Wombat - 2016

Simple tools for complex problems: making molehills out of mountains

Dr Zoe van Havre

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.
- *Key areas*:

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.
- *Key areas*:
  - Bayesian statistics

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.
- *Key areas*:
  - Bayesian statistics
  - Mixture and hidden Markov models,

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.
- *Key areas*:
  - Bayesian statistics
  - Mixture and hidden Markov models,
  - Bio-statistics/informatics/security,

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.
- *Key areas*:
  - Bayesian statistics
  - Mixture and hidden Markov models,
  - Bio-statistics/informatics/security,
- *Research interests*

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.
- *Key areas*:
  - Bayesian statistics
  - Mixture and hidden Markov models,
  - Bio-statistics/informatics/security,
- *Research interests*
  - data driven, accessible, intuitive tools

Who am I?

- PhD in statistics, from QUT & Paris-Dauphine
- I live in Brisbane, by way of Canada, New Zealand, and various places in between.
- *Key areas*:
  - Bayesian statistics
  - Mixture and hidden Markov models,
  - Bio-statistics/informatics/security,
- *Research interests*
  - data driven, accessible, intuitive tools
  - **making data analysis easier**

What drives me?

The most common question asked since I started to pursue Statistics has been
"Why...?"

- I have three reasons:

What drives me?

The most common question asked since I started to pursue Statistics has been
"Why...?"

- I have three reasons:

1. A sense of urgency,

What drives me?

The most common question asked since I started to pursue Statistics has been
"Why...?"

- I have three reasons:

1. A sense of urgency,
2. tantalizing hope,

What drives me?

The most common question asked since I started to pursue Statistics has been
"Why...?"

- I have three reasons:

1. A sense of urgency,
2. tantalizing hope,
3. boundless excitement.

Urgency?

Can we keep up?

- The exponential growth of computing has not slowed down.

to do: DIA-GRAM 1

to do: DIA-GRAM 2

Can we keep up?

- The exponential growth of computing has not slowed down.
- New types of data and new challenges require new approaches

to do: DIA-GRAM 1

to do: DIA-GRAM 2

Can we keep up?

- The exponential growth of computing has not slowed down.
- New types of data and new challenges require new approaches
- in 10 years, expecting to see 1000 times growth

to do: DIA-GRAM 1

to do: DIA-GRAM 2

Hope...

Not all change is bad

- Everyone is coming onboard! amazing advances

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now
- We are standing on a methodological goldmine...

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now
- We are standing on a methodological goldmine...
  - **the traditional way**:

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now
- We are standing on a methodological goldmine...
  - **the traditional way**:
    - Develop methods based on large sample theory.

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now
- We are standing on a methodological goldmine...
  - **the traditional way**:
    - Develop methods based on large sample theory.
    - Adapt / make assumptions to deploy on realistic sample sizes

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now
- We are standing on a methodological goldmine...
  - **the traditional way**:
    - Develop methods based on large sample theory.
    - Adapt / make assumptions to deploy on realistic sample sizes
  - **the future...?**

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now
- We are standing on a methodological goldmine...
  - **the traditional way**:
    - Develop methods based on large sample theory.
    - Adapt / make assumptions to deploy on realistic sample sizes
  - **the future...?**
    - Rework common methods to be closer to underlying theory

Not all change is bad

- Everyone is coming onboard! amazing advances
- Data-science is a thing now
- We are standing on a methodological goldmine...
  - **the traditional way**:
    - Develop methods based on large sample theory.
    - Adapt / make assumptions to deploy on realistic sample sizes
  - **the future...?**
    - Rework common methods to be closer to underlying theory
    - This might mean going Bayesian, yes. Sorry.

Excitement!

Better tools make data analysis easier

Amazing things happen when data analysis combines clear research questions, appropriate data, and suitable, accessible tools.

- Accessibility: usability, and understanding what the tool does.

It doesnt have to be just "analysis", it can be exploration, discovery, and a little bit magical.

Better tools make data analysis easier

Amazing things happen when data analysis combines clear research questions, appropriate data, and suitable, accessible tools.

- Accessibility: usability, and understanding what the tool does.
  - *A screwdriver doesn't come with an instruction manual.*

It doesnt have to be just "analysis", it can be exploration, discovery, and a little bit magical.

Better tools make data analysis easier

Amazing things happen when data analysis combines clear research questions, appropriate data, and suitable, accessible tools.

- Accessibility: usability, and understanding what the tool does.
  - *A screwdriver doesn't come with an instruction manual.*
- People can do more with less (ie. *bricks VS cement*)

It doesnt have to be just "analysis", it can be exploration, discovery, and a little bit magical.

Better tools make data analysis easier

Amazing things happen when data analysis combines clear research questions, appropriate data, and suitable, accessible tools.

- Accessibility: usability, and understanding what the tool does.
  - *A screwdriver doesn't come with an instruction manual.*
- People can do more with less (ie. *bricks VS cement*)
- Simple models are less likely to be wrongly used

It doesnt have to be just "analysis", it can be exploration, discovery, and a little bit magical.

A short story about Alzheimer's Disease| featuring... overfitted mixture models!

Key background

Alzheimer's Disease (AD) currently affects over 342,800 Australians, and this number is expected to rise to 900,000 by 2050.
Cognitive changes indicated something is amiss, but these occur late in the disease ($\geq$ 20 years).
During this time, AD causes irreversible damage to the brain:

- accumulation of **amyloid** β,

To better research and treat AD, we need to be able to treat it earlier.

Key background

Alzheimer's Disease (AD) currently affects over 342,800 Australians, and this number is expected to rise to 900,000 by 2050.
Cognitive changes indicated something is amiss, but these occur late in the disease ($\geq$ 20 years).
During this time, AD causes irreversible damage to the brain:

- accumulation of **amyloid** β,
- neurofibrillary tangles,

To better research and treat AD, we need to be able to treat it earlier.

Key background

Alzheimer's Disease (AD) currently affects over 342,800 Australians, and this number is expected to rise to 900,000 by 2050.
Cognitive changes indicated something is amiss, but these occur late in the disease ($\geq$ 20 years).
During this time, AD causes irreversible damage to the brain:

- accumulation of **amyloid** β,
- neurofibrillary tangles,
- overall atrophy.

To better research and treat AD, we need to be able to treat it earlier.

What you need to know

- Alzheimer's Disease (AD) is something we need to address

What you need to know

- Alzheimer's Disease (AD) is something we need to address
- disease development is very slow

What you need to know

- Alzheimer's Disease (AD) is something we need to address
- disease development is very slow
- cognitive changes indetectable for $\geq$ 20 years

What you need to know

- Alzheimer's Disease (AD) is something we need to address
- disease development is very slow
- cognitive changes indetectable for $\geq$ 20 years
- Tests which assess physical change are $ $ $ and intrusive

How can we help improve early detection

To better tackle AD, we need to be able to treat it earlier.

- we know little about how AD behaves in its early stage

How can we help improve early detection

To better tackle AD, we need to be able to treat it earlier.

- we know little about how AD behaves in its early stage
- could compare known cases to controls,

How can we help improve early detection

To better tackle AD, we need to be able to treat it earlier.

- we know little about how AD behaves in its early stage
- could compare known cases to controls,
  - does not target early stage of AD

How can we help improve early detection

To better tackle AD, we need to be able to treat it earlier.

- we know little about how AD behaves in its early stage
- could compare known cases to controls,
  - does not target early stage of AD
- **would like to identify individuals likely to be in early stage of AD**

How?

- large repository of data exists thanks to AIBL study

How?

- large repository of data exists thanks to AIBL study
- many data types, potential variables, time points, and sources

How?

- large repository of data exists thanks to AIBL study
- many data types, potential variables, time points, and sources
- possibilites = *endless* (thousands of potential approaches)

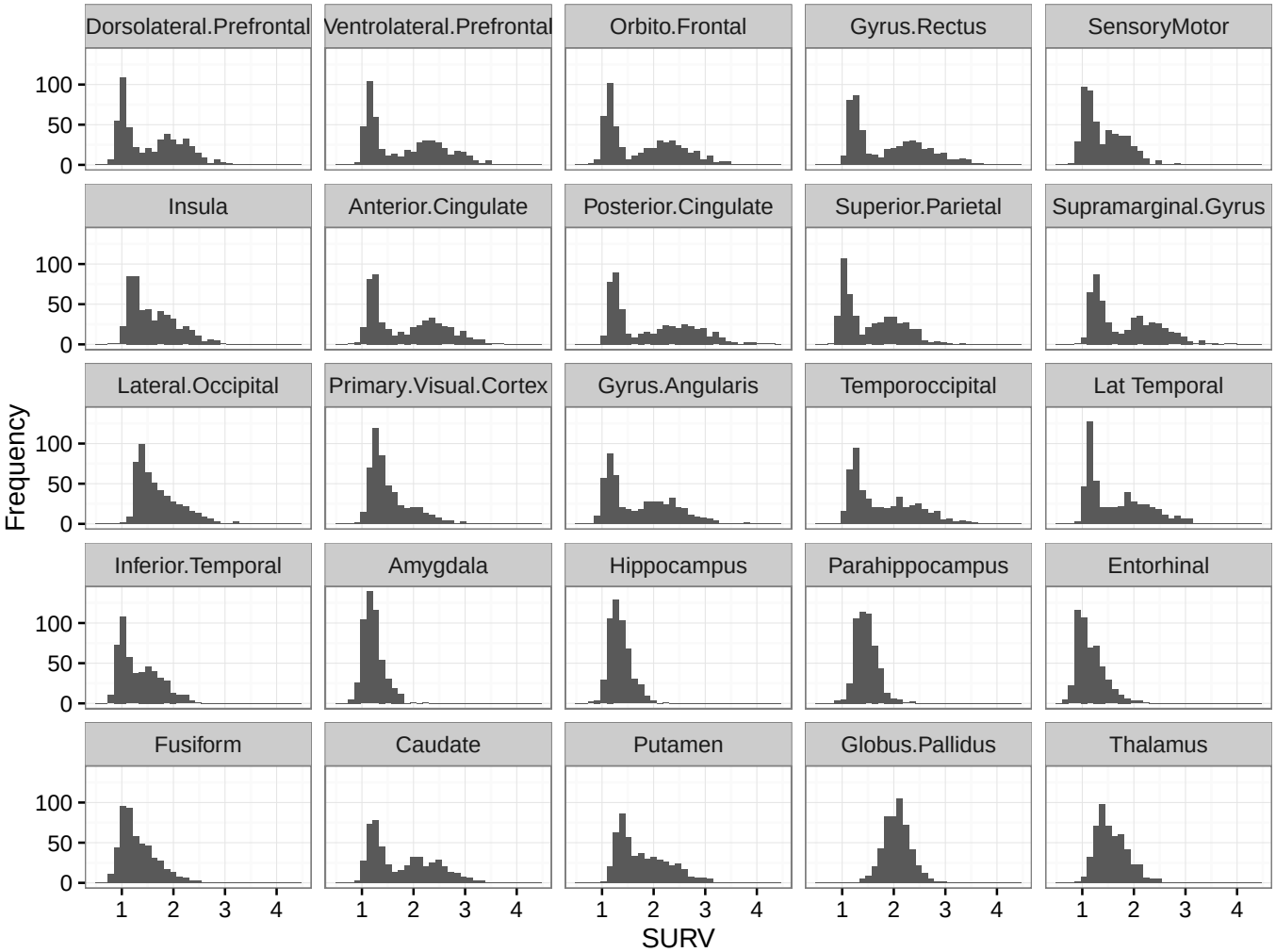
How?

- large repository of data exists thanks to AIBL study
- many data types, potential variables, time points, and sources
- possibilites = *endless* (thousands of potential approaches)
- What now...?

The Data

The study consists of 507 individuals, composed of Healthy Controls (HC), MCI, and AD patients.

```
##
##  AD MCI  HC
## 103 114 290
```


Dorsolateral.Prefrontal
Ventrolateral.Prefrontal
Orbito.Frontal
Gyrus.Rectus
SensoryMotor
Insula
Anterior.Cingulate
Posterior.Cingulate
Superior.Parietal
Supramarginal.Gyrus
Lateral.Occipital
Primary.Visual.Cortex
Gyrus.Angularis
Temporoccipital
Lat Temporal
Inferior.Temporal
Amygdala
Hippocampus
Parahippocampus
Entorhinal
Fusiform
Caudate
Putamen
Globus.Pallidus
Thalamus
Frequency
0
50
100
1
2
3
4
SURV

Overfitting with Zmix

Overfitted mixture models

We can model an unknown number of groups using **overfitted mixture models**, a Bayesian method found in the R package "Zmix".

- too many groups are included in a mixture model

Overfitted mixture models

We can model an unknown number of groups using **overfitted mixture models**, a Bayesian method found in the R package "Zmix".

- too many groups are included in a mixture model
- extra groups **empty out**

Overfitted mixture models

We can model an unknown number of groups using **overfitted mixture models**, a Bayesian method found in the R package "Zmix".

- too many groups are included in a mixture model
- extra groups **empty out**
- probability of number of occupied groups

Overfitted mixture models

We can model an unknown number of groups using **overfitted mixture models**, a Bayesian method found in the R package "Zmix".

- too many groups are included in a mixture model
- extra groups **empty out**
- probability of number of occupied groups
- data driven and fully parametric

Overfitted mixture models

We can model an unknown number of groups using **overfitted mixture models**, a Bayesian method found in the R package "Zmix".

- too many groups are included in a mixture model
- extra groups **empty out**
- probability of number of occupied groups
- data driven and fully parametric
- Bayesian, but straightforward

Overfitted mixture models

We can model an unknown number of groups using **overfitted mixture models**, a Bayesian method found in the R package "Zmix".

- too many groups are included in a mixture model
- extra groups **empty out**
- probability of number of occupied groups
- data driven and fully parametric
- Bayesian, but straightforward
- Assumes only that up to K groups are normally distributed with an unknown mean and variance.

How it's done

Install the package

```
install_github('zoevanhavre/Zmix') # Thank you Hadley!
library(Zmix)
```

Run the model with $K = 5$ groups

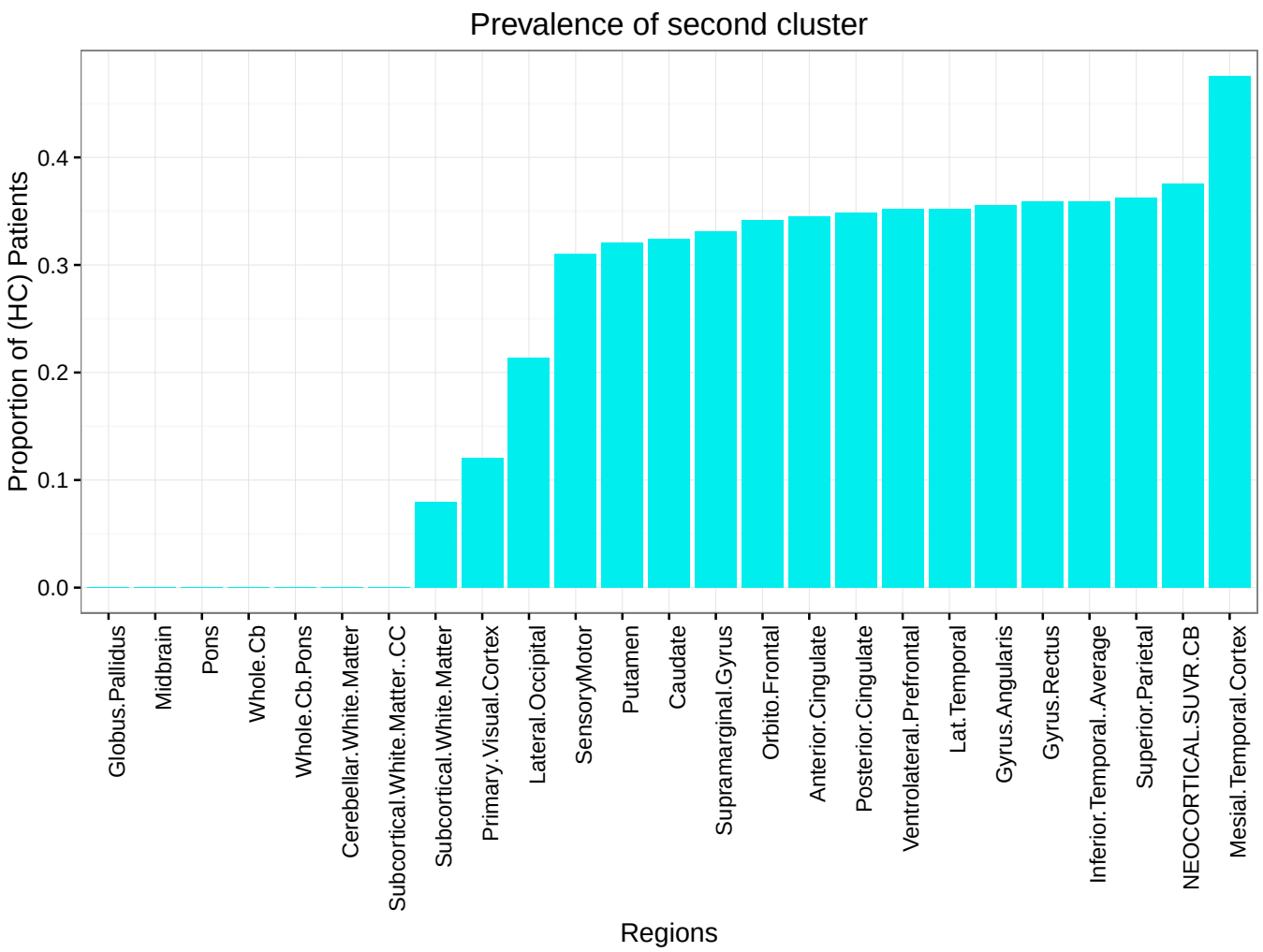
```
### <b>
Zmix.Y<-Zmix_univ_tempered (Y, iter=50000, k=5)
### </b>
```

Process the results

```
Proc.Zmix.Y<-Process_Output_Zmix(Zmix.Y, Burn=25000)
```

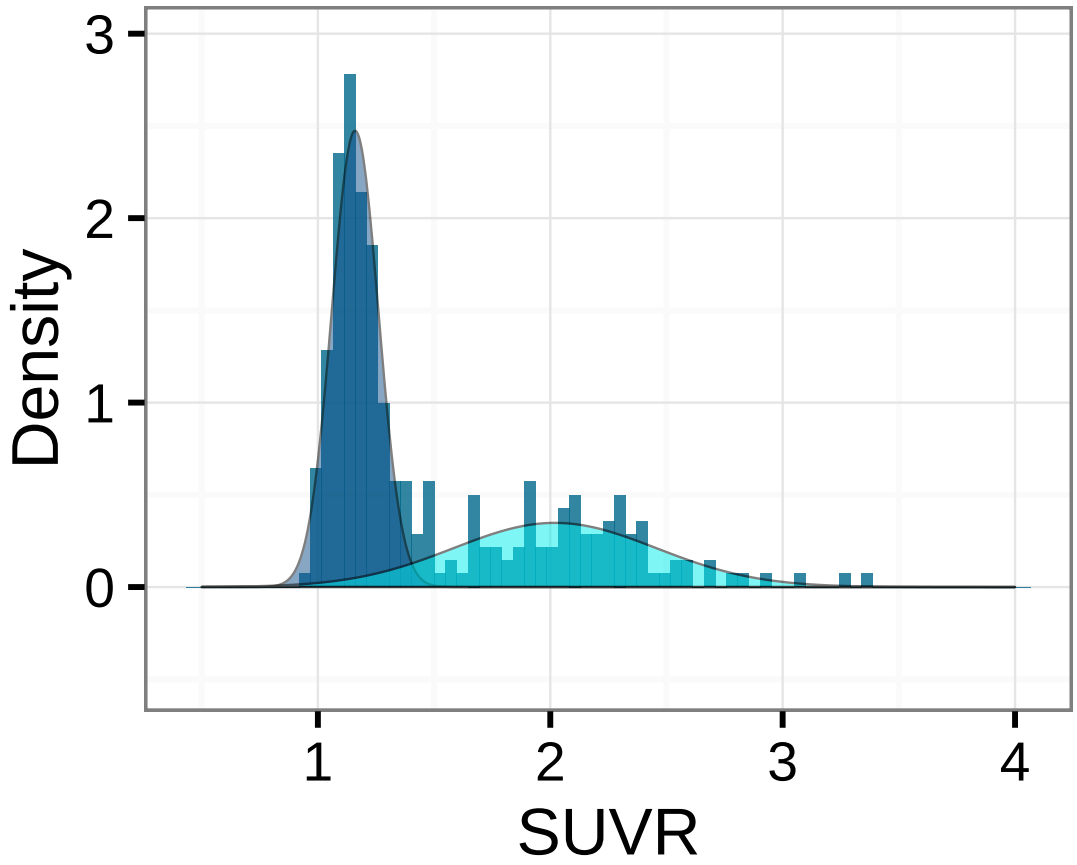
Check out the README for more examples

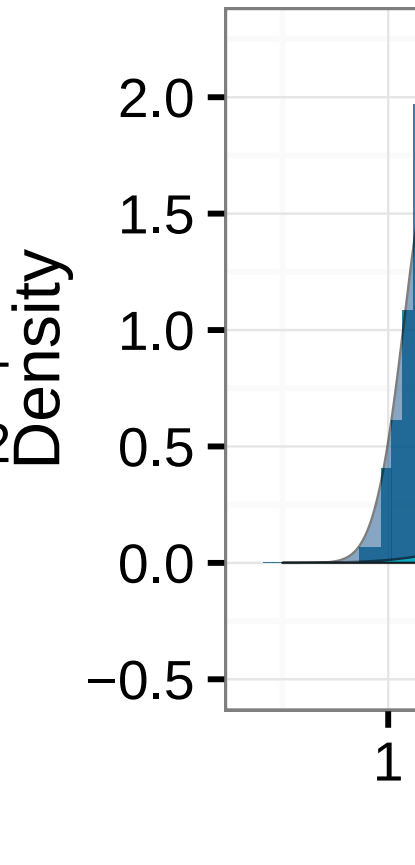
Results


Prevalence of second cluster
Proportion of (HC) Patients
0.0
0.1
0.2
0.3
0.4
Globus.Pallidus
Midbrain
Pons
Whole.Cb
Whole.Cb.Pons
Cerebellar.White.Matter
Subcortical.White.Matter..CC
Subcortical.White.Matter
Primary.Visual.Cortex
Lateral.Occipital
SensoryMotor
Putamen
Caudate
Supramarginal.Gyrus
Orbito.Frontal
Anterior.Cingulate
Posterior.Cingulate
Ventrolateral.Prefrontal
Lat.Temporal
Gyrus.Angularis
Gyrus.Rectus
Inferior.Temporal..Average
Superior.Parietal
NEOCORTICAL.SUVR.CB
Mesial.Temporal.Cortex
Regions

Ventrolateral.Prefrontal

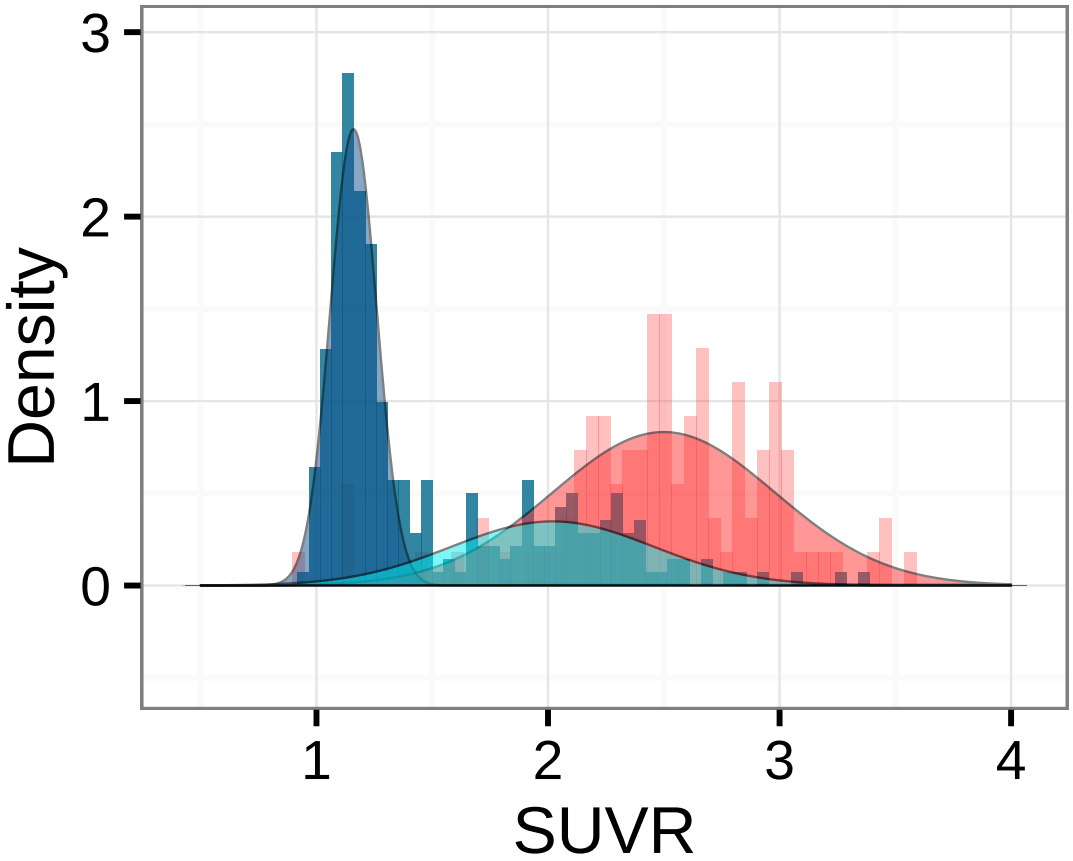
HC (Blue)

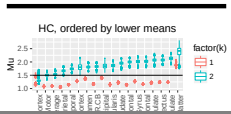


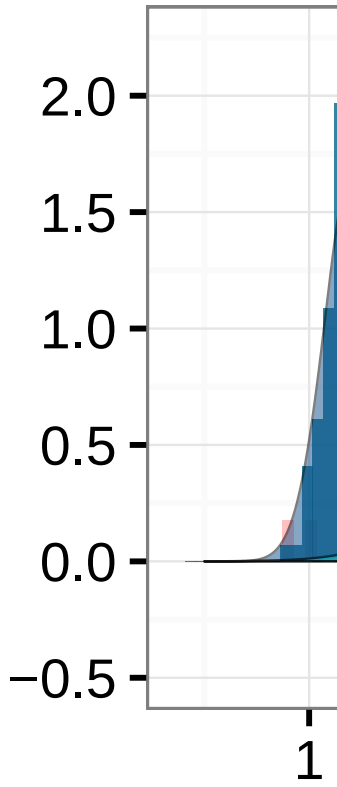


Ventrolateral.Prefrontal

HC (Blue) & AD (Red)







Results overview

- Zmix found either **one** or **two** groups

Results overview

- Zmix found either **one** or **two** groups
- Majority of regions result in two clusters,

Results overview

- Zmix found either **one** or **two** groups
- Majority of regions result in two clusters,
- Prevalence of 2nd group similar across regions,

Results overview

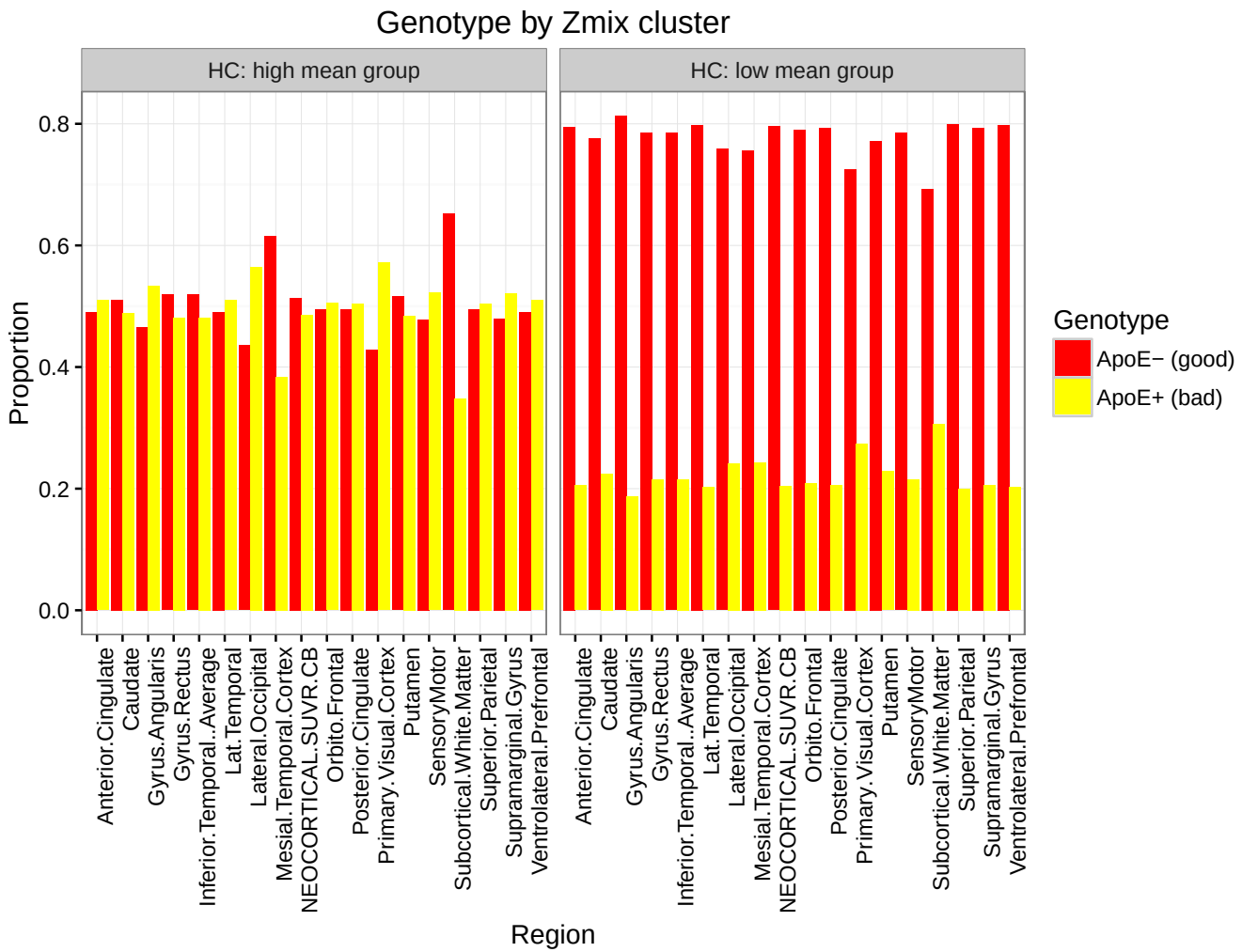
- Zmix found either **one** or **two** groups
- Majority of regions result in two clusters,
- Prevalence of 2nd group similar across regions,
- Allocations to 2nd group highly correlated (across individuals)

Results overview

- Zmix found either **one** or **two** groups
- Majority of regions result in two clusters,
- Prevalence of 2nd group similar across regions,
- Allocations to 2nd group highly correlated (across individuals)
- The HC clusters with larger means resemble the distribution of SUVR in AD, shifted to a lower mean, (as would be expected in early stages of the disease).

Results overview

- Zmix found either **one** or **two** groups
- Majority of regions result in two clusters,
- Prevalence of 2nd group similar across regions,
- Allocations to 2nd group highly correlated (across individuals)
- The HC clusters with larger means resemble the distribution of SUVR in AD, shifted to a lower mean, (as would be expected in early stages of the disease).
- They also follow a similar pattern across regions to AD


Genotype by Zmix cluster
HC: high mean group
HC: low mean group
Proportion
0.0
0.2
0.4
0.6
0.8
Anterior.Cingulate
Caudate
Gyrus.Angularis
Gyrus.Rectus
Inferior.Temporal..Average
Lat.Temporal
Lateral.Occipital
Mesial.Temporal.Cortex
NEOCORTICAL.SUVR.CB
Orbito.Frontal
Posterior.Cingulate
Primary.Visual.Cortex
Putamen
SensoryMotor
Subcortical.White.Matter
Superior.Parietal
Supramarginal.Gyrus
Ventrolateral.Prefrontal
Region
Genotype
ApoE− (good)
ApoE+ (bad)


Memory Status by Zmix cluster
HC: high mean group
HC: low mean group
Proportion
0.0
0.2
0.4
0.6
Anterior.Cingulate
Caudate
Gyrus.Angularis
Gyrus.Rectus
Inferior.Temporal..Average
Lat.Temporal
Lateral.Occipital
Mesial.Temporal.Cortex
NEOCORTICAL.SUVR.CB
Orbito.Frontal
Posterior.Cingulate
Primary.Visual.Cortex
Putamen
SensoryMotor
Subcortical.White.Matter
Superior.Parietal
Supramarginal.Gyrus
Ventrolateral.Prefrontal
Region
Memory
Complainer
Non−Complainer