Introduction Distributional Semantic Models

LOT School Winter 2019

Antske Fokkens

January 14th 2019 (all cited sources verified on this date)

Acknowledgements

- Baroni & Boleda. Distributional Semantic Models https://www.cs.utexas.edu/~mooney/cs388/slides/dist-sem-intro-NLP-class-UT.pdf
- Pia Sommerauer. What is in a word embedding vector?

Meaning = Use

Meaning = Use

- Marco saw a furry little **wampimuk** hiding in the tree

source: http://www.aclweb.org/anthology/P14-1132

from Lazaridou et al. (2014)

Meaning = Use

- Can meaning be deducted from text?

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

- a similar category as black coffee:
  a (hot) beverage?

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island.

- a similar category as black coffee:
  a (hot) beverage?
- different degrees of strength: mixed/drawn/brewed
- color can be used to describe a river: transparent, blue, green, brown tone

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island

X comes from the leaves that have been withered and dried immediately after picking

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island.

X comes from the leaves that have been withered and dried immediately after picking

- a similar category as black coffee:
  a (hot) beverage?
- different degrees of strength: mixed/drawn/brewed
- color can be used to describe a river: transparent, blue, green, brown tone
- made from dried leaves

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island

X comes from the leaves that have been withered and dried immediately after picking

It is a large leaf X with a very delicate flavor

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island.

X comes from the leaves that have been withered and dried immediately after picking

It is a large leaf X with a very delicate flavor

- a similar category as black coffee:
  a (hot) beverage?
- different degrees of strength: mixed/drawn/brewed
- color can be used to describe a river: transparent, blue, green, brown tone
- made from dried leaves
- can have delicate flavor; probably variations in flavor exist

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island

X comes from the leaves that have been withered and dried immediately after picking

It is a large leaf X with a very delicate flavor

It may be black or X flavored with jasmine flowers, is very fragrant and is always drunk without milk

adapted from Sommerauer

What do we know about X?

Whereas traditional politicians offer visitors X, the Reform of Heisei serves black coffee.

The river Neckinger, “the colour of strong X”, flowed round Jacob’s Island.

X comes from the leaves that have been withered and dried immediately after picking

It is a large leaf X with a very delicate flavor

It may be black or X flavored with jasmine flowers, is very fragrant and is always drunk without milk

- a similar category as black coffee:
  a (hot) beverage?
- different degrees of strength: mixed/drawn/brewed
- color can be used to describe a river: transparent, blue, green, brown tone
- made from dried leaves
- can have delicate flavor; probably variations in flavor exist
- there is something similar that is black

adapted from Sommerauer

What is X?

distributional hypothesis

- false?
- weak?
- strong?

Creating Embeddings

Creating Embeddings

- Basic idea:
  - words with similar meaning have similar context
  - represent the context a word occurs in as a vector (e.g. by counting how often it co-occurs with a specific term)
  - compare the vectors: similar vectors = similar meaning

Context of *dog*

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	++
park	+
owner	+
leash	+

from Baroni & Boleda

Context of *dog*

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	++
park	+
owner	+
leash	+

from Baroni & Boleda

Context of *dog*

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

from Baroni & Boleda

Context of *dog*

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	++
park	+
owner	+
leash	+

from Baroni & Boleda

Context of *dog*

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	++
park	+
owner	+
leash	+

from Baroni & Boleda

Context of *dog*

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	++
park	+
owner	+
leash	+

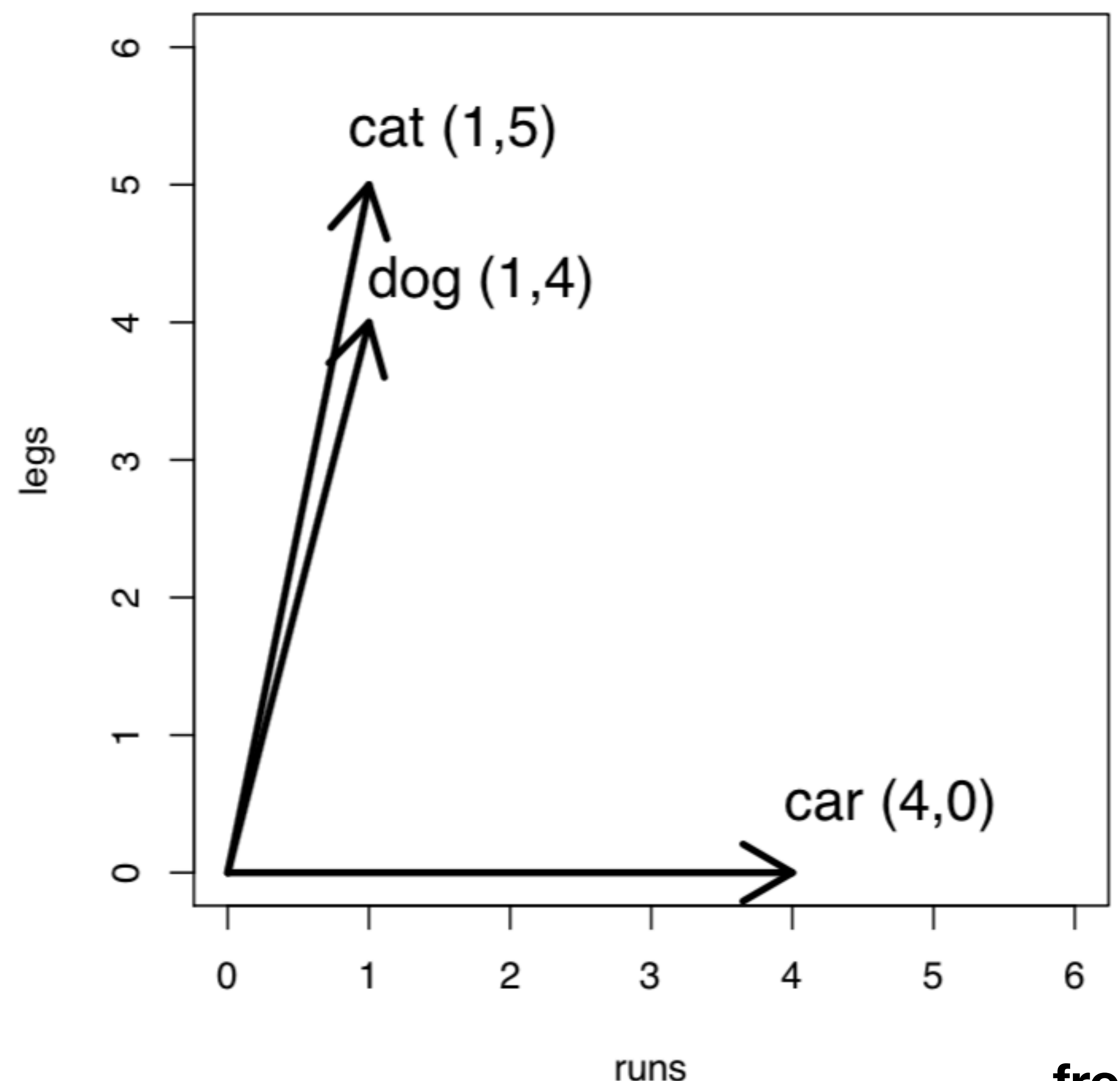
from Baroni & Boleda

Context representations

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
lion	0	3	2	0	1	0
light	0	0	0	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

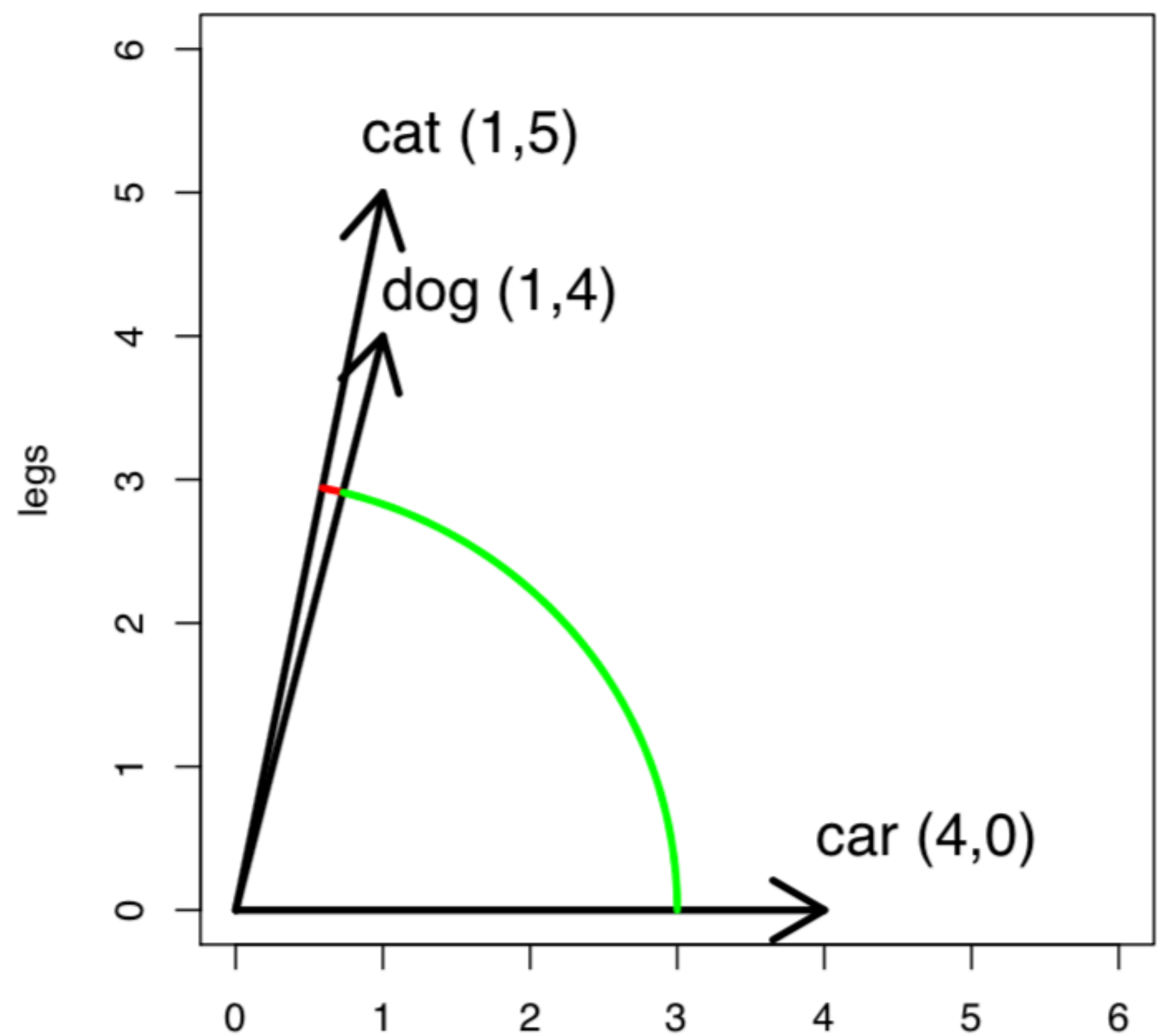
from Baroni & Boleda

Vector representation



from Baroni & Boleda

Vector representation



from Baroni & Boleda

Creating Embeddings

1. Preprocess data

2. Select contexts

3. Count contexts & transform counts

   or learn to predict them

   => Vector representations of meaning

What & How

- What to count/predict?

  => what context are we looking at?

- How to count/predict?

  => how are we using this context?

What?

- Possible contexts:
  - document
  - sentence
  - n-nearest words
  - syntactically related words

Selecting Context…

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Selecting context

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It’s midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Selecting context

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Selecting context

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Selecting context

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Selecting context

DOC1: The silhouette-n of the sun beyond a wide-open-a bay-n on the lake-n; the sun still glitter-v although evening-n has arrive-v in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Selecting context

DOC1: The silhouette-n of the sun beyond a wide-open bay on the lake; the sun still glitter-v although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Selecting context

DOC1: The silhouette-n_ppdep of the sun beyond a wide-open bay on the lake; the sun still glitter-v_subj although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

from Baroni & Boleda

Common selection criteria

- Window size
- Removing stop-words
- Filtering out low & high frequency terms

Selecting context

- What impact do you think context selection has?
  - window-size
  - syntactic restrictions
  - syntactic/pos-tag encoding
  - filtering low frequency terms
  - filtering high frequency terms

Most similar to *dog* (based on vector comparison)

2-word window

- cat
- horse
- fox
- pet
- rabbit
- pig
- animal
- mongrel
- sheep
- pigeon

30-word window

- kennel
- puppy
- pet
- bitch
- terrier
- rottweiler
- canine
- cat
- to bark
- Alsatian

from Baroni & Boleda

How?

- Count models: PPMI, SVD
- Predict models: word2vec, ELMO software packages

Preprocessing

- Minimum: tokenizing
- Frequently done:
    - remove punctuation/non-alphanumeric symbols
    - lowercase
    - low-frequency cut-off
    - stop-word removal
- Further analysis:
    - lemmatization, pos-tagging, dependency parsing

Counting

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
lion	0	3	2	0	1	0
light	0	0	0	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

adapted from Baroni & Boleda

Count models

- Just counting is not ideal:
  - frequent words have a lot of high numbers in their representations
  - frequent, not-so meaningful representations are overemphasized

Count models: PPMI

- Common solution to frequency problem:

  Pointwise mutual information:

$$log(\frac{P(w1,w2)}{P(w_1)P(w_2)})$$

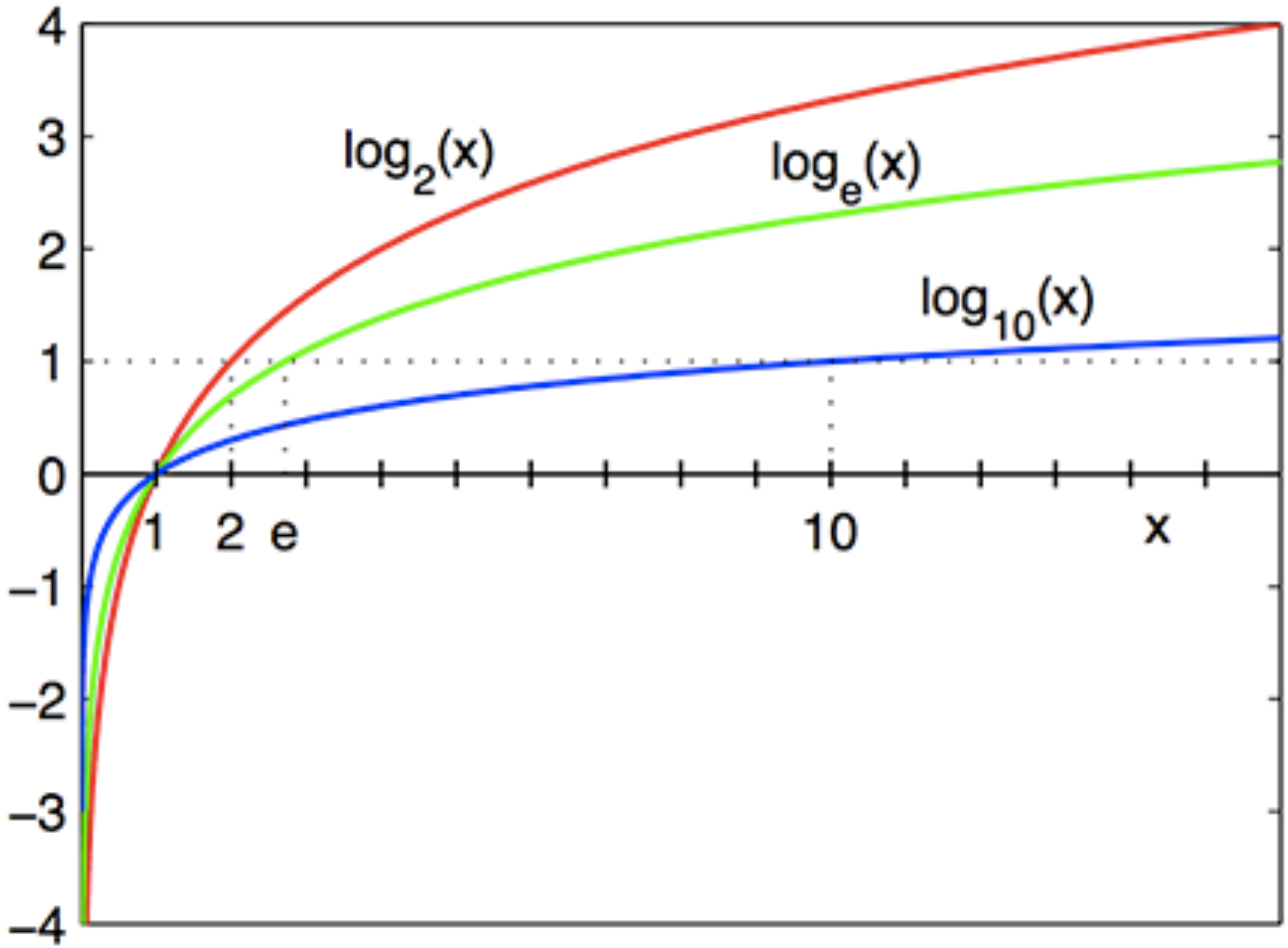
Count models: PPMI

Probability of w1 & w2 occurring together in corpus

$$log(\frac{P(w1,w2)}{P(w_1)P(w_2)})$$

Probability of w1 occurring with w2 by chance

plots of logarithms



from: https://commons.wikimedia.org/wiki/File:Logarithm_plots.png

PPMI

- Problem with PMI?

$$log(\frac{P(w1,w2)}{P(w_1)P(w_2)})$$

PPMI

- Problem with PMI?

$$log(\frac{P(w1,w2)}{P(w_1)P(w_2)})$$

=> zero co-occurrences?
=> low cooccurrences approximate minus-infinity

PPMI

- PPMI = $argmax(0, log(\frac{P(w1,w2)}{P(w1)P(w2)}))$

PPMI

- High dimensional (size of vocabulary)
- Low density (zeros for all context-words occurring less than by chance)
- Relatively high impact of low frequency words

Singular Value Decomposition (SVD)

- Method to reduce the number of dimensions:

  given a m x n matrix, construct a m x k matrix,
  where $k << n$

- Uses linear algebra to reduce the number of dimensions, preserving most of the variance of the original matrix

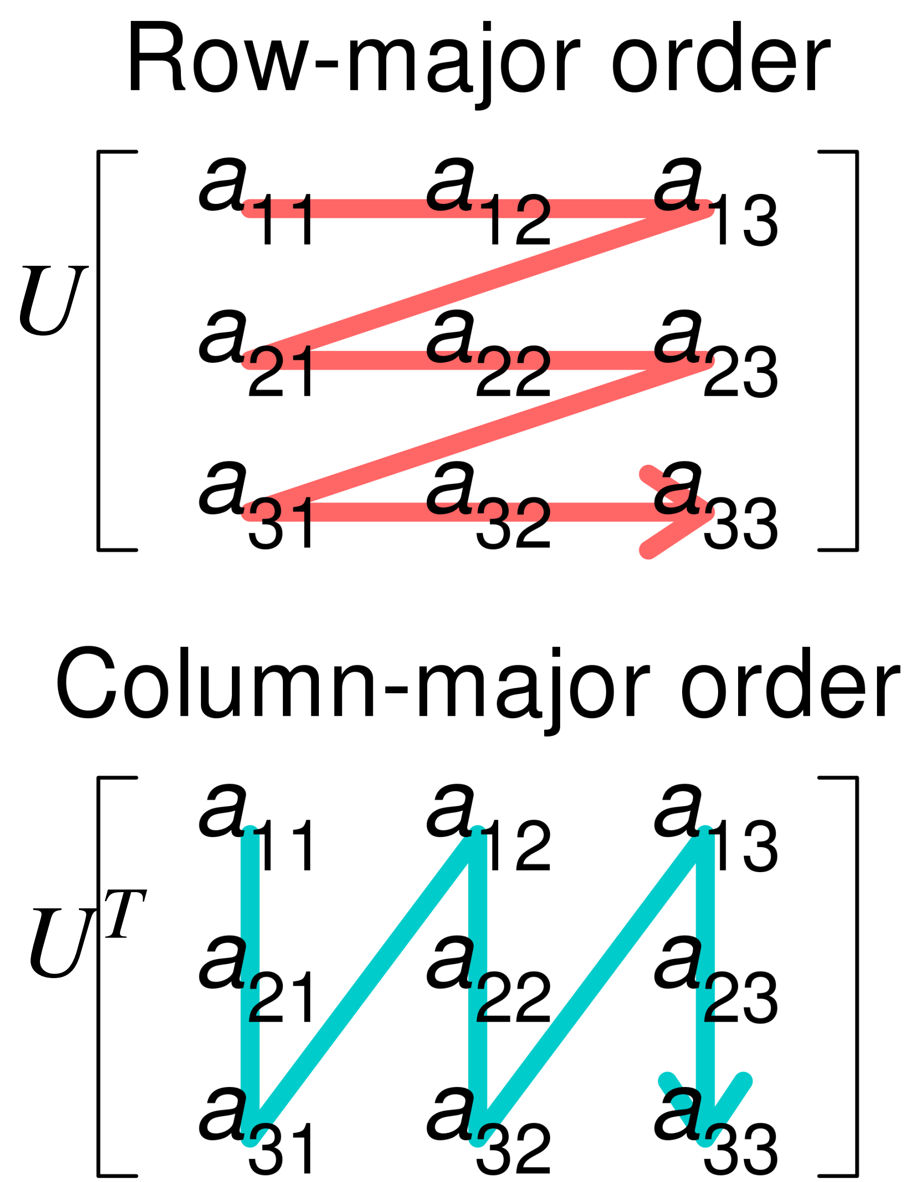
SVD

- A matrix *A* can be broken down (decomposed) into the product of three matrices:

$$A = U\Sigma V^T$$

- where *U* and *V* are orthogonal
- The columns of *U* are orthonormal eigenvectors of AA^T
- The columns of *V* are orthonormal eigenvectors of A^TA
- Σ is a diagonal matrix containing square roots of eigenvalues from *U* or *V* in descending order

U and V are orthogonal

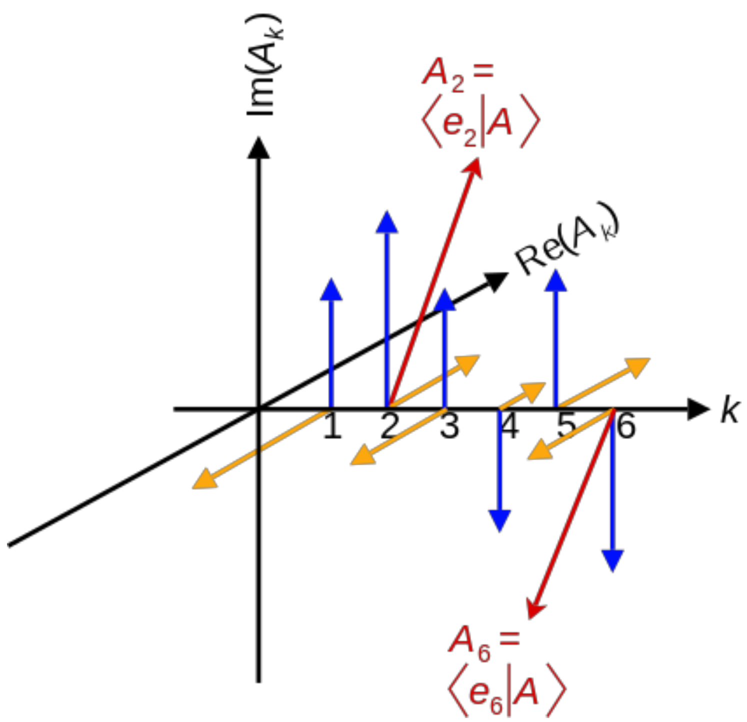


$$UU^T = I$$

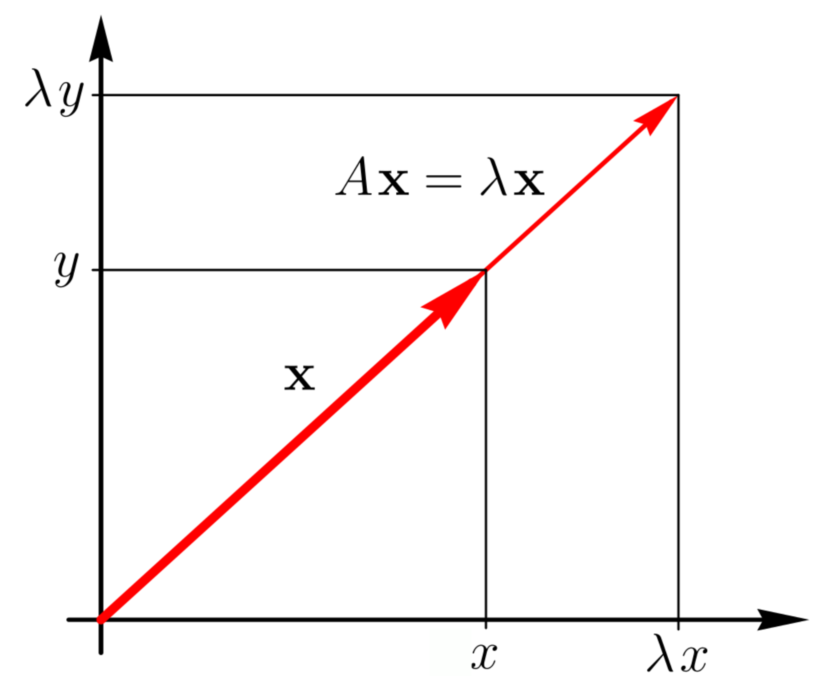
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Left image from: <u>https://en.wikipedia.org/wiki/Matrix_representation</u>

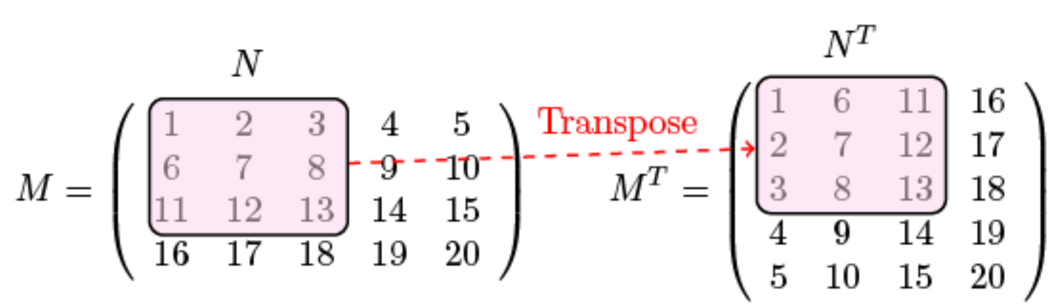
Columns of U are orthonormal eigenvectors of AA^T



source: https://en.wikipedia.org/wiki/File:Discrete_complex_vector_components.svg

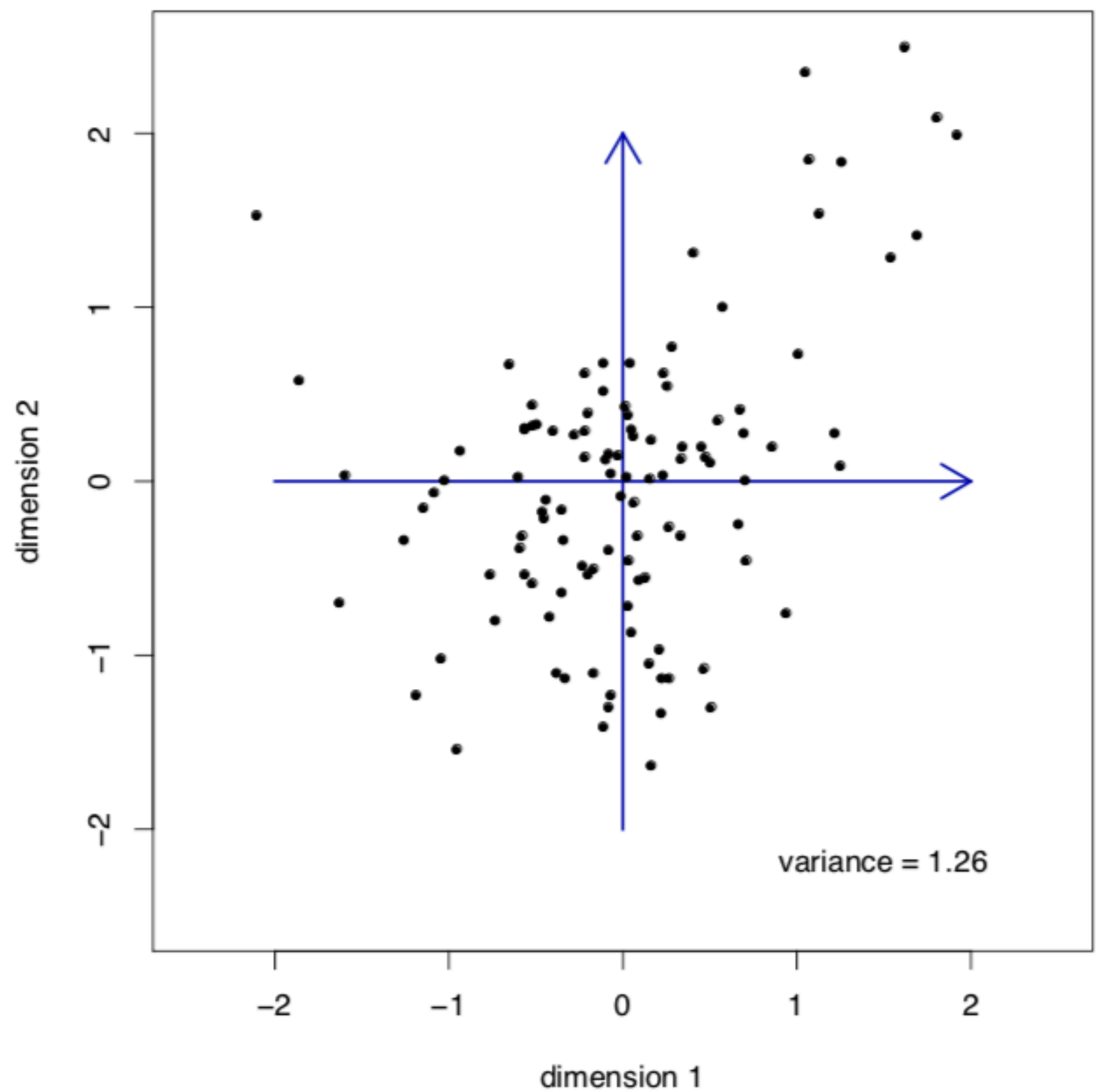


source: https://commons.wikimedia.org/wiki/File:Eigenvalue_equation.svg



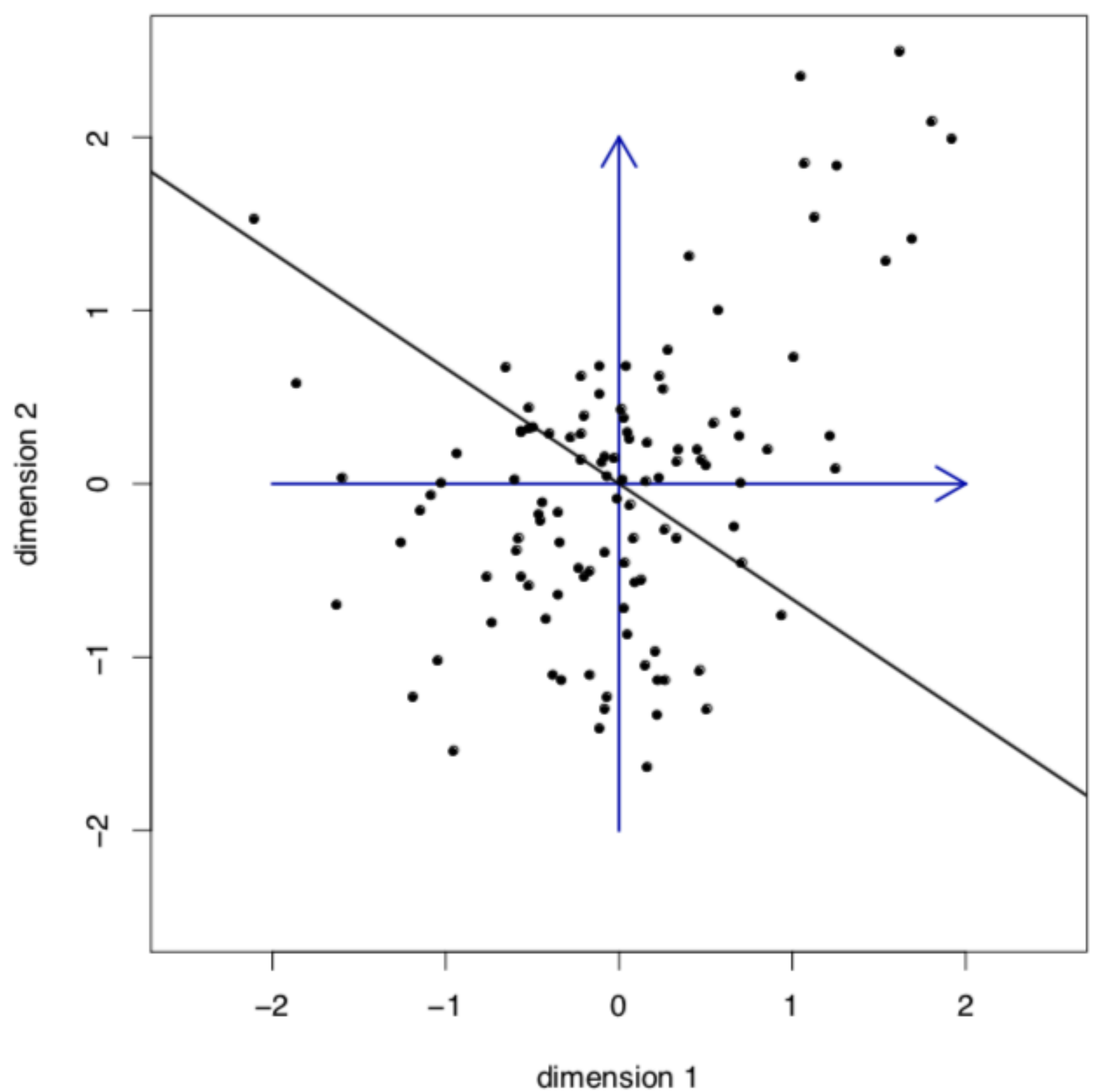
http://www.texample.net/tikz/examples/highlighting-matrix/

Capturing most variance



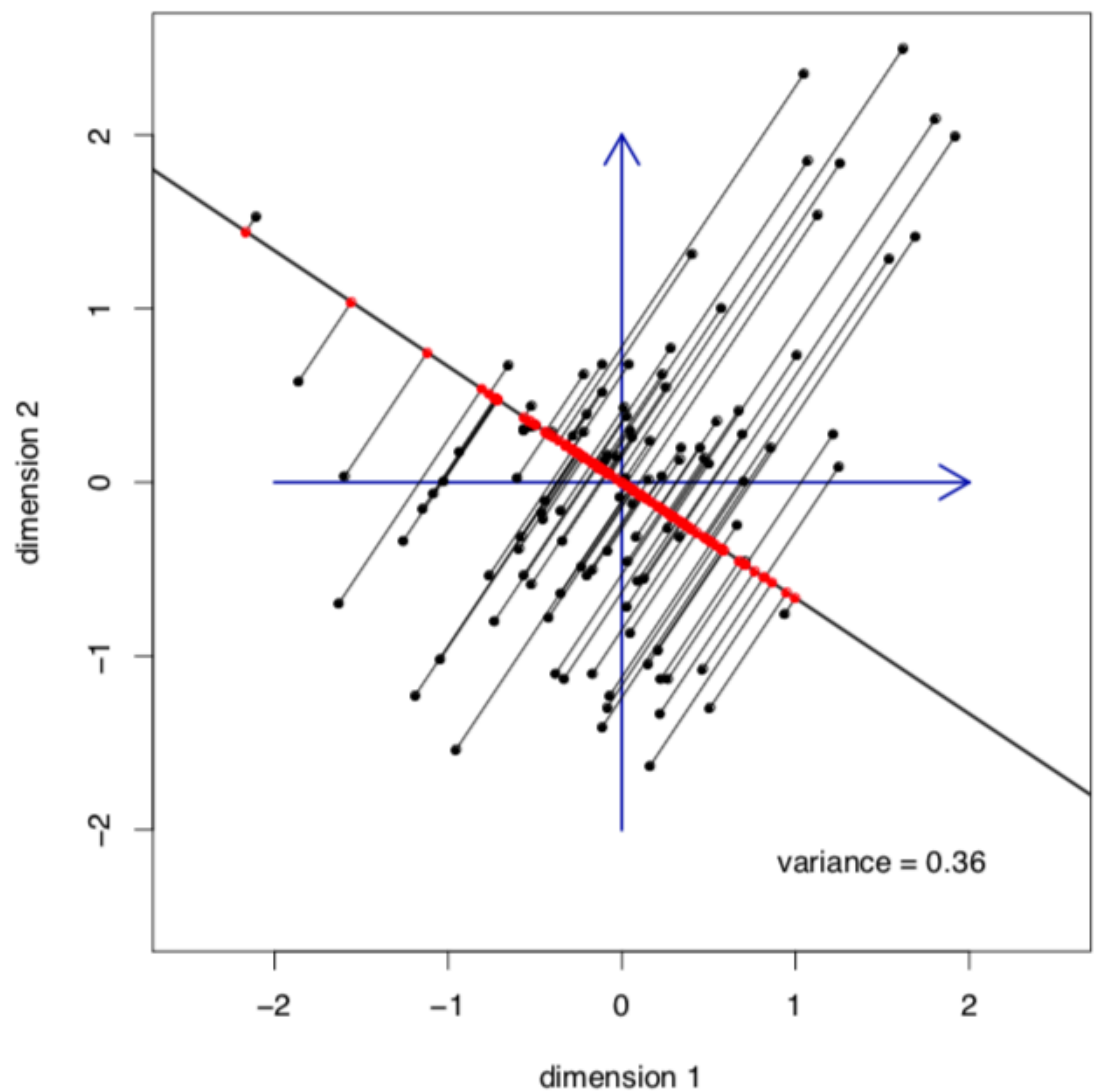
from Baroni & Boleda

Capturing most variance



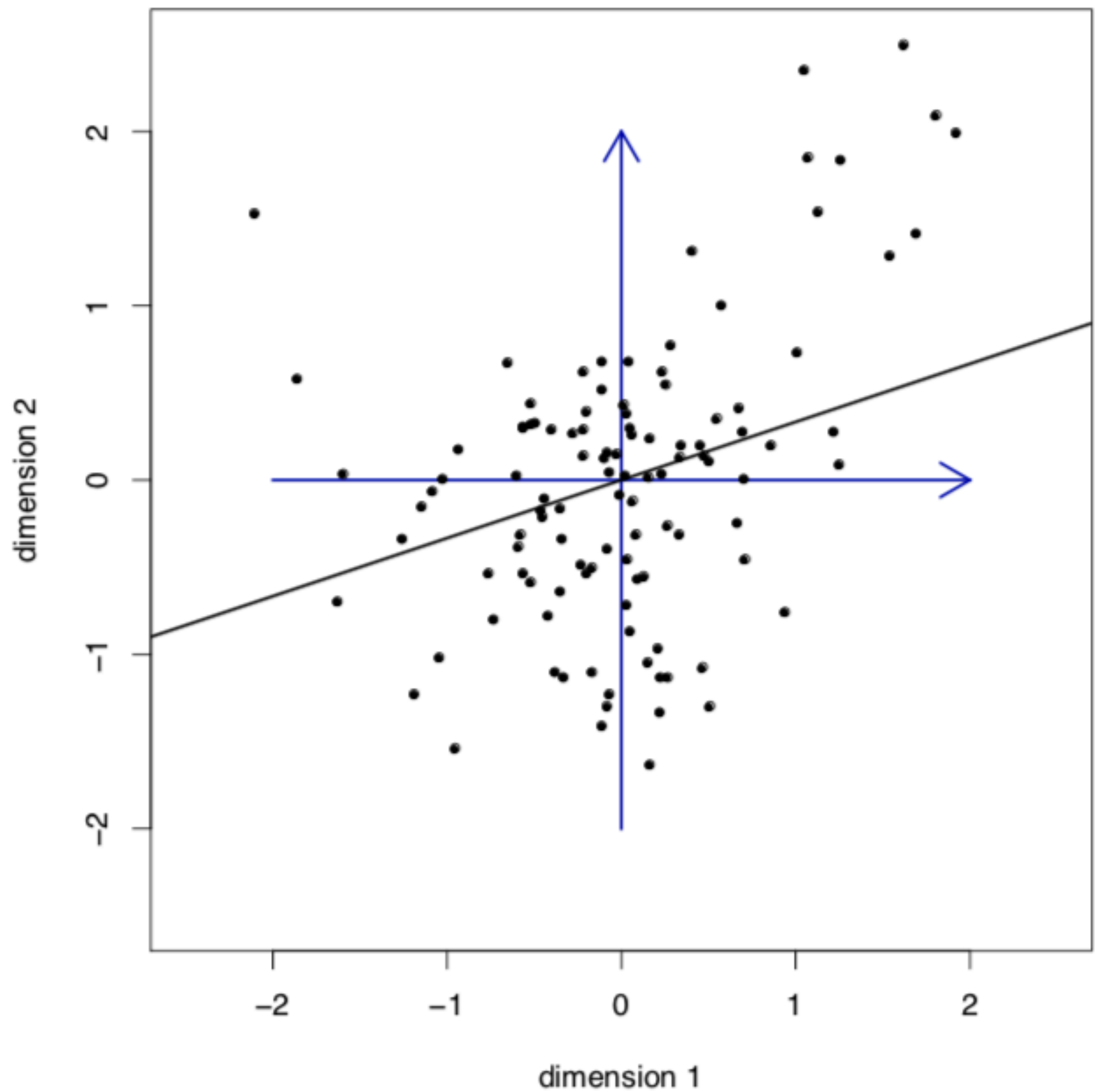
from Baroni & Boleda

Capturing most variance



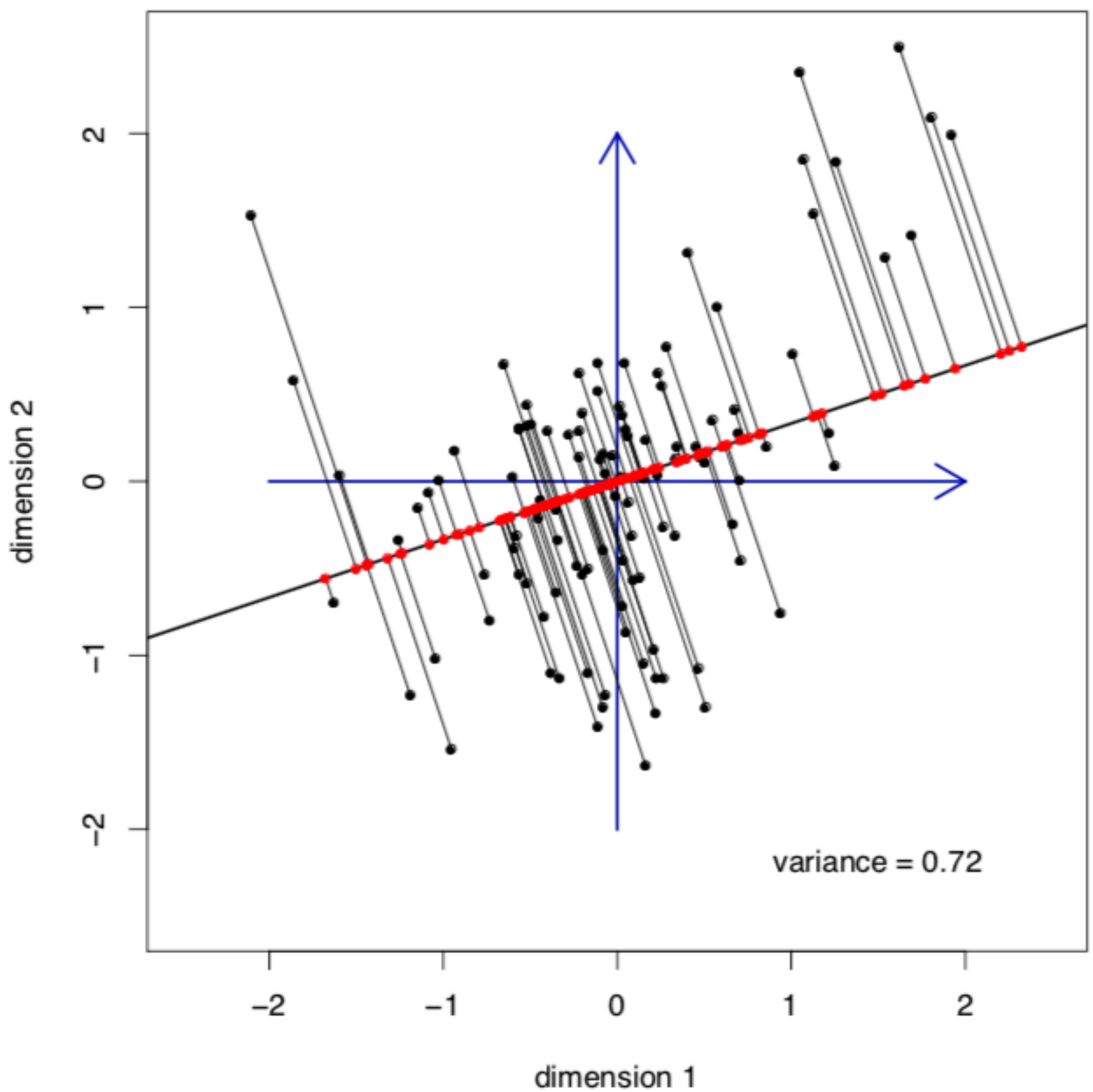
from Baroni & Boleda

Capturing most variance



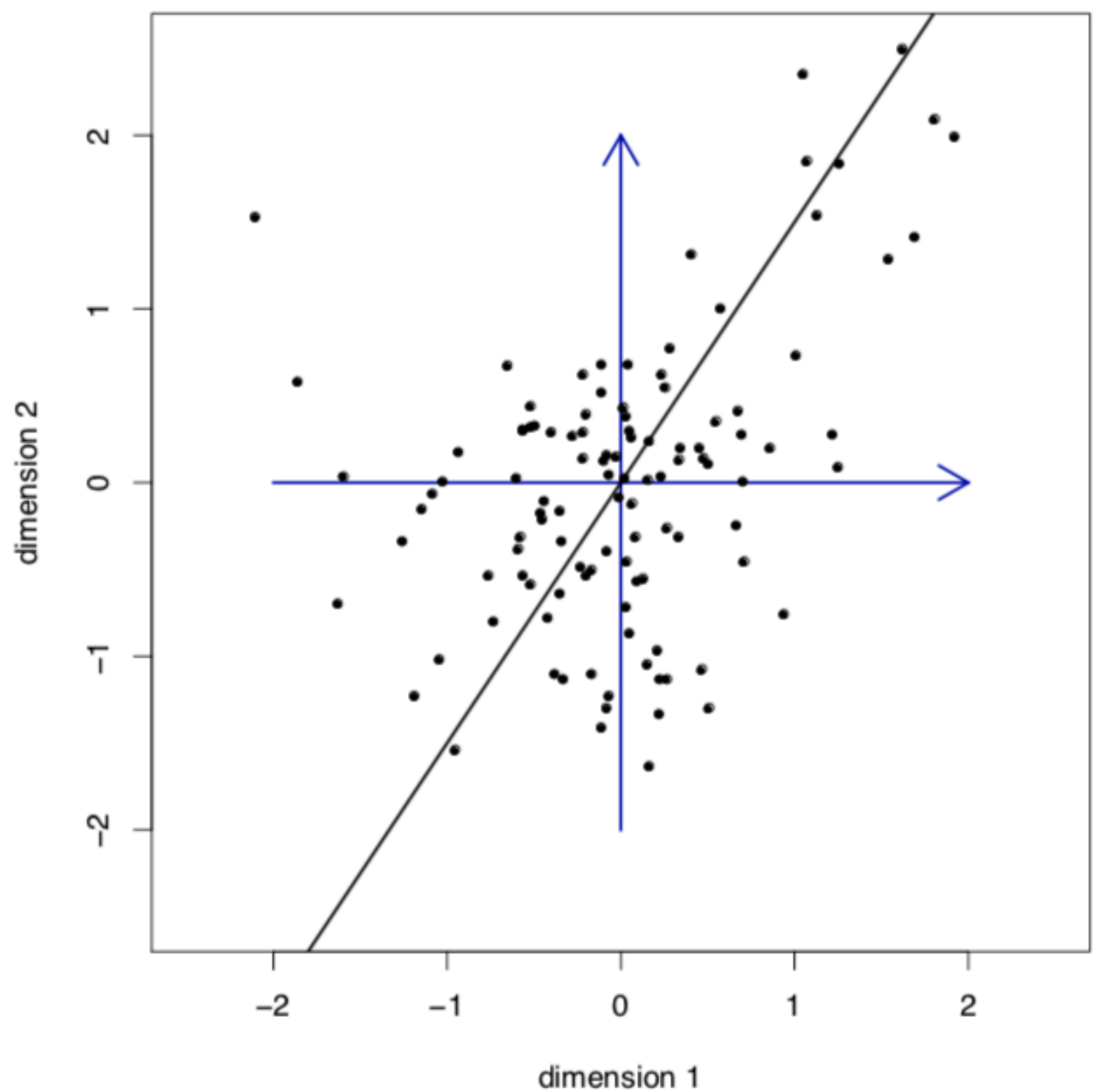
from Baroni & Boleda

Capturing most variance



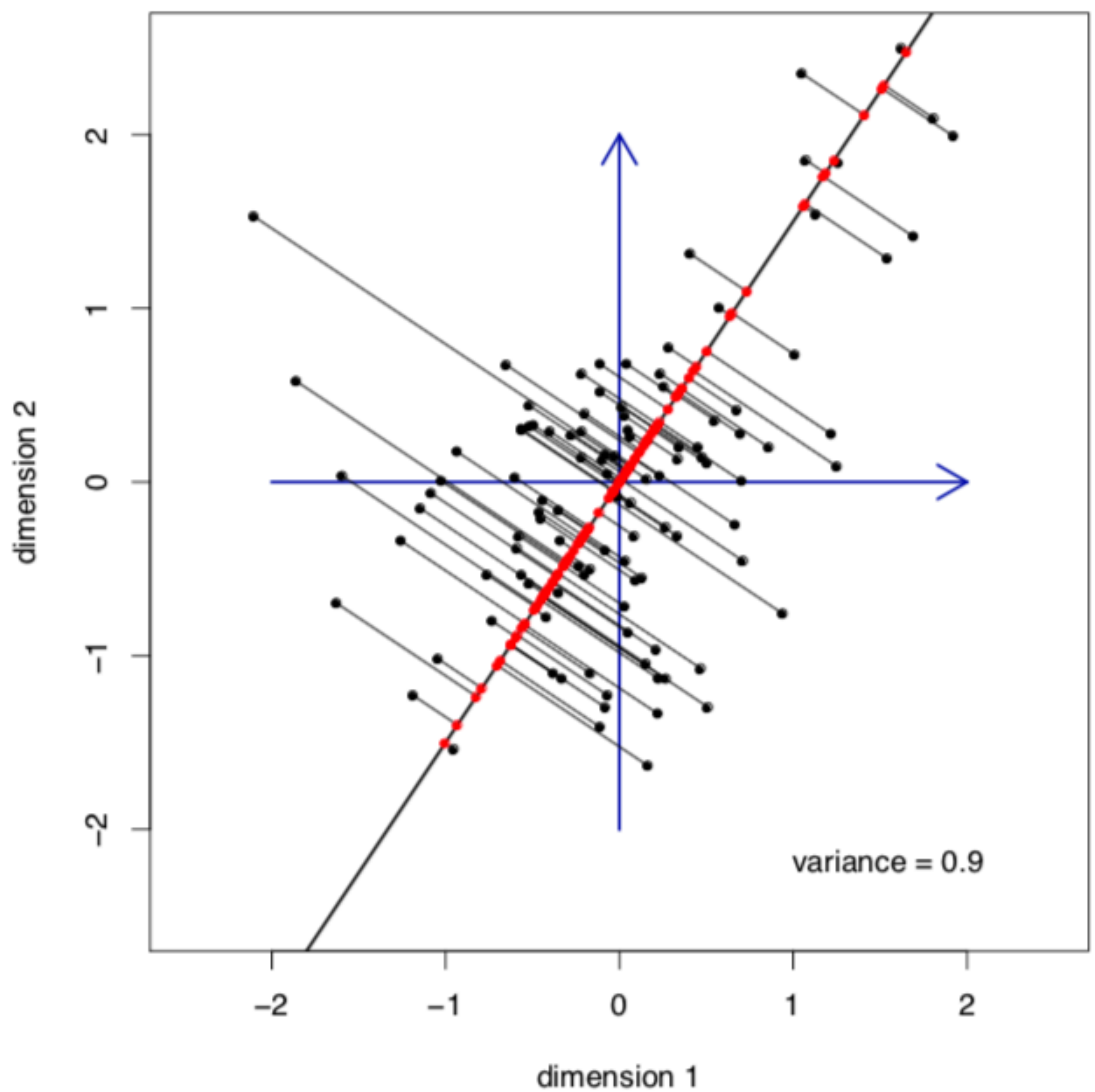
from Baroni & Boleda

Capturing most variance



from Baroni & Boleda

Capturing most variance



from Baroni & Boleda

Reducing dimensions

- Columns of *U* and *V* are ordered according to highest associated eigenvalue

- Diagonal values of Σ are ordered starting with highest (root of) eigenvalues

  => Using the first *d* rows of *U,* the first *d* columns of V^T and *d*x*d* rows and columns of Σ guarantees that we end up with those values that provide the highest variance.