Are we there yet?

Testing the effectiveness of graphics as MCMC diagnostics

Johns Hopkins University

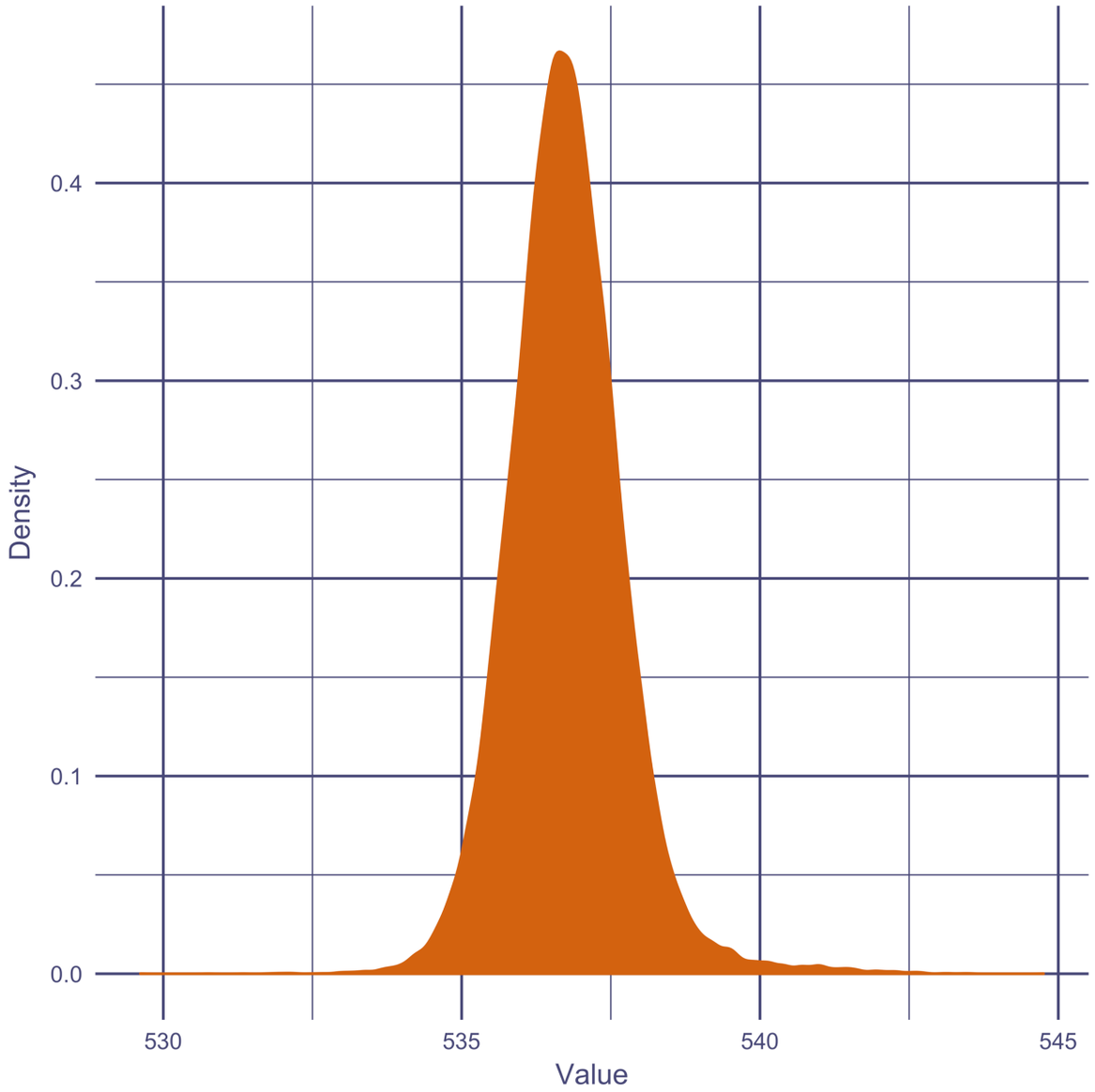
Wednesday 9th January, 2019

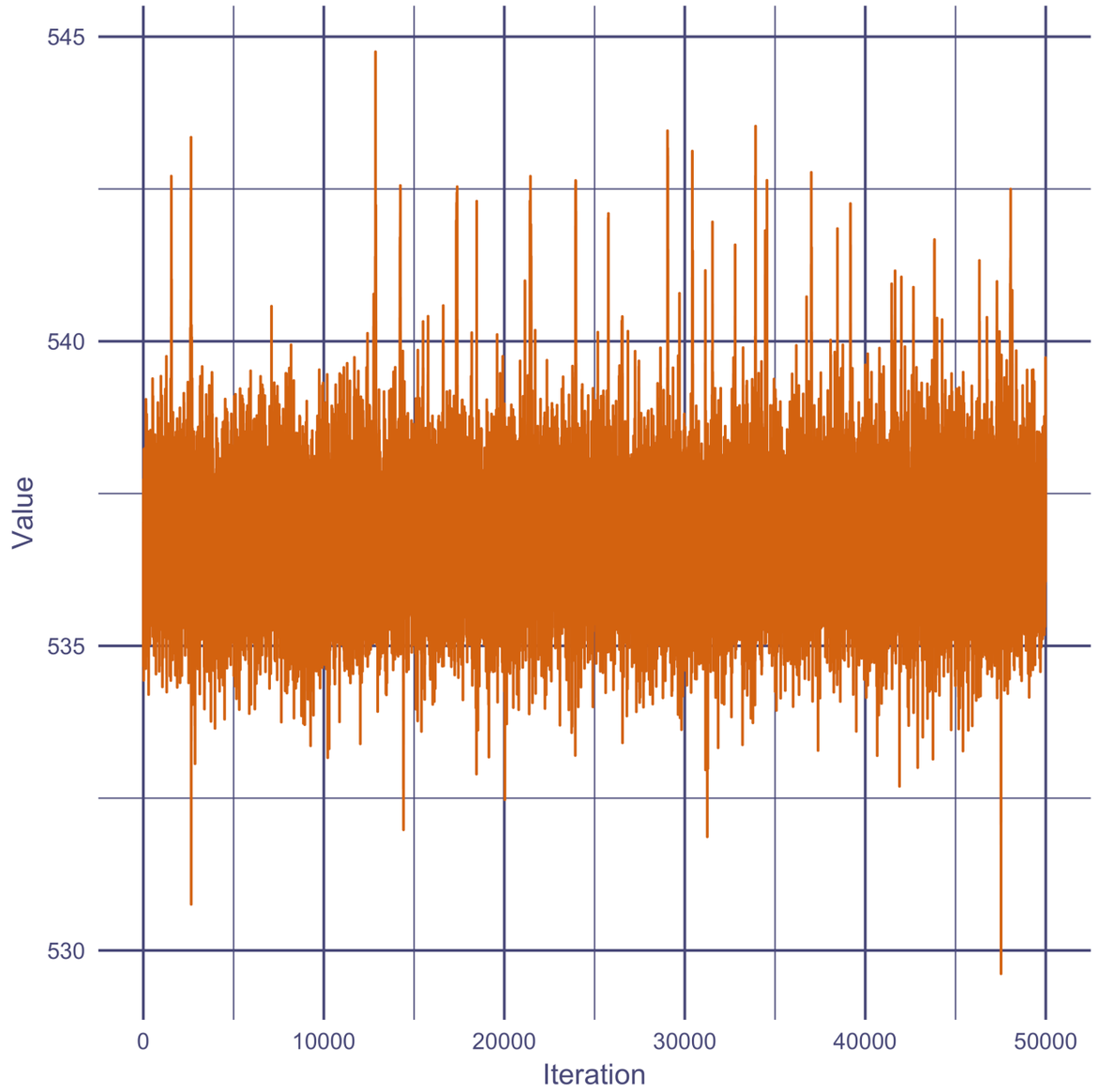
Nicholas Tierney, Monash University, Australia

nj_tierney

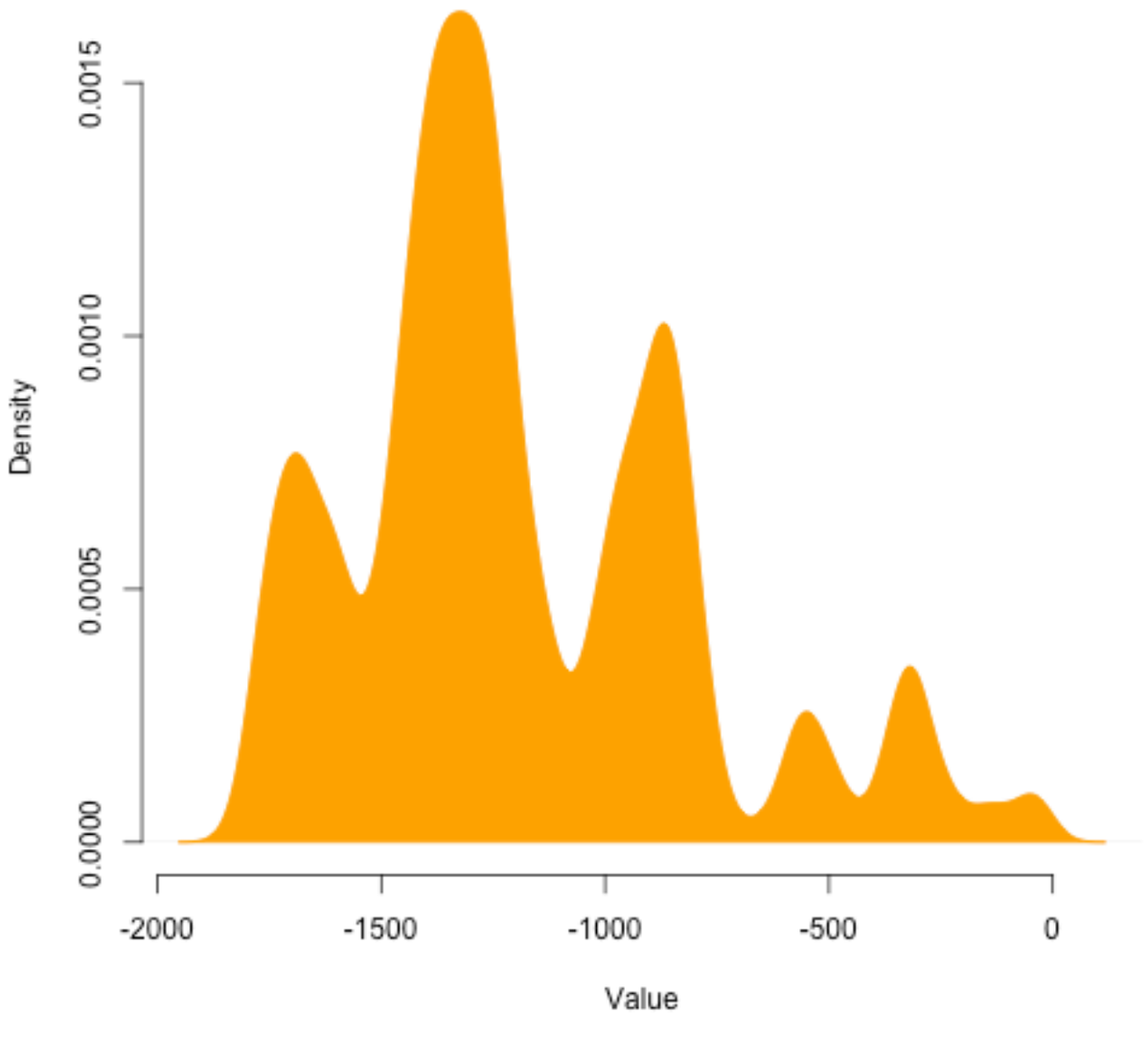
Why care about convergence?

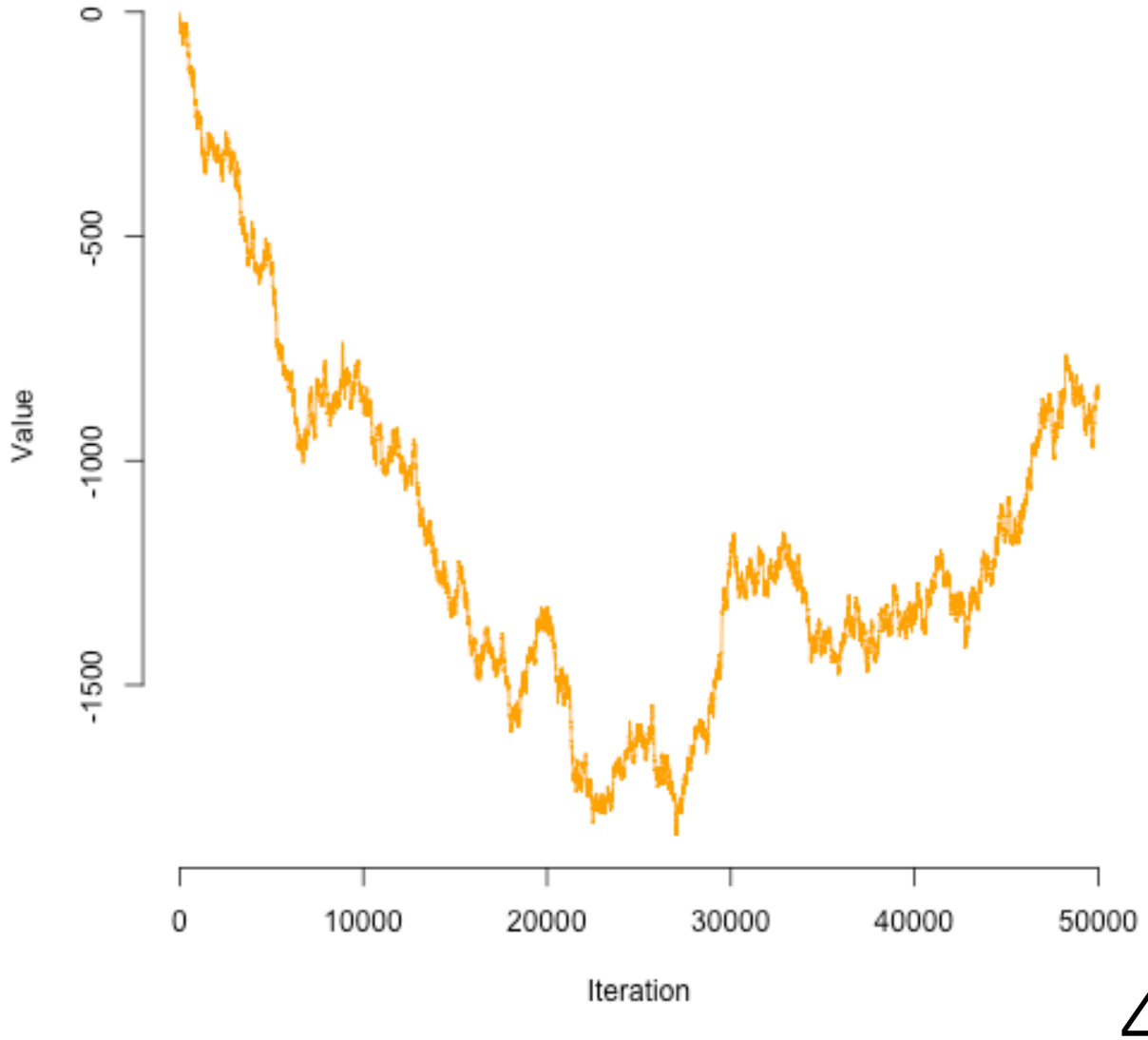
Converged?


0.4
0.3
0.2
0.1
0.0
Density
530
535
540
545
Value


545
540
535
530
Value
0
10000
20000
30000
40000
50000
Iteration

Converged?


Density
0.0000
0.0005
0.0010
0.0015
-2000
-1500
-1000
-500
0
Value


Value
0
-500
-1000
-1500
0
10000
20000
30000
40000
50000
Iteration

We *know* when it looks converged.

Right?

Yes, and no.

Research Question

Can visualisations be used to detect convergence?

Research Question

Can visualisations be used to detect convergence?

Are some visualisations better than others for detecting convergence?

Is confidence related to detecting convergence?

Is experience with Bayesian statistics related to detecting convergence?

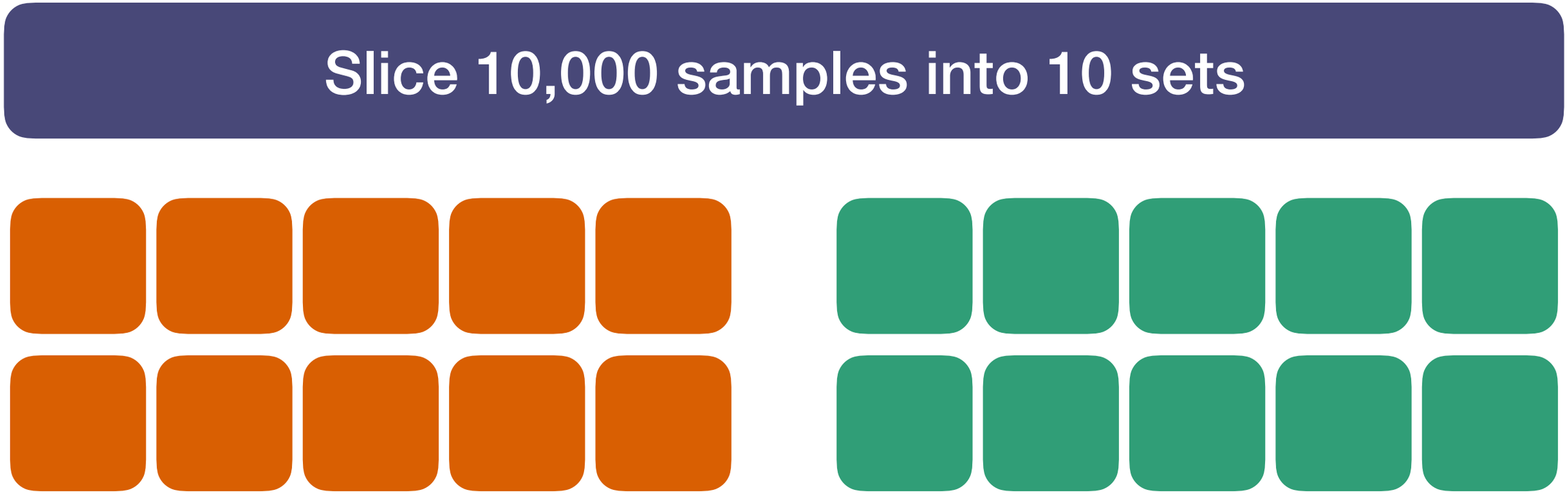
Experimental Method

Model

Bad Specification	Good Specification
60,000 Burn in	60,000 Burn in
10,000 Samples	10,000 Samples

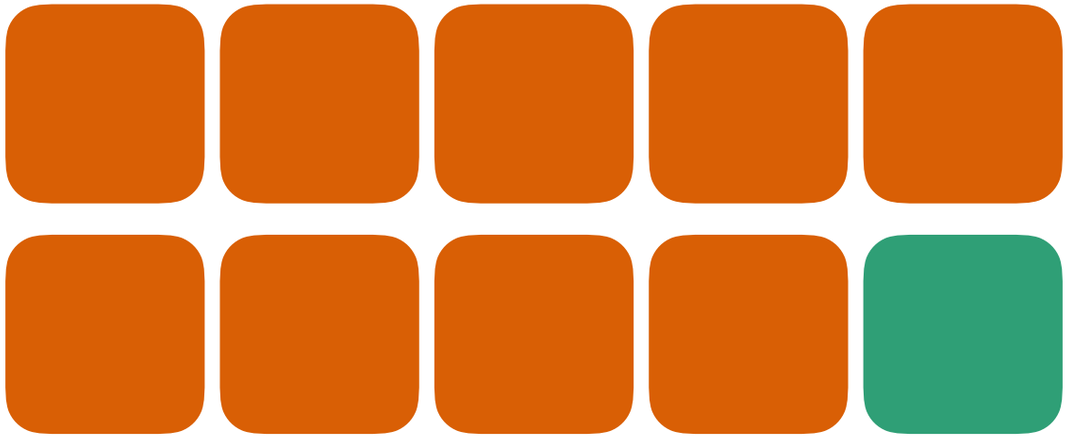
Experimental Method

Slice 10,000 samples into 10 sets



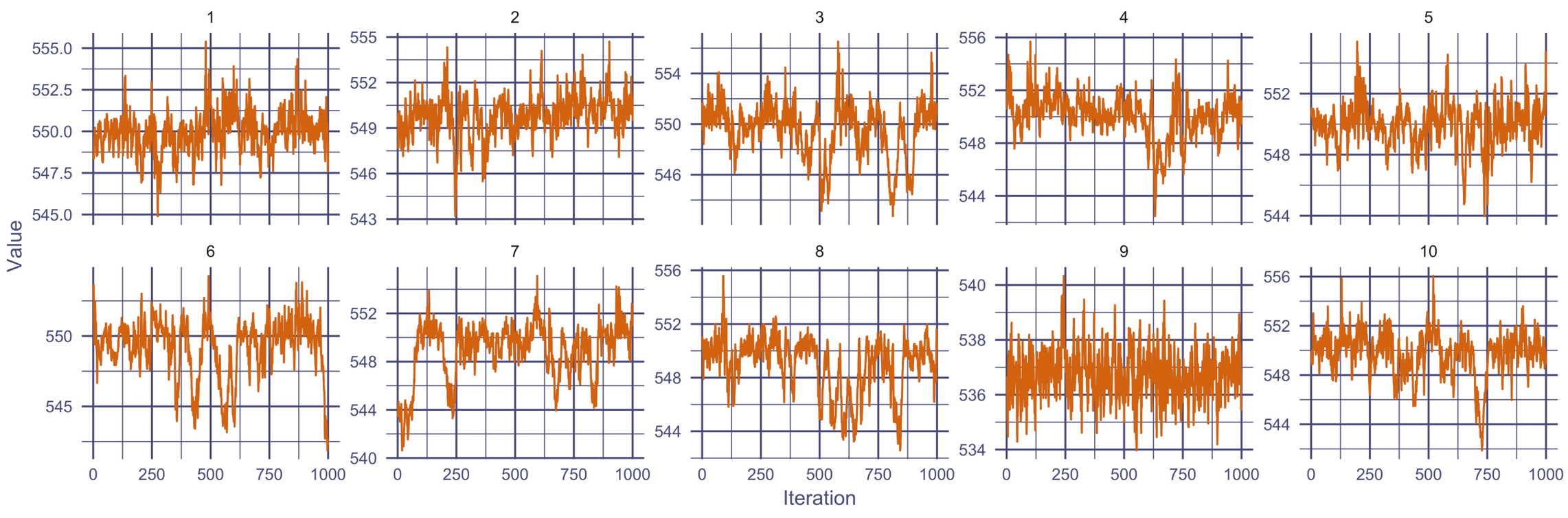
Experimental Method

Select nine bad and one good

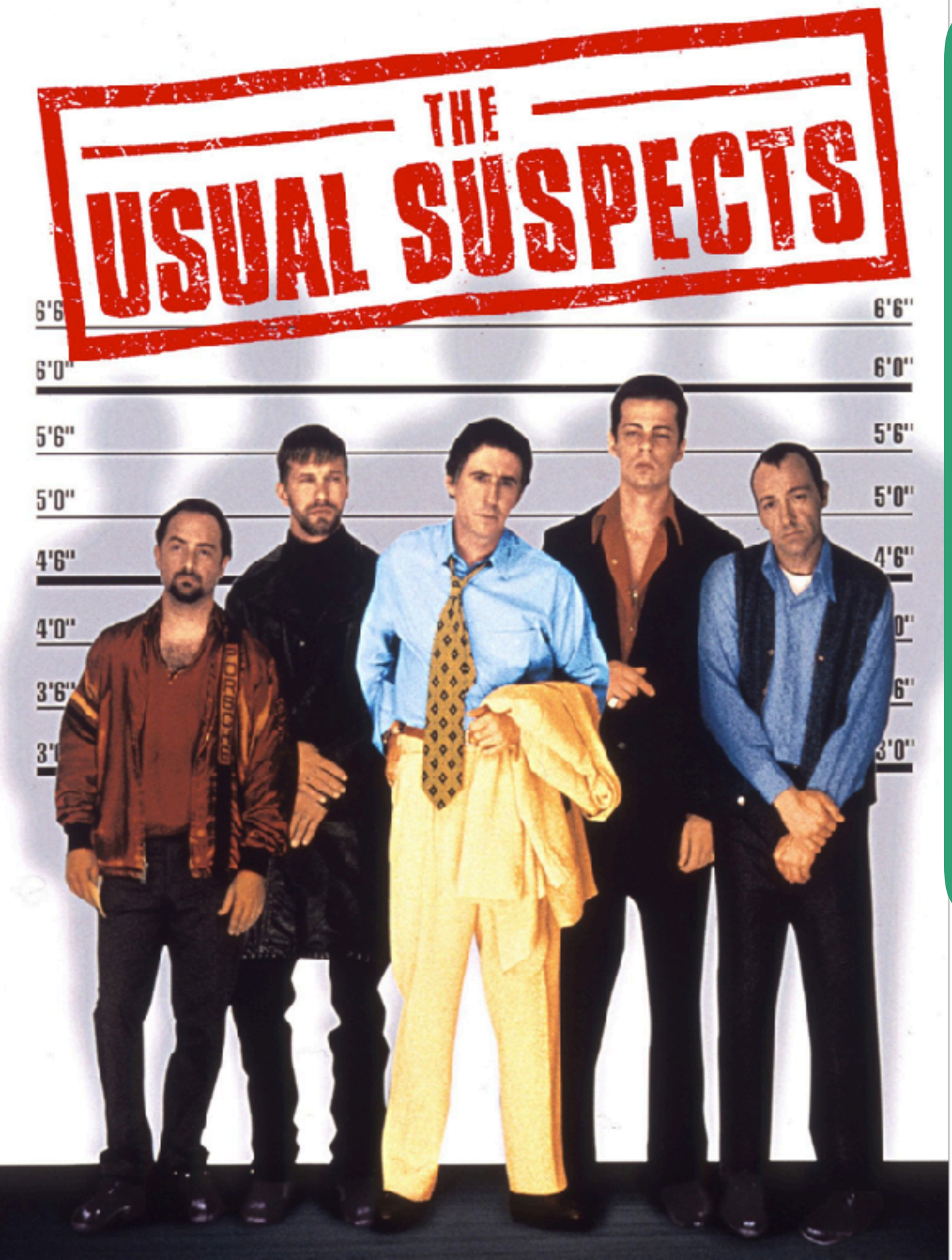


Experimental Method

Generate graphics



Select the one that is most converged

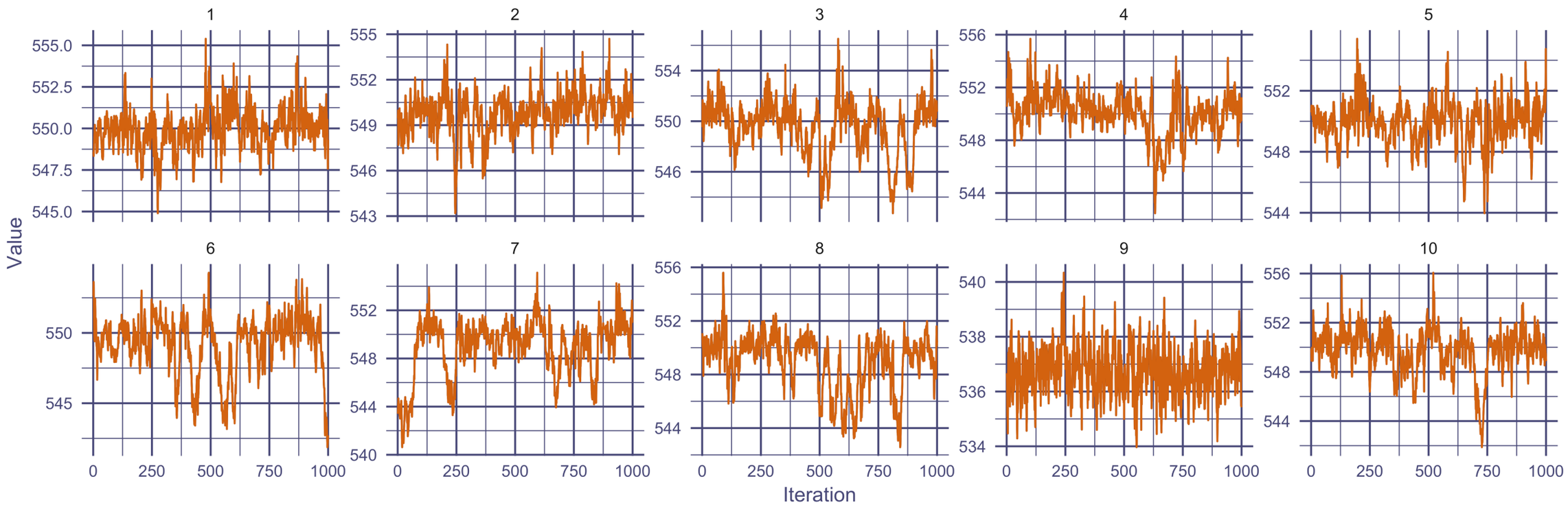


Statistical inference for exploratory data analysis and model diagnostic

Buja, A.et al (2009).

Experimental Method

Generate graphics



Select the one that is most converged

BAYES ON THE BEACH 2017

Miles McBain @MilesMcBain · 14 Nov 2017

"@#&$ I have to log on" The poster innovations at #bayesonbeach2017 continue @nj_tierney

1 2 21

Demo

Experience (years)

Please provide an estimate of the number of years experience you have with Bayesian Statistics

- (•) 0 - 1 years
- ( ) 1 - 2 years
- ( ) 2 - 3 years
- ( ) 3 - 4 years
- ( ) 4 - 5 years
- ( ) 5 plus years

< Previous Next >

Experience (Rank)

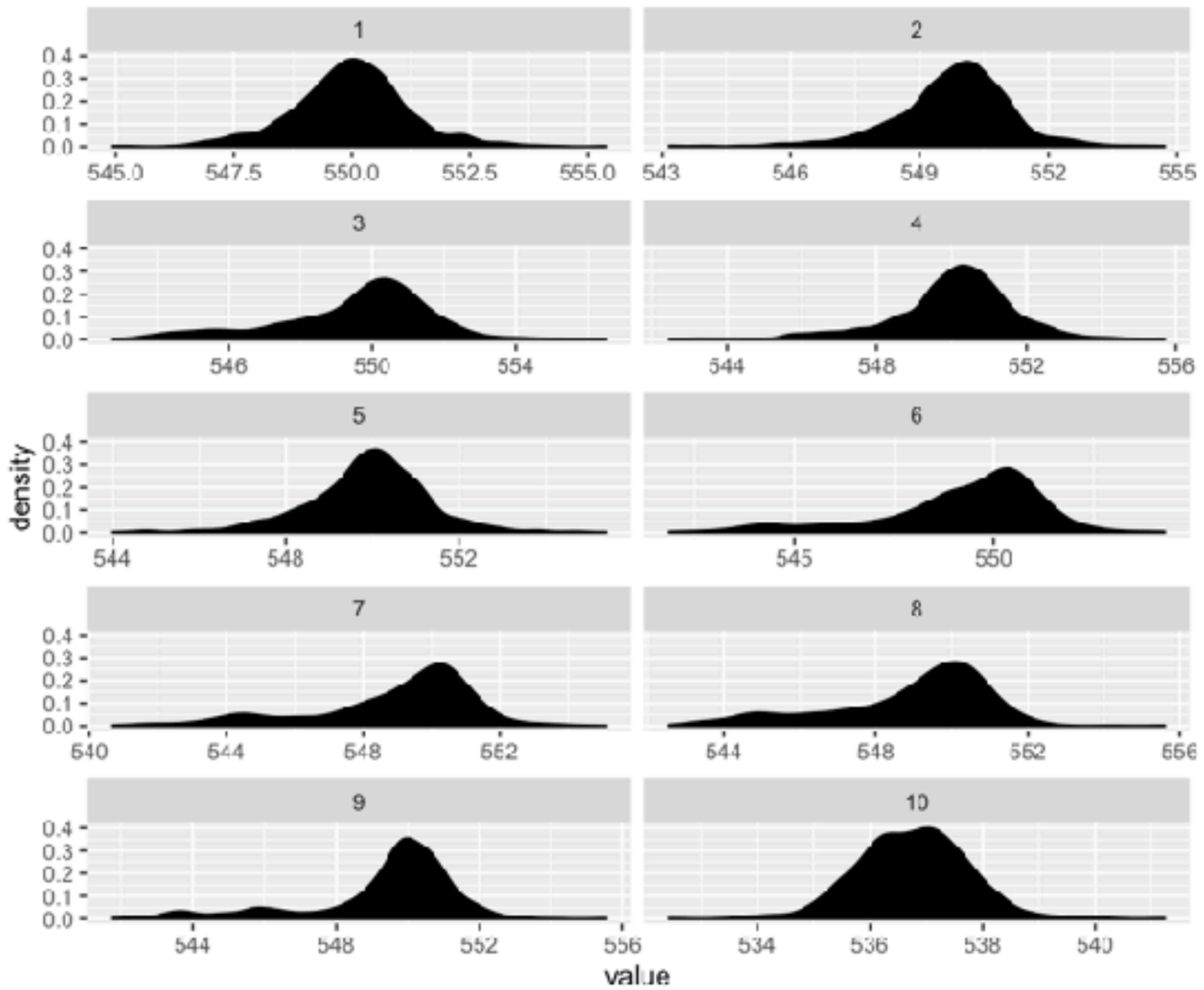
Please evaluate your experience with Bayesian Statistics, selecting one of the following

- ◉ No experience at all
- ○ Some experience
- ○ Moderate experience
- ○ Very experienced
- ○ Expert

< Previous | Next >

Density

Identify which graphic is the **most** converged

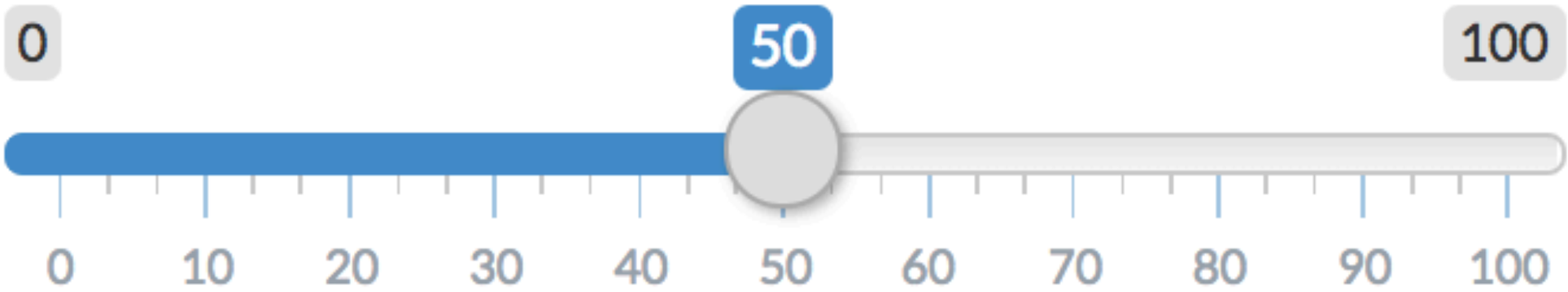


Choose one:

1 2 3 4 5 6 7 8 9 10

< Previous Next >

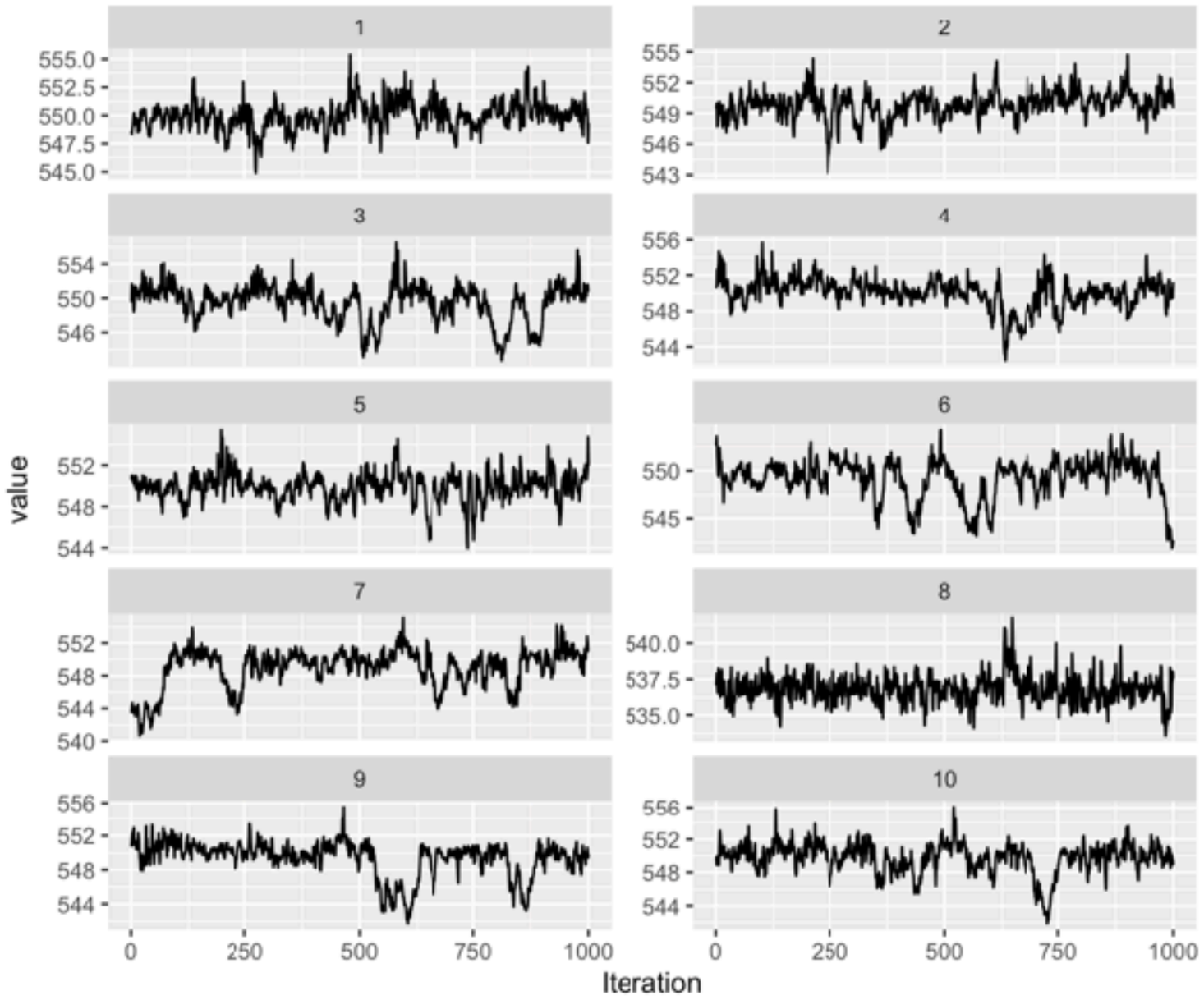
From 0 to 100, with 0 being 'no confidence at all' and 100 being 'absolutely confident', rank your confidence with this decision



< Previous

Next >

Identify which graphic is the **most** converged

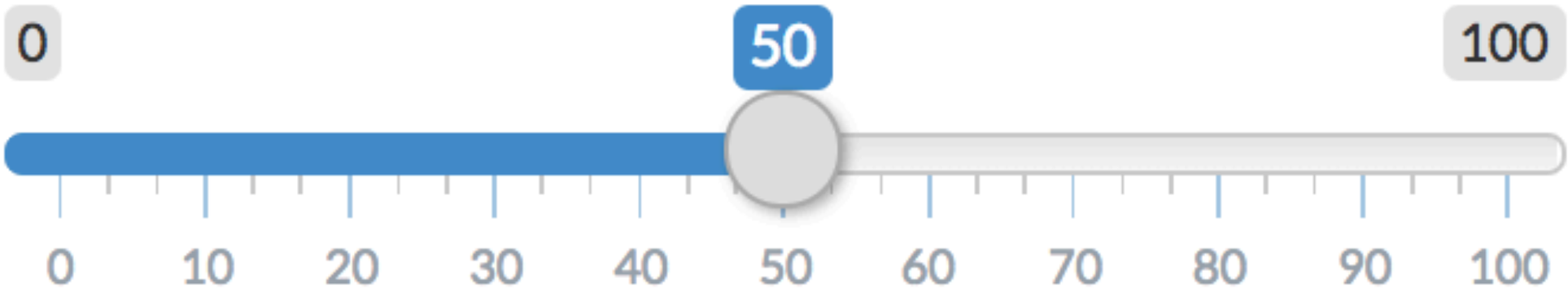


Choose one:

1 2 3 4 5 6 7 8 9 10

< Previous Next >

From 0 to 100, with 0 being 'no confidence at all' and 100 being 'absolutely confident', rank your confidence with this decision

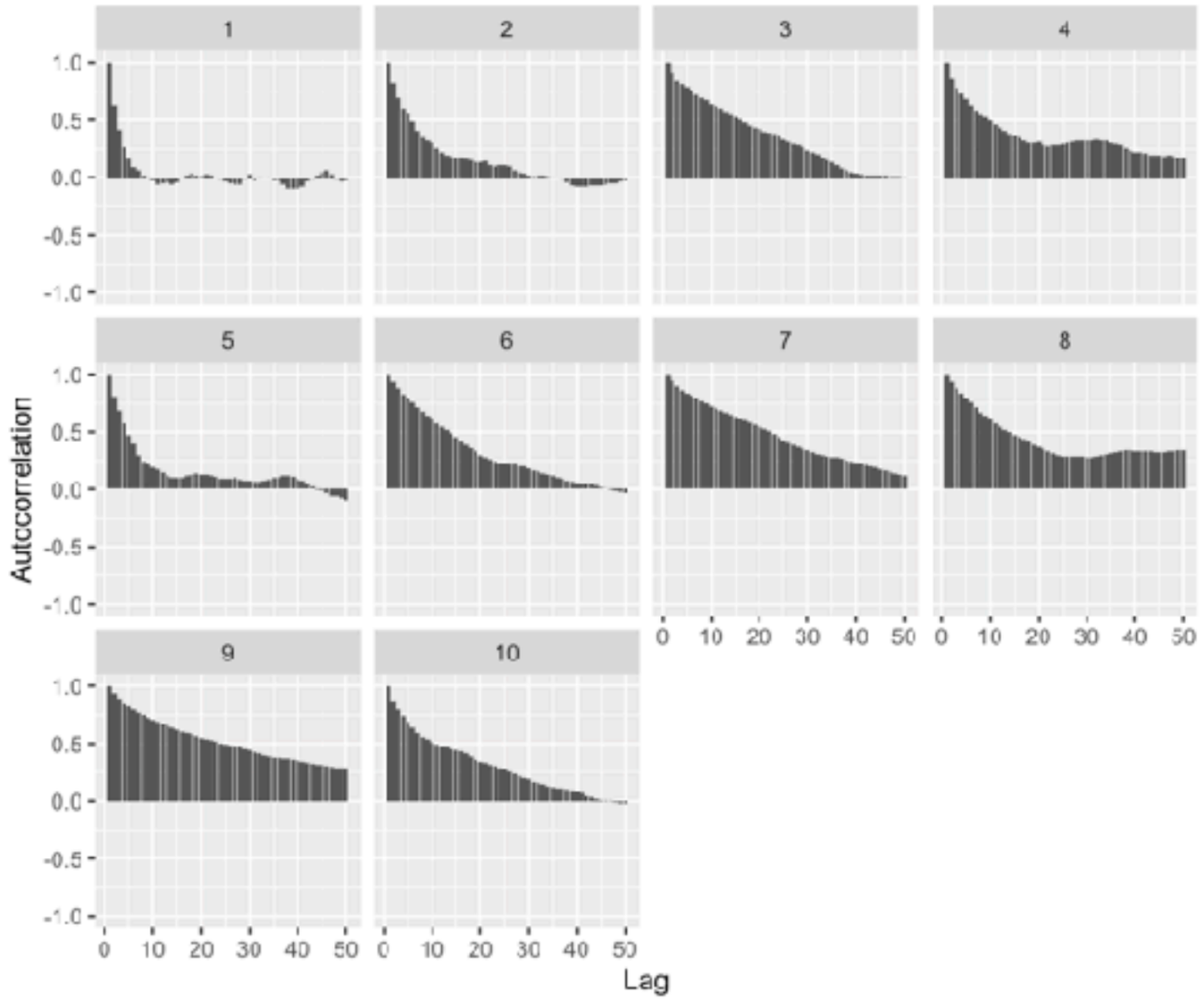


< Previous

Next >

Autocorrelation

Identify which graphic is the **most** converged

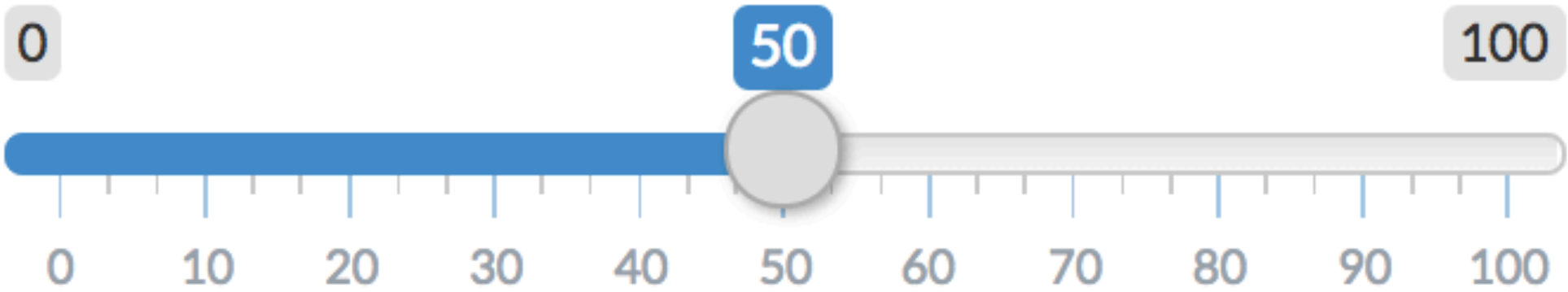


Choose one:

1 2 3 4 5 6 7 8 9 10

< Previous Next >

From 0 to 100, with 0 being 'no confidence at all' and 100 being 'absolutely confident', rank your confidence with this decision



< Previous

Next >

```
model{
   for( i in 1 : N ) {
      y[i] ~ dnorm(mu[i], tau)
      mu[i] <- lambda[T[i]]
      T[i] ~ dcat(P[])
      }
   P[1:maxT] ~ ddirch(alpha[])
   lambda[1] ~ dnorm(0.0, 1.0E-6)
   for (j in 2:maxT){
      lambda[j] <- lambda[j-1] + theta[j-1]
      theta[j-1] ~ dnorm(0.0, 1.0E-6)I(0.0, )
      }
   tau ~ dgamma(0.001, 0.001)
   sigma <- 1 / sqrt(tau)
}
```

Bad Model

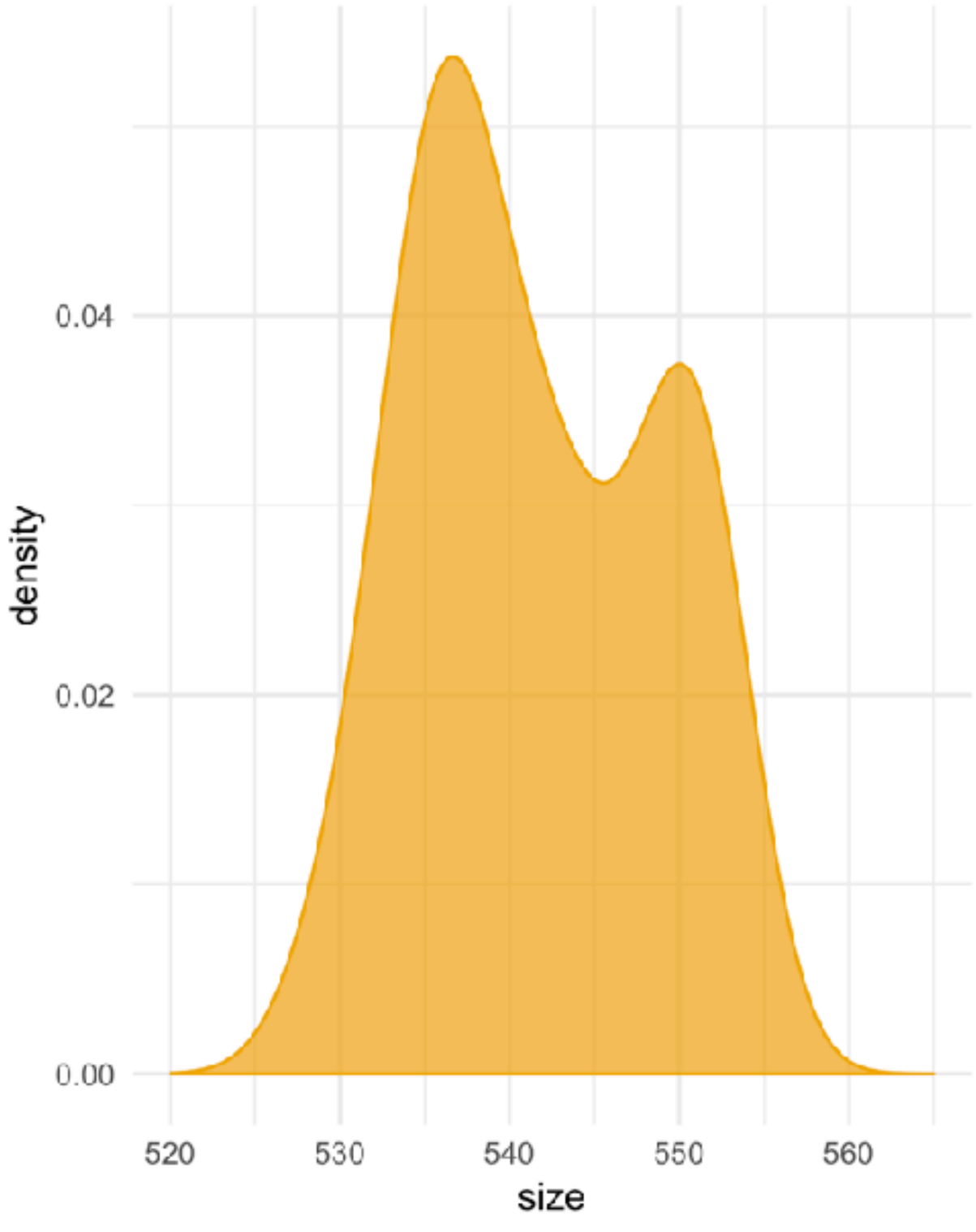
22 Mixtures

Good Model

2 Mixtures

The Data: “eyes”

Size	Group
529	NA
530	NA
532	NA
533	NA
534	NA
534	NA
534	NA
535	NA
535	NA



Results

Descriptive statistics

N = 16

Expertise	N
Some Experience	3
Moderate Experience	6
Very Experienced	4
Expert	1

Results

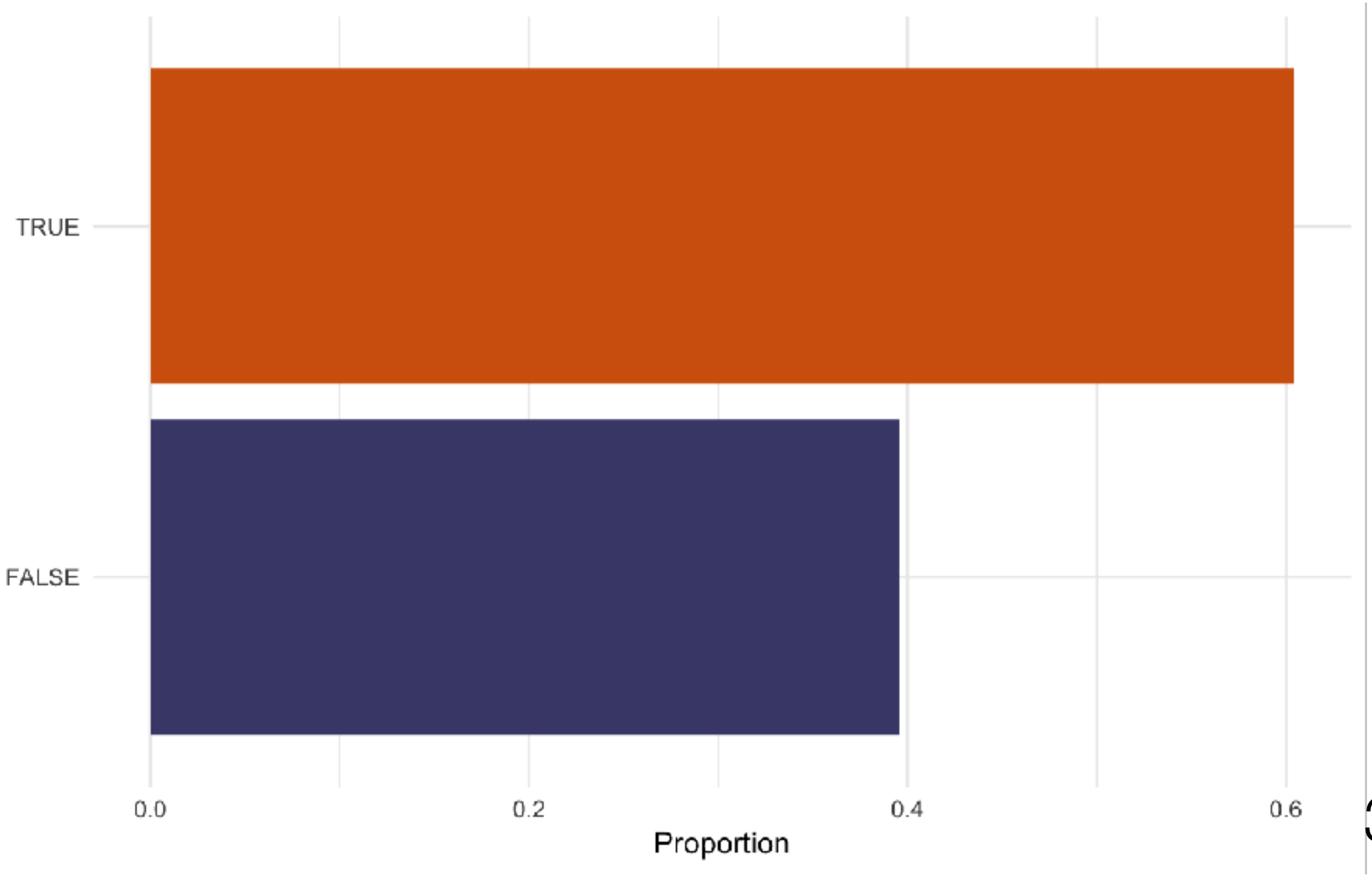
Descriptive statistics

N = 16

Expertise	N
0 - 1 years	2
1 - 2 years	1
2 - 3 years	2
3 - 4 years	3
5 plus years	8

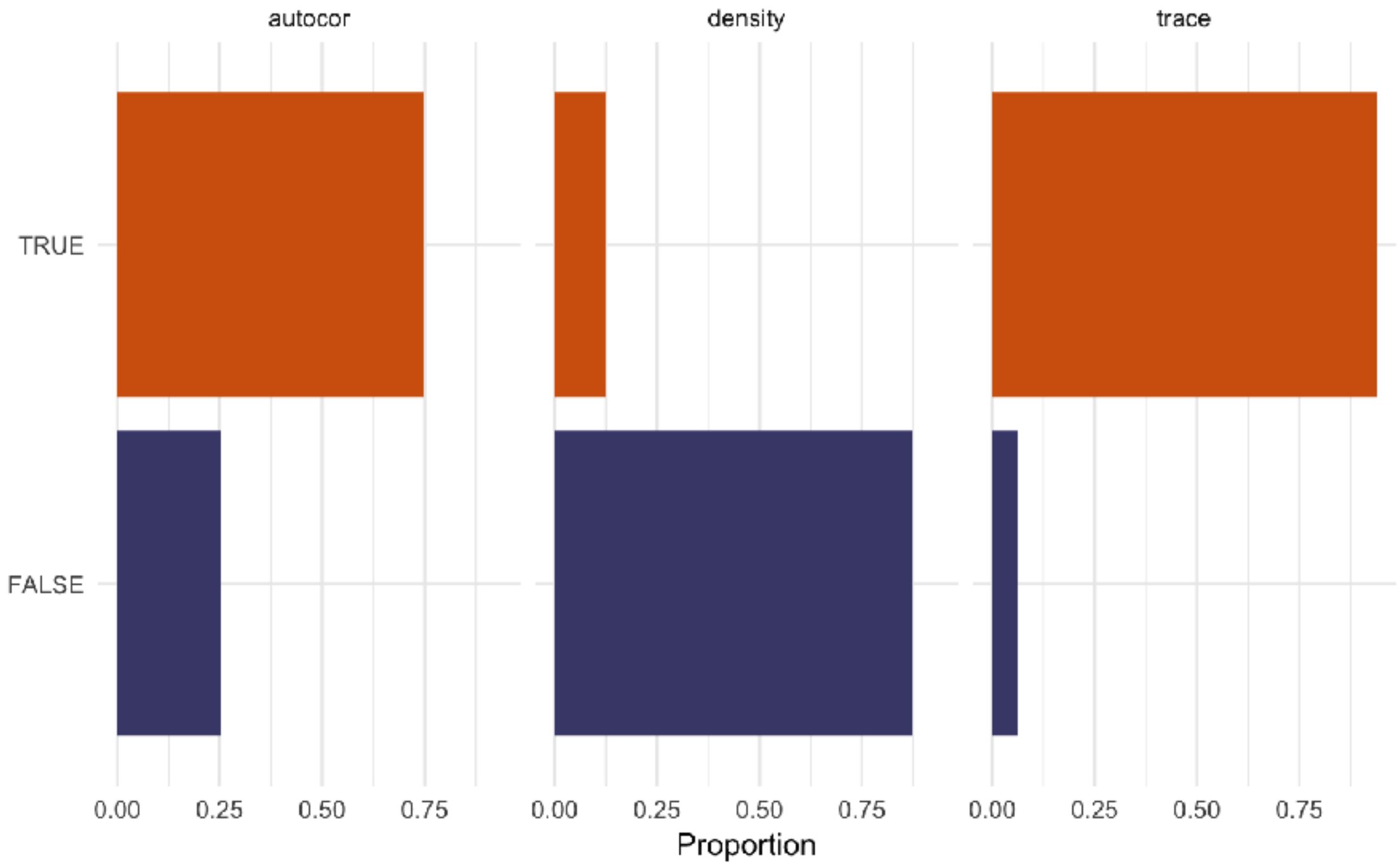
Results

Can visualisations be used to detect convergence?



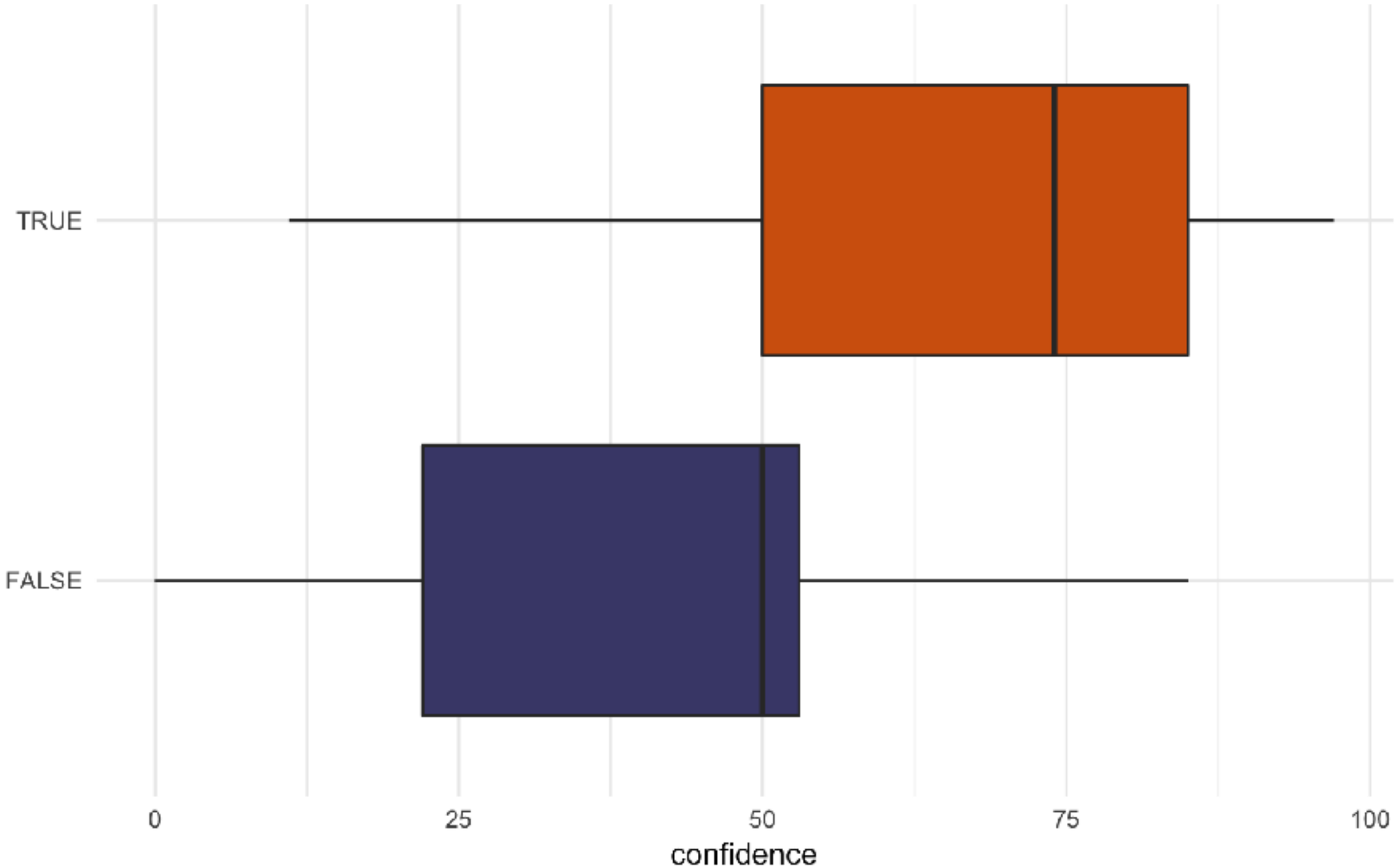
Results

Are some visualisations better than others?



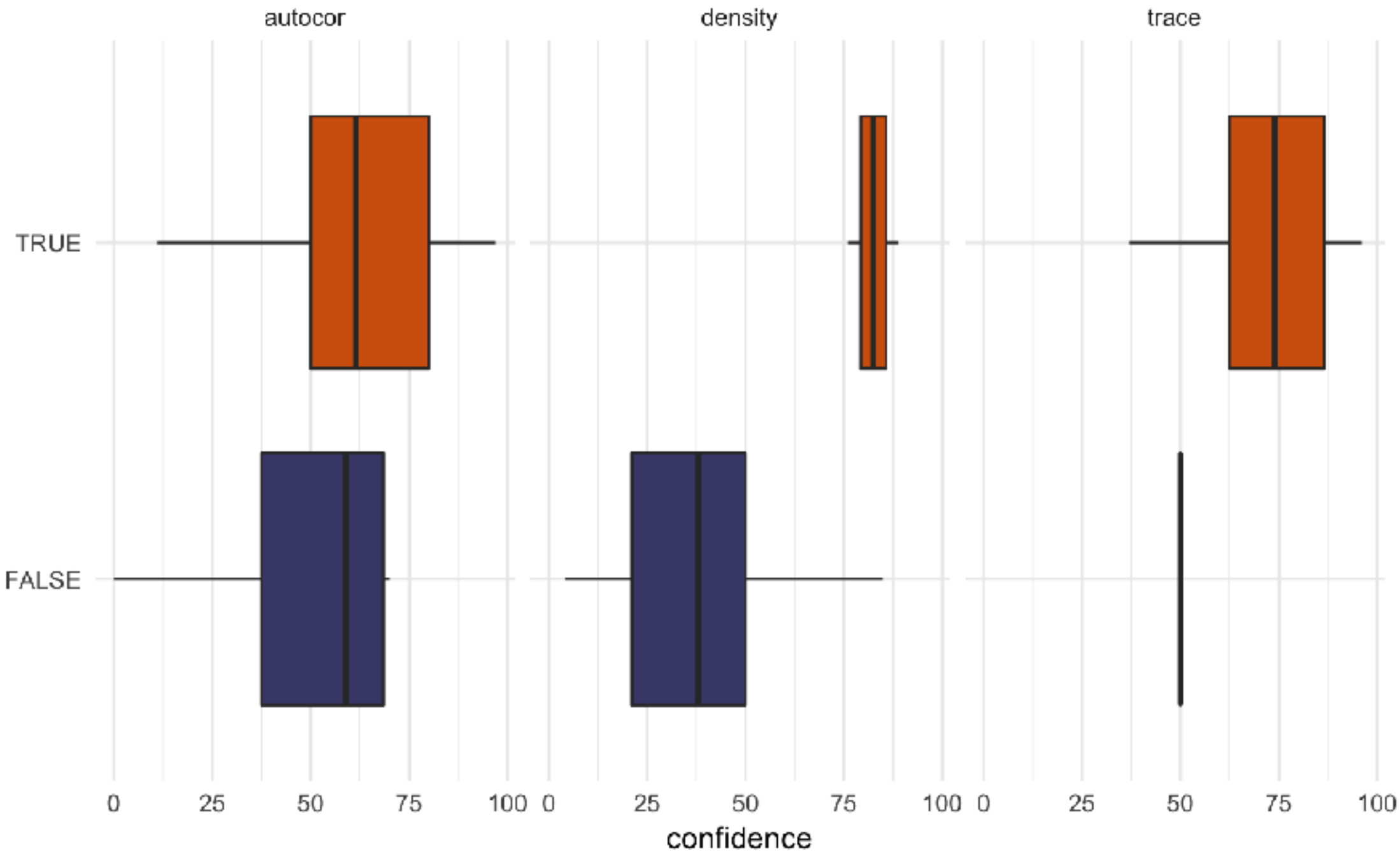
Results

Is confidence related to detecting convergence?



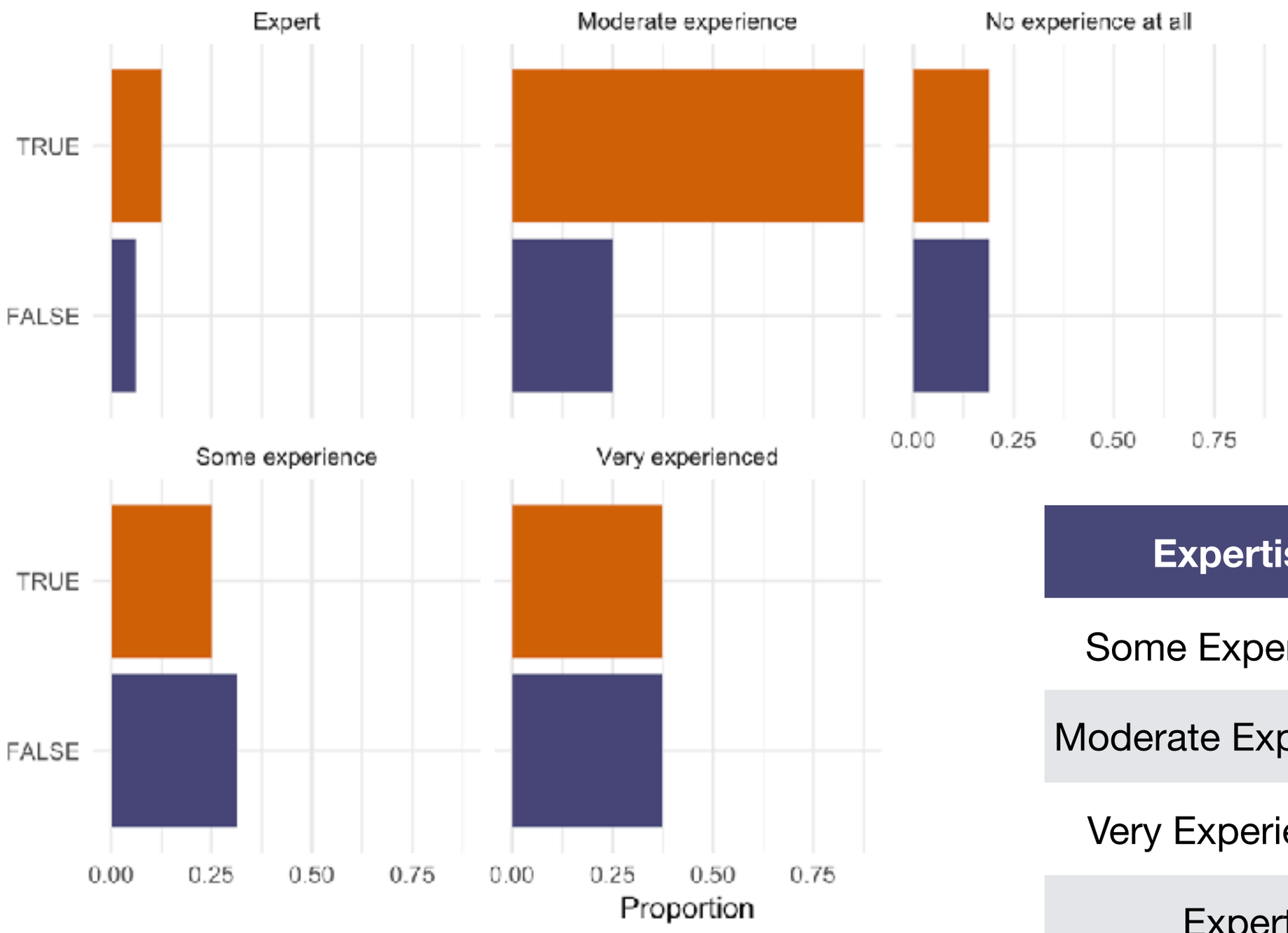
Results

Is confidence related to detecting convergence?



Results

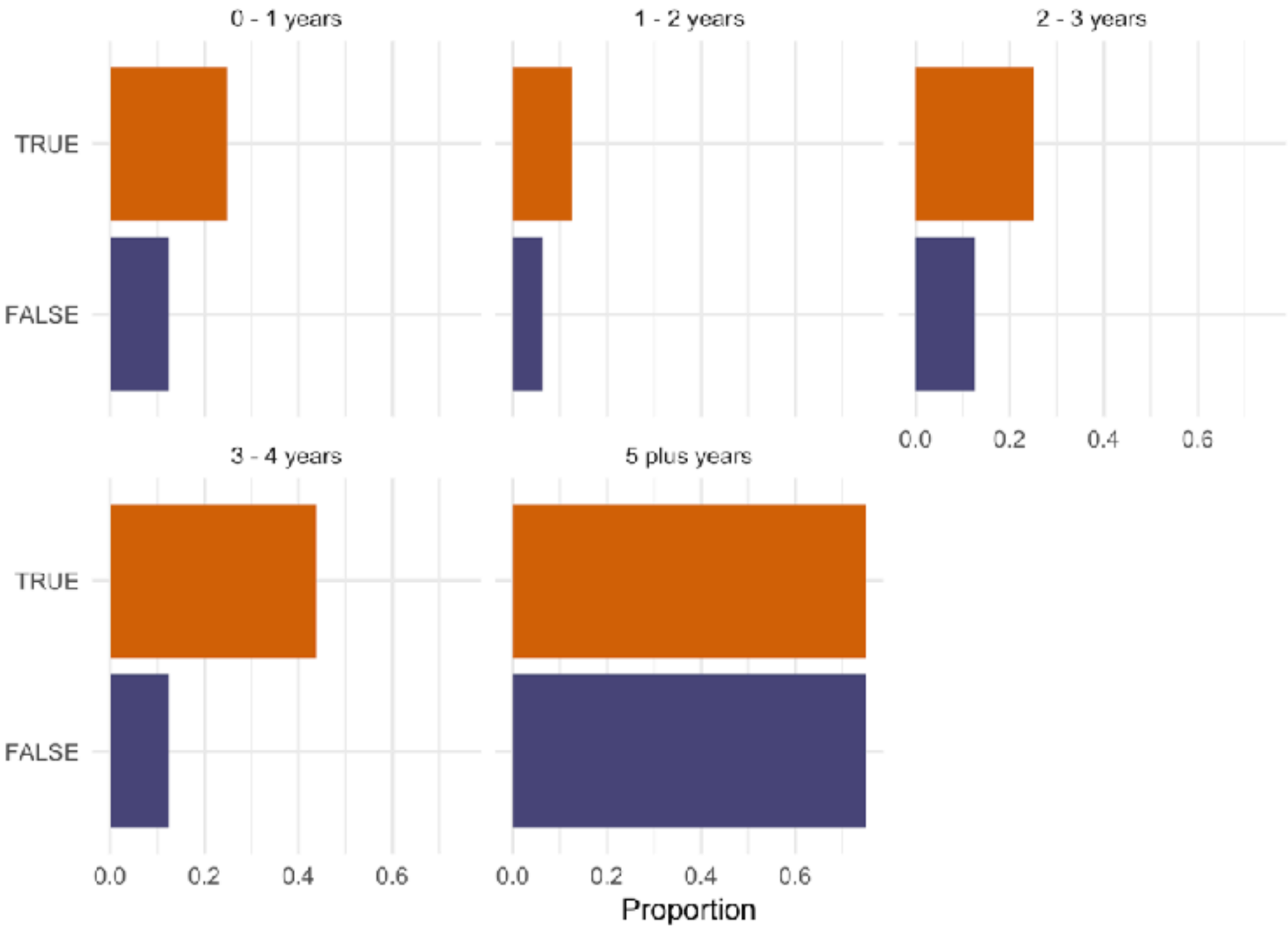
How Is Experience Related?



Expertise	N
Some Experience	3
Moderate Experience	6
Very Experienced	4
Expert	1

Results

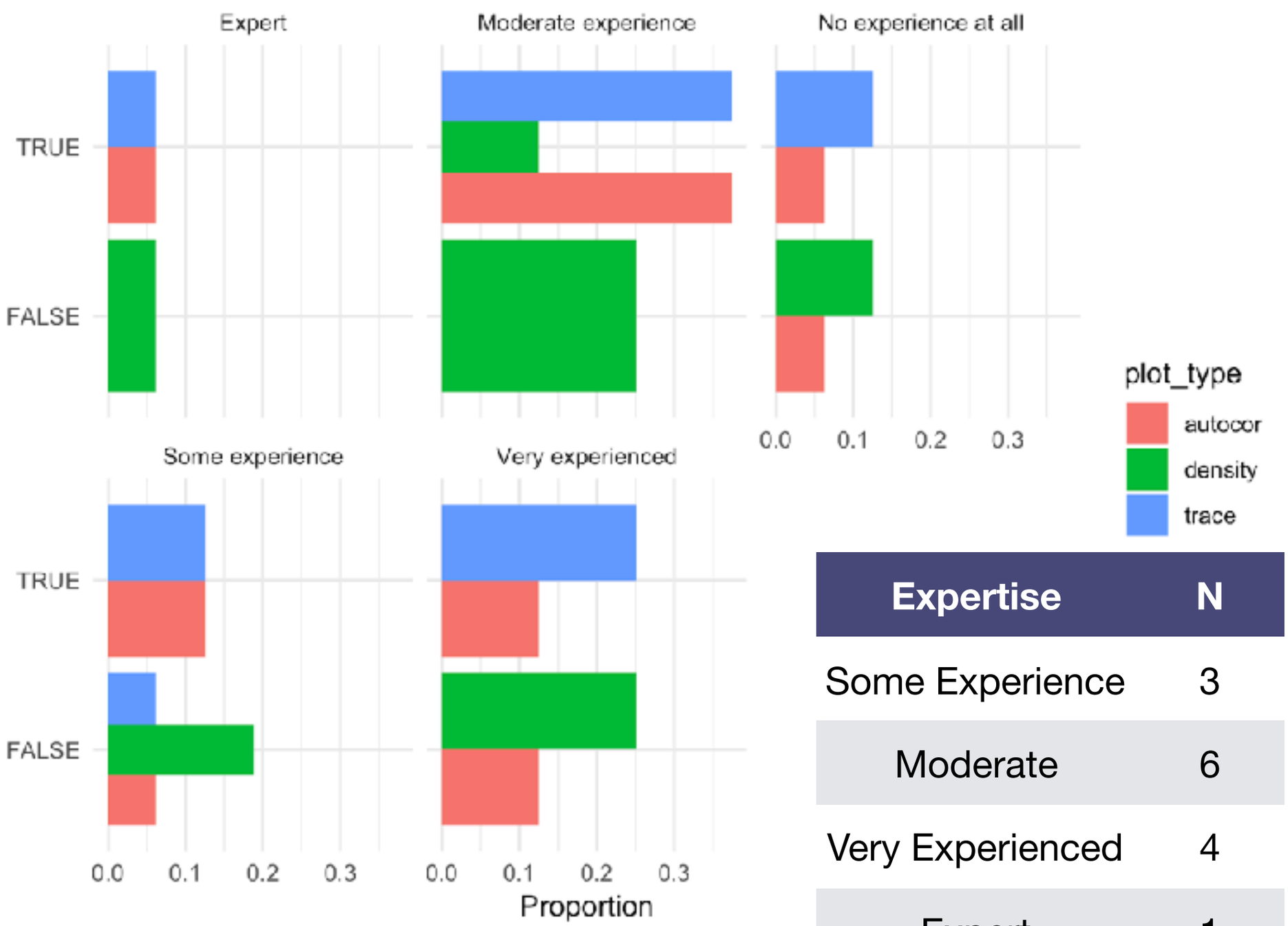
How Is Experience Related?



Expertise	N
0 - 1 years	2
1 - 2 years	1
2 - 3 years	2
3 - 4 years	3
5 plus years	8

Results

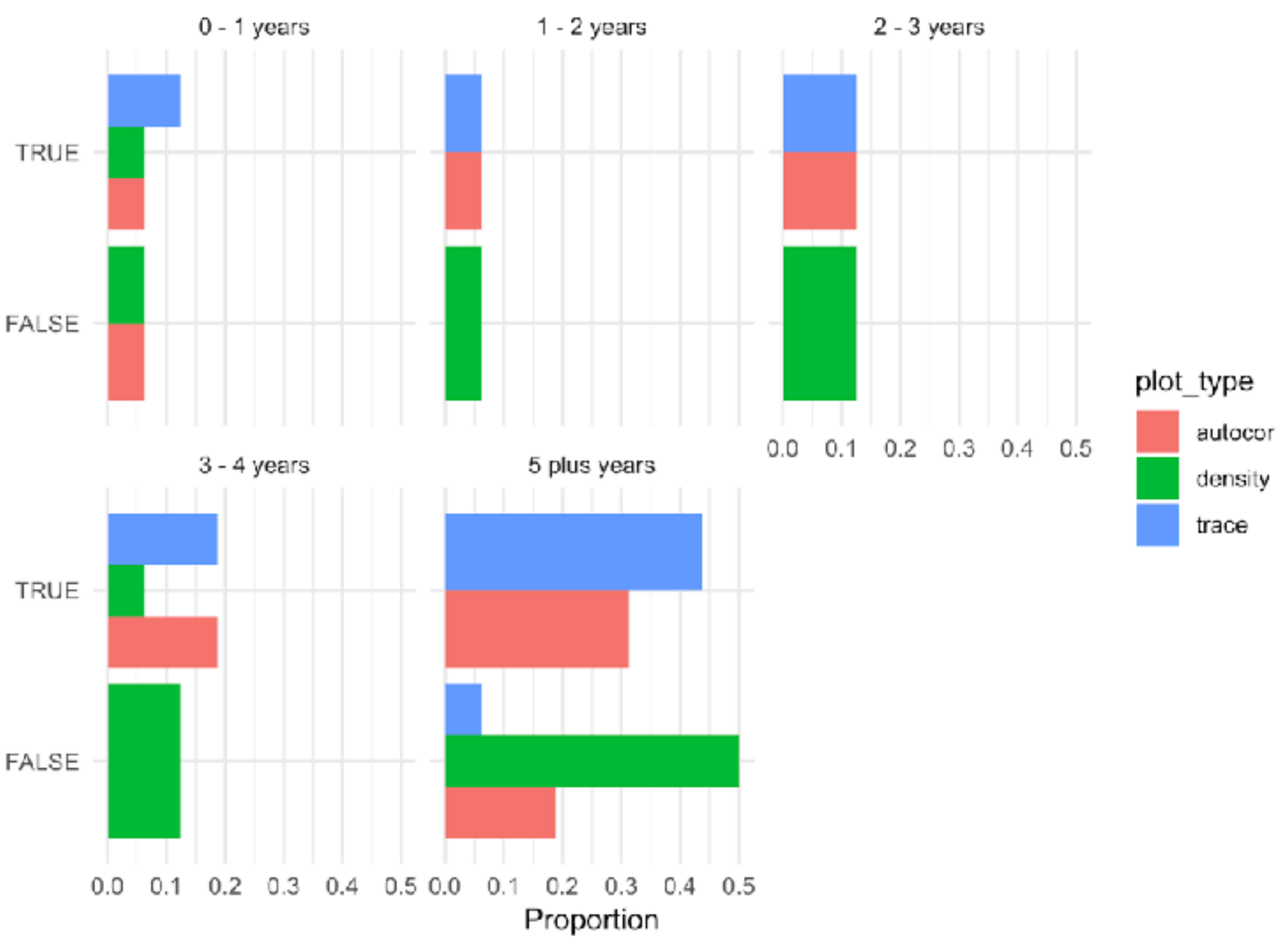
How Is Experience Related?



Expertise	N
Some Experience	3
Moderate	6
Very Experienced	4
Expert	1

Results

How Is Experience Related?



Expertise	N
0 - 1 years	2
1 - 2 years	1
2 - 3 years	2
3 - 4 years	3
5 plus years	8

Summary

Can visualisations be used to detect convergence?

Are some visualisations better than others for detecting convergence?

Is confidence related to detecting convergence?

Is experience with Bayesian statistics related to detecting convergence?

Summary

Can visualisations be used to detect convergence?

Are some visualisations better than others for detecting convergence?

Is confidence related to detecting convergence?

Is experience with Bayesian statistics related to detecting convergence?

Summary

Can visualisations be used to detect convergence?

Are some visualisations better than others for detecting convergence?

Is confidence related to detecting convergence?

Is experience with Bayesian statistics related to detecting convergence?

Summary

Can visualisations be used to detect convergence?

Are some visualisations better than others for detecting convergence?

Is confidence related to detecting convergence?

Is experience with Bayesian statistics related to detecting convergence?

Summary

Are they effective?

Are some visualizations more effective for convergence?

Is confidence related to detection of convergence?

Is experience with Bayesian statistics related to ability to assess convergence?

Summary

Are the

Are we there yet?

Is experience with Bayesian statistics related to

ce?

Summary

Are the

Are

Yeah, Nah

yet?

ive?

is experience wit stics related to u ce?

10 things I hate about ~~you~~ MCMC

10 things I ~~hate~~

learned about

~~you~~

~~MCMC~~

...Experiments?

#1: Shiny is not optimised for experiments

Sending/storing responses to a cloud is not trivial

Efficiently presenting many experiments and tracking responses is challenging

Do we track the time taken? Is that interesting? Is that difficult?

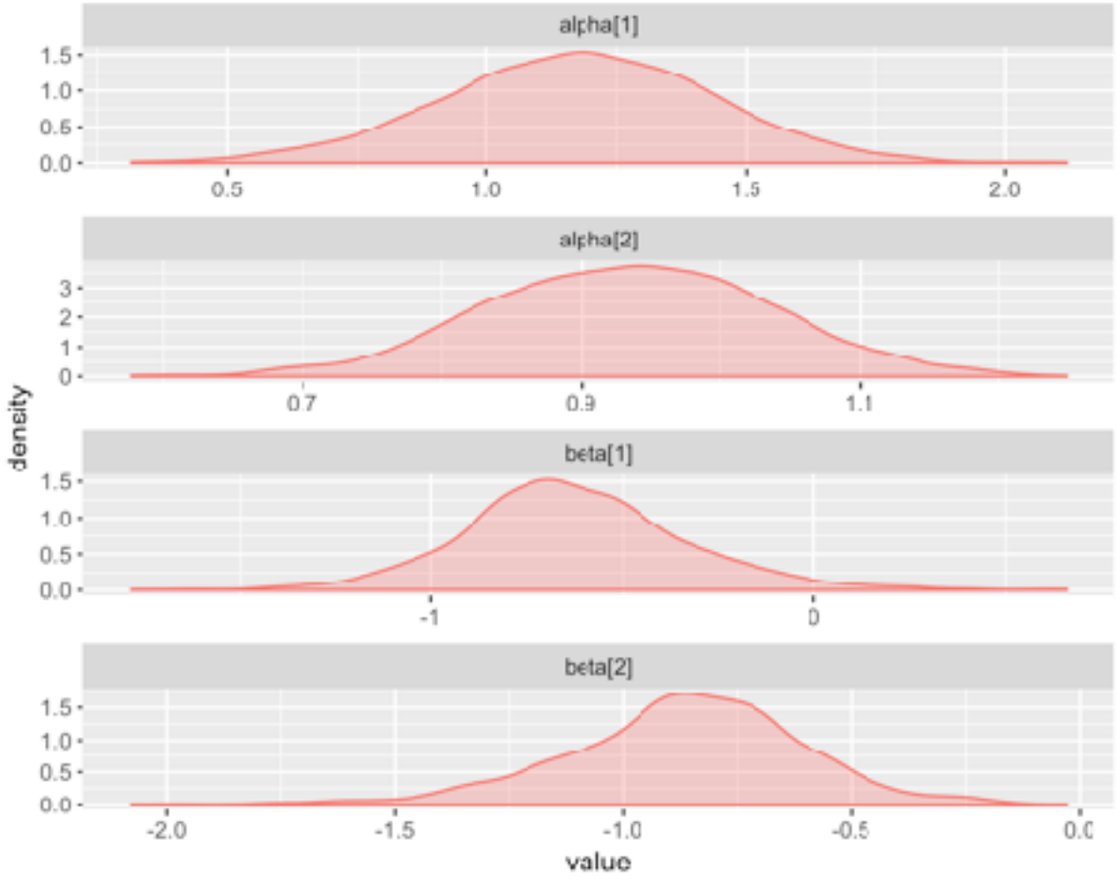
#2: Creating good and bad convergence?

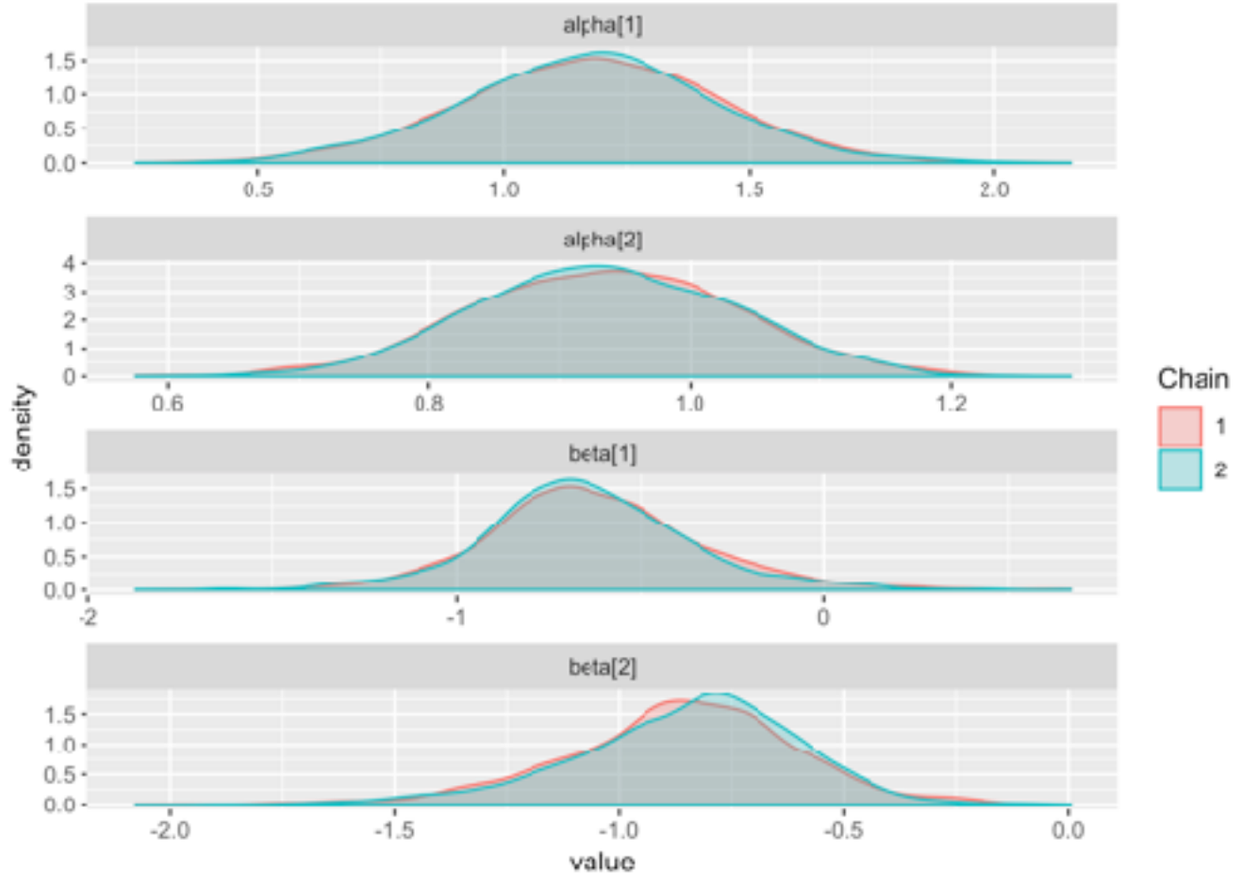
I have definitely created bad samples accidentally

I found it hard to create bad samples AND understand the implication

#3: More Chains

Looking at only one chain requires you to know more about the distribution





#4: Design and simulate the data first

We only had one observation per person per plot

Designing the model and simulating collected data would have helped

It also helps you understand what you are most interested in

#5: No context diagnostics are weird

But explaining and comprehending an entire Bayesian model takes time

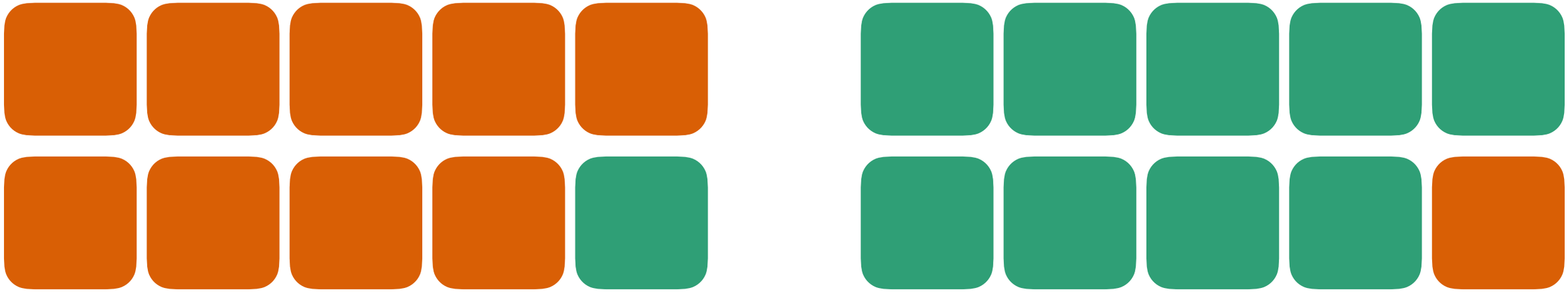
#6: Are these diagnostics real?

If a diagnostic fails and no one sees it, did it even diagnose anything?

Can we design better, more obvious / usable diagnostics?

#7: Are we measuring the right thing?

Should people be picking the good out of the bad, or the bad out of the good?



We are actually asking people to pick the mis-specified model, is that useful?

#8: We should do some training/testing

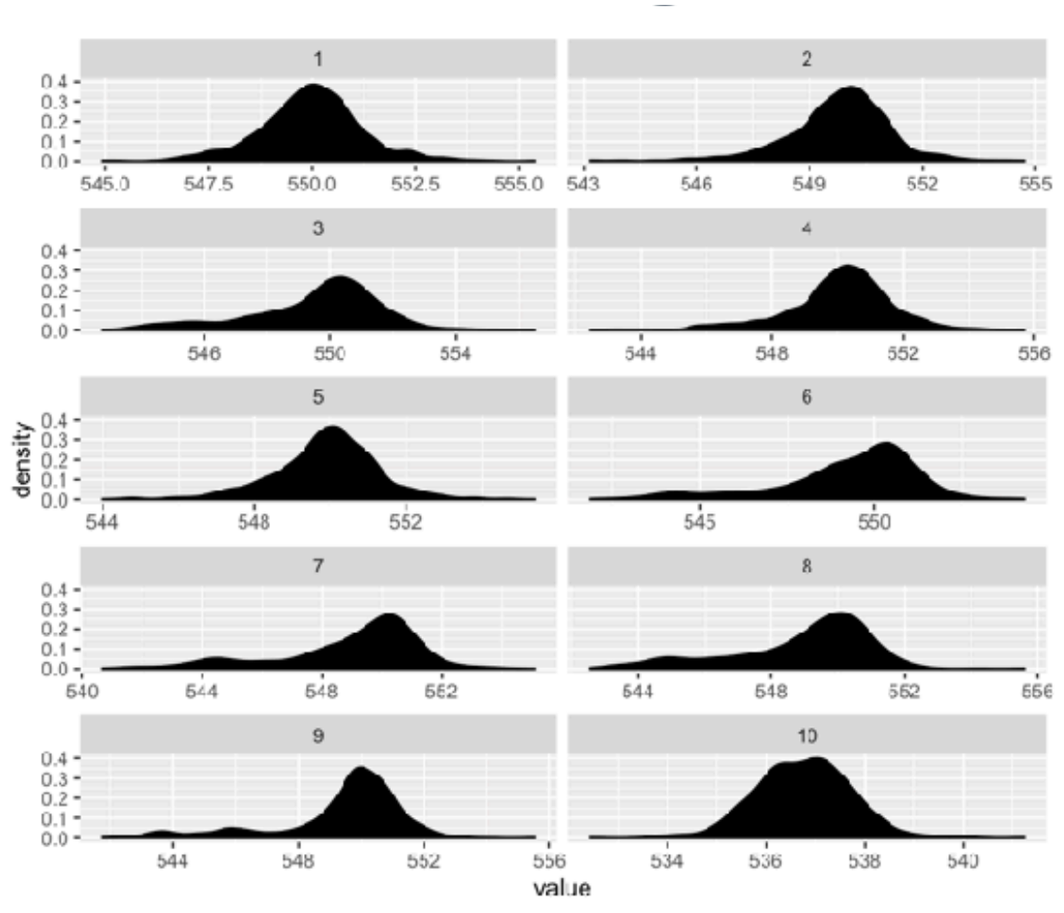
Explaining what convergence is

So that users can understand what is good and bad

#9: Comparing diagnostics?

Can we compare a single number diagnostic to vis?

0.99, 0.98
0.97, 0.96,
1.00, 0.95,
0.93, 0.92,
0.89, 0.88



#10: Your turn

Acknowledgements

Dr. David Frazier for his helpful suggestions on the methodology

Dr. Sam Clifford for discussions on the model implementation, and writing the JAGS code to create the “Good” and “Bad” models.

Colophon

Colours generated from the ochRe package

github.com/ropenscilabs/ochRe

Fonts used: Helvetica, impact.

References

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4361-4383.x

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of Markov Chain Monte Carlo*.

Loy, A., Hofmann, H., & Cook, D. (2017). Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners. *Journal of Computational and Graphical Statistics*

Learn more

njtierney/dualchain

nj_tierney

njtierney.com

nicholas.tierney@gmail.com