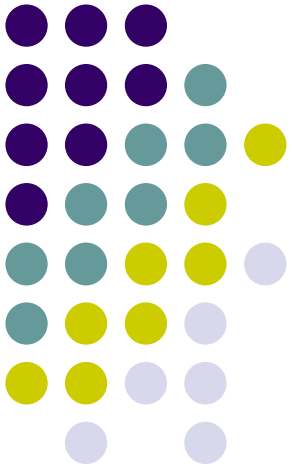
Big Data Analytics

Lecture 3
EXTRA: Wordshoal



Wordshoal

Wordshoal algorithm(Lauderdale and Herzog 2016) is based on 2 stages:

The first stage uses Wordfish to scale word use variation in **each debate separately**. By doing that, we estimate the **topic-specific positions** of MPs

In the second stage, it uses **Bayesian factor analysis** to construct a **common scale** from the debate specific positions estimated in the first stage, i.e., it unifies the multiple topic-specific positions by applying factor analysis to the topic-specific positions estimated in the first stage

What do you mean by that?

Wordshoal

Any factor analysis (FA) is used to **reduce the number of dimensions** within a data set **by choosing** only those “factors” (1 or more) that account for **most of the variation in the original multivariate data** and to summarize the data with little loss of information by projecting them onto a lower dimensional subspace

This can result in a **good approximation** of the original data

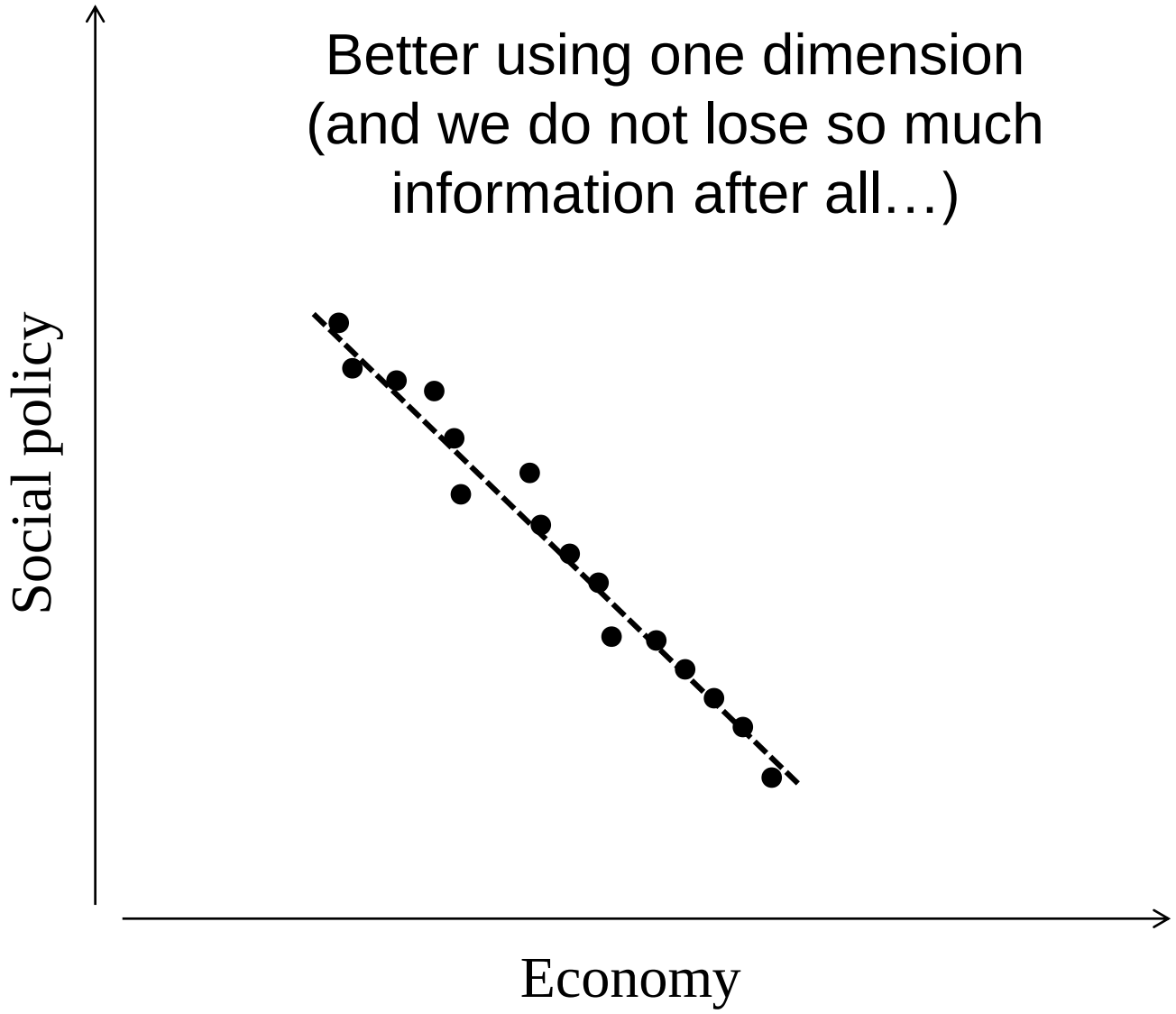
FA is especially useful when there is a **high-degree of correlation** present in our dataset (if correlation in your data is zero, there is no way to do any reasonable dimension reduction!!!)

Wordshoal

Do we really need two dimensions to capture all the variance in party positions?

Wordshoal

Better using one dimension (and we do not lose so much information after all…)



Wordshoal

In our case, the (Bayesian) FA allows to select out those debate-specific dimensions that **reflect a common dimension**, while down-weighting the contribution of those debates where the word usage variation across individuals seems to be idiosyncratic

This framework can be eventually extended to a 2-dimensional framework

Why a Bayesian FA? Its advantage is that it allows missing values in observed values. Going back to our example, this property is desirable if each MP does not necessarily speak in all legislative debates; thus, considerable debate-specific positions may be unobservable

Wordshoal

Wordshoal is therefore **attractive everytime** you want to analyze several different speeches/documents per-speaker/actor taken in very different contexts (over possible different topics) – as long of course there is *some correlation* about authors' positions across the contexts…

Lauderdale, Benjamin E., and Alexander Herzog (2016). Measuring Political Positions from Legislative Speech, *Political Analysis* (2016) 24:374–394

To install Wordshoal:

```
devtools::install_github("kbenoit/wordshoal")
```

Quanteda command: `textmodel_wordshoal`

Wordshoal

See an example here

N.B. the Quanteda command allows you to estimate only a 1-dimensional world. If you are interested to estimate a 2-dimensional world, write me!