interpretation

frequenc

probabil MI

measure

Motivatio

Goals

- -----

Experimen

Data

rechnic

Problem

Commean

Noisy dat

more v

Plots

....

# The effects of N-gram probabilistic measures on the recognition and production of four-word-sequences.

Eberhard Karls Universität Tübingen
Course: Frequency effects of multi-word sequences
Lecturer: Hendrix-Sun, June, Ph.D.
Student: Johannes Krämer
SS 2017

#### How to read the title?

## Title interpretation

frequency

What t

Motivation

IVIOLIVALIOI

Goal

\_

Ехренине

Date

Problems

more ,

FreaB

• Effects of [...] probabalistic measures [...]

#### How to read the title?

#### Title interpretation

frequency probabilit

What to

Motivation

iviotivatio

Goal

Sponer

Experimen

Data

\_ . . .

Problem

Collinearity

......

more v

Plots

- Effects of [...] probabalistic measures [...]
  - $\rightarrow$  it's about the methods:

#### How to read the title?

• Effects of [...] probabalistic measures [...]

 $\rightarrow$  it's about the methods:

- how was it measured?
- which probabalitic methods were used?
- which are the best methods?

#### Which methods were used?

Title interpretation

#### Method

probabili MI

measure

Motivation

. .

Spoile

Experimen

Data

Technic

Problem

Collinearit

more v

Plots

ггефь

Methods = Predictors

- frequency of occurance
- log probability of occurance
- mutual information

litle nterpretation

frequency probabilit

measure

Motivation

Goals

Spoller

Experimen

Data

Problems

Noisy dat

more v

Plots

FreqB

onclusion

frequency of occurance

• frequency of occurance

$$freq-of-event$$

 $\frac{\textit{freq} - \textit{of} - \textit{event}}{\textit{all} - \textit{possible} - \textit{events}}$ 

Title interpretation

Methods frequency probability

What to measure

Motivation

. ..

Evporimor

Technics

Problems

Collinearity

Noisy dat

more v

FreaR

probability of occurance

Title interpretation

Methods frequency probabilit

measure

Motivation

C . . . 1

Experime

Problem

Collinearit

......

Plots

FreqB

Canalinatan

probability of occurance

•  $P(W_i|W_{i-2},W_{i-1})$ 

Title interpretatior

Methods frequency probability

measure

Motivation

Goals

Spoile

Experimer

Б.

Technic

Problen

Collinearit

Noisy dat

more v

FreaB

FreqB

probability of occurance

•  $P(W_i|W_{i-2},W_{i-1})$ 

• what is more propable?

wit OR with

given that previous words are...

Title interpretation

frequency probability MI

measure

Motivation

Goals

Experimer

Dat

Techni

Problen

Collinearit

more v

FreqB

probability of occurance

- $P(W_i|W_{i-2},W_{i-1})$
- what is more propable?
  - wit OR with
  - given that previous words are...
  - I like pizza

Title interpretatior

Methods frequency probability MI

measure

Motivation

Goals

Spoile

Experimer

Techi

Droblon

Collinearit

Noisy dat

more v

Plots

FreqB

probability of occurance

•  $P(W_i|W_{i-2},W_{i-1})$ 

• what is more propable?

wit OR with

• given that previous words are...

I like pizza

(this example is from Mr. Cöltekins SNLP Course)

probability of occurance

•  $P(W_i|W_{i-2},W_{i-1})$ 

• what is more propable?

wit OR with

given that previous words are...

I like pizza

(this example is from Mr. Cöltekins SNLP Course)

log is for scaling

 $\rightarrow$  fitting more on the same plot

interpretation

Methods frequency probability

What to

Mativation

iviotivatioi

Goals

\_

Experimen

Dat

Problems

Noisy data

more v

Plots

mutual information

- mutual information
- how much do we know about a fact given another fact
- correlation of variables

Title interpretatior

Methods frequency probabilit

measure

Motivation

Goals

E.....

Experimen

Dat

Technie

Problem

Collinearit

Noisy da

more v

Plots

FreqB

- mutual information
- how much do we know about a fact given another fact
- correlation of variables
- I(X; Y)

Title interpretatior

Methods frequency probability

measure

Motivation

Goals

-----

Experimen

Dat

recillic

Problem

Collinearit

Plots

FreqB

- mutual information
- how much do we know about a fact given another fact
- correlation of variables
- I(X; Y)
- how much do we now about the gender given the name of a person?

Title interpretation

Methods frequency probability

What to measure

Motivation

. .

Spoil

Experimer

\_ `

Technic

Problem

Collinearity

...

Plots

Plots FreqB

- mutual information
- how much do we know about a fact given another fact
- correlation of variables
- I(X; Y)
- how much do we now about the gender given the name of a person?
- how much do we know about FreqAB given FreqA

#### What is measured?

Title interpretatior

Methods frequency probability

What t

Mativation

Motivatio

Goals

Experimen

Data

1 TODICITIS

Noisy data

more var

Plots

FreqB

onclucior

onset latency

#### What is measured?

Title interpretation

Methods frequency probability

measure

Motivation

Caala

experiment

D. . . .

Technic

Problems

Collinearity

Moiey dat

more v

Plots

FreqB

onclusion

onset latency

production duration

# Why do such a experiment in the first place? What inspired them?

• no one did it before

# Why do such a experiment in the first place? What inspired them?

Title interpretation

frequency probabilit

measure

Motivation

\_ .

Spoile

Experimen

Dat

Technic

Problem

Collinearit

Noisy dat

more \

Plots

FreqB

- no one did it before
- Baayen, Kuperman, Bertram (2010):
  - found interaction between measures of different linguistic units
  - $\rightarrow$  claim: multiple sources of linguistic information are processed in parallel

### This study: Find out the following questions

Title interpretation

frequency probabili

measure

Motivatio

Goals

Data

a ... .

Commeanty

INDISY UA

more v

FreaR

. . .

• whether freq., log prob., MI of 1-4grams affect onset latency and production duration

#### This study: Find out the following questions

Title interpretation

Methods frequency probabilit

measure

Motivatio

Goals

experimen

Data

Lechnic

Problem

Collinearit

Moisy dat

more v

Plots

FreqB

- whether freq., log prob., MI of 1-4grams affect onset latency and production duration
- Which is the better predictor for:
  - onset latency
  - production duration

#### This study: Find out the following questions

Title interpretation

frequency probabilit

Metivetie

Motivatio

Goals

Experimen

Data

rechnics

riobieilis

Collinearit

worsy dat

more v

Plots

FreqB

- whether freq., log prob., MI of 1-4grams affect onset latency and production duration
- Which is the better predictor for:
  - onset latency
  - production duration
- Are there (linear) interactions between predictors?

Title interpretation

Methods frequency probability

What t

Motivatio

Goals

Spoiler

Tankaia

Problems

Collinearity

more v

Plots

FreqB

Title interpretation

Methods frequency probabilit

measure

Motivation

Goals

Experimen

Data

rechnic

Problem

Collinearit

Noisy dat

more v

Plots

FreqB

C---l....

- $\rightarrow$  leaves us with:
- frequency of occurance
- log probability of occurance

Title interpretatior

frequency probabilit MI

measure

Motivatio

Goals

\_

Experimer

Α.

Technic

Problem

Collinearit

rvoisy de

more v

FreqB

Complete

- $\rightarrow$  leaves us with:
- frequency of occurance
- log probability of occurance
- prob. measures up to 4-grams all interact with each other

Title interpretation

frequency probabilit

measure

IVIOLIVALIO

Evperime

recimies

. ...

Collinearity

Plots

FreqB

- $\rightarrow$  leaves us with:
- frequency of occurance
- log probability of occurance
- log probability of occurance
- prob. measures up to 4-grams all interact with each other
  - but how strong?
  - for which N-gram does it work best?
  - what does it mean if they interact?

Title interpretation

frequency probabilit

measure

Motivation

Goals

:xperimer

Data

Collinearit

Noisy da

more v

FreaR

· .

• participants:

- 17 Students (10 female, 7 male)
- University of Alberta
- paid: yes

- participants:
  - 17 Students (10 female, 7 male)
  - University of Alberta
  - paid: yes
- material:
  - extracted 112 most freq. 4grams from BNC (100 mio words)

Title interpretation

Methods frequency probabilit

measure

iviotivatioi

Goals

Lxperimer

Data

Collinearity

Plots

FreqB

- participants:
  - 17 Students (10 female, 7 male)
  - University of Alberta
  - paid: yes
- material:
  - extracted 112 most freq. 4grams from BNC (100 mio words)
    - frequeny: 12 117 per million words (?)

participants:

17 Students (10 female, 7 male)

University of Alberta

paid: yes

material:

 extracted 112 most freq. 4grams from BNC (100 mio words)

• frequeny: 12 - 117 per million words (?)

• random selection of 320 4grams:

frequeny: 0.3 - 11 per million words

Title interpretatior

frequency probabilit

measure Motivation

o .

c "

Experime

\_

Technic

Problem

Commean

Noisy da

Dista

more va

participants:

• 17 Students (10 female, 7 male)

University of Alberta

paid: yes

material:

extracted 112 most freq. 4grams from BNC (100 mio words)

• frequeny: 12 - 117 per million words (?)

random selection of 320 4grams:

frequeny: 0.3 - 11 per million words

total: 432 4grams

#### Where did the data come from?

interpretation

frequency probabilit

What t

Motivatio

. .

Spoiler

experimen

Dat

Гесhnic

Problems

Collinearity

Noisy data

more vai

Plots

FreqD

• most freq. items from BNC

#### Where did the data come from?

most freq. items from BNC

frequencies of the items taken from COCA

 COCA = The Corpus of Contemporary American and English

COCA size: 385 mio. words

#### Where did the data come from?

Title interpretation

Methods frequency probabilit

What to

Motivation

\_ .

Spoile

Experimen

Technic

Problem

Collinearit

reorsy data

more v

Plots FreaB most freq. items from BNC

- frequencies of the items taken from COCA
  - COCA = The Corpus of Contemporary American and English
  - COCA size: 385 mio. words
- ullet don't, you've, wasn't o one word
- data was randomized for every participant

Title interpretation

Methods frequency probability

What to measure

Motivation

C 1

Spoller

Experimen

Data

Problems

Collinearity

Noisv data

more

Plots

Пець

• 4gram preceded by a fixation cross

Title interpretation

frequency probabilit

measure

Motivatio

Goals

experimen

Dat

#### Technic

Problems

Collinearit

Noisy dat

more v

Plots

FreqB

4gram preceded by a fixation cross

ullet cross for 500ms o black screen 20ms o 4gram 1500ms

Title interpretation

frequency probability MI

measure

Motivatio

Goals

Experimer

Data

Technic

Problem

Collinearit

Noisy dat

more v

Plots

ггефь

- 4gram preceded by a fixation cross
- ullet cross for 500ms o black screen 20ms o 4gram 1500ms
- interstimulus 1000ms

Title interpretatior

frequency probability MI

measure

Motivation

Coole

Spoile

Experimen

Dat

Technic

Problem

Collinearit

Noisy dat

more \

Plots

ггефь

- 4gram preceded by a fixation cross
- ullet cross for 500ms o black screen 20ms o 4gram 1500ms
- interstimulus 1000ms
- read out loud aqap ;)

Title interpretation

frequency probability MI

measure

Motivation

Goals

-----

Experimer

Tochnic

Problem

Collinearit

Noisy data

more v

FreqB

- 4gram preceded by a fixation cross
- ullet cross for 500ms o black screen 20ms o 4gram 1500ms
- interstimulus 1000ms
- read out loud aqap ;)
- 2 possible breaks (most of the participants didn't make use of it)

4gram preceded by a fixation cross

- ullet cross for 500ms o black screen 20ms o 4gram 1500ms
- interstimulus 1000ms
- read out loud agap;)
- 2 possible breaks (most of the participants didn't make use of it)
- 2 microphones

4gram preceded by a fixation cross

- ullet cross for 500ms o black screen 20ms o 4gram 1500ms
- interstimulus 1000ms
- read out loud agap;)
- 2 possible breaks (most of the participants didn't make use of it)
- 2 microphones
  - onset latency recording
  - speech recording

I itle interpretation

Methods frequency probabilit

What to measure

Motivation

Cools

:xperimer

Data

Problems

Collinearity

Confidently

more ver

IIIOIC Va

FreqB

. . . . . . . . . . . . .

W is our outcome

Title interpretation

frequenc probabili

What to

Motivotio

....

Spoile

Experimen

Collingarita

Collinearity

more v

Plots

- W is our outcome
- X and Z are predictors

Title interpretation

frequenc probabili

measure

Motivatio

Spoil

Experimer

ZAPCITICI

Techni

Problem

Collinearit

NI - 1 - - I - -

more v

Plots

FreqD

- W is our outcome
- X and Z are predictors
- $\bullet$  if X,Z correlated  $\to$  can't tell wether W is due to X or to Z

Title interpretation

frequence probabili

What to measure

Motivatio

. .

Spoile

Experimen

Technic

Problem

Collinearity

Noisy dat

more v

Plots

FredD

- W is our outcome
- X and Z are predictors
- $\bullet$  if X,Z correlated  $\to$  can't tell wether W is due to X or to Z
- draw on blackboard!

Title interpretation

Methods frequency probability

What to measure

Motivation

Goals

- .

\_\_\_\_\_

Dat

Problems

Collinearity

Noisy dat

more var

Plots

C .......

detect collinearity

Title interpretation

Methods frequency probability

measure

Motivation

Goals

Experimen

Collinearity

TVOISY Ga

more va

FreaB

Canalusian

- detect collinearity
- reduce collinearity

Title interpretation

Methods frequency probabilit

What to measure

Motivation

....

Count

Experimen

Data

Technic

Problems

Collinearit

Ninia. das

more v

Plots

FreqB

Conclusion

- detect collinearity
- reduce collinearity
  - centering

Title interpretation

Methods frequency probabilis

What to measure

Motivation

C . . . . . . .

Experimen

Data

Technic

Problem:

Collinearity

Noisy dat

more \

Plots

Constant

- detect collinearity
- reduce collinearity
  - centering
  - residualization

interpretation

frequency probabilit

What t measur

Motivatio

Эропсі

xperimen

Data

Moiey dat

more va

FreaB

Conclusion

ullet total: 7344 data points (17 imes 432)

Title interpretation

frequency probabilit

measure

Motivatio

Goals

Lxperimer

----

Commeant

ivoisy dai

more v

FreqE

....

total: 7344 data points (17 x 432)

• removed: 2251 = 30,7%

total: 7344 data points (17 x 432)

• removed: 2251 = 30.7%

WHY?

• 5 items had p = 0 in COCA  $\rightarrow$  85 data points | 1,2%

Title interpretation

frequency probabilit MI

measure

Motivation

Coole

C .... : I .

Experimer

Dat

Technic

Problem

Collineari

more var

FreqB

FreqB

- total: 7344 data points (17 x 432)
- removed: 2251 = 30,7%
- WHY?
  - $\bullet$  5 items had p = 0 in COCA  $\rightarrow$  85 data points | 1,2%
  - $\bullet~1$  Participant had to many errors  $\rightarrow~35$  data points |~0.5%

total: 7344 data points (17 x 432)

• removed: 2251 = 30.7%

WHY?

• 5 items had p = 0 in COCA  $\rightarrow$  85 data points | 1,2%

• 1 Participant had to many errors  $\rightarrow$  35 data points | 0.5%

• not triggered sound  $\rightarrow$  869 data points | 11,8%

Title interpretatior

frequency probabilit

measure

Motivation

Goals

Spoile

Experimer

Data

Technic

Problems

Collinearit

more var

FreqB

total: 7344 data points (17 x 432)

• removed: 2251 = 30,7%

• WHY?

• 5 items had p = 0 in COCA  $\rightarrow$  85 data points | 1,2%

ullet 1 Participant had to many errors ightarrow 35 data points | 0.5%

ullet not triggered sound ightarrow 869 data points  $\mid$  11,8%

ullet research assistants removed ightarrow 1258 data points | 17,1%

Title interpretation

frequency probabilit

What to

Motivation

....

Experimen

Technic

Problems

Collinearit

Noisy data

more va

Eroal

· . .

trials 1-432

Title interpretation

Methods frequency probability

measure

Motivation

Goals

\_ .

Experimen

Dat

Technic

Problems

Collinearit

more va

FreqE

C . . .

• trials 1-432

Manner

- stop
- approximant
- vowel
- fricative
- nasal

Title interpretation

Methods frequency probability

measure

Motivation

Goals

Lxperimen

Data

Technic

Problem

Collinearit

more va

FreqE

Completion

• trials 1-432

Manner

- stop
- approximant
- vowel
- fricative
- nasal
- NumSyll

Title interpretation

Methods frequency probabilit

measure

Motivation

Goals

Spoile

Experimen

Data

Technic

Collinearit

more va

Freql

FreqD

- trials 1-432
- Manner
  - stop
  - approximant
  - vowel
  - fricative
  - nasal
- NumSyll
- PhraseABCD

- trials 1-432
- Manner
  - stop
  - approximant
  - vowel
  - fricative
  - nasal
- NumSyll
- PhraseABCD
  - 117 phrase
  - 310 non-phrase

#### Interpretation of data

Title interpretation

frequency probabili

MI What t

. . . . .

MOLIVALIC

Goals

Experimer

Б...

Lechnic

Problem:

Collinearit

Plots

11040

• We will look together at the plots and figger out whats going on there.

## Why does FreqB have such an big impact?

Title interpretation

frequency

What t

Motivatio

iviotivatio

Spoil

Experimen

Dat

Problems

Collinearit

Noisy dat

more v

Plots

FreqB

. . . . . . . . . . . . . . . . . .

B-gram appeared at position of fixation cross

## Why does FreqB have such an big impact?

Title interpretation

frequency probabili

What t

Motivatio

Evporimor

\_\_\_\_\_

Data

Problem

Collinearit

Noisy da

more v

Plots

11040

B-gram appeared at position of fixation cross

WordTypeB as influencer?

Title interpretation

frequency probability

What t measur

Motivation

Goals

-

Experimen

Data

Problems

Commeanty

Noisy data

more var

FreqB

Conclusion

look at the results together!

Title interpretation

frequency probabili

What to measure

Motivation

Spoiler

xperimer

Data

\_\_\_\_

Problems

Collinearity

Moiey data

more v

Plots

Complexion

ullet onset latency o log probability of occurance

Title interpretatior

frequency probabilit MI

measure

Motivatio

Goals

Evnorimo

Collingarit

Commeanty

moro

Plots

. .

- $\bullet$  onset latency  $\rightarrow$  log probability of occurance
  - $\rightarrow$  'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'

Title interpretation

frequency probability

What to measure

Motivatio

c .

Spoile

Experimer

Data

Problen

Collinearit

Noisy da

more \

Plots

Conclusion

- ullet onset latency o log probability of occurance
  - ightarrow 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- ullet production duration o frequency of occurance (amount of experience)

Title interpretatior

Methods frequency probability

measure

Motivatio

Goals

Spoile

Experimen

Dat

Technic

Collinearit

moro

Plots

Plots FreqB

- ullet onset latency o log probability of occurance
  - ightarrow 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- ullet production duration o frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams

ullet onset latency o log probability of occurance → 'recognition is mainly underpinned by a mechanism

whereby a target N-gram competes with its family members'

- $\bullet$  production duration  $\rightarrow$  frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?

Title interpretatior

Methods frequency probability

What to measure

Motivatio

Goals

Spoile

Experimen

Data

**.** . .

Collineari

Noisy dat

more v

Plots FreqB

- ullet onset latency ightarrow log probability of occurance
  - $\rightarrow$  'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- ullet production duration o frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?
    - ightarrow trigrams

Title interpretatior

frequency probability

What to measure

Motivatio

Goals

Spoile

Experimen

Data

Droblom

Collineari

Noisy da

more va

Plots FreqB ullet onset latency ightarrow log probability of occurance

 $\rightarrow$  'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'

- ullet production duration o frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?
    - $\rightarrow \mathsf{trigrams}$
  - production duration?

Title interpretatior

frequency probabilit

What to measure

Motivatio

Goals

Spoiler

Experimen

Technic

Problems

----, ---

Plots

- ullet onset latency ightarrow log probability of occurance
  - ightarrow 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- ullet production duration o frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?
    - $\to \mathsf{trigrams}$
  - production duration?
    - ightarrow unigrams

Title interpretatior

frequency probabilit

What to measure

Motivatio

Goals

Spoile

Experimen

Technic

Problems

Connean

loisy da

more var Plots FreqB

- ullet onset latency ightarrow log probability of occurance
  - $\rightarrow$  'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- ullet production duration o frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?
    - $\rightarrow \mathsf{trigrams}$
  - production duration?
    - $\rightarrow$  unigrams
- BUT they still interact → what does it mean?

## The very conclusion

Title interpretation

frequency probabilit

measure

Motivatio

Spoil

Experime

Data

Technic

Problem

Collinearit

Noisy dat

Dlote

Plots FreaB

Conclusion

'Finally, the finding that probabilistic measures tied to N-grams up to four-words long **interacted** with each other in the onset latency and production duration analyses suggests that they are **processed in parallel** in both recognition and production.'

#### Reference:

Title interpretation

frequency probabilit

measure

Motivation

Spoil

Experimer

Data

Technic

Problem

Collinearit

reorsy da

more \

Plots FreaB

Camalunian

Antoine Tremblay and Benjamin V. Tucker. 'The effects of N-gram probabilistic measures on the recognition and production of four-word sequences.' IWK Health Center / University of Alberta.