Shrinkage priors

Dr. Jarad Niemi

Iowa State University

August 26, 2021

Normal model with normal prior

Consider the model

$$Y \sim N(\theta, V)$$

with prior

$$\theta \sim N(m, C)$$

Then the posterior is

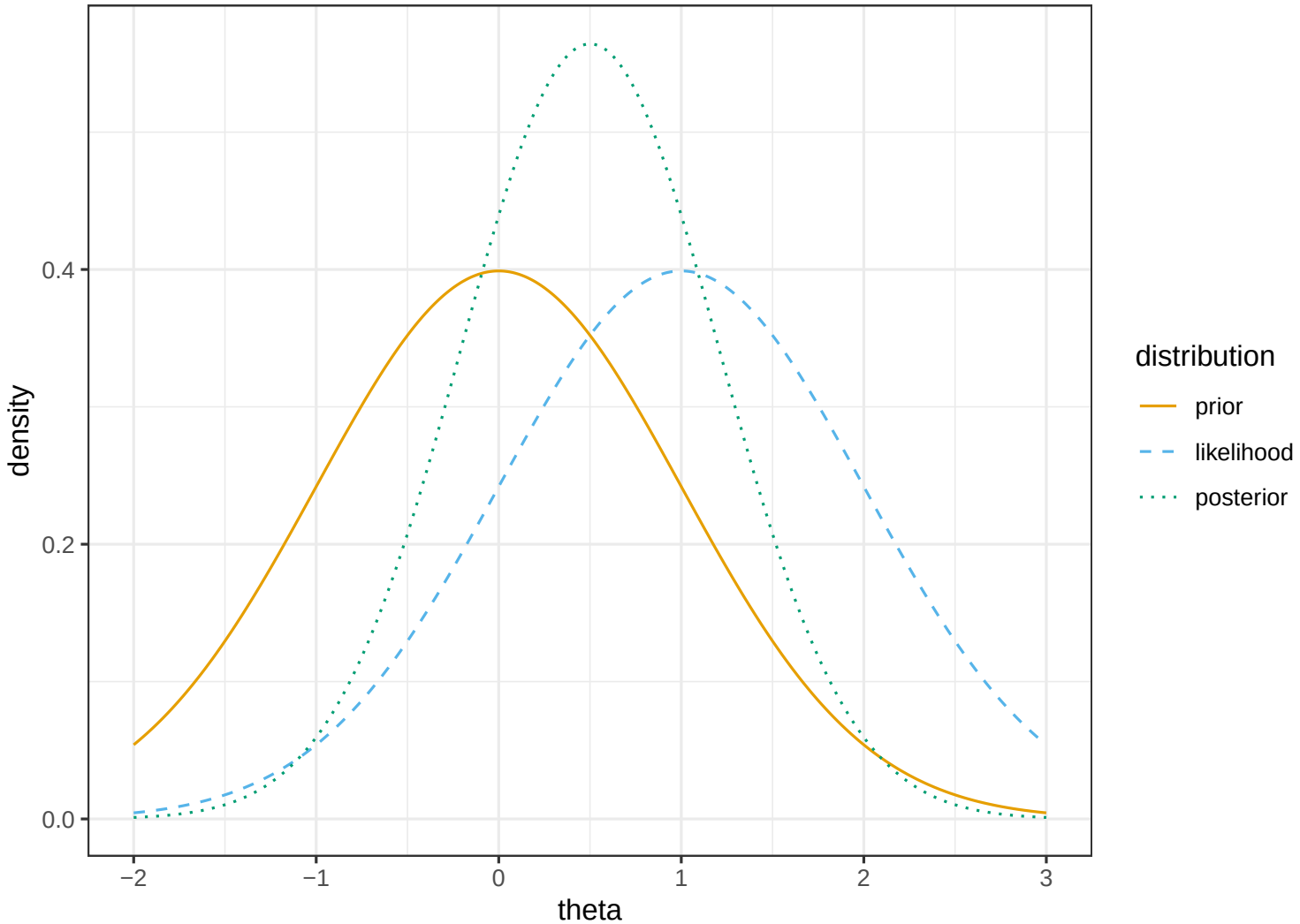
$$\theta|y \sim N(m', C')$$

where

$$\begin{aligned} C' &= 1/(1/C + 1/V) \\ m' &= C'[m/C + y/V] \end{aligned}$$

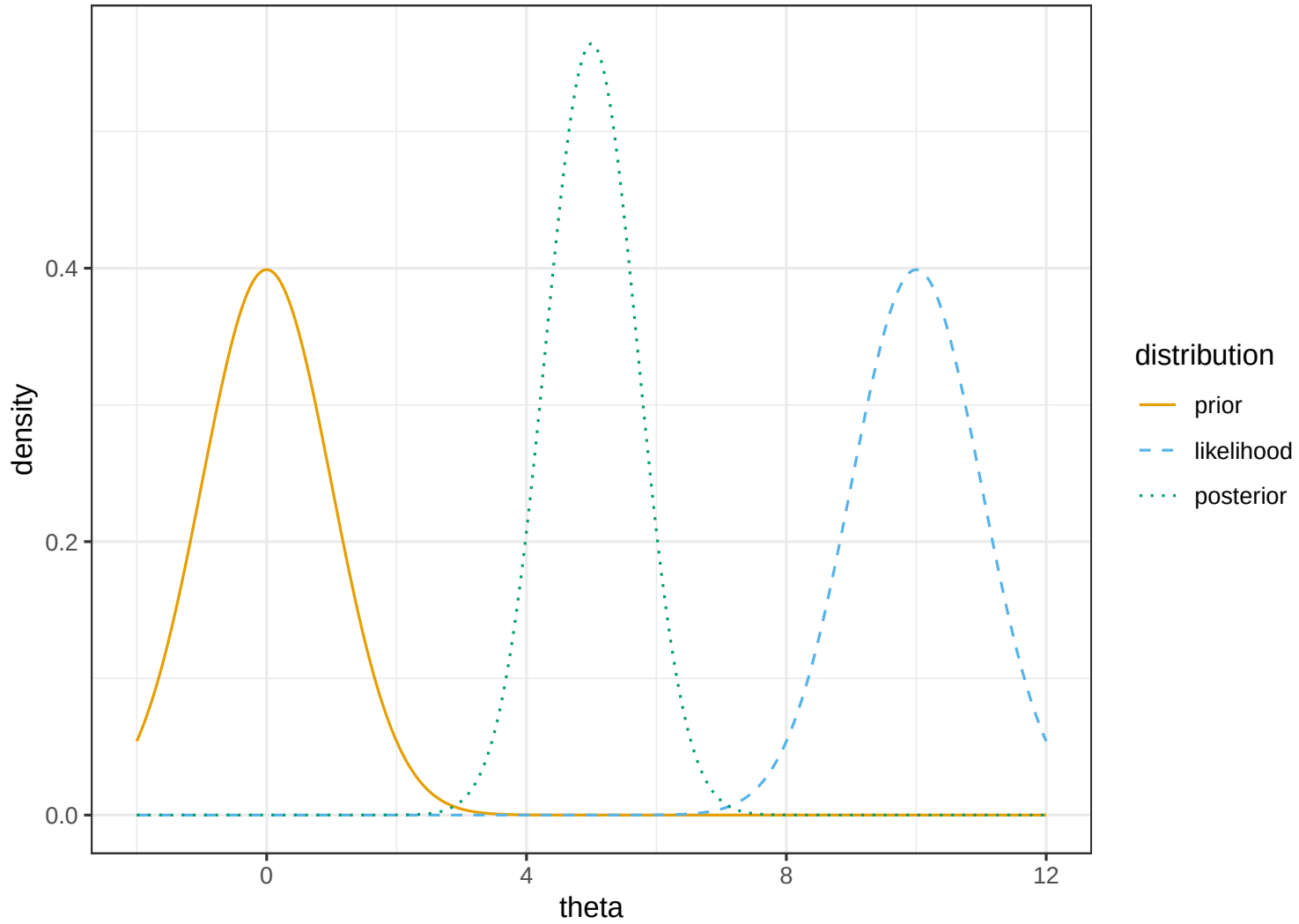
Normal model with normal prior (cont.)

For simplicity, let $V = C = 1$ and $m = 0$, then $\theta|y \sim N(y/2, 1/2)$.
Suppose $y = 1$, then we have



Normal model with normal prior (cont.)

Now suppose $y = 10$, then we have



Summary - normal model with normal prior

- If the prior and the likelihood agree, then posterior seems reasonable.
- If the prior and the likelihood disagree, then the posterior is ridiculous.
- The posterior precision is always the sum of the prior and data precisions and therefore the posterior variance always decreases relative to the prior.
- The posterior mean is always the precision weighted average of the prior and data.

Can we construct a prior that allows the posterior to be reasonable always?

Normal model with t prior

Now suppose

$$Y \sim N(\theta, V)$$

with

$$\theta \sim t_v(m, C),$$

where $E[\theta] = m$ for $v > 1$ and $Var[\theta] = C\frac{v}{v-2}$ for $v > 2$.

Now the posterior is

$$p(\theta|y) \propto e^{-(y-\theta)^2/2V}\left(1 + \frac{1}{v}\frac{(\theta - m)^2}{C}\right)^{-(v+1)/2}$$

which is not a known distribution, but we can normalize via

$$p(\theta|y) = \frac{e^{-(y-\theta)^2/2V}\left(1 + \frac{1}{v}\frac{(\theta-m)^2}{C}\right)^{-(v+1)/2}}{\int e^{-(y-\theta)^2/2V}\left(1 + \frac{1}{v}\frac{(\theta-m)^2}{C}\right)^{-(v+1)/2} d\theta}$$

Normal model with t prior (cont.)

Alternatively, we can calculate the marginal likelihood

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta \\ &= \int N(y;\theta,V)t_v(\theta;m,C)d\theta \end{aligned}$$

where

- $N(y;\theta,V)$ is the normal density with mean θ and variance V evaluated at y and
- $t_v(\theta;m,C)$ is the t distribution with degrees of freedom v, location m, and scale C evaluated at θ.

and then find the posterior

$$p(\theta|y) = N(y;\theta,V)t_v(\theta;m,C)/p(y).$$

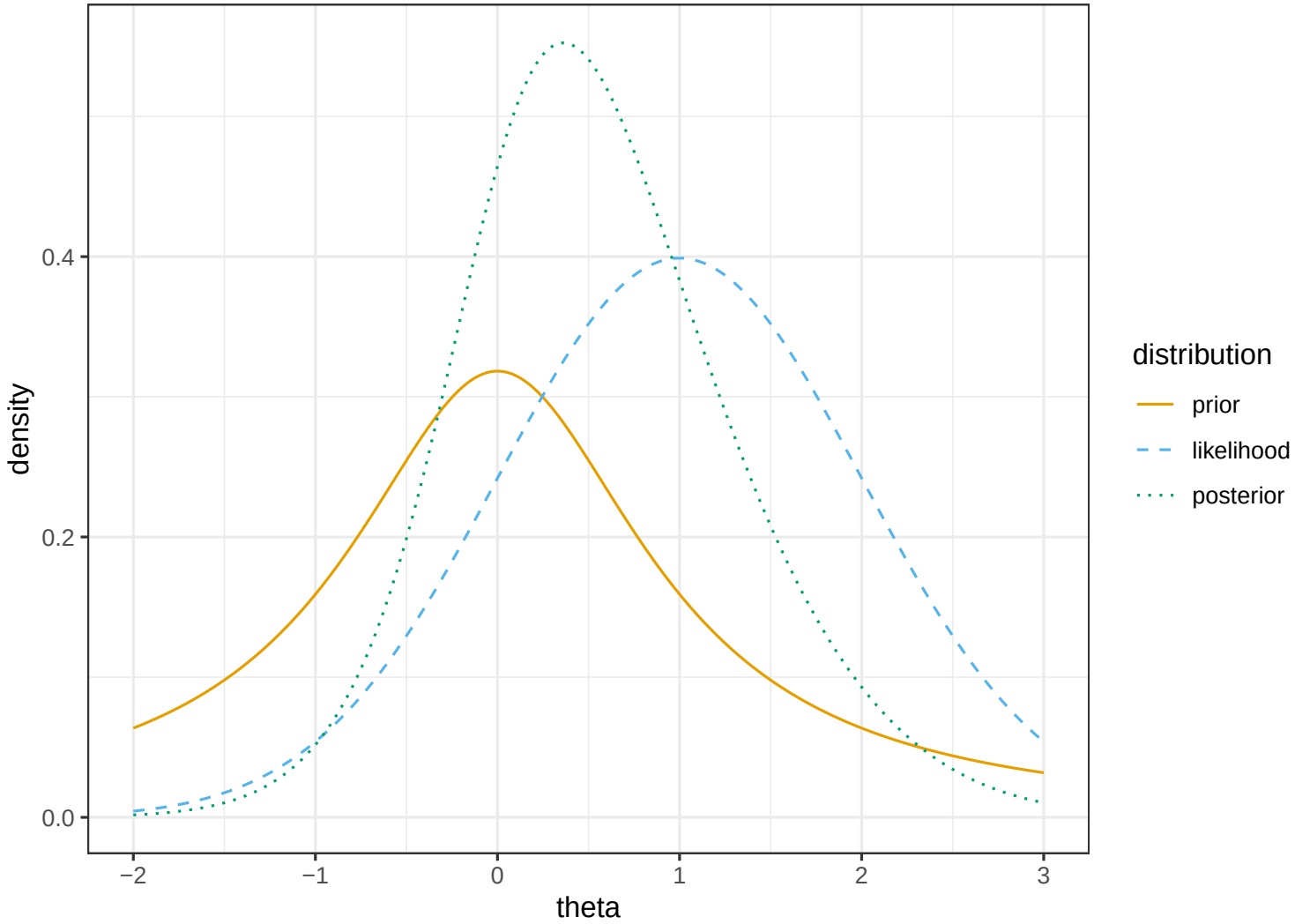
Normal model with t prior (cont.)

Since this is a one dimensional integration, we can easily handle it via the `integrate` function in R:

```
# A non-standard t distribution
my_dt = Vectorize(function(x, v=1, m=0, C=1, log=FALSE) {
  logf = dt((x-m)/sqrt(C), v, log=TRUE) - log(sqrt(C))
  if (log) return(logf)
  return(exp(logf))
})

# This is a function to calculate p(y|\theta)p(\theta).
f = Vectorize(function(theta, y=1, V=1, v=1, m=0, C=1, log=FALSE) {
  logf = dnorm(y, theta, sqrt(V), log=TRUE) + my_dt(theta, v, m, C, log=TRUE)
  if (log) return(logf)
  return(exp(logf))
})

# Now we can integrate it
(py = integrate(f, -Inf, Inf))

## 0.1657957 with absolute error < 1.6e-05
```

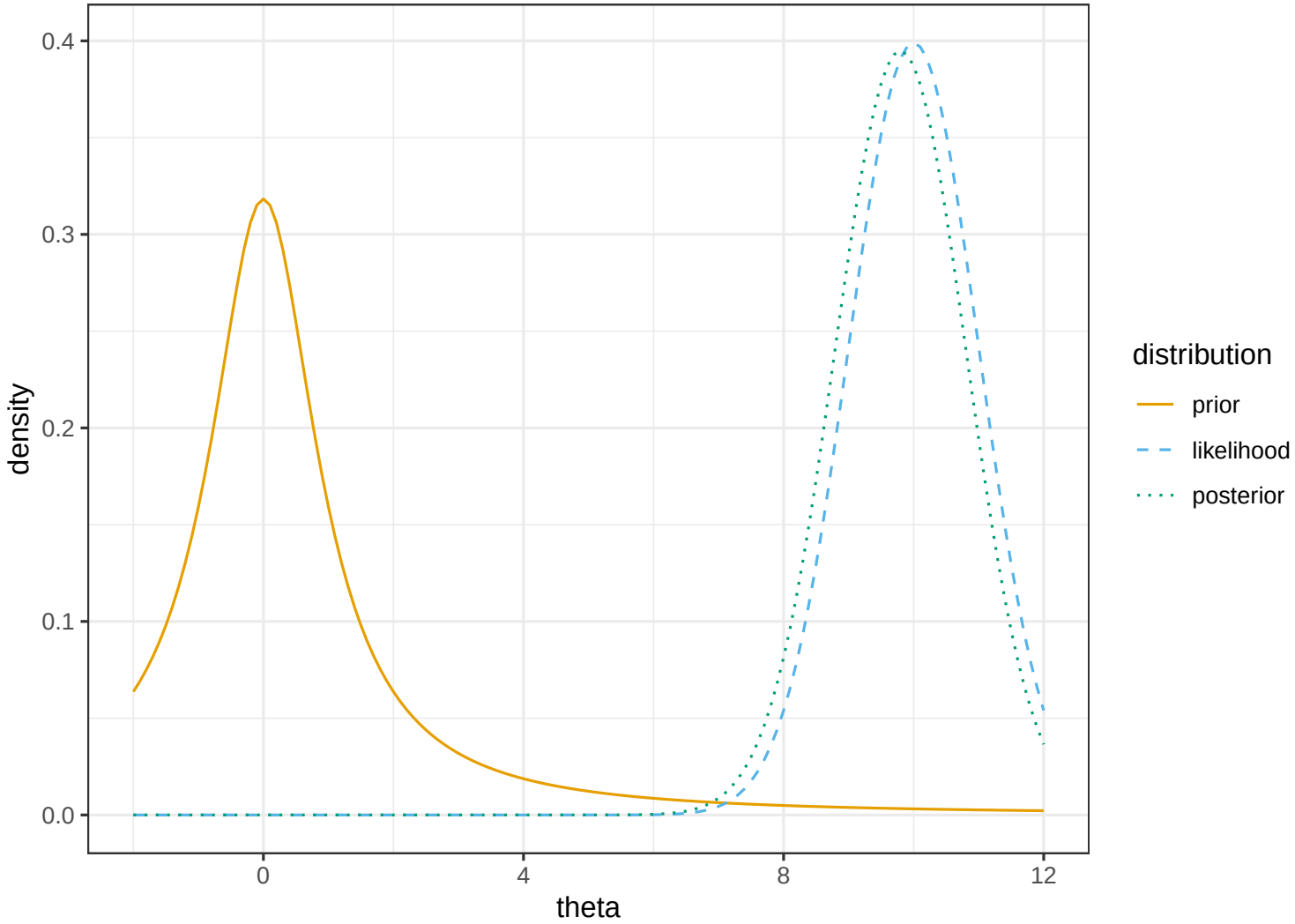
Normal model with t prior (cont.)

Let $v = 1$, $m = 0$, $V = C = 1$ and $y = 1$. then



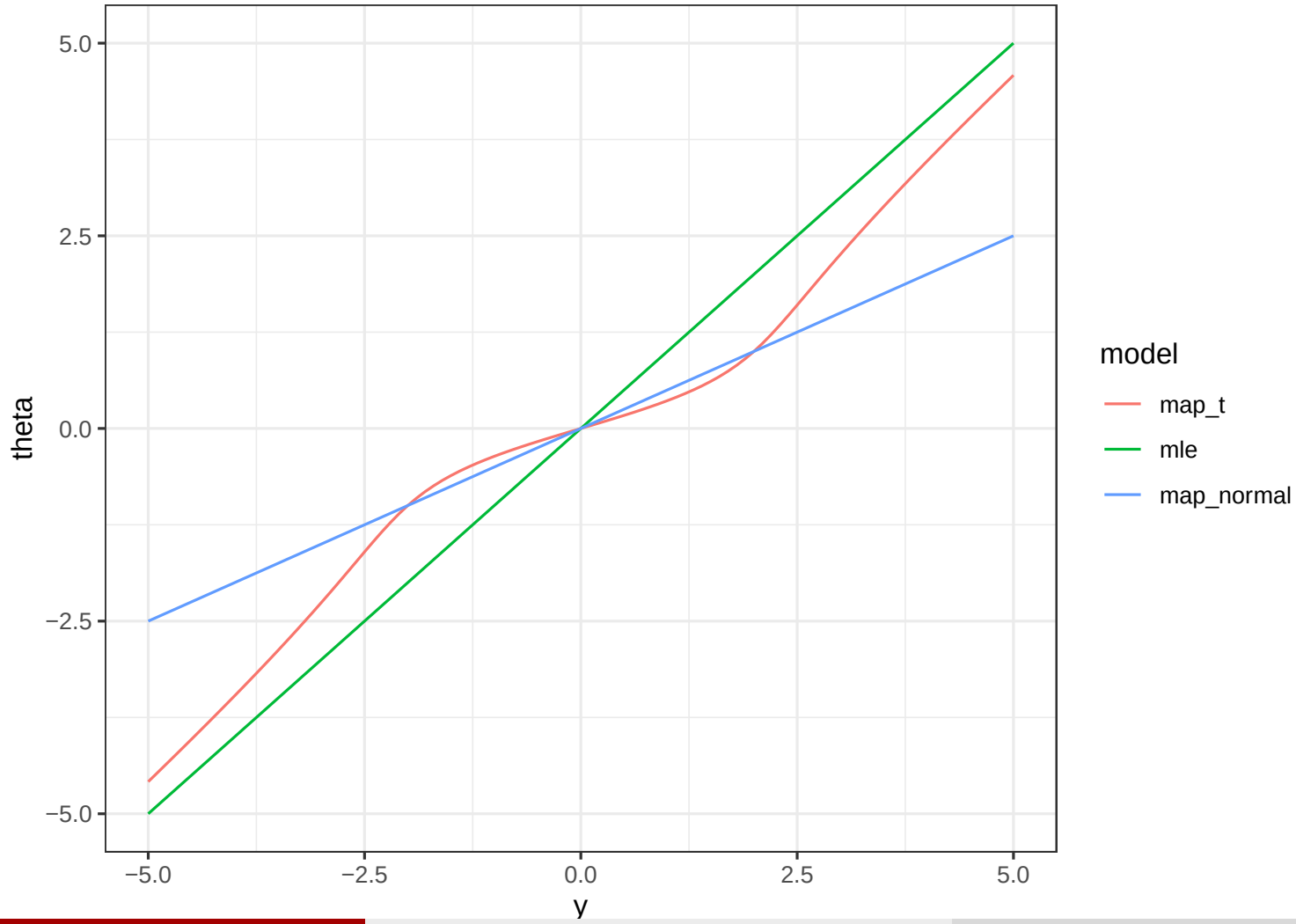
Normal model with t prior (cont.)

Let $\nu = 1$, $m = 0$, $V = C = 1$, and $y = 10$. then



Shrinkage of MAP as a function of signal

Let's take a look at the *maximum a posteriori* (MAP) estimates as a function of the signal (y) for the normal and t priors.



Summary - normal model with t prior

- A t prior for a normal mean provides a reasonable posterior even if the data and prior disagree.
- A t prior provides similar shrinkage to a normal prior when the data and prior agree, but provides little shrinkage when the data and prior disagree.
- The posterior variance decreases the most when the data and prior agree and decreases less as the data and prior disagree.

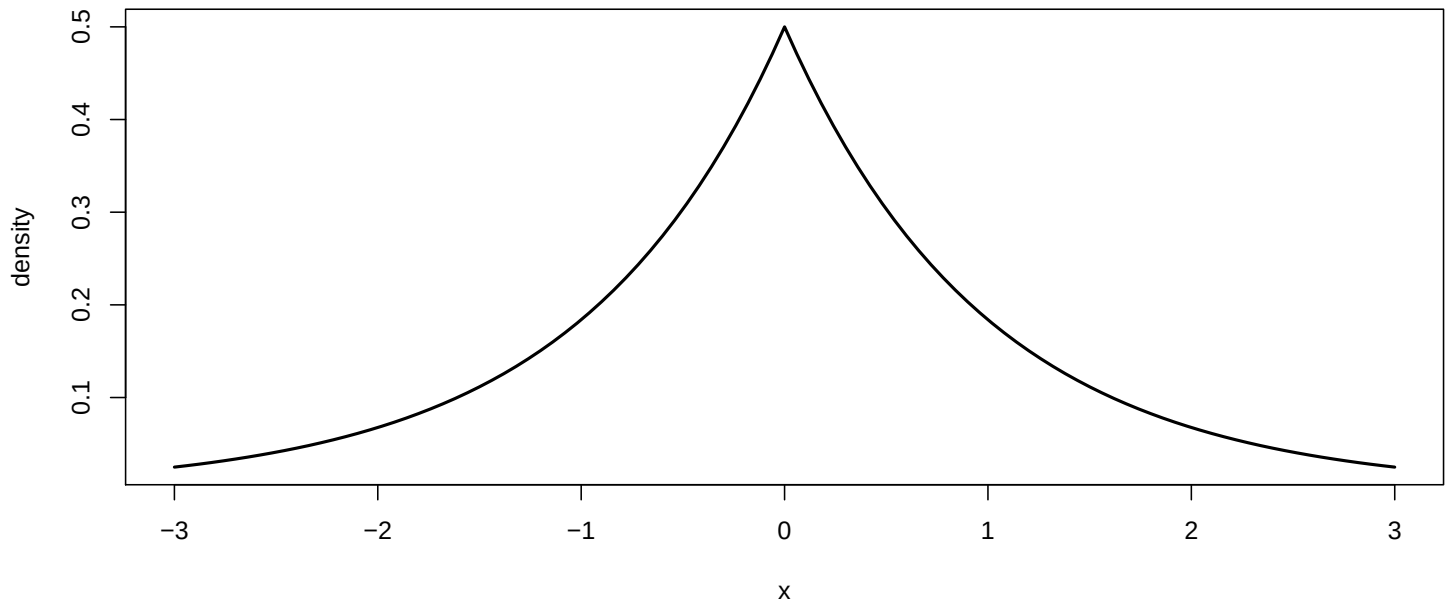
There are many times that we might believe the possibility of $\theta = 0$ or, at least, $\theta \approx 0$. In these scenarios, we would like our prior to be able to tell us this.

Can we construct a prior that allows us to learn about null effects?

Laplace distribution

Let $La(m, b)$ denote a Laplace (or double exponential) distribution with mean m, variance $2b^2$, and probability density function

$$La(x; m, b) = \frac{1}{2b} \exp\left(-\frac{|x - m|}{b}\right).$$



Laplace prior

Let

$$Y \sim N(\theta, V)$$

and

$$\theta \sim La(m, b)$$

Now the posterior is

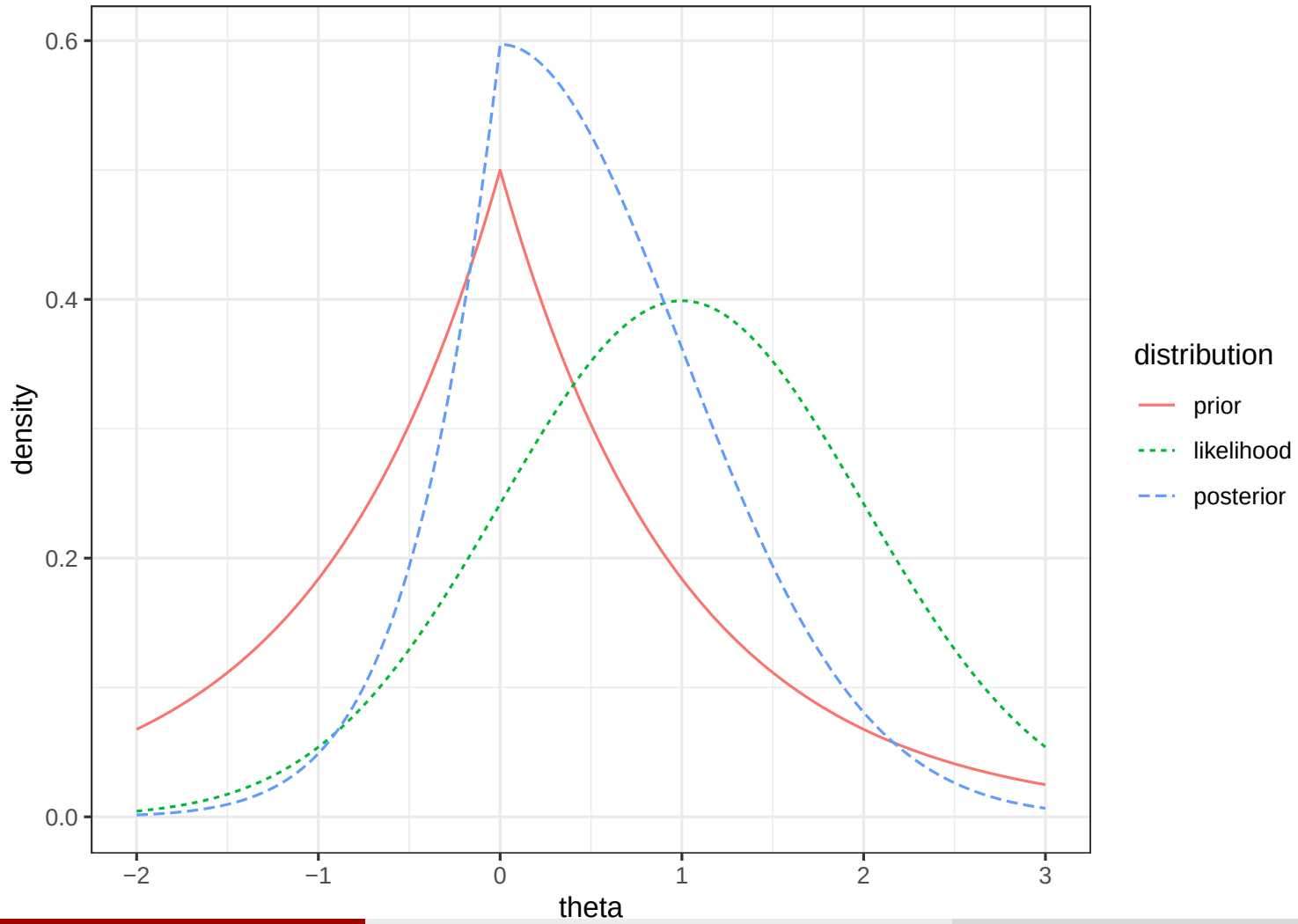
$$p(\theta|y) = \frac{N(y;\theta,V)La(\theta;m,b)}{p(y)} \propto e^{-(y-\theta)^2/2V} e^{-|\theta-m|/b}$$

where

$$p(y) = \int N(y;\theta,V)La(\theta;m,b)d\theta.$$

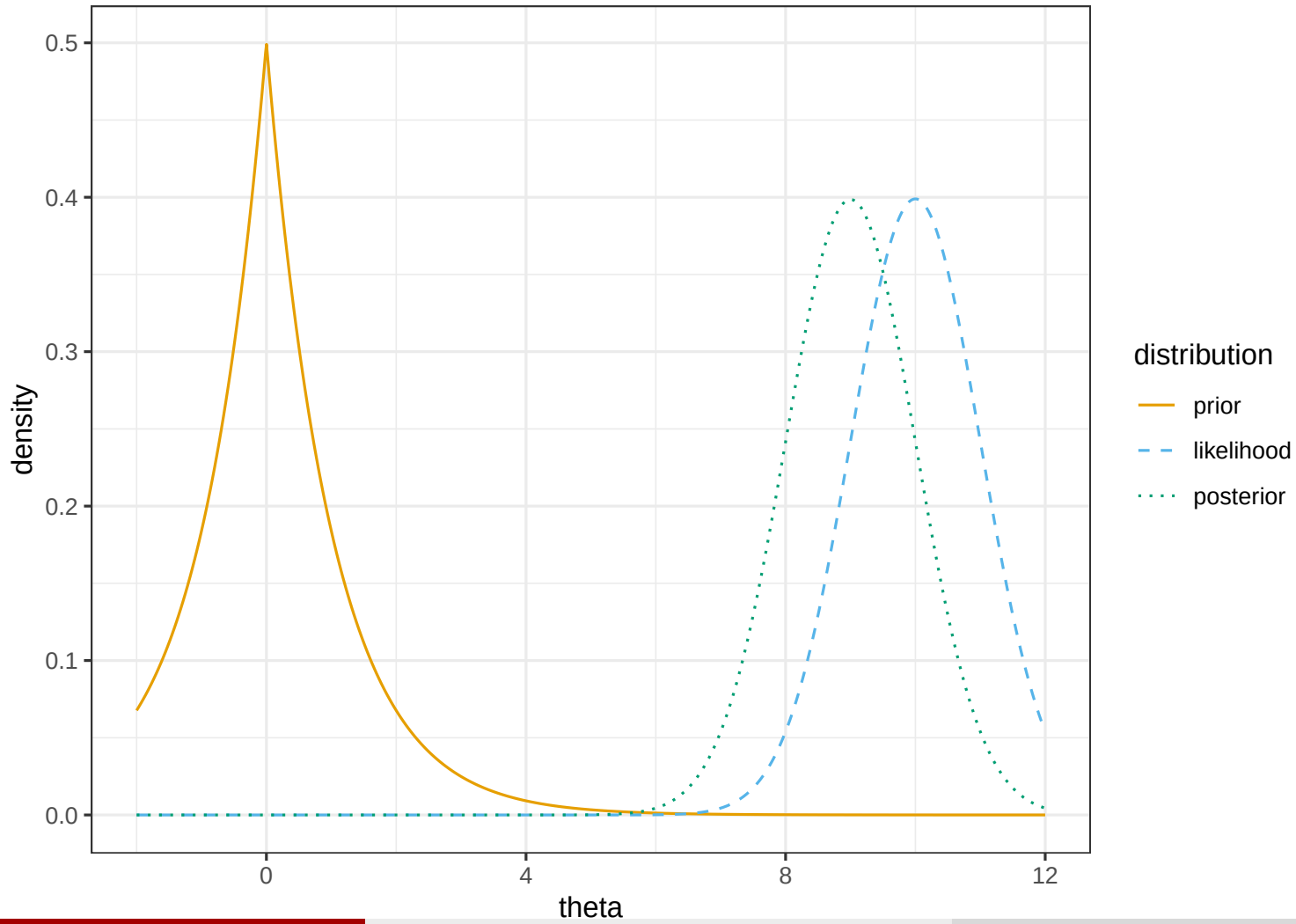
Laplace prior (cont.)

For simplicity, let $b = V = 1$, $m = 0$ and suppose we observe $y = 1$.

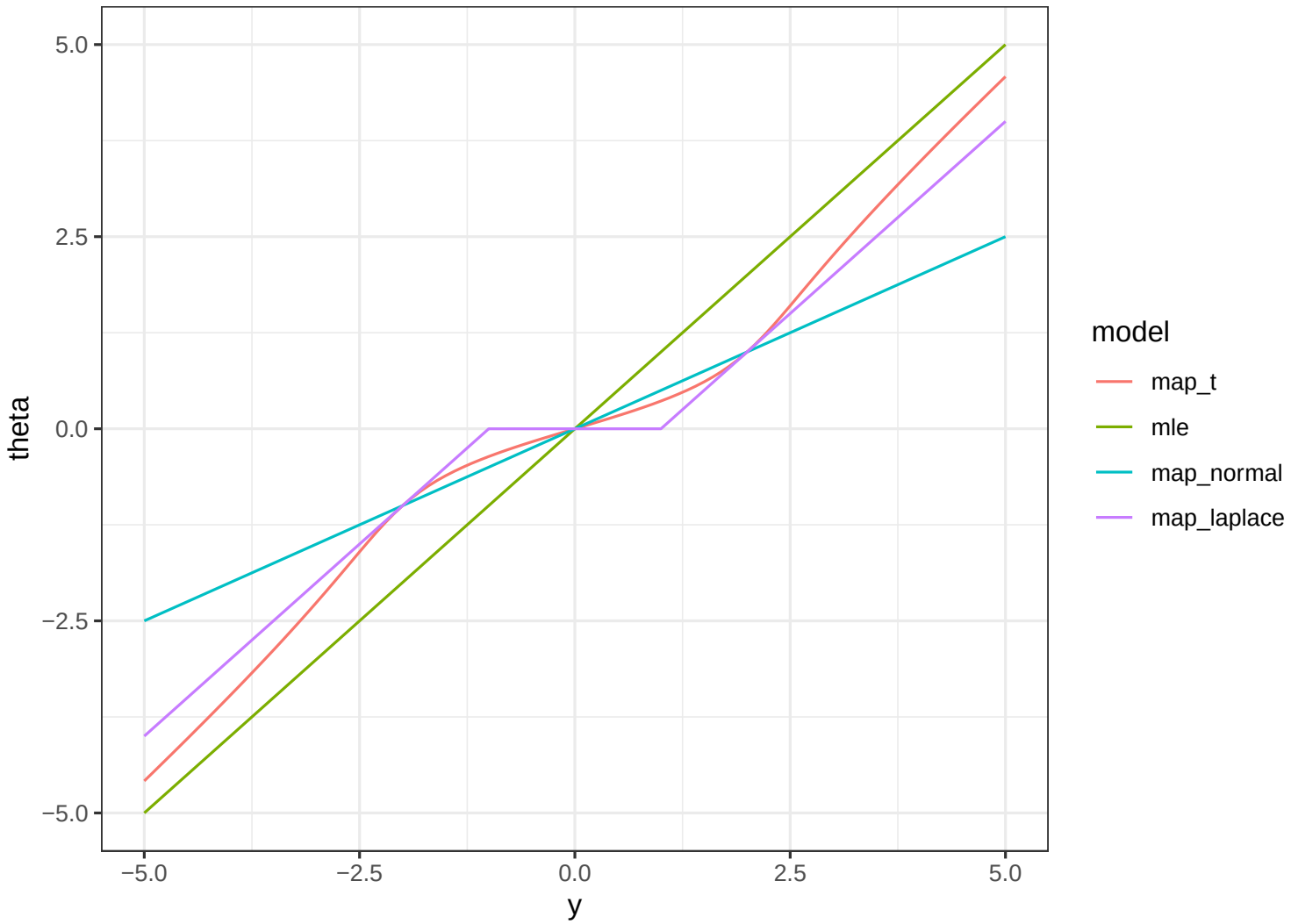


Laplace prior (cont.)

For simplicity, let $b = V = 1$, $m = 0$ and suppose we observe $y = 10$.



Laplace prior - MAP as a function of signal



Summary - Laplace prior

- For small signals, the MAP is zero (or m).
- For large signals, there is less shrinkage toward zero (or m) but more shrinkage than a t distribution.
- For large signals, the shrinkage is constant, i.e. it doesn't depend on y.

It's fine that the MAP is zero, but since the posterior is continuous, we have $P(\theta = 0|y) = 0$ for any y.

Can we construct a prior such that the posterior has mass at zero?

Dirac δ function

Let $\delta_c(x)$ be the Dirac δ function, i.e. formally

$$\delta_c(x) = \begin{cases} \infty & x = c \\ 0 & x \neq c \end{cases}$$

and

$$\int_{-\infty}^{\infty} \delta_c(x) dx = 1.$$

Thus $\theta \sim \delta_c \stackrel{d}{=} \delta_c(\theta)$ indicates that the random variable θ is a degenerate random variable with $P(\theta = c) = 1$.

Point-mass distribution

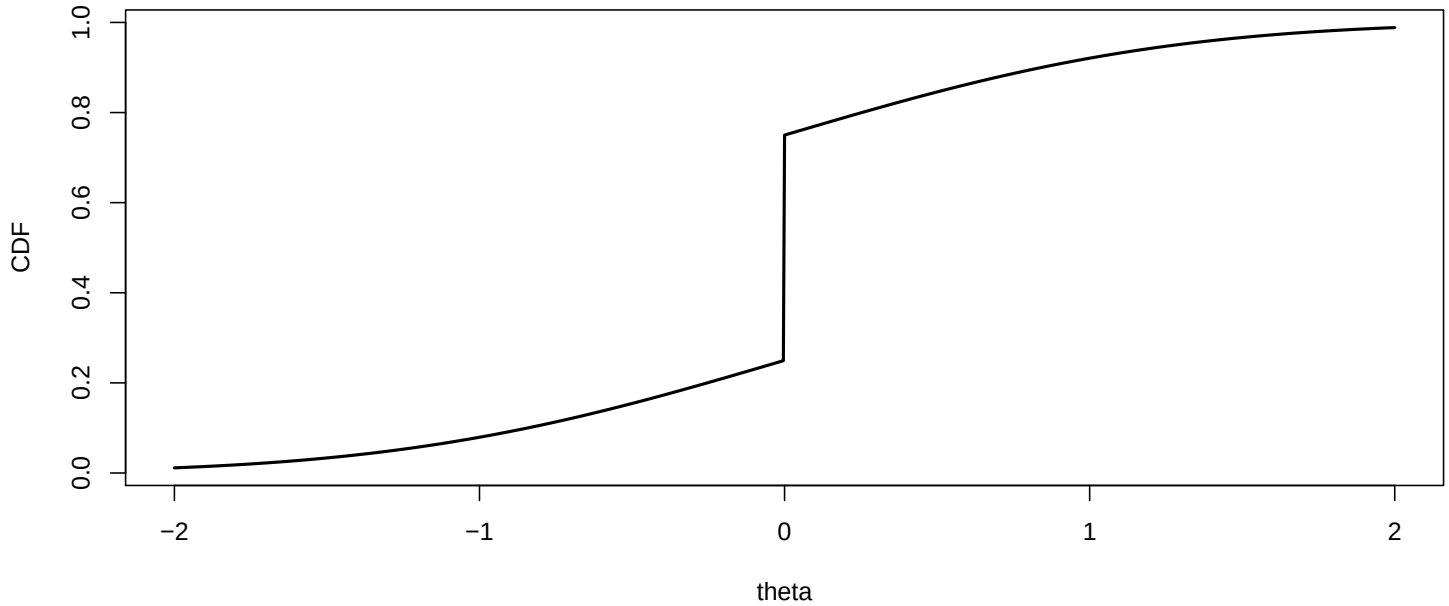
Let

$$\theta \sim p\delta_0 + (1-p)N(m, C)$$

be a distribution such that the random variable θ

- is 0 with probability p and
- a normal random variable with mean m and variance C with probability $(1-p)$.

If $p = 0.5$, $m = 0$, and $C = 1$, it's cumulative distribution function is



Point-mass prior

Suppose

$$Y \sim N(\theta, V)$$

and

$$\theta \sim p\delta_0 + (1-p)N(m, C).$$

Then

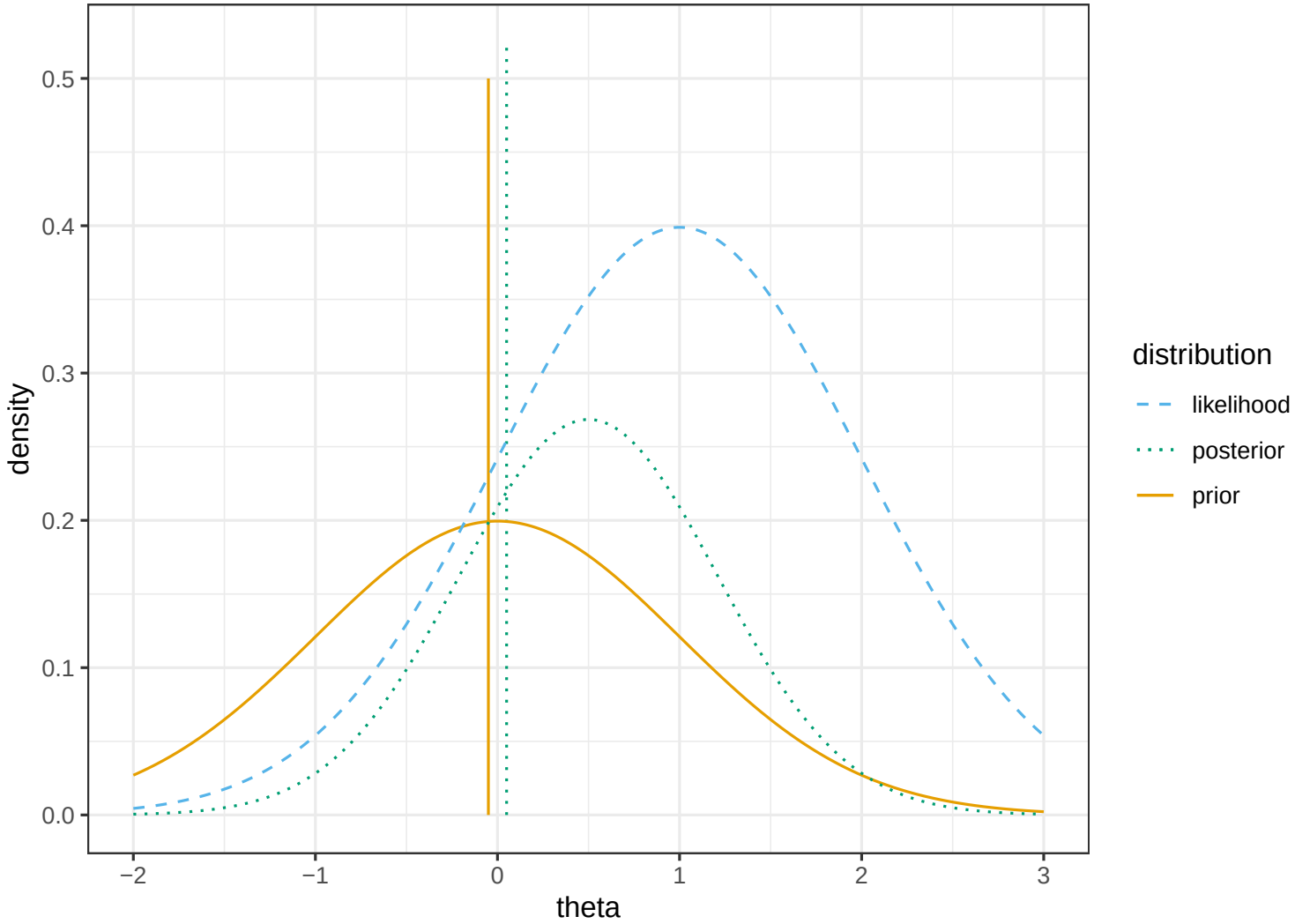
$$\theta|y \sim p'\delta_0 + (1-p')N(m', C')$$

where

$$\begin{aligned} p' &= \frac{pN(y;0,V)}{pN(y;0,V)+(1-p)N(y;m,C+V)} = \left(1 + \frac{(1-p)}{p}\frac{N(y;m,C+V)}{N(y;0,V)}\right)^{-1} \\ C' &= 1/(1/V + 1/C) \\ m' &= C'(y/V + m/C) \end{aligned}$$

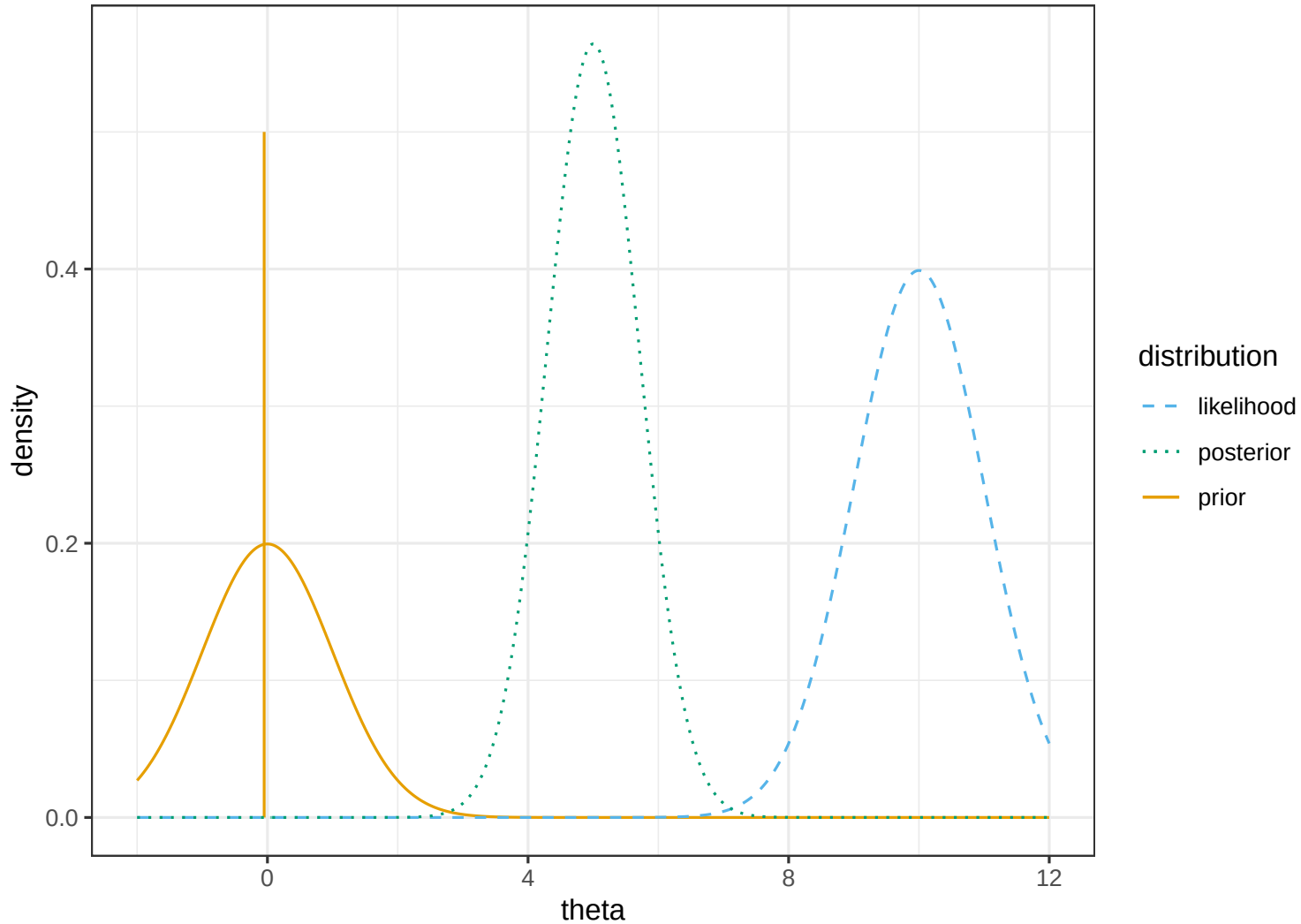
Point-mass prior (cont.)

For simplicity, let $V = C = 1$, $p = 0.5$, $m = 0$ and $y = 1$. Then



Point-mass prior (cont.)

For simplicity, let $V = C = 1$, $p = 0.5$, and $m = 0$. Suppose we observe $y = 1$.



Summary - point-mass prior

- For small signals, the posterior puts most of its mass at zero (or m).
- For large signals, the posterior puts most of its mass away from zero (or m) and therefore has the same problems that a normal prior has.

Can we create a prior that 1) puts most of the posterior mass at zero for small signals and 2) leaves large signals unshrunk?

Point-mass prior with t distribution

Suppose

$$Y \sim N(\theta, V)$$

and

$$\theta \sim p\delta_0 + (1-p)t_v(m, C).$$

Then

$$\theta|y \sim p'\delta_0 + (1-p')\,?$$

where

$$p' = \left(1 + \frac{(1-p)\int N(y;\theta,V)t_v(\theta;m,C)d\theta}{pN(y;0,V)}\right)^{-1}$$

and

$$? \propto N(y;\theta,V)t_v(\theta;m,C).$$

But we already calculated this posterior earlier in the lecture, i.e. normal model with t prior.

Point-mass prior with t distribution (cont.)

Suppose $v = V = C = 1$, $p = 0.5$, $m = 0$, and $y = 1$.

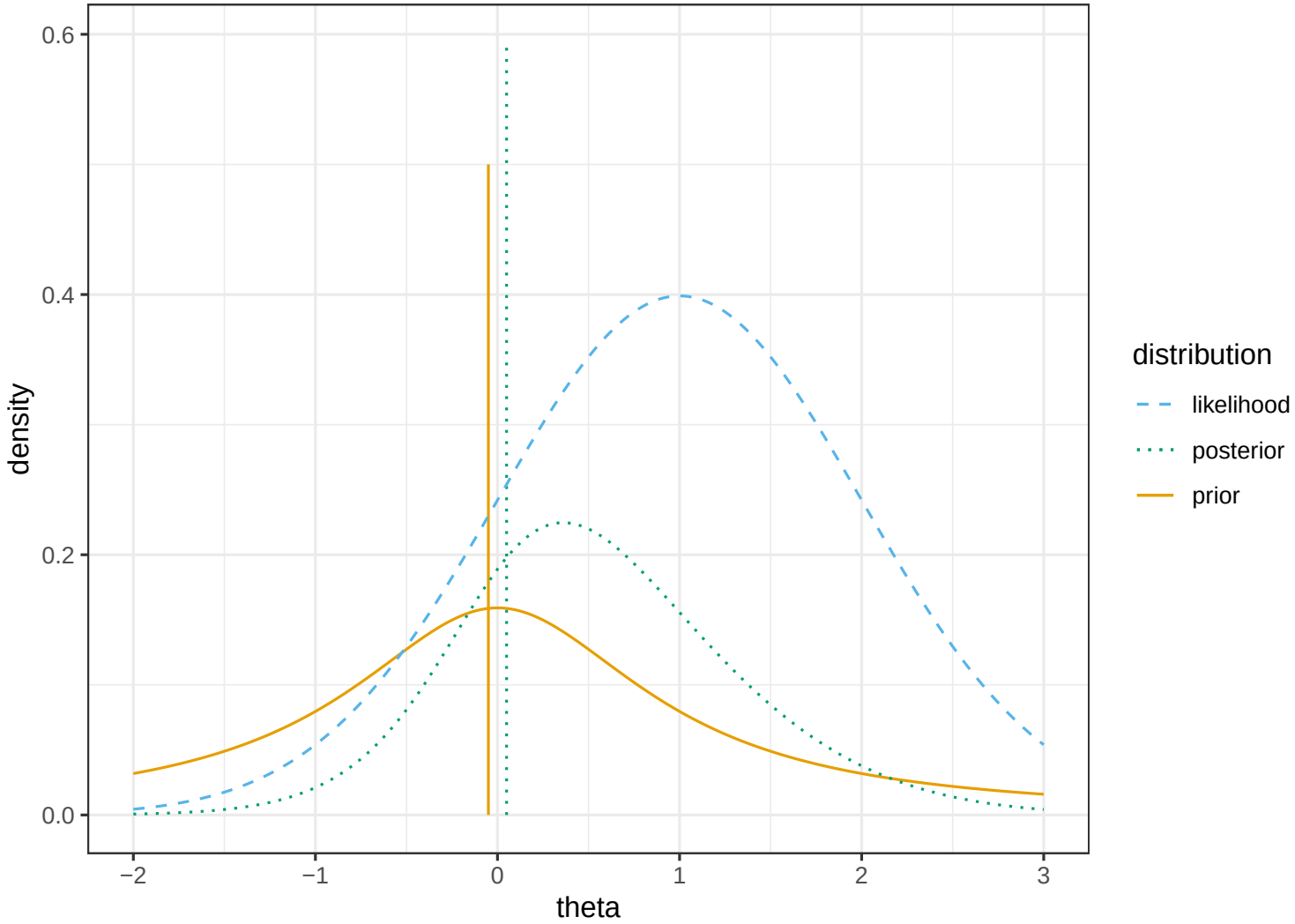
Then, we can calculate the following integral (marginal likelihood) numerically

$$\int N(y; \theta, V) t_v(\theta; m, C) d\theta$$

```
v = C = V = 1; p = 0.5; m = 0; y=1
(int = integrate(function(x) dnorm(y,x,sqrt(V))*my_dt(x), -Inf, Inf))

## 0.1657957 with absolute error < 1.6e-05

(int0 = dnorm(y,0,sqrt(V)))

## [1] 0.2419707

(pp = 1/(1+(1-p)*int$value/(p*int0)))

## [1] 0.5934053
```

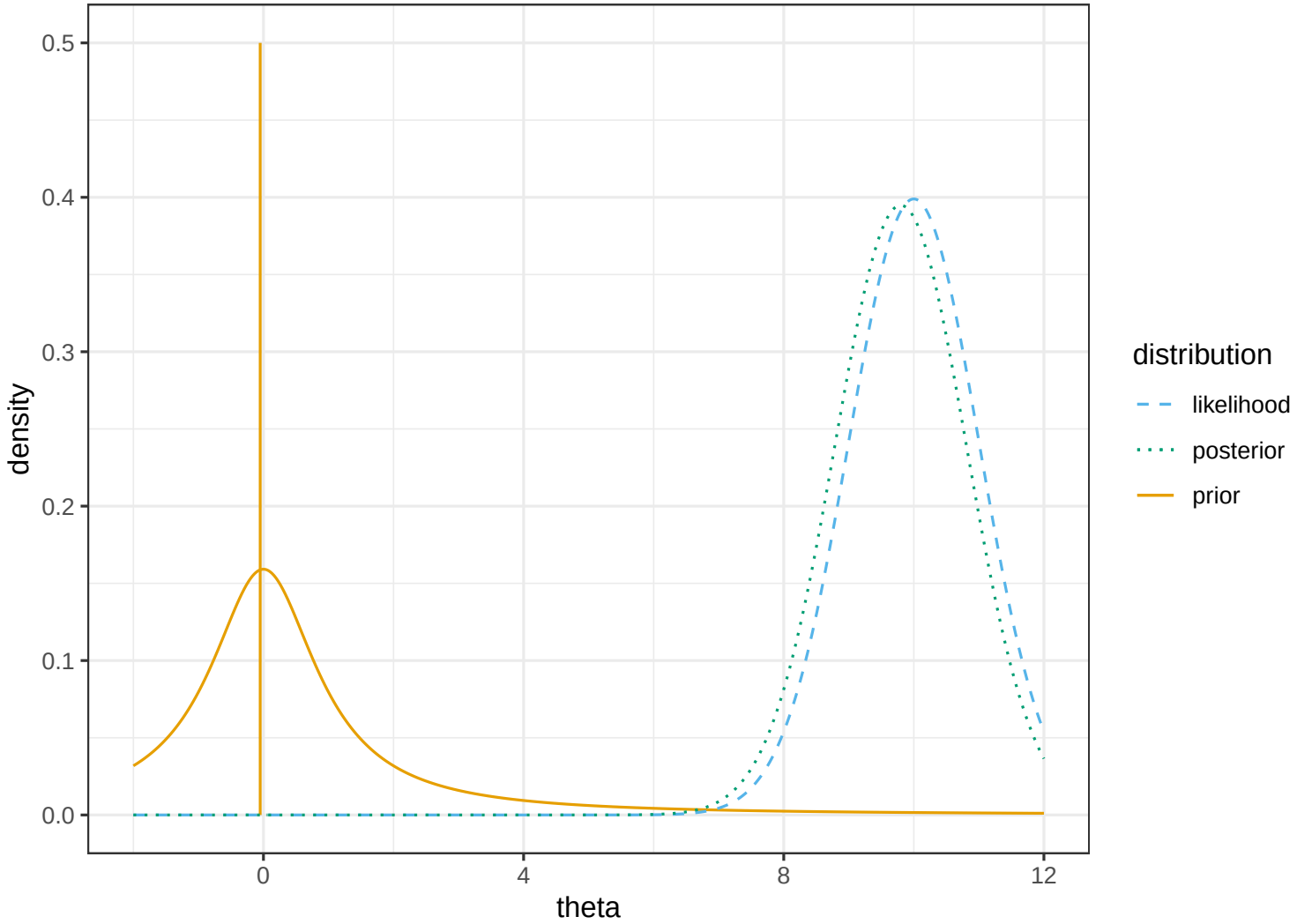
Point-mass prior with t distribution (cont.)

Suppose $v = V = C = 1$, $p = 0.5$, and $m = 0$. And we observe $y = 1$.



Point-mass prior with t distribution (cont.)

Suppose $v = V = C = 1$, $p = 0.5$, and $m = 0$. And we observe $y = 10$.



Summary

- Heavy tails allow the likelihood to easily overwhelm the prior.
- A peak allows "complete" shrinkage.

Discussion questions

- What would happen if we tried to take this idea to the logical extreme by having a point-mass prior with an improper distribution for the non-point mass portion?
- Why do the phrases "random effects" or "mixed effects" imply a normal distribution for the random effects?