POIR 613: Computational Social Science

Pablo Barberá

University of Southern California
pablobarbera.com

Course website:

pablobarbera.com/POIR613/

Dimensionality reduction

Dimensionality reduction

Goal: reduce number of features / variables to a smaller set

- When to use it?
  1. Multiple variables
  2. (potentially) Highly correlated
- Output: a smaller set of *principal components* or *latent variables*
- For example:
  - Survey items and a latent psychological measure
  - Stock prices for companies in similar industries
  - Country-level political or economic aggregates
  - Range of emotions that an image can generate
- Many techniques:
  - Principal component analysis (focus here)
  - Factor analysis
  - Item-response theory models

Principal Components Analysis (PCA)

- **Intuition**:
  - Combine multiple numeric features into a smaller set of variables (*principal components*), which are linear combinations of the original set
  - Principal components explain most of the variability of the full set of variables, reducing the dimensionality of the data
  - Key: fewer variables but information is not lost
  - Weights used to form PCs reveal relative contributions of the original variables
- **Mathematically**: assume several variables (X_1, X_2, ... X_K):

$$Z_i = w_{i,1}X_1 + w_{i,2}X_3 + \ldots + w_{i,K}X_N$$

where w_1 to w_K are known as the *component loadings* and Z_i (PC) is the linear combination that best explains variance in X_1 to X_K. We can have as many PCs as variables ($N \leq K$)

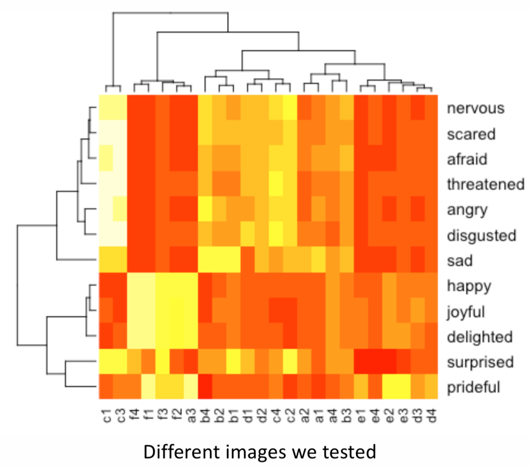
Example: dimensionality reduction of emotions attached to pictures

- Study on emotional responses to images about immigration
- Asked a sample of 100 respondents to rate a set of 24 pictures






Example: dimensionality reduction of emotions attached to pictures



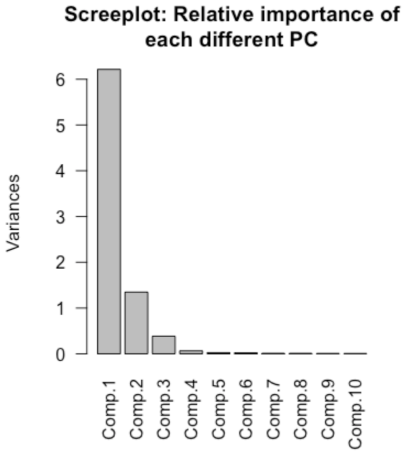
- **Coders were asked**: "Do you think this image would generate the following emotion to most people?"
- In graph, shade indicates average rating (darker = more likely)

Example: dimensionality reduction of emotions attached to pictures

	Comp.1	Comp.2
afraid	-0.32	-0.22
angry	-0.28	-0.18
delighted	0.30	-0.45
disgusted	-0.26	-0.18
happy	0.34	-0.48
joyful	0.32	-0.46
nervous	-0.33	-0.21
prideful	0.10	-0.07
sad	-0.27	-0.01
scared	-0.35	-0.24
surprised	-0.07	-0.16
threatened	-0.36	-0.33

- **Factor loadings (w_i)**: weights that transform predictors into the components (here only first 2 components shown)
- **How to interpret them?**
  - High values with same sign are positively correlated (covary together)
  - High values with opposite sign are negatively correlated (as one goes up, the other goes down)
- **Findings**: PCs correspond to
  1. Negative to positive emotion
  2. Emotion intensity

Example: dimensionality reduction of emotions attached to pictures



How many components should we keep?

- We can use a **screeplot**: plot of the variances of each of the components, showing their relative importance
- Here, 1st component explains a large proportion of the variance. 2nd component is also somewhat relevant. Rest of components do not seem important.
- **Conclusion:** we can reduce the dimensionality of all emotions to two components:
  1. Negative vs positive emotion
  2. Low vs high emotional response

Summary: principal component analysis (PCA)

- Each PC is a linear combination of the variables (numeric features only)
- Calculated so as to minimize correlation between components, limiting redundancy
- A small number of components will typically explain most of the variance in the outcome variable
- The limited set of PCs can be used in place of the (more numerous) original predictors, reducing dimensionality

Latent space network models

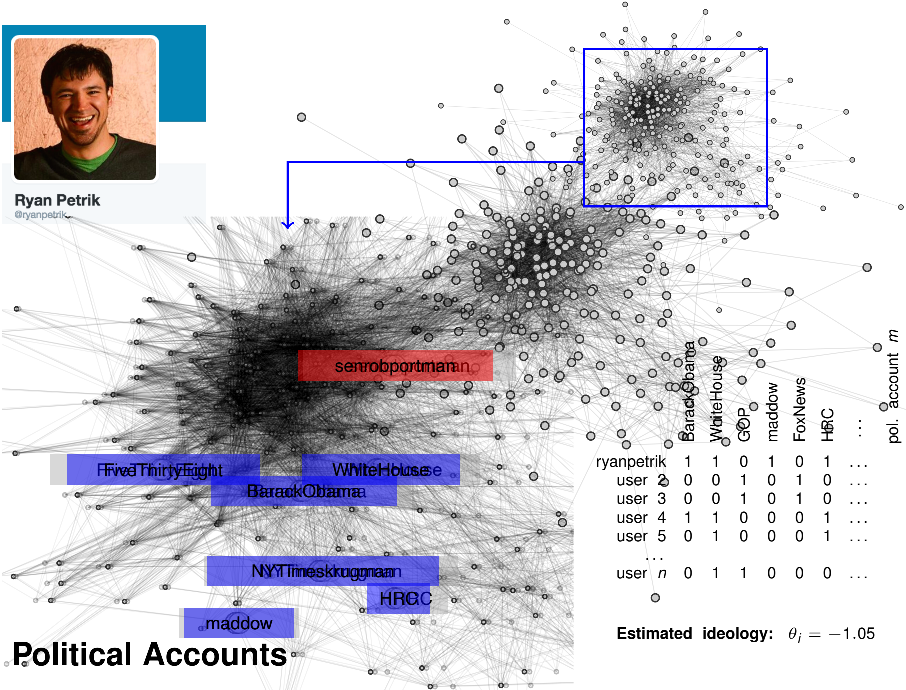
Latent space models

Spatial models of social ties (Enelow and Hinich, 1984; Hoff *et al*, 2012):

- Actors have unobserved positions on latent scale
- Observed edges are costly signal driven by similarity

Spatial *following* model (Barberá, 2015):

- **Assumption**: users prefer to *follow* political accounts they perceive to be ideologically close to their own position.
- Following decisions contain information about allocation of scarce resource: attention
- Selective exposure: preference for information that reinforces current views
- Statistical model that builds on assumption to estimate positions of both individuals and political accounts



Political Accounts

	BarackObama	WhiteHouse	GOP	maddow	FoxNews	HRC	...	pol. account m
ryanpetrik	1	1	0	1	0	1	...	
user 2	0	0	1	0	1	0	...	
user 3	0	0	1	0	1	0	...	
user 4	1	1	0	0	0	1	...	
user 5	0	1	0	0	0	1	...	
...								
user n	0	1	1	0	0	0	...	

Estimated ideology: $\theta_i = -1.05$

Spatial following model

- Users' and political accounts' ideology (θ_i and ϕ_j) are defined as latent variables to be estimated.
- Data: "following" decisions, a matrix of binary choices (**Y**).
- Probability that user *i* follows political account *j* is

$$P(y_{ij} = 1) = \text{logit}^{-1}\left(\alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2\right) \quad ,$$

- with latent variables:

  θ_i measures *ideology* of user *i*

  ϕ_j measures *ideology* of political account *j*
- and:

  α_j measures *popularity* of political account *j*

  β_i measures *political interest* of user *i*

  γ is a normalizing constant