YES, BUT DOES IT WORK?

Daniel Simpson
Department of Statistical Sciences
University of Toronto

SHE WORKS HARD FOR THE MONEY

1. This will be an index

WHY BAYES?
WHITHER BAYES?
WHEREFORE BAYES?

BAYESIAN JUSTIFICATION (FT. NATE DOGG)

BAYESIAN JUSTIFICATION (NOT FT. NATE DOG)

- **?:** If regularisation isn't the key word, what is the advantage of Bayesian thinking here?
- **!:** Building a Bayesian model **forces** you to build a model for how the data is generated
- We often think of Bayesian modelling as specifying a **prior** and a **likelihood** as if these are two separate things.
- **They. Are. Not.**

A BAYESIAN MODELLER COMMITS TO AN A PRIORI JOINT DISTRIBUTION

Latent Gaussian
(Finn's stuff + covariates +
design effects +++
all shoved into one vector)

$$p(y, \eta, \theta) = \cancel{p(y \mid \eta)p(\eta \mid \theta)p(\theta)}$$

Data

Parameters

HIDING ALL AWAY

- This decomposes the joint distribution into three parts:
  - The marginal likelihood (ie the density of the data under the prior model)

$$p(y)$$

  - The marginal posterior for the parameters

$$p(\theta \mid y)$$

  - The full conditional for the latent field

$$p(\eta \mid \theta, y)$$

- The last of these is almost Gaussian

LEWIS TAKES OFF HIS SHIRT

- The most important distribution is the marginal likelihood $p(y)$, which tells us how well the model can capture the data
- Simulations from the marginal likelihood are the *prior predictions*
- If none of these look like plausible data, there's trouble
- But wait: We don't know it!

THE MAJESTY OF GENERATIVE MODELS

- If we disallow improper priors, then Bayesian modelling is generative.
- In particular, we have a simple way to simulate from $\tilde{y} \sim p(y)$
  - Simulate $\tilde{\theta} \sim p(\theta)$
  - Simulate $\tilde{\eta} \sim N(0, Q(\tilde{\theta})^{-1})$
  - Simulate $\tilde{y} \sim p(y \mid \tilde{\eta}, \tilde{\theta})$

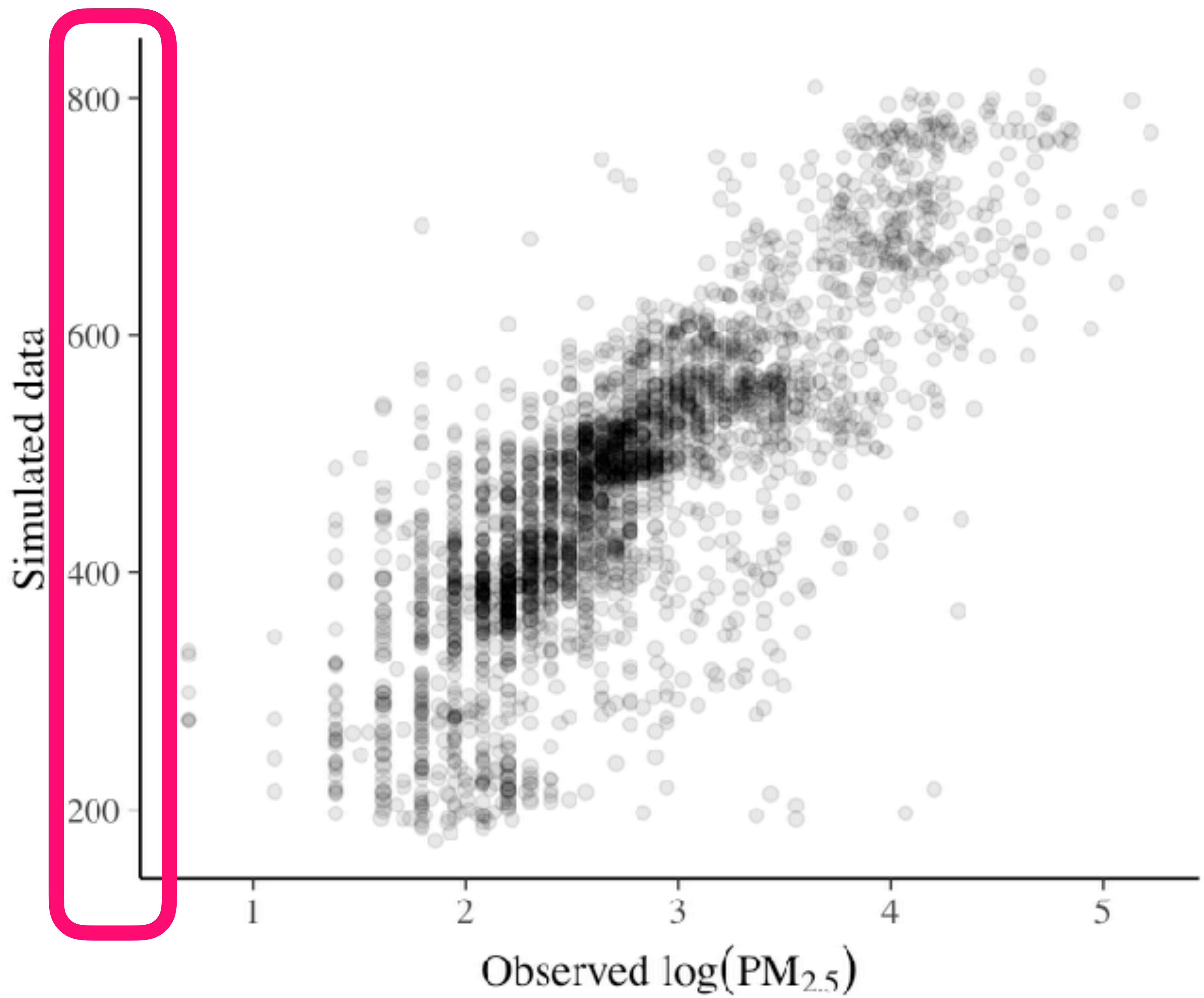
WHY DO WE CARE?

- Consider a cartoon model for estimating global PM2.5 concentration based on (good) Ground Monitor measurements and (noisy) satellite estimates

$$\log(\mathrm{PM}_{2.5})_i = \beta_0 + \beta_{0\mathrm{region}(i)} + (\beta_1 + \beta_{1\mathrm{region}(i)}) \log(\mathrm{SAT})_i + \epsilon_i$$

- Consider the following priors (we'll fix the observation noise for now:
  - $\beta_j \sim N(0, 10)$
  - $\beta_{jr} \sim N(0, \sigma_j^2)$
  - $\sigma_j^{-2} \sim \mathrm{Exp}\left(10^{-2}\right)$

WHAT DOES THIS LOOK LIKE?



WHAT DO WE NEED IN OUR PRIORS?

- This suggests we need *containment*: Priors that keep us inside sensible parts of the parameter space
- The prior for the **range:**
  - Needs to not have too much mass on smaller ranges than the data observations
  - A *inverse-Gamma* tuned so that $\Pr(\text{range} < L) = \alpha$ is good
- The prior for the **standard deviation:**
  - **Not the variance or the precision!**
  - Again, an exponential or half-t so that $\Pr(\sigma > U) = \alpha$

LESSON FOR BAYESIAN UNCERTAINTY QUANTIFICATION

- You need to check how your priors interact with each other and the likelihood in order to assess if they're sensible.
- Hence, an important step in any sort of data assimilation / backwards uncertainty quantification is ***forwards*** **uncertainty quantification**
- It alerts us if we've accidentally put too much weight on unphysical model configurations

CAN WE EVEN DO BAYES?

WHAT DO WE DO ABOUT PARAMETERS?

- We need to construct a principled way to deal with the parameters θ
- In theory this is straightforward. If

$$u \mid \theta \sim \mathrm{N}\left[0, Q(\theta)^{-1}\right]$$

- Then the to the log-posterior is

$$\log \pi(y \mid u) + \frac{1}{2}\log|Q(\theta)| - \frac{1}{2}u^T Q(\theta) u + \log \pi(\theta)$$

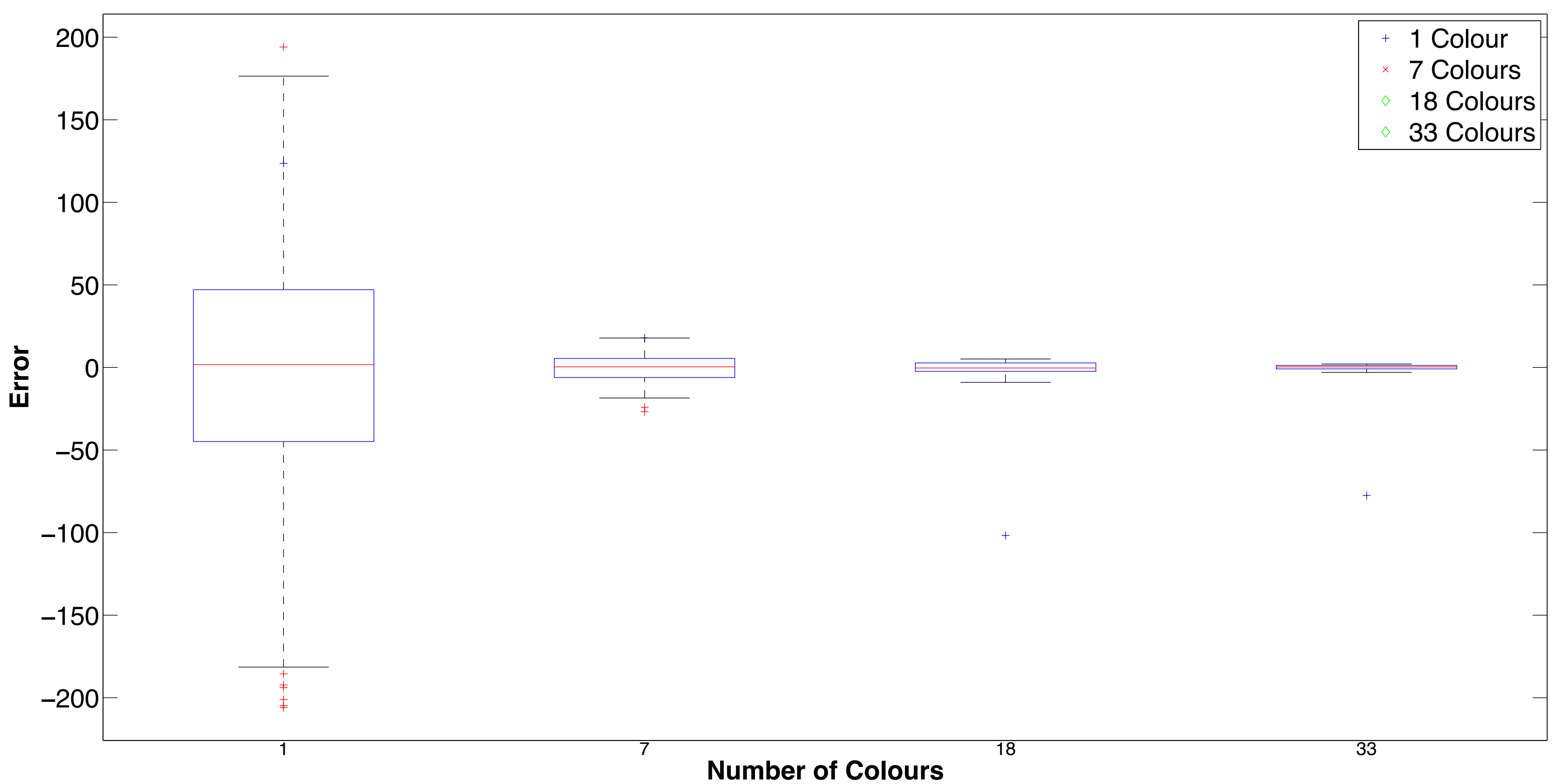
(Red is the colour of pain)

HOW DO YOU COMPUTE A DETERMINANT?

- With a Cholesky factorization.
  - If $Q = LL^T$ then $\log|Q| = \sum_i \log(L_{ii})$
  - This only works if you can actually compute the the Cholesky
    - For a dense matrix, this costs $\mathcal{O}(n^3)$
    - For a sparse matrix this costs $\mathcal{O}(n^{3/2})$—$\mathcal{O}(n^2)$
    - If you can write your model in state space form it's $\mathcal{O}(n)$
  - This really hurts!

ONE POSSIBLE WAY THROUGH

- Note that $\log|Q| = \mathbb{E}\left(z^T \log(Q) z\right)$
- *z* is a vector of iid zero mean, unit variance random variables
- This requires the computation of a matrix logarithm
- There are some **clever tricks!**
- **In the name of all that is holy, do not re-sample *z*!**



REAL TALK

- Honestly, I’ve never got this stable.
- Michael Jordan (and others) may be extolling the virtues of Stochastic optimization, but that only works when you can control the noise
- We found that really hard to do
- So, the point where you can no longer compute a Cholesky (or something similar) is the point where you can’t compute the likelihood
- (Let us not speak of pseudomarginal methods. They do not work for this problem)

THE THREE STAGES OF MODELLING

- Formulation
  - *Hi Finn!*
- Approximation
  - *SPDEs*
  - *Other dimension-reduction techniques*
- **Desperation**

EMPIRICAL BAYES: THE LAST HOPE OF THE HOPELESS

- Replace the good thing with the cheap thing:

$$p(u \mid y) = \int p(u, \theta \mid y)\, d\theta \overset{?}{\approx} p(u \mid y, \theta^*)$$

- This is a one-point integration rule, so it's pretty important to choose the one point correctly!
- You want

$$\theta^* = \arg\max_{\theta} \pi(\theta \mid y) \neq \arg\max_{\theta} \pi(u, \theta \mid y)$$

- (or some appropriate approximation to it)

BUT SHIRLEY THIS IS JUST AS BAD

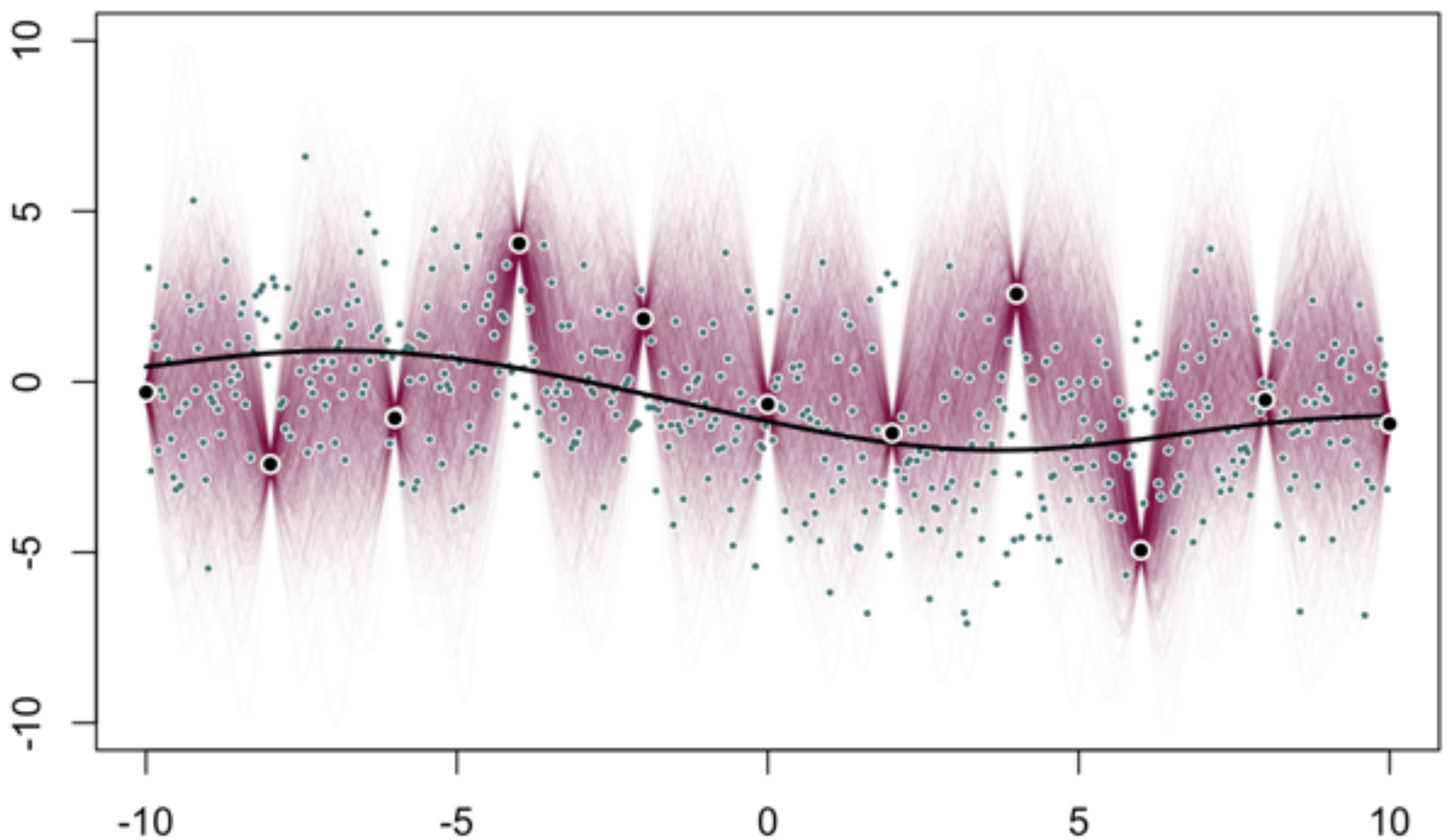
- Instead of computing a log-determinant, this requires its derivative

$$\operatorname{tr}\left(Q(\theta)^{-1}\frac{\partial Q}{\partial \theta_j}\right)$$

- This is **much** easier to compute!
- And amenable to the tricks Finn mentioned!
- You can use all your fancy linear solvers here!

WHAT HAVE WE LOST?

- The uncertainty intervals for u will be wrong
- When there isn't very much information about θ in the data, you will sometimes over-fit.
- This is kinda common.



ALL THIS WORK, BUT DID I ACTUALLY COMPUTE THE RIGHT THING?

WE HAVE COMPUTED SOME THINGS

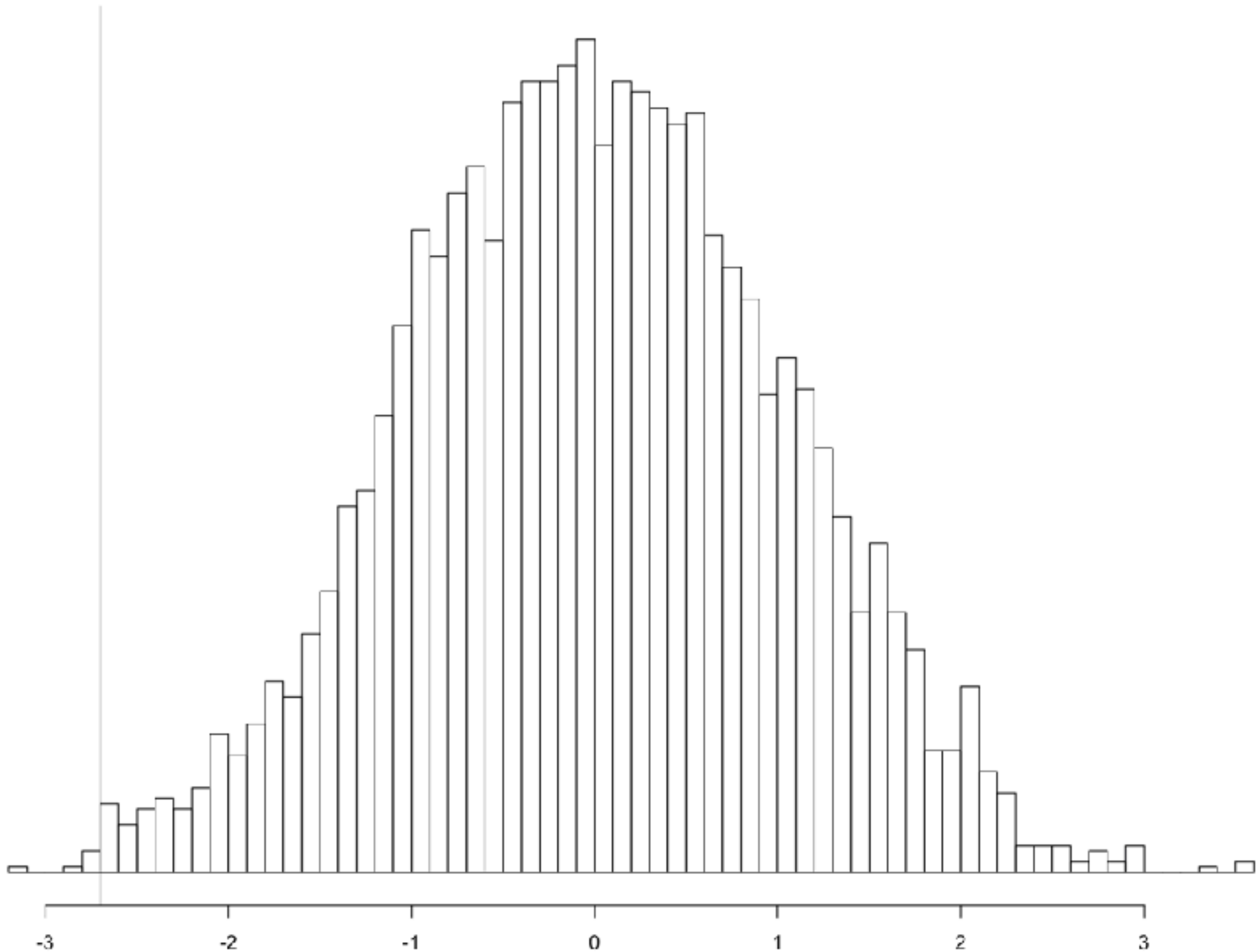
- Depending on what is possible, we've computed one of these approximate posteriors:
  - $p(\eta, \theta \mid y)$
  - $p(\theta \mid y)$
  - $p(\eta \mid \theta, y)$
- One thing to ask is "did we do a good job?"

HOW CAN WE TELL IF AN ALGORITHM ACTUALLY WORKS?

- Idea: Run the algorithm on simulated data.

1. Pick a parameter value θ_0
2. Generate data from $p(\mathbf{y} \mid \boldsymbol{\theta}_0)$
3. Fit model to data
4. Compare the posterior to the known true value

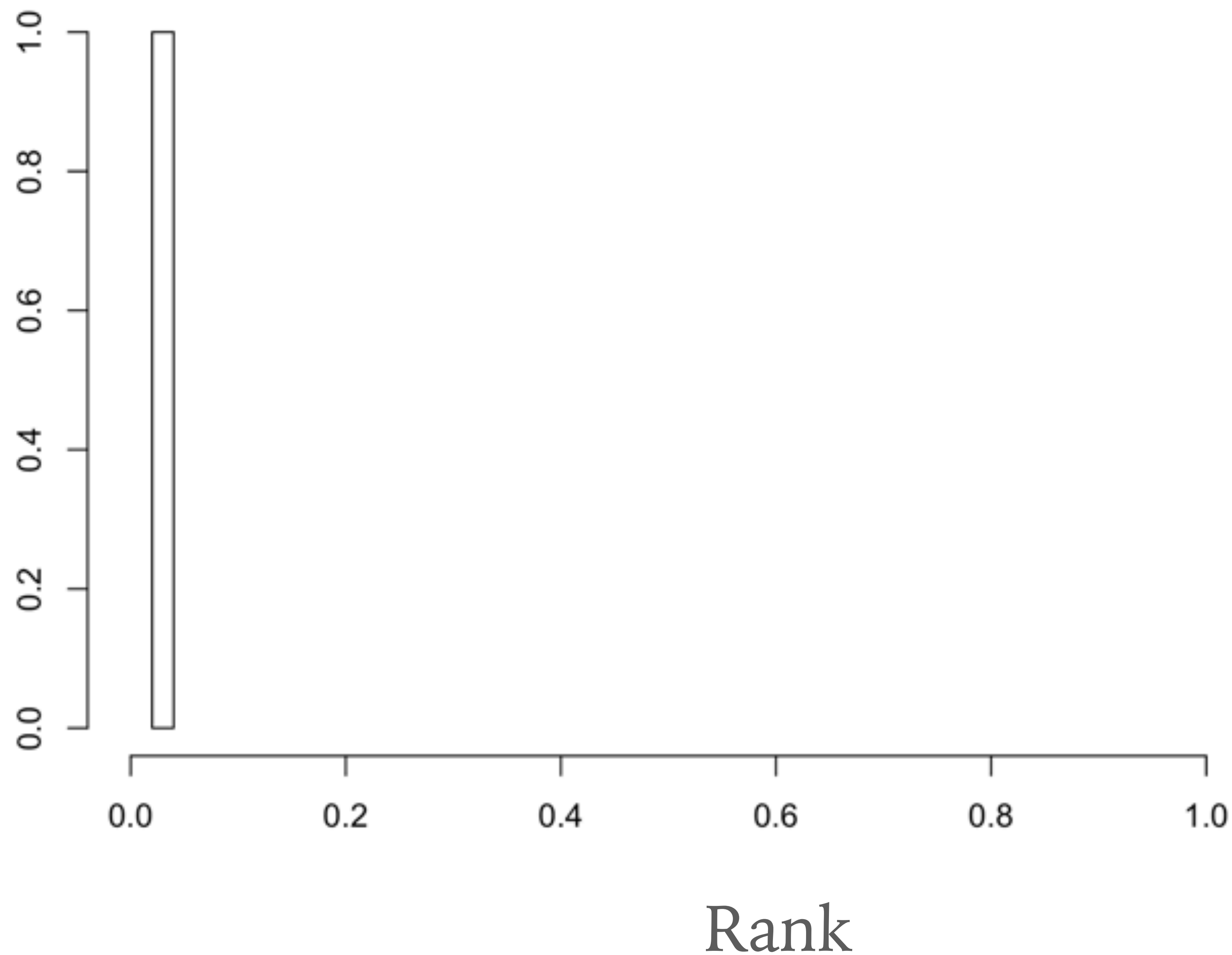
OKAY! IS THIS RIGHT?



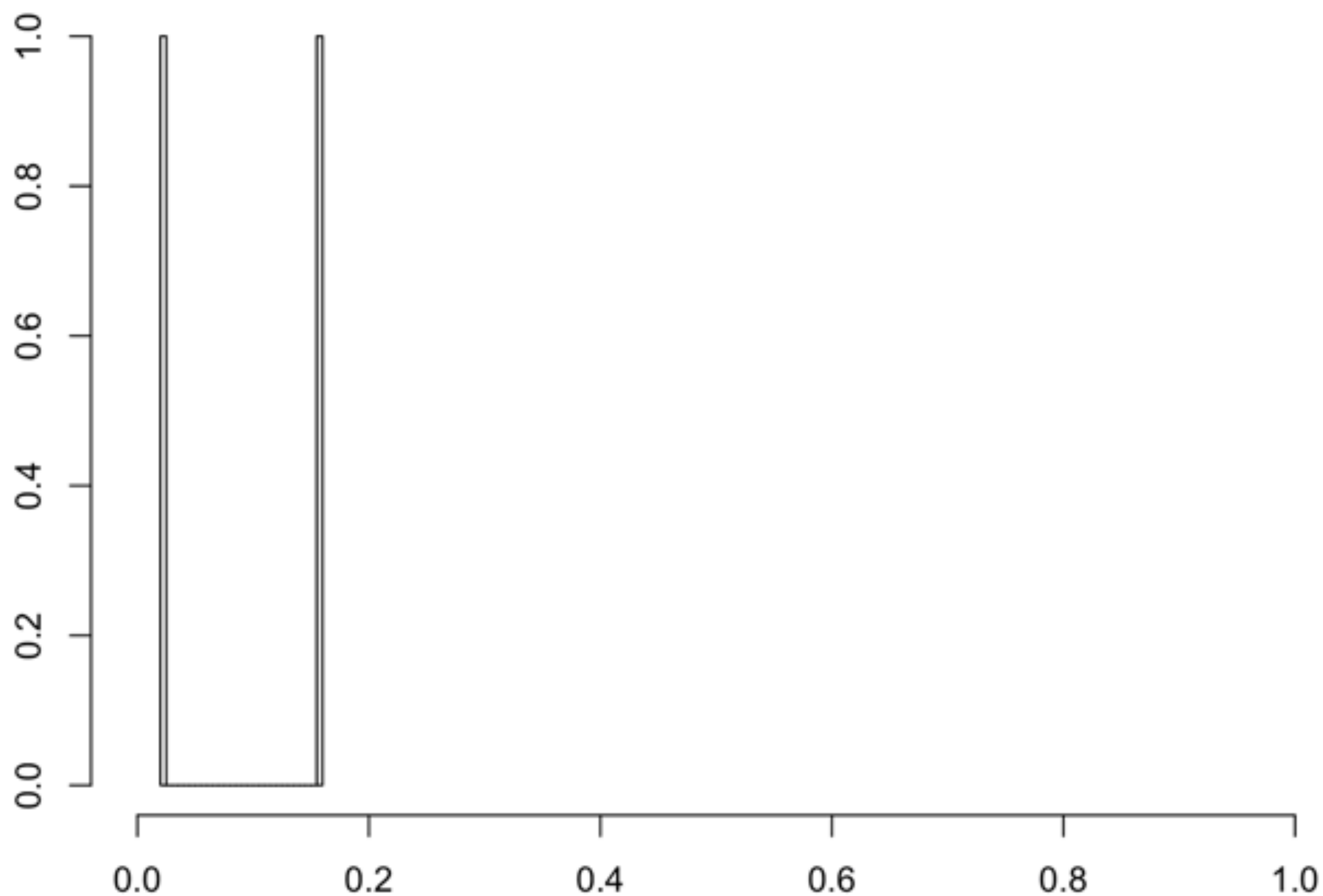
ONCE MORE WITH FEELING

- Maybe we should check more than one point!
- How do we do that?
- We want to check all reasonable values of θ
- Idea: Simulate multiple $\theta \sim p(\theta)$ and check the fit
- How do we check the fit?
- **Big idea:** Look at where the true parameter lies in a bag of L posterior samples

SINGLE RECOVERY

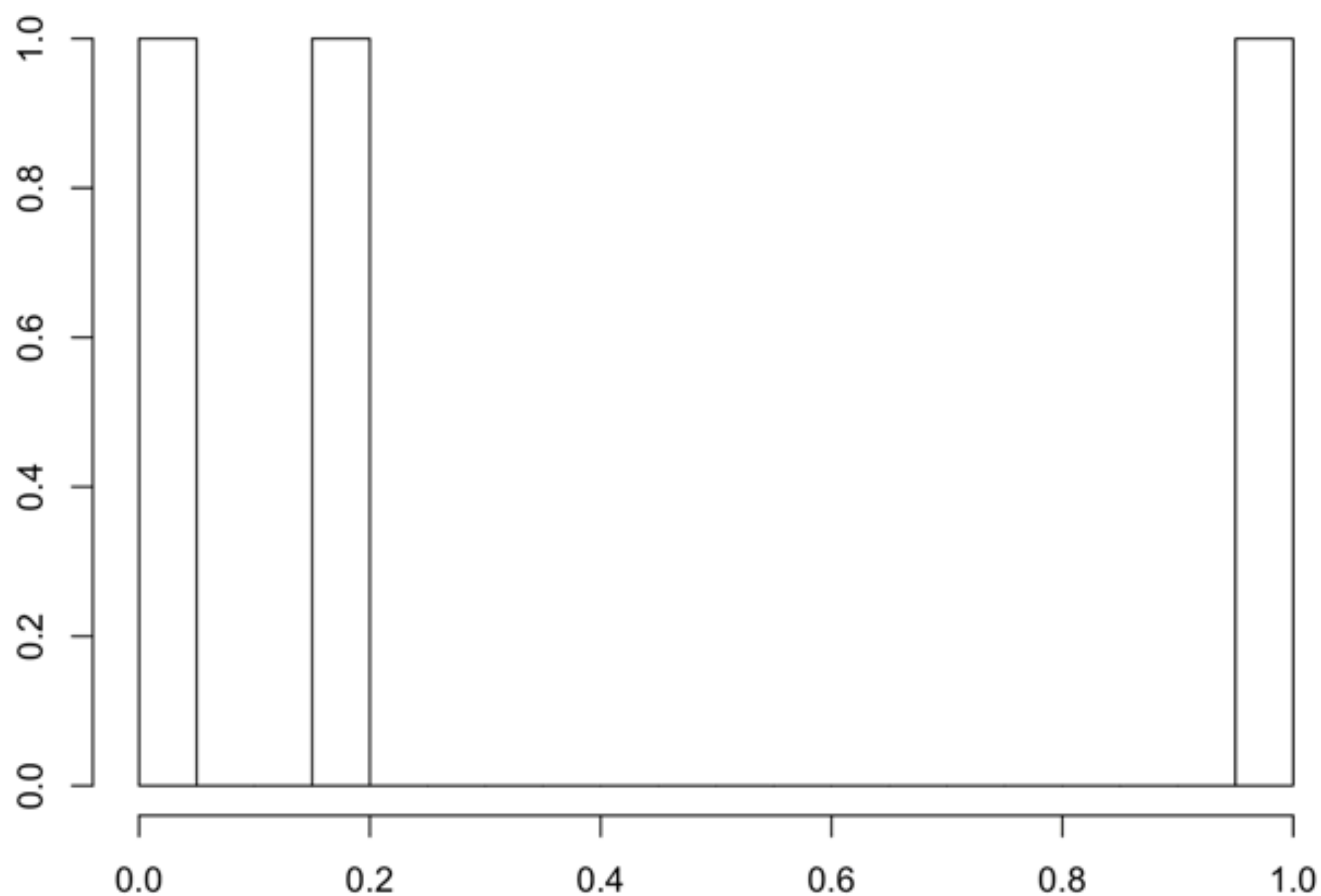


MULTIPLE RECOVERY



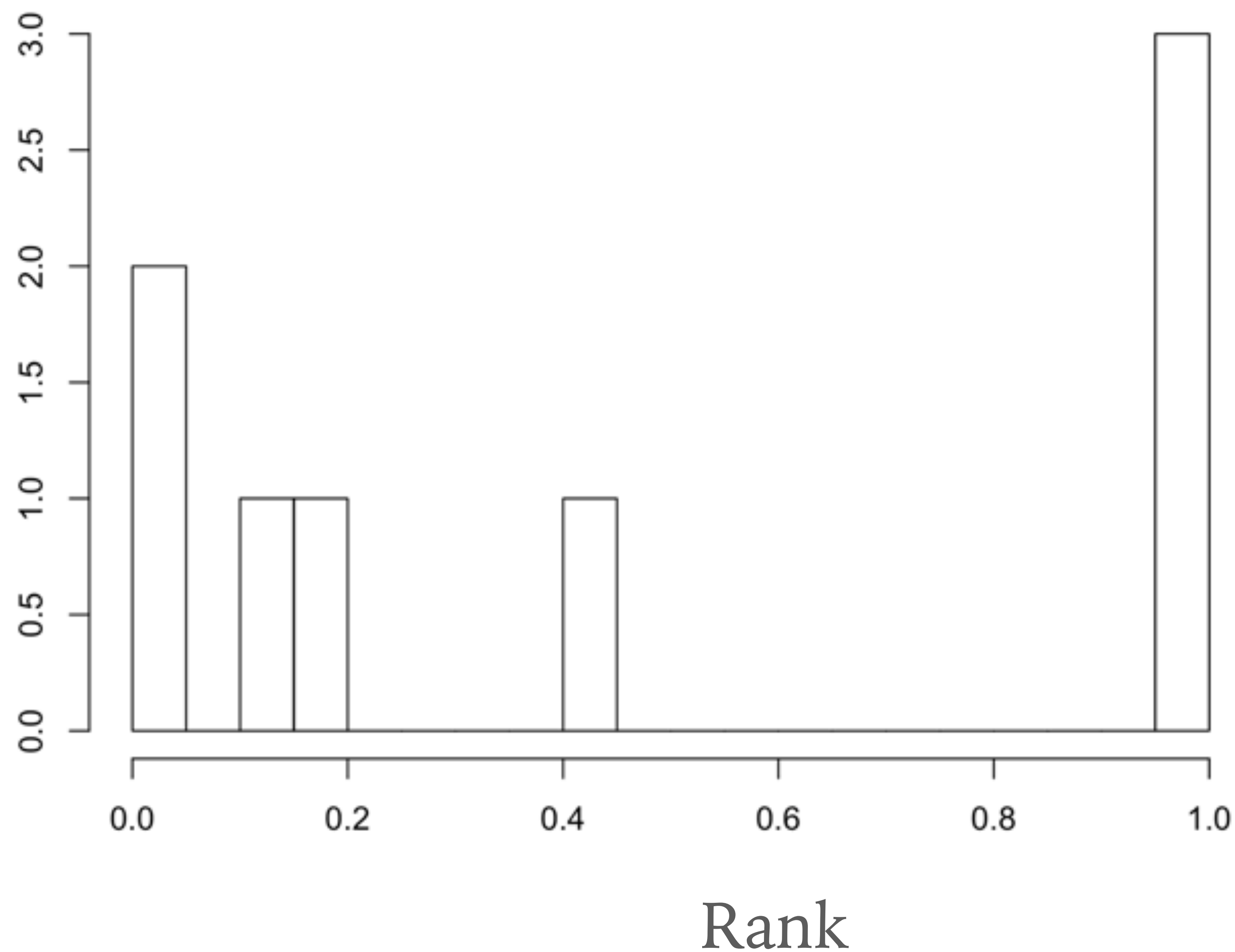
Rank

MULTIPLE RECOVERY

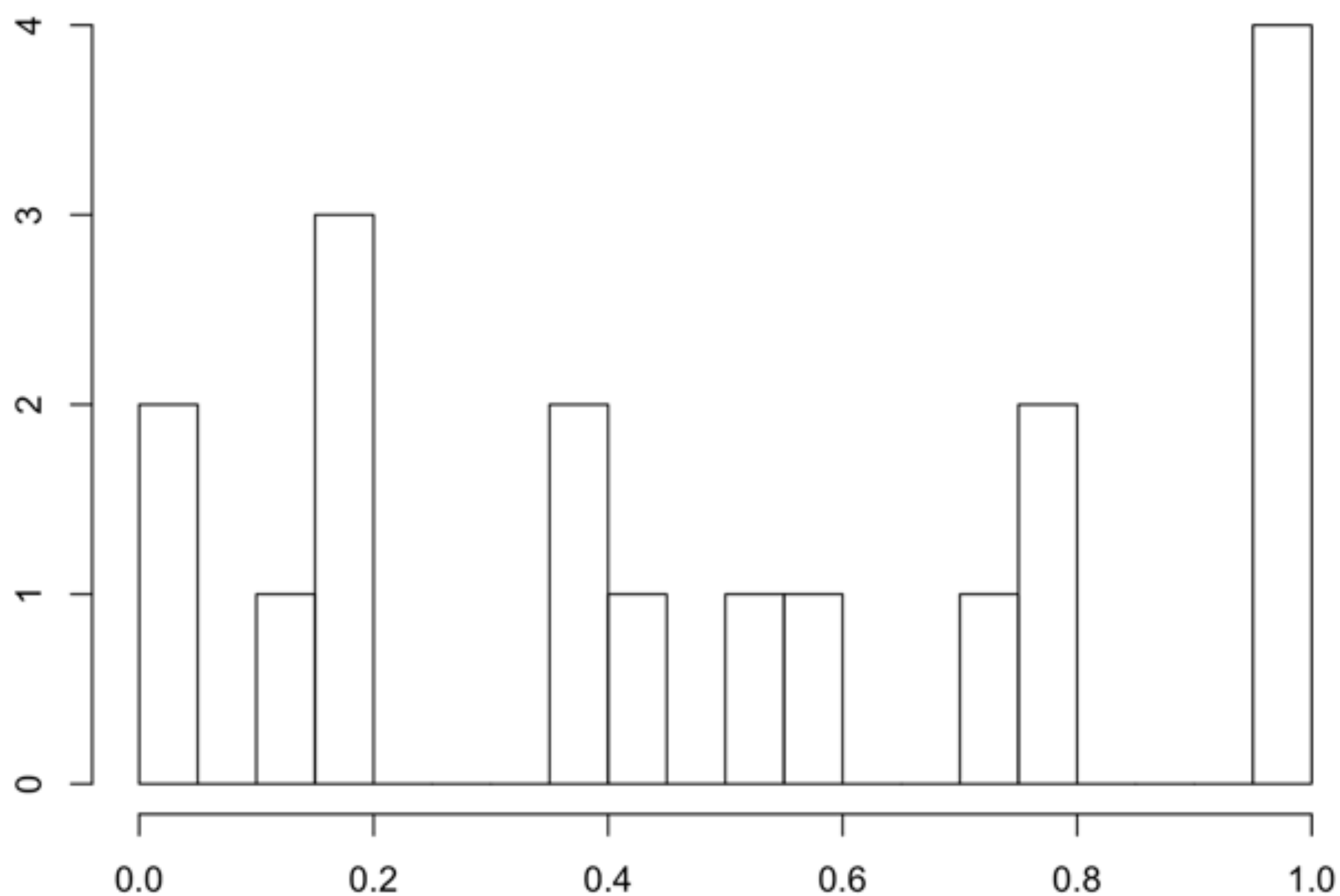


Rank

MULTIPLE RECOVERY

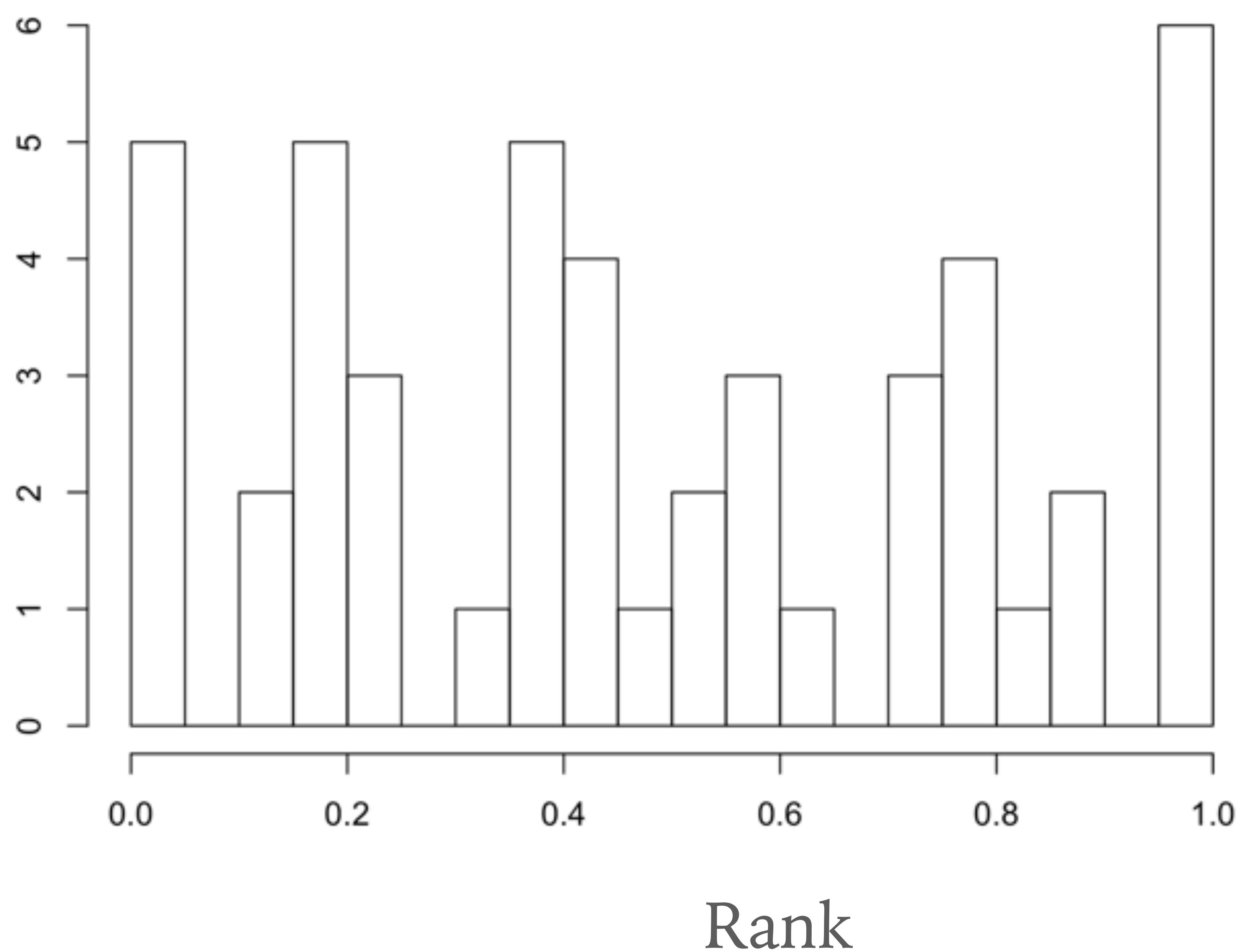


MULTIPLE RECOVERY

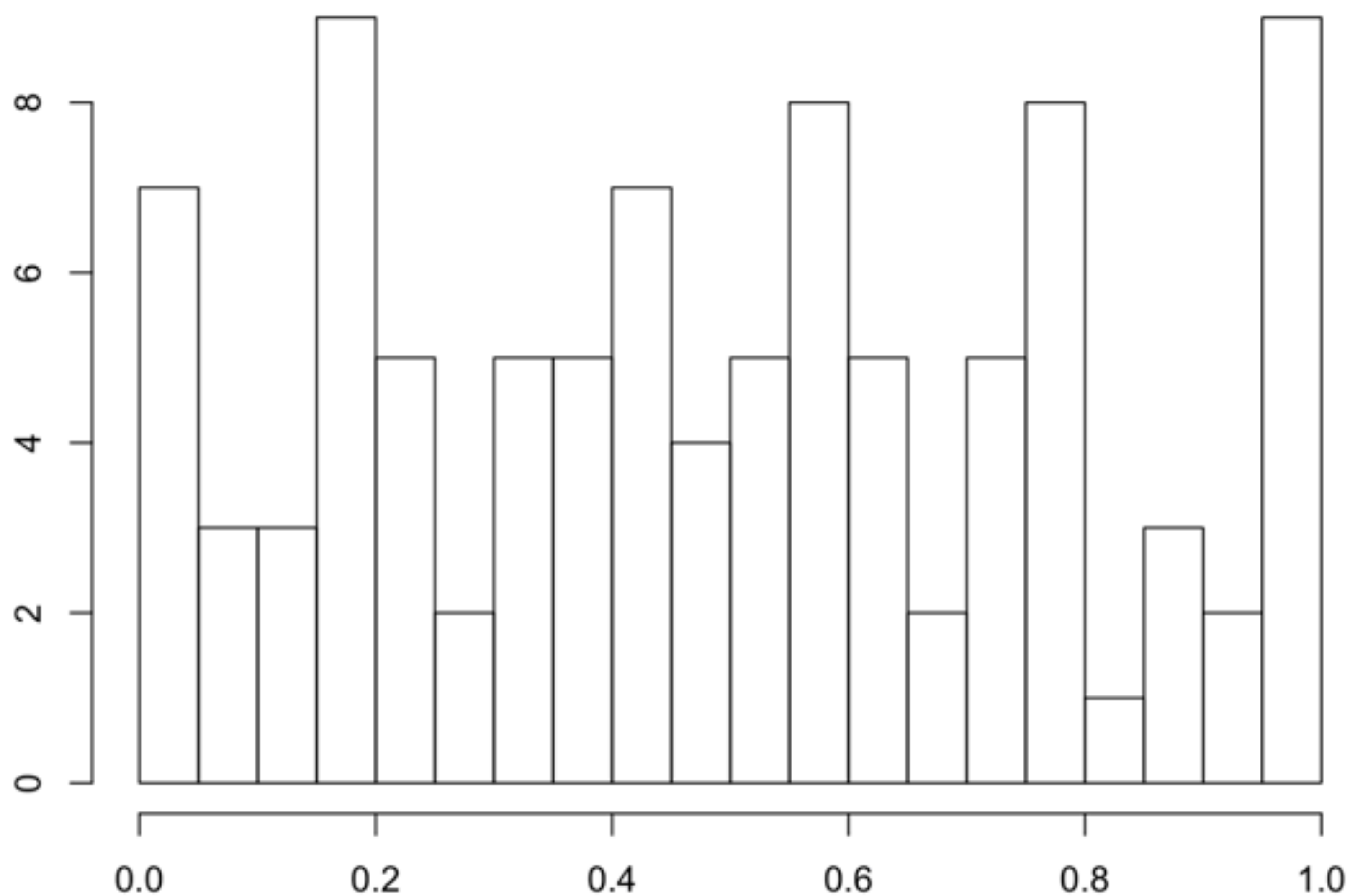


Rank

MULTIPLE RECOVERY

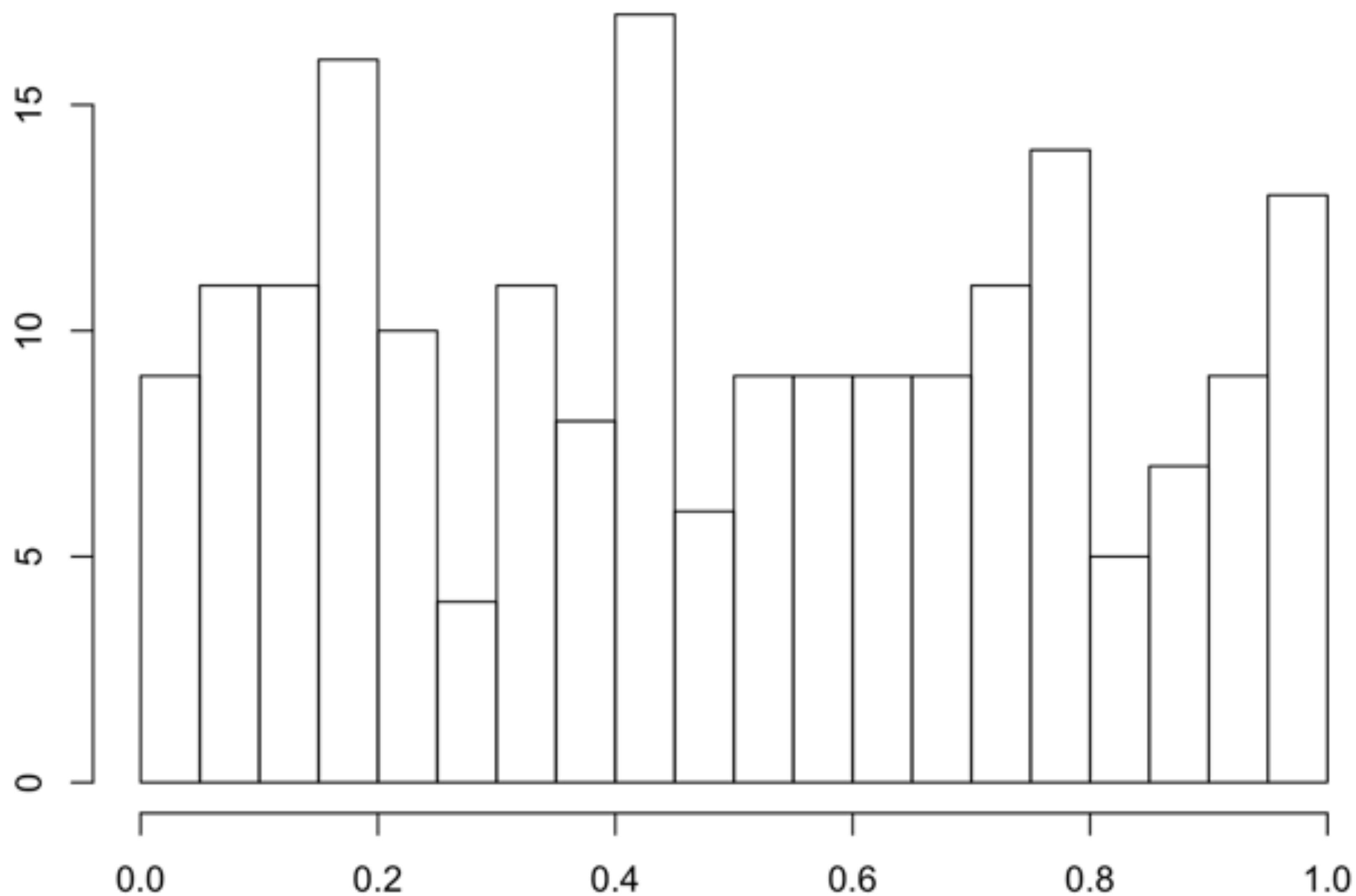


MULTIPLE RECOVERY



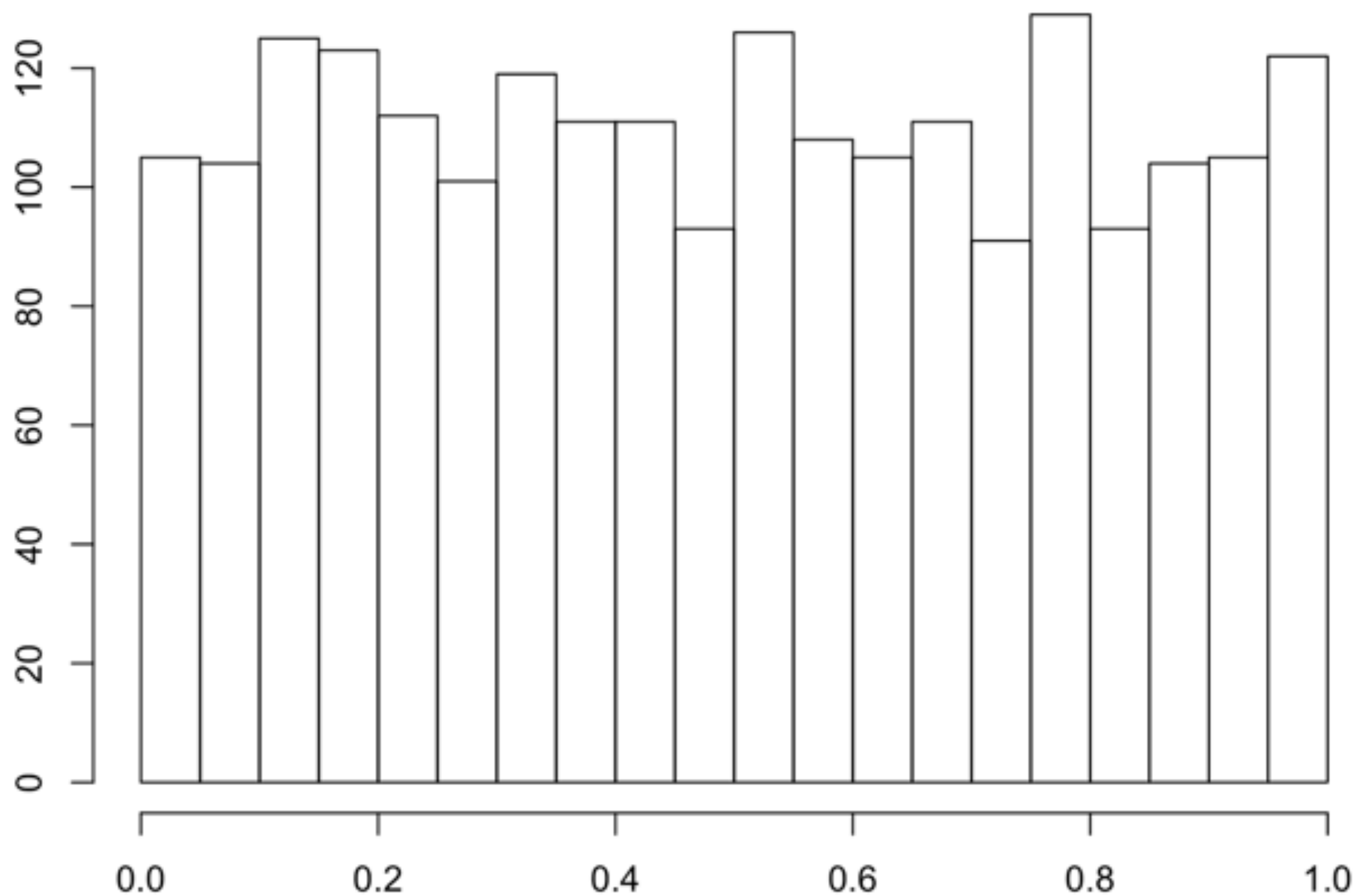
Rank

MULTIPLE RECOVERY



Rank

MULTIPLE RECOVERY

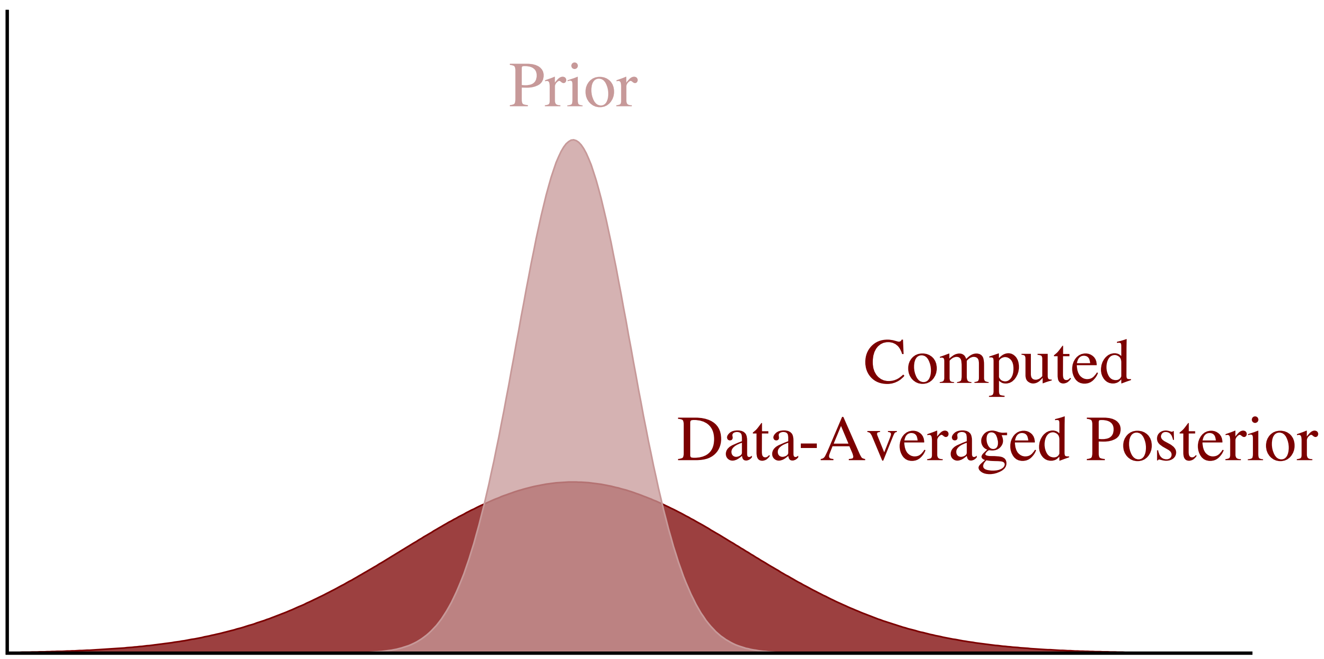


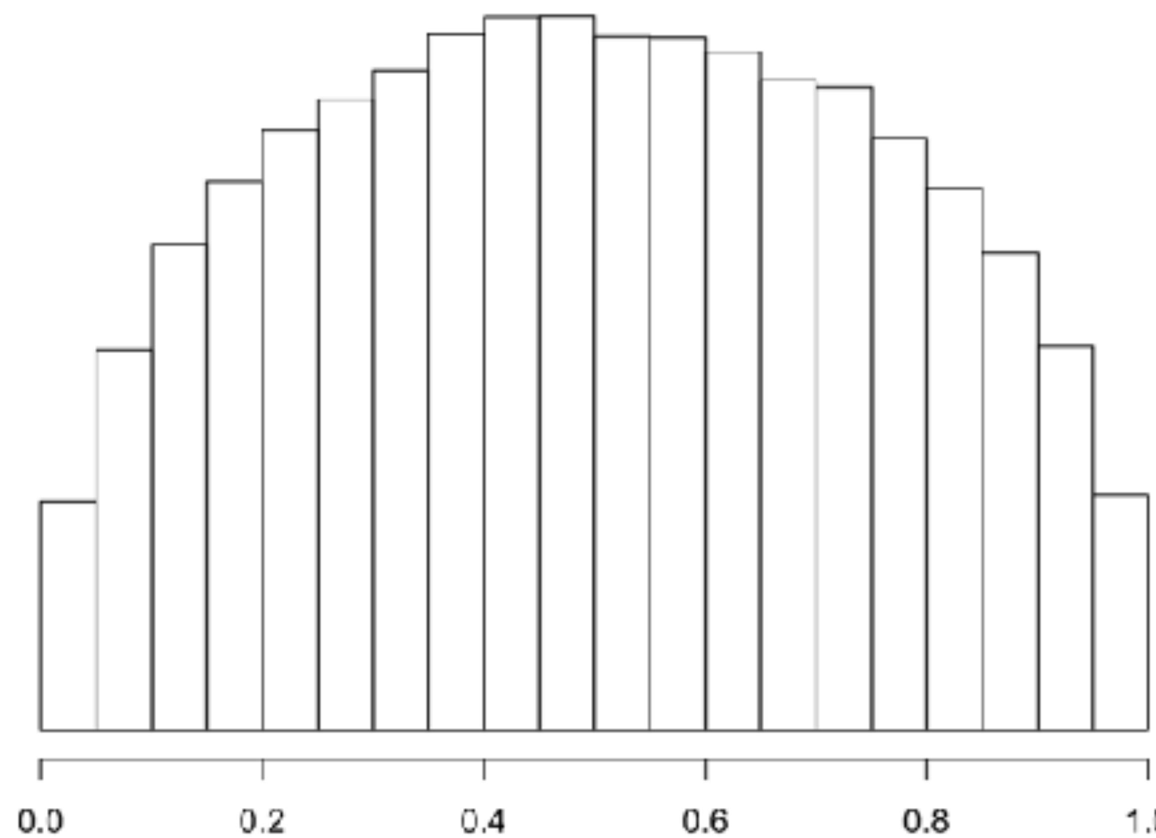
Rank

THAT LOOKS MIGHTY UNIFORM…

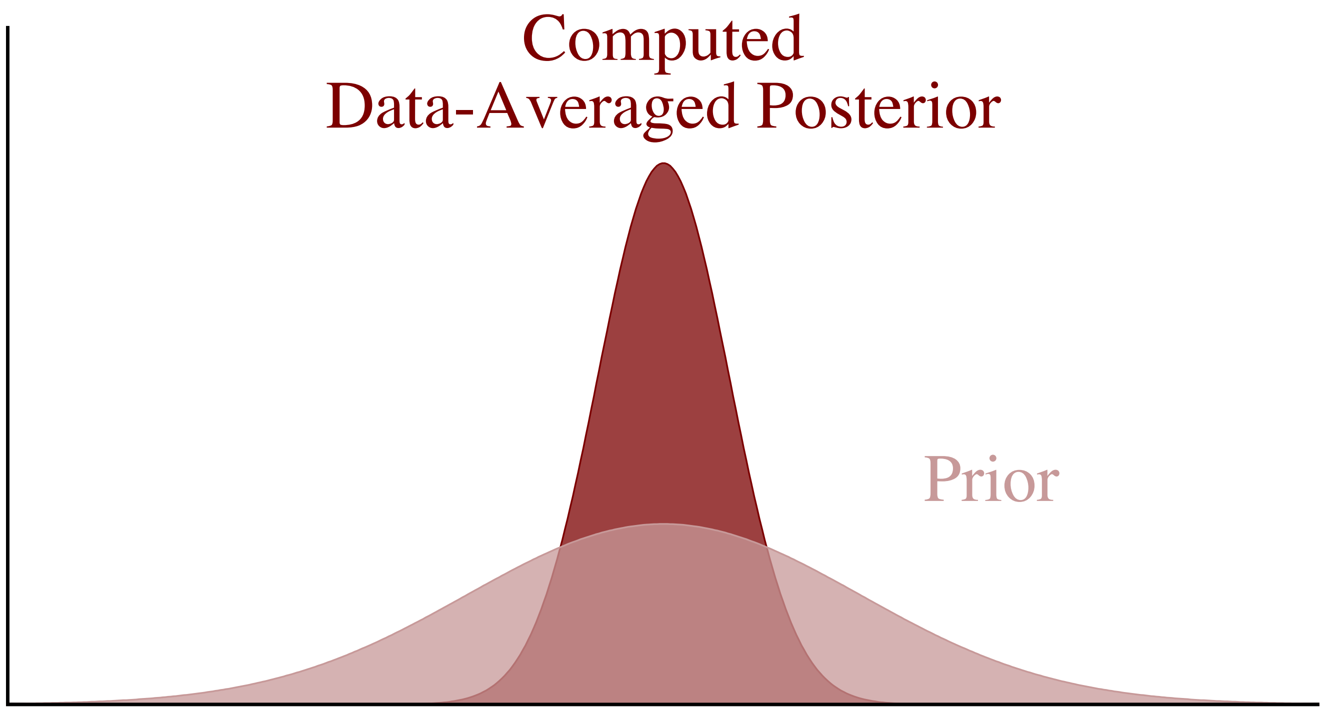
- Why is this uniform?
  - Maths.
  - It turns out that ranks are uniformly distributed *because* when you average the posterior over data generated from $p(y)$, you get the prior back!
- Better yet, deviations from uniformity are meaningful!

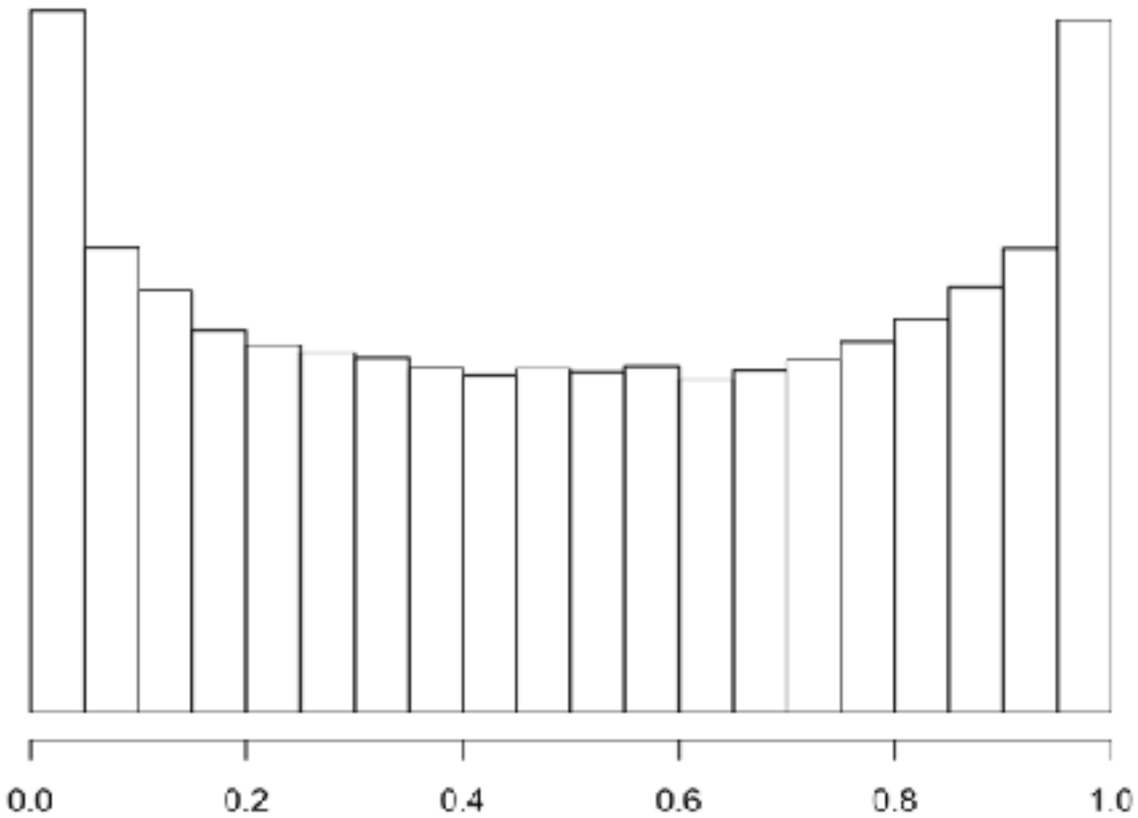
POSTERIOR TOO WIDE



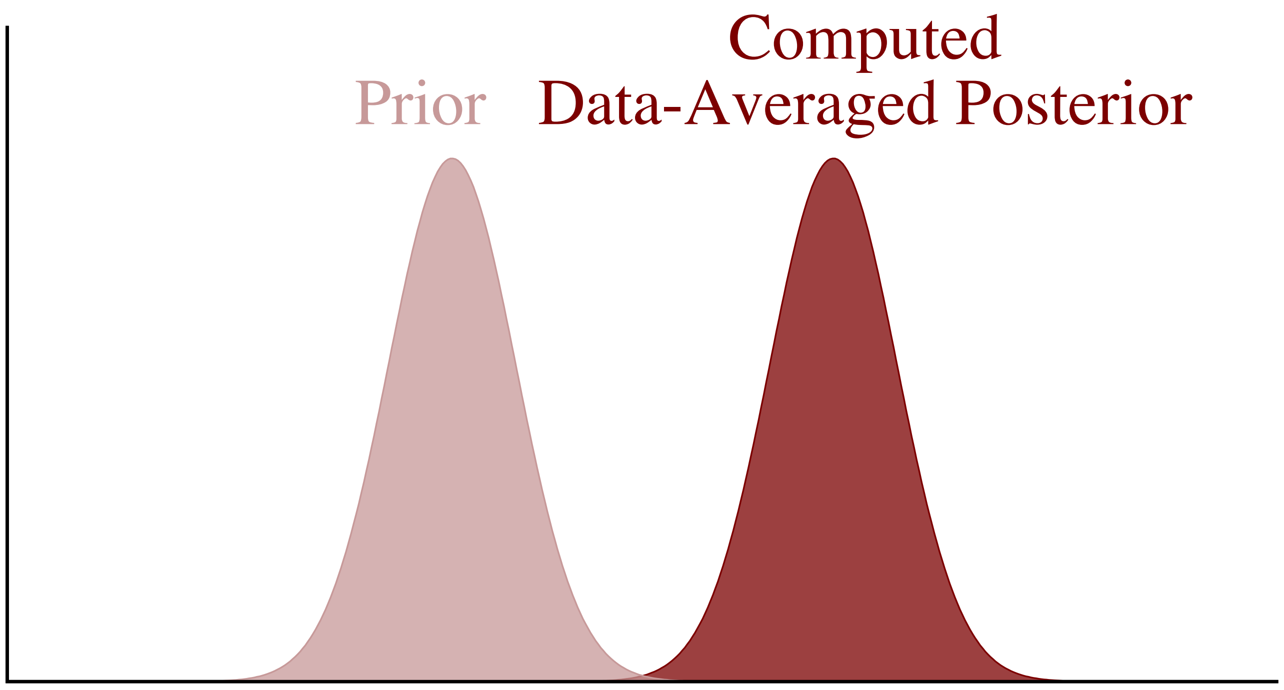


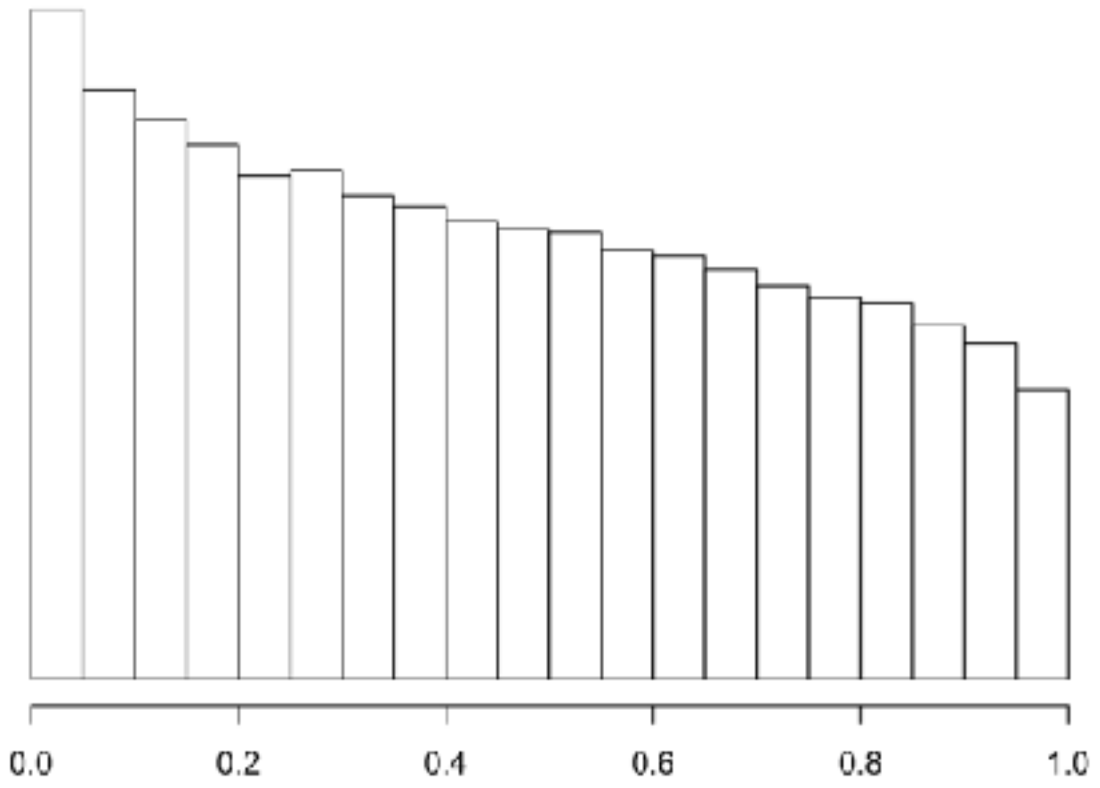
POSTERIOR TOO NARROW





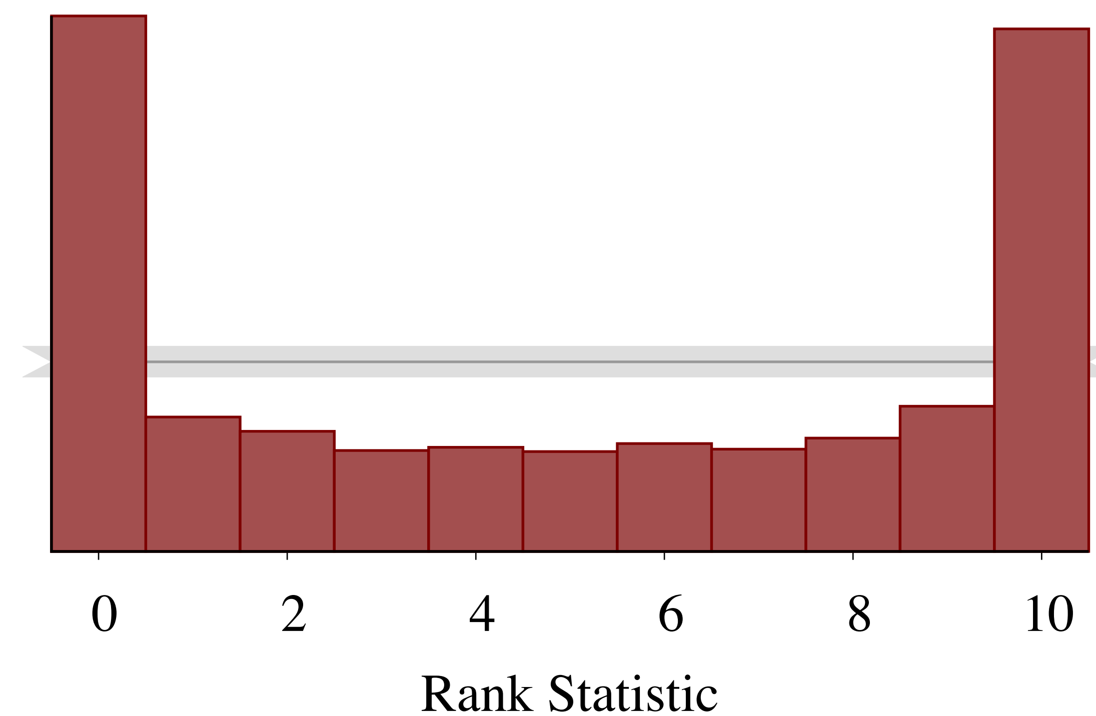
POSTERIOR BIASED TOWARDS LARGER VALUES





IT'S IMPORTANT TO USE ALMOST INDEPENDENT SAMPLES

- Draws from MCMC will usually be strongly correlated.
- This is bad!
- The theory only works for independent posterior samples
- **Solution:** Thin your Markov chain



THIS IS ALL A BIT ONE-DIMENSIONAL

- Everything here has been predicated on a one-dimensional parameter
- If we can compute the marginal posterior quantiles, we can check the univariate calibration for each parameter
- The system still works for functions $f(\theta)$
- We recommend checking the marginals, functionals of interest, and a collection of random linear functionals
- This should be sufficient to see if things have worked
- (NB: The cost of checking a new functional is usually dominated by computing the posterior, so the more the merrier)

YES BUT DOES YOUR MODEL ACTUALLY FIT?

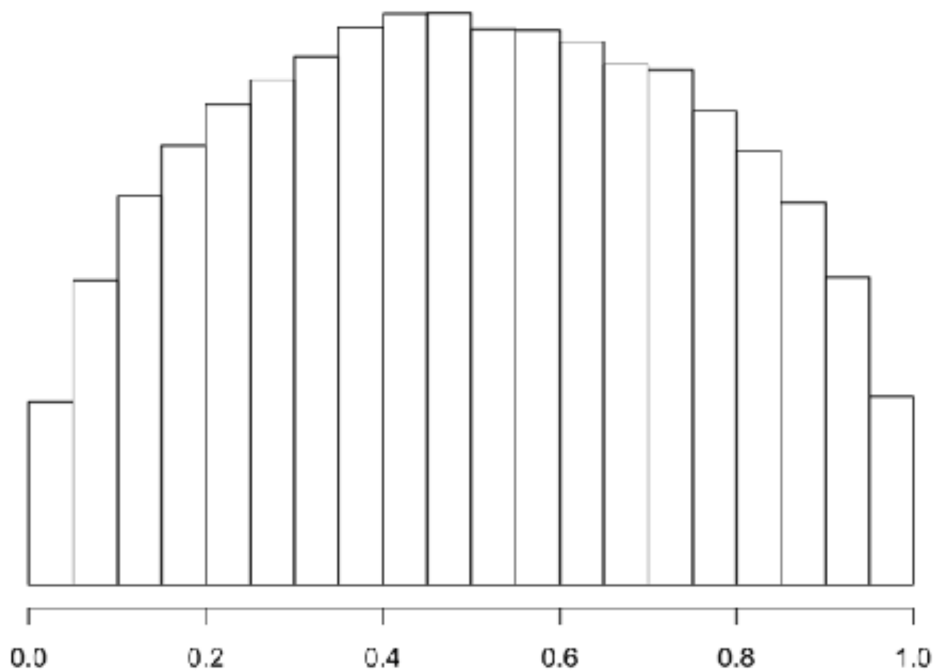
LOOKING AT SIMULATED DATA WAS USEFUL, WHAT ABOUT THE REAL STUFF?

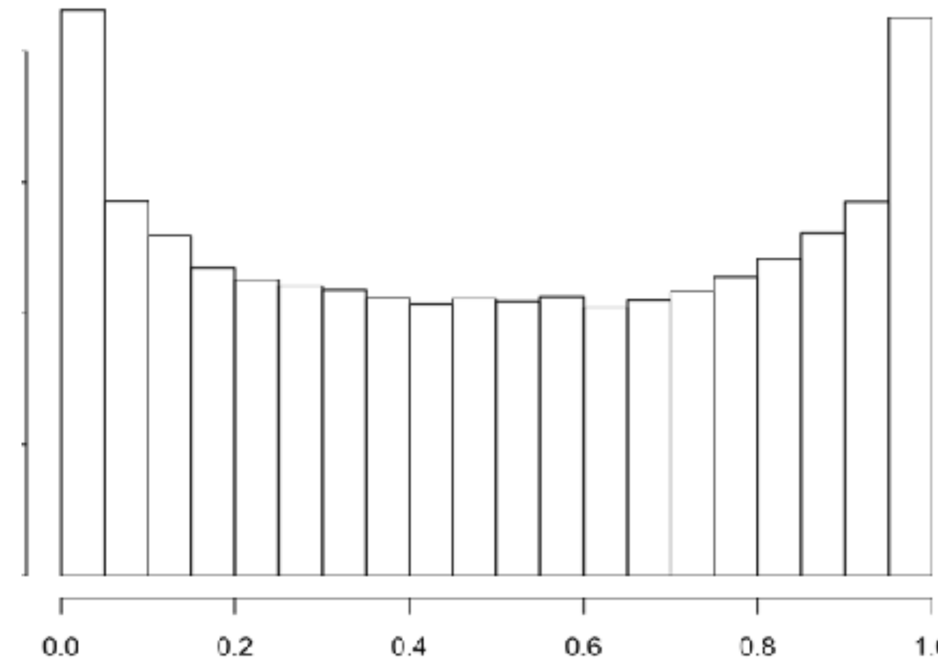
- Looking at simulated data was a good "sense check" for our algorithms.
- But if we want to see if our model has actually done an ok job, we need to do something similar for *real* data
- Idea: What if we look at the rank of a single data point y_i in a bag of samples from the posterior predictive

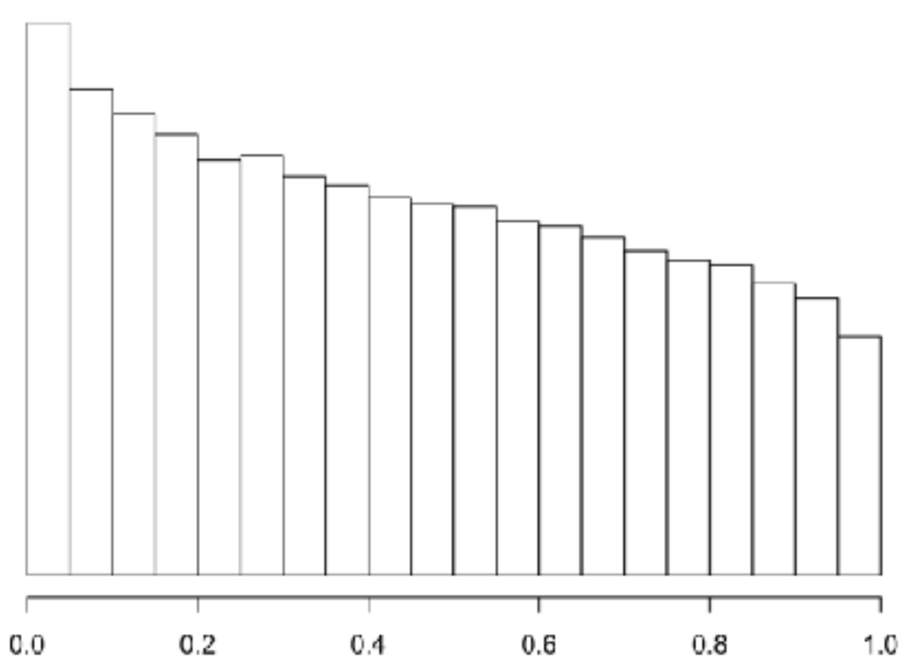
$$p(y_i \mid y_{-i}) = \int p(y_i \mid \eta, \theta) p(\eta, \theta \mid y_{-i}) \, d\eta d\theta$$

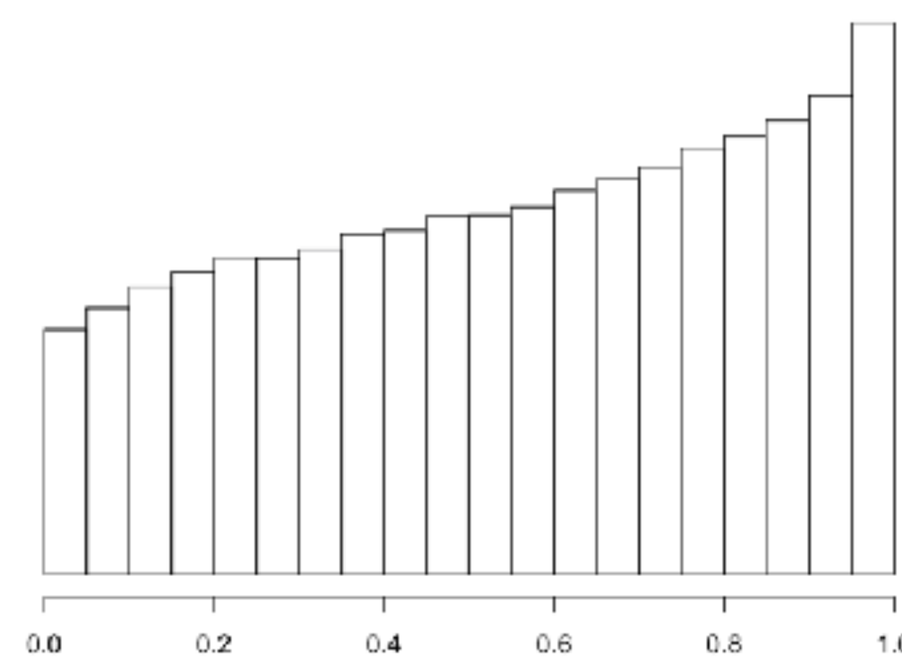
- Here y_{-i} is all of the data points *except* y_i

WE GET THE SAME HISTOGRAMS!









SOME CONCLUDING THOUGHTS

FINAL THOUGHTS

- Complex, multiresolution space-time models are hard to formulate and harder to fit
- There are a lot of traps you can fall into
- **Meaningful** priors are important. Don't just slap any old gaff on
- We really don't know how to compute big likelihoods, but empirical Bayes will fail for uniformed parameters
- Finally, it's best to think of Bayesian analysis as a **workflow** rather than a single magical thing that you only do once. Check your model before, during, and after your analysis!

REFERENCES

- J Gabry, D Simpson, A Vehtari, M Betancourt, A Gelman. *Visualization in Bayesian workflow. Journal of the Royal Statistical Society Series A (to Appear with Discussion), 2018*. arXiv: 1709.01449, 2018
- A Gelman, D Simpson, M Betancourt. *The prior can generally only be understood in the context of the likelihood. Entropy, 2018*.
- S Talts, M Betancourt, D Simpson, A Vehtari, A Gelman. *Simulation-Based Calibration: Validation of Bayesian Inference Algorithms*. ArXiv. 2018.