# Bayesian Inference Part II: Samplers Strike Back

DS6040 Fall 2024
Teague R. Henry


SCHOOL of DATA SCIENCE

# Outline

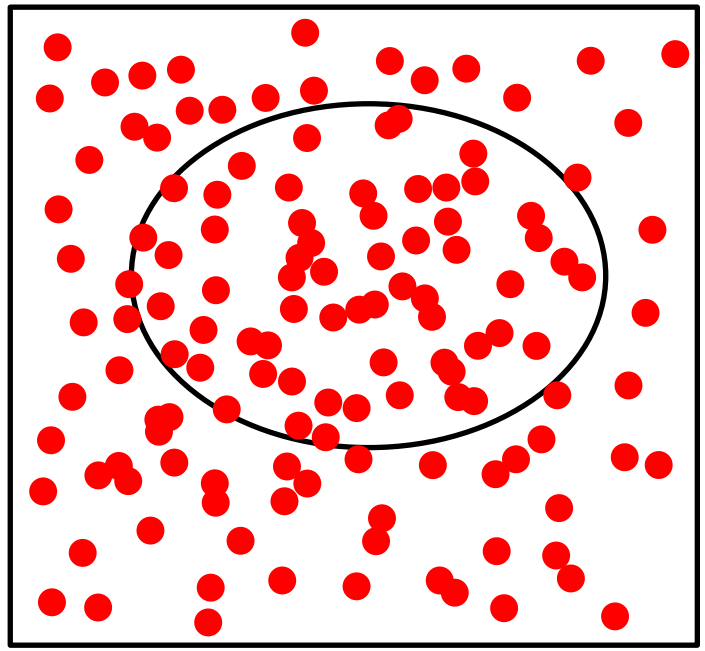
- Sampler Review
- What could Possibly Go Wrong? (With Samplers)
- What a Sampler Should Do
- Eyeballing Your Sampler
- Formal Diagnostics
- Sampling Categorical Variables

# Sampler Review

The basic ideas behind samplers are:

- Find where in the parameter space the posterior is.
- Sample from that region of parameter space.

If all goes well, your set of samples can be used to characterize your posterior distribution.



**Important Aside:** Remember you can mix and match sampling steps…

# Sampler Review – Types of Samplers

Gibbs Sampler –

- Iterate through sampling from proper conditional distributions
  - i.e. Sample from $\theta_1|\theta_2, X$ then from $\theta_2|\theta_1, X$
- Requires that you have *proper* conditional distributions
  - i.e. $\theta_1|\theta_2, X \sim Normal(\theta_2 X, 1)$ for an arbitrary example
- Always use Gibbs if possible, it's very powerful and has guaranteed good performance!

Random Walk Metropolis –

- Use if the conditional distributions are improper.
- Requires that you specify a proposal distribution for each parameter
  - Important choice, as it impacts convergence and performance
- Sampler walks through parameter space trying to find the posterior

# Sampler Review – Types of Samplers

Random Walk Metropolis-Hastings –

- Slight modification to the acceptance ratio step of the Metropolis sampler to account for asymmetric proposal distributions
  - Symmetry here refers to the probability of proposing values, not necessarily what the proposal distribution looks like.

Hamiltonian MC –

- Just a much better version of Metropolis-Hastings
- Uses gradient information
- Very powerful for use in high dimensional problems
  - Not appropriate for use on categorical variables

No U-Turns Sampler -

- Fixes issues in Hamiltonian MC

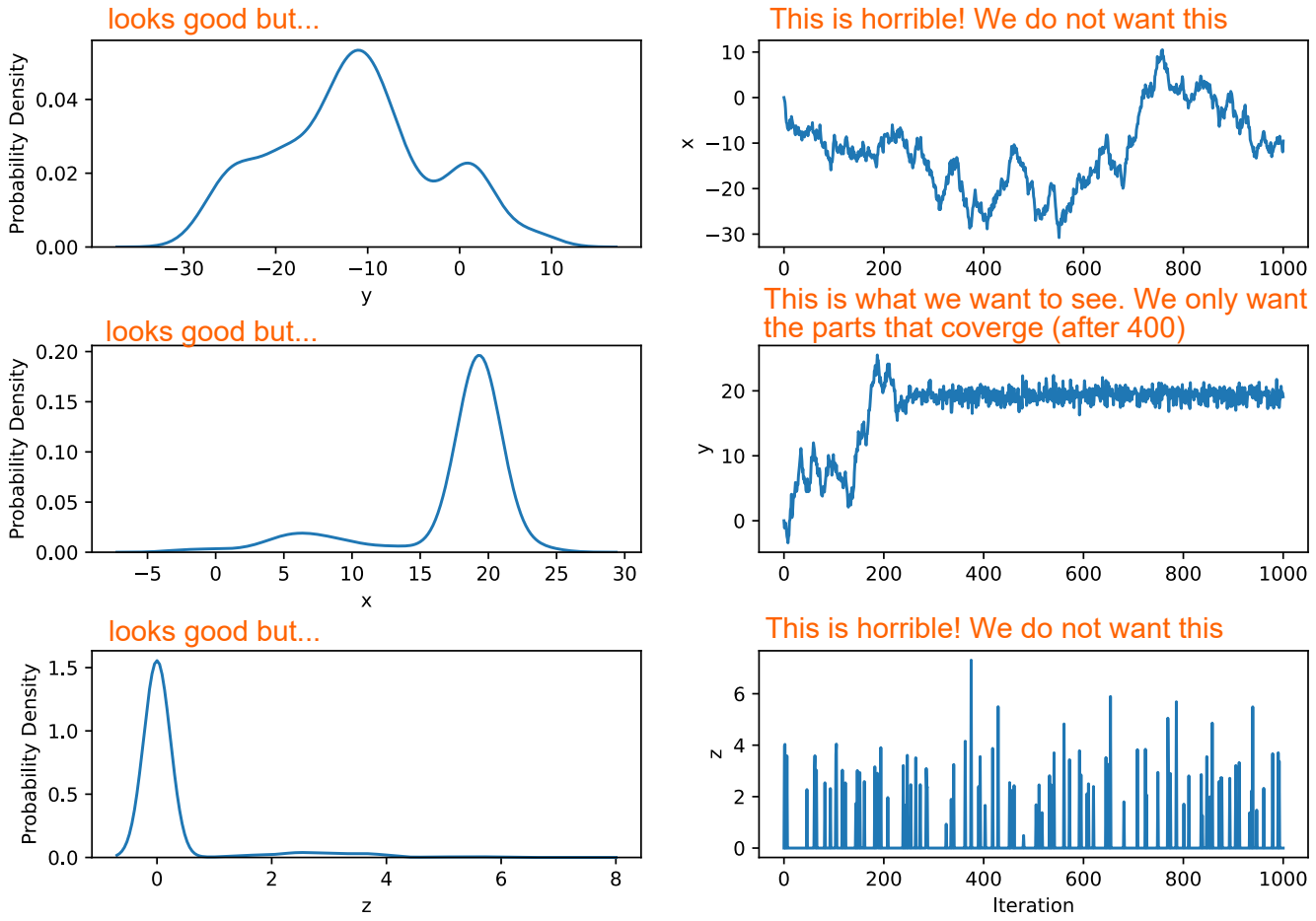
# What Samplers Should Do

If everything goes right, samplers should converge to the posterior and **stay there forever.**

Well constructed samplers are *ergodic*

- They cover the whole space
- Which means they tend to find the posterior.

**You can have an appropriate model, and still have sampler issues**



# What Goes Wrong with Samplers

Samplers are highly sensitive to the topology of the parameter space.

- They can get lost in valleys
- They can drop down holes
- They can't climb massive cliffs

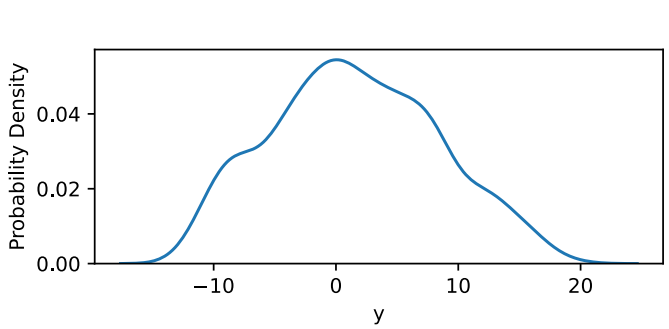
If the space is nice, the sampler is **ergodic**, in that it will nicely ramble all around the space.

**Like data itself, samplers can and will lie to you..**

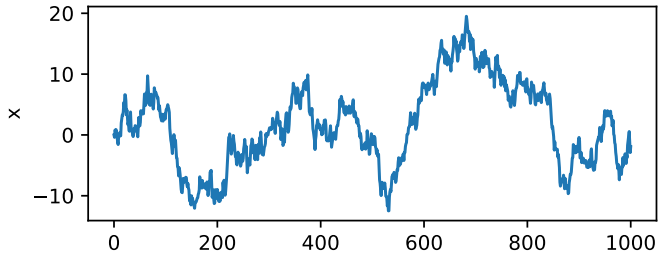
a good example of this is on the last slide with the samplers. you NEED to look at the trace plots

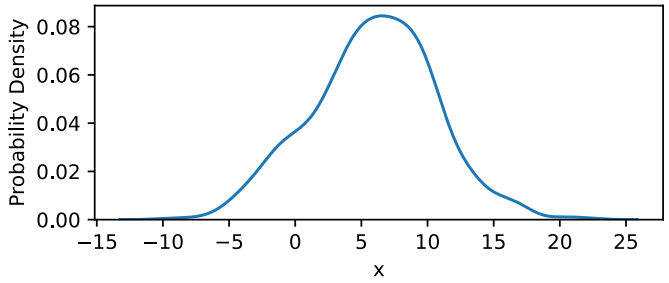
- The results you get can look completely reasonable, and be completely incorrect

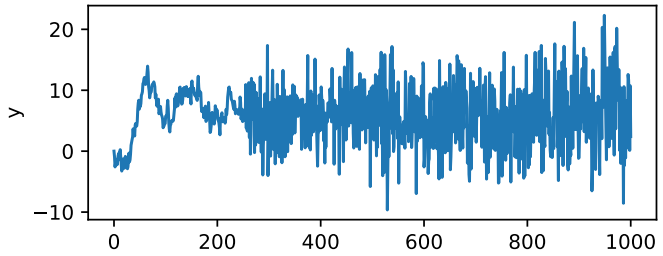
# Sampler Pathologies – An Informal Guide

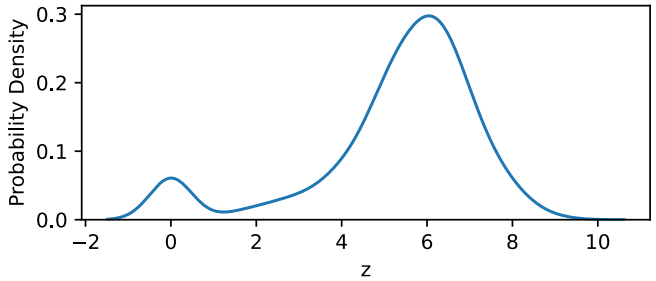


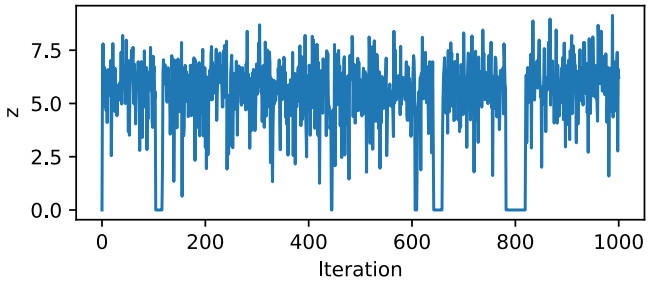

trace plots




## Random Walk Pathology –

- I've seen this behavior when two parameters are "doing" the same thing (i.e. are not uniquely identified)
- Think solving for $x$ in $x^2 + y^2 = 1$ when $y$ is unknown.

## Funnel Pathology –

- The walk's variance is increasing (or changing more broadly)
- Usually something do to with bad identification in **other parts of the model.**

we get a convergence, but our variance is everywhere

## Switching Pathology –

- Usually a PEBCAK, you probably misspecified a mixture somewhere
- e.g. the marginal mean of a mixture

# Geweke's Z-score Diagnostic

we want convergence...

**Convergence** – When the sampler has settled into the posterior, and will remain for the rest of time.

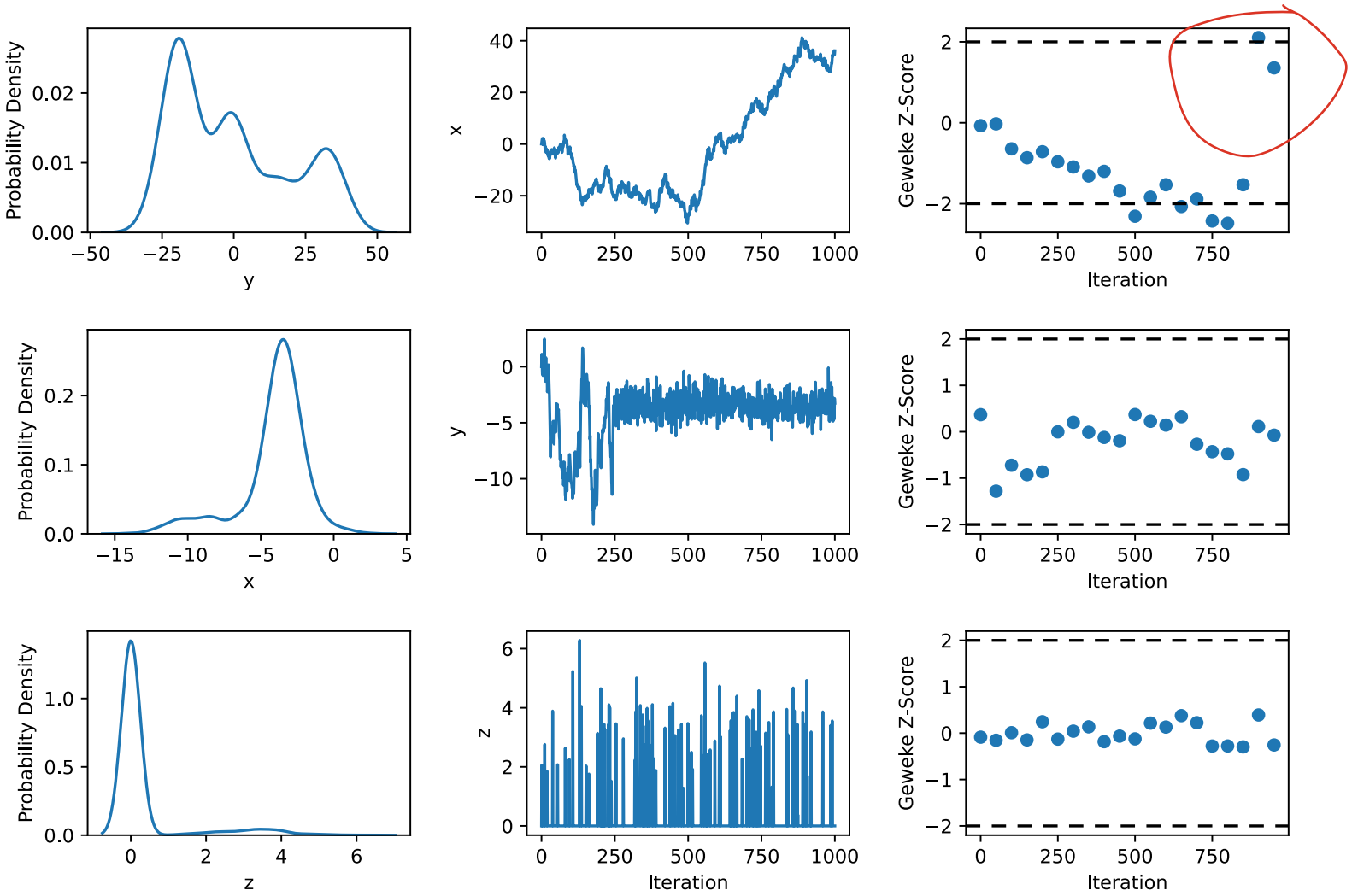
What does "remain till the rest of time" entail?

In a converged chain, the expected values of various segments should be very close.

**The Geweke Z-Score** –

A two-sample t-test calculated on $i$th percentile bins at the beginning vs. $j$th percentile bin at the end of the chain. --> does the beginning of your chain look like the end of your sample chain?

# Geweke's Z-score Diagnostic




Top Panel – Yes, bad convergence.

we see points that lie outside of our standard z-score numbers of our statistic (-2, 2)

Mid Panel – Not an issue

this is fine, they converge

if you're just looking at your z-score, you are going to be missing this info. You ALWAYS need to look at your trace plots.

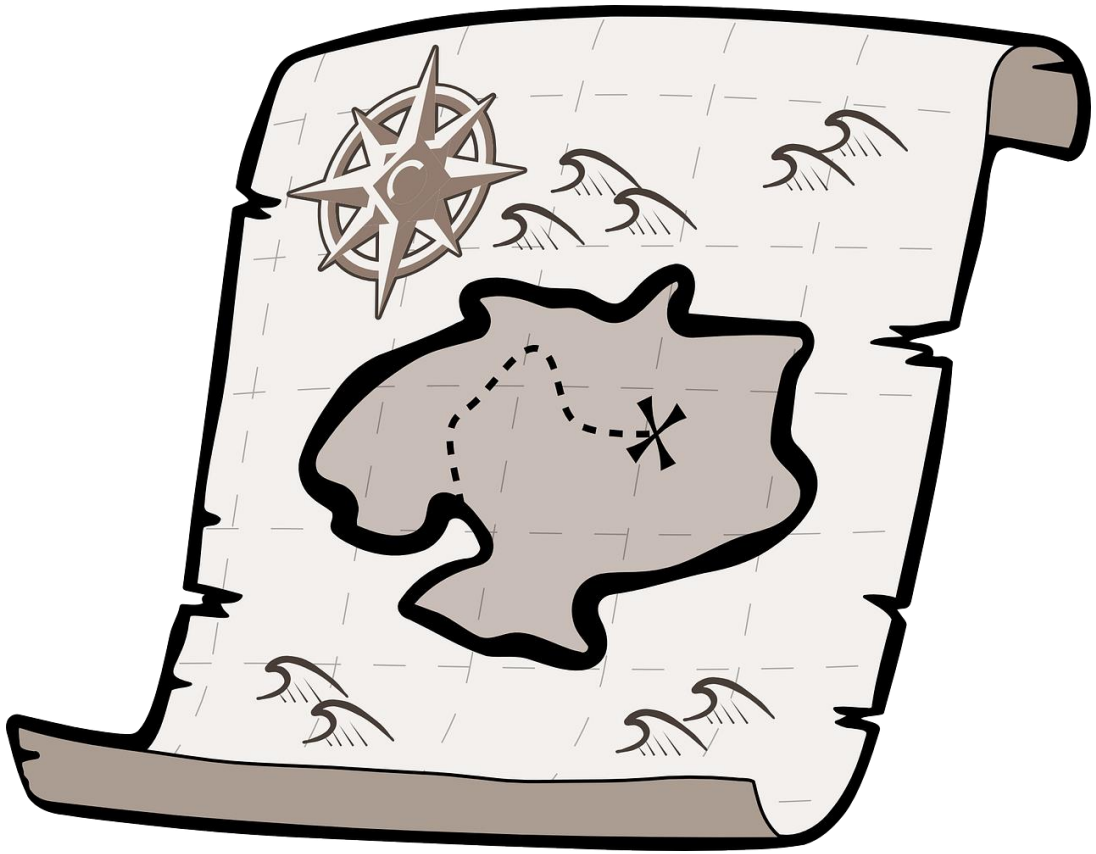
Last Panel – No evidence, but bad chain.

# Sampler Diagnostics – An Informal Guide

**Q:** If you are digging for buried treasure, what's better?

A. Sending out 1 digger with a month's supply of food?

B. Sending out 30 diggers with a day's supply of food? This one b/c you're able to cover more ground

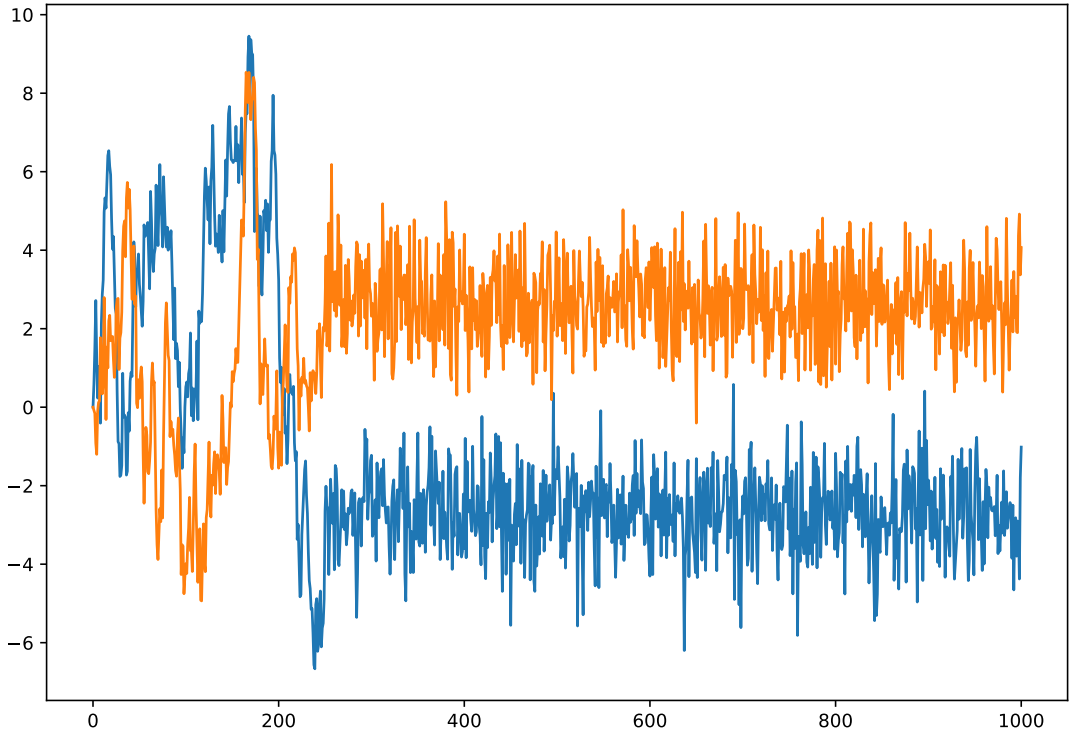


Our sampler is **in** the space, and our goal is to map it. One powerful method is simply to have multiple chains going from different starting points. When we rely on samplers, we rely on multiple chains. Let's look at the next slide

# Sampler Diagnostics – An Informal Guide

Multiple Chains – more exploreres!

- By starting multiple chains from different starting locations, you can more effectively explore the space.
- Also provide powerful diagnostics.

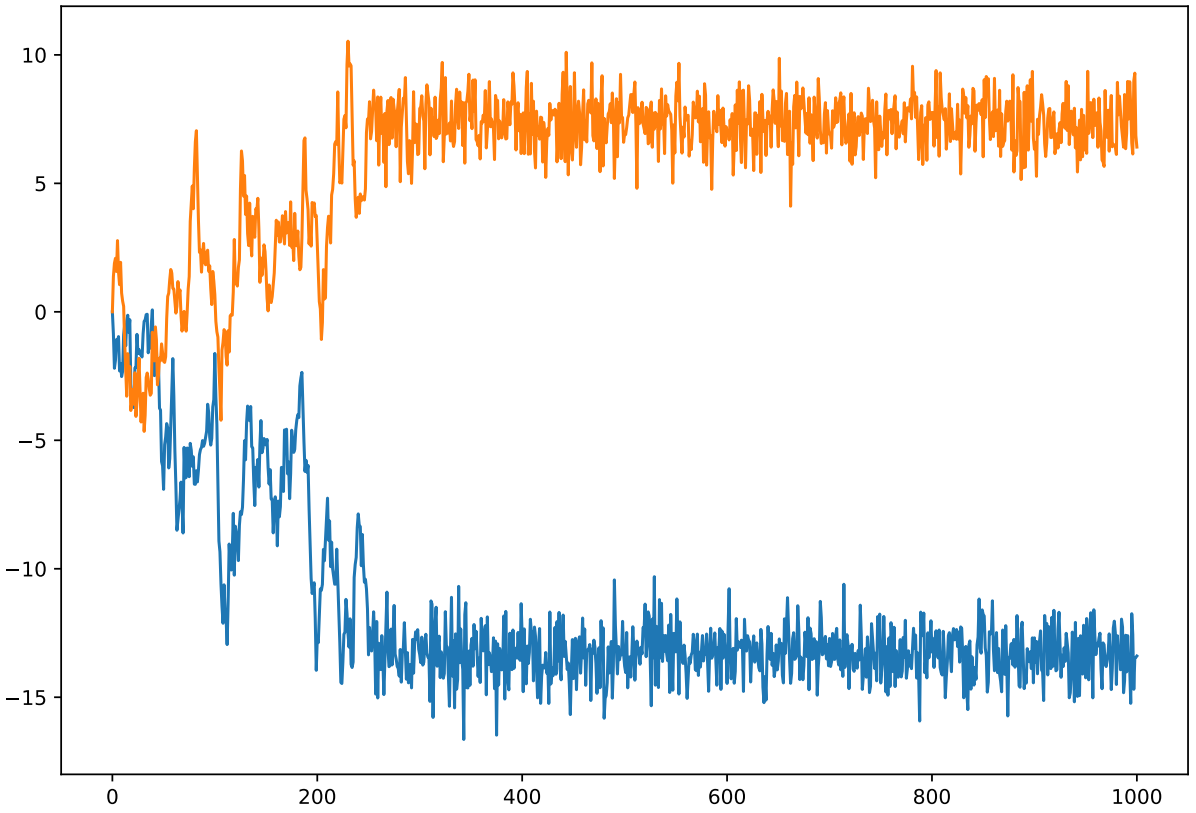



Gelman-Rubin $\hat{R}$ statistic –

- Converged chains should, by definition, look very similar.
- $\hat{R}$ looks at the within- vs. between- variance.
- We want $\hat{R}$ near 1.

if our chains have converged, then our between variance should be the same as the within variance

# Sampler Diagnostics – An Informal Guide




Gelman-Rubin $\hat{R}$ statistic –

- 4.76! <-- HORRIBLE!

Another effective measure is...

Effective Sample Size (ESS) –

- A converged chain has low autocorrelation.
- ESS measures how bad that autocorrelation is.
- Solution to this can be to thin the chains. E.g. run for 10,000, take every $10^{th}$ sample.

instead of regularly running all these models, we can just take every 10th sample and run that instead

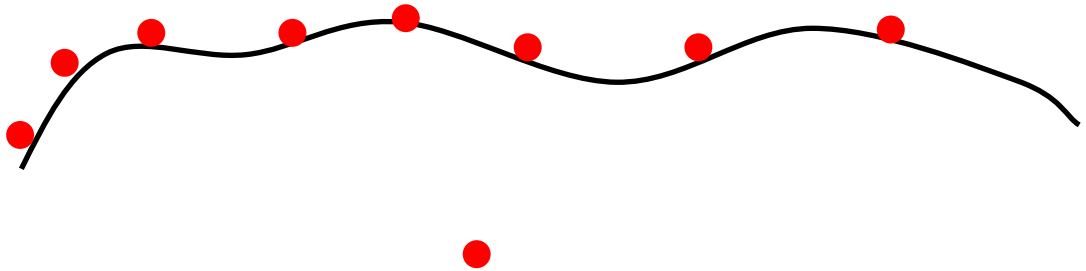
# NUTS Diagnostics– An Informal Guide

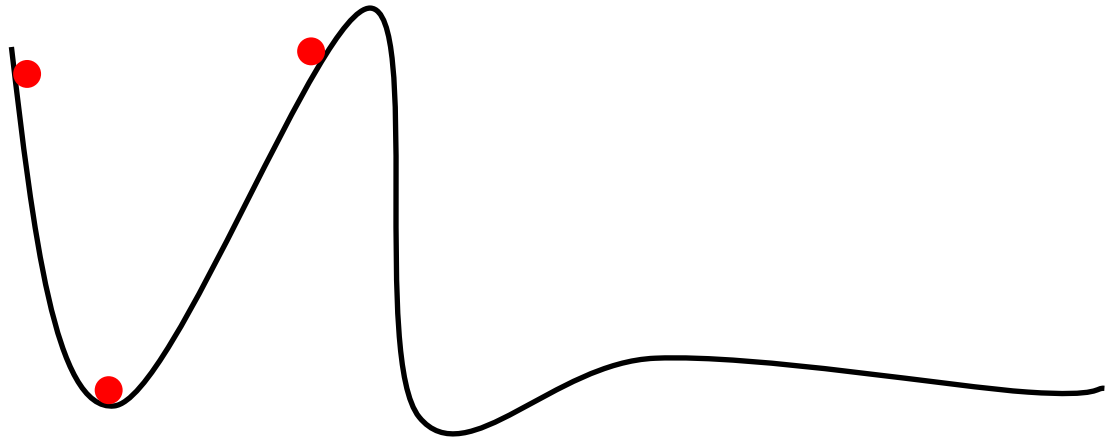
Recall that the NUTS sampler is using a “physical” simulation to, in essence, rollerblade around the space more effectively.

Divergence – When the sampler jumps too far from the target set.

- Indicates that the space has some pathological curvature.
- Possibly fixed by model reparameterization.

This is fine and exactly how NUTS is supposed to work





This is a divergence. This only happens when dealing with Hamiltonian stuff and NUTS samples. This indicates that there's some weird curvature in the space. This is caused by the very steep hill in the curve.

# NUTS Diagnostics– An Informal Guide

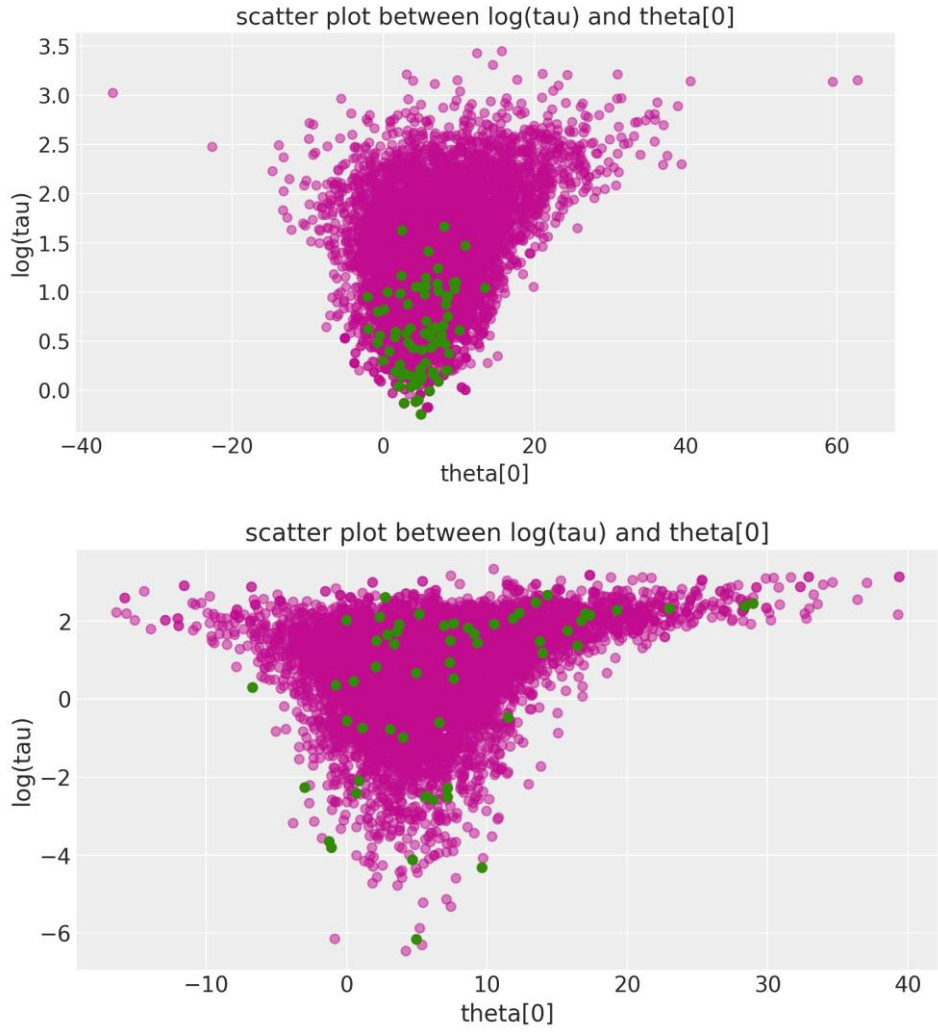
Divergences indication where in the parameter space there might be problems.

- Clustered divergences are indicators!
  this is the fault of the slopes of the model

But divergences can happen because the step size is too large.

- Then the divergences are more uniformly distributed. the fault of the stepsize (you're going to fast or moving too quickly)

**Guideline:** More divergences, more issues.



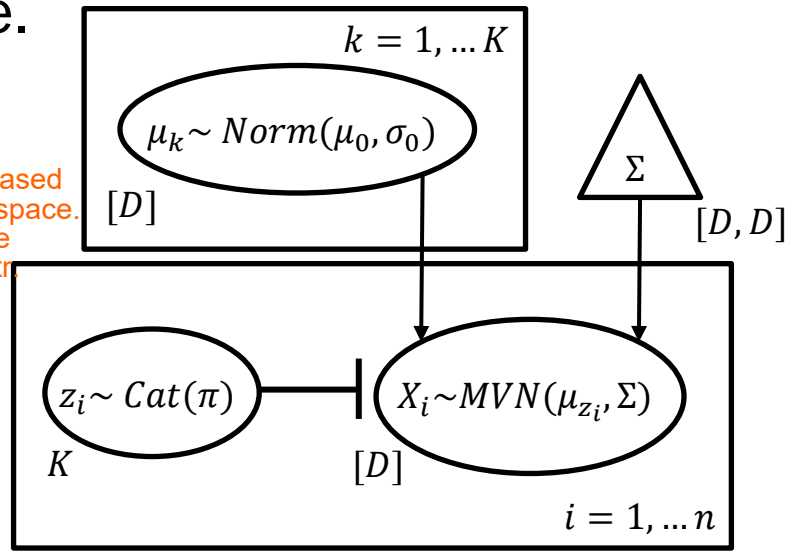

Plots taken from the excellent tutorial (Author: Michael Betancourt)
https://docs.pymc.io/notebooks/Diagnosing_biased_Inference_with_Divergences.html

# Sampling Categorical Variables

Simple Gaussian Mixture Model –

- Hyperparameters suppressed in figure.

a mixture model is a way of parameterizing a cluster model. For ex. K-means is estimated and evaluate based on where values are in the space. A mixture model says I have two clusters w/ different distr and I want to make distr. assumptions about them




Unsupervised Clustering –

- $\mu_k$ is estimated from the data.

Consider a mixture model of 2 groups

- A and B, each having different means.

When you build your sampler, you create a variable `z_i`, which takes on values *A* and *B*.

Question: Are *(*A and B), and (*A* and *B*) the same?

Are the data generating clusters (A and B) the same thing as the estimated labels? Am I guranteed that my estimate A is going to be my cluster A? NO, b/c the model has no idea what we labeled them as. It could be that the estimate of (itilisized B) refers to cluster A.

# Label Switching

**Label switching** – When the meaning of category (i.e. discrete classes) differs between **AND WITHIN** sampling runs/chains.

These are just two stats that you calculate on cluster statistics. They calculate if things are the same or are different

**Why?** – In unsupervised clustering, a label is just that, a label.

- The definition of each cluster depends on the observations in it
- In one chain, True Group A could be labeled as Group 1, in another chain, Group 2.
- If the clusters are not well defined, then this switch can happen mid sample.

like if the clusters are really close, we can see a switch up stating that the meaning is switching.

**Problems** – Can't run label specific between chain diagnostics.

The chains will converge on different means b/c they're meaning different things. How do we fix this?

**Solution** – Evaluate convergence using a label switch immune similarity measure like **Adjusted Rand Index (ARI)** or **Variation of Information (VI)**

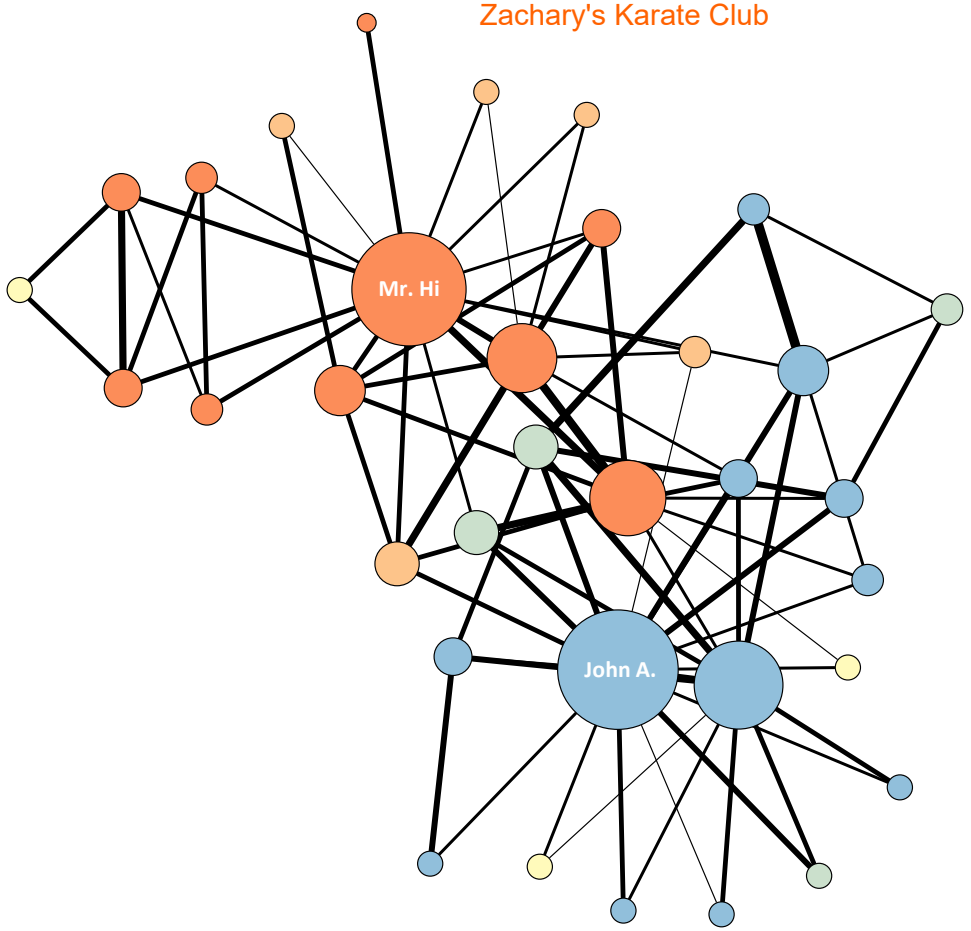
# Combinatorics and Networks

Consider a (real) social network:

- You make friends based on common interests.
- You also make friends from the friends of your current friends. (Transitivity).

**Two related problems:**

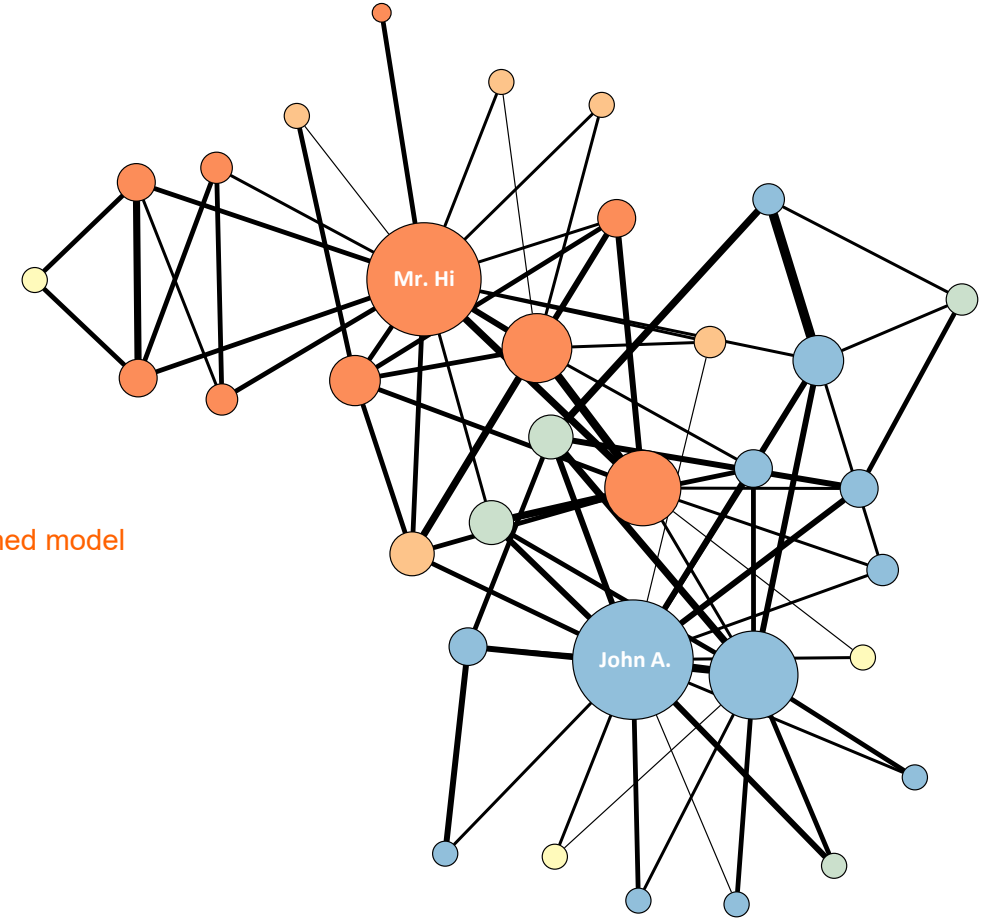
- Clustering the network
- Estimating a parametric model on the network.



# Combinatorics and Networks

## Clustering –

- When you move a node from one cluster to another, you change more than one nodes' most probable cluster.
- You can't update blocks of nodes, you need to update one at a time.




## Parametric Modeling –

this will take forever to run, even w/ a well defined model

- In transitivity respecting models of networks, the normalizing constant is over all possible combination of edges.
- More networks than particles in the universe...

our sampling methods for this specific problem relies on hope :D, b/c it's too difficult to do

# Combinatorics and Networks

## Clustering –

- When categorical variables where the meaning of the categorical variable depends on the values across the entire sample (i.e. clustering), think carefully about how you update your cluster labels, and definitions of the clusters themselves.

## Parametric Modeling –

- Consider estimating multiple categorical variables within an observation…
  - It pains me to my core, but I'll use the Myers-Briggs personality "test" as an example.
- If the meaning of the combination is highly specific to the exact combination, then each combination needs to be evaluated.
- Assumptions help:
  - For example, define the combination of categories as an additive sum, e.g. mental health diagnoses.

Fairly niche topic yes, but when it comes up, it comes up bad…

# Summary

Diagnosing samplers is akin to a vacation post-mortem

- Where did everybody go? Was that place good? Did you stay there the whole time? Oh, your friends have gone there? They stayed the whole time? Nobody ran around too fast and flung themselves into the atmosphere? No? Great!

Within Chain Diagnostics –

- Where did it go, and did it stay there (Geweke's Z)
- Did it get where it needed to go, and is now wandering aimlessly? (ESS)

you want semi-random wandering

Between Chain Diagnostics –

- $\hat{R}$- Looks at differences between chains, needs to be close to 1 (not infallible)

you don't want the chains to be different

**Takeaway: No statistic is an infallible test. Don't be swayed by rules of thumb. Know what each test is testing and estimate your own mental posterior.**

**And most importantly: Always Eyeball Your Plots**

# Next Time: Mode Approximation

**Mode Approximation**

- When we have more then enough data, and all we want is a point estimate

**EM Algorithm**

- Kinda like Gibbs, but even more useful

**Variational Inference**

- When you have a billion observations, everything looks normal