CS 1675

Spring 2020

Cathedral of Learning: G24

Week 03

Normal with unknown standard deviation: Introduction

Last week, we walked through all the details of the normal-normal model…

- Our goal was to learn the unknown mean, μ, given N observations, $\mathbf{x}$, and ASSUMING the likelihood noise, σ, was known.

- We compared the MLE to the Bayesian approach.

- Discussed asymptotic behavior, when the prior standard deviation, τ_0, approaches infinity and when the sample size, N, approaches infinity.

Normal-normal model with known σ

$$p(\mu|\mathbf{x}, \sigma) \propto \prod_{n=1}^{N} \{\text{normal}(x_n|\mu, \sigma)\} \times \text{normal}(\mu|\mu_0, \tau_0)$$

But now…let's relax the assumption on a known σ…

- Perhaps instead of weighing a 25 dumbbell, I want to weigh myself.

- I no longer feel as confident in being able to just assume a fixed value for σ.

- We now have TWO unknown parameters to learn: μ and σ

The posterior is now a joint distribution between μ and $\boldsymbol{\sigma}$ conditioned on $\mathbf{x}$

- We denote the joint posterior distribution as:

$$p(\mu, \sigma \mid \mathbf{x})$$

Does the likelihood change?

- The measurement process is essentially the same. I’m still weighing something.

- So a Gaussian still seems appropriate as the likelihood for this case:

$$p(\mathbf{x} \mid \mu, \sigma) = \prod_{n=1}^{N} \left(\mathrm{normal}(x_n \mid \mu, \sigma) \right)$$

What about the prior?

- We now have to encode our belief not just about μ or not just about σ…

- We have to encode our prior belief about their joint distribution!

$$p(\mu, \sigma)$$

The joint posterior is proportional to the likelihood times the prior

- Bayesian formulation for the unknown mean and unknown likelihood noise:

$$p(\mu, \sigma | \mathbf{x}) \propto \prod_{n=1}^{N} \{\mathrm{normal}(x_n | \mu, \sigma)\} \times p(\mu, \sigma)$$

How do we handle the joint prior?

- Most statistics textbooks focus on factoring the prior as:

$$p(\mu|\sigma)p(\sigma)$$

- Allows making use of a conjugate prior on σ.

- Makes use of the math presented last time, and leads to analytic solutions for the marginal posterior on σ.

However, let’s consider a different formulation

- We will assume a-priori the parameters are independent:

$$p(\mu, \sigma) = p(\mu)p(\sigma)$$

- Represents that knowing something about σ does not tell me anything about μ.

The un-normalized posterior distribution can then be written:

$$p(\mu, \sigma | \mathbf{x}) \propto \prod_{n=1}^{N} \{\mathrm{normal}(x_n | \mu, \sigma)\} \times p(\mu) p(\sigma)$$

- Even though a-priori the parameters are assumed independent, the posterior may have a relationship between the parameters via the observations!

Independent prior specification

- We will continue to use a normal prior on the unknown mean:

$$p(\mu \mid \mu_0, \tau_0) = \text{normal}(\mu \mid \mu_0, \tau_0)$$

- For σ let's use a uniform prior defined between a lower bound, l, and an upper bound, u:

$$p(\sigma \mid l, u) = \text{uniform}(\sigma \mid l, u)$$

The un-normalized joint posterior:

$$p(\mu, \sigma|\mathbf{x}) \propto \prod_{n=1}^{N} \{\mathrm{normal}(x_n|\mu, \sigma)\} \times \mathrm{normal}(\mu|\mu_0, \tau_0) \times \mathrm{uniform}(\sigma|l, u)$$

The joint posterior distribution does **NOT** have an analytic or closed form expression!

Let’s start putting some numbers to this problem

- For the prior on my unknown mean weight…
  - I will use $\mu_0 = 250$ and $\tau_0 = 2$.
  - Thus, a-priori there is a $\approx$99% probability that I weigh less than 255 pounds!

- For the prior on the unknown likelihood noise:
  - Set the lower bound to $l = 0.5$ and the upper bound to $u = 5.5$.
  - The units on the bounds are pounds.

Why those bounds on σ?

- In this example, σ is the sampling noise or measurement error…remember the likelihood!

$$x_n \mid \mu, \sigma \sim \mathrm{normal}(x_n \mid \mu, \sigma)$$

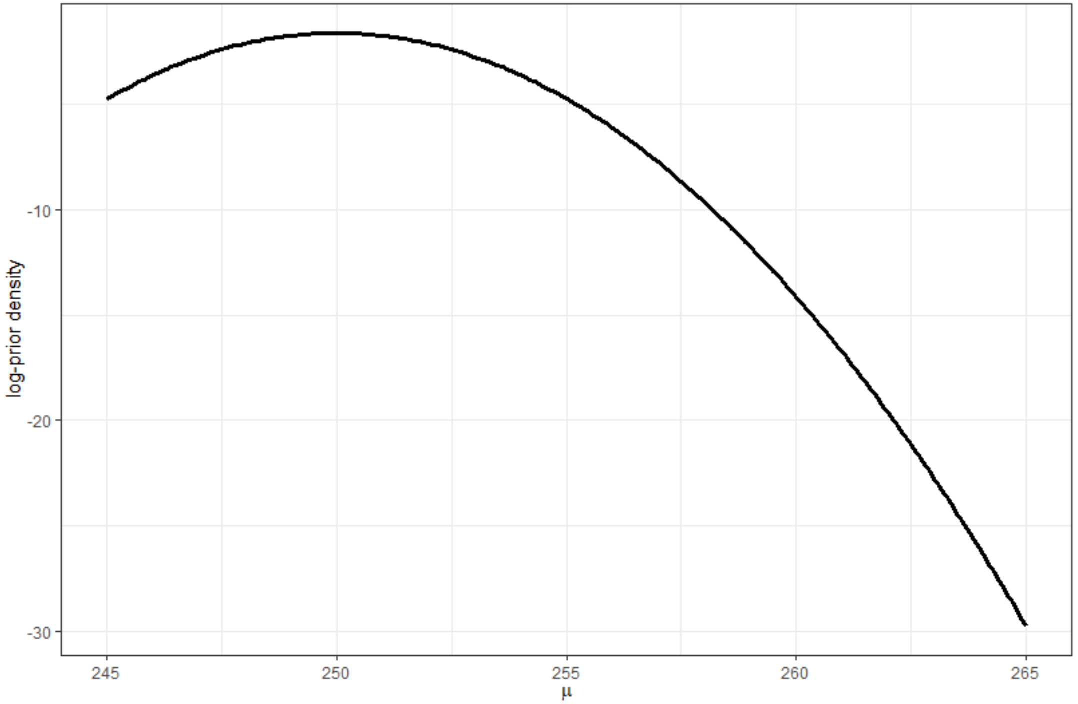
- σ is the lack of repeatability of a measurement.

Why those bounds on σ?

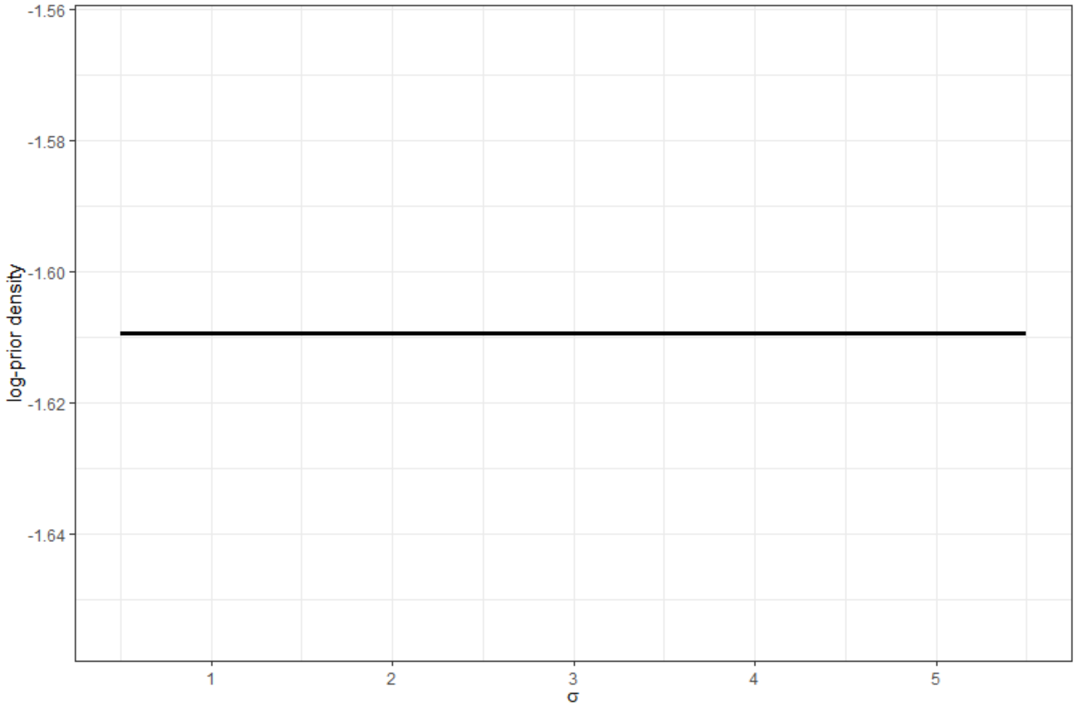
- We are not considering day-to-day variation in my weight.
- The noise represents variability in repeated measurements. For example, I step on the scale 10 times in a row.
- If I would weigh 250 pounds, $\sigma = 5$ pounds means that there’s $\approx$95% chance that the scale would read between 240 and 260 across those 10 measurements!

What do the marginal log-prior densities look like?

a-priori μ is centered at $\mu_0 = 250$ with $\tau_0 = 2$

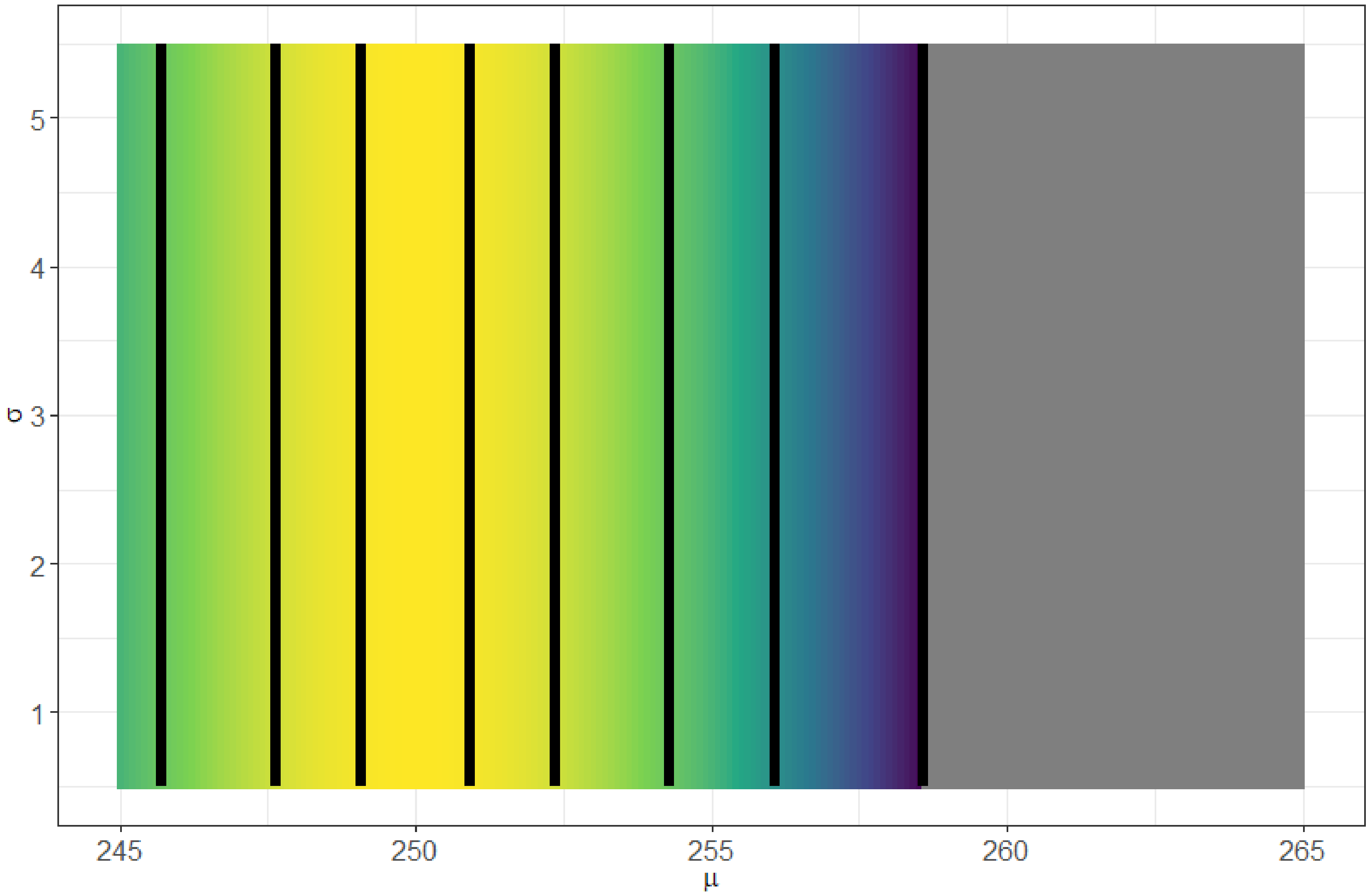


a-priori σ favors no interval in particular between $l = 0.5$ and $u = 5.5$



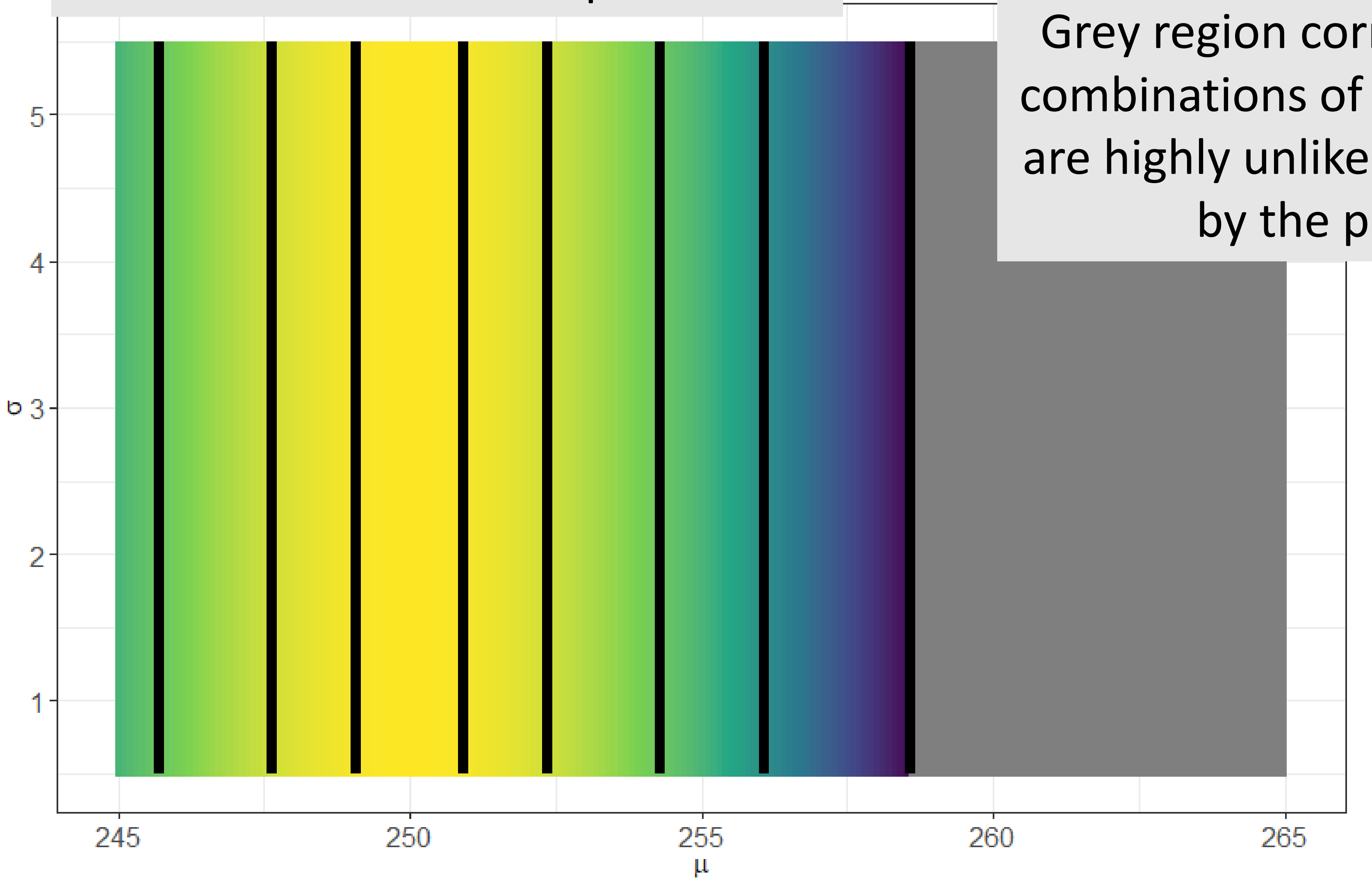
But, σ is constrained to exist ONLY between the assumed lower and upper bounds!

Putting the two together reveals the log-prior surface



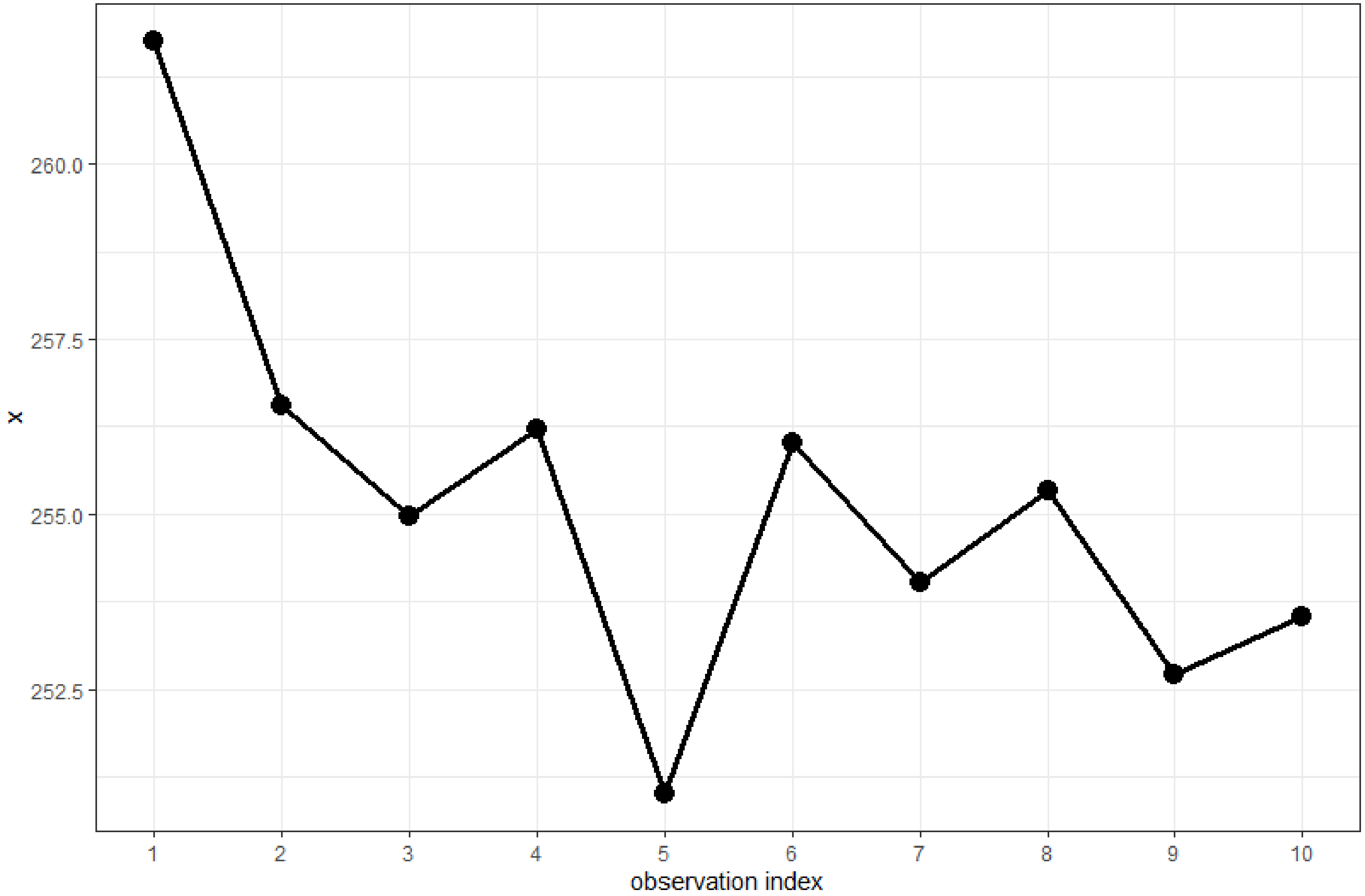
Putting t[illegible] the log-prior surface

Vertical contours are displayed because of the uniform prior on σ!



Grey region corresponds to combinations of μ and σ that are highly unlikely, as viewed by the prior!

Let's take a look at 10 measurements



Without an analytic expression…how can we proceed?

- Since we have just two unknowns, we can actually visualize the posterior, or rather the log-posterior, surface!

- We know how to write down the un-normalized log-posterior:

$$\log[p(\mu, \sigma \mid \mathbf{x})] \propto \sum_{n=1}^{N} (\log[\mathrm{normal}(x_n \mid \mu, \sigma)]) + \log[\mathrm{normal}(\mu \mid \mu_0, \tau_0)] + \log[\mathrm{uniform}(\sigma \mid l, u)]$$

- So if we specify some candidate values for (μ, σ) we can just calculate the un-normalized log-posterior!

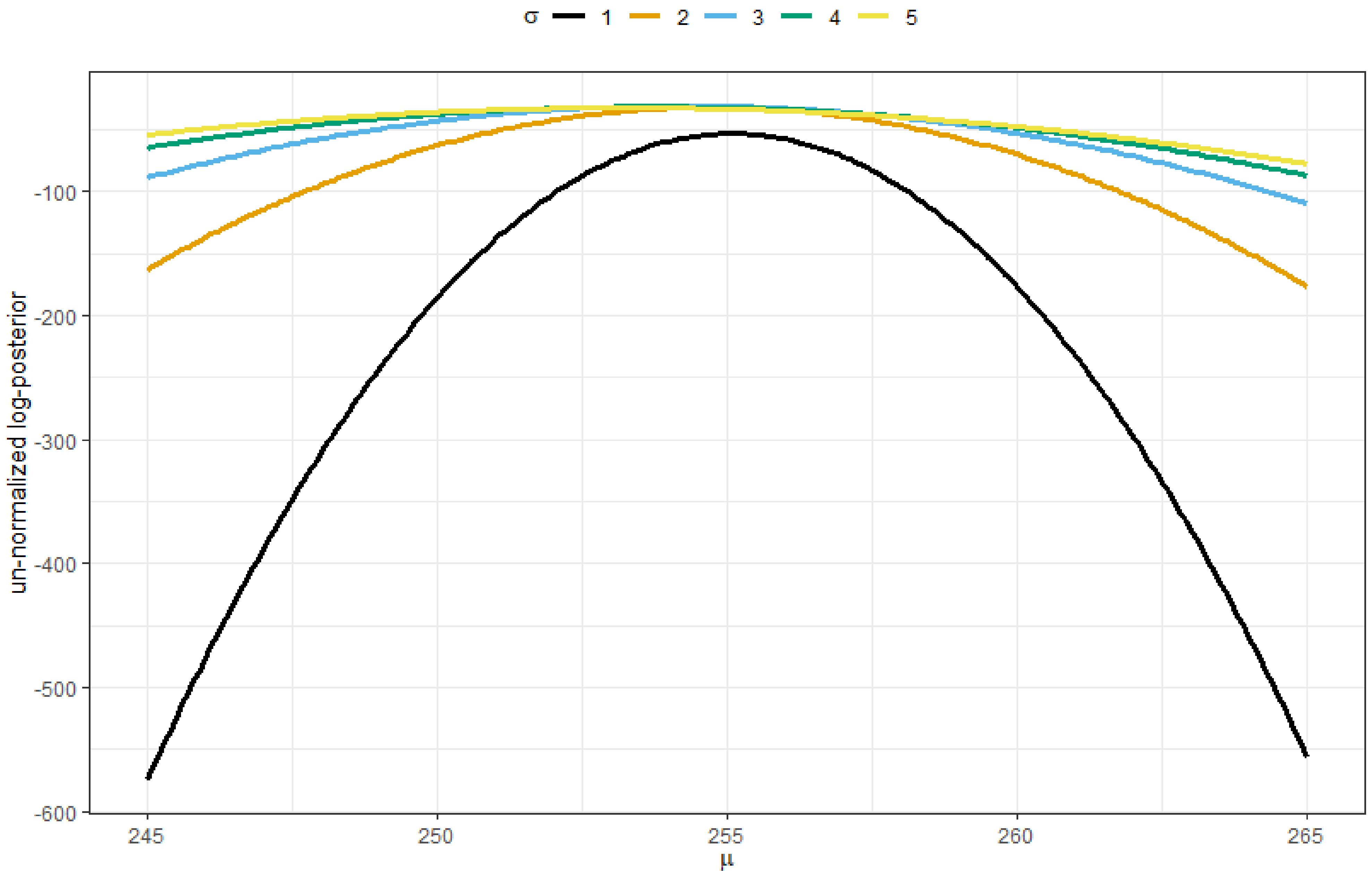
Before visualizing the surface, let's focus on one parameter at a time

- First, plot the log-posterior with respect to μ at a few specific values of σ.

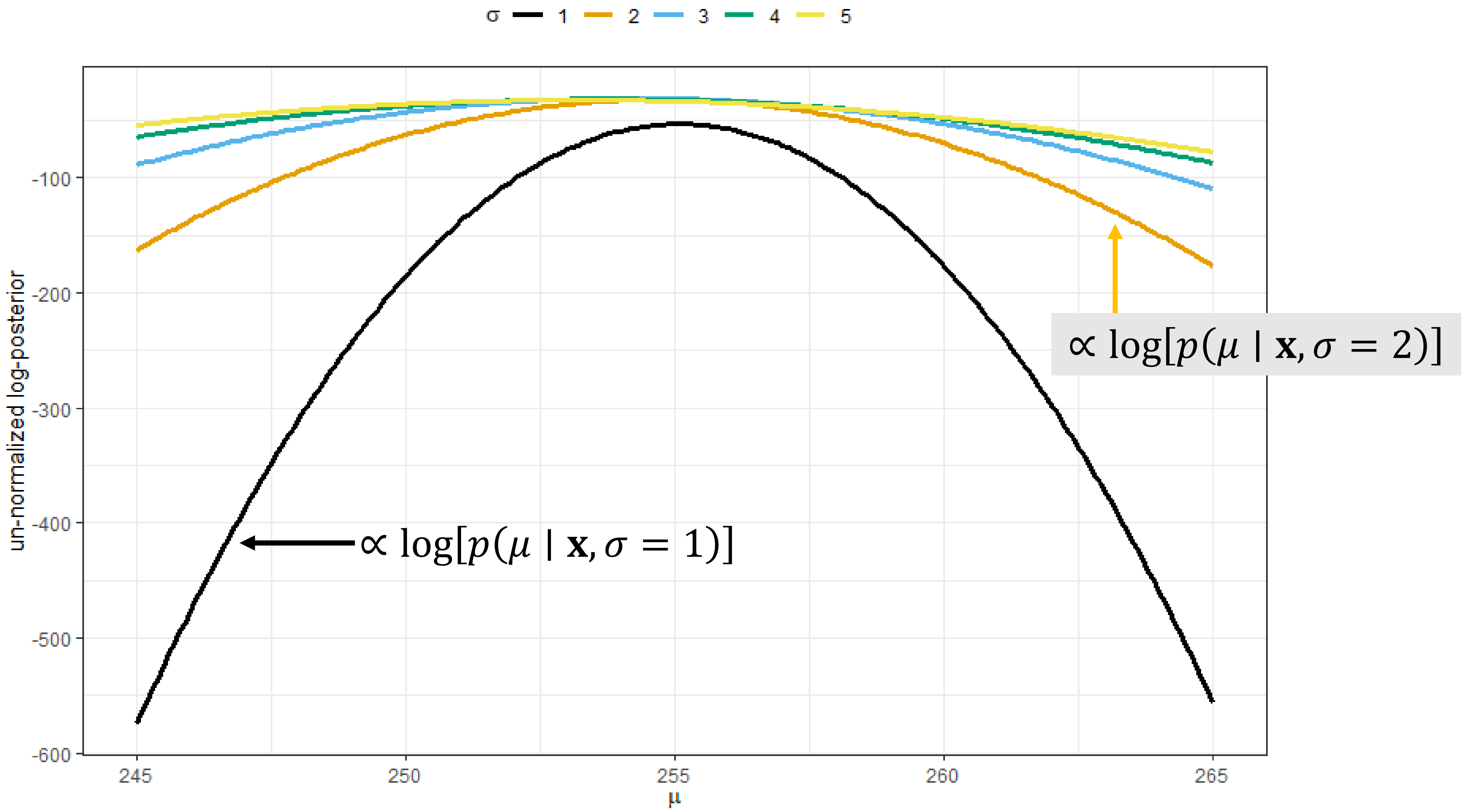
- Then we will plot the log-posterior with respect to σ at a few specific values of μ.

- Lastly, we will examine the log-posterior surface with respect to both μ and σ.

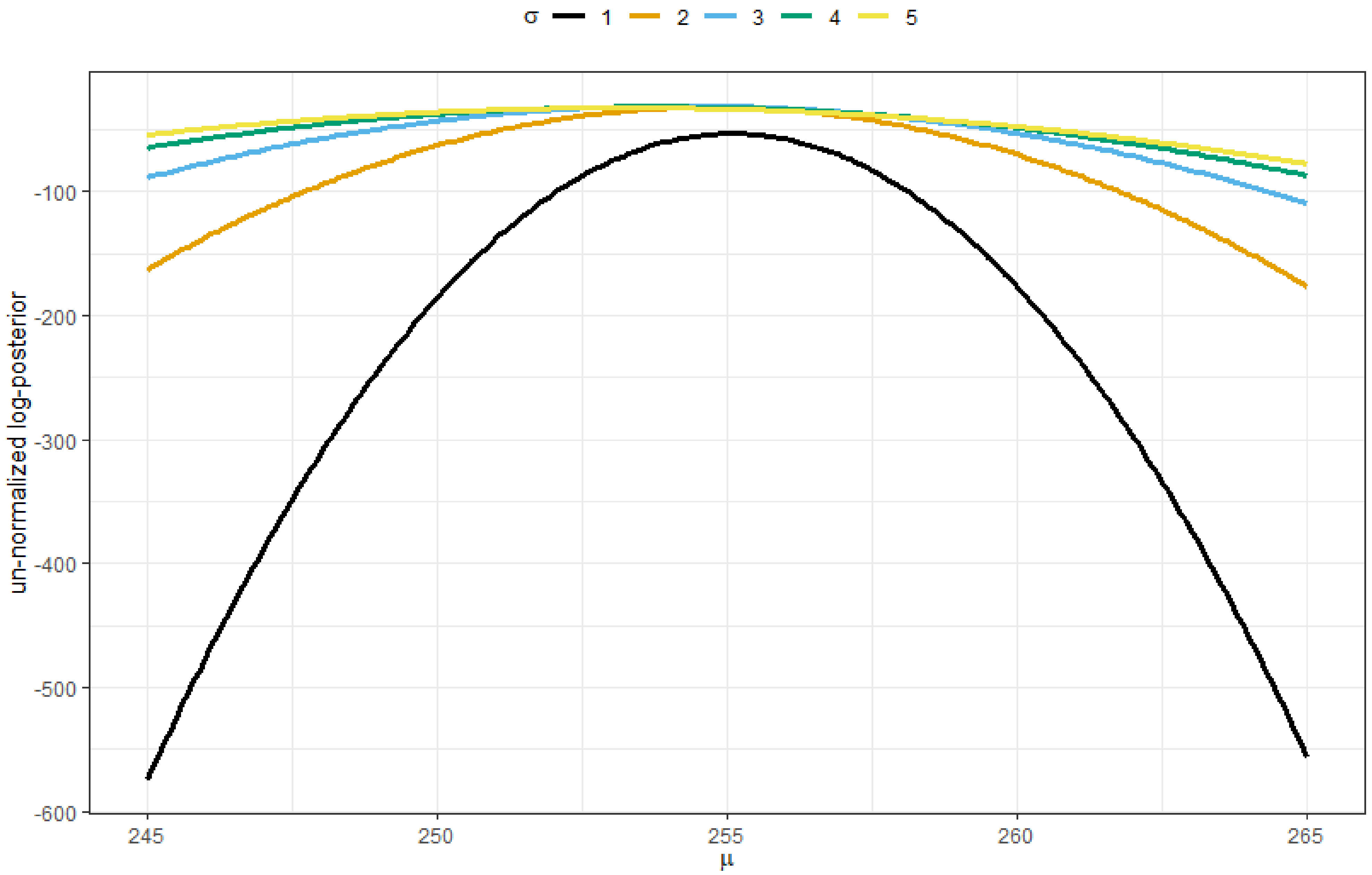
The log-posterior with respect to μ given specific values of σ should look familiar…



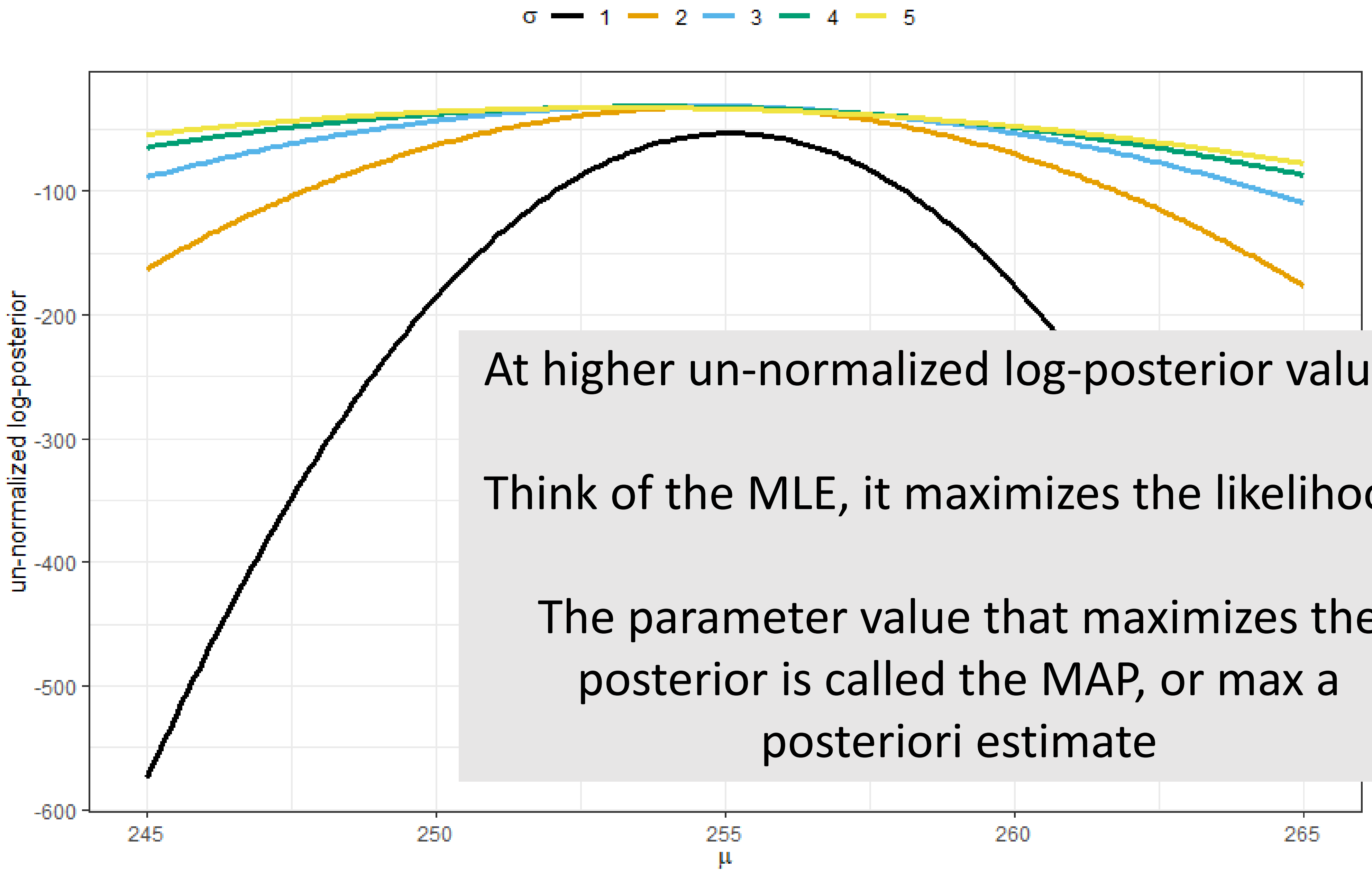
The log-posterior with respect to μ given specific values of σ should look familiar…



Posterior plausible values of μ are located where?



Posterior plausible values of μ are located where?

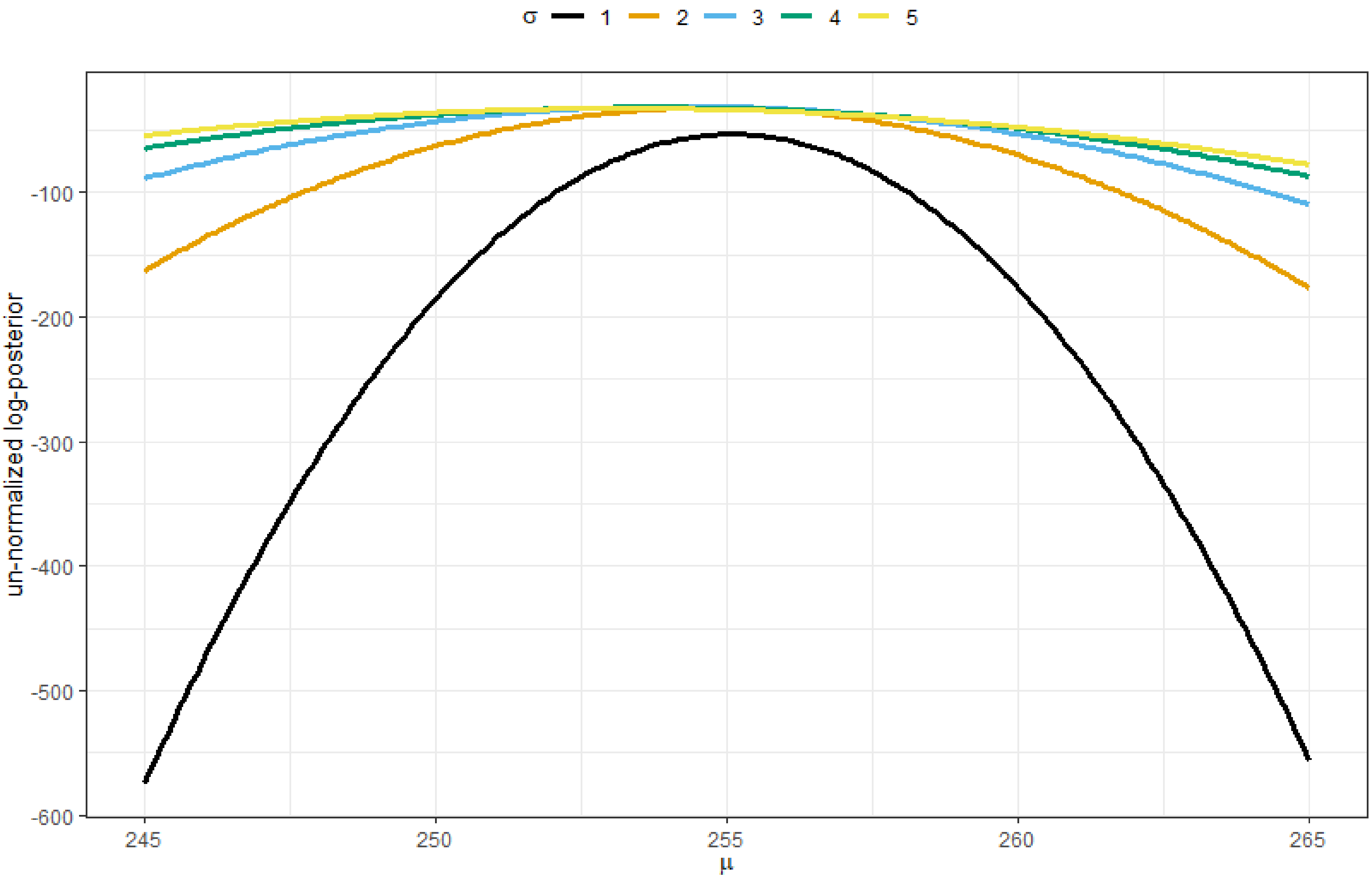


At higher un-normalized log-posterior values!

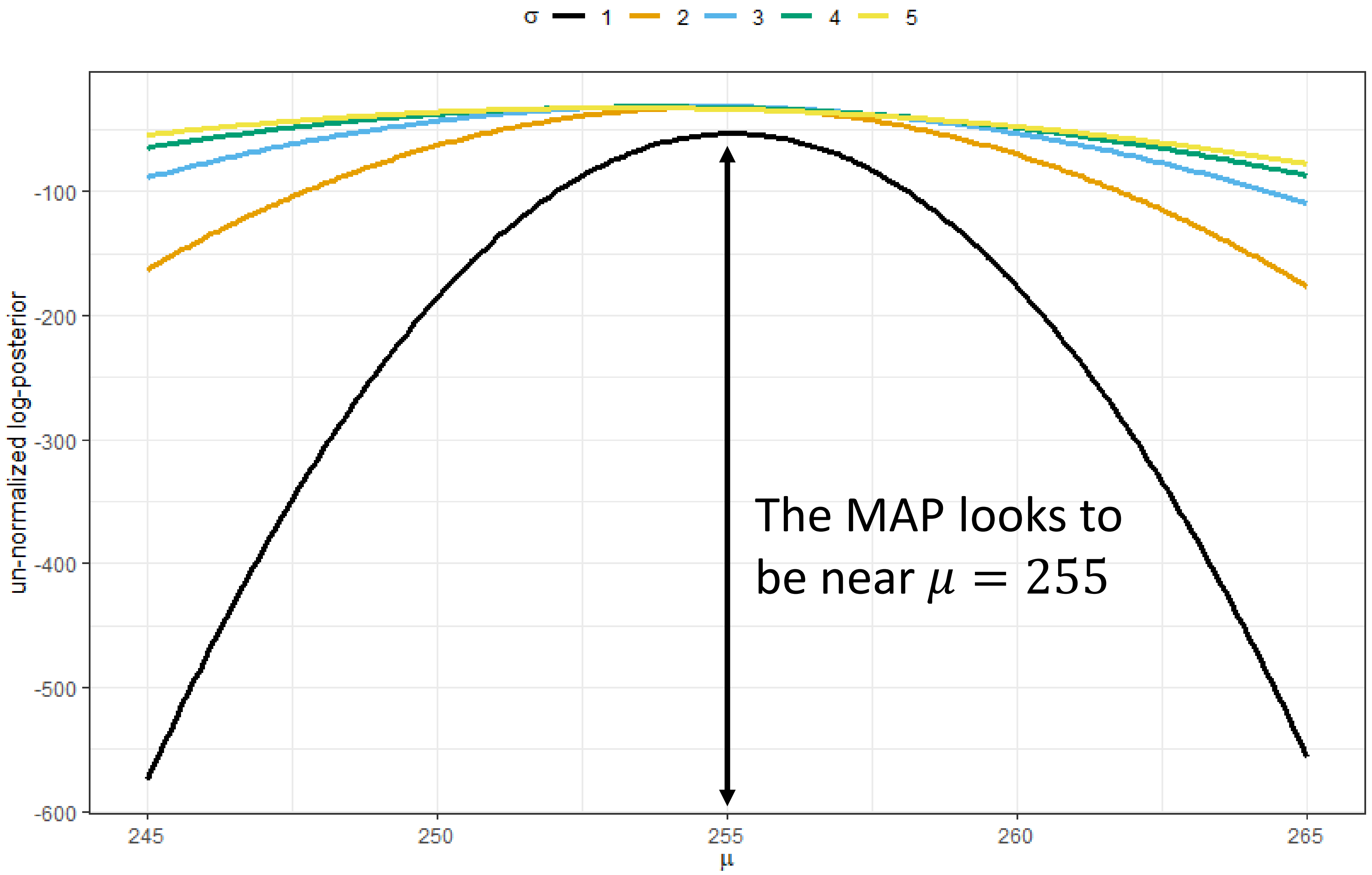
Think of the MLE, it maximizes the likelihood!

The parameter value that maximizes the posterior is called the MAP, or max a posteriori estimate

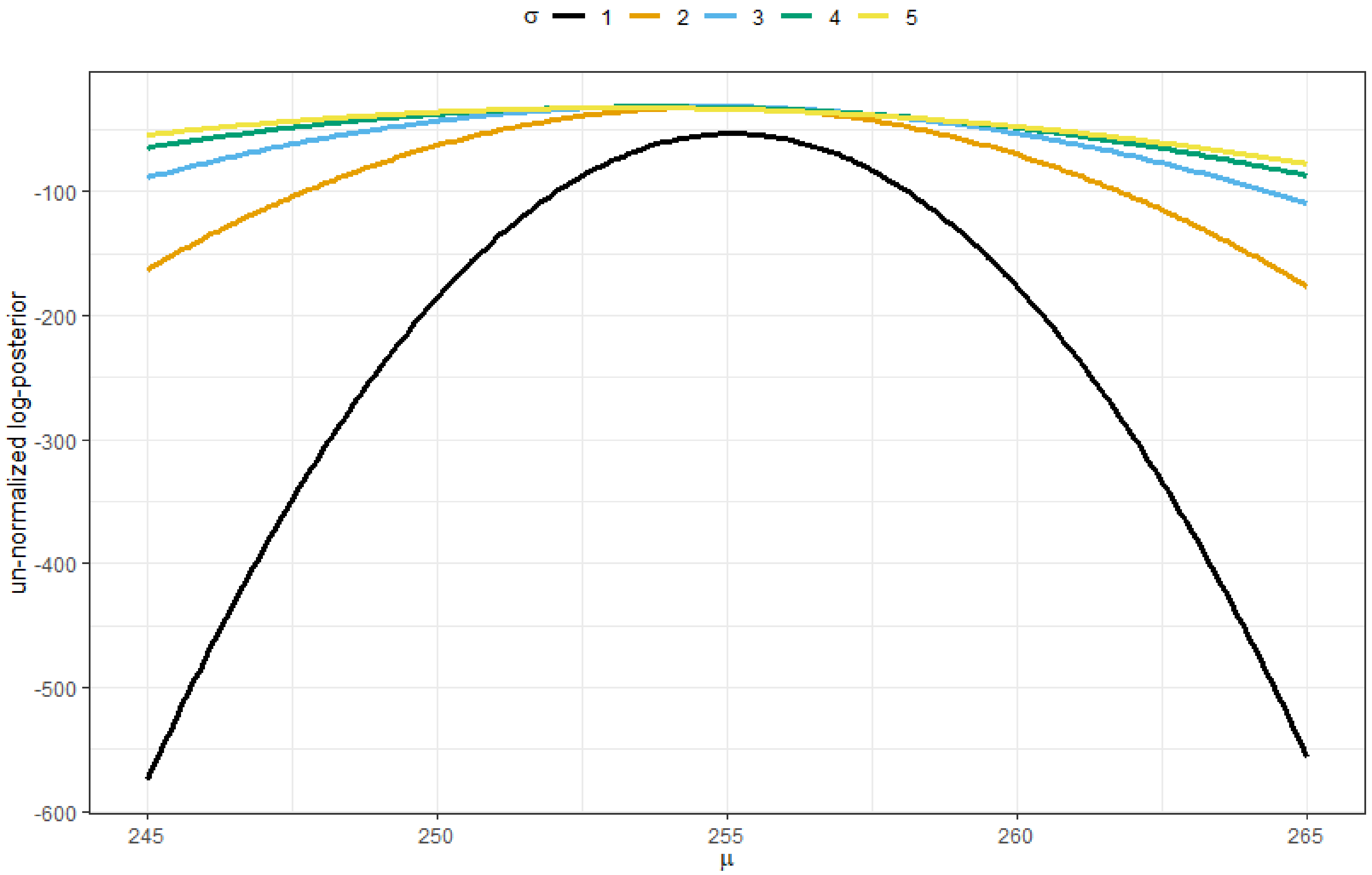
For $\sigma = 1$, what is the MAP on μ?



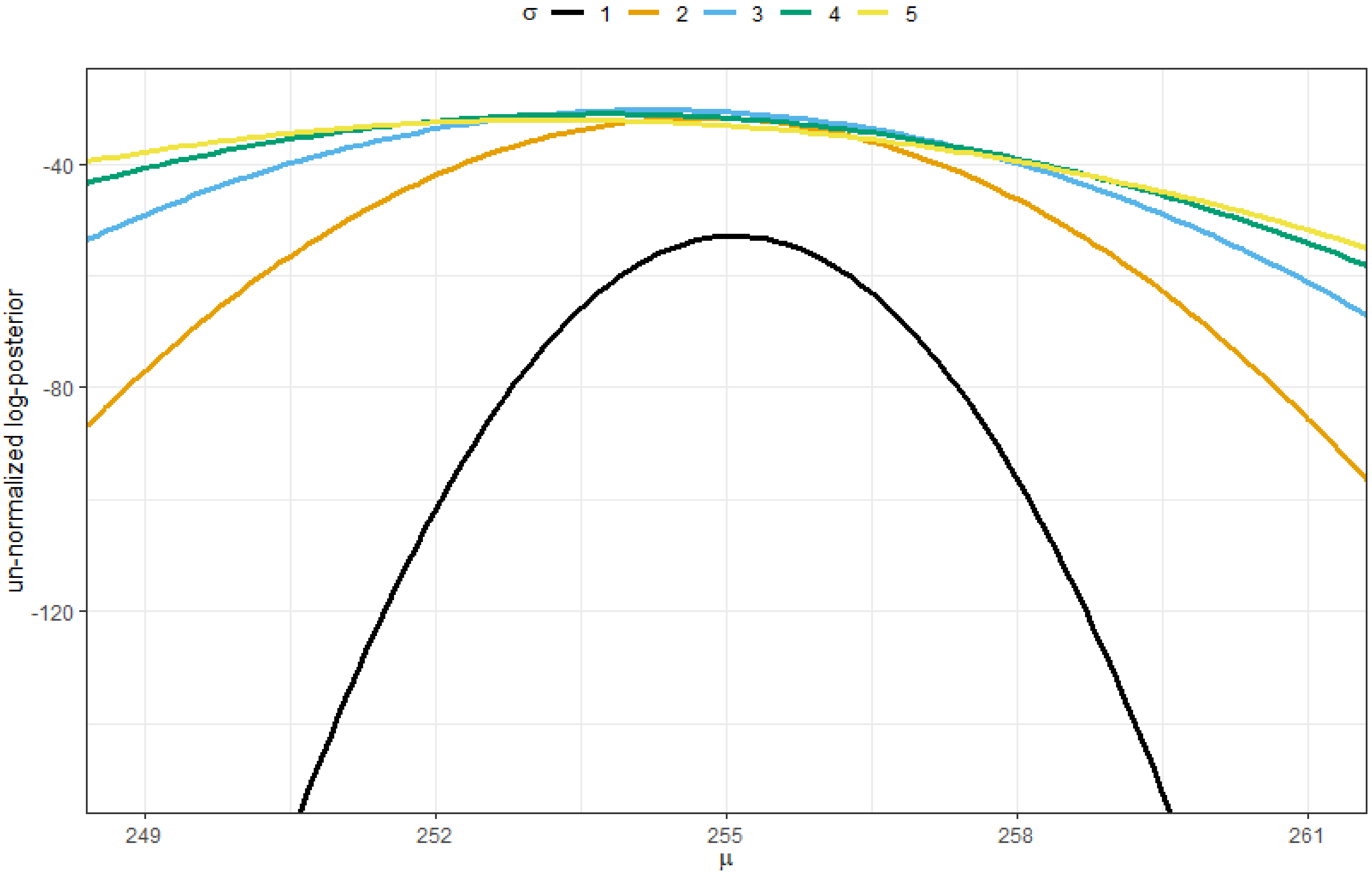
For $\sigma = 1$, what is the MAP on μ?



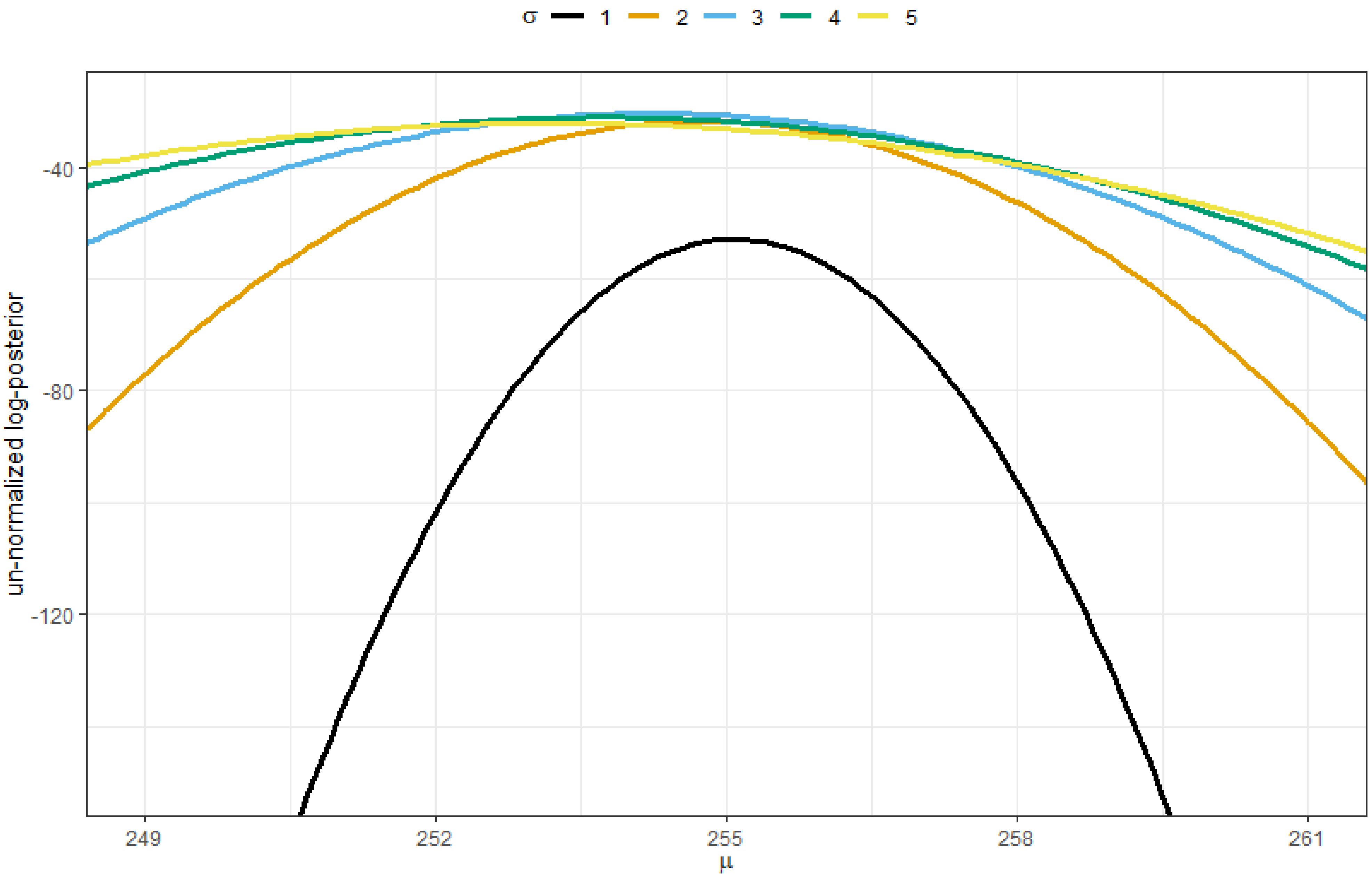
But…is $\sigma = 1$ more plausible than other values of σ?



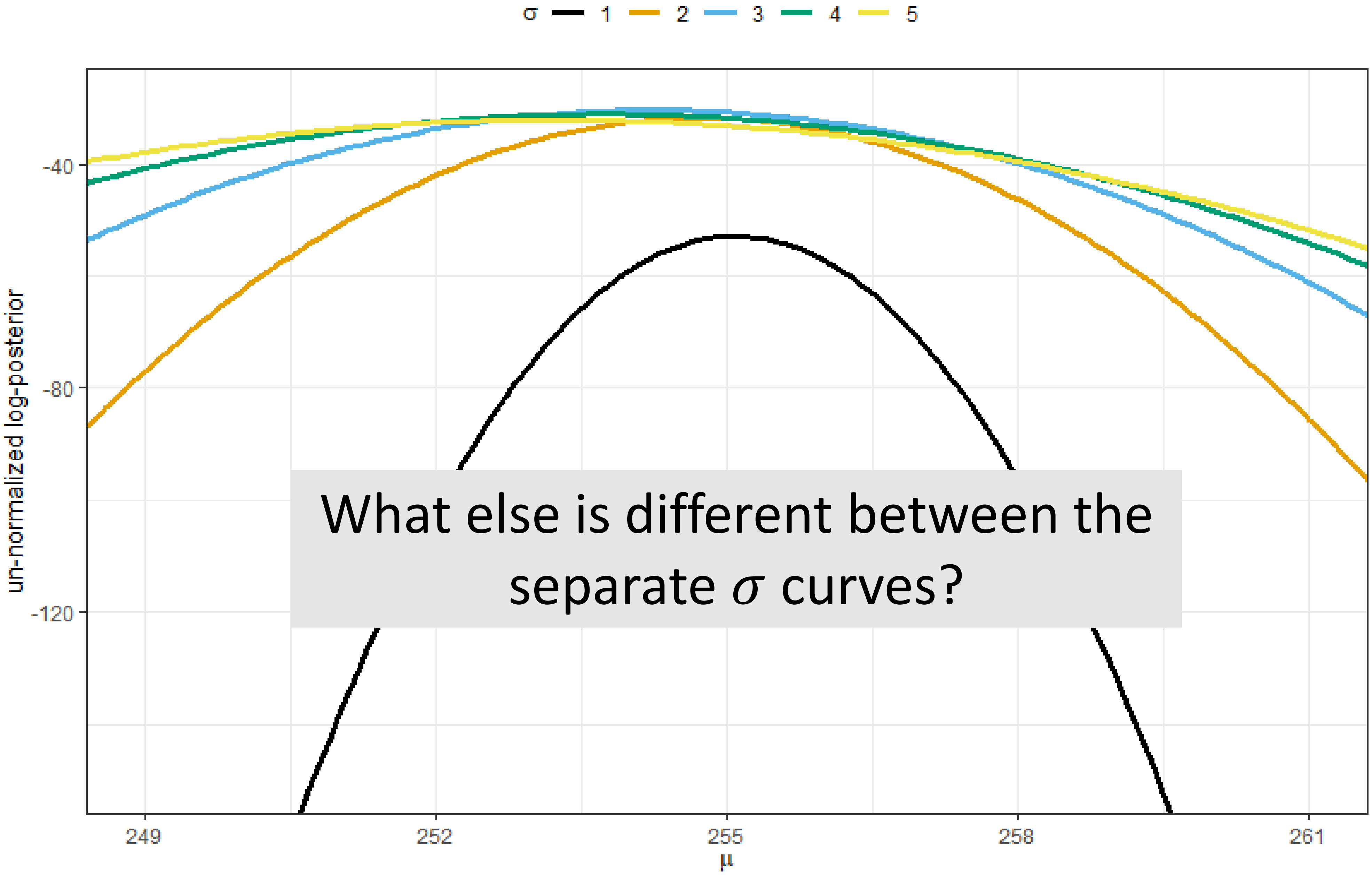
Zooming in, we see that the other curves have higher log-posterior values!



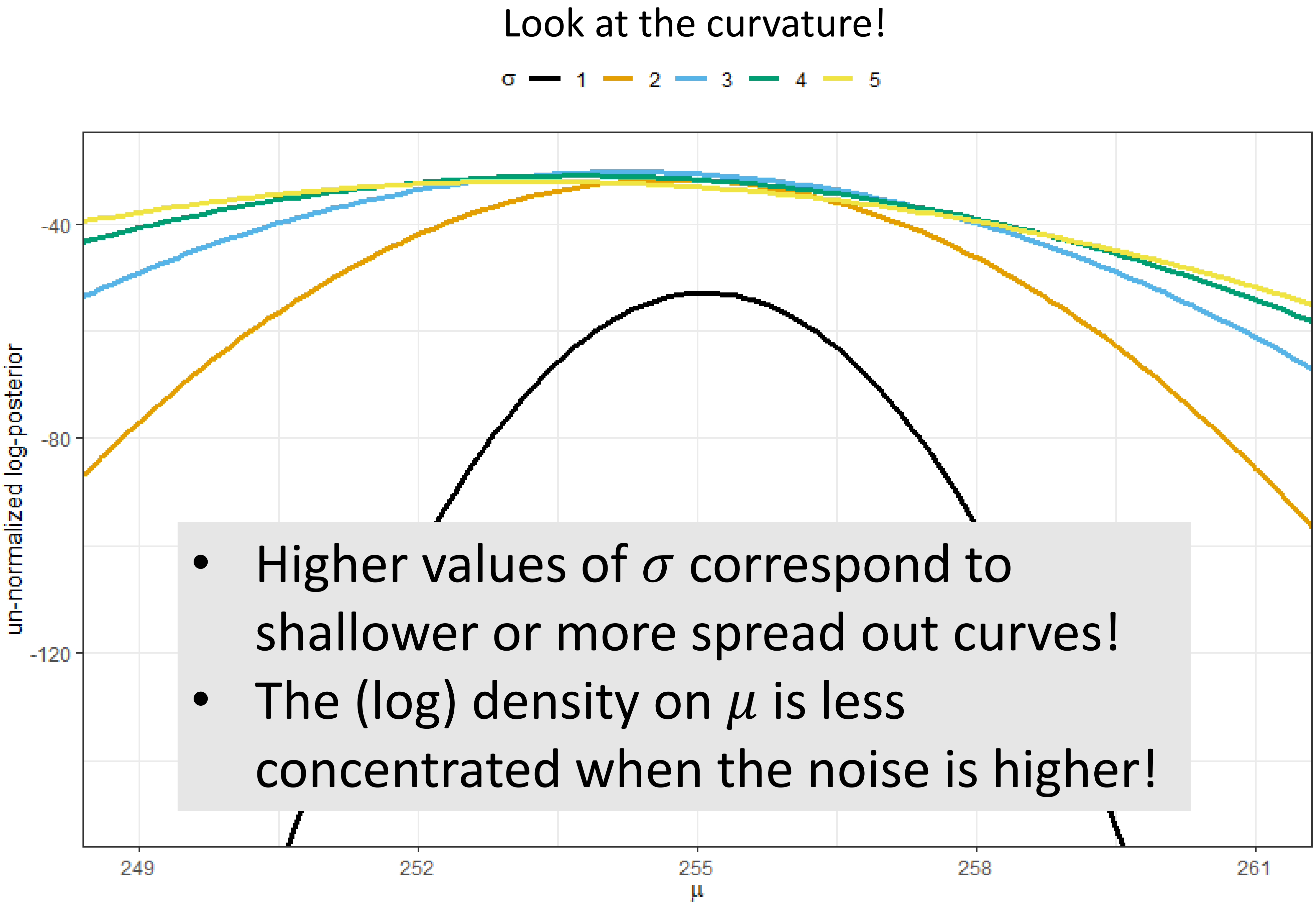
$\sigma > 1$ appears to be more plausible!



$\sigma > 1$ appears to be more plausible!

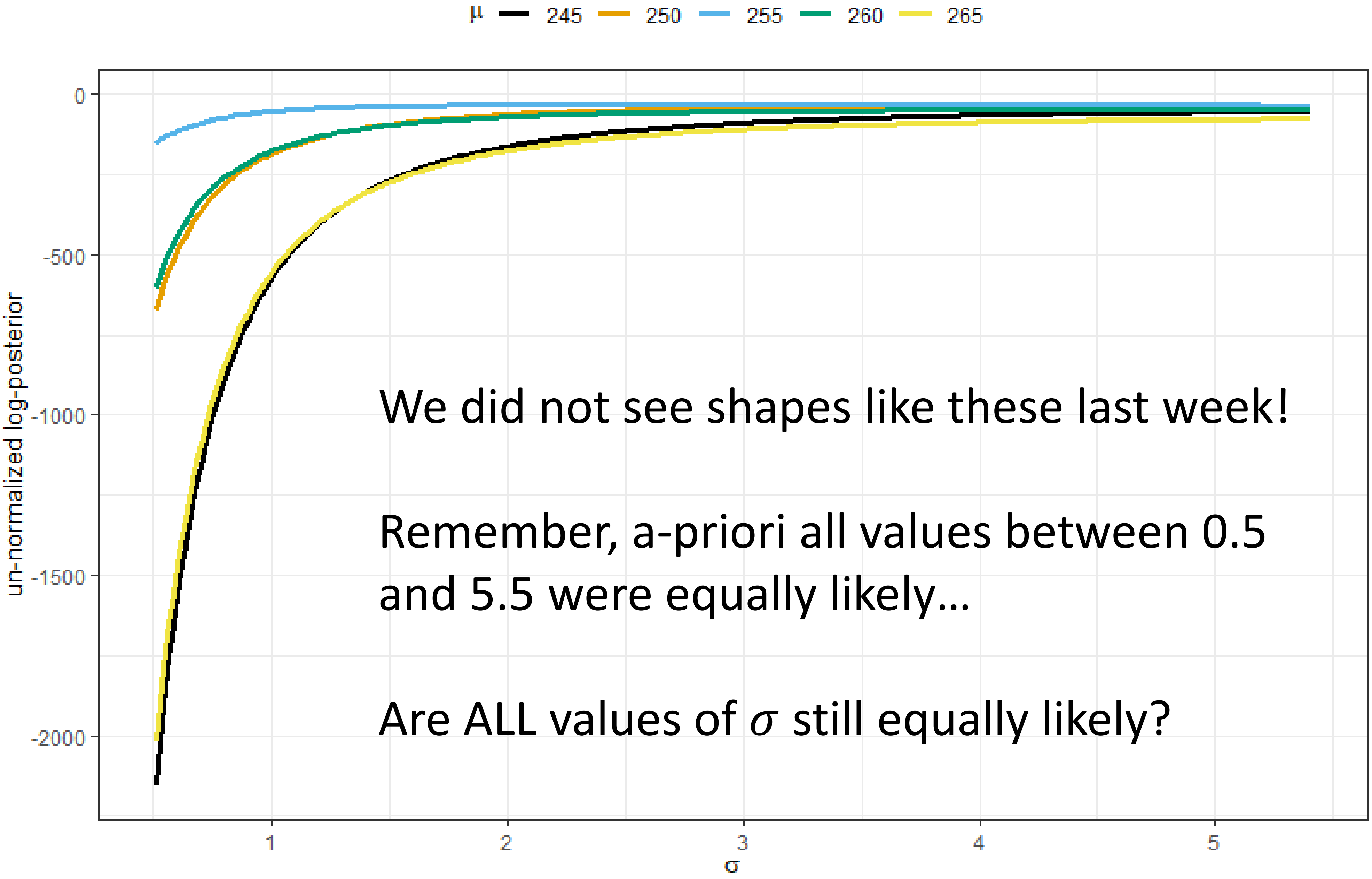


What else is different between the separate σ curves?



- Higher values of σ correspond to shallower or more spread out curves!
- The (log) density on μ is less concentrated when the noise is higher!

Now consider the log-posterior with respect σ

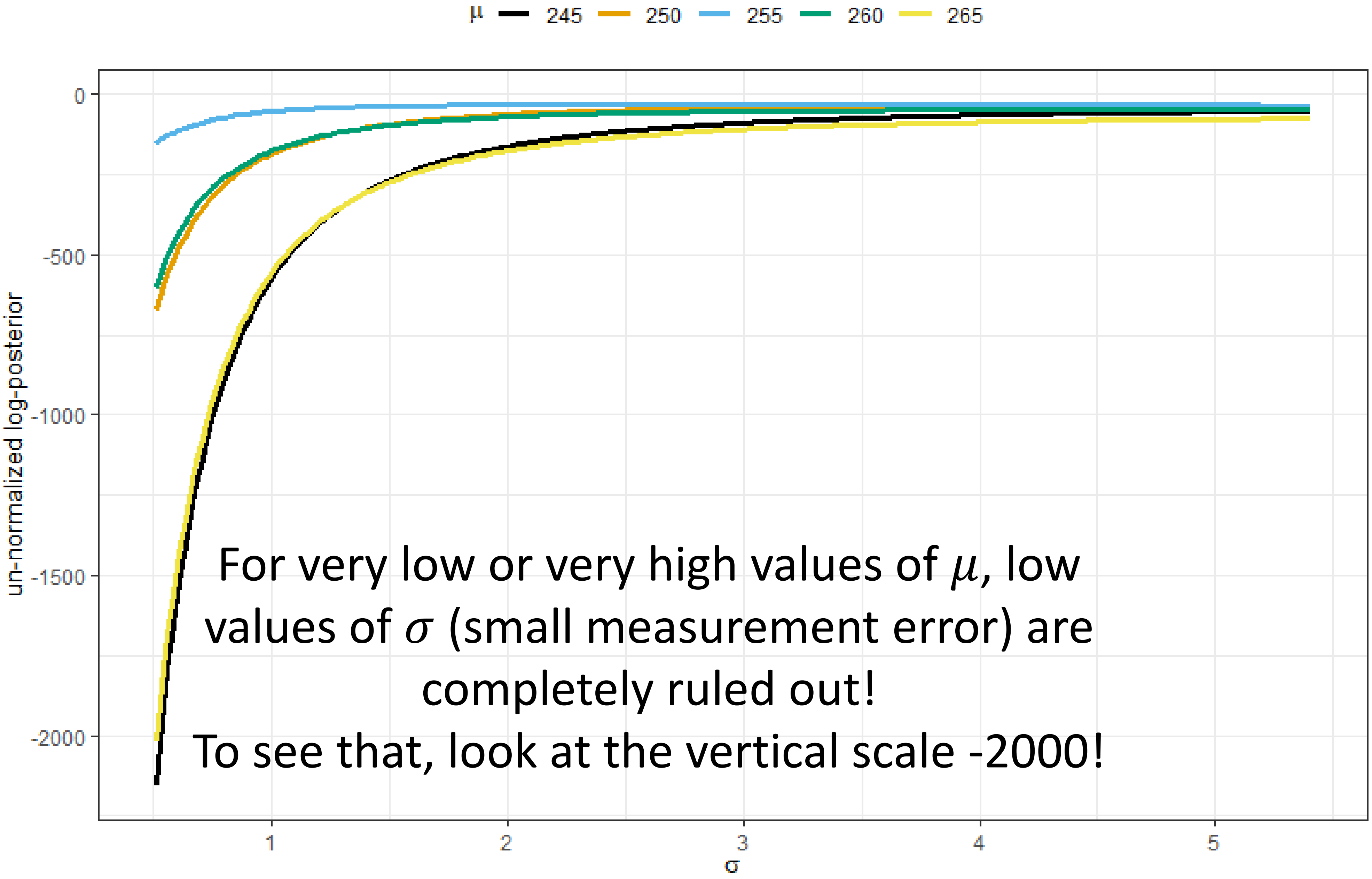


We did not see shapes like these last week!

Remember, a-priori all values between 0.5 and 5.5 were equally likely…

Are ALL values of σ still equally likely?

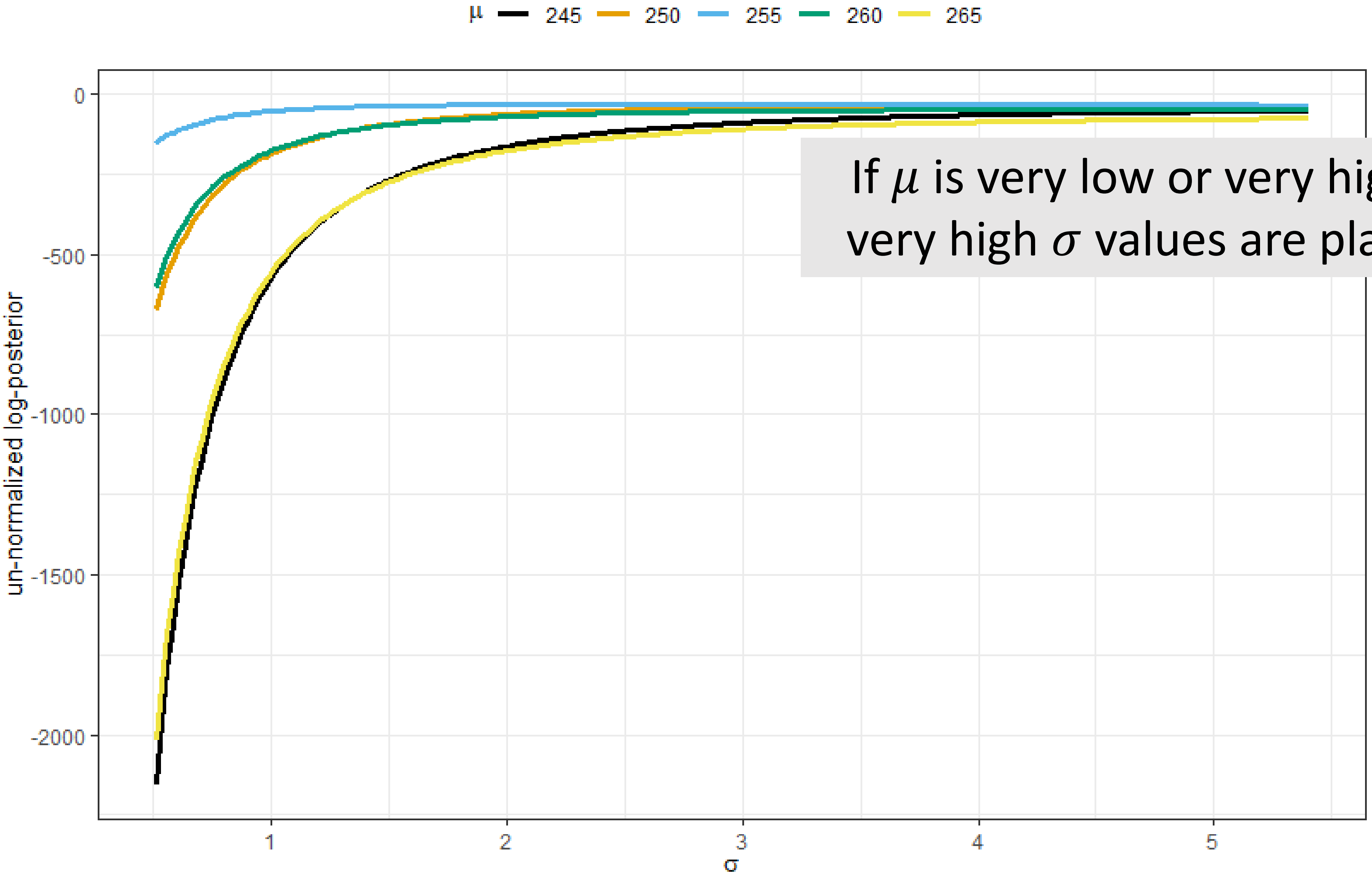
Are all values of σ equally likely? The answer depends on μ.



For very low or very high values of μ, low values of σ (small measurement error) are completely ruled out!

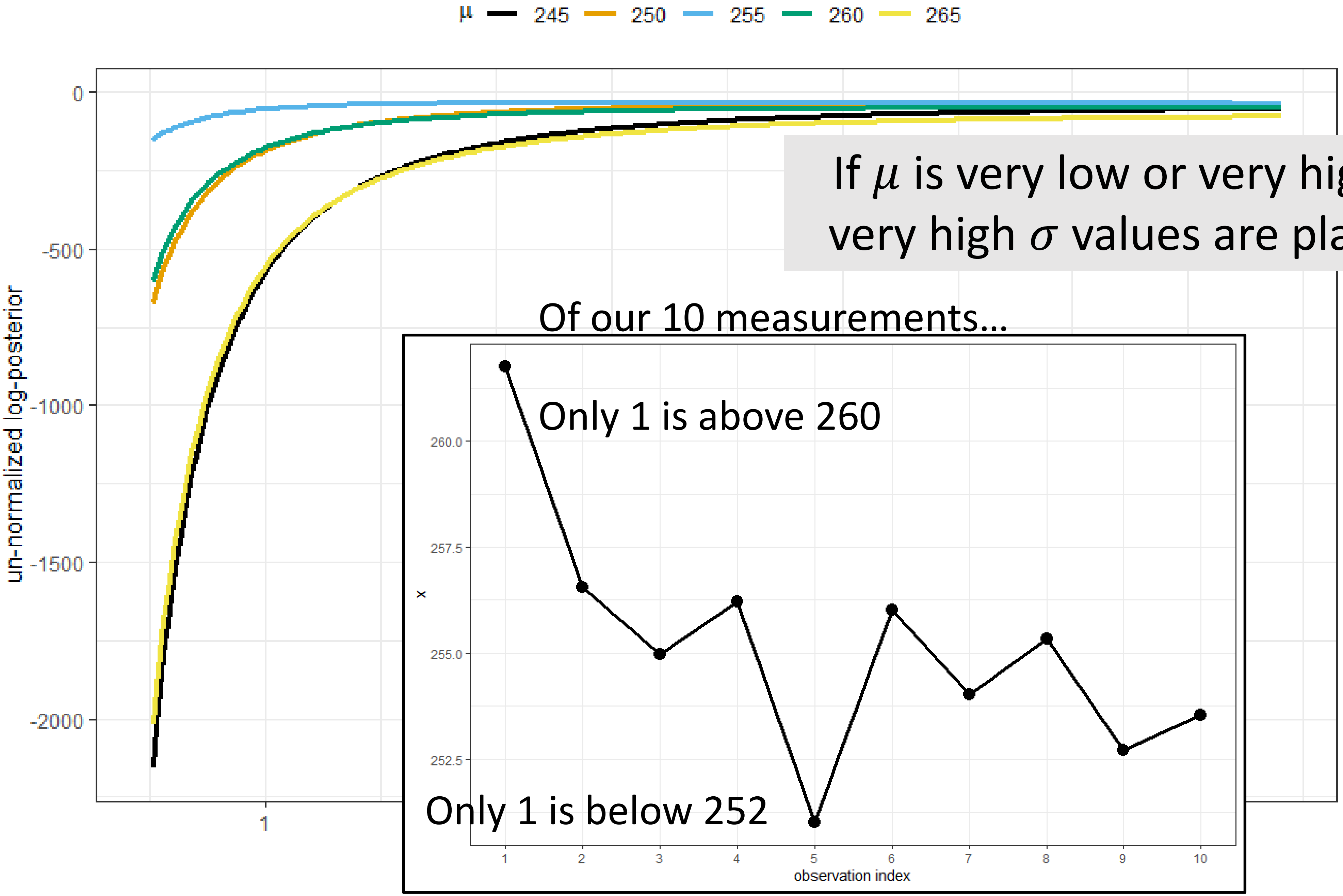
To see that, look at the vertical scale -2000!

Are all values of σ equally likely? The answer depends on μ.



If μ is very low or very high only very high σ values are plausible!

Are all values of σ equally likely? The answer depends on μ.



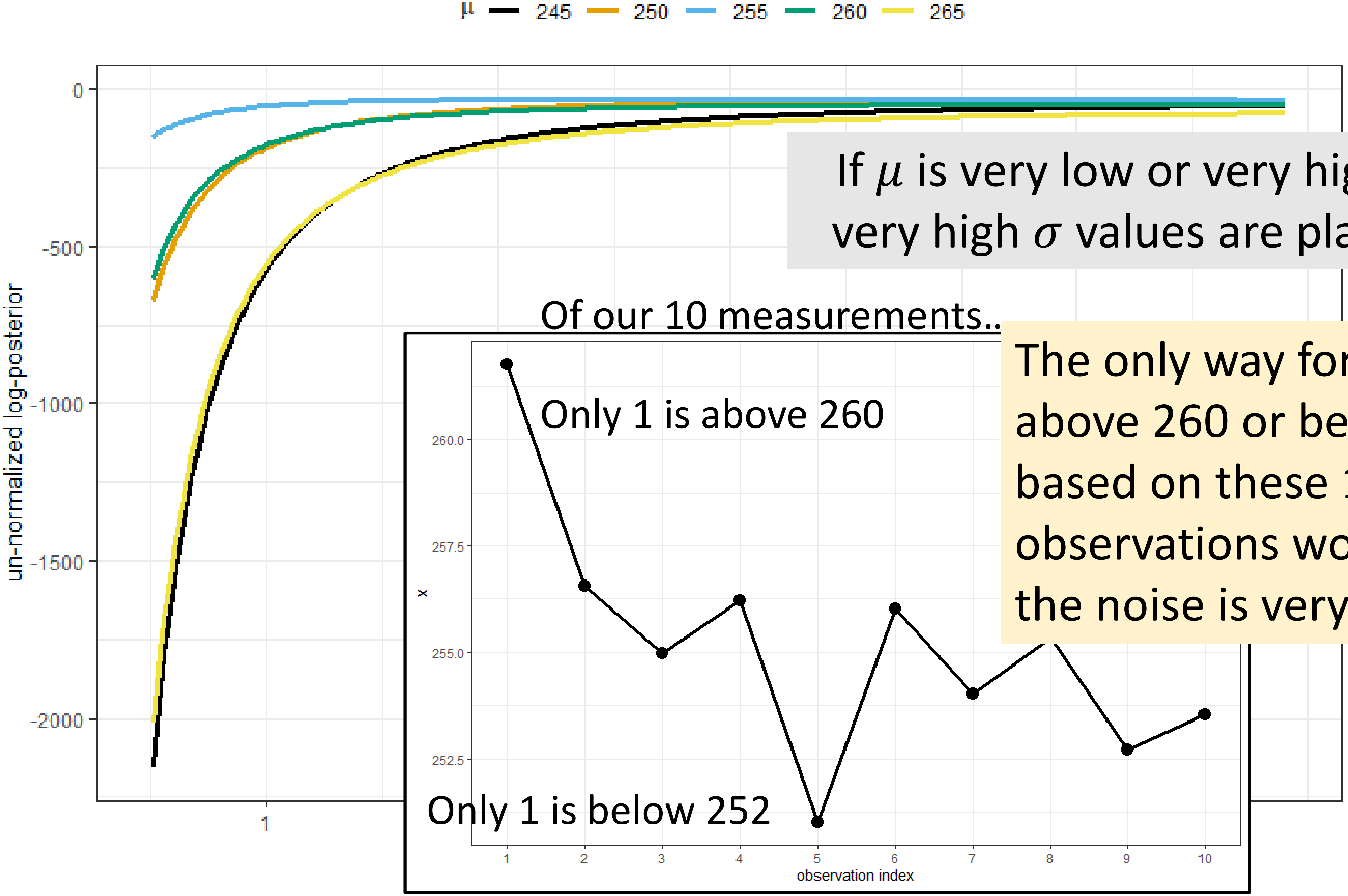
If μ is very low or very high only very high σ values are plausible!

Of our 10 measurements…

Only 1 is above 260

Only 1 is below 252

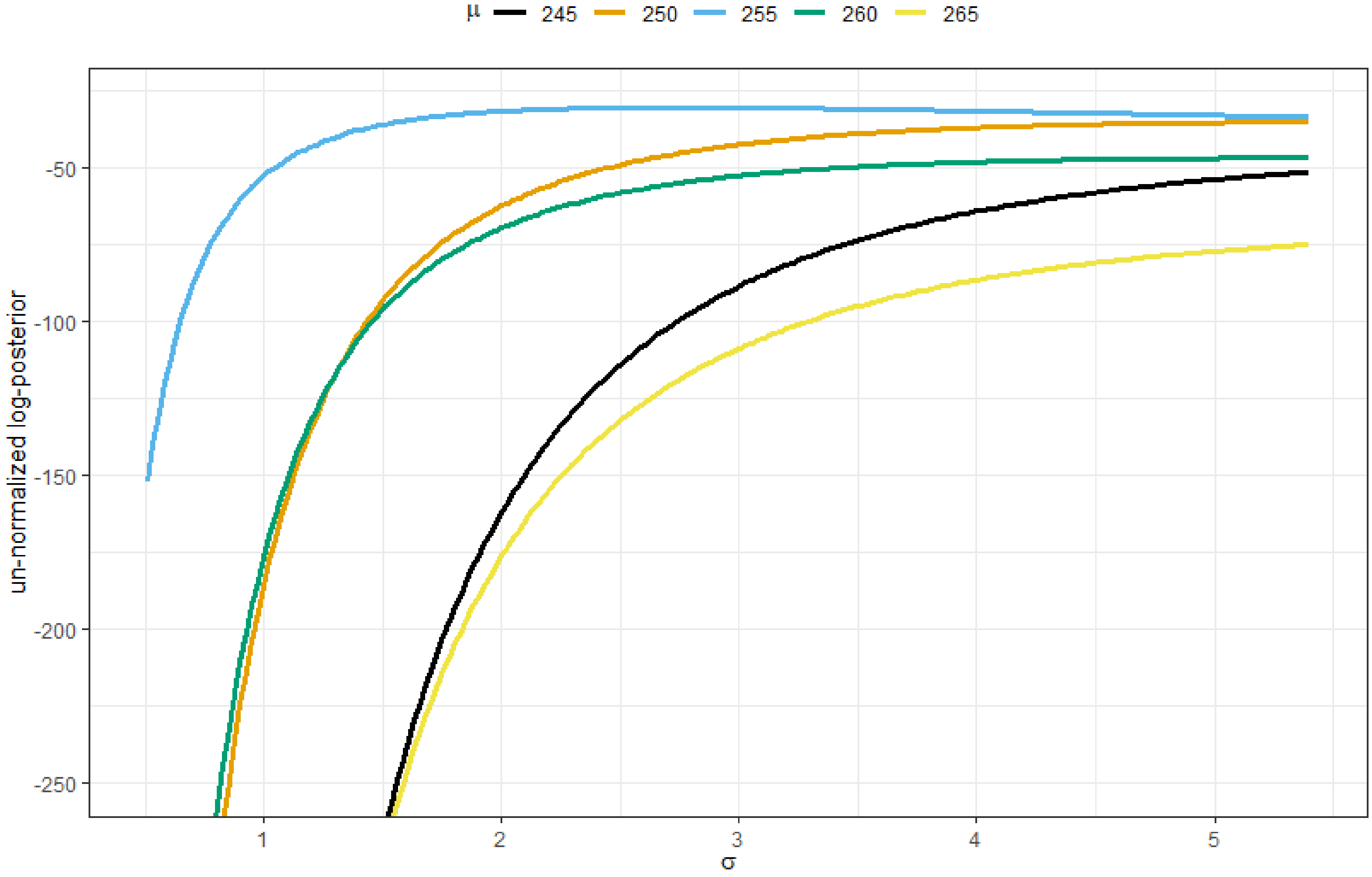
Are all values of σ equally likely? The answer depends on μ.



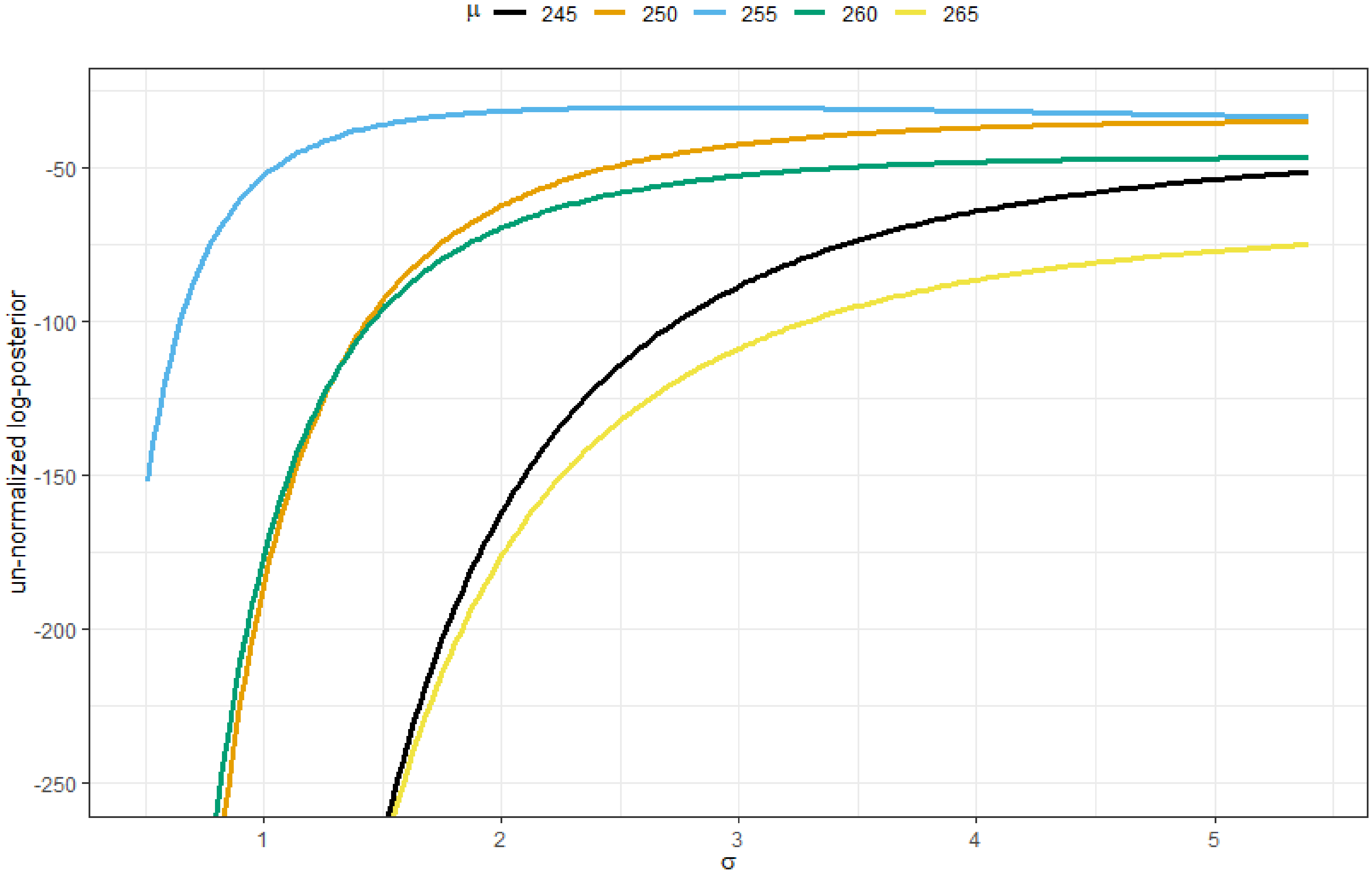
If μ is very low or very high only very high σ values are plausible!

The only way for μ to be above 260 or below 250 based on these 10 observations would be if the noise is very high!

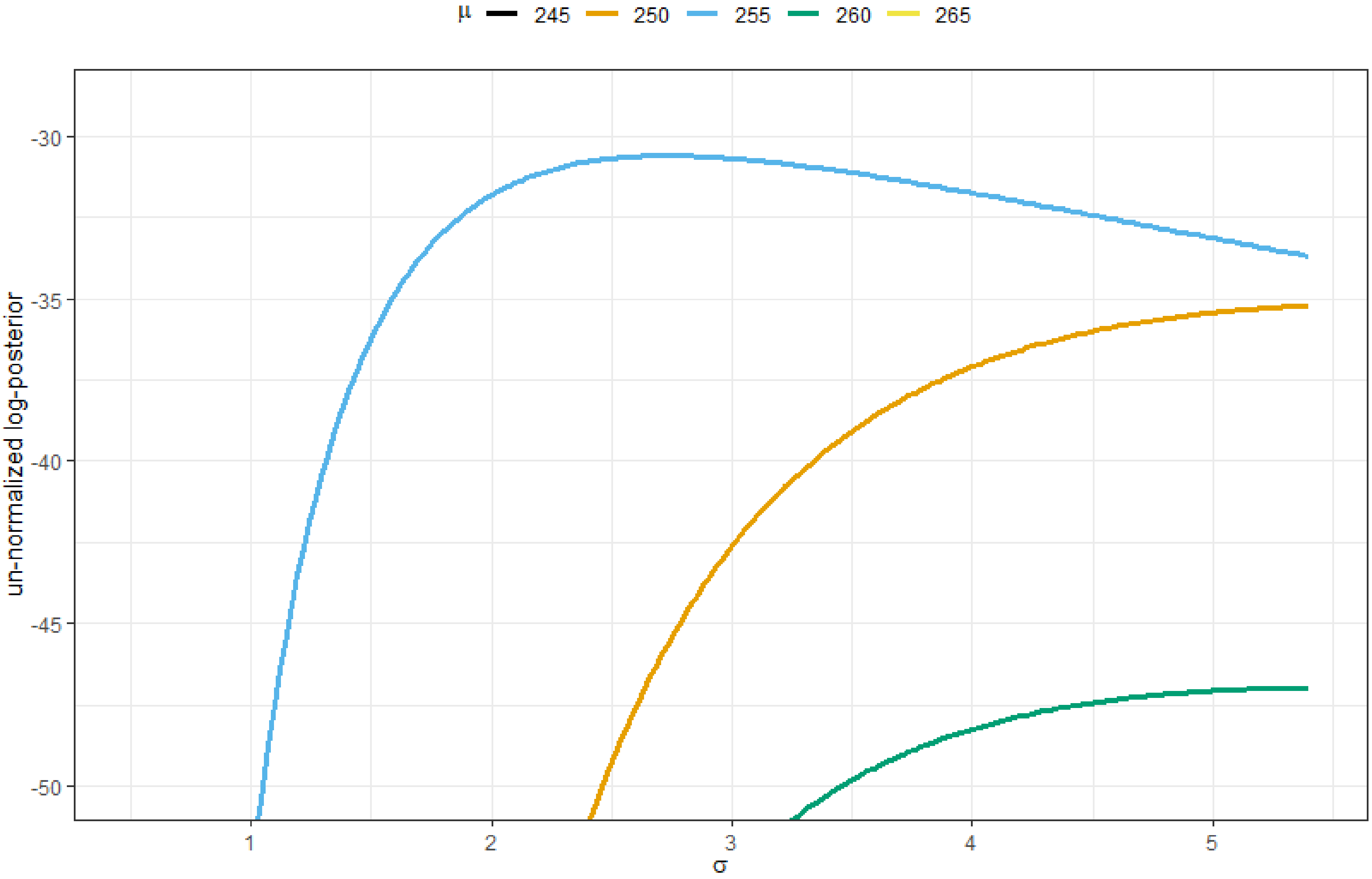
Zooming in we can see the very low or very high μ values are not as plausible as others!



σ does not seem to have as sharply defined MAP as we saw on μ…



BUT…remember the vertical scale! Zoom in further to reveal the MAP!



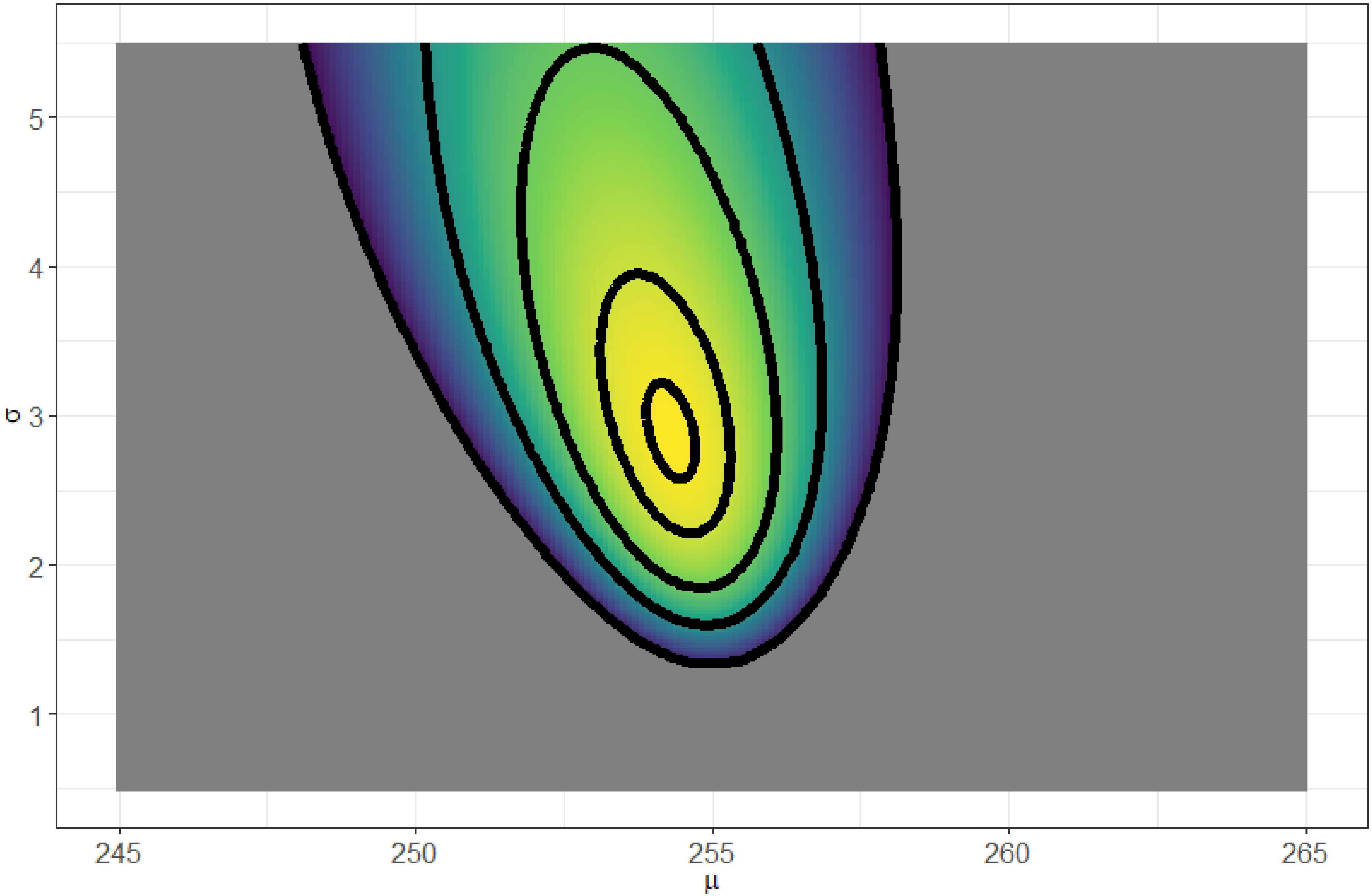
After all of that…let's now look at the surface

- The log-posterior surface with respect to μ and σ will be displayed similar to how the log-prior surface was displayed.

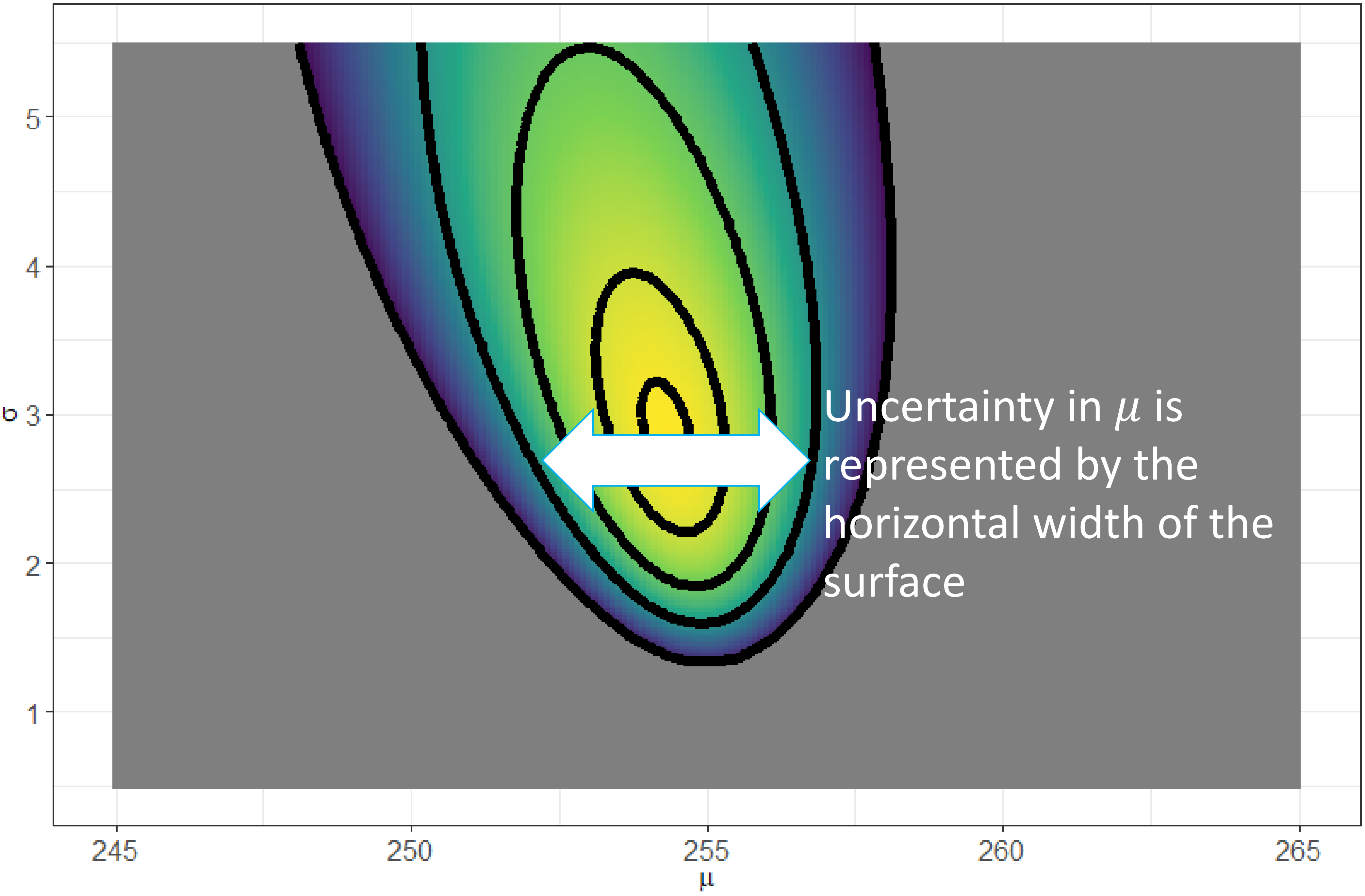
- Fill represents the value un-normalized log-posterior, bright yellow are higher values and dark blue are lower values.

- Combinations that are very implausible are greyed out.

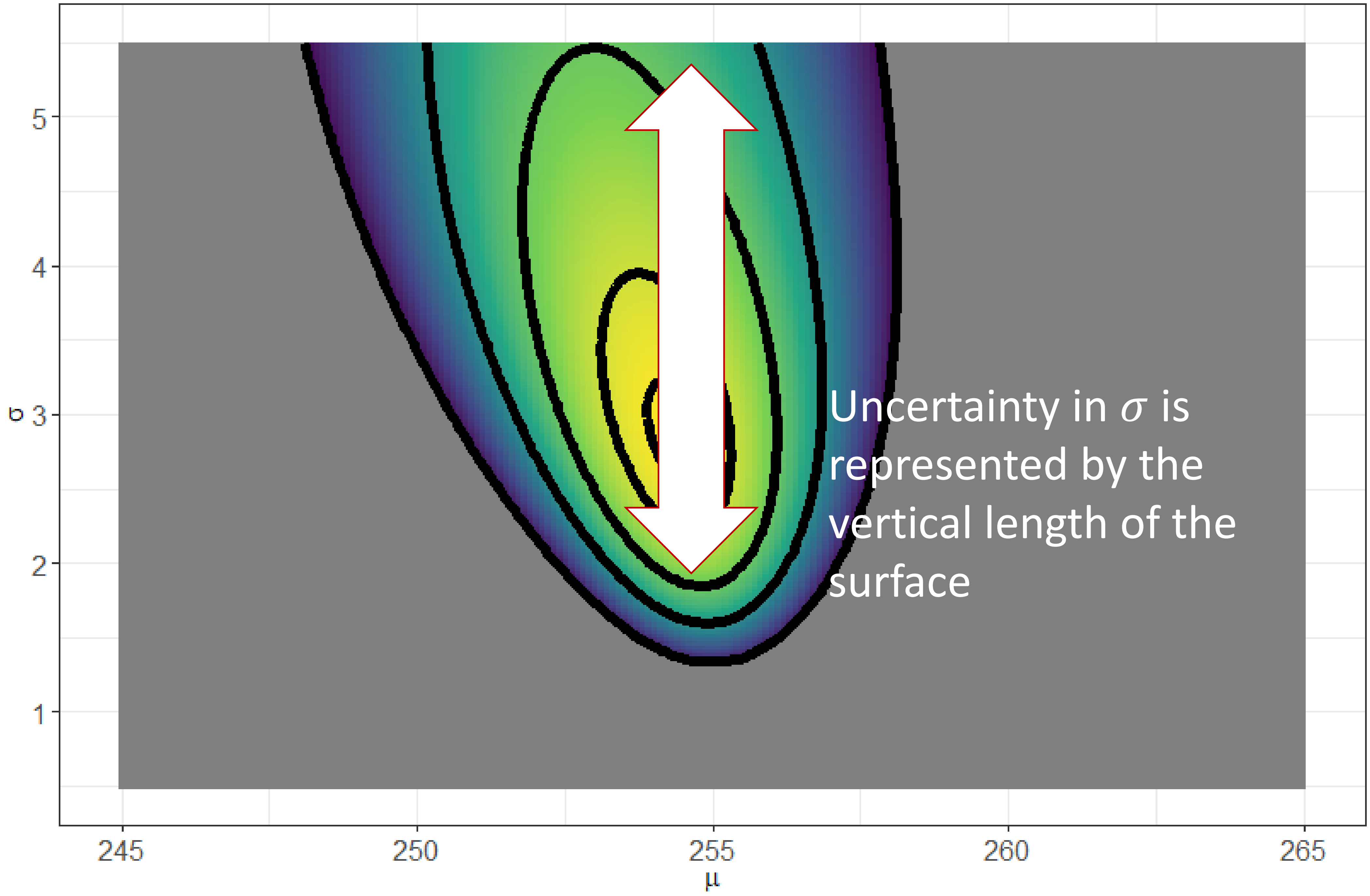
- Contour lines represent the top 10%, 50%, 90%, 99%, and 99.99% of combinations.

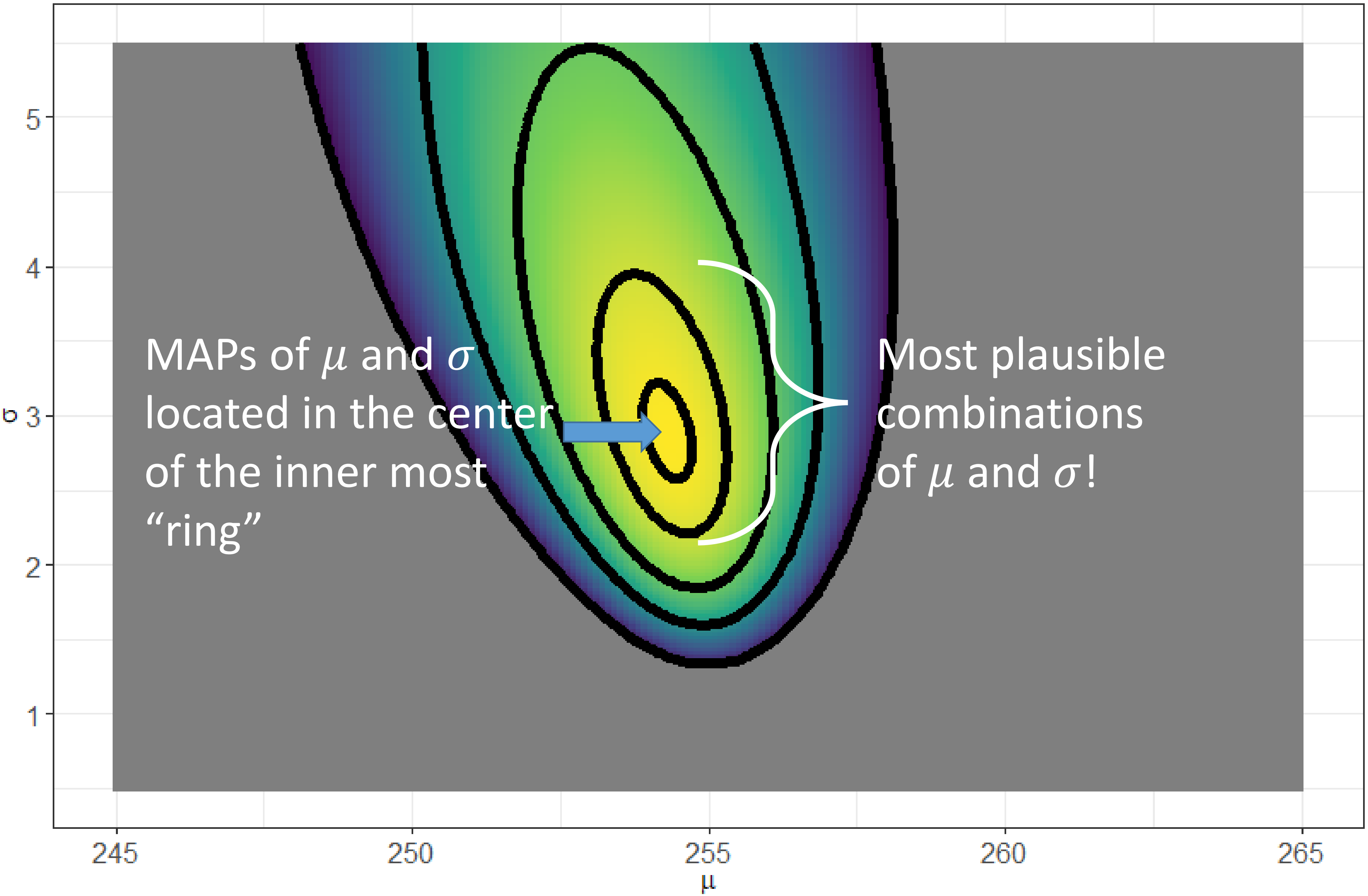


5
4
σ 3
2
1
245
250
255
260
265
μ

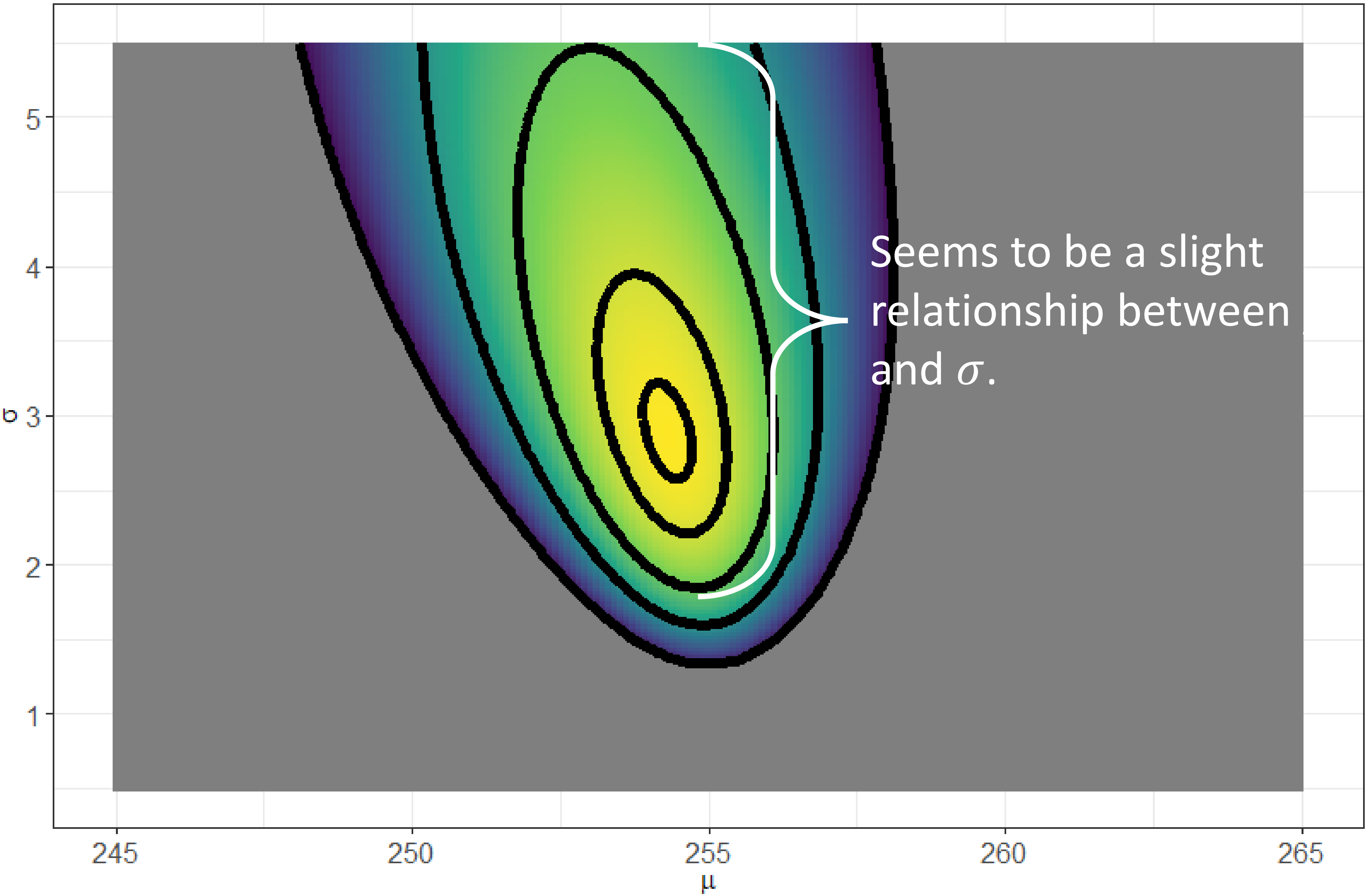


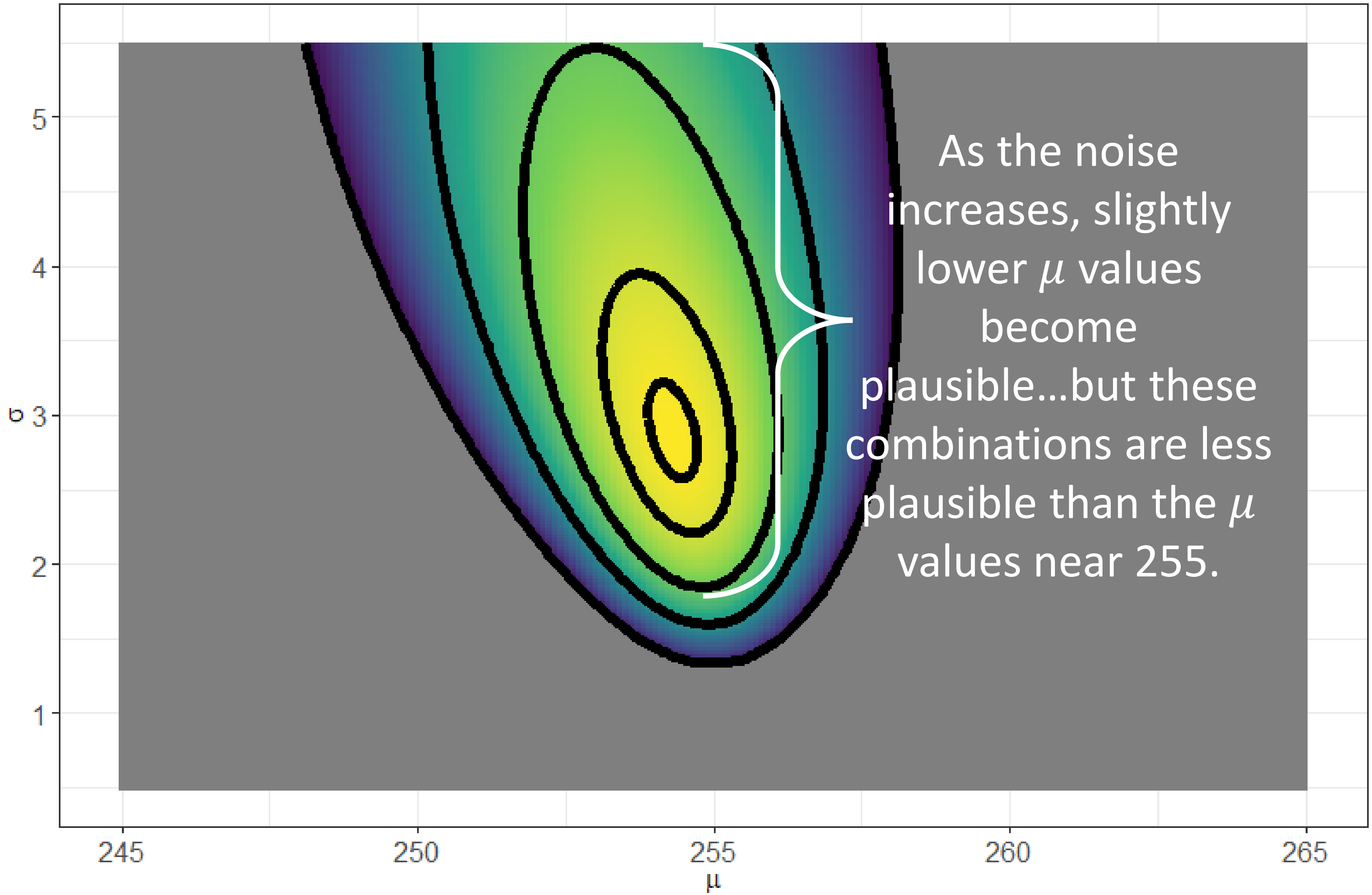
Uncertainty in μ is represented by the horizontal width of the surface
5
4
3
2
1
σ
245
250
255
260
265
μ


Uncertainty in σ is represented by the vertical length of the surface
σ
μ
5
4
3
2
1
245
250
255
260
265



MAPs of μ and σ located in the center of the inner most "ring"
Most plausible combinations of μ and σ!
σ
5
4
3
2
1
245
250
255
260
265
μ

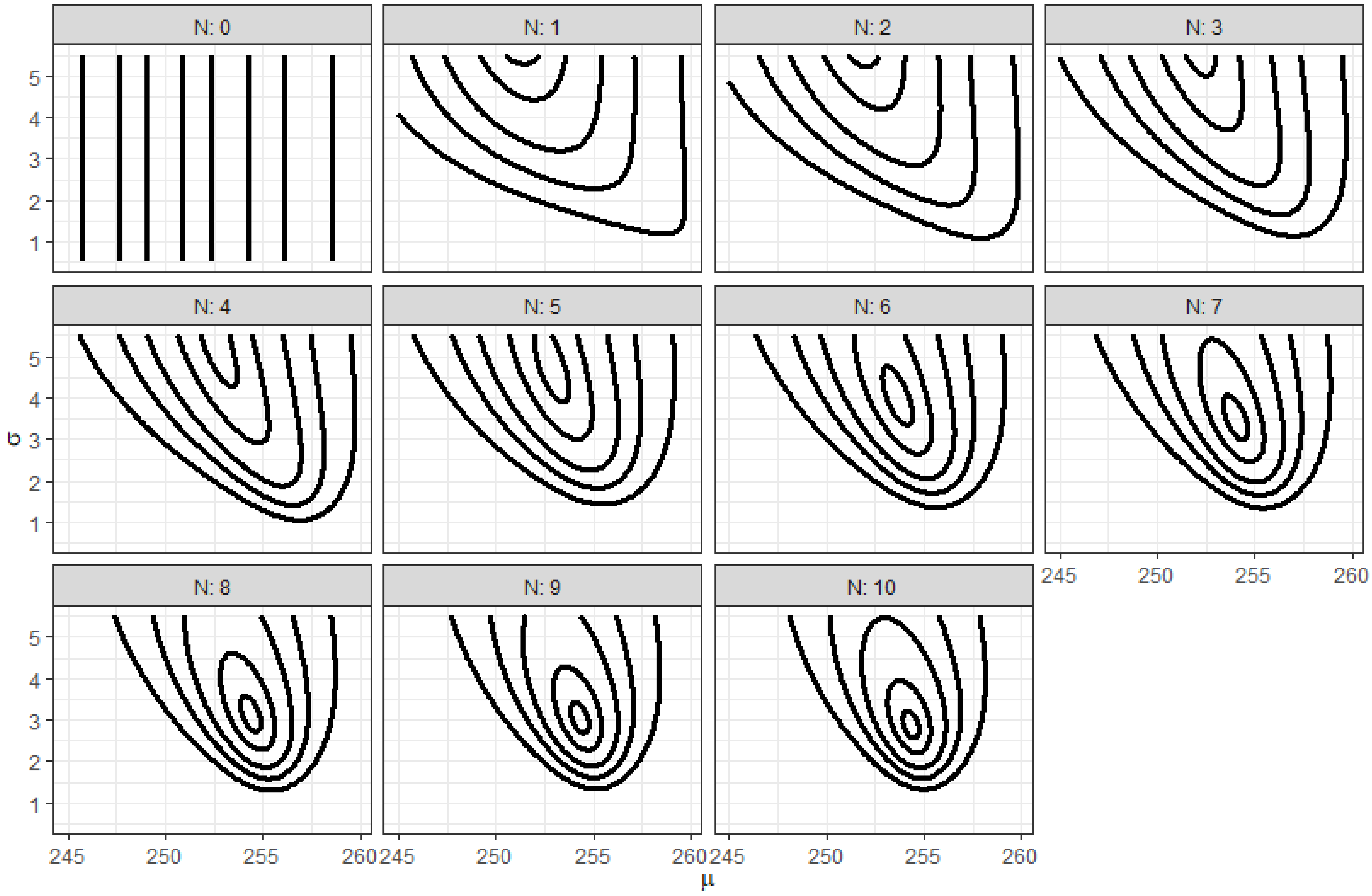

Seems to be a slight relationship between
and σ.
5
4
σ 3
2
1
245
250
255
260
265
μ



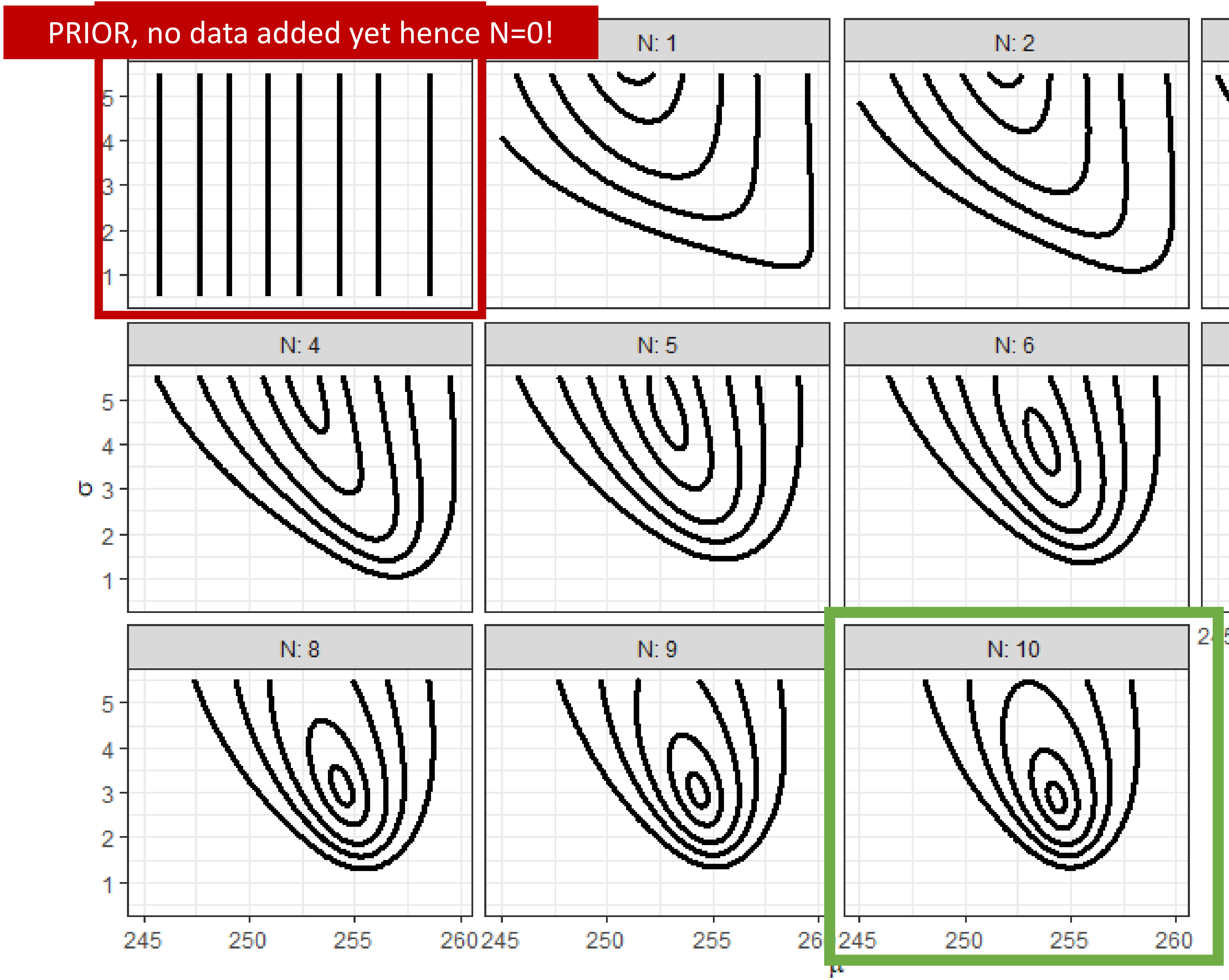
As the noise increases, slightly lower μ values become plausible…but these combinations are less plausible than the μ values near 255.
5
4
3
2
1
σ
245
250
255
260
265
μ

We just visualized the joint posterior distribution of 2 unknown parameters!

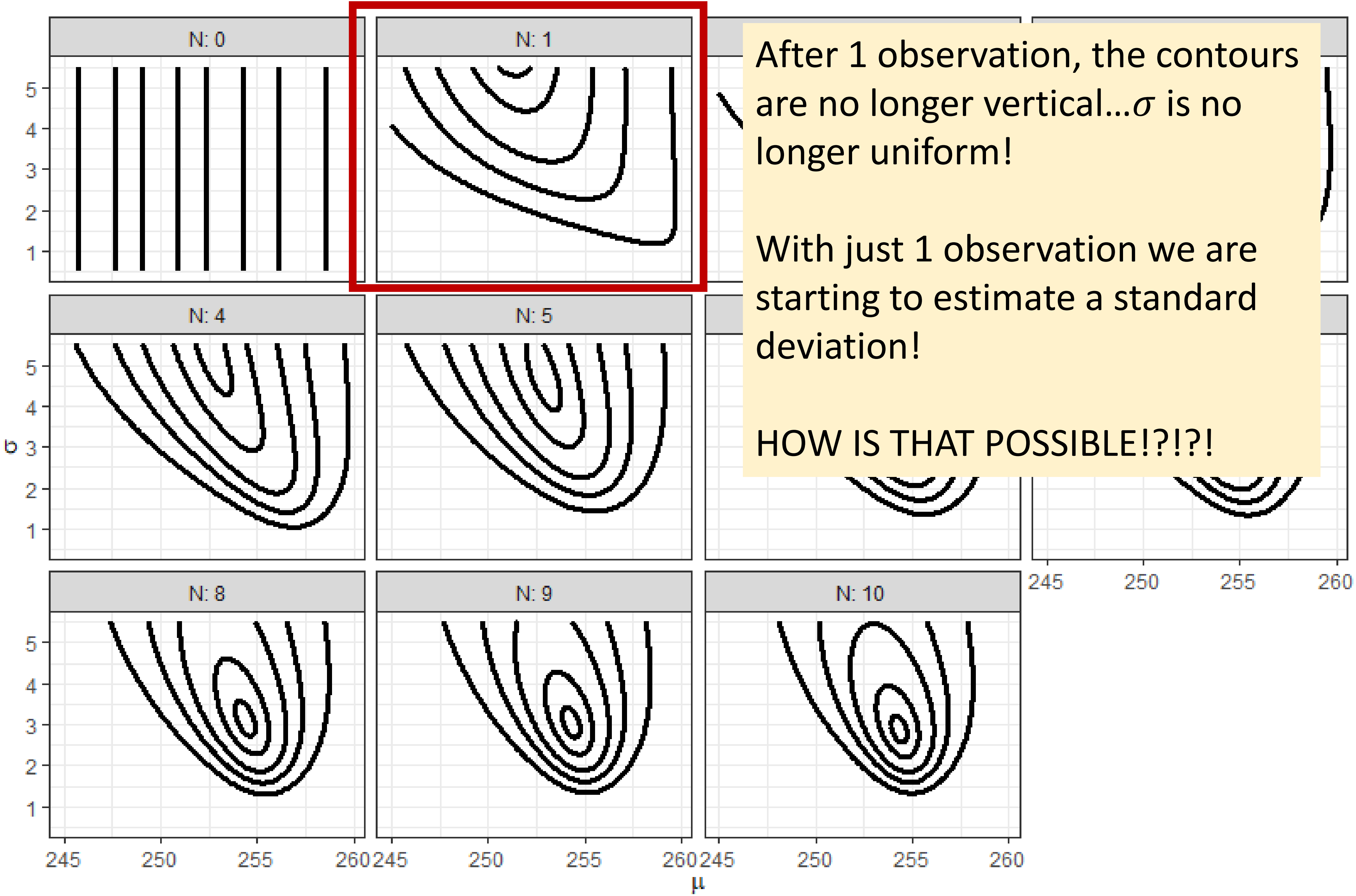
- It was based on our prior beliefs and 10 observations.
- How did the prior morph or evolve into the posterior we just visualized?
- Let’s now step through how the log-posterior surface changes as we sequentially add each data point.
- The following figures show the contours, but do not fill in the color.
- The contours represent the same iso-probability contours as the previous figure.



N: 0
N: 1
N: 2
N: 3
N: 4
N: 5
N: 6
N: 7
N: 8
N: 9
N: 10
σ
μ
1
2
3
4
5
245
250
255
260



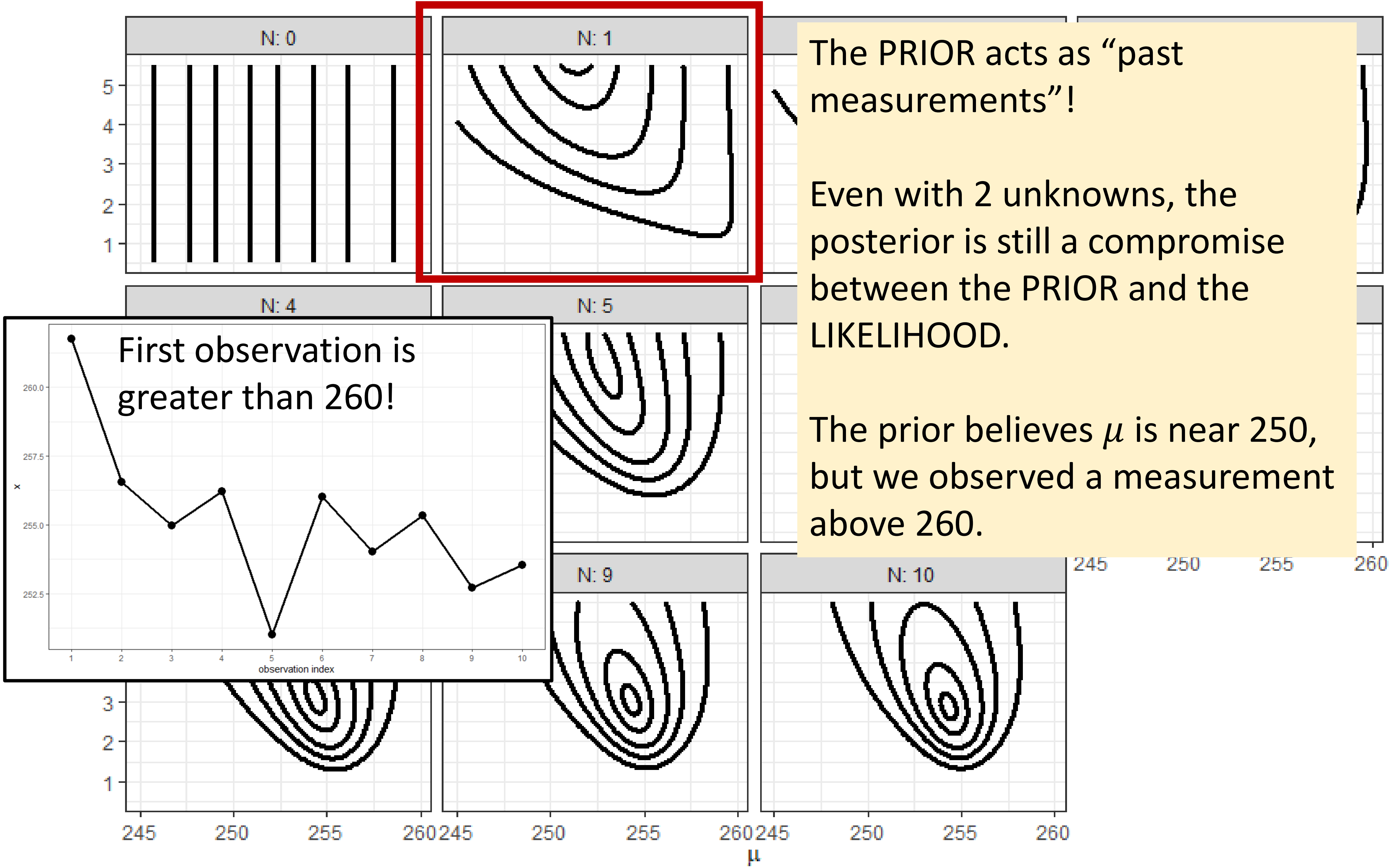
The surface we described based on 10 observations just without the fill color.



After 1 observation, the contours are no longer vertical…σ is no longer uniform!

With just 1 observation we are starting to estimate a standard deviation!

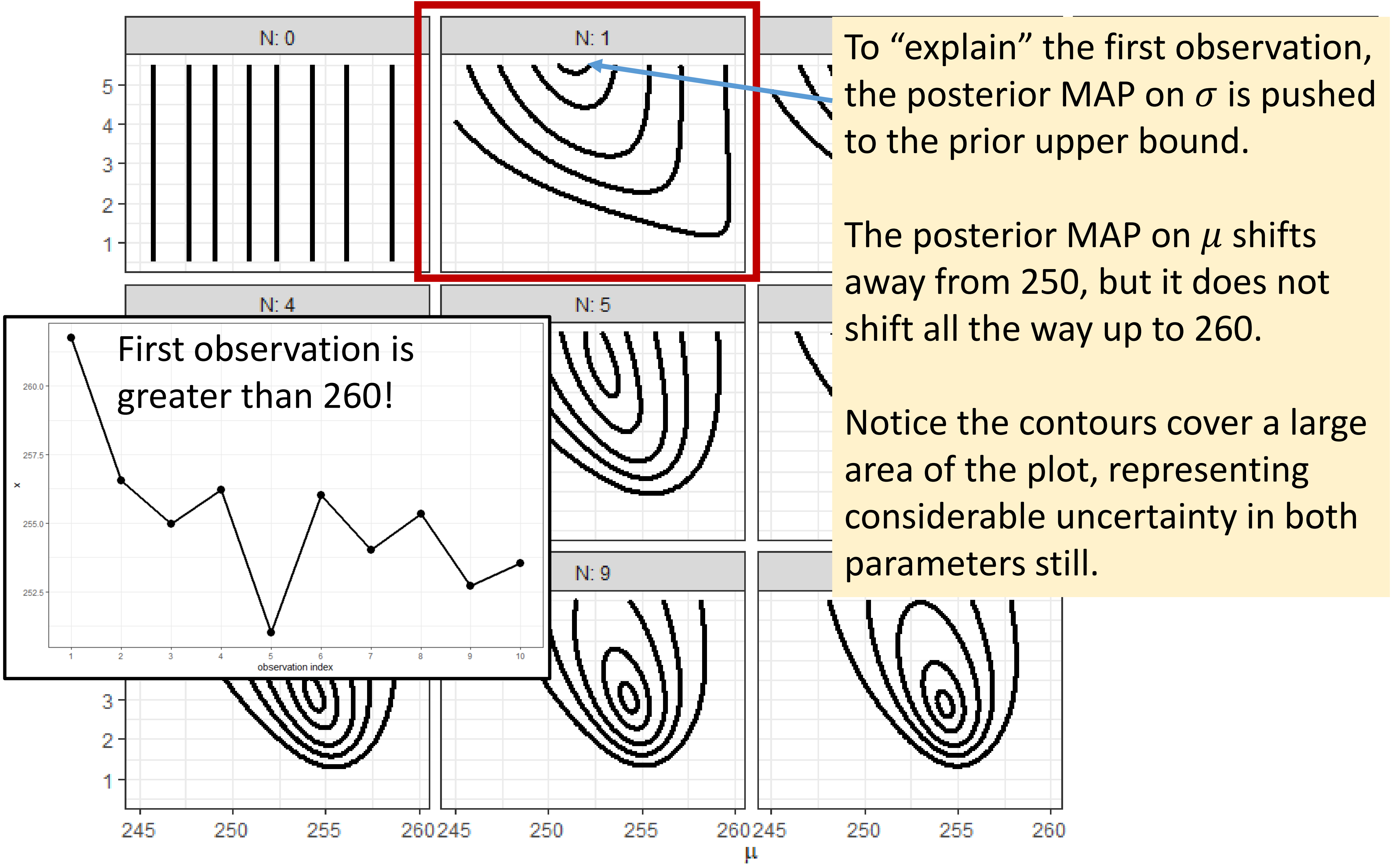
HOW IS THAT POSSIBLE!?!?!



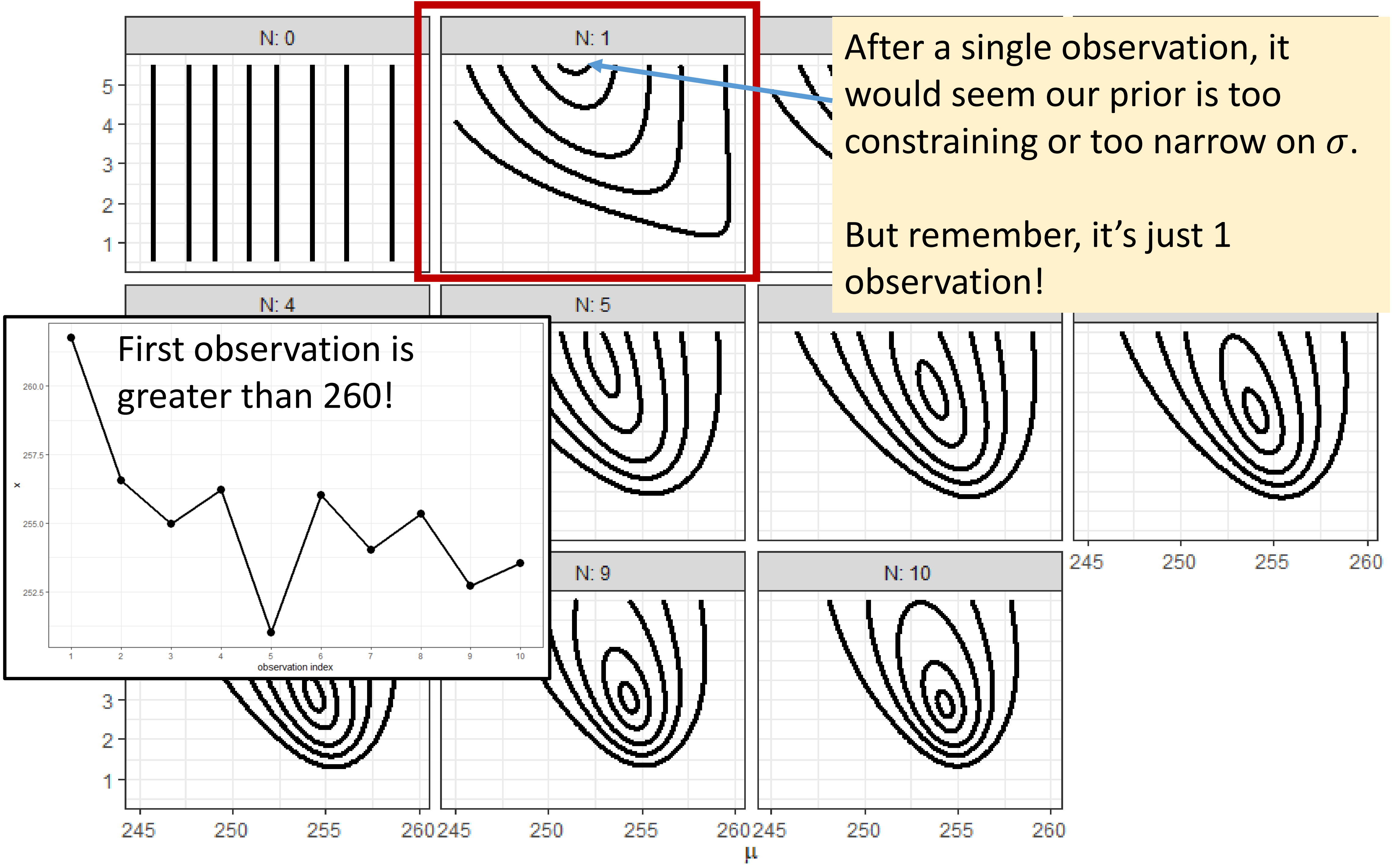
The PRIOR acts as "past measurements"!

Even with 2 unknowns, the posterior is still a compromise between the PRIOR and the LIKELIHOOD.

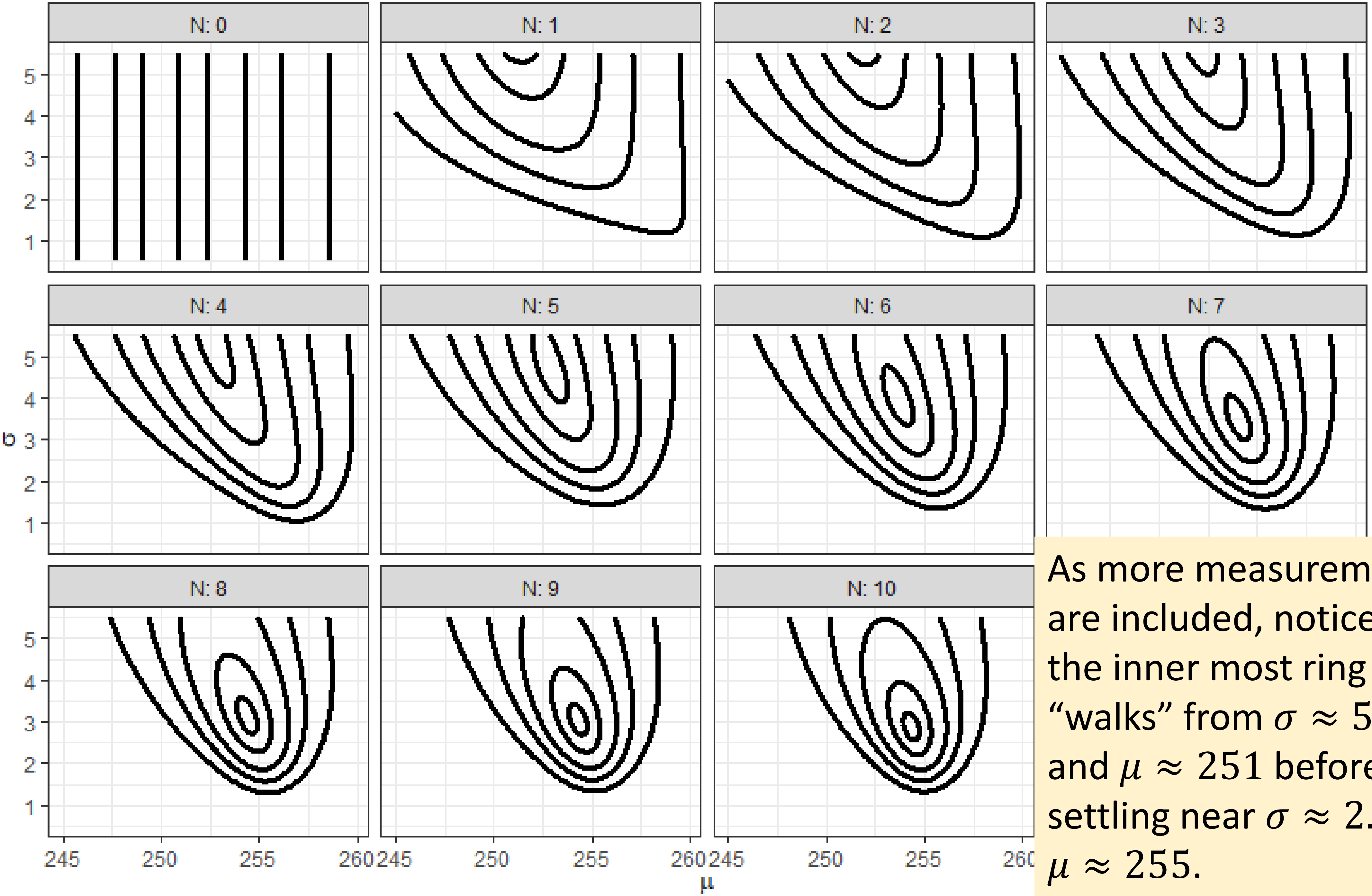
The prior believes μ is near 250, but we observed a measurement above 260.



N: 0
N: 1
N: 4
N: 5
N: 9
5
4
3
2
1
245
250
255
260
μ
First observation is greater than 260!
260.0
257.5
255.0
252.5
x
1
2
3
4
5
6
7
8
9
10
observation index
To "explain" the first observation, the posterior MAP on σ is pushed to the prior upper bound.
The posterior MAP on μ shifts away from 250, but it does not shift all the way up to 260.
Notice the contours cover a large area of the plot, representing considerable uncertainty in both parameters still.

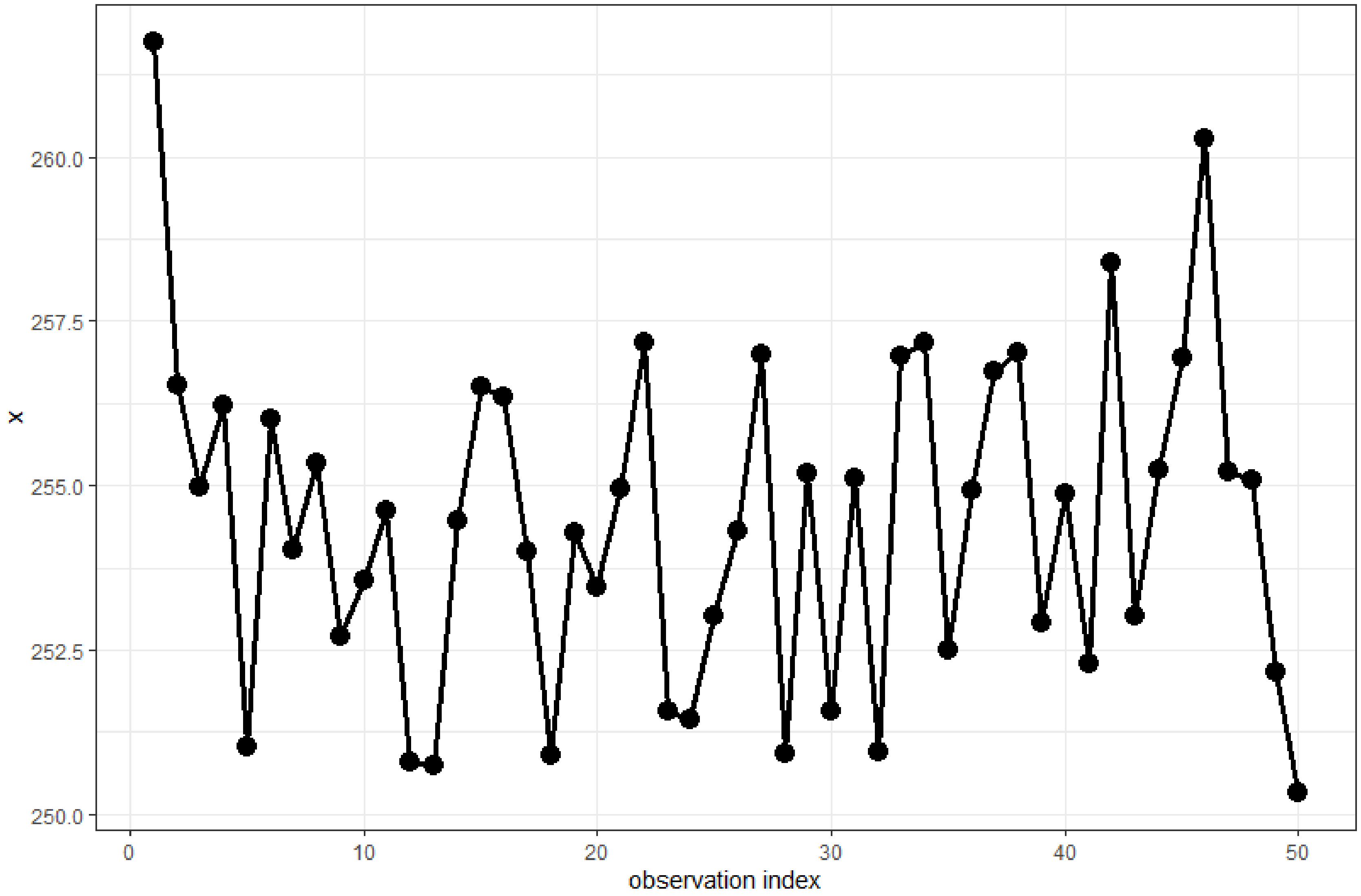


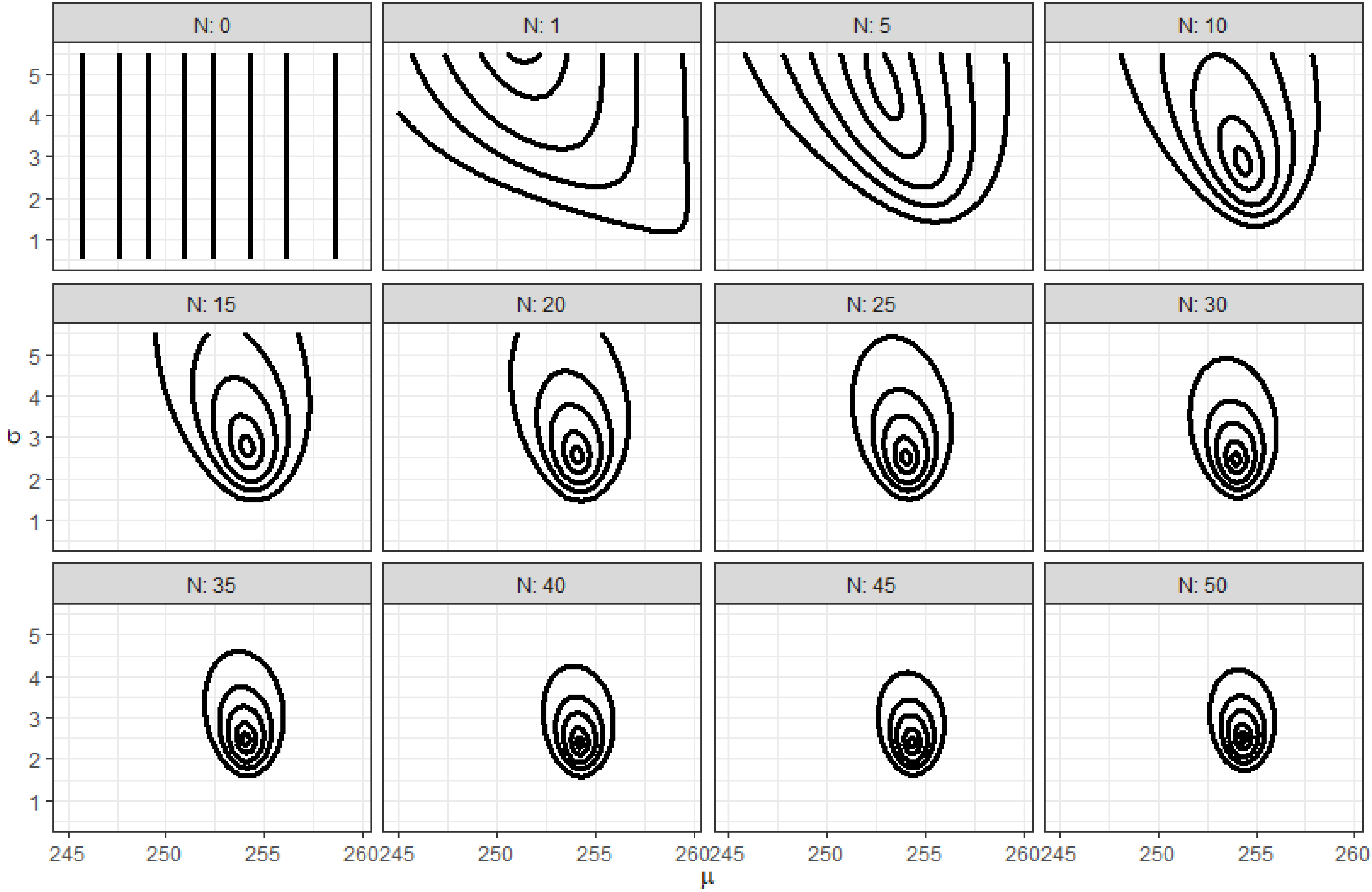
After a single observation, it would seem our prior is too constraining or too narrow on σ.
But remember, it’s just 1 observation!
First observation is greater than 260!
N: 0
N: 1
N: 4
N: 5
N: 9
N: 10
1
2
3
4
5
245
250
255
260
μ
260.0
257.5
255.0
252.5
x
1
2
3
4
5
6
7
8
9
10
observation index



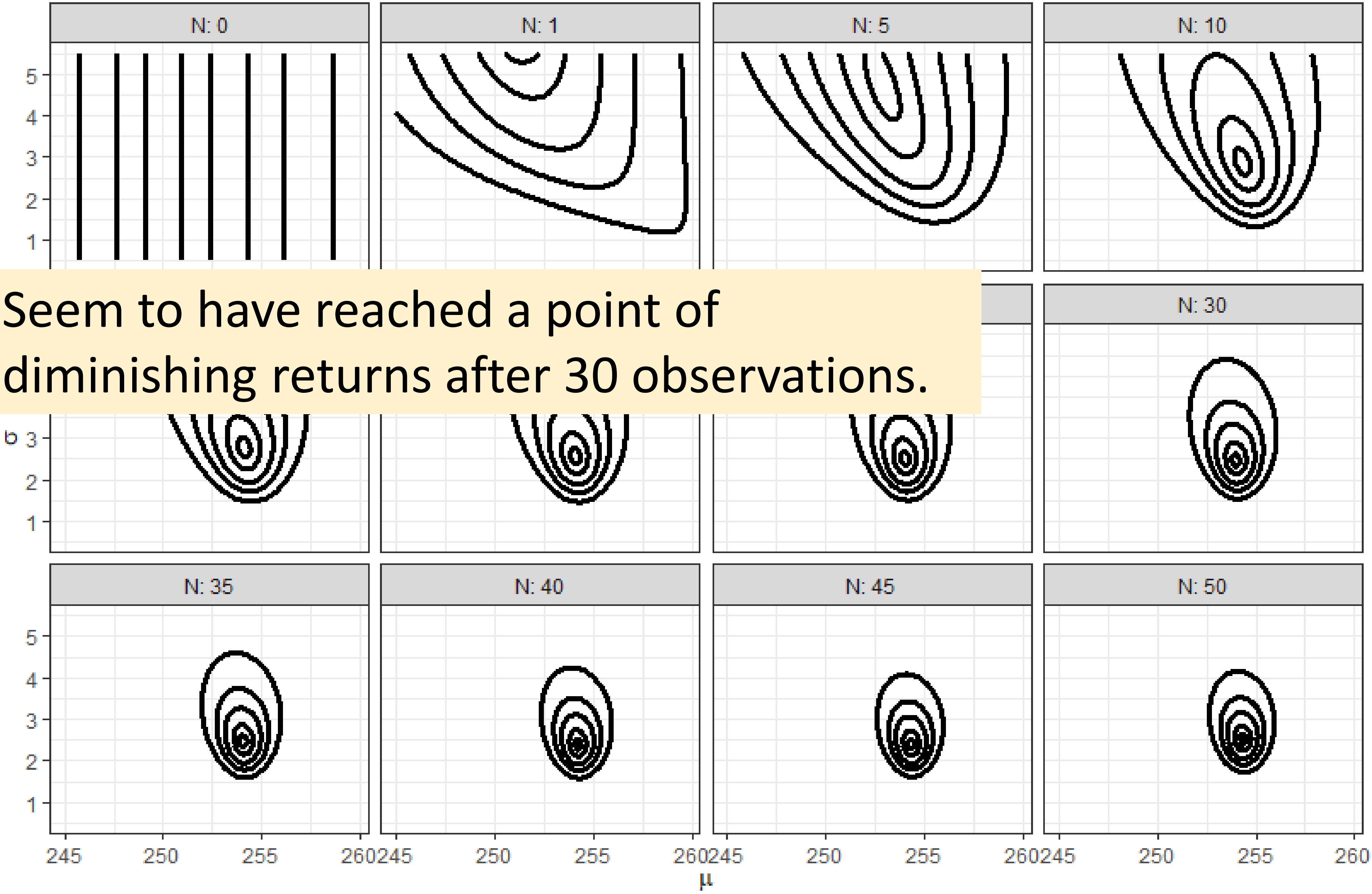
As more measurements are included, notice how the inner most ring "walks" from $\sigma \approx 5.5$ and $\mu \approx 251$ before settling near $\sigma \approx 2.5$ and $\mu \approx 255$.

Now consider up to 50 measurements



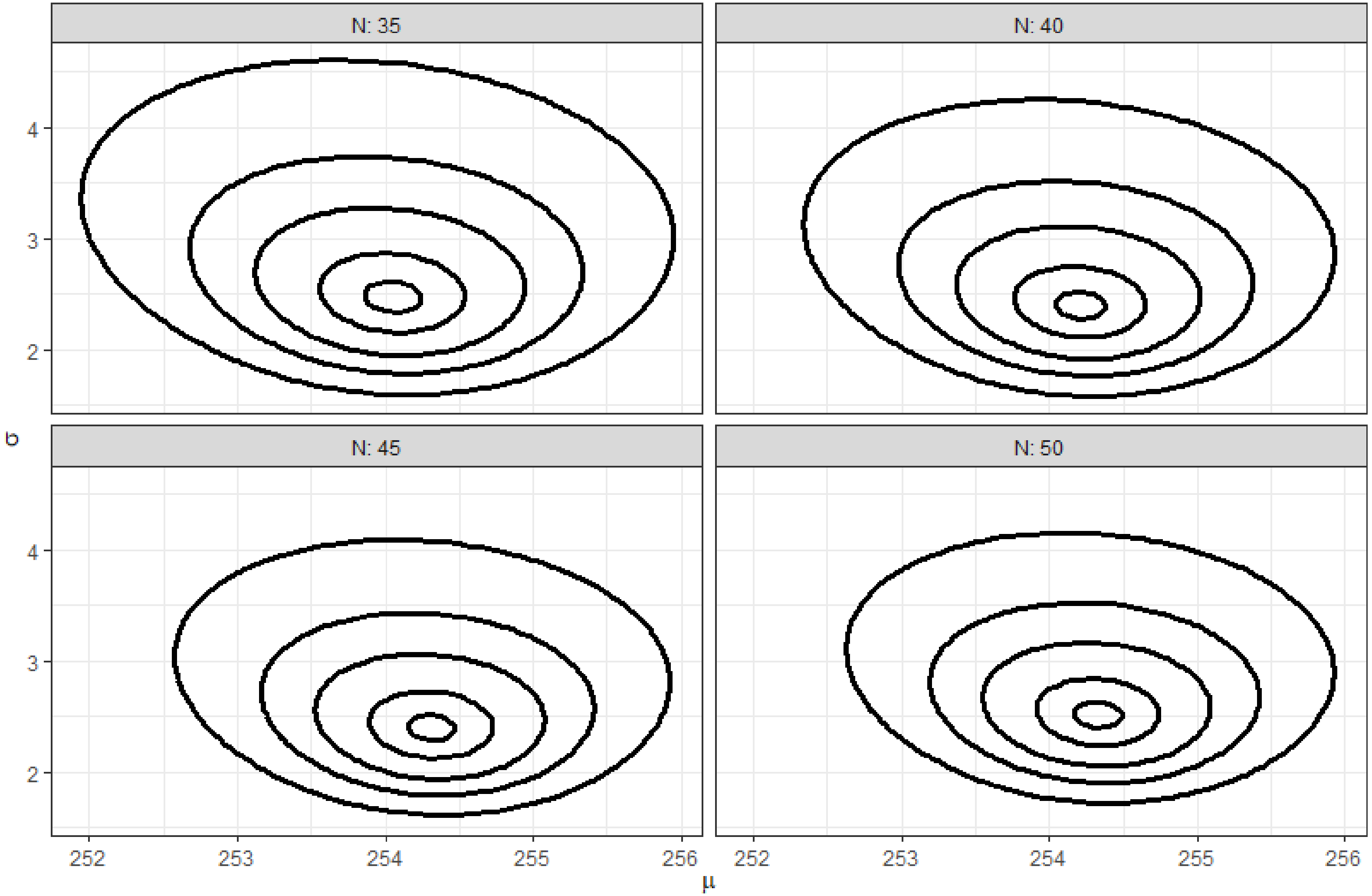


N: 0
N: 1
N: 5
N: 10
N: 15
N: 20
N: 25
N: 30
N: 35
N: 40
N: 45
N: 50
σ
1
2
3
4
5
μ
245
250
255
260



Seem to have reached a point of diminishing returns after 30 observations.

How would you describe the contour shapes?



What’s happening to the surface as the sample size increases?

- The contours are starting to become…more regular…or more *normal* looking…around the posterior mode.

- What if we could approximate the posterior as a Gaussian centered around the mode?

Laplace, Quadratic, or Normal approximation

- Approximate the joint posterior distribution with a **Multivariate Normal** (MVN) centered on the **MAP**.

- Benefits:
  - Straightforward to implement.
  - Relatively fast to execute.
  - Scales to a moderate number of variables.
- Cons:
  - Let’s see with an example later…

First things first…what's a MVN?

- Generalization of the Gaussian distribution to more than 1 dimension.

- Each dimension (variable) is a Gaussian and each subset of variables are MVN.

MVN density function

- There are D elements to the vector of variables:

$$\mathbf{x} = \{x_1, x_2, \dots x_d, \dots, x_D\}$$

- IMPORTANT: D refers to the number of variables, NOT the number of observations!

$$p(x_1, x_2, \dots, x_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

Break down the terms in the density

$$p(x_1, x_2, \dots, x_d | \boxed{\boldsymbol{\mu},} \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

Vector of means associated with each element in the x-vector:
$\boldsymbol{\mu} = \{\mu_1, \mu_2, \dots, \mu_d, \dots, \mu_D\}$

Break down the terms in the density

$$p(x_1, x_2, \ldots, x_d | \boldsymbol{\mu}, \boxed{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

$D \times D$ (variance-) covariance matrix between all elements of the x-vector.

Off-diagonal elements store the covariance between the variables.
Main-diagonal elements store the variance of each element in the x-vector

Break down the terms in the density

$$p(x_1, x_2, \ldots, x_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D \boxed{|\boldsymbol{\Sigma}|}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

Determinant of the covariance matrix.

Break down the terms in the density

$$p(x_1, x_2, \ldots, x_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boxed{\boldsymbol{\Sigma}^{-1}} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

Inverse of the covariance matrix.

Break down the terms in the density

$$p(x_1, x_2, \dots, x_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2} \boxed{(\mathbf{x}-\boldsymbol{\mu})^T} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

Transpose of the $(\mathbf{x}-\boldsymbol{\mu})$ vector

Break down the terms in the density

$$p(x_1, x_2, \ldots, x_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2} \boxed{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}\right\}$$

What's this?

Break down the terms in the density

$$p(x_1, x_2, \ldots, x_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\boxed{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}\right\}$$

Multidimensional generalization of the 1-D Gaussian term:

$$\left(\frac{x-\mu}{\sigma}\right)^2$$

Break down the terms in the density

$$p(x_1, x_2, \ldots, x_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\boxed{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}\right\}$$

$\sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$ is a generalized distance known as the Mahalanobis distance

Bivariate Gaussian – 2D case

- $D = 2$ the vector of elements becomes: $\mathbf{x} = \{x_1, x_2\}$
- Define the correlation coefficient between the two variables as, ρ.
- The mean vector and covariance matrix are:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Bivariate Gaussian – marginal distributions

- Each variable has a marginal Gaussian distribution:

$$x_1|\mu_1, \sigma_1 \sim \text{normal}(x_1|\mu_1, \sigma_1)$$

$$x_2|\mu_2, \sigma_2 \sim \text{normal}(x_2|\mu_2, \sigma_2)$$

Holds for higher dimensions!

Bivariate Gaussian – conditional distribution

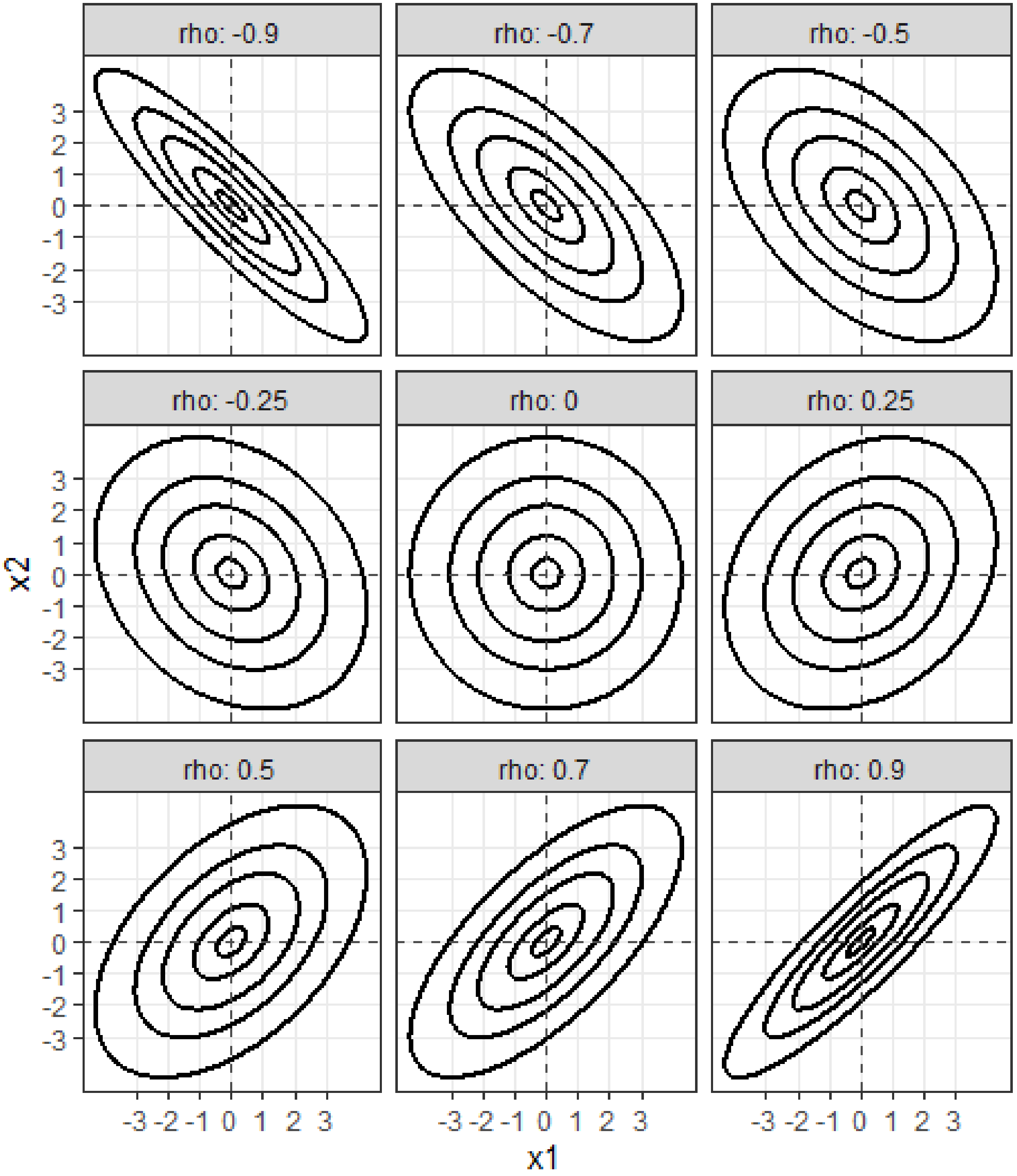
- The conditional distribution of one variable given the other…is also a Gaussian!

$$x_1|x_2, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), (1 - \rho^2)\sigma_1^2\right)$$

Holds for higher dimensions!

When ρ is near zero, the bivariate Gaussian density looks like circles.

As $|\rho|$ increases away from zero the circles look more and like ellipses.



As $|\rho|$ increases away from 0, knowing information about one variable changes what we think about the other!

Consider if $x_1 = 3$, the conditional mean of $x_2 \mid x_1$ depends on the correlation coefficient!

Compare the uncertainty in x_2 if $x_1 = 3$ as ρ changes.

When $|\rho| > 0$ specifying one variable reduces the uncertainty in the value of the other.