Bayes on the Beach 2024

COOLANGATTA GOLD

Bayesian mixture modelling...
of delicious mixtures for Bayesians.

DARE

ARC Training Centre in Data Analytics for Resources & Environments

Australian Government

Australian Research Council

Acknowledgement of Country

Today we begin by acknowledging the Traditional Custodians of the lands and waters of the Gold Coast; the Yugambeh people, and pay our respects to their Elders past, present and emerging.

As we share our own knowledge, teaching, learning and research practices within this conference, may we also pay respect to the knowledge embedded forever in the Aboriginal Custodianship of Country.

The University of Sydney | UNSW Sydney | The University of Western Australia | Macquarie University Sydney · Australia | Australian National University

DARE
ARC Training Centre in
Data Analytics for Resources & Environments

Australian Government
Australian Research Council

Bayes on the Beach Challenge:
Coolangatta Gold

Dr. Travis Stenborg
Research Engineer
DARE / University of Sydney

The University of Sydney | UNSW Sydney | The University of Western Australia | Macquarie University Sydney · Australia | Australian National University

Coolangatta Gold

Ironman Race

- Iconic Gold Coast surf lifesaving race.
- Organised by Surf Lifesaving Australia.

Australian Film

- 1984 film starring actors and real ironmen.

Bayes on the Beach Cocktail

- A fictional cocktail invented for this challenge!
- Based on a particular specialty Martini.

Alcohol Awareness

Moderation

- Alcohol can be fun, but…
- Consider your health.
- Drink responsibly.

Prerequisites

R + Stan

- Ensure R is installed.
  https://cran.csiro.au/

- Invoke Stan via the rstan package.
  https://mc-stan.org/users/interfaces/rstan

- RStudio may be helpful too.
  https://posit.co/download/rstudio-desktop/

Secret Ingredient

Coolangatta Gold

- 1.5 ounces *SECRET INGREDIENT*.
- 0.5 ounces yellow Chartreuse.
- 1 dash orange bitters.
- Garnish: pineapple slice.

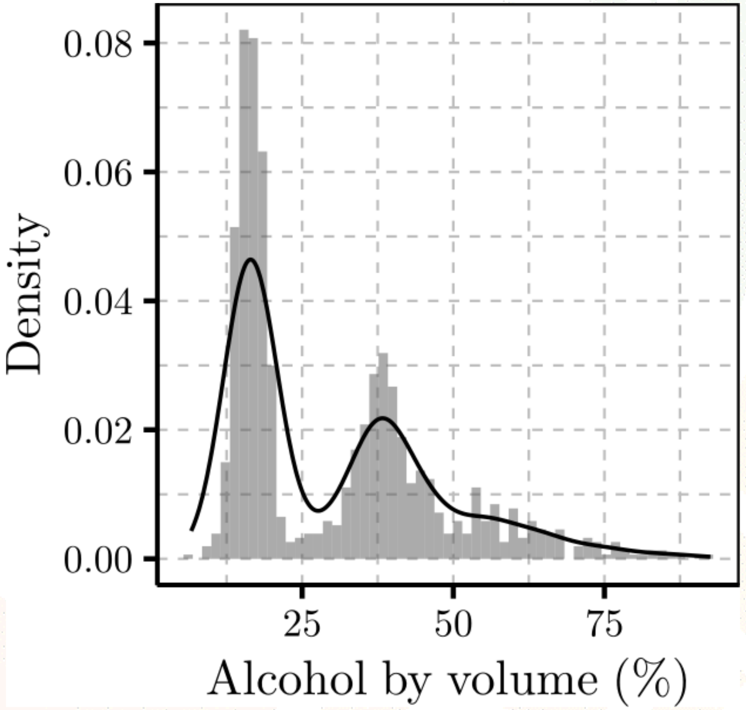
Challenge

- Find the secret ingredient.
- Use Bayesian mixture modelling.

Gaussian Mixture

Cocktail Testing

- A mixture of ≈1,000 test cocktails.
- They used a common base ingredient set.
- Each one also had a secret ingredient added.
- The secret (alcoholic) ingredient could vary between test cocktails.
- Small number of distinct secret ingredients.



Probability density function of the alcohol by volume for the secret ingredient in ≈1,000 test cocktails.

Imprecise Alcohol by Volume

- Different **inter**ingredient means.
- Dispersed **intra**ingredient values.

Possible Secret Ingredients

Ingredients *potentially* in the mixture model.

Ingredient	Relative σ (alcohol by vol %)
Gin	8.3
Pisco	1.0
Sherry	1.7
Vermouth	5.7
Vodka	2.3

The secret ingredient selected from amongst the candidates was the one with greatest mean alcohol by volume. What's the secret ingredient?

Problem Solving Process

Assumptions

- Gaussian mixture model.
- ≤ 5 components.

Please Answer…

- How many components are there?
  (And how did you determine the number?)
- What are their means, standard deviations and mixture weights? (And uncertainties?)
- What's the secret ingredient?

Solution Rigour

MCMC Convergence

- Describe your convergence assessment method(s).
- Traditional $\hat{R}$?

  Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci. 1992;7(4):457-72.
- Split-$\hat{R}$?

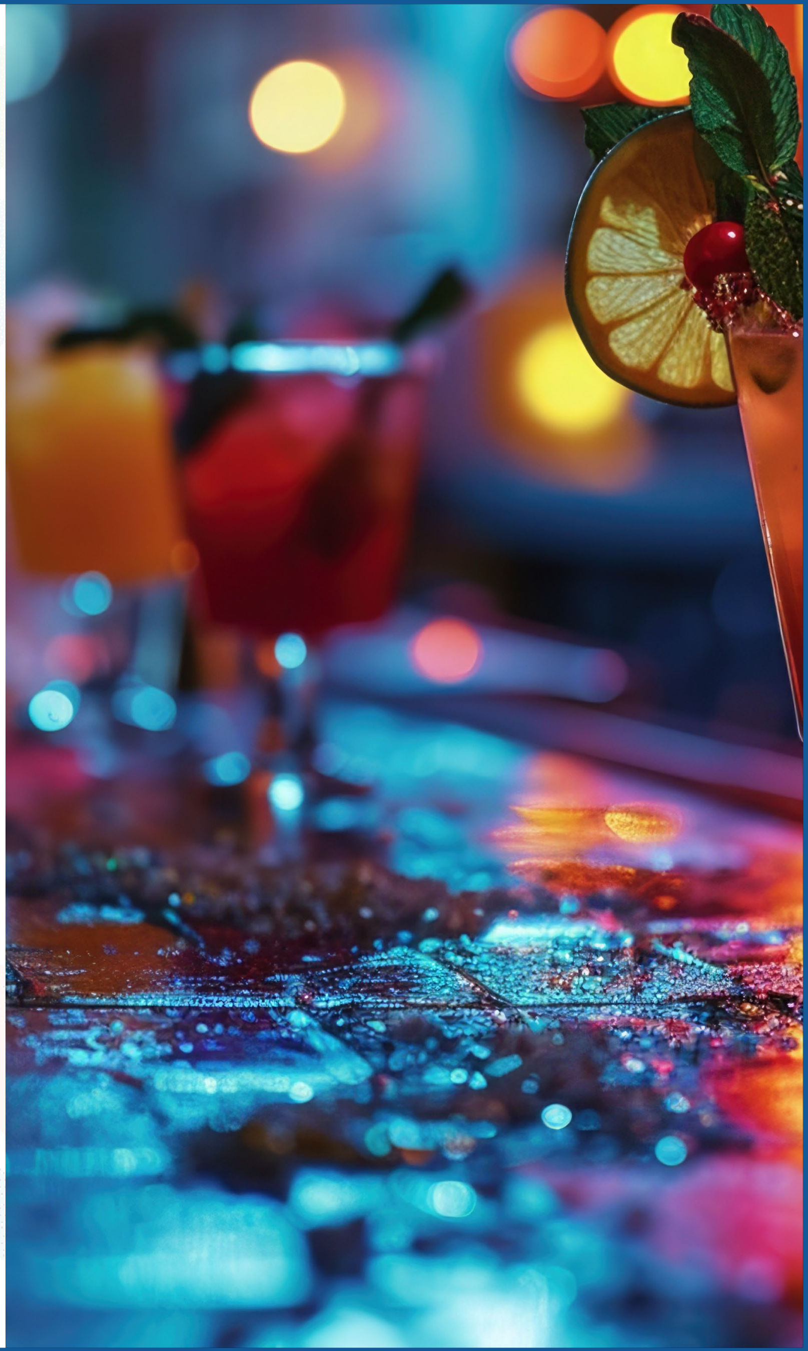
  Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed. Chapman & Hall/CRC; 2013.
- Is there something even better?

  Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC (with Discussion). Bayesian Anal. 2021;16(2):667-718.

Solution Rigour

MCMC Convergence

- What about effective sample size?
  (Is Markov chain convergence uniform across parameter space?)
  Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC (with Discussion). Bayesian Anal. 2021;16(2):667-718.
- Have you considered a graphical metric?



Performance Tuning

Solution Speed

- Points will be awarded for fast code.
- Efficient choice of cores / chains to use is typically a function of model complexity and data volume.
- Can that choice be automated?

  Stenborg TN. Adaptive MCMC parallelisation in Stan. In: MODSIM2023, 25th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand; 2023. p. 913. [Extended Abstract].

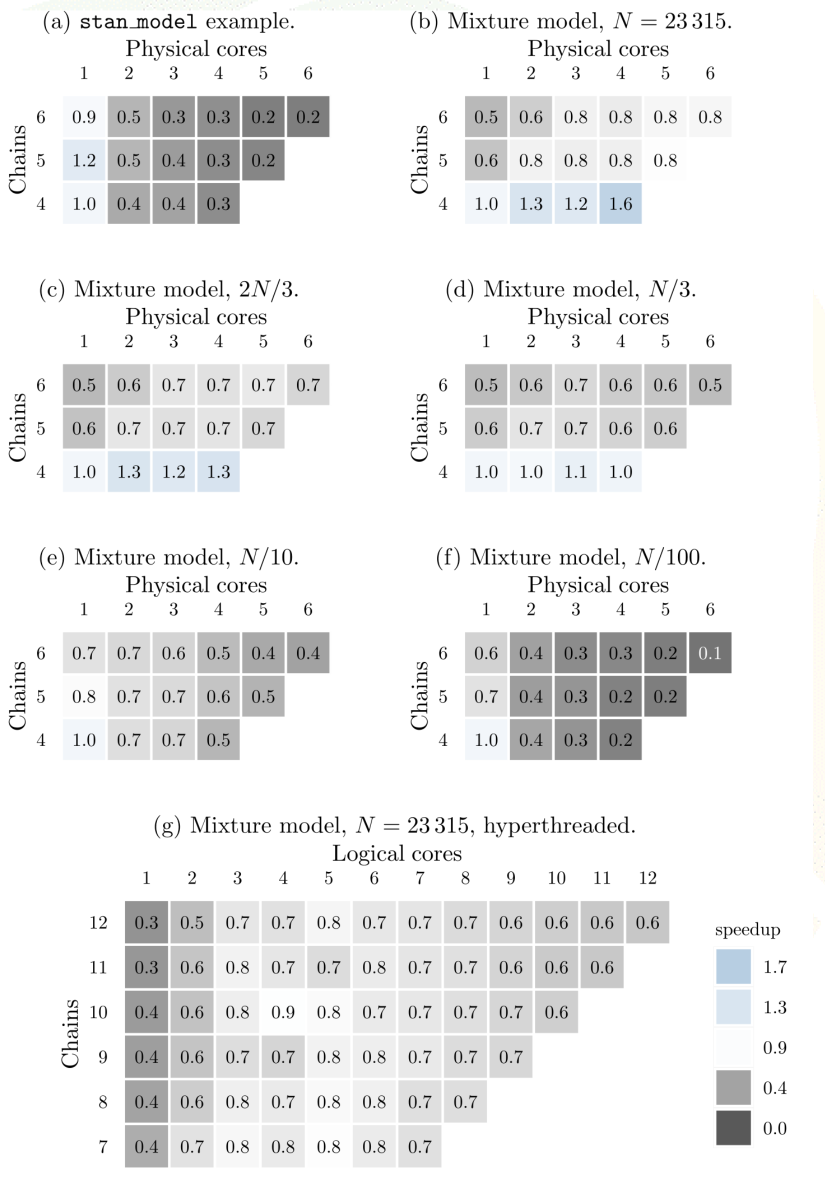


Figure 14: **MCMC parallelisation** and `Stan` model inference speedup. Sampling used 50 warmup and $50/m$ post-warmup iterations/chain.

Performance Tuning

Challenge Results

Windows 11, on an Intel “Comet Lake” Core i7-10750H, 16GB RAM system.

Physical cores

Chains	1	2	3	4	5	6
6	0.7	1.0	1.5	1.5	2.1	1.4
5	0.9	1.3	1.6	1.7	1.5	
4	1.0	1.5	1.6	2.2		

Logical cores

Chains	1	2	3	4	5	6	7	8	9	10	11	12
12	0.4	0.7	0.8	1.0	1.0	1.1	1.1	1.5	1.3	1.0	1.0	1.0
11	0.4	0.6	0.8	0.8	1.1	1.1	1.1	1.1	1.5	1.3	1.0	
10	0.4	0.6	0.9	0.9	1.1	1.1	1.1	1.1	1.1	1.5		
9	0.5	0.9	1.0	1.1	1.4	1.3	1.3	1.1	1.1			
8	0.6	1.0	1.0	1.4	1.4	1.3	1.4	1.3				
7	0.6	1.0	1.5	1.5	1.4	1.6	1.7					

speedup

2.2

1.7

1.1

0.6

0.0

Solution Implementation

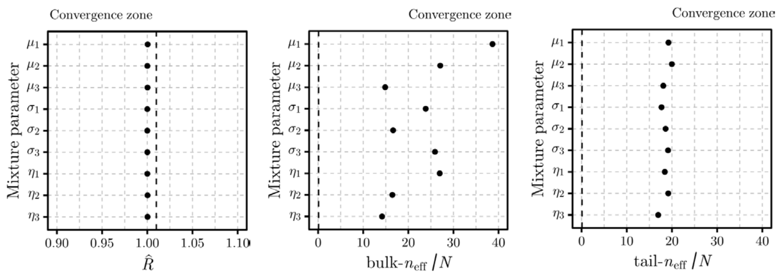
- The team tried:
  Stan, brms, PyMC.

- Auto-optimised
  Stan convergence...

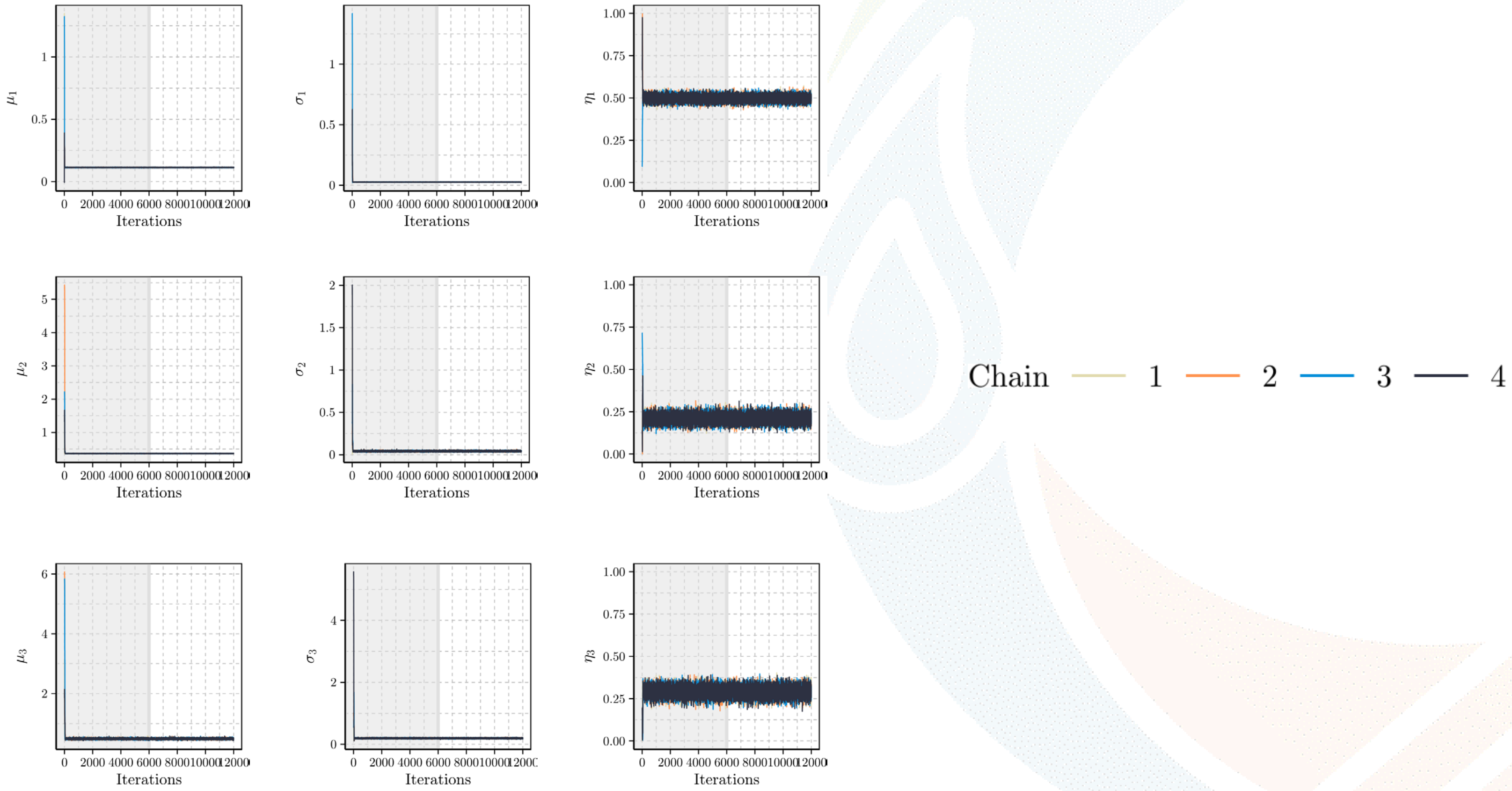
```
> print(fit_synth)
Inference for Stan model: anon_model.
4 chains, each with iter=12000; warmup=6000; thin=1;
post-warmup draws per chain=6000, total post-warmup draws=24000.

          mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
mu[1]     0.11    0.00 0.00   0.11   0.11   0.11   0.11   0.12 38656    1
mu[2]     0.37    0.00 0.00   0.36   0.36   0.37   0.37   0.37 27091    1
mu[3]     0.48    0.00 0.02   0.45   0.47   0.48   0.50   0.52 14671    1
sigma[1]  0.03    0.00 0.00   0.02   0.03   0.03   0.03   0.03 23637    1
sigma[2]  0.04    0.00 0.01   0.03   0.04   0.04   0.04   0.05 16598    1
sigma[3]  0.19    0.00 0.01   0.17   0.19   0.19   0.20   0.21 25870    1
eta[1]    0.50    0.00 0.02   0.47   0.49   0.50   0.51   0.53 27036    1
eta[2]    0.21    0.00 0.02   0.17   0.19   0.21   0.23   0.26 16267    1
eta[3]    0.29    0.00 0.03   0.24   0.27   0.29   0.31   0.34 14101    1
lp__    738.96    0.02 2.01 734.14 737.87 739.30 740.43 741.87 11252    1

Samples were drawn using NUTS(diag_e) at Wed Feb  7 10:37:57 2024.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
> print(summarise_draws(fit_synth))
# A tibble: 10 × 10
   variable     mean   median      sd     mad      q5     q95  rhat ess_bulk ess_tail
   <chr>       <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>    <dbl>    <dbl>
 1 mu[1]       0.114    0.114 0.00122 0.00121   0.112   0.116  1.00   38622.   19208.
 2 mu[2]       0.365    0.365 0.00422 0.00422   0.359   0.372  1.00   27048.   19962.
 3 mu[3]       0.485    0.484 0.0180  0.0177    0.457   0.515  1.00   14853.   18069.
 4 sigma[1]    0.0258   0.0258 0.00101 0.00101  0.0242  0.0275 1.00   23829.   17678.
 5 sigma[2]    0.0415   0.0413 0.00513 0.00503  0.0334  0.0503 1.00   16611.   18575.
 6 sigma[3]    0.192    0.192 0.00973 0.00974   0.177   0.209  1.00   25889.   19118.
 7 eta[1]      0.498    0.498 0.0168  0.0169    0.471   0.526  1.00   26933.   18399.
 8 eta[2]      0.210    0.209 0.0234  0.0234    0.172   0.249  1.00   16442.   19163.
 9 eta[3]      0.292    0.292 0.0271  0.0273    0.247   0.336  1.00   14164.   16942.
10 lp__      739.     739.   2.01    1.85     735.     742.     1.00   11402.   15427.
```



Solution Implementation



DARE
ARC Training Centre in
Data Analytics for Resources & Environments

Australian Government
Australian Research Council

https://github.com/dare-centre/BOTB-2024-data-challenge

THE UNIVERSITY OF SYDNEY
UNSW SYDNEY
THE UNIVERSITY OF WESTERN AUSTRALIA
MACQUARIE University SYDNEY · AUSTRALIA
Australian National University