Big Data for Public Policy

6. Unsupervised Learning

Elliott Ash & Malka Guillot

Where we are

- Past weeks:
  - w1: Overview and motivation
  - w2: Finding datastests using webcrawling and API
  - w3: Intro to supervised Machine Learning (ML) - regressions
  - w4: Text analysis fundamentals
  - w5: Supervised learning - classifications
- This week (w6):
  - Unsupervised ML
  - Corresponding references: Geron chap 8 and 9
- Next:
  - w7: ensemble explanations [Ash]
  - w8: econometrics and ML [Guillot]
  - w9: AI policies [Ash]

Today: Unsupervised ML

- Slides:
  1. Principal Components Analysis (PCA)
  2. Clustering
  3. Topic models
- Notebook:
  1. Crime

Outline

Unsupervised learning

- So far, supervised learning methods such as classification and regression
- Unlike for supervised learning, there is no known goal
  - No labeled data,
  - Not a prediction exercise
- The algorithm **discovers** patterns in the data
- We (human) **interpret** the results

  $\rightarrow$ More subjective than supervised learning
- Hard to assess the results

Setting

- Features $X_1, X_2, ...X_p$ measured on n observations, but no associated labeled variable Y
- **Dimensionality reduction** methods are needed
  - Especially in the case of **text data**, ML pbs often involve thousands of features.
  - not just for computational tractability, but also to help find a good solution.
  - Can be use as a **descriptive tool**
    - Can we extract information from the data and vizualise it?
    - Can we discover subgroups among th evariables or among the observations?

Examples

- Costumer segmentation in marketing
- Topic models widely used in political science
- Dimension reduction for pre-processing

Outline

Main idea

PCA (principal components analysis), a

- popular dimension reduction technique.
- identifies the axis that accounts for the largest amount of variance in the training set.
- finds a second axis, orthogonal to the first, that accounts for the largest amount of the remaining variance, and so on
- The unit vector defining the ith axis is called the ith principal component.

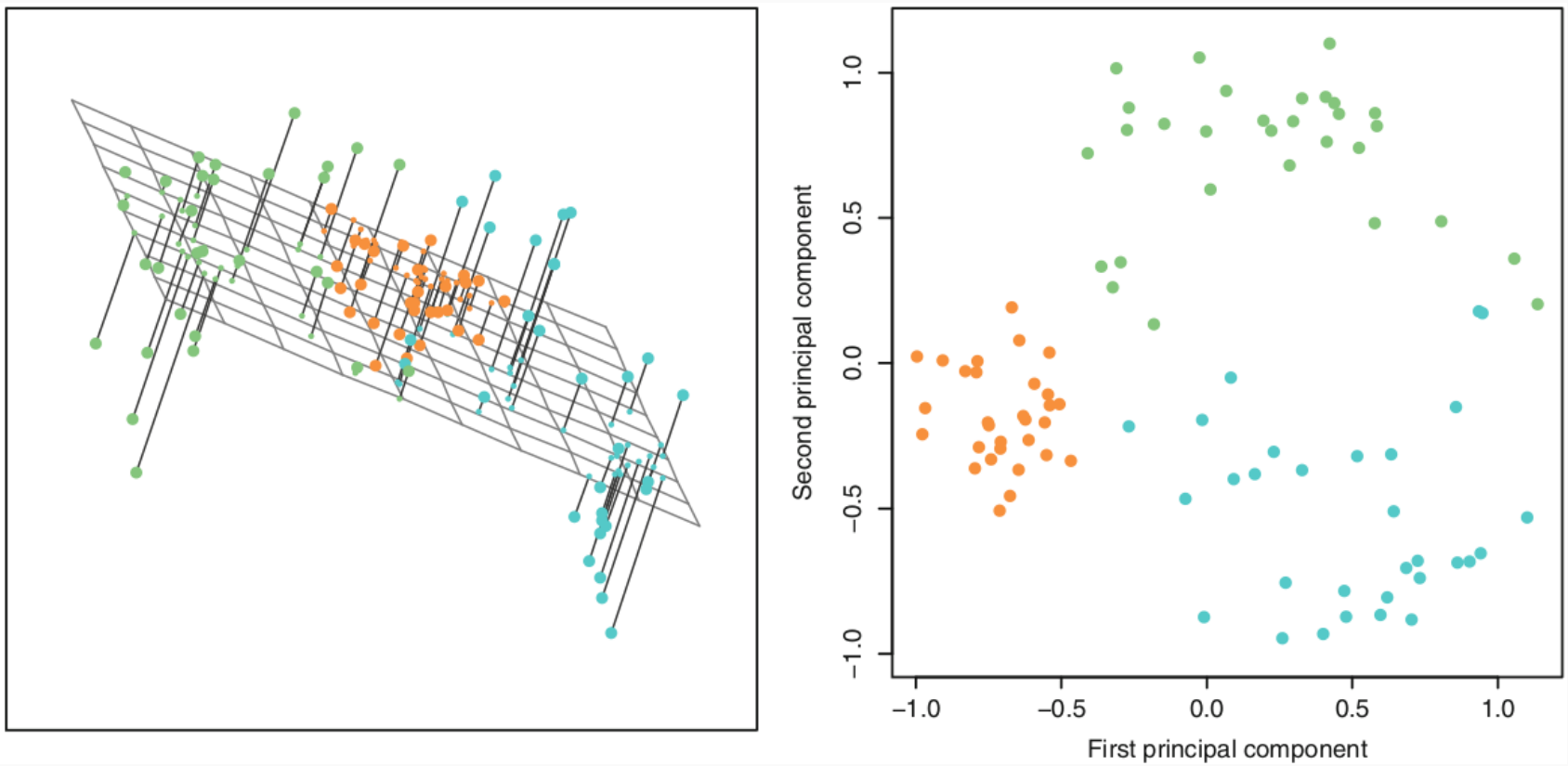
Objective

- Summarize a large set of feature variables with a smaller number of respresentative variables that collectively explain most of the variability in the original dataset

$\rightarrow$ Find a **low-dimensional representation** of the data that captures as much of the information possible

- If we can obtain a 2-dimensional representation, then we can plot the observations in 2D.

Illustration in 3D, projected on a 2D space



Left: simulated data in 3 dimensions

Right: projection on the first 2 principal components (plan represented on the left)

Principal Component

- What we are after
- Each of the dimensions found by the PCA is a linear combination of the p features.
- The *First Principal Component* of a set of features $X_1, X_2, ..., X_p$ is the normalized linear combination of the features:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + ... + \phi_{p1} X_p$$

  that has the largest variance
- Normalized means that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
- $\phi_1 = (\phi_{11}, \phi_{21}, ...\phi_{p1})^T$ is the *loading vectors* of the first principal component

Computing the first principal component

- We solve:

$$\max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \left\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2}_{\text{Sample variance of } Z_1} \right\} \text{ subject to} \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

- Re-writen :

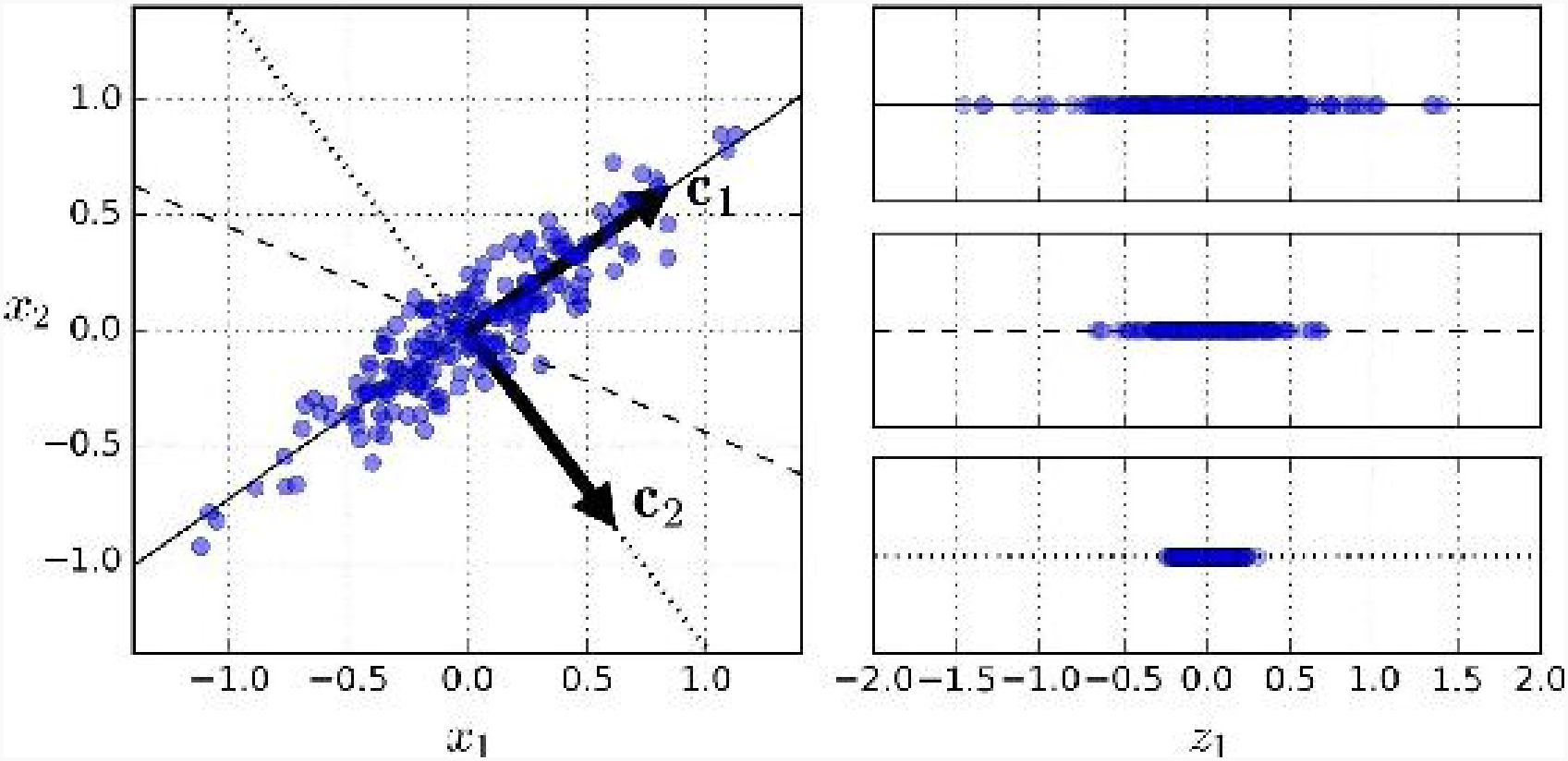
$$\max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \text{ subject to} \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

- as $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ (mean zero property)
- using eigen decomposition (outside the scope of the class)
- $z_{11}, \ldots, z_{n1}$ are the *scores* of the first principal component
- Solved using Singular Value Decomposition (SVD) [a standard linear algebra tool]

Finding the second principal component Z_2

- Z_2 is the linear combination of $X_1, X_2, ..., X_p$ that has maximal variance out of all linear combinations that are uncorrelated with $Z1$
- Z_2 uncorrelated with $Z_1 \Leftrightarrow \phi_2$ is orthgonal to ϕ_1

Illustrative example



Pre-processing the variables

- Variables should
  - be centered to have mean zero
  - have the same variance 1
- the results obtained depend on whether the variables have been individually scaled
- Step done by default in python

The Proportion of the Variance Explained (PVE)

How much of the information in a given data set is lost by projecting the observatrions onto the first few PC?

$\rightarrow$ plot the proportion of the variance explained by each PC

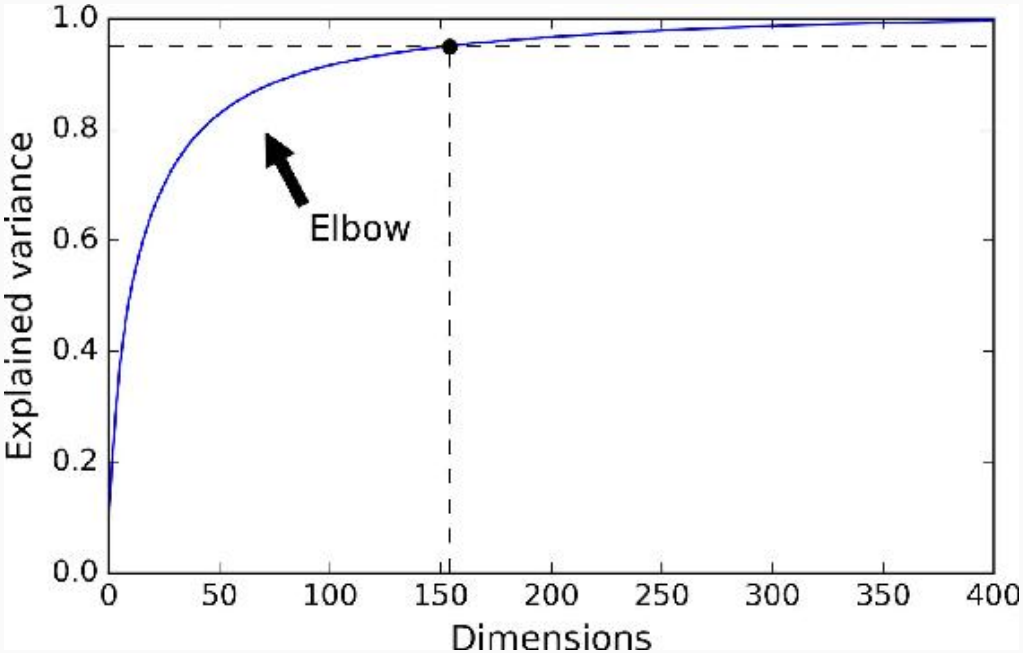
$$PVE_m = \frac{\text{Variance exeplained by the m}th\text{ component}}{\text{Total variance}}$$

Choosing the number of dimensions

There is no criteria for deciding how many principal components are required, but some rules of thumb:

- Choose the smallest number of PC required to explain a **sizable amount** of the variation in the data
- For dimensionality reduction
  - Explaining 95% of the variance is a good objective
- For data visualization:
  - Focus on a small number of axis that you can interpret
  - Do not interpret the components explaining less than 10%

Choosing the number of dimensions



An extension of PCA with categorical variables in sociology

Figure 1 The space of social positions and the space of lifestyles (the dotted line indicates probable orientation toward the right or left)

Source: *La raison pratique*, Bourdieu

Outline

Objective

When performing clustering, the aim is to group data into subsets so that:

- the objects grouped in each subset are similar, close to one another, homogeneous
- and different from, relatively more distant to the objects in other groups.

$\Rightarrow$ find some structure in the data.

K-means clustering

k-means clustering consists in **partitioning** the data into K disjoint clusters, setting the desired value of K, based on the p variables.

K-means clustering

k-means clustering consists in **partitioning** the data into K disjoint clusters, setting the desired value of K, based on the p variables.

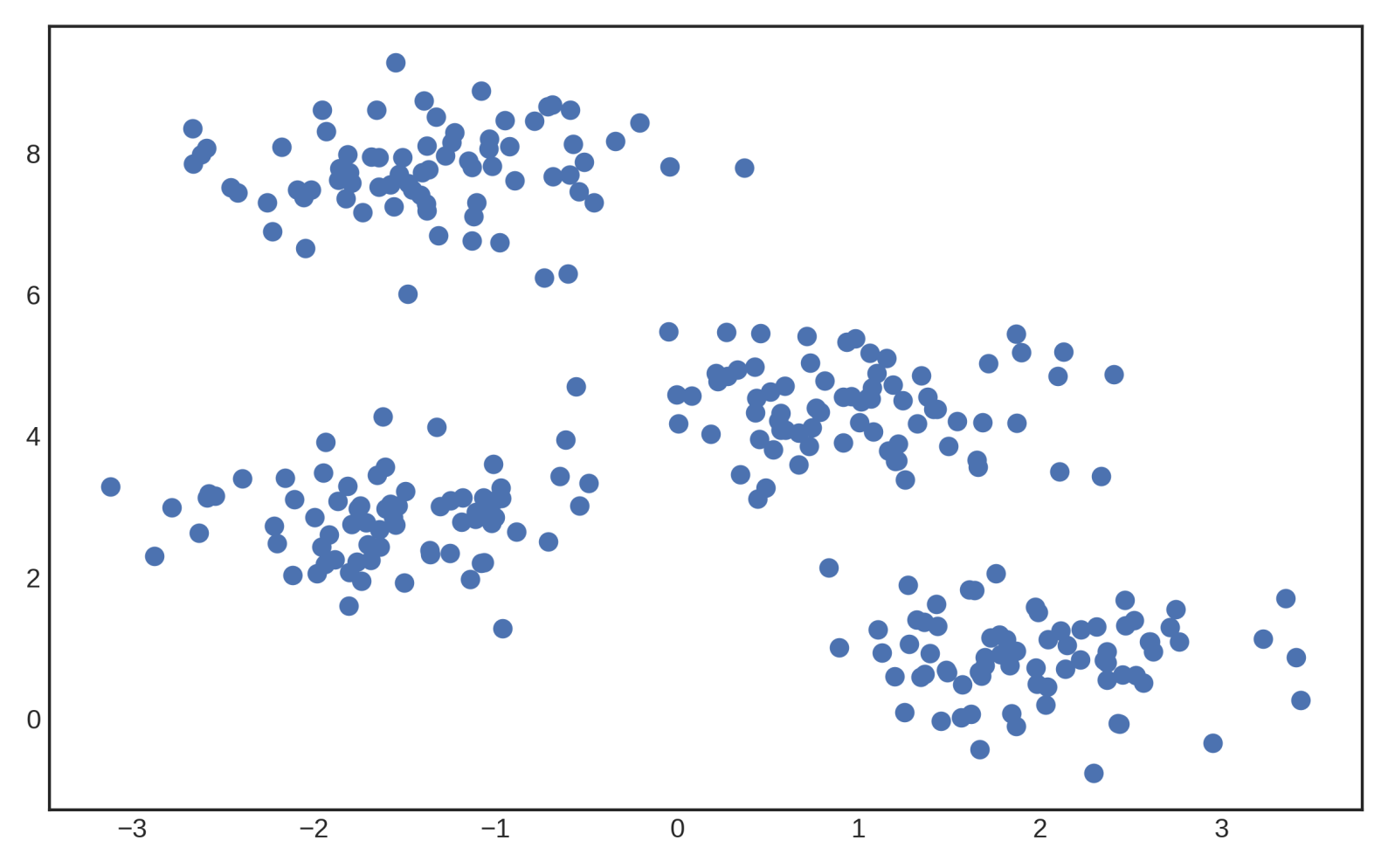
The partitioning corresponds to an **optimization problem** which consists in:

- partitioning the data into k clusters of equal variance
- so that it minimizes the within-cluster sum-of-squares [**inertia**]:

$$\sum_{i=0}^{k} \min_{\mu_j}(||x_i - \mu_j||^2)$$

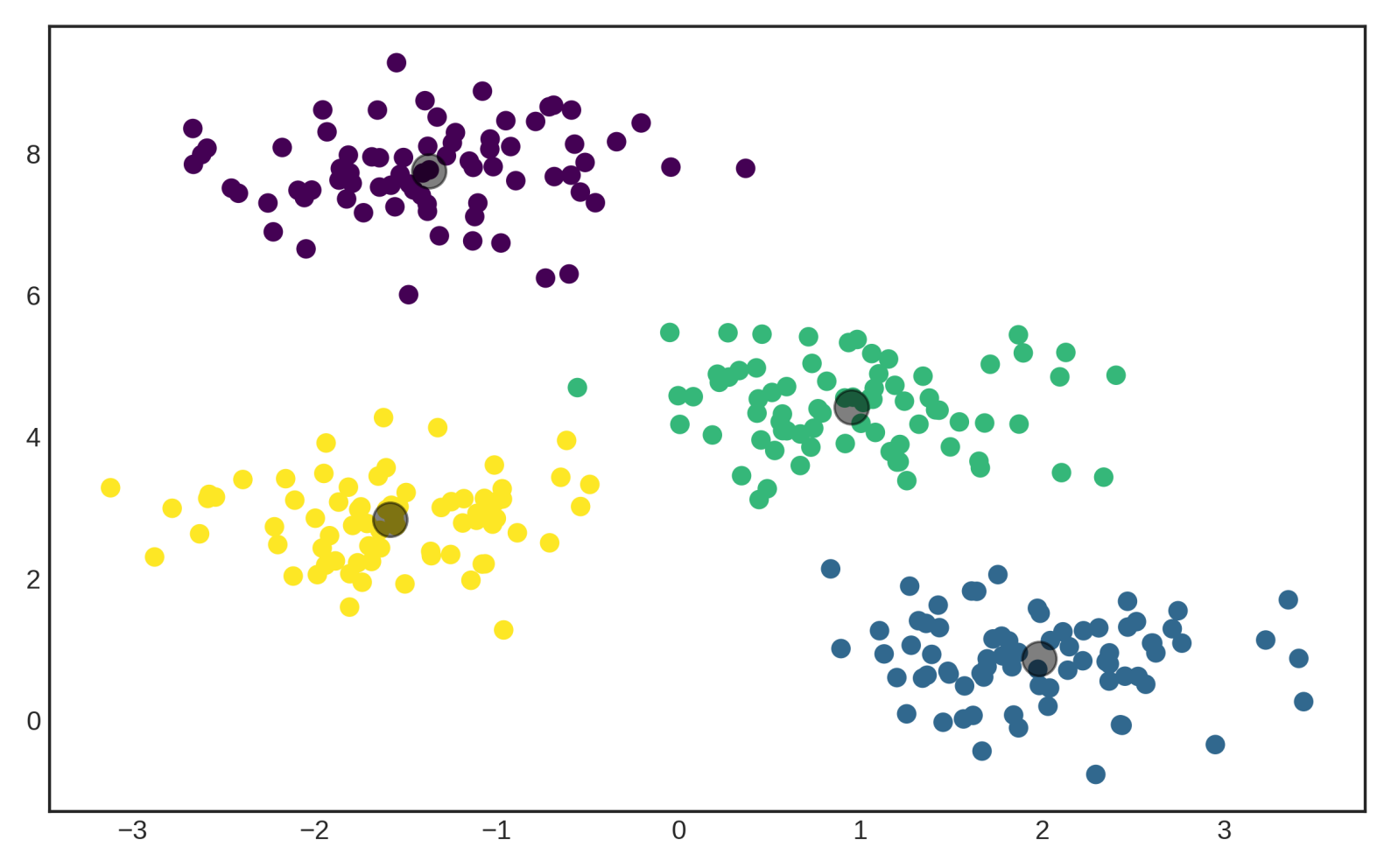
- Each cluster is represented by the **central vector** or centroïd μ_j

K-means clustering



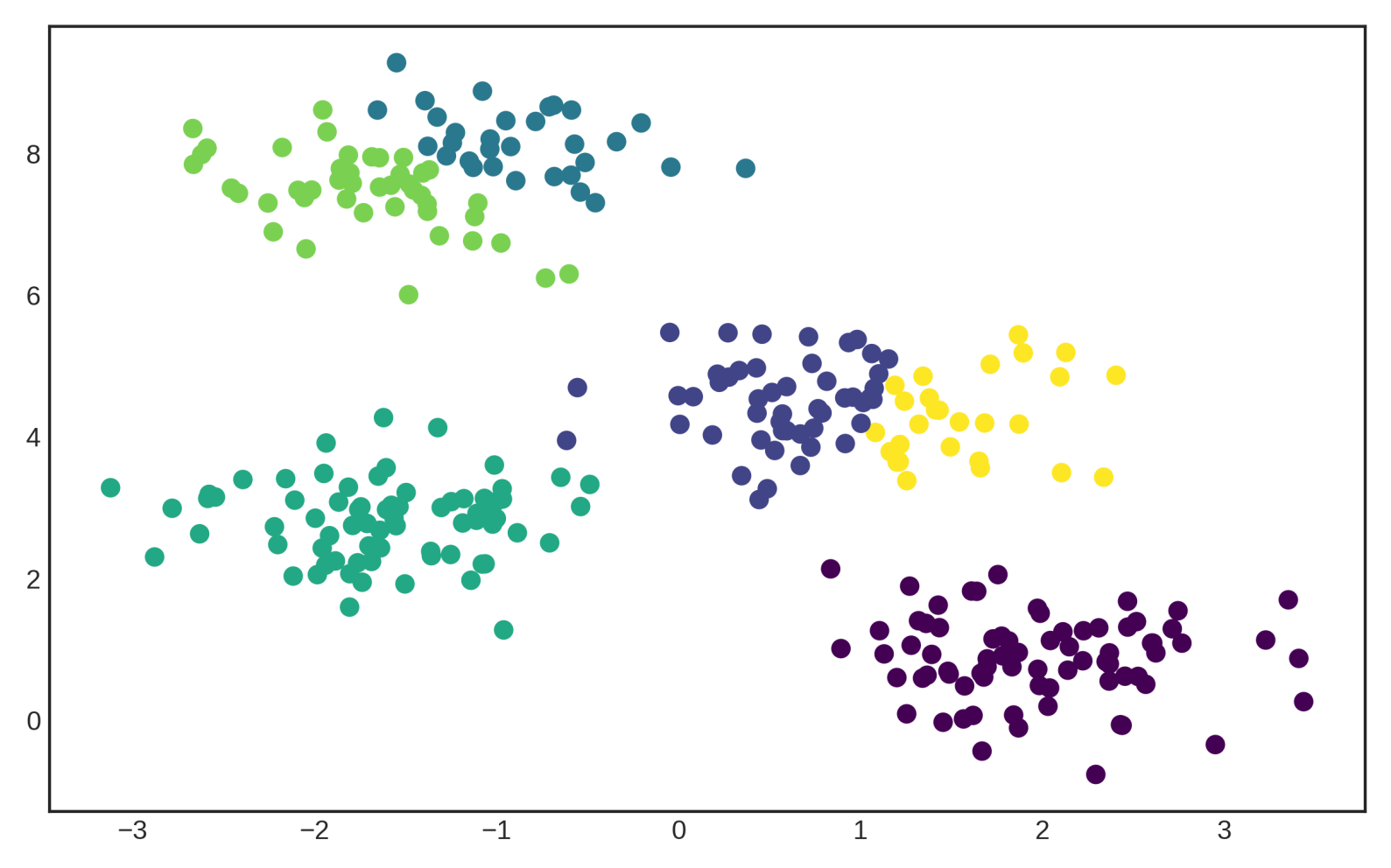
Simulated data

K-means clustering



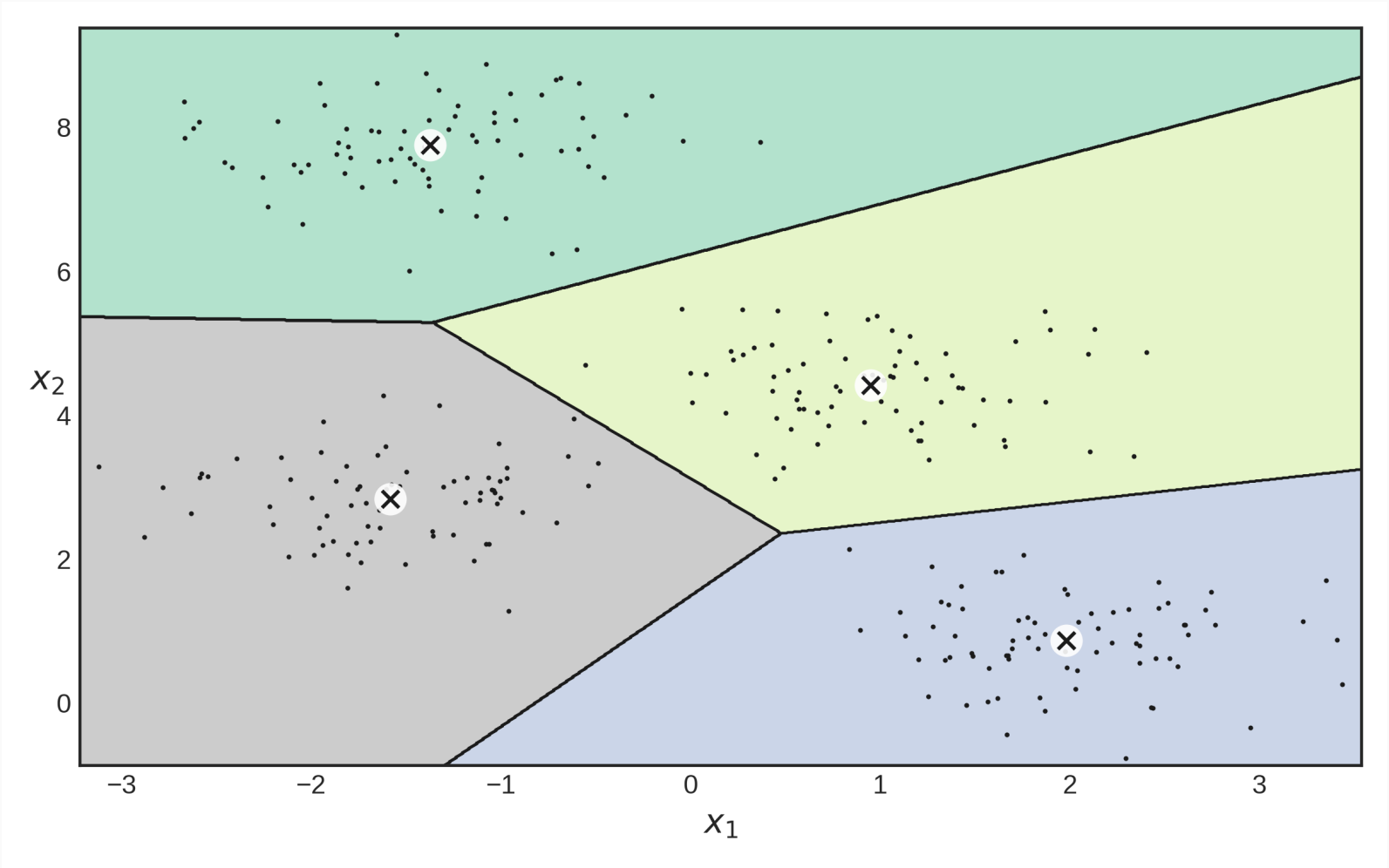
4 clusters and their centroïds

K-means clustering



6 clusters

K-means clustering



Decision boundaries for 4 clusters

K-means algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

K-means algorithm

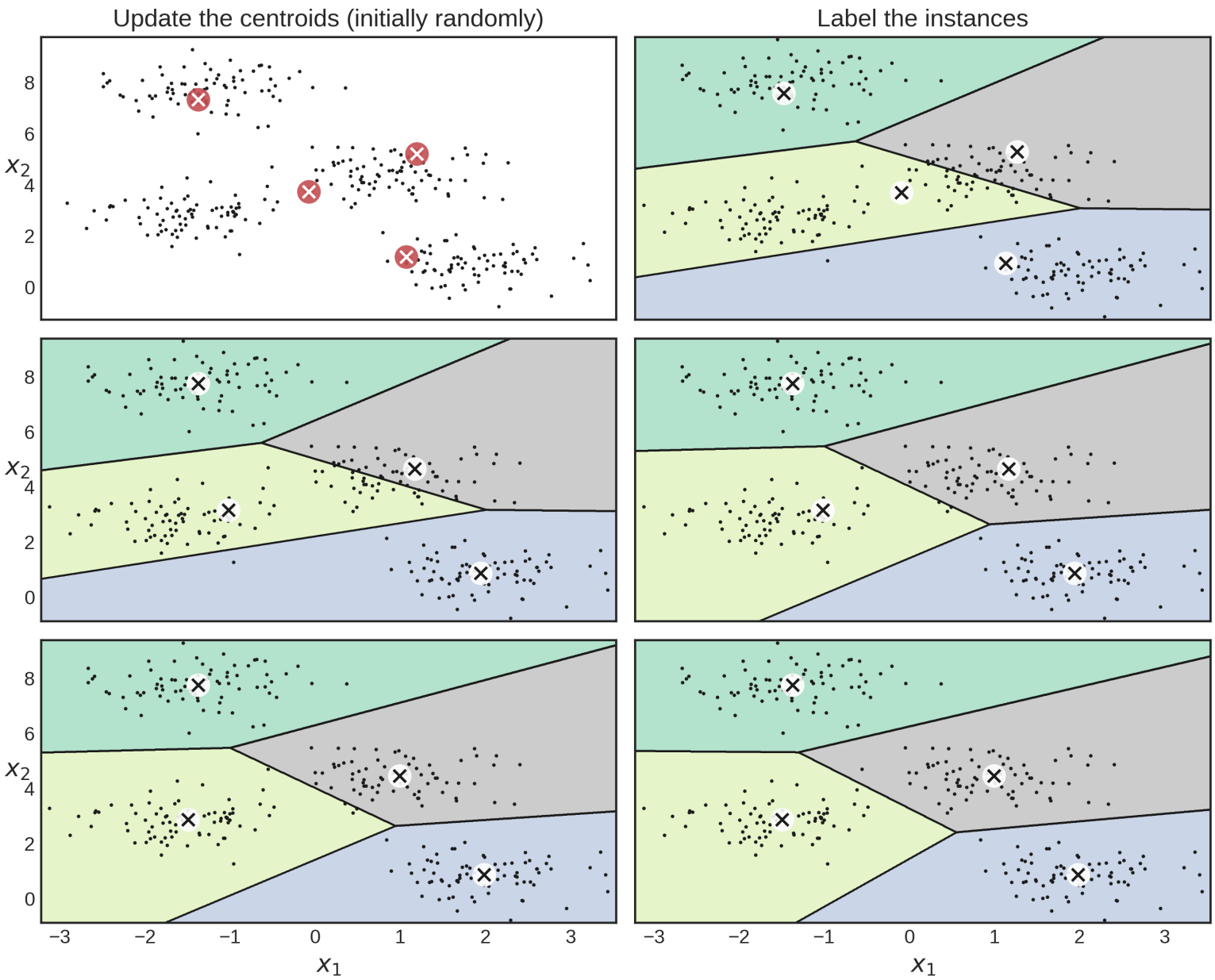
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   - 2.1 For each of the K clusters, compute the cluster **centroid**. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   - 2.2 Assign each observation to the cluster whose centroid is closest (where **closest** is defined using Euclidean distance)

K-means algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   - 2.1 For each of the K clusters, compute the cluster **centroid**. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   - 2.2 Assign each observation to the cluster whose centroid is closest (where **closest** is defined using Euclidean distance)

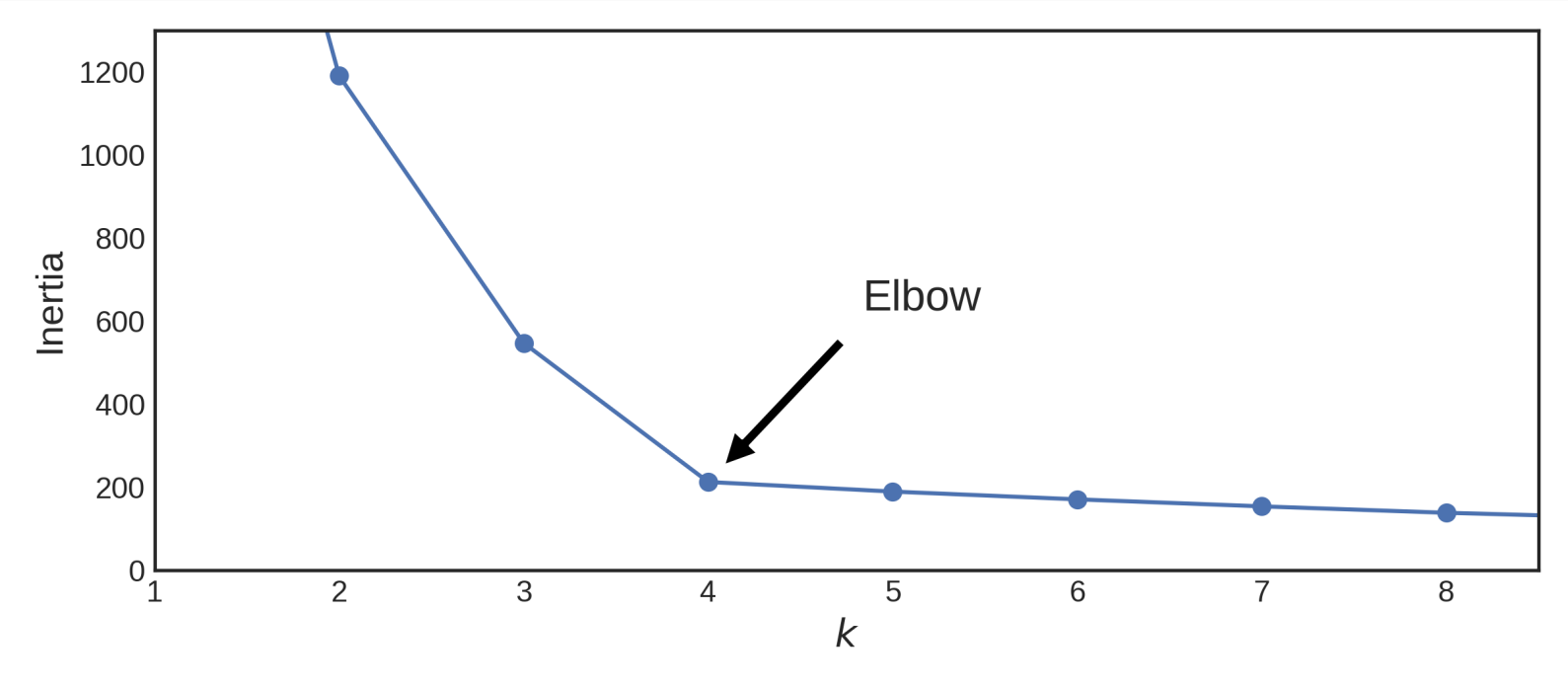
$\rightarrow$ The algorithm aims to choose centroids that minimise the inertia, or within-cluster sum-of-squares criterion

K-means algorithm: example



Finding the optimal number of clusters

- Most of the time, the number of clusters does not stand out from looking at the data
- Why not picking the model with the lowest inertia?
- Because inertia decreases with the number of clusters
- Rule of thumb: choose the number of clusters at the "elbow"

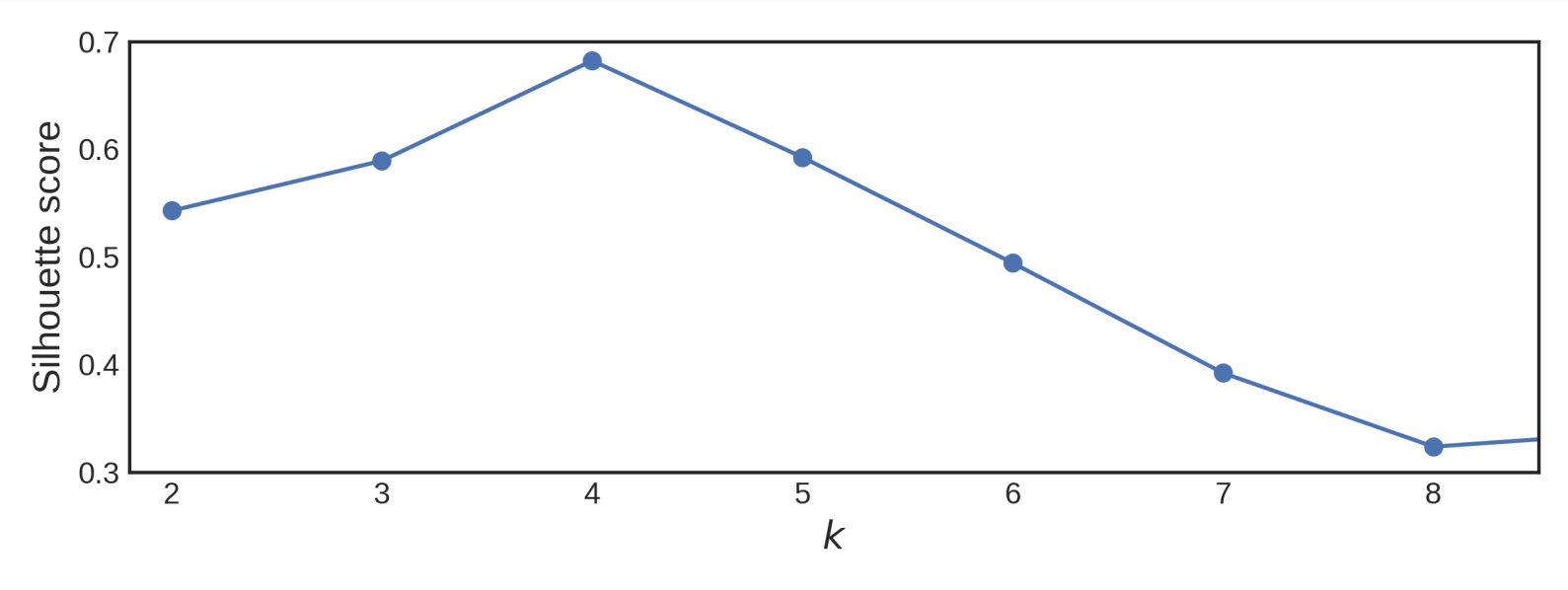


Finding the optimal number of clusters

- Can pick the optimal number of clusters with the **silhouette score**:

$$\frac{b_i - a_i}{max(a_i, b_i)}$$

  - a_i mean distance to members of i's cluster
  - b_i mean distance to members of i's second-closest cluster



Other clustering algorithms

- **DBSCAN** defines clusters as continuous regions of high density
  - Works well if all the clusters are dense enough and if they are very well separated by low-density regions
  - Detects and excludes outliers automatically

Other clustering algorithms

- **DBSCAN** defines clusters as continuous regions of high density
  - Works well if all the clusters are dense enough and if they are very well separated by low-density regions
  - Detects and excludes outliers automatically
- Agglomerative **hierarchical clustering** makes nested clusters:

  - we do not know in advance how many clusters we want
  - tree-like visual representation of the observations, called a dendrogram,

  $\rightarrow$ show at once the clusterings obtained for each possible number of clusters, from 1 to n.

Outline

Topic Models in Social Science

- Topic models developed in computer science and statistics:
  - summarize unstructured text using words within document
  - useful for dimension reduction:
    - rather than representing a text T in its feature space as $\{word_i : count(word_i, T) \text{ for } word_i \text{ in } Vocabulary\}$,
    - you can represent it in a topic space as $\{Topic_i : Weight(Topic_i, T) \text{ for } Topic_i \text{ in } Topics\}$
- Social scientists use topics as a form of measurement
  - how observed covariates drive trends in language
  - tell a story not just about what, but how and why
  - topic models are more interpretable than other methods, e.g. principal components analysis.

Latent Dirichlet Allocation (LDA)

- Idea: documents exhibit each topic in some proportion.
  - Each document is a distribution over topics.
  - Each topic is a distribution over words.
- **Latent Dirichlet Allocation** (e.g. Blei 2012) is the most poular topic model in this vein because it is easy to use and (usually) provides great results.
  - Maintained assumptions: Bag of words/phrases, fix number of topics ex ante.

Latent Dirichlet Allocation (LDA)

- A document is a **mixture of topics** and topics are probability distributions over tokens in the vocabulary.
- The (normalized) frequency of word j in document i can be written as:

$$q_{ij} = v_{i1} * \theta_{1j} + v_{i2} * \theta_{2j} + ... + v_{iK} * \theta_{Kj}$$

  - K: the total number of topics
  - θ_{kj}: the probability that word j shows up in topic k
  - v_{ik}: the weight assigned to topic k in document i.
- The model treats v and θ as generated from Dirichlet-distributed priors and can be estimated through Maximum Likelihood or Bayesian methods.

LDA in practice

- Topic modeling is not really appropriate for very short texts
- The input for the topic model is a **Document-term matrix** with either counts or TF-IDF scores:
  - Try both
- `sklearn` has a `LatentDirichletAllocation` function (doc)
- Hyper parameters of LDA:
  - $\alpha =$ document-topic density
    higher $\alpha \rightarrow$ documents are made up of more topics $\rightarrow$ more specific topic distribution per document.
  - $\beta =$ topic-word density
    higher $\beta \rightarrow$ topics are made up of more words $\rightarrow$ more specific word distribution per topic.
- Choosing the number of topics

LDA implementation

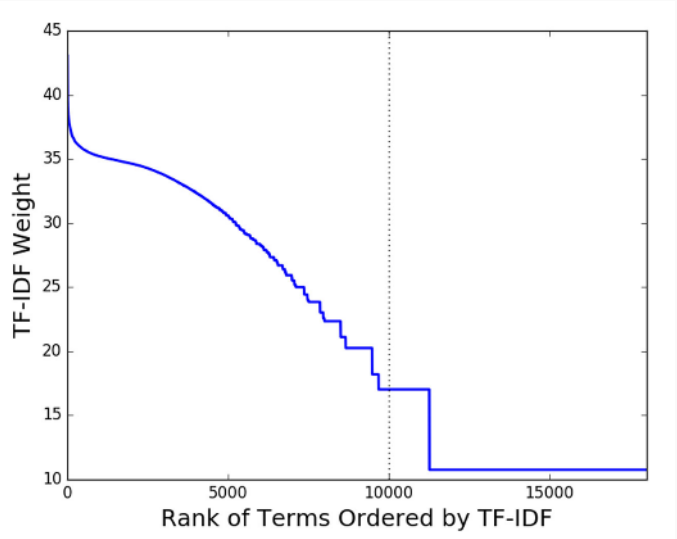
1. Loading and cleaning data
2. Text pre-processing
    - remove any punctuation, lowercase the text, remove stopwords
    - build Document-term matrix (counts or TF-IDF score)
3. Model Evaluation and parameter tuning
    - LDA model tranining with a defined number of topics
    - Analyzing LDA model with a visualization of the most common words by topics

Topic modeling Federal Reserve Bank transcripts

- Use LDA to analyze speech at the FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - transcripts: 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.

Topic modeling Federal Reserve Bank transcripts

- Use LDA to analyze speech at the FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - transcripts: 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.
- Pre-processing:
  - drop stopwords, stems, etc.
  - Drop words with low TF-IDF weight



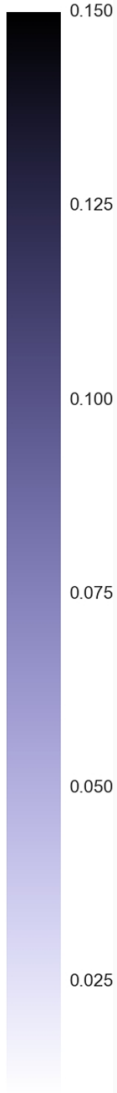
LDA Training

- $K = 40$ topics selected for interpretability / topic coherence.
  - the "statistically optimal" $K = 70$, but these were less interpretable.

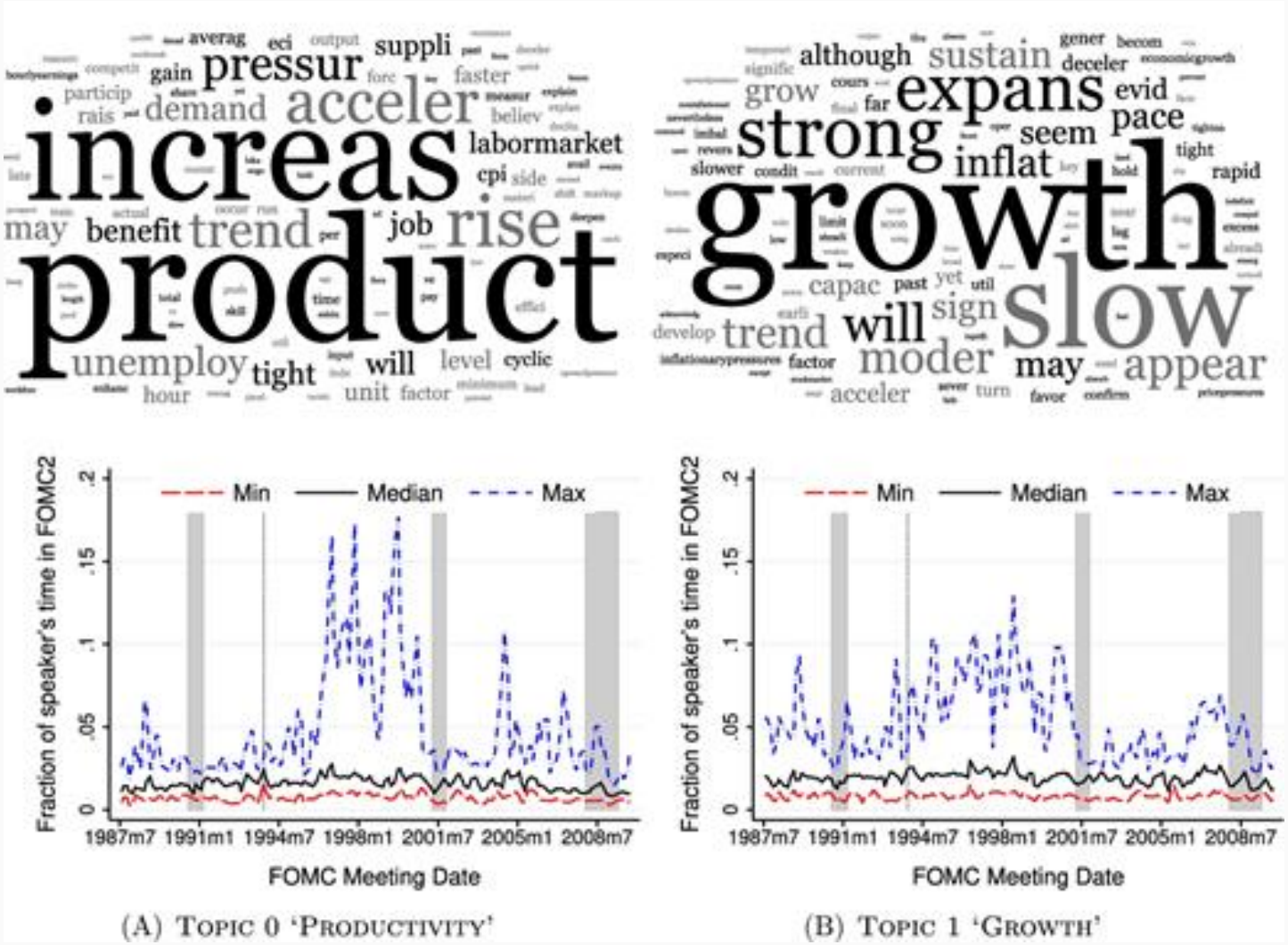
LDA Training

- $K = 40$ topics selected for interpretability / topic coherence.
  - the "statistically optimal" $K = 70$, but these were less interpretable.
- hyperparemeters $\alpha = 50/K$ and $\eta = .025$ to promote sparse word distributions (and more interpretable topics).

													Pro-cyclicality
Topic0[1]	product	increas	wage	price	cost	labor	rise	acceler	inflat	pressur	trend	compens	0.024
Topic1[1,2]	growth	slow	economi	continu	expans	strong	trend	inflat	will	recent	slowdown	moder	0.023
Topic2[2]	inflat	expect	core	measur	higher	path	slack	gradual	continu	remain	view	suggest	0.017
Topic3[1]	percent	year	quarter	growth	month	rate	last	next	state	averag	california	employ	0.007
Topic4	number	data	look	chang	measur	use	point	show	revis	estim	gdp	actual	0.007
Topic5[1,2]	polici	inflat	monetarpol	need	time	can	monetari	move	tighten	view	action	believ	0.005
Topic6[2]	rate	term	expect	real	lower	increas	rise	level	declin	short	nomin	year	0.005
Topic7	statement	word	chang	meet	languag	discuss	issu	want	read	sentenc	view	use	0.005
Topic8[2]	chairman	support	mr	direct	recommend	agre	asymmetr	prefer	symmetr	move	toward	favor	0.004
Topic9[1]	employ	continu	growth	job	nation	region	seem	state	manufactur	greenbook	busi	bit	0.004
Topic10	dollar	unitedstates	export	countri	import	foreign	japan	growth	abroad	trade	develop	currenc	0.003
Topic11	model	use	simul	shock	effect	scenario	nairu	differ	rule	chang	baselin	altern	0.003
Topic12[2]	risk	may	balanc	seem	side	uncertainti	possibl	economi	probabl	reason	upsid	much	0.003
Topic13	forecast	greenbook	staff	project	differ	assumpt	littl	assum	somewhat	lower	end	period	0.002
Topic14	period	committe	consist	econom	run	maintain	futur	read	slightli	stabil	expect	develop	0.002
Topic15	invest	incom	spend	capit	household	consum	busi	hous	consumpt	sector	stock	stockmarket	0.002
Topic16[1]	month	report	increas	survey	expect	indic	remain	continu	last	recent	data	activ	0.002
Topic17[1]	project	forecast	year	quarter	expect	will	percent	revis	anticip	growth	next	recent	0.002
Topic18	question	ask	issu	let	want	answer	rais	discuss	don	start	without	okay	0.001
Topic19	peopl	talk	lot	much	comment	around	differ	number	realli	look	thing	hear	0.001
Topic20	presid	ye	governor	parri	stern	vice	hoenig	minehan	kelley	jordan	moskow	mcteer	0.001
Topic21	move	can	evid	signific	stage	inde	will	issu	economi	may	quit	clearli	0.001
Topic22[2]	chairman	thank	mr	time	meet	laughter	comment	let	will	point	call	may	0.0
Topic23[1]	year	panel	line	shown	right	chart	expect	project	percent	middl	left	next	0.0
Topic24	district	nation	area	continu	sector	construct	manufactur	report	activ	region	economi	remain	0.0
Topic25	know	someth	happen	right	thing	want	look	sure	can	realli	anyth	els	0.0
Topic26[1,2]	polici	might	committe	market	may	tighten	eas	risk	action	staff	possibl	potenti	-0.001
Topic27	year	continu	product	price	level	industri	will	sale	increas	auto	last	district	-0.001
Topic28[1]	inventori	product	sale	level	order	will	sector	come	good	quarter	much	adjust	-0.001
Topic29	price	oil	increas	energi	effect	import	suppli	product	demand	will	market	oilprices	-0.002
Topic30	term	might	point	can	sens	run	short	probabl	time	longer	tri	someth	-0.002
Topic31	seem	may	time	certainli	bit	littl	quit	much	far	perhap	better	might	-0.003
Topic32	money	aggreg	borrow	seem	rang	reserv	rate	target	time	altern	suggest	million	-0.003
Topic33[2]	move	market	point	will	fundsrate	rate	basispoints	need	fed	today	basi	time	-0.004
Topic34[1]	report	busi	compani	year	contact	firm	sale	worker	expect	plan	director	industri	-0.004
Topic35	will	fiscal	ta	budget	cut	govern	effect	billion	state	spend	deficit	year	-0.005
Topic36	will	economi	world	rather	problem	believ	can	situat	much	seem	view	good	-0.008
Topic37	realli	look	side	thing	lot	problem	concern	littl	pretti	situat	kind	much	-0.012
Topic38	bank	credit	market	loan	financi	debt	lend	fund	concern	financ	problem	spread	-0.018
Topic39[1,2]	economi	weak	recoveri	recess	confid	eas	neg	econom	will	turn	declin	period	-0.059



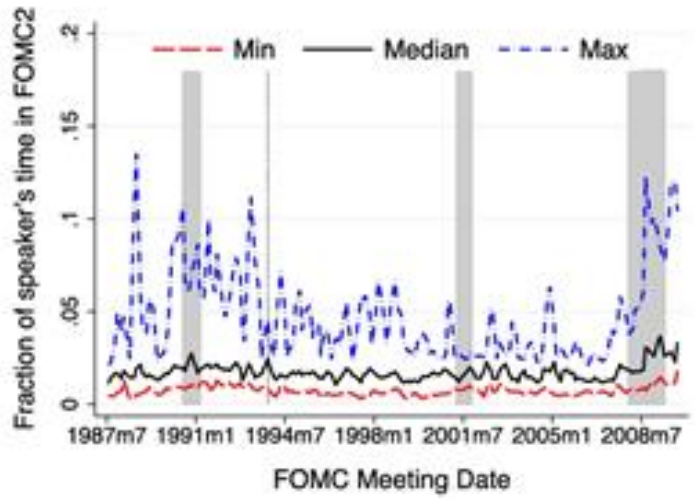
Pro-Cyclical Topics



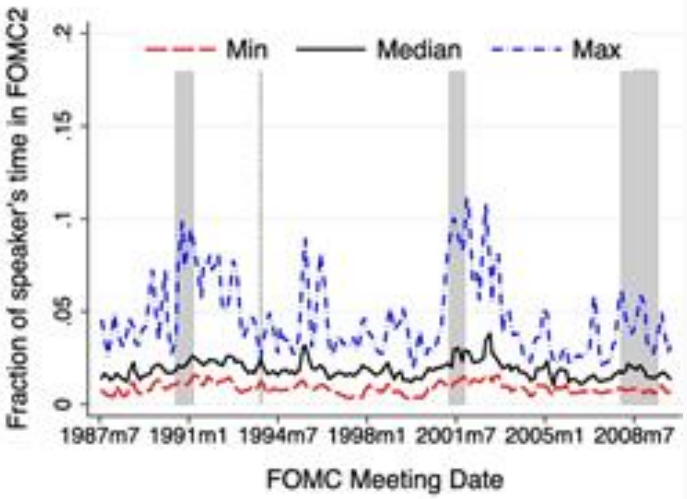
(A) Topic 0 'Productivity'

(B) Topic 1 'Growth'

Counter-Cyclical Topics

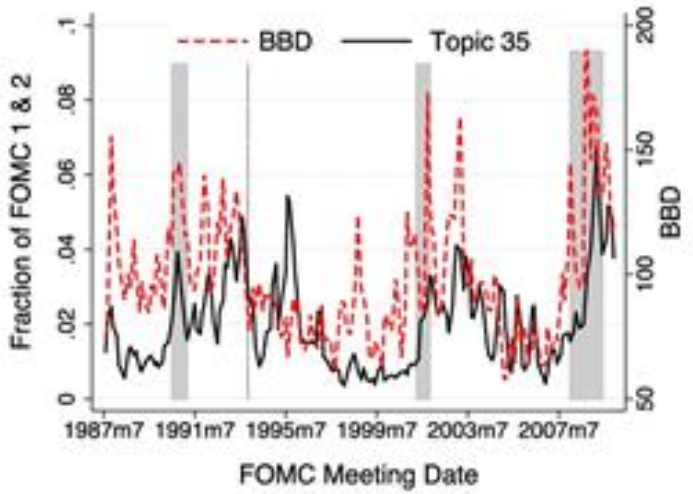


(A) Topic 38 'Financial sector'

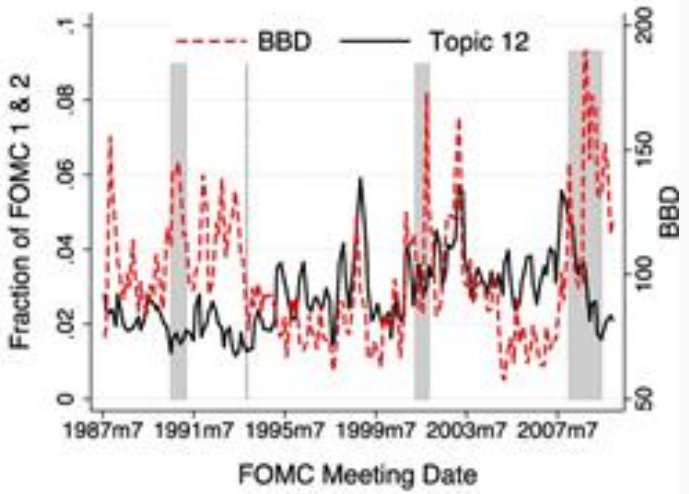


(B) Topic 39 'Economic weakness'

FOMC Topics and Policy Uncertainty



(A) Topic 35 'Fiscal issues'



(B) Topic 12 'Risk'

Effect of Transparency

- In 1993, there was an unexpected transparency shock where transcripts became public.

Effect of Transparency

- In 1993, there was an unexpected transparency shock where transcripts became public.
- Increasing transparency results in:
  - higher discipline / technocratic language (probably beneficial)
  - higher conformity (probably costly)
- Highlights tradeoffs from transparency in bureaucratic organizations.

Clusters vs. Topics

- The number of topics, like the number of clusters, is an output parameter.
- Each cluster is a set of documents that are close to each other in the vector space (normally, they will be topically related)
- With topic modeling, we build clusters of words rather than clusters of texts. A text is thus a mixture of all the topics, each having a specific weight
- The advantage of clusters, rather than topics or embeddings, is that they provide discrete groups.
    - This might be useful depending on your research task.