Advances in independent component analysis with applications to data mining

Ella Bingham

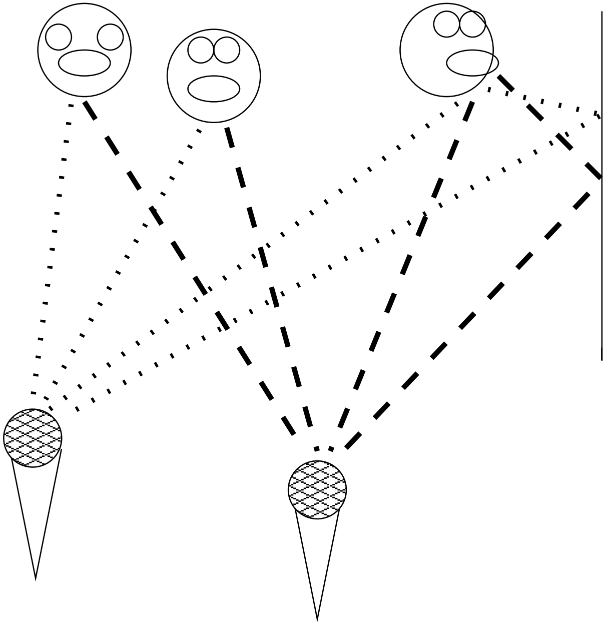
⇐ ← → ⇒ ■ ? ✕

Problem setting

- There are so much of data that we need efficient methods to process it.
- Data: measurements or observations that can be presented in numerical format. In this thesis, the data are presented in a tabular form.
- We assume that the data consists of unknown components and we wish to find out what they are.

- The components often give a simpler and clearer view of the data:
    - a large text document corpus contains several topics, and listing the topics gives an overview of the corpus
    - at a cocktail party one observes a mixture of several speakers; now the components are the voices of individual speakers
- The components are independent of each other, hence the name “independent component analysis”.
- Next I will present an advance to the method.

Analysis of complex valued signals



- Human voices arrive at the measurement locations (microphones) with different delays and might be reflected from nearby walls. The same applies to mobile phone signals arriving at base stations.
- We transform the signals into the complex domain which makes the delays easier to work with.
- We then solve the independent components from the complex valued measurements.

Data mining

- Data are automatically gathered into data bases. The data are not originally aimed for analysis purposes.
- Analyzing the data may give us information about the phenomenon that has created the data.
- Data mining is a broad set of mathematical, statistical and data base methods for efficient analysis of large data sets.

- An example of the application of independent component analysis to data mining: Dynamical textual discussion.
- In a chat line discussion in the Internet there are several discussions going on simultaneously, and the topics of discussion change in time.

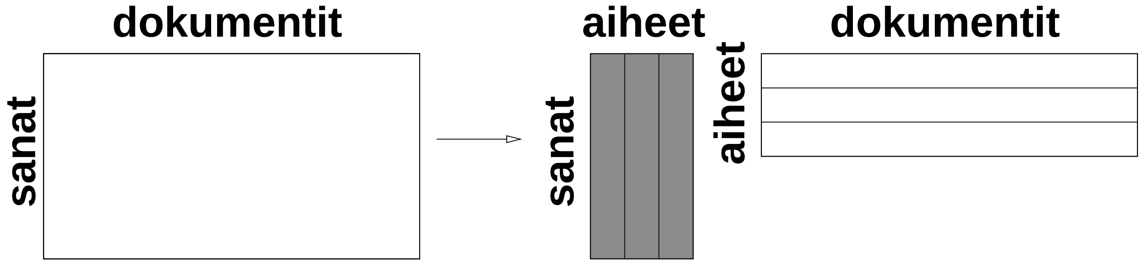
```
<Haley-CNN>  Gallup Poll’s  Newport will join the #CNN_Newsroom
at 10:00 a.m. EST to discuss:  the public views the Bush transition
<TomM--_> Fatboy....smart!!!
<Gary> ya bush at 60% and Gore at 78% still
<Nailheads> ........I thought not.
<Whitikau-CNN> jeter: Does that mean that Bush will need to work harder
than previous Presidents to gain the respect and trust of the people?
<alfie> jeter, according to Louis Lapham, neither candidate was worth
dooky.
<DUFUS> That’s because when asked why they voted for him.....the answer
was: "He is a good old boy."
<DUFUS> Silvertone........HELLO?
<Nailheads> ....ha ha. if that’s all you got...b.s....then okay.  sling away.
<stella> DUFUS MAKE SURE YOU HAVE THE RIGHT FACT OK BEFORE YOU BLAME HIM
EVERY TIME YOU BE IN THIS CHANNEL
<Moro^^^> daddy Bush promised Powell Sec of State...if he sat this election
out. This is a grand conspiracy head by the elder Bush. It is his second
administration.
<jeter> Nailheads, 60%...LOLOLOLOL!!! Are you nuts? Maybe that was a question
on him personally. But on handling the job...thus far, it’s less than 50%.
```

- The discussion is split into short pieces, each of which is considered a document.

- The document corpus is expressed in numerical format:

  - form a vocabulary consisting of the most important terms in the corpus
  - at each document, list the frequencies of the terms

- Then solve the independent components from the term by document matrix. As a result we get two matrices, a term by topic matrix and a topic by document matrix.

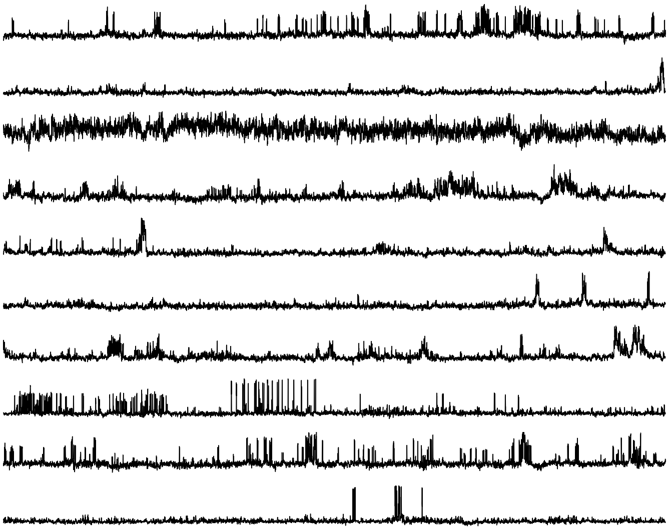


In the following we show the term by topic matrix as list of keywords, and the topic by document matrix as signals in time.

Topic 1 deals with Jesse Jackson and his illegitimate child, topic 2 is about parental control over children's web usage and topic 3 is a general discussion about G.W. Bush. Topic 4 is a religious discussion, topic 5 deals with problems of the youth such as violence and drug abuse, and topic 6 is about the controversial flag of the state of Georgia, US, due to which the NCAA basketball games risked cancellation in Atlanta. Topic 7 involves the energy shortage in California, topic 8 corresponds to comments given by the chat line moderator, topic 9 is about taxation and topic 10 is a short discussion dealing with the values of the politicians in the US

Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
jackson	site	bush	religion	violenc
sharpton	web	ashcroft	god	report
child	net	vote	jesu	youth
stori	word	kennedi	bibl	children
drudg	parent	presid	religi	gun
rainbow	nanni	cnn	life	point
monei	internet	time	follow	home
mistress	block	gore	read	drug
coalition	kid	question	stori	famili
tonight	system	elect	univers	satcher
pregnant	access	god	exist	health
affair	child	senat	faith	risk
black	base	power	man	factor
chenei	chat	thing	book	surgeon
jessi	page	fact	earth	prevent

Aihe 6	Aihe 7	Aihe 8	Aihe 9	Aihe 10
flag	california	join	tax	free
move	power	discuss	cut	liber
citi	electr	est	exempt	opinion
ncaa	energi	tonight	monei	religion
offici	blackout	room	gop	form
atlanta	state	studio	hous	polit
count	deregul	cnn	congress	conserv
game	compani	conserv	pay	birth
night	crisi	american	interest	philosophi
georgia	price	nea	recess	establish
chang	plant	union	payer	narrow
lose	util	keen	secur	restrict
confeder	order	type	henri	independ
hehe	home	chat	hypocrit	orthodox
chenei	cost	newsroom	hyde	bound



The horizontal axis shows the time (24 hours).
The uppermost signal corresponds to topic 1, the second signal to topic 2 and so on.
The signal has a peak when the topic is being discussed.

We notice that different topics are active at different points in time.

Why did we find these topics? Because different topics of discussion are statistically independent of each other.

We might have looked for a different number than 10 topics — often there is no “correct” number of topics or independent components in real world data.

Please note that we are not just clustering the terms, but instead a term may belong to several components, such as “power” in 3 and 7, “religion” in 4 and 10, “home” in 5 and 7. Also, at one point in time there may be several discussions going on simuultaneously. There are a lot of methods for clustering the variables (here the terms) but usually the variables only belong to one cluster. ICA is very different in this sense.

Thank you