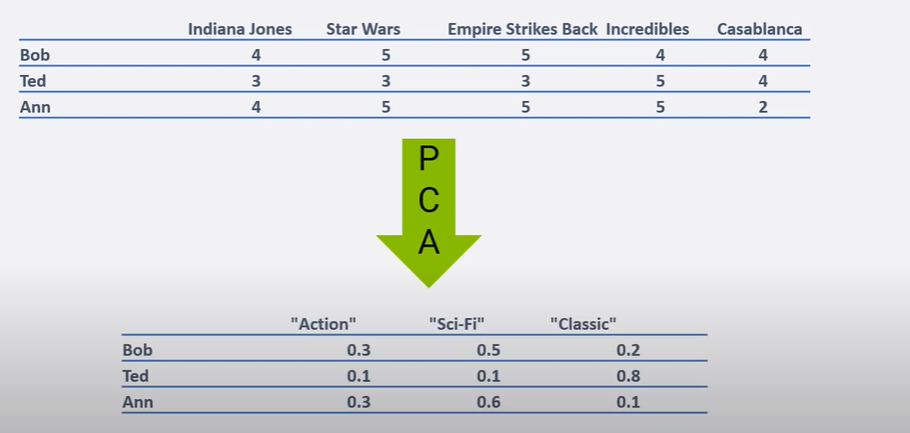
Between users = find correlations between rating of different users and multiply the other users rating by the correlation.

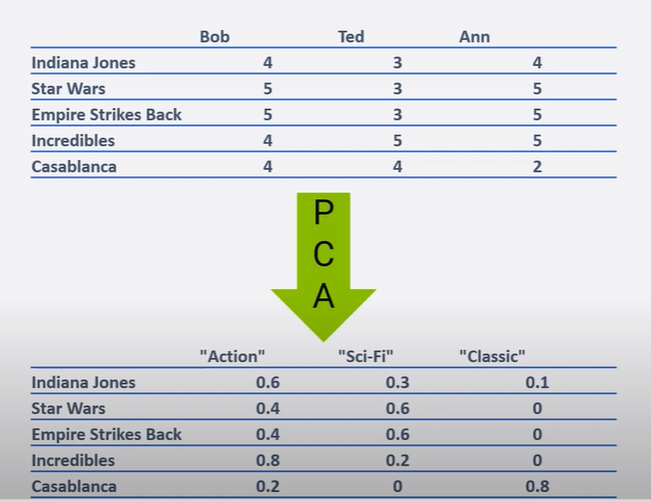
Between item = find correlations between items across users. Then multiply the users own score with the item correlation.

### SVD and SVD Funk

**SVD =** singular value dicomposition.

Her kigger vi altså på en matrixe med user rating samt en beskrivelse af hvilken kategori en film hører til. Vi kan således lave PCA på vores film scores, således at kan definere folk som værende til 0.3 action, 0.5 scifi og 0.2 classic.



Hvis vi transponerer vores matrix kan vi også lave det på film niveau, således at hver film beskrives som at bestå af en hvis mængde af hver faktor:  


Når vi har en rating for bob og en for the empire strikes back kan vi således udregne hans forventede score.

Altså kan vi ved SVM splitte vores originale matrix i en for film og en for kunderne. Dog under constraint at når vi ganger dem sammen skal vi få den originale matrix. For at predicte ukendte observationer kører vi denne algoritme indtil vi får en matrixe som mindre om den originale og derved en lav error.

SVM funk prøver at undgå at modellen over fitter ved man looper igennem matrixen mange gange med en learning rate.

ALM = er at man løser en ligning med 2 ubekendte.

Modeller med bias er faktisk bare at vi korrigerer for gennemsnitlig score.

regularinzation = har vi meget data bliver vores model kompleks.

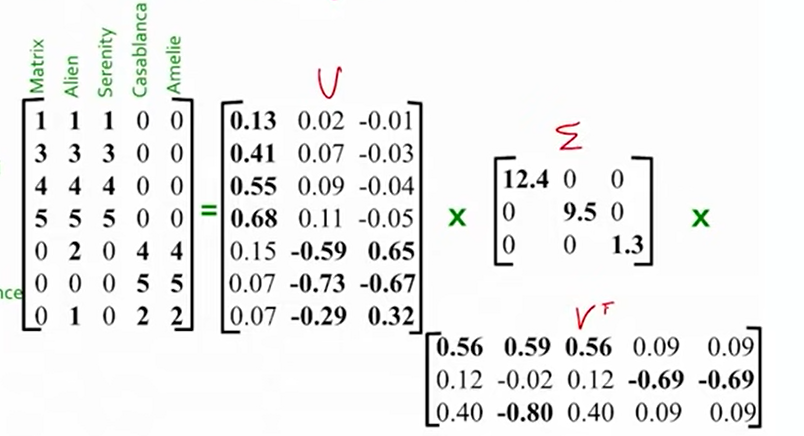
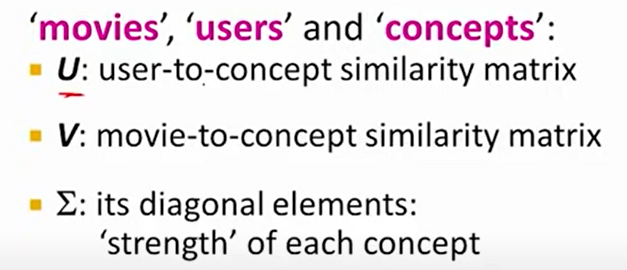
Har vi ikke meget data forbliver den simple for bedre at kunne generaliserer.

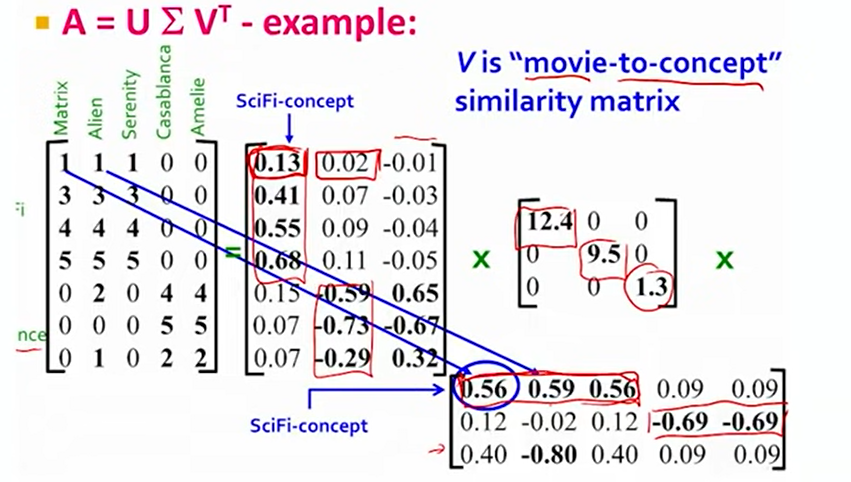
Bias = vi standardiserer for folk generelle score.

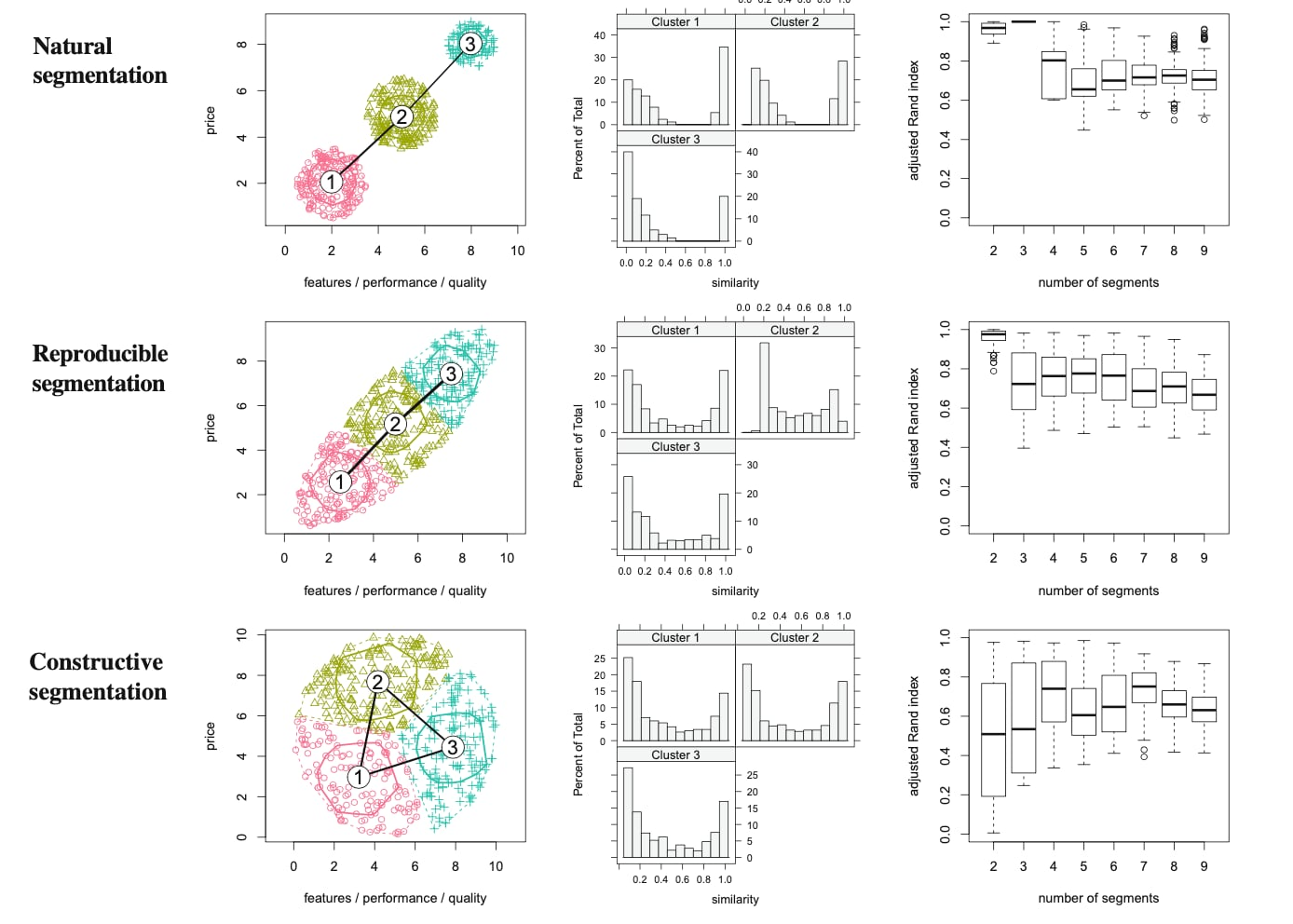
For film standardiserer vi for popularitet.

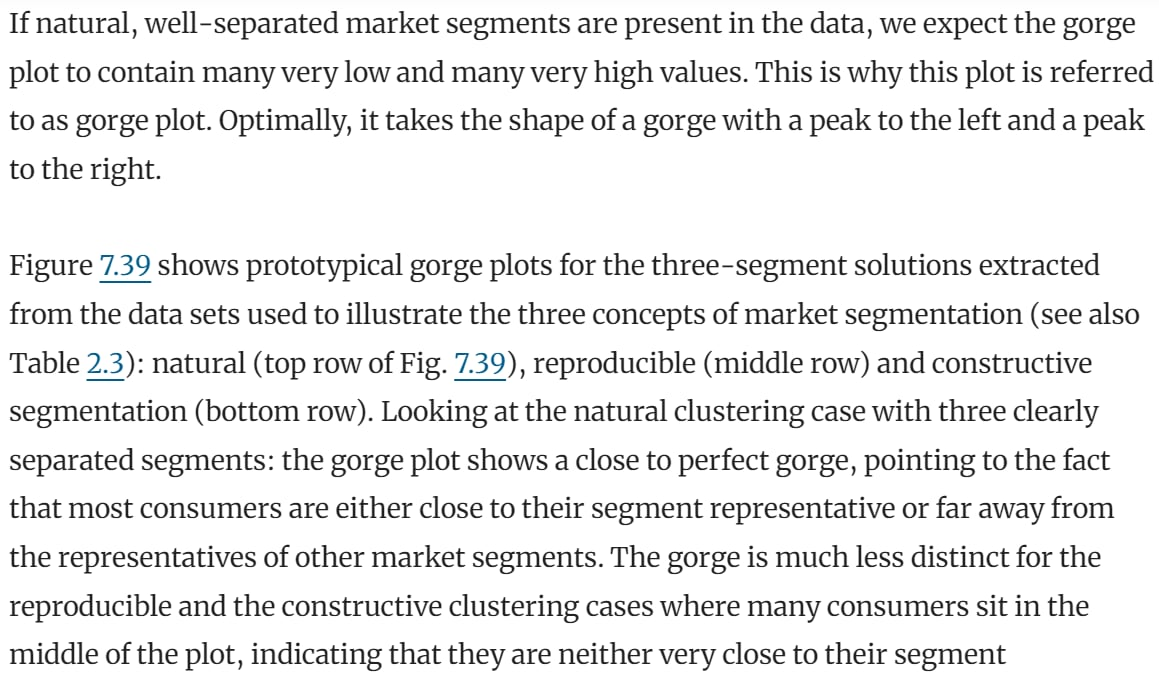
Igen - Vores matrix med folk og deres score af forskellige film kan blive skåret over i 3 matrixer. Den ene viser hvor meget filmene høre til et konstrukt eks. romantiske og scifi film. Den anden matrixe viser hvor meget hvor kunder kan lide hver konstrukt romantiske og scifi film. Vores sidste matrix er en diagonal som viser styrken af hvert konstrukt. Når vi ganger alt det her sammen få vi således vores originale matrixe igen.

Vi optimerer således at vores 3 matrizer bedst muligt approksimerer hinanden.





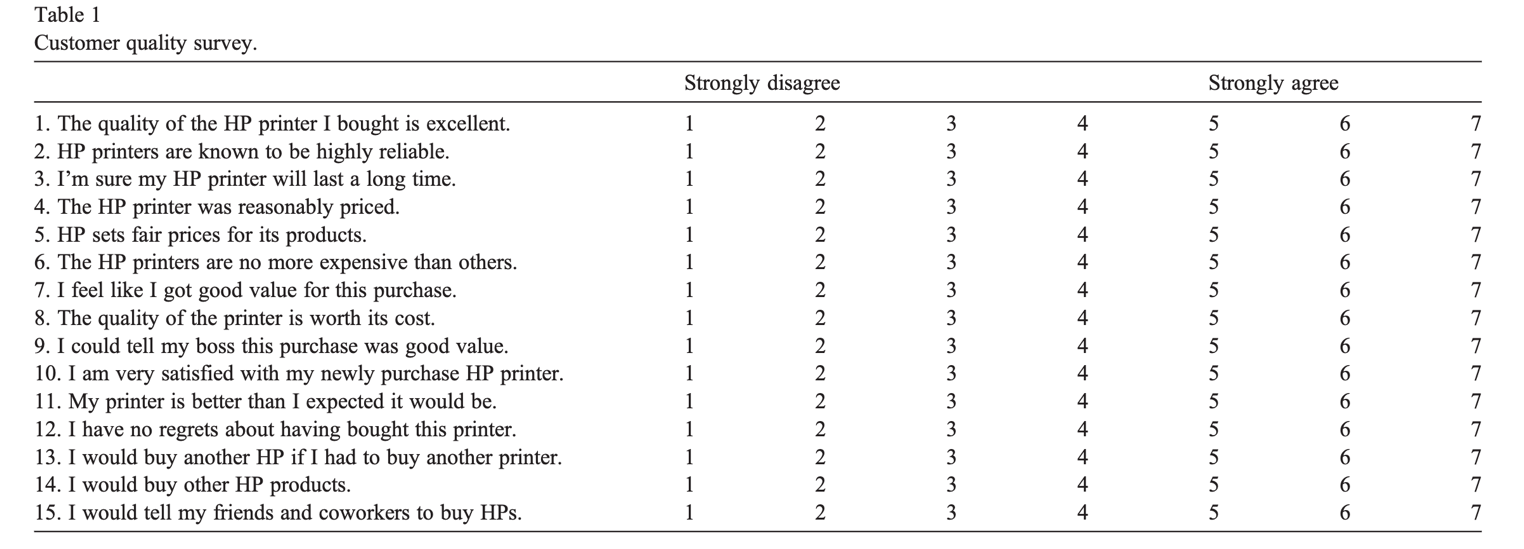




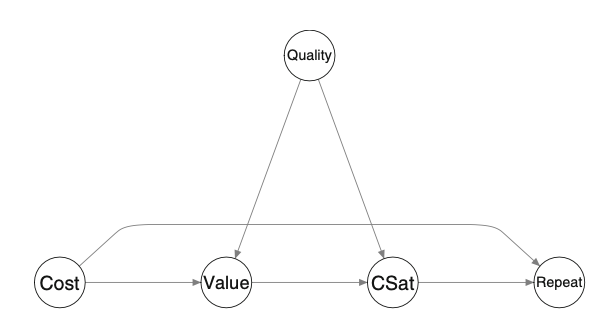
# Previous Exams

Exercise 1.

*Table 1 contains a survey of 15 statements based on which customers have provided their assessments of Hewlett-Packard printers. The structure of the survey is such that the first three statements are intended to tap quality, the next three measure cost perceptions, and then the last nine statements can be divided into three concerning customers' perceptions of value, three highlighting satisfaction, and three indicating behavioural intentions (adapted from Iacobucci (2009); Chapman and Feit (2015)).*

**

*Below is a SEM model of customer repeat purchase intent based on this survey. In this model, the cost of a product is associated with both perception of value and intent to repurchase, while perception of quality relates to both perceived value and satisfaction, which is then associated with repurchase.*



***1) Considering the model path diagram above:***

***a) Identify the exogenous, endogenous, and mediator variables. Identify a serial, diverging, and V-structure connection, and give an example of conditional independence and its practical implication for explaining customer repeat purchase. (10 pts)***Definitionen af en mediator variable: Effekten fra en anden variable går igennem dem.  
Cost har en effekt på value som har en effekt på CSat, Value fungere derfor som full mediator, for costs-effekt på CSat. Quality har en direkte effekt på CSat men samtidigt en effekt på value, som har en en effekt på CSat i denne situation er Value en mediator variable.

***b) Discuss under which circumstances this model is causal. Justify your answer. (10 pts)***

This model would be causal under the circumstances: If the four types of evidence are satisfied we would have the foundation to discuss whether the model is causal or not. The four types are Theoretical support, Non-spurious correlations (Making sure that there is no omitted variables bias, by controlling for the right variables or factors), The sequence of variables and factors, covariation the path coefficients should be significant.

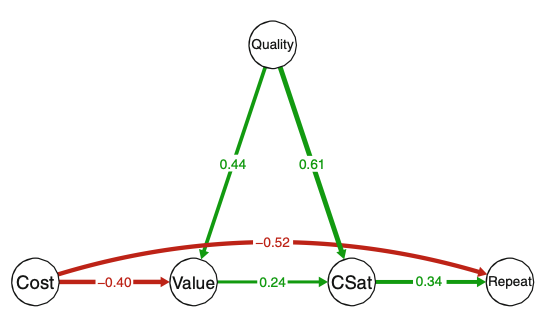
**3) Based on a model with five latent variables and 15 observed variables, how can one determine if the sample size is sufficient for testing the model using SEM? (5 pts)**

Et billede, der indeholder tekst, kvittering, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

or number of indicators \* 10

**4) Based on the coefficient estimates displayed in the path diagram below (all are statistically significant at 5%, and the data was standardized before analysis), estimate the total effect of the cost on repeat purchase. (10 pts)**

****

-0.52 + (-0.40\*0.24\*0.34) = XX.

Direct effect + (The product of Indirect effect)

**The influencer model is described on pages 70-73 in your curriculum textbook (Hair Jr et al. 20211). An online survey was carried out in 2023, building on this model. Compared to the description of the data collection in the textbook, the respondents in the 2023 online survey only saw the “real” influencer and used a different scale (rate how much you agree with each statement on a scale from 0 = strongly disagree to 100 = strongly agree). However, those responsible for the online survey forgot to include the “sic\_7” question when they developed the survey. Thus, only “sic\_1” to “sic\_6” are available (and they used “global\_sic” as the variable name for the global single item for redundancy analysis). See Tables 3.9 and 3.10 in the textbook for an overview of the variables. To solve this assignment, you must use the dataset in the file called “PLS\_data\_exam.csv”.**

**1) Discuss briefly how to treat missing values in PLS analyses in general. (5 pts)**

There’s different methods to impute the variables, such at KNN etc., however these methods haven’t been properly investigated. Missing values are therefore treated using mean replacement as often done in the litterature or just dropping the variables.

**2) Specify and run the influencer model. Next, evaluate the reflective measurement model. Be specific: report your implementation, relevant results, and comment on your output. (15 pts)**

Loadings, path coefficients, bootstrapped, overall R-Squared for the last identified variable

**Assuming a segmentation of the respondents based on the answers to variables “sic\_1” to “sic\_6” using cluster analysis:**

**3) Discuss considerations regarding the use of the three types of cluster analyses: classical cluster analysis, model-based cluster analysis, and latent class analysis. (10 pts)**

**4) Discuss the suitability of this dataset for the segmentation**

There is many problems with the dataset, as the six variables sic\_1 - sic\_6 is correlated.

**1) A new online sports store has over 100,000 items of sportswear and sports equipment, but its rating database has only 1,000 ratings. Which of the following would be a better recommendation system?**

**a) User-user collaborative filtering**

**b) Item-item collaborative filtering**

**c) User-item collaborative filtering**

**d) Content-based recommendation**

**Justify your answer. (7.5 pts)**

**2) In recommender systems, how can you balance the support and confidence to achieve accurate and relevant recommendations? (7.5 pts)**

In general then we want a high support, as we can use it on more items. Additionally, we also want a high confidence to ensure that our decision rule is strong (significant)

**CONTENT-BASED FILTERING**

In content-based recommendation systems, the focus is on the attributes of the items themselves rather than user behavior or interaction patterns. Here's a scenario where content-based recommendations might not perform as well as collaborative filtering:

### Scenario: Music Streaming Services

\*\*Context:\*\*

Music streaming services, like Spotify or Apple Music, have extensive catalogs with millions of songs across various genres. Users have unique and often rapidly changing tastes in music, which can be influenced by factors like current trends, new releases, or even social influences that might not be readily apparent through content analysis alone.

\*\*Limitations of Content-Based Recommendations in This Scenario:\*\*

1. \*\*Homogeneity of Recommendations:\*\*

- Content-based systems recommend items similar to those a user has liked before, based on characteristics such as genre, artist, or instrumentation. This approach can lead to a lack of diversity in the recommendations, as users are repeatedly presented with songs that are similar to their past preferences.

- In contrast, collaborative filtering can introduce users to a broader range of music by leveraging the varied tastes of similar users, potentially including trending or emergent genres that the user has not yet explored but may find appealing.

2. \*\*Dynamic User Preferences:\*\*

- Musical tastes can change quickly and are often influenced by external social and cultural factors that content-based systems cannot easily capture. For example, a user’s interest might shift towards a new genre following a live concert experience or a viral social media event.

- Collaborative filtering adapts more dynamically to such shifts, as it can quickly integrate changes in user behavior across the platform and update recommendations accordingly.

3. \*\*New Music Discovery:\*\*

- Content-based recommendations struggle with new songs that have few historical data points on their attributes. This makes it difficult for these systems to accurately place new tracks in the context of a user's preferences.

- Collaborative filtering, especially the user-based type, can more effectively recommend new songs that are quickly gaining popularity within certain user segments, thus aiding discovery.

4. \*\*Over-specialization:\*\*

- Content-based systems can cause the "filter bubble" effect, where users are only exposed to music that matches their previous behaviors. This can limit their exposure to a wider array of music they might potentially like.

- Collaborative filtering mitigates this by incorporating diverse user interactions, thus offering a more varied set of recommendations that could enhance user engagement and satisfaction.

### Conclusion

In the context of music streaming services, collaborative filtering generally outperforms content-based recommendation systems. It better addresses the dynamic, diverse, and socially influenced patterns of music consumption, leading to more personalized, engaging, and satisfying user experiences.

# Interpretation of R Output

## EFA & CFA

**Chi-Square for Model Test User Model:**

- A non-significant p-value (> 0.05) suggests a good fit, meaning the model is not significantly different from the data.

- A significant p-value (<0.05) indicates a significant different between model and data, hence a poor fit.

\* Chi-Square test is sensitive to sample size, often leading to rejection of the model when sample sizes are large.

**CFI and TLI**

Values closer to 1 indicates a better fit. Our acceptable threshold would be a value > 0.9

**Root Mean Square Error of Approximation (RMSEA).**

- Values < 0.05 indicate a close fit.

- Values between 0.05 and 0.08 indicate a reasonable fit.

- Values > 0.10 indicate a poor fit.

**Standardized Root Mean Square Residual (SRMR)**

- Values < 0.08 are generally considered a good fit.

- Values < 0.10 are sometimes acceptable

**Latent Variables (Loadings):**

- P-value < 0.05 indicates a significant loading

- Standardized loadings (above 0,6)

**Variances**

- Check if the variance estimates are negative.

o Negative variables indicate potential issues with model specification or data quality.

**Residuals Matrix**

Look for large residuals, indicating areas where the model doesn't fit the data well. Large residuals are above 0.10-0.15

## SEM

#### **Interpretation of R-Output**

**Output**

- Summary of the Fit Indexes (Chi-Square, RMSEA, CFI, TLI …) for SEM are interpreted as in EFA and CFA

- The structural coefficients are interpreted (Regressions: )

# Concluding Customers perceptions about CS, PV and MK are positively and significantly correlated with satisfaction. TS (Technical Service) perceptions is not significantly related to consumer satisfaction.

- Check modification indices if relevant