What is Latent Structure Analysis?

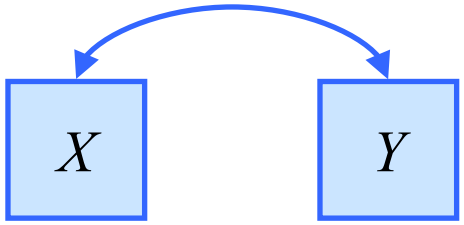
Nick Shryane
Social Statistics Discipline Area
University of Manchester

nick.shryane@manchester.ac.uk

What is Latent Structure Analysis?

- A family of statistical models.
- It explains the correlations among observed variables by making assumptions about the hidden ('latent') causes of those variables.
- Older models force us to choose between latent groups (classes) and latent dimensions (factors).
- Newer models allow a mixture of both.

Correlation

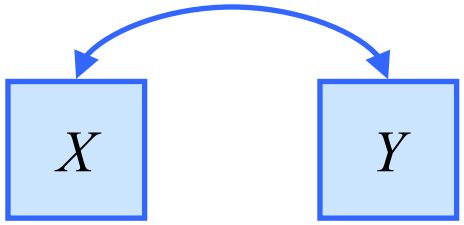


We observe a correlation between two variables. For example:

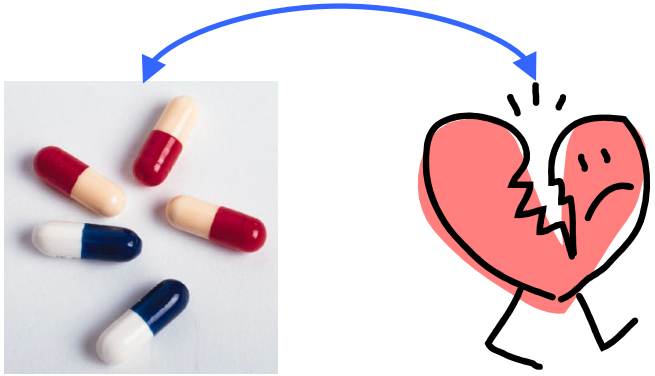
Number of fire engines and cost of fire damage

Shoe size and driving skill

Correlation



We observe a correlation between two variables. For example:



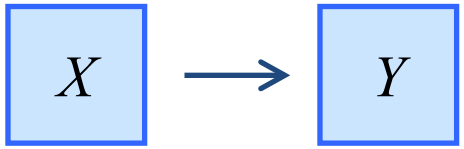
Hormone replacement therapy and reduction in heart disease.

Why do the correlations exist?

Could just be random?
No, these relationships have been observed repeatedly.

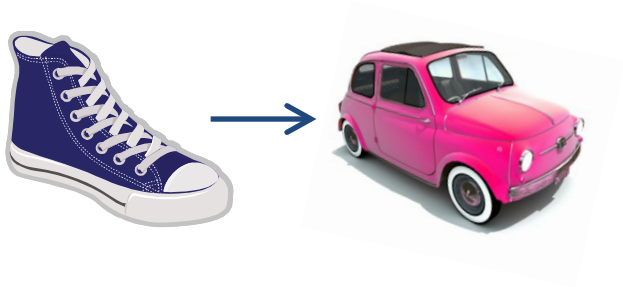
Causal relationships? X causes Y?

Direct Causation



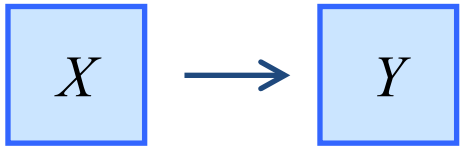
Correlation is consistent with causation. For example:

Firefighters cause the damage. Send fewer, less damage.

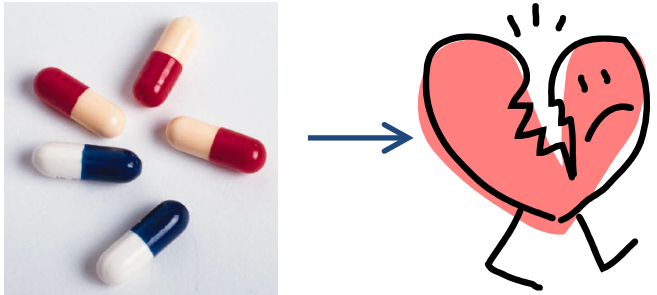


Big feet reflect a 'driving' gene. No need for driving test, measure feet instead.

Direct Causation



Correlation is consistent with causation. For example

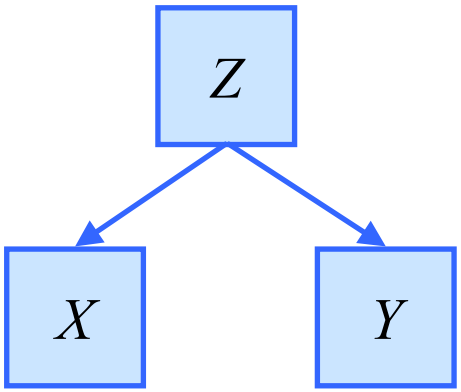


Hormone replacement therapy causes the reduction in heart disease. Give it to all women.

Some of these don't sound right as causal relationships.

What else might be going on?

Confounding

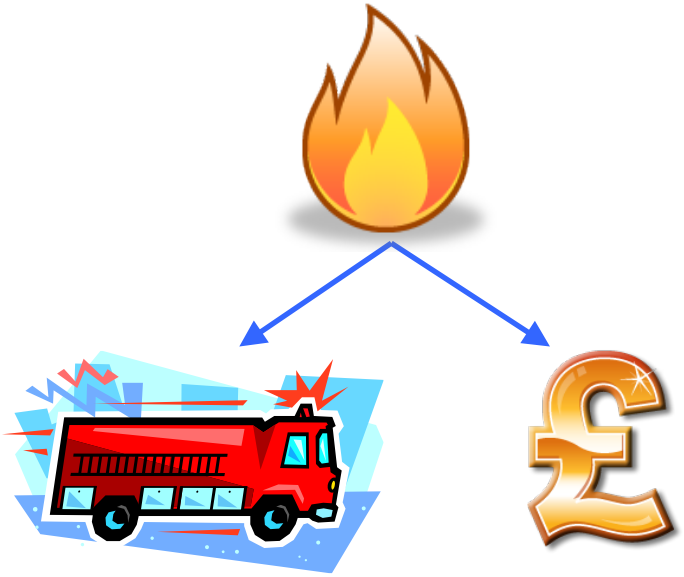


Correlation is also consistent with the effects of an unmeasured, third variable – a confounder.

A confounder causes both X and Y, resulting in their correlation.

After taking the confounder into account, the correlation will be different, usually weaker or absent.

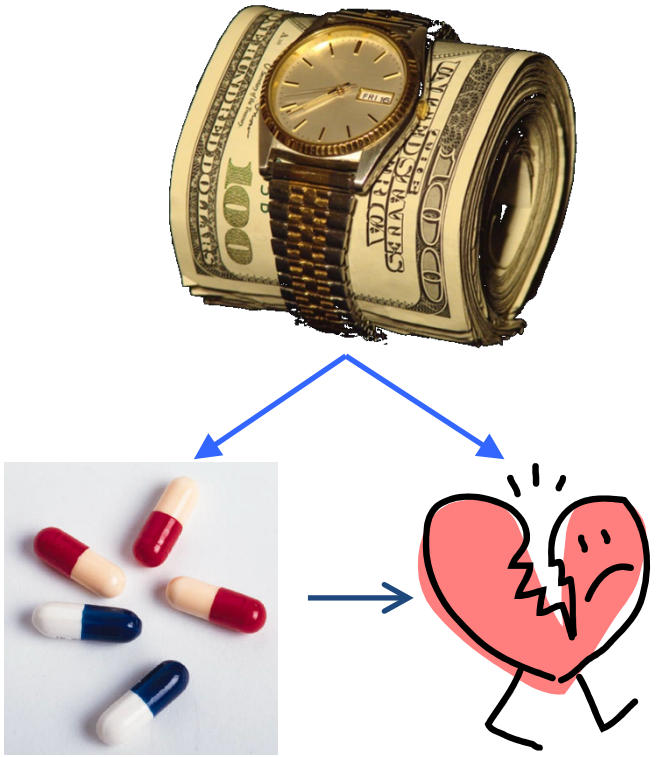
Confounding



Confounder: Size of fire

Confounder: Age

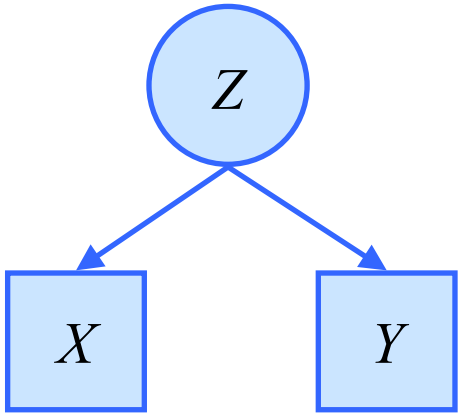
Confounding

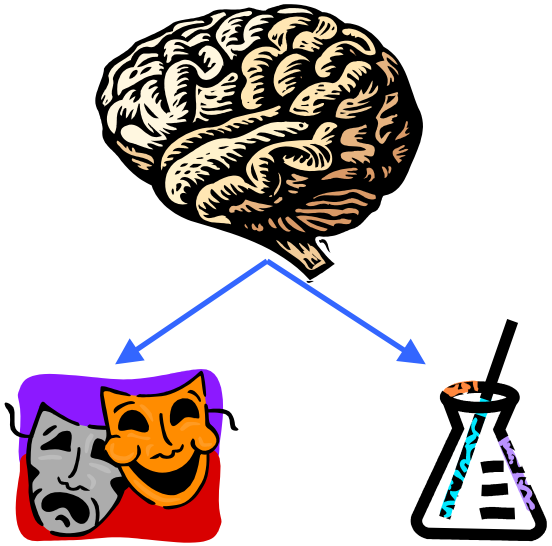


Confounder: Socio-economic status.
Middle class healthier and more likely to be prescribed HRT than working class.

Actually, after accounting for class (and lots of other things), HRT actually gives **increased** risk for infarct and stroke (Lawlor et al., 2004)

Latent variable





The third variable isn't always a confounder, i.e. a nuisance to deal with.

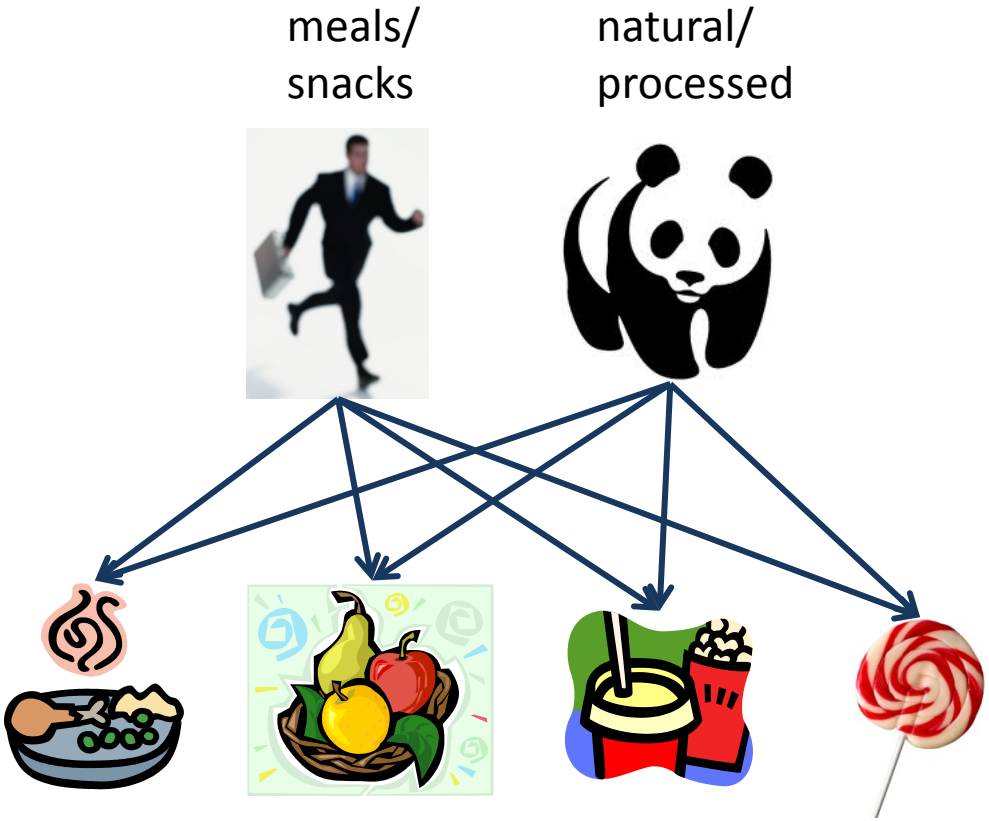
The third variable might be a key variable of interest that we do not, or cannot, measure directly.

In this case, we can call it a latent variable. ('Latent' - from the Greek for 'hidden' or 'dormant'.)

A classic example of a latent variable is **Intelligence**. We don't see intelligence directly – we infer it from performance (e.g. in arts and sciences).

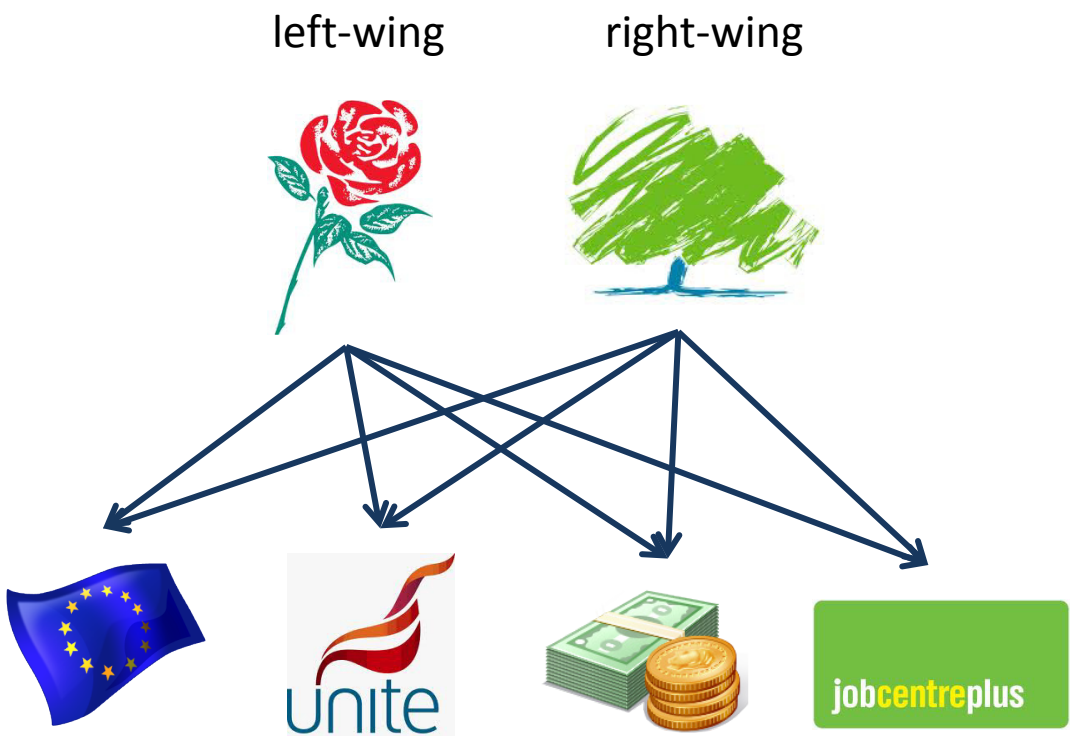
Latent variable examples

Food Preference



Latent variable examples

Political Preference



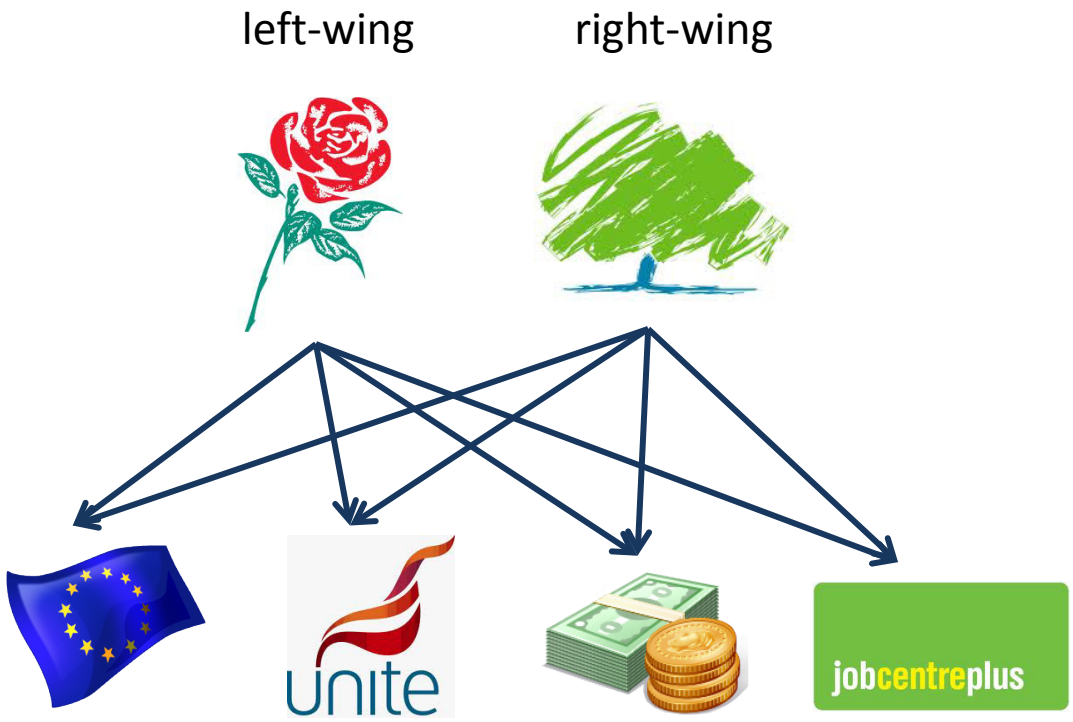
Conditional (local) independence

How do we find / model the latent variable?

Remember, **the latent variable acts like a confounder** -

Without assuming the presence of a latent variable, the indicators should be **correlated**.

Conditional (local) independence

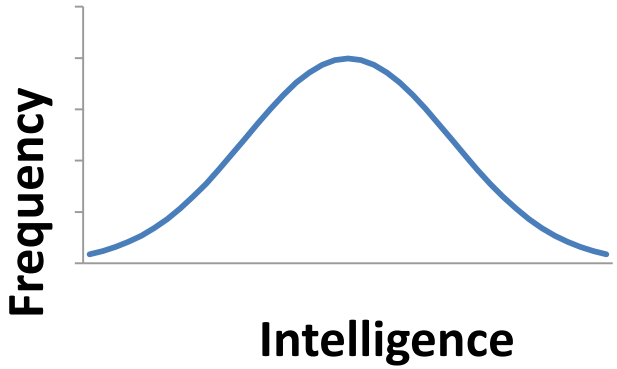


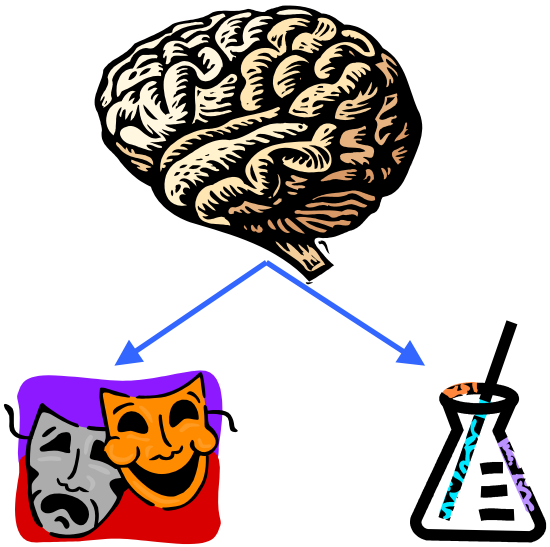
After conditioning on the latent variable(s) these **correlations should disappear**.

Assumptions

- The presence of a latent variable is an assumption, a conjecture.
  - The data may or may not be consistent with this.
- Statistical models have been devised to test these latent variable assumptions on data.
- These models require assumptions about the **distribution** of the latent variable

Latent dimensions - Factors





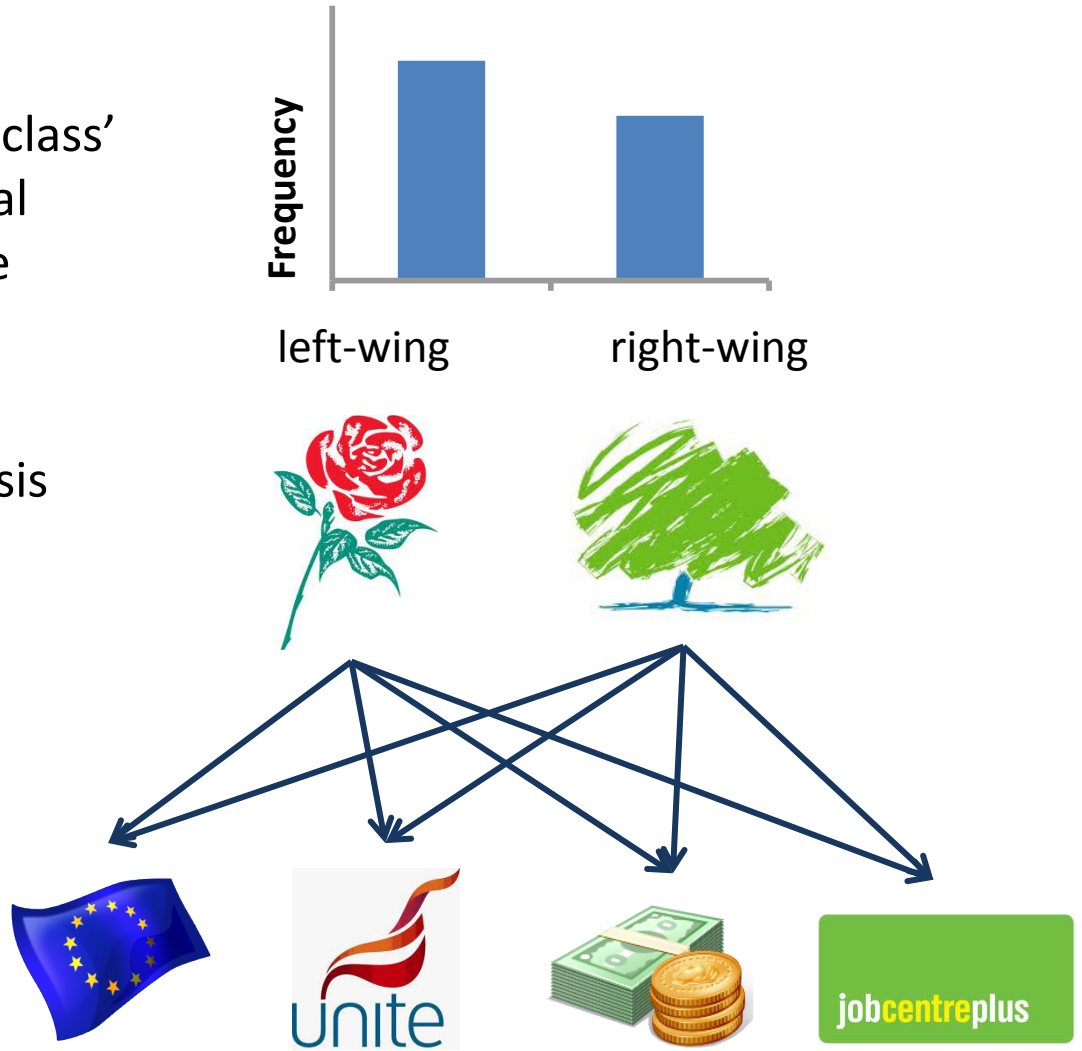
The classic model for intelligence assumed the latent variable was a continuous, normally distributed variable – a 'factor'.

Common Factor Analysis
(Spearman, 1904; Thurstone, 1947)

Latent groups / classes

What we (now) call a ‘latent class’ model assumes a multinomial latent distribution – separate groups / classes.

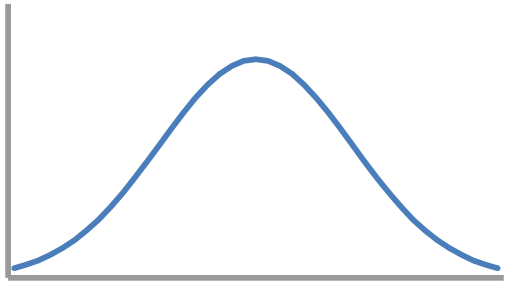
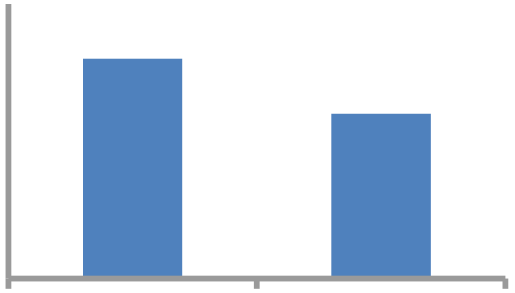
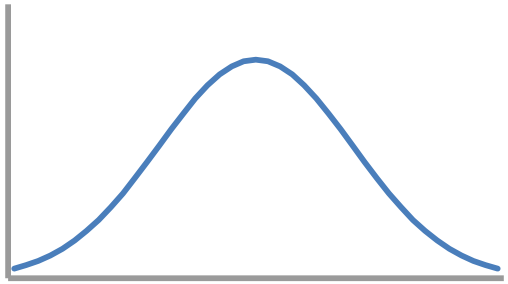
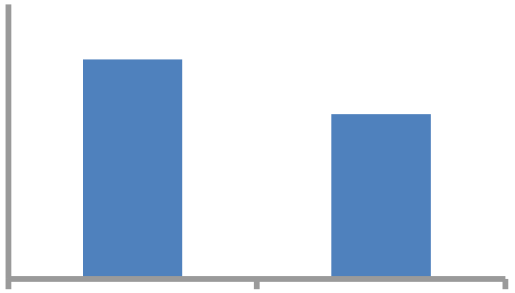
Latent Structure/Class Analysis (Lazarsfeld, 1950)



Observed vs. Latent distributions

- A word about the distribution of the observed variables
- It's sometimes thought that the distribution of the observed variable dictates what sort of latent structure analysis should be conducted
  - Continuous observed variables – Factor Analysis
  - Discrete observed variables – Latent Class Analysis

Latent vs. observed distributions

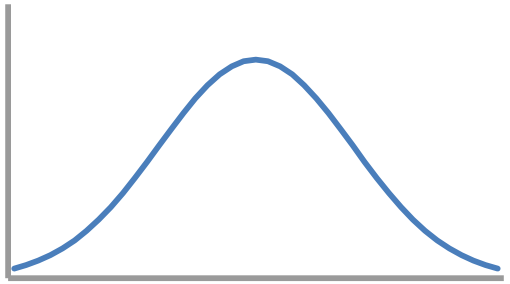
	Factor Analysis	Latent Class Analysis
Latent Variables		
Observed Variables		

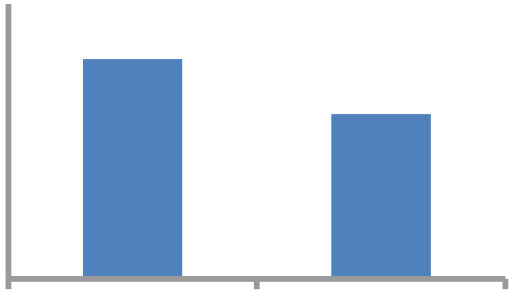
Latent vs. observed distributions

Latent Trait Analysis / Item Response Theory (Lord, 1952)

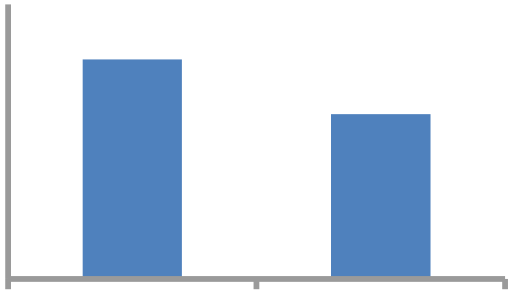
Latent Profile Analysis (Gibson, 1959)

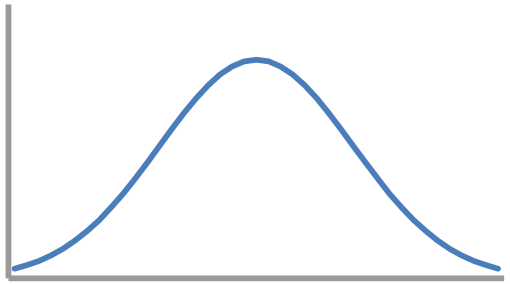
Latent Variables





Observed Variables



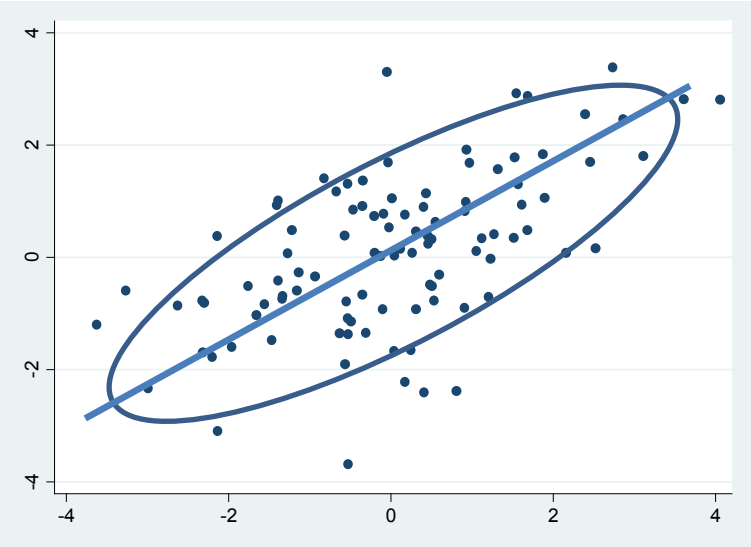


Observed vs. Latent distributions

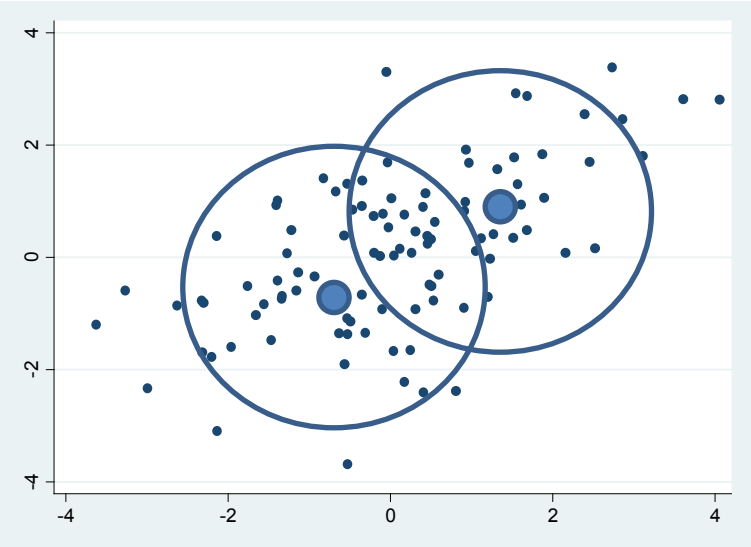
- Latent structure models can nowadays cope with just about any type of observed variable you care to mention (see Muthen, 1984):
  - Binary, e.g. yes/no, correct/incorrect
  - Ordinal, e.g. 3-,4-,5-, whatever-point Likert scale
  - Counts, e.g. frequencies
  - Censored, e.g. reservation wage
- The most important thing in Latent Structure Modelling is the distribution of the latent variables.

Factors vs. Classes – fundamentally different?

Latent Factor Model



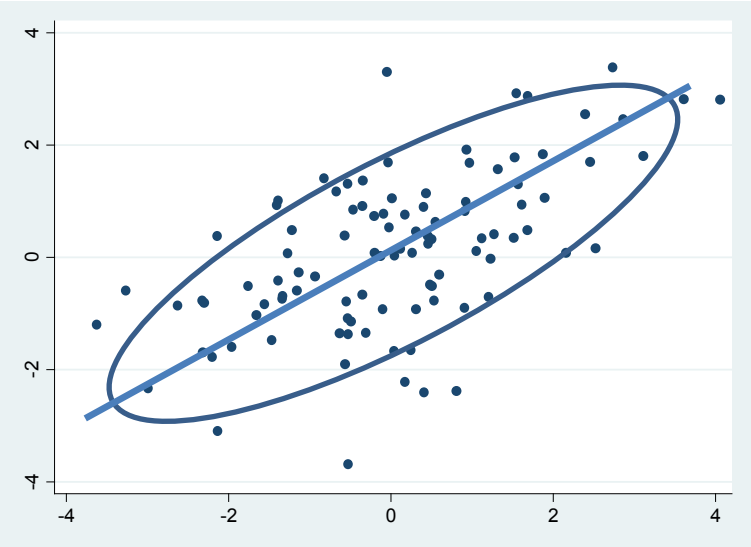
Latent Profile Model



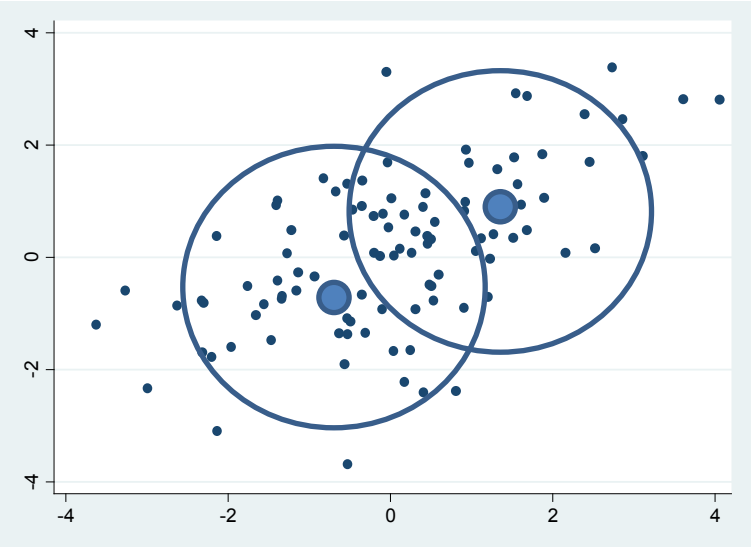
For **normally distributed indicators**, a Latent Factor Model and a Latent Profile Model can account for the observed correlations **equally well**.

Factors vs. Classes – fundamentally different?

Latent Factor Model



Latent Profile Model



If the indicators are **not** normally distributed, then extra factors or classes may be required to achieve local independence (i.e. a lack of significant correlations among indicators)

Choosing the right Latent Structure

- Which to use? Factor Analysis? Latent Class Analysis?
  - Continuous or discrete latent variables?

- Can use theory to guide the choice

Example: Acculturation; Schwartz & Zamboanga (2008)

Berry (1997) hypothesized four 'Acculturation' groups.

		Receiving Culture	
		Accept	Reject
Heritage Culture	Retain	**Integration**	**Separation**
	Discard	**Assimilation**	**Marginalization**

Berry (1997)

Schwartz & Zamboanga (2008) used questionnaire responses to test this model, by specifying a Latent Profile Analysis model with four latent profiles.

Example: Acculturation; Schwartz & Zamboanga (2008)

Stephenson Multigroup Acculturation Scale (SMAS); Stephenson (2000). SMAS questionnaire items, showing 4 out of 32.

Acculturation Profile (Class).	1. I know how to speak my native language.	2. I'm informed about current affairs in my native country.	3. I like to eat [American] food.	4. I regularly read an [American] newspaper.
Integration	☑	☑	☑	☑
Assimilation	☒	☒	☑	☑
Separation	☑	☑	☒	☒
Marginalization	☒	☒	☒	☒

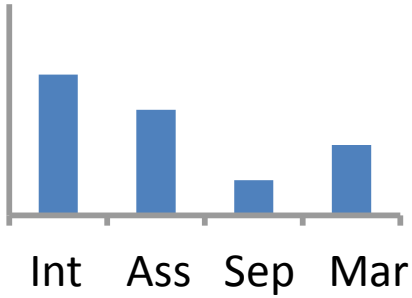
Theory demands classes?

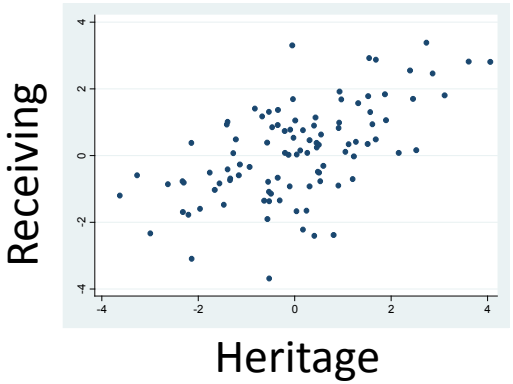
- Alternative view: the theory posits two dimensions of variation.
  - More or less attachment to Heritage culture.
  - More or less rejection of Receiving culture.

		Receiving Culture	
		Accept	Reject
Heritage Culture	Retain	**Integration**	**Separation**
	Discard	**Assimilation**	**Marginalization**

Berry (1997)

Theory-equivalent classes and factors





Acc4

A discrete Latent Class variable with 4 latent classes.

Rec *Her*

Two continuous Latent Factor variables.

Berry's theory admits a dimensional (factors) interpretation as well as a categorical (classes) interpretation. Take your pick.
Other theories may **demand** a **particular** latent structure.

How many factors / classes?

- How much latent structure do we need?
  - First, are you conducting an exploratory or confirmatory analysis?
- **Confirmatory** – hypothesis testing.
  - Theory tells me how many classes / factors to model.
  - Also, theory should tell me the **expected nature** of the factors/classes (cf. Berry's acculturation theory.)
  - Then compare the model to the data – how well does the theoretical model fit?

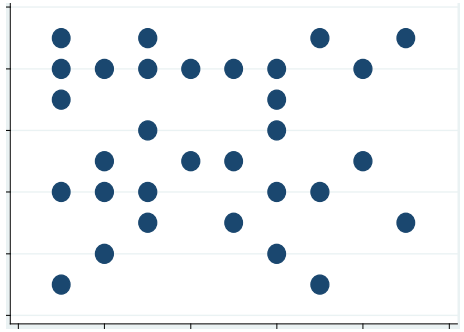
How many factors / classes?

- **Exploratory** – hypothesis generating.
  - Some automatic procedure selects the optimum number of factors/classes, and the nature of those factors/classes.
  - Vulnerable to distributional assumptions regarding the observed data (e.g. Normally distributed indicators).
    - Less problematic these days – wide range of different models available (for binary, ordinal, count, etc. data).
    - See Bauer & Curran (2004) for a discussion.

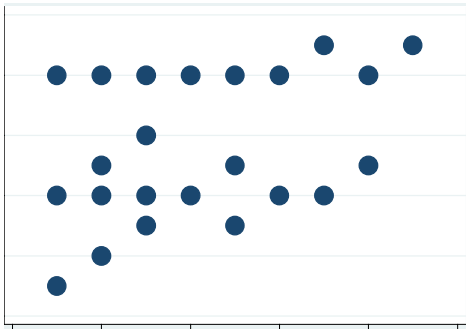
How many factors / classes?

- **Exploratory** – hypothesis generating.
  - Vulnerable to opportunistic capitalisation on sampling variation.
    - Samples vary by chance; might find factors / classes (i.e. pockets of correlation in the data) that are not characteristic of the population.

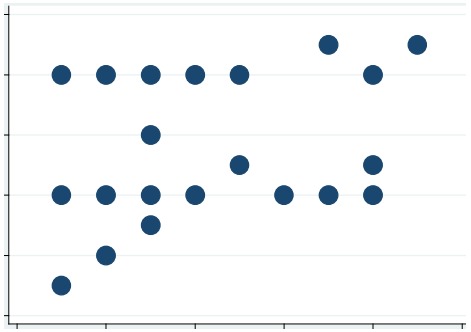
How many lines?



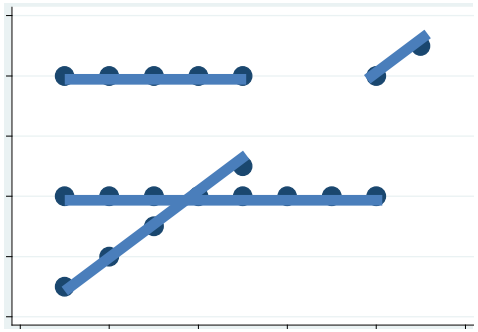
25% error



10% error

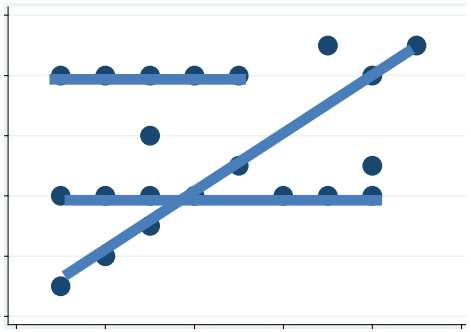


5% error



0% error

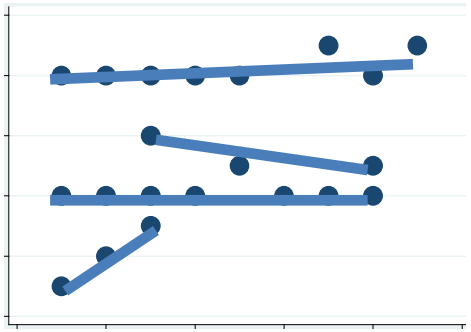
'true' 4 lines



5% error

3 lines?

Captures main structure



5% error

4 lines?

Distorts main structure

Model fit

- What do we mean by model fit?
  - Ability of the model to achieve **conditional independence**?
  - Ability of the model to **reproduce observed data**?
    - Both can be done by just making the model **more and more complicated**, by including more parameters.
  - Ability of the model to produce a **parsimonious description** of the data?
    - A trade off between model complexity (no. of parameters) and model fit.

Information criteria

- Can use the concept of **information** to decide upon the adequacy of the model-based description
  - The model uses parameters to describe the data
    - Factor loadings, latent class thresholds, etc.
  - How well does the model fit the data, compared to the number of parameters in the model?
    - How much extra **information** does each **additional parameter** give us?

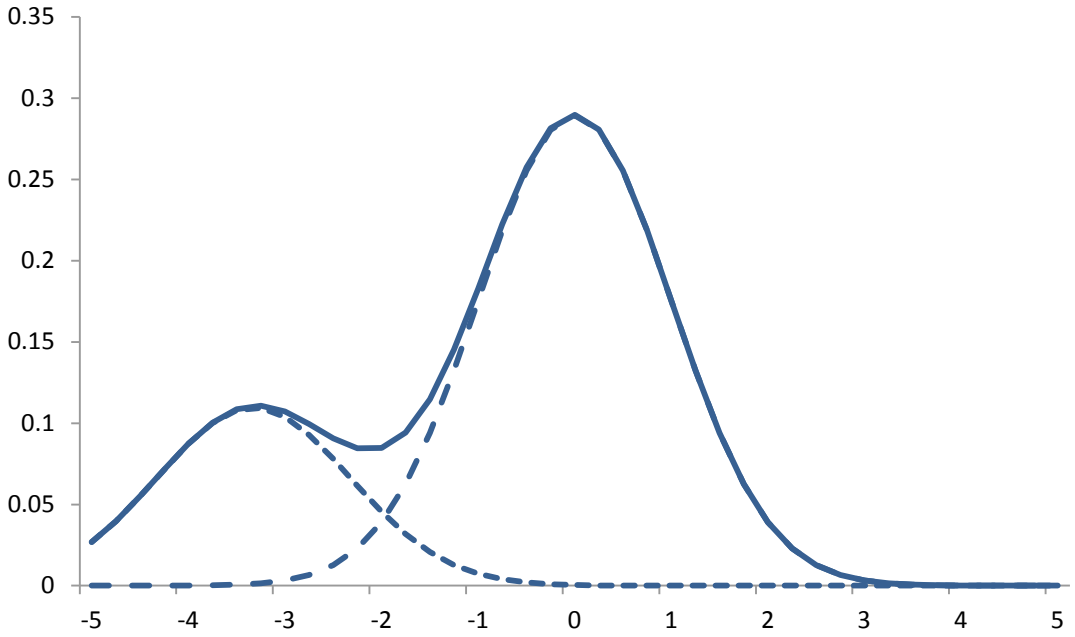
Information criteria

- Can use the concept of **information** to decide upon the adequacy of the model-based description
  - The Bayesian Information Criterion, BIC.
    - Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6(2): 461–464.
    - A measure of relative misfit (i.e. lowest is best), weighted by the complexity (parameters) in the model.
    - Useful to compare models with the same DVs but different model structure.

A new, unified approach

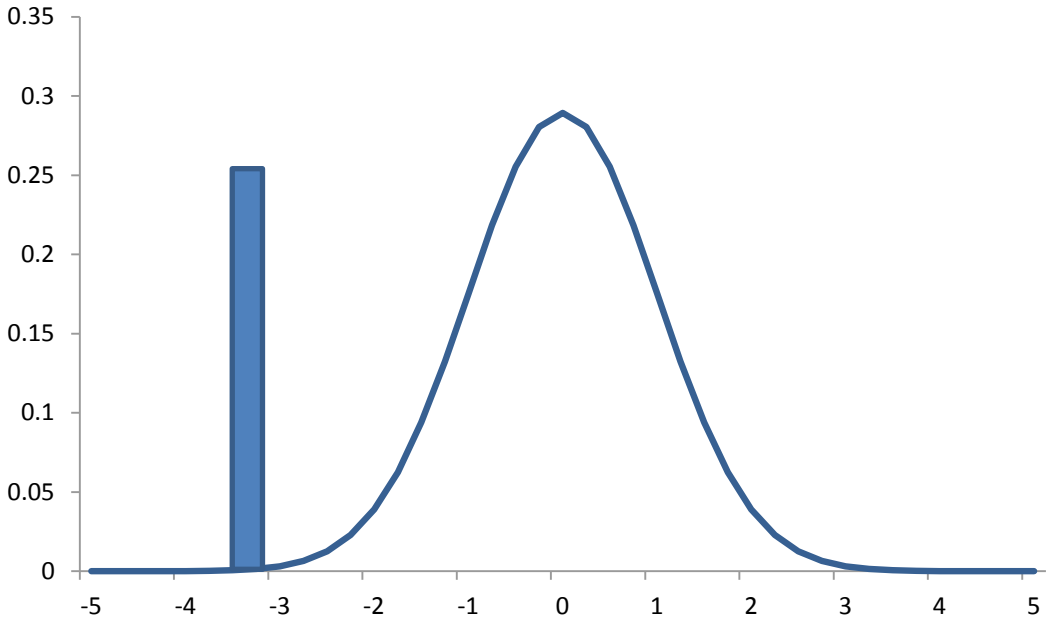
- Slowly, a unified approach to latent structure has emerged.
  - General(ized) Latent Mixture Models
    - E.g. Anderson, 1959; Bartholomew, 1984; Muthen, 1984, 2002; Skrondal & Rabe-Hesketh, 2004.
- Here, latent structure is conceived as a **mixture** of distributions.
  - A model can now contain factors **and** classes

Mixture distributions



A severely non-normal distribution is here produced by assuming a mixture of two normal distributions

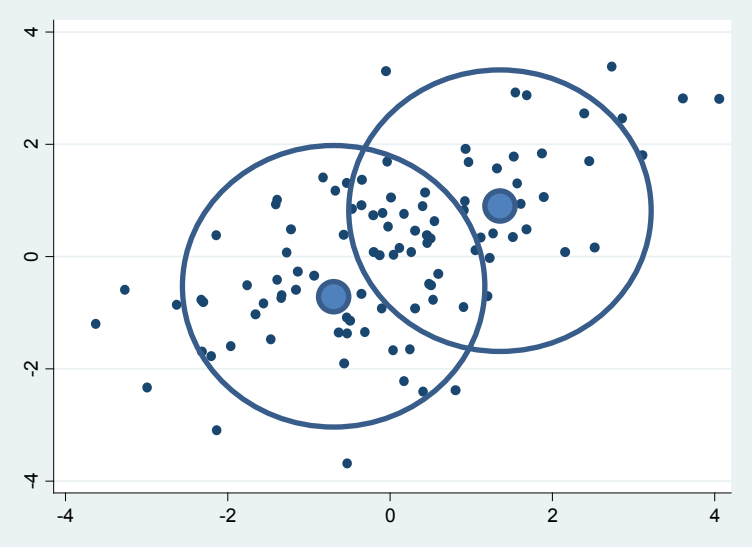
Mixture distributions



By assuming a mixture component has **zero variance**, a **latent class** is produced.

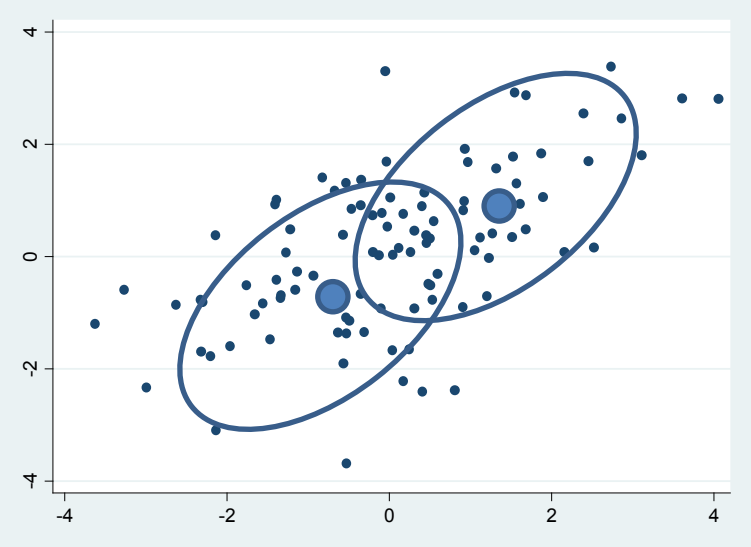
Factor Mixture distributions

Latent Profile Model



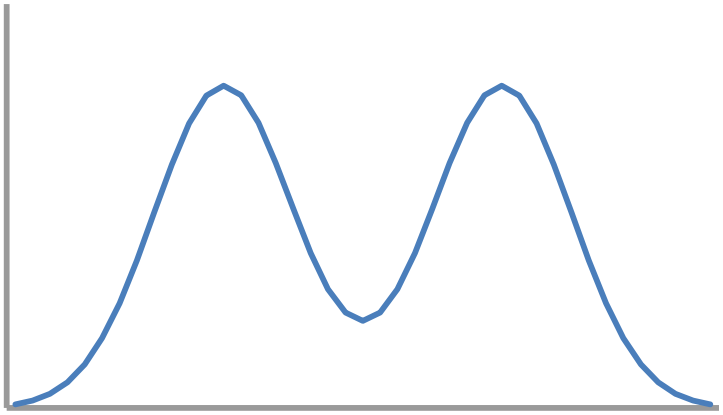
A latent profile model assumes zero correlations within classes (local independence)

Latent Factor Mixture Model



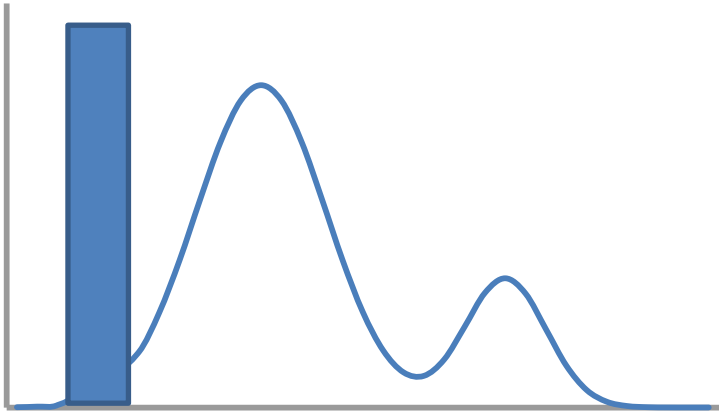
Adding a factor model within latent classes allows us to relax the local independence assumption within classes.

Flexible latent distributions



E.g. Latent **political left-right.**

People can be more or less left- or right-wing, but tend to cluster. Groups are not sharply defined, but have within-group variation.



E.g. Latent **mental health**.

A distinct, homogeneous, asymptomatic class on the left.

Two clusters of heterogeneous, symptomatic people, low and high.

Why does latent structure matter?

- Factors, classes, mixtures, blah blah, who cares?
- Scientific understanding, causality
  - Normally distributed traits? Normal distribution implies many, many small elements combining addititively.
  - Highly skewed traits? Might imply multiplicative combination of such elements.
  - Distinct classes? Some groups may contain different ‘active’ elements to others.

Why does latent structure matter?

- Factors, classes, mixtures, blah blah, who cares?
- Policy. For example:
  - If qualitatively different acculturation classes exist, they may respond differently to interventions, e.g. language classes.
  - Many mental health problems used to be seen exclusively as classes – well/ill – but now it is becoming appreciated that gradual variation exists
    - Basis for providing early support for at-risk groups.

Texts

Bartholomew & Knott (1999). *Latent Variable Models and Factor Analysis* (2nd Ed.). London: Arnold.

Skrondal & Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. Boca-Raton, Fla.: Chapman-Hall.

Software

- Mplus
  - Uses a very flexible approach to fitting generalized latent variable models to all types of observed data (including longitudinal, multilevel, etc.) .
  - Download free demo version of Mplus from:
    - www.statmodel.com
  - Download introductory tutorial from:
    - http://tinyurl.com/shryane-mplus-manual
    - http://tinyurl.com/shryane-mplus-examples

Software

- Stata
  - The **gllamm** command can fit a wide range of Generalized Linear, Latent and Mixed models (hence *gllamms*).
  - Download the manual and lots of worked examples from
    - www.gllamm.org

References 1

Anderson, T. W. (1959). Some scaling models and estimation procedures in the latent class model. In U. Grenander (Ed.), *Probability and Statistics*, pp. 9-38). New York: Wiley.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9(1), 3-29.

Gibson, W. A. (1959). 3 multivariate models - factor-analysis, latent structure-analysis, and latent profile analysis. *Psychometrika*, 24(3), 229-252.

Lawlor, et. al. (2004). The HRT-coronary heart disease conundrum: Is this the death of observational epidemiology? *Int. J. Epidemiology*, 33, 464.

Lazarsfeld, Paul F. (1950a) "The Logical and Mathematical Foundations of Latent Structure Analysis", in S. A. Stouffer (Ed.) (1950) *Measurement and Prediction, Volume IV of The American Soldier: Studies in Social Psychology in World War II*. Princeton University Press.

References 2

Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 17(2), 181-193.

Muthen, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115-132.

Muthen, B. O. (2002). Beyond SEM: general latent variable modeling. *Behaviormetrika*, 29, 81-117.

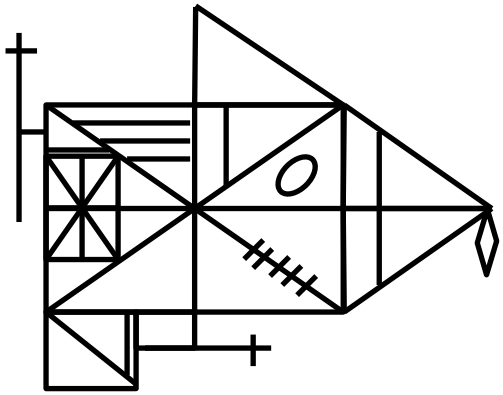
Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201–293

Schwartz, S. J., & Zamboanga, B. L. (2008). Testing Berry's Model of Acculturation: A Confirmatory Latent Class Approach. *Cultural Diversity & Ethnic Minority Psychology*, 14(4), 275-285.

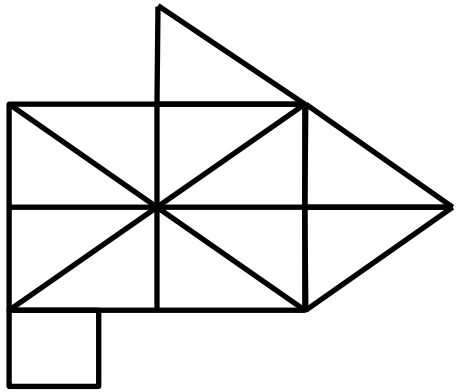
Thurstone, L.L. 1947. *Multiple-factor Analysis: A Development and Expansion of The Vectors of the Mind*. Chicago, IL: University of Chicago

Thanks for listening

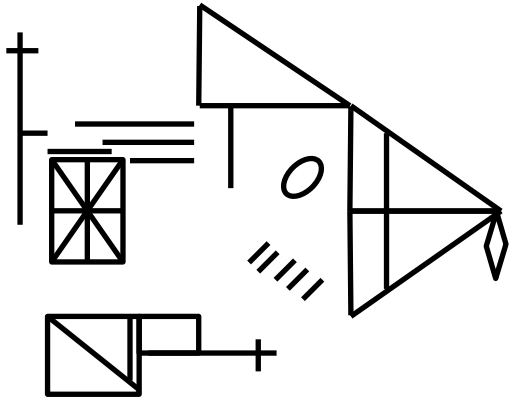
Original data



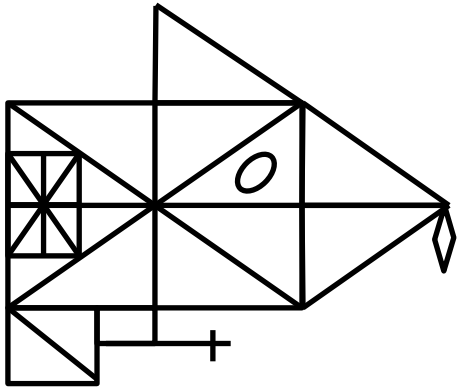
Model 1

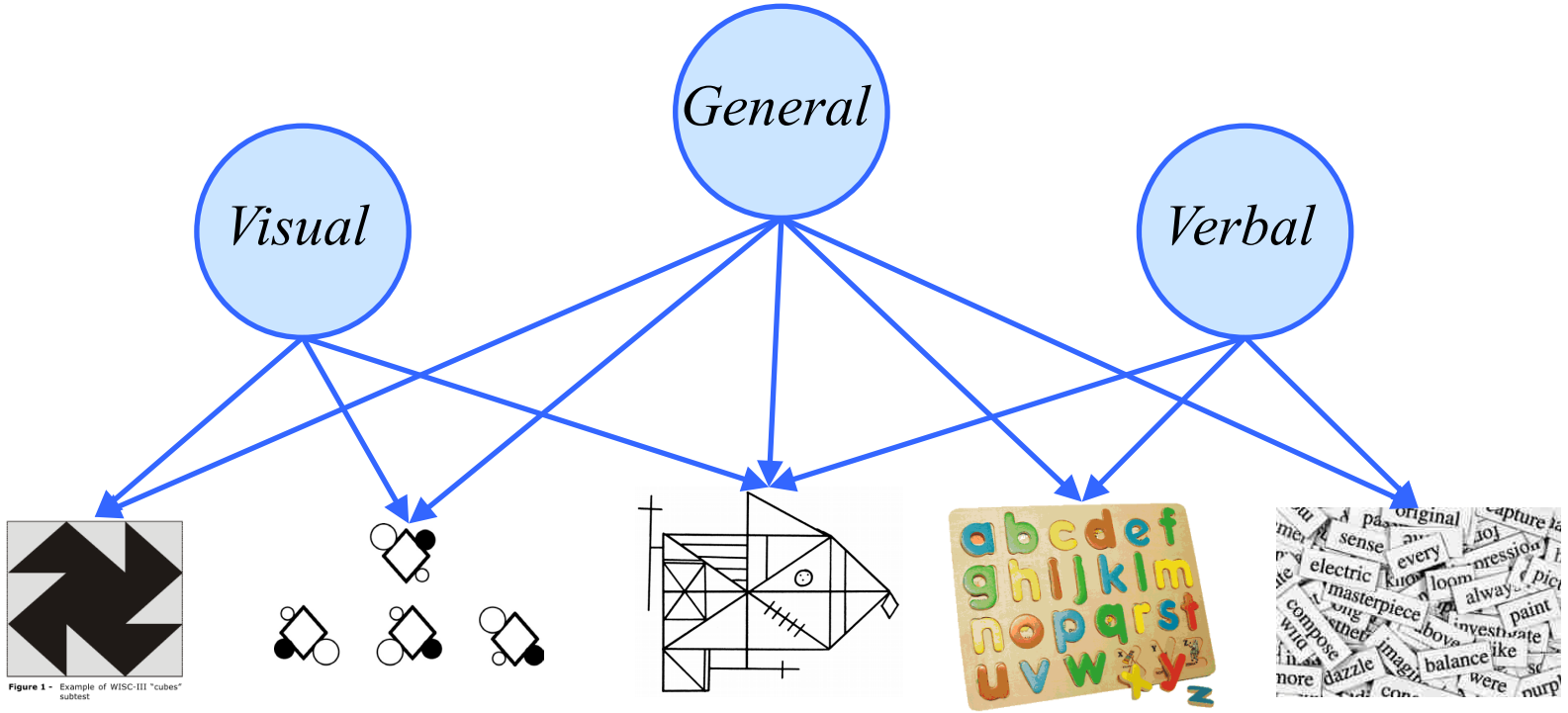


Model 2



Model 3




General
Visual
Verbal
Figure 1 - Example of WISC-III "cubes" subtest