A Factor Model with Paired Comparison Indicators

Joshua N. Pritikin

Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University

May 2019

Acknowledgment

Some mentors and collaborators

- Mike Neale
- Rob Kirkpatrick

Funded in part by NIDA R25 DA-26119

How to plan a study

A pre-registered analysis plan

- Model chosen
- Fake data simulated
- Analyses scripts written

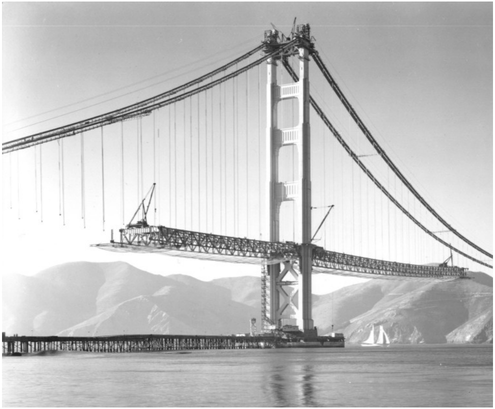
Not pre-registered but

- Model chosen
- Fake data simulated
- Analyses scripts somewhat written

Thorough and complete mental plan

Delegated to an expert

Vague Plan

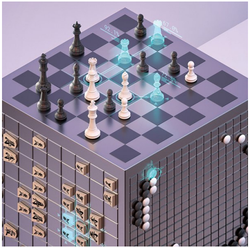


- Collect data
- Figure out the model later

Magical thinking

Reality sets in

Ranking skill



Chess, Go, shogi, etc

$$i, j \in \{1, 2, \ldots, N\} \quad (1)$$

$$\pi(i > j) = \frac{\theta_i}{\theta_i + \theta_j} \quad (2)$$

where N is the number of players and θ is some measure of skill[1]

Also similar to Thurstone (1927)

[1] Bradley and Terry (1952); Luce (1959)

A factor model

RESEARCH

COMPUTER SCIENCE

A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

David Silver[1,2*†], Thomas Hubert[1*], Julian Schrittwieser[1*], Ioannis Antonoglou[1], Matthew Lai[1], Arthur Guez[1], Marc Lanctot[1], Laurent Sifre[1], Dharshan Kumaran[1], Thore Graepel[1], Timothy Lillicrap[1], Karen Simonyan[1], Demis Hassabis[1†]

The game of chess is the longest-studied domain in the history of artificial intelligence. The strongest programs are based on a combination of sophisticated search techniques, domain-specific adaptations, and handcrafted evaluation functions that have been refined by human experts over several decades. By contrast, the AlphaGo Zero program recently achieved superhuman performance in the game of Go by reinforcement learning from self-play. In this paper, we generalize this approach into a single AlphaZero algorithm that can achieve superhuman performance in many challenging games. Starting from random play and given no domain knowledge except the game rules, AlphaZero convincingly defeated a world champion program in the games of chess and shogi (Japanese chess), as well as Go.

Prior work is mostly univariate

Thurstone (1927) ≈ 6000 citations

Bradley and Terry (1952) ≈ 2500 citations

Luce (1959) ≈ 6500 citations

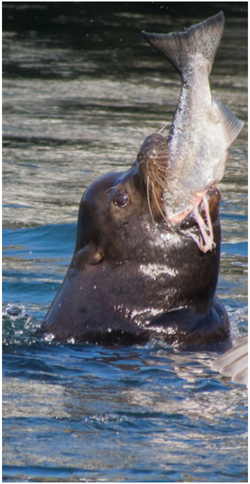
Association/correlational model (not a factor model)

- Bockenholt (1988) ≈ 15 citations
- Dittrich, Francis, Hatzinger, and Katzenbeisser (2006) ≈ 15 citations

Ship's manifest

- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

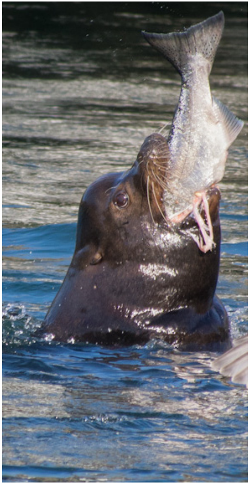
† Two prominent
methods for Bayesian model validation



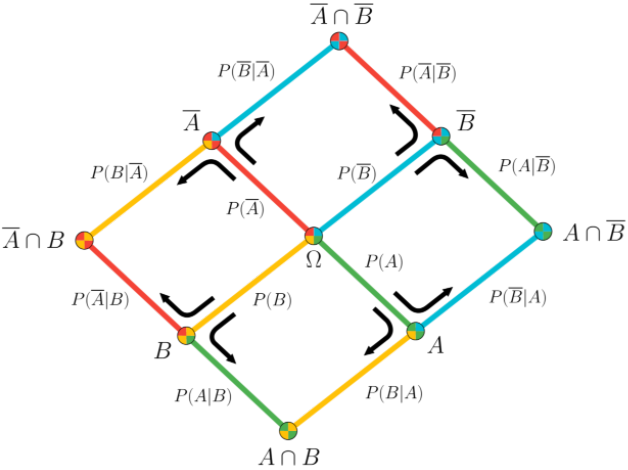
Ship's manifest

- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

† Two prominent
methods for Bayesian model validation



Bayes' Theorem



$$P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$$

Bayesian Inference

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \tag{3}$$

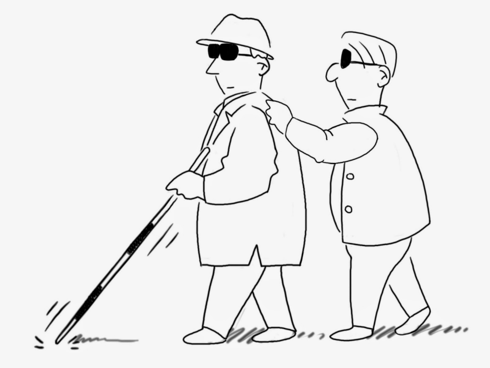
- y is the observed data
- θ is the parameter vector
- $\pi(\theta)$ is the prior
- $\pi(y|\theta)$ is the likelihood
- $\pi(\theta|y)$ is the posterior

Bayesian Inference

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta) \tag{4}$$

- y is the observed data
- θ is the parameter vector
- $\pi(\theta)$ is the prior
- $\pi(y|\theta)$ is the likelihood
- $\pi(\theta|y)$ is the posterior

Posterior navigation



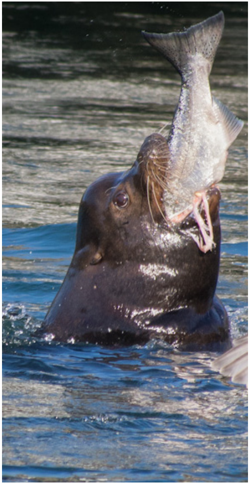
Efficient sampling from $\pi(y|\theta)\pi(\theta)$ is hard

`https://chi-feng.github.io/mcmc-demo/app.html`

Ship's manifest

- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

† Two prominent
methods for Bayesian model validation



A tournament

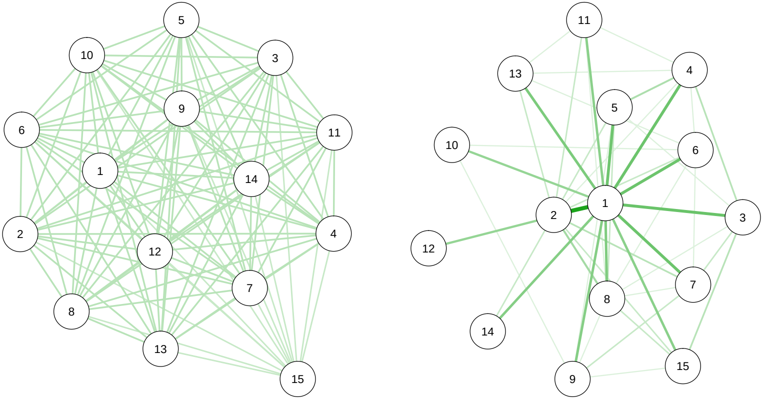
Item Response Model

```
softmax <- function(y) exp(y) / sum(exp(y))
cmp_probs <- function(scale, pa1, pa2, thRaw) {
  th <- cumsum(thRaw)
  diff <- scale * (pa1 - pa2);
  unsummed <- c(0, c(diff - rev(th)), c(diff + th),
    use.names = FALSE)
  softmax(cumsum(unsummed))
}
```

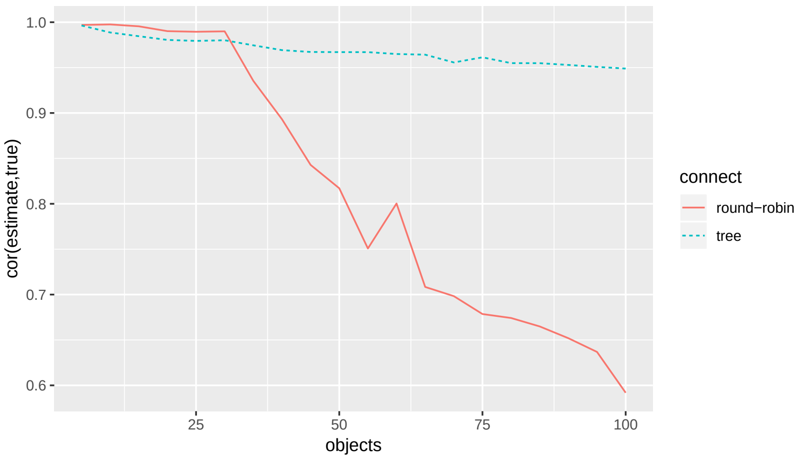
- `scale` is a scalar number
- `pa1`, `pa2` are latent scores for two different objects
- `thRaw` is a vector of thresholds
- Similar to the partial credit model[2]

[2] Masters (1982)

Which is more efficient?



Parameter recovery by connectivity



More than win/lose

Participant picks: $\boxed{\text{running}}$, $\boxed{\text{golf}}$

How predictable is the action?

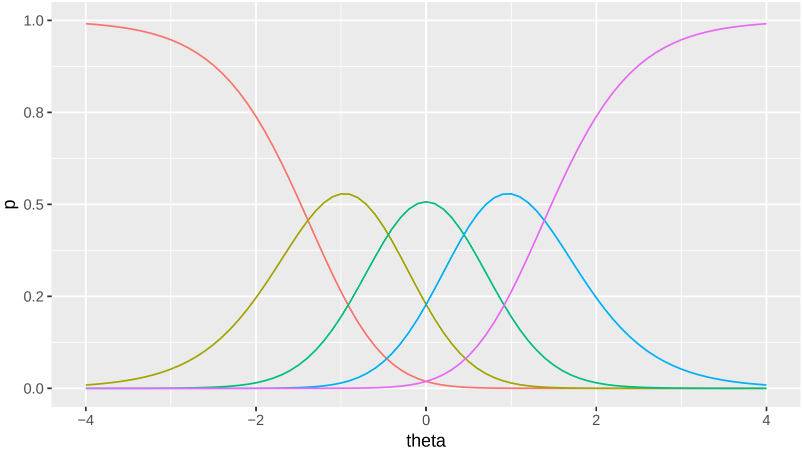
- $\boxed{\text{golf}}$ is much more predictable than $\boxed{\text{running}}$.
- $\boxed{\text{golf}}$ is somewhat more predictable than $\boxed{\text{running}}$.
- Both offer roughly equal predictability.
- $\boxed{\text{running}}$ is somewhat more predictable than $\boxed{\text{golf}}$.
- $\boxed{\text{running}}$ is much more predictable than $\boxed{\text{golf}}$.

Item Parameters

Item model explorer demo

- Can regard `scale` as arbitrary
- For any `scale`, thresholds can be correspondingly scaled
- Object variance and item discrimination are confounded

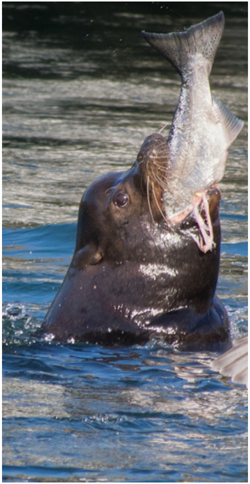
Response curve for simulations



Ship's manifest

- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

† Two prominent
methods for Bayesian model validation



Posterior predictive check

Define

$$\pi(y^{\text{rep}}|y) = \int d\theta \; \pi(y^{\text{rep}}|\theta) \; \pi(\theta|y) \; \pi(\theta). \tag{5}$$

A good likelihood satisfies[3]

$$\pi(y^{\text{rep}}|\theta) = \pi(y^{\text{rep}}|\theta, y) \tag{6}$$

[3] Gelman et al. (2013, p. 146)

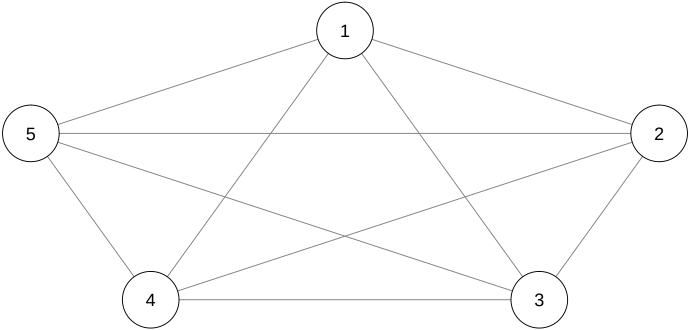
Method

- 2 thresholds
- 5 objects with round-robin connectivity
- 1000 observations (100 per object pair)
- $\chi^2 = \frac{(O-E)^2}{E}$
- Minimum frequency of 5 per expected cell

For each object pair i, j where $i < j$, collect statistics

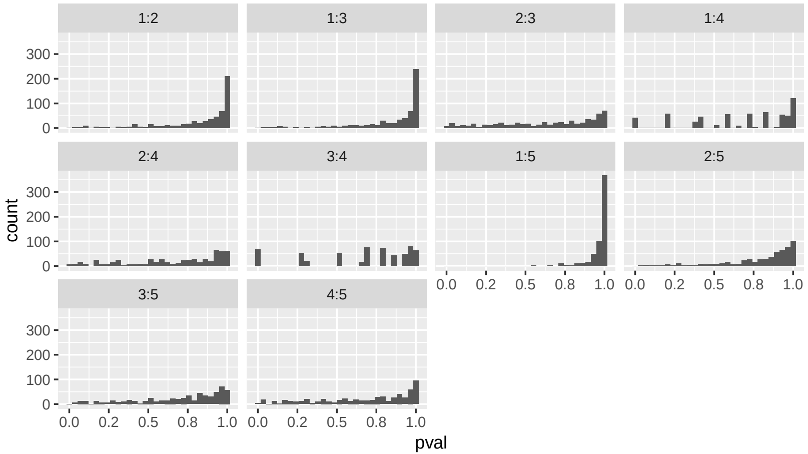
$$\int \mathrm{d}y^{\mathrm{rep}}\ \chi^2_{ij}(y, y^{\mathrm{rep}})\ \pi(y^{\mathrm{rep}}|y) \tag{7}$$

Object connectivity



Note: Can't check datasets with sparse, tree-like connectivity

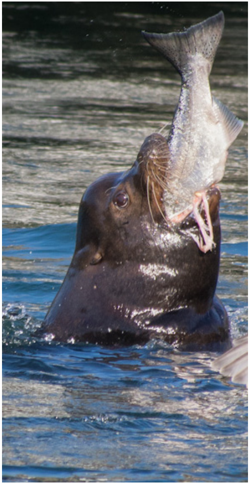
Results



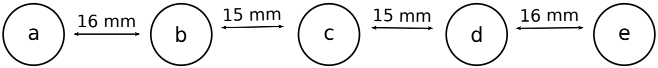
Ship's manifest

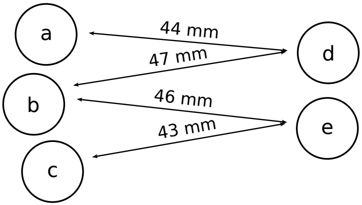
- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

† Two prominent
methods for Bayesian model validation



Distribution of objects





- Nice if objects are standard normally distributed
- How to find the best scaling constant?

Intuition

```
cmp_probs <- function(scale, pa1, pa2, thRaw) {
  th <- cumsum(thRaw)
  diff <- scale * (pa1 - pa2);  # <------------------
  unsummed <- c(0, c(diff - rev(th)), c(diff + th),
    use.names = FALSE)
  softmax(cumsum(unsummed))
}
```

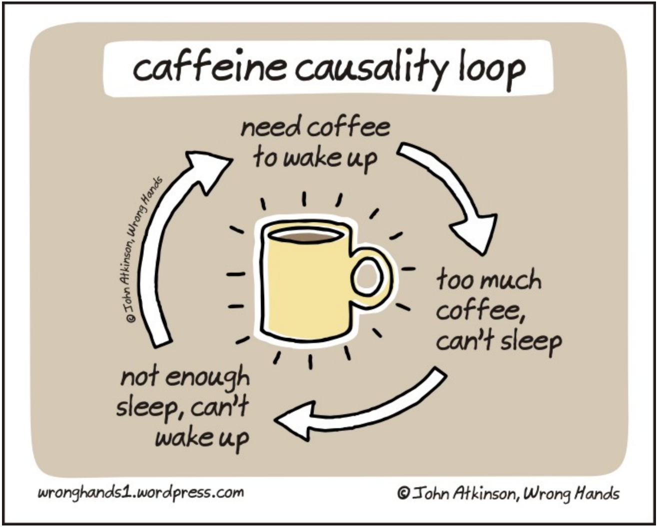
- The scale of `diff` is fixed; thresholds have prior $\mathcal{N}(0, 2)$
- If `scale` increases then `(pa1 - pa2)` decreases to keep the product (`diff`) constant; Object variance is reduced
- If `scale` decreases then `(pa1 - pa2)` increases to keep the product (`diff`) constant; Object variance is increased

Negative feedback loop

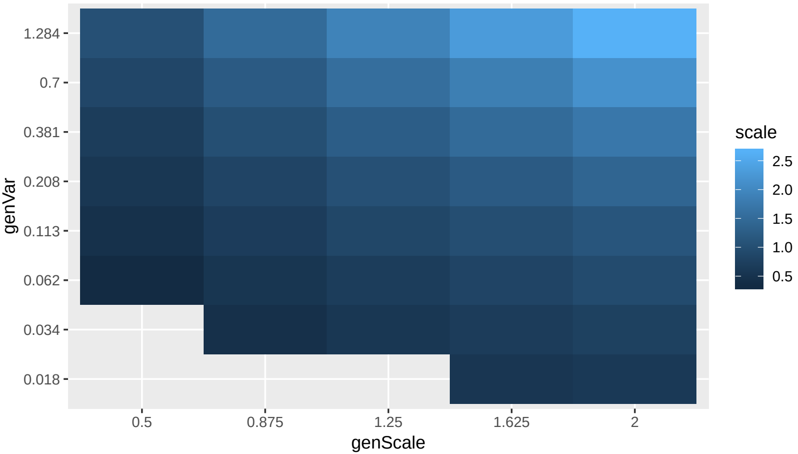
```
transformed parameters {
  real scale = (sigma * sigma) ^ varCorrection;
}
model {
  sigma ~ lognormal(1, 1);
  theta ~ normal(0, sigma);
  ...   // remaining lines omitted
}
```

- `sigma` is an estimate of object standard deviation
- `theta` is a vector of object locations
- `varCorrection` is some constant ≥ 1
- Model fails if object variance is too small

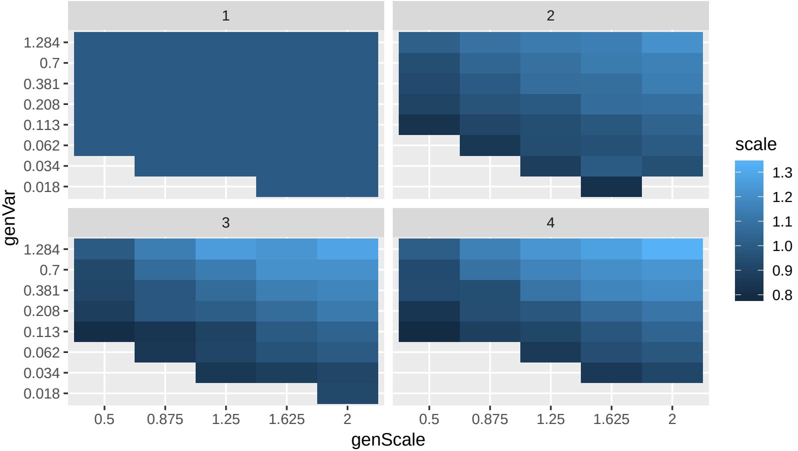
Illustration



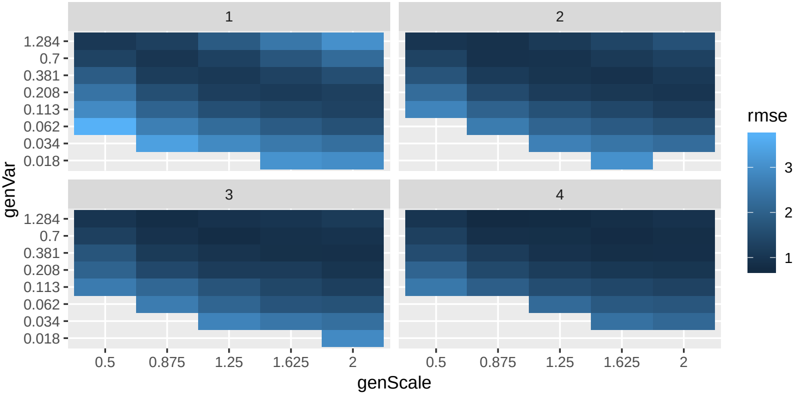
Scale correction



Boosted scale correction



RMSE with true scores

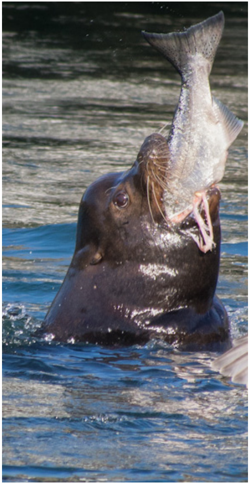


Excessive measurement error signaled by model failure

Ship's manifest

- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

† Two prominent
methods for Bayesian model validation



Simulation-Based Calibration (SBC)

Procedure

1. $\tilde{\theta} \sim \pi(\theta)$
2. $\tilde{y} \sim \pi(y|\tilde{\theta})$
3. $\{\theta_1, \ldots, \theta_L\} \sim \pi(\theta|\tilde{y})$ or $\pi(\tilde{y}|\theta)\ \pi(\theta)$
4. $\sum_{l=1}^{L} \mathcal{I}(\theta_l < \tilde{\theta})$ is uniformly distributed in $[0, L]$

Integrating the posteriors over the joint distribution returns the prior,[4]

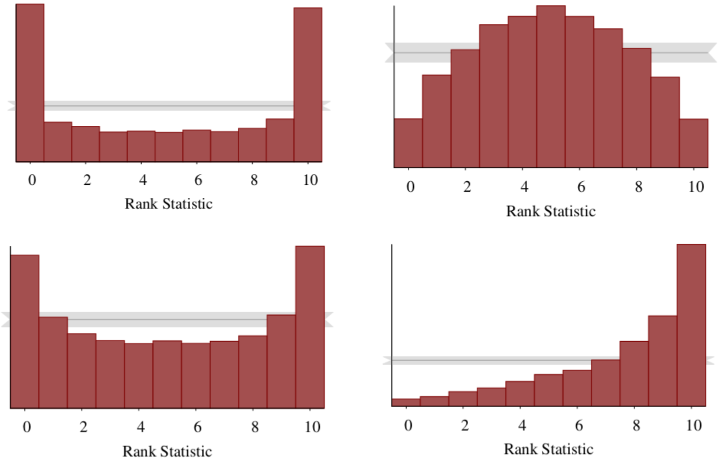
$$\pi(\theta) = \int \mathrm{d}\tilde{y}\mathrm{d}\tilde{\theta}\ \pi(\theta|\tilde{y})\ \pi(\tilde{y}|\tilde{\theta})\ \pi(\tilde{\theta}) \tag{8}$$

[4] Talts, Betancourt, Simpson, Vehtari, and Gelman (2018)

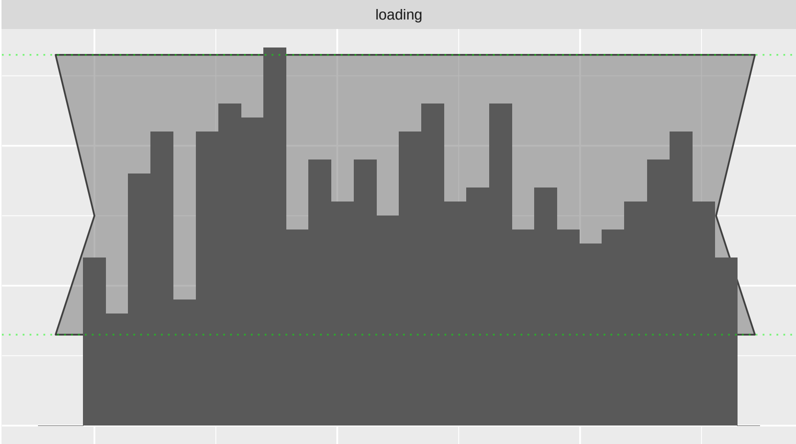
Example #1

```
data {
  int<lower=1> N;  // number of observations
}
transformed data {
  real loading_ = normal_rng(0,3);  // sim prior
  vector[N] obs;
  for (n in 1:N) obs[n] = normal_rng(loading_, 1); // sim data
}
parameters {
  real loading;  // theta
}
model {
  loading ~ normal(0, 3);  // prior
  for (n in 1:N) {
    obs[n] ~ normal(loading, 1.0);  // likelihood
  }
}
```

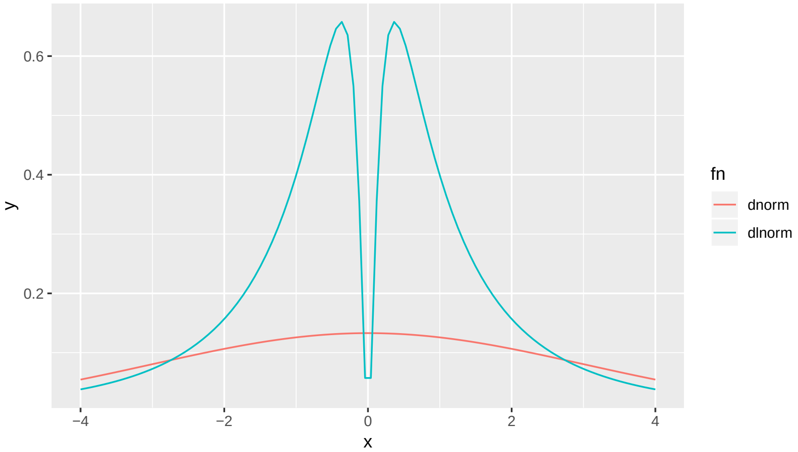
Examples of bad histograms



Results



Support compatibility



Bending the rules

Procedure

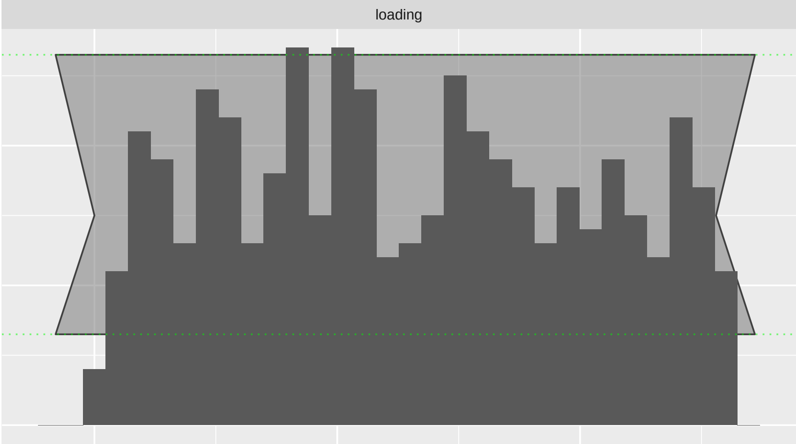
1. $\tilde{\theta} \sim \pi_1(\theta)$
2. $\tilde{y} \sim \pi(y|\tilde{\theta})$
3. $\{\theta_1, \ldots, \theta_L\} \sim \pi(\tilde{y}|\theta)\ \pi_2(\theta)$
4. $\sum_{l=1}^{L} \mathcal{I}(\theta_l < \tilde{\theta})$

- $\pi_1(\theta)$ generates parameters
- $\pi_2(\theta)$ recovers parameters
- $\pi_1(\theta) \neq \pi_2(\theta)$

Example #2

```
data {
  int<lower=1> N;  // number of observations
}
transformed data {
  real loading_ = (lognormal_rng(0,1) *
    (bernoulli_rng(0.5) * 2 - 1));
  vector[N] obs;
  for (n in 1:N) obs[n] = normal_rng(loading_, 1); // sim data
}
parameters {
  real loading;  // theta
}
model {
  loading ~ normal(0, 3);  // prior
  for (n in 1:N) {
    obs[n] ~ normal(loading, 1.0);  // likelihood
  }
}
```

Results



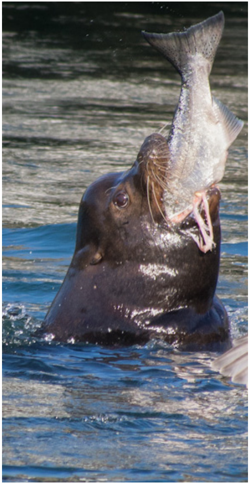
Huh?

Validated a subset of the parameter space

Ship's manifest

- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

† Two prominent
methods for Bayesian model validation

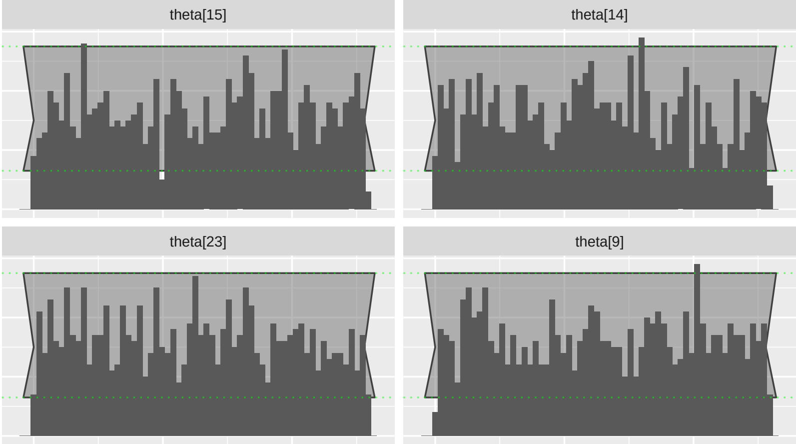


Univariate model

SBC

- $\tau_1, \tau_2 \sim \mathcal{N}(0, 2)$
- 25 objects $\sim \mathcal{N}(0, 1)$
- 375 pairwise comparisons
- random tree connectivity
- 1000 draws from the prior, with 1023 draws per prior

Worst univariate histograms

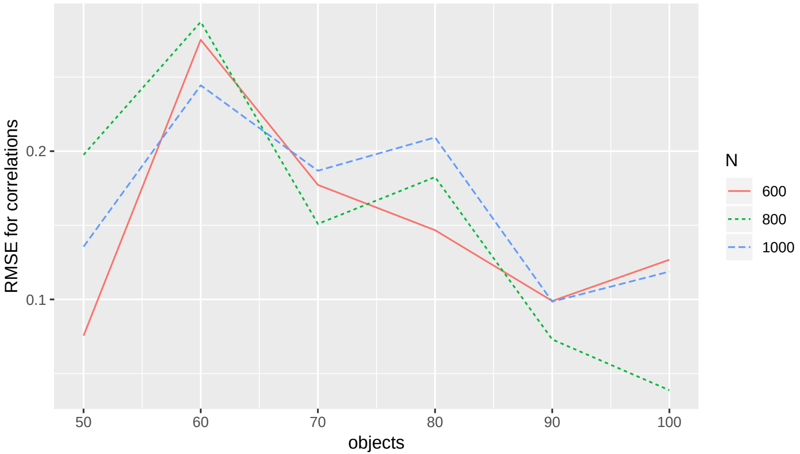


Multivariate workflow

Screen indicators for good behavior,

- Fit univariate models
- Discard items with extreme scaling factors
- Fix model-wise scale at the mean item scale

Covariance, sample size

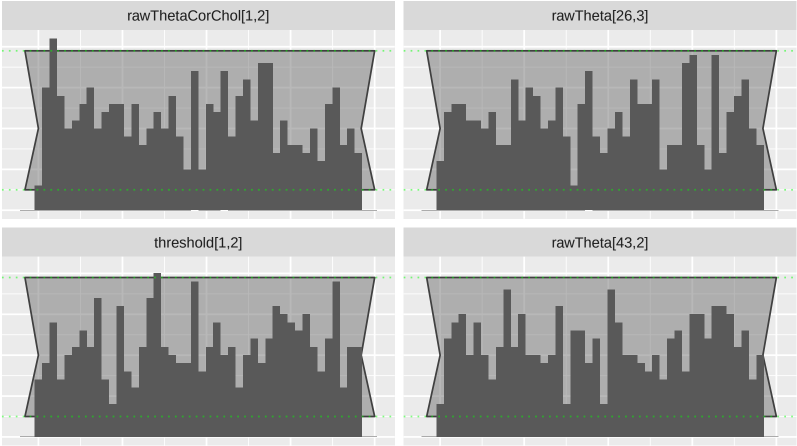


Method

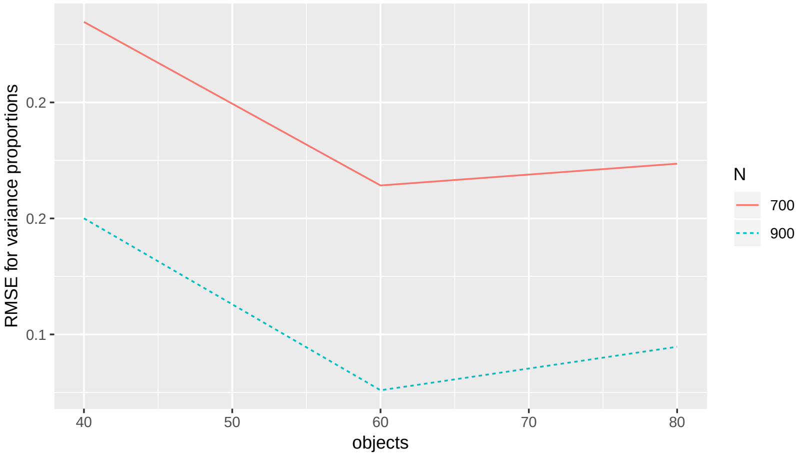
SBC

- 3 items
- 50 objects $\sim \mathcal{N}(0, 1)$
- $\tau_1 \sim \mathcal{N}(-0.8, 0.2), \tau_2 \sim \mathcal{N}(-1.7, 0.2)$
- Variances $\sim \log\mathcal{N}(0.3^2, 0.3)$
- Correlations $\sim$ lkj(2.0)
- 700 pairwise comparisons
- Random tree connectivity
- 500 draws from the prior, with 1535 draws per prior

Worst covariance model histograms



Factor model, sample size

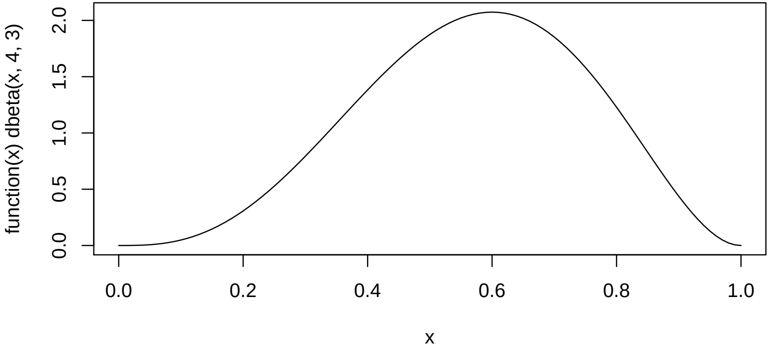


Method

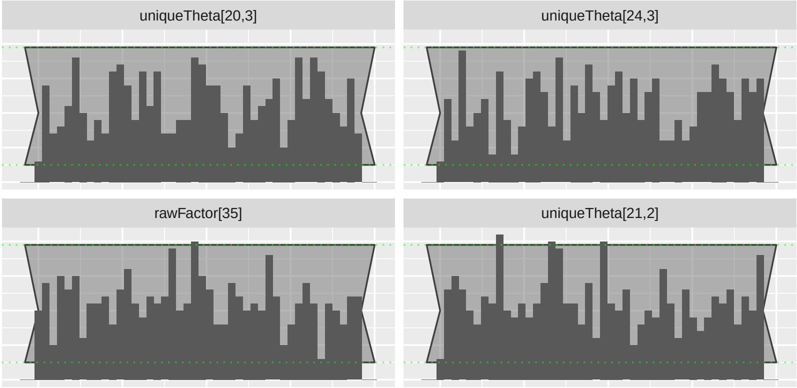
SBC

- 4 items
- 50 objects $\sim \mathcal{N}(0,1)$
- $\tau_1 \sim \mathcal{N}(-0.8, 0.2), \tau_2 \sim \mathcal{N}(-1.7, 0.2)$
- Proportions $\sim$ Beta$(4.0, 3.0)$
- 800 pairwise comparisons
- Random tree connectivity
- 500 draws from the prior, with 1535 draws per prior

Beta(4.0, 3.0)



Worst factor model histograms

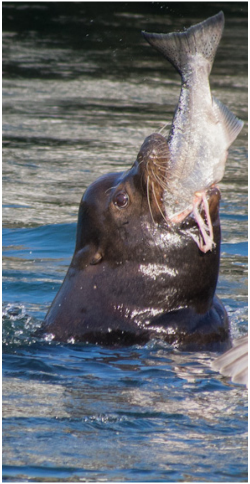


Note: 68 chains had divergent transitions after warm-up. There were a total of 127 divergent transitions across all chains.

Ship's manifest

- Bayes
- Item response model
- Posterior predictive check †
- Scale recovery
- Simulation-based Calibration †
- Validation
- Close

† Two prominent
methods for Bayesian model validation



Caveat emptor

- Figure out your analysis plan before data collection
- Simulate data and write the analyses scripts
- Preregister it

Future work

CRAN package

- Stan models are tricky
- Connections are exogenous, non-stochastic; Need SBC

Submit manuscript

Questions?

Bockenholt, U. (1988). A logistic representation of multivariate paired-comparison models. *Journal of Mathematical Psychology*, *32*(1), 44–63. doi: 10.1016/0022-2496(88)90037-5

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. doi: 10.2307/2334029

Dittrich, R., Francis, B., Hatzinger, R., & Katzenbeisser, W. (2006). Modelling dependency in multivariate paired comparisons: A log-linear approach. *Mathematical Social Sciences*, *52*(2), 197–209. doi: 10.1016/j.mathsocsci.2006.06.001

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed. ed.). CRC press.

Luce, R. D. (1959). *Individual choice behaviour: A theoretical analysis*. New York: Wiley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi: 10.1007/BF02296272

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... others (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. doi: 10.1126/science.aar6404

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating bayesian inference algorithms with simulation-based calibration.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273. doi: 10.1037/h0070288