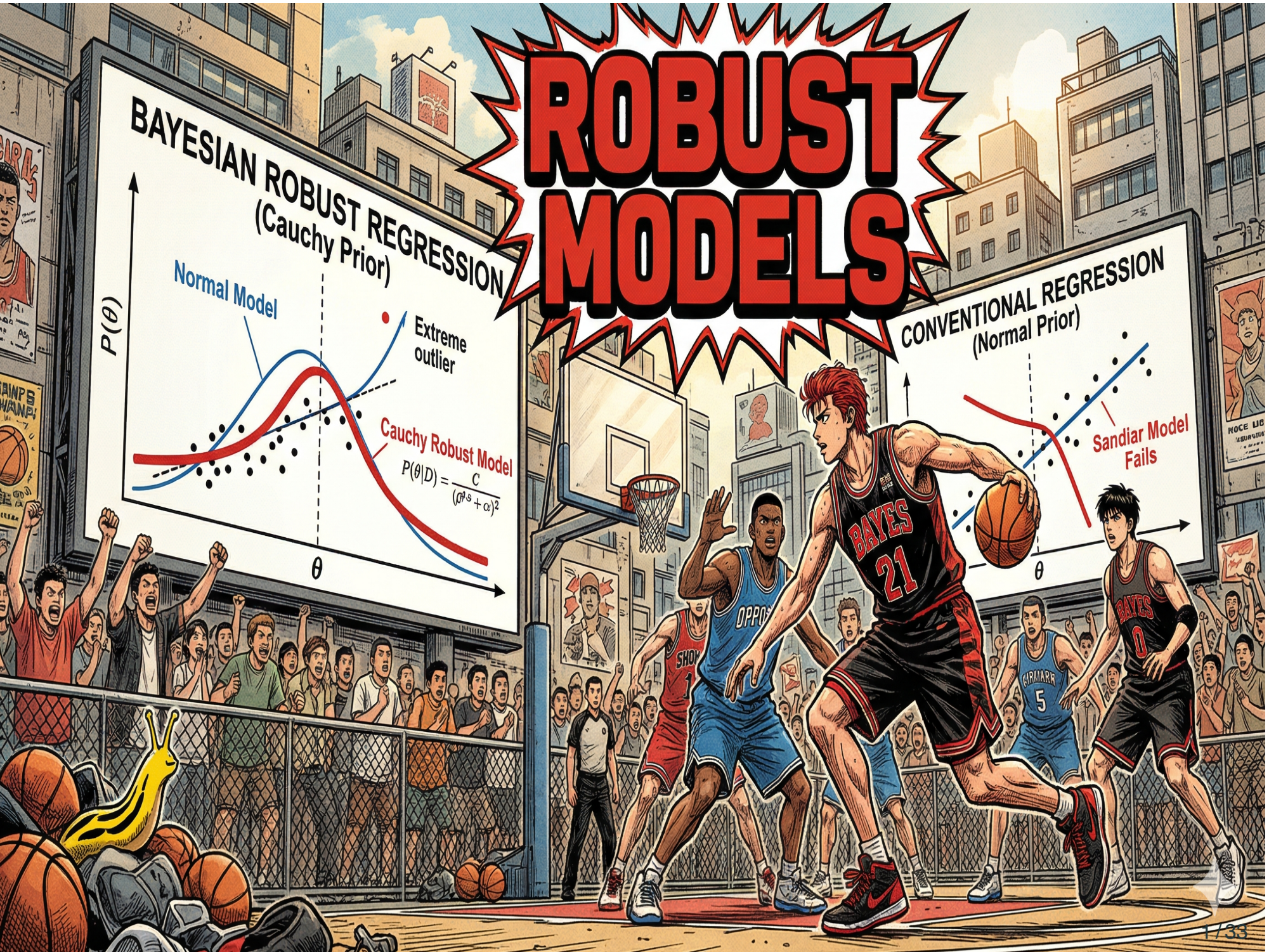

ROBUST MODELS
BAYESIAN ROBUST REGRESSION
(Cauchy Prior)
Normal Model
Extreme outlier
Cauchy Robust Model
$P(\theta|D) = \frac{C}{(\theta^2 + \alpha)^2}$
$P(\theta)$
θ
CONVENTIONAL REGRESSION
(Normal Prior)
Sandiar Model Fails
θ
BAYES 21
BAYES 0

Robust inference

- What do we mean by robustness? We may want model inference to be less sensitive to
  - outliers ,i.e., data that are surprising;
  - perturbations on the model specification, e.g., prior choice, auxiliary variables;
  - model misspecification;
- Many other aspects of robustness in the more general sense that we will not discuss here:
  - Confounding, omitted variables, ...
  - Distribution shift
  - ...

Case Study 1: outliers and tail of the normal distribution

- Normal distribution is notoriously not robust to outliers.
- A common alternative to normal distribution that allows outliers to arise is the (location-scale) Student's t-distribution

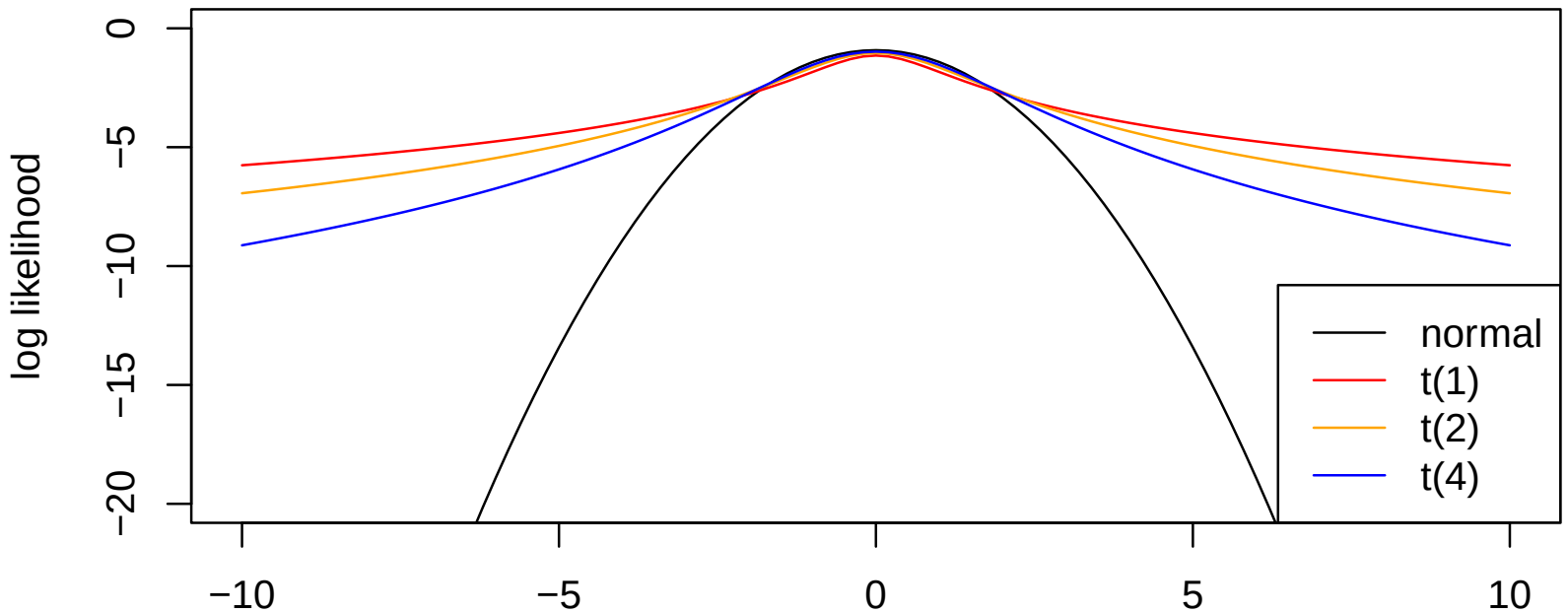
$$p(x \mid \nu, \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}}\left(1 + \frac{1}{\nu}(\frac{x-\mu}{\sigma})^2\right)^{-(\nu+1)/2}$$

  with $\mathbb{E}(x) = \mu$ for $\nu > 1$, and $var(x) = \sigma^2 \frac{\nu}{\nu-2}$ for $\nu > 2$.
- Notice that Cauchy distribution is a special case when the degree of freedom $\nu = 1$,

$$p(x \mid \mu, \sigma) = \frac{1}{\pi\sigma}\left(1 + (\frac{x-\mu}{\sigma})^2\right)^{-1}$$

Tail behavior of t-distribution

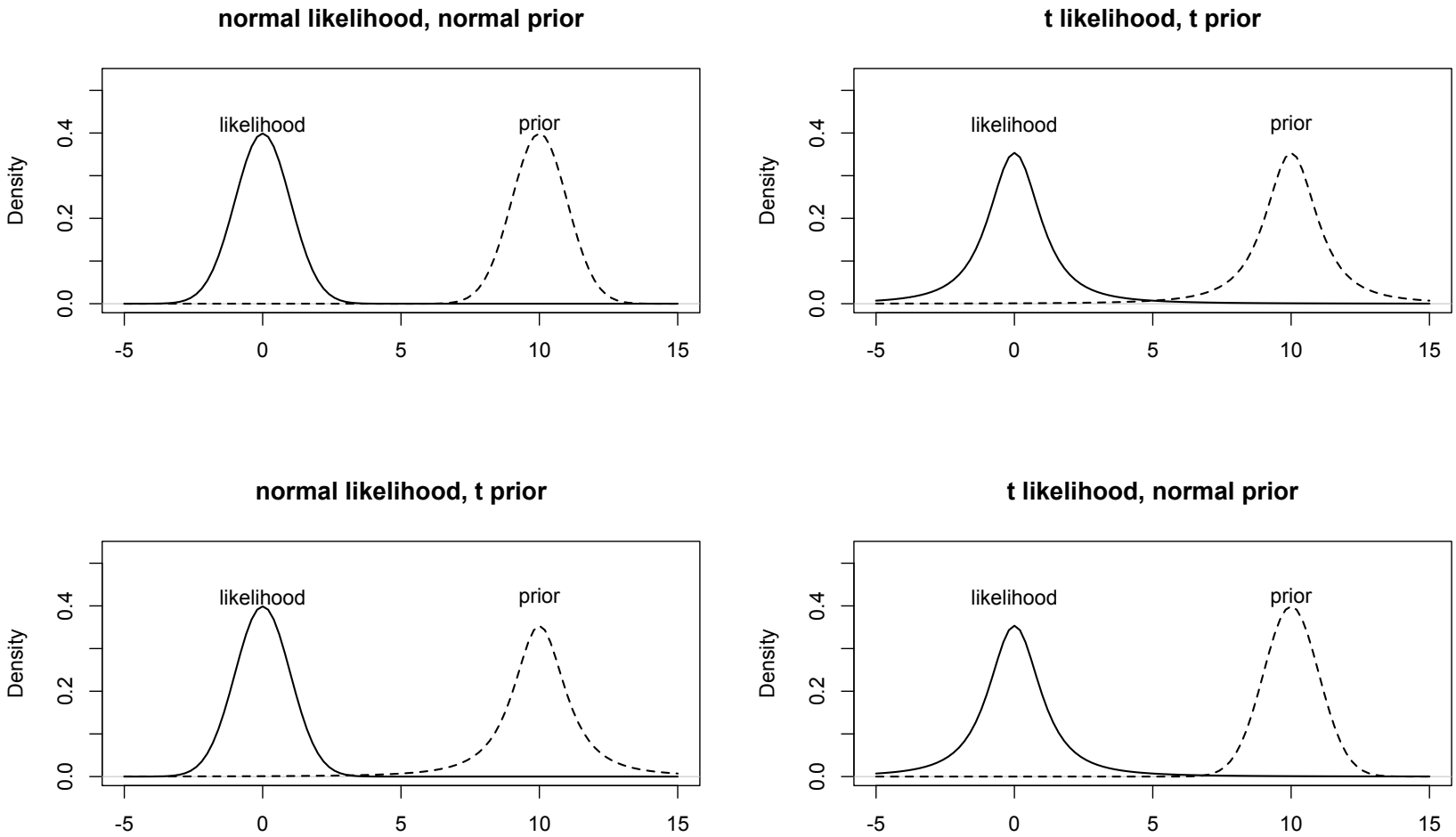
- Comparing $N(0, 1)$ and t-distribution with different degrees of freedom (mean 0 and scale 1).



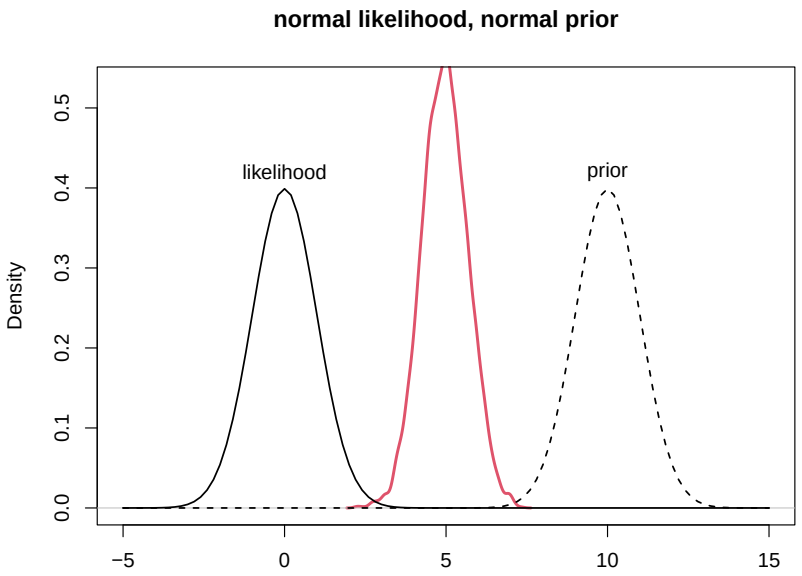
Implication for Bayesian modeling

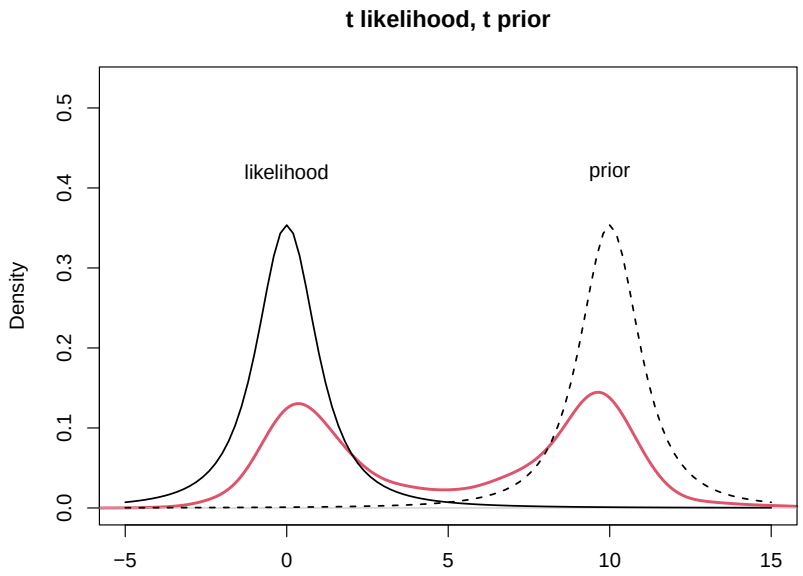
- How do we make use of the wide tail?
- Consider a single observation $y = 0$, and the combination of the following prior and likelihood models.
- Likelihood:
  - $y \sim N(\mu, 1)$
  - $y \sim t_2(\mu, 1)$
- Prior:
  - $\mu \sim N(10, 1)$
  - $\mu \sim t_2(10, 1)$
- The data point of $y = 0$ conflicts with both priors centered at 10, and thus an outlier from the model's point of view.
- This is a toy example of course. But can you sketch what the posterior of μ will look like based on intuition?

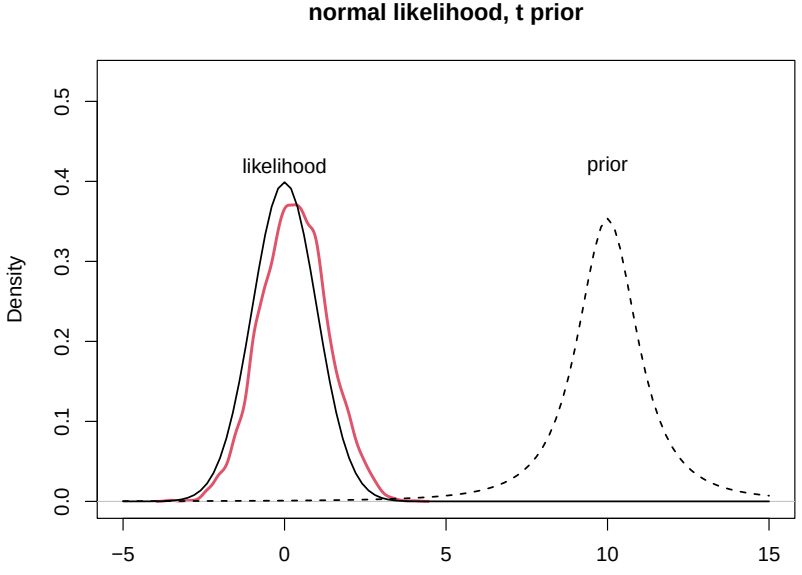
An outlier

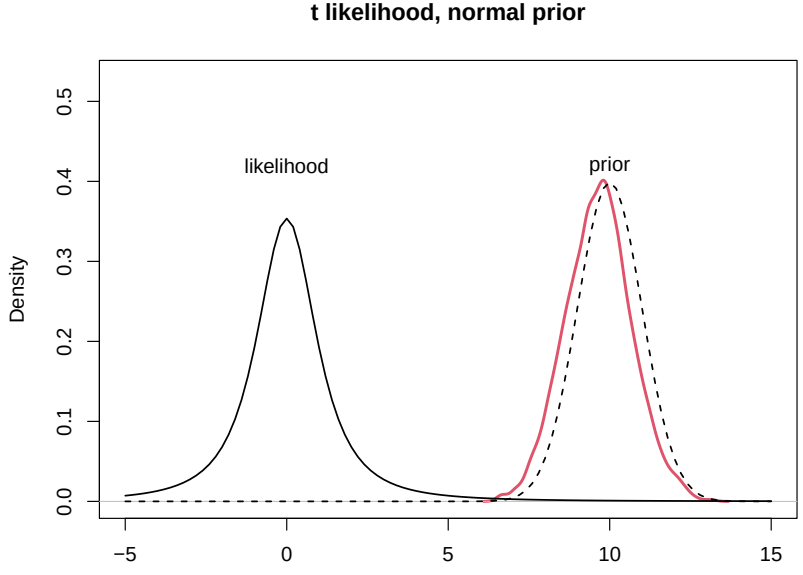

normal likelihood, normal prior
t likelihood, t prior
normal likelihood, t prior
t likelihood, normal prior
likelihood
prior
Density
0.0
0.2
0.4
-5
0
5
10
15

An outlier

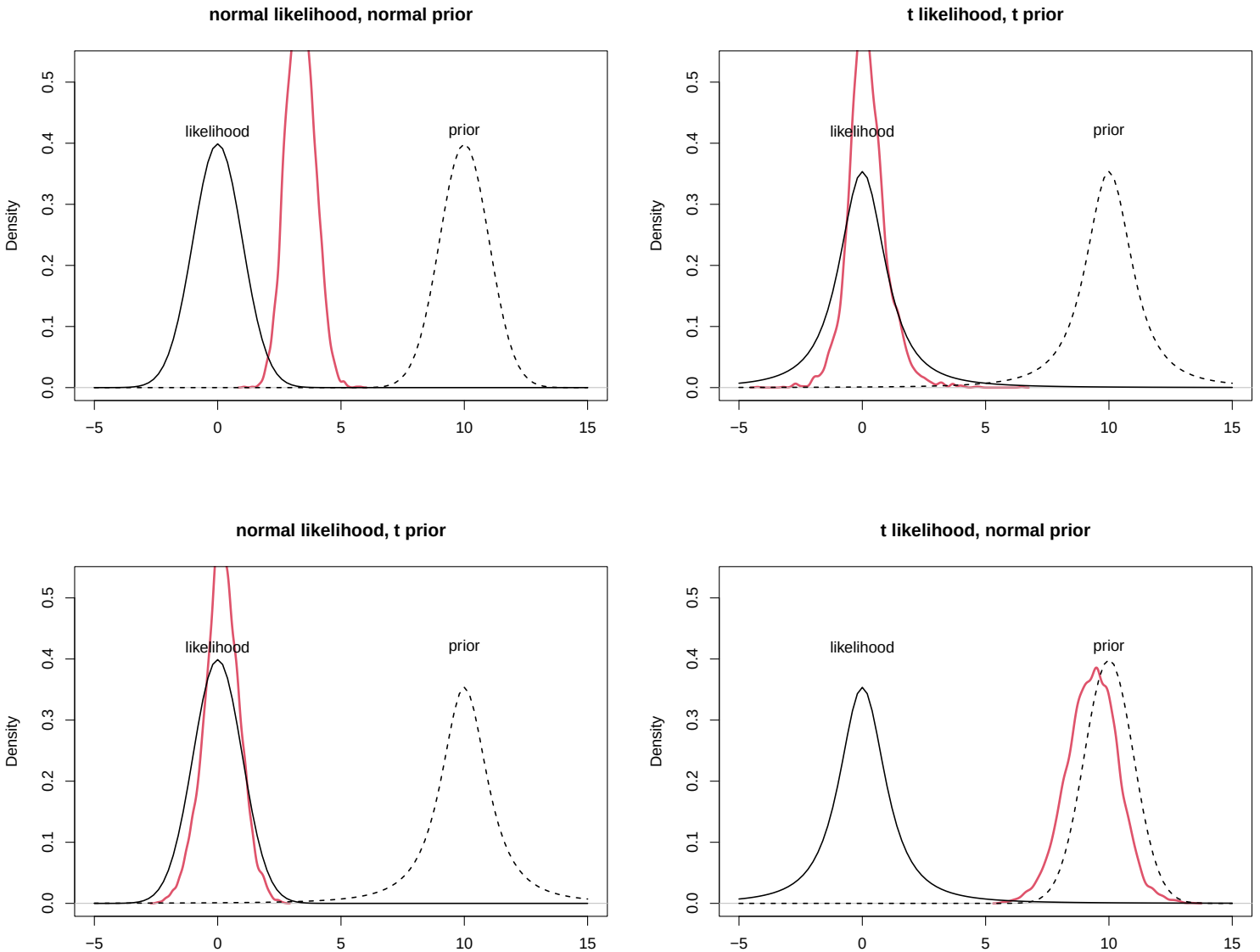

normal likelihood, normal prior
likelihood
prior
Density
0.0
0.1
0.2
0.3
0.4
0.5
−5
0
5
10
15


t likelihood, t prior
likelihood
prior
Density
0.0
0.1
0.2
0.3
0.4
0.5
−5
0
5
10
15

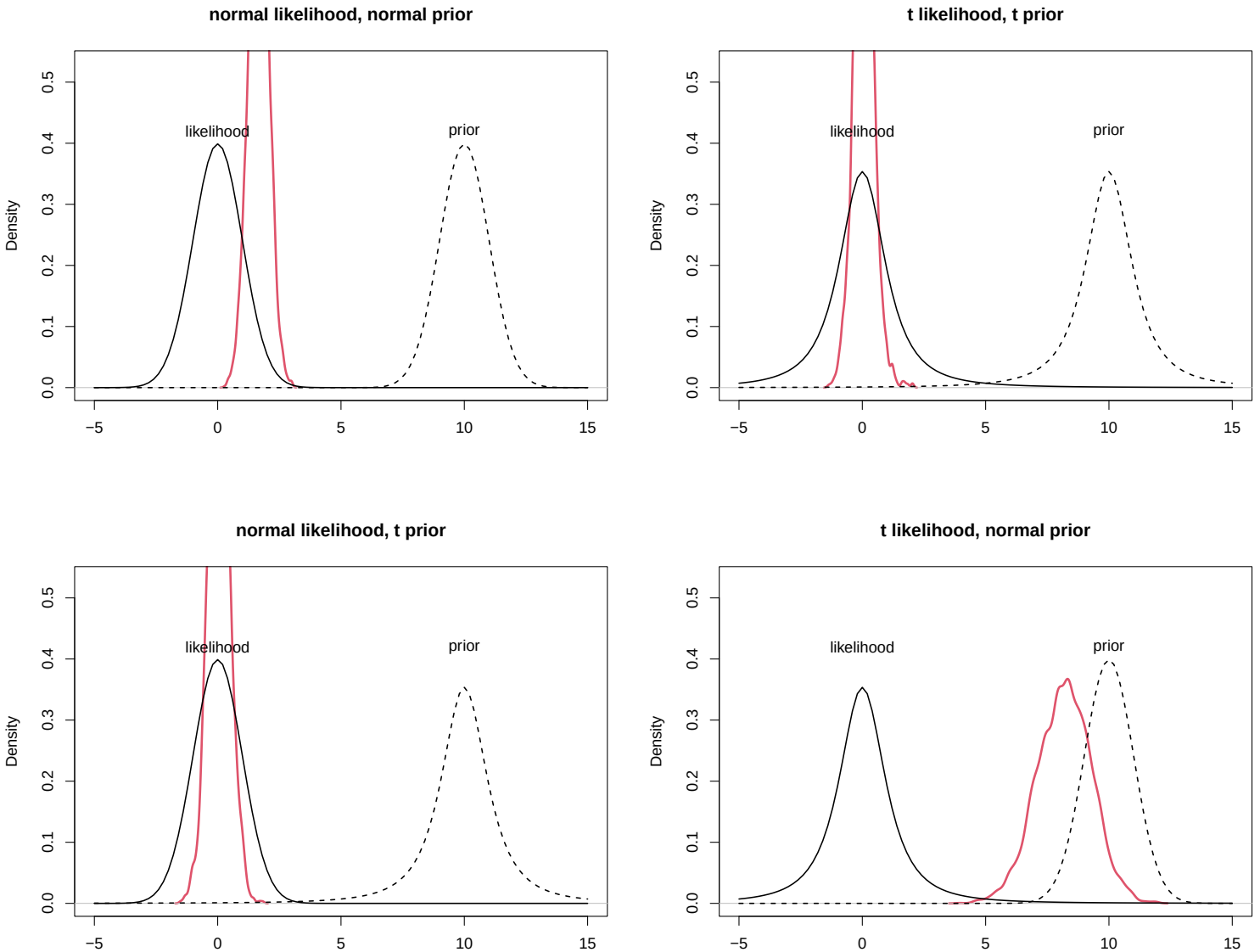

normal likelihood, t prior
likelihood
prior
Density
0.0
0.1
0.2
0.3
0.4
0.5
−5
0
5
10
15


t likelihood, normal prior
likelihood
prior
Density
0.0
0.1
0.2
0.3
0.4
0.5
−5
0
5
10
15

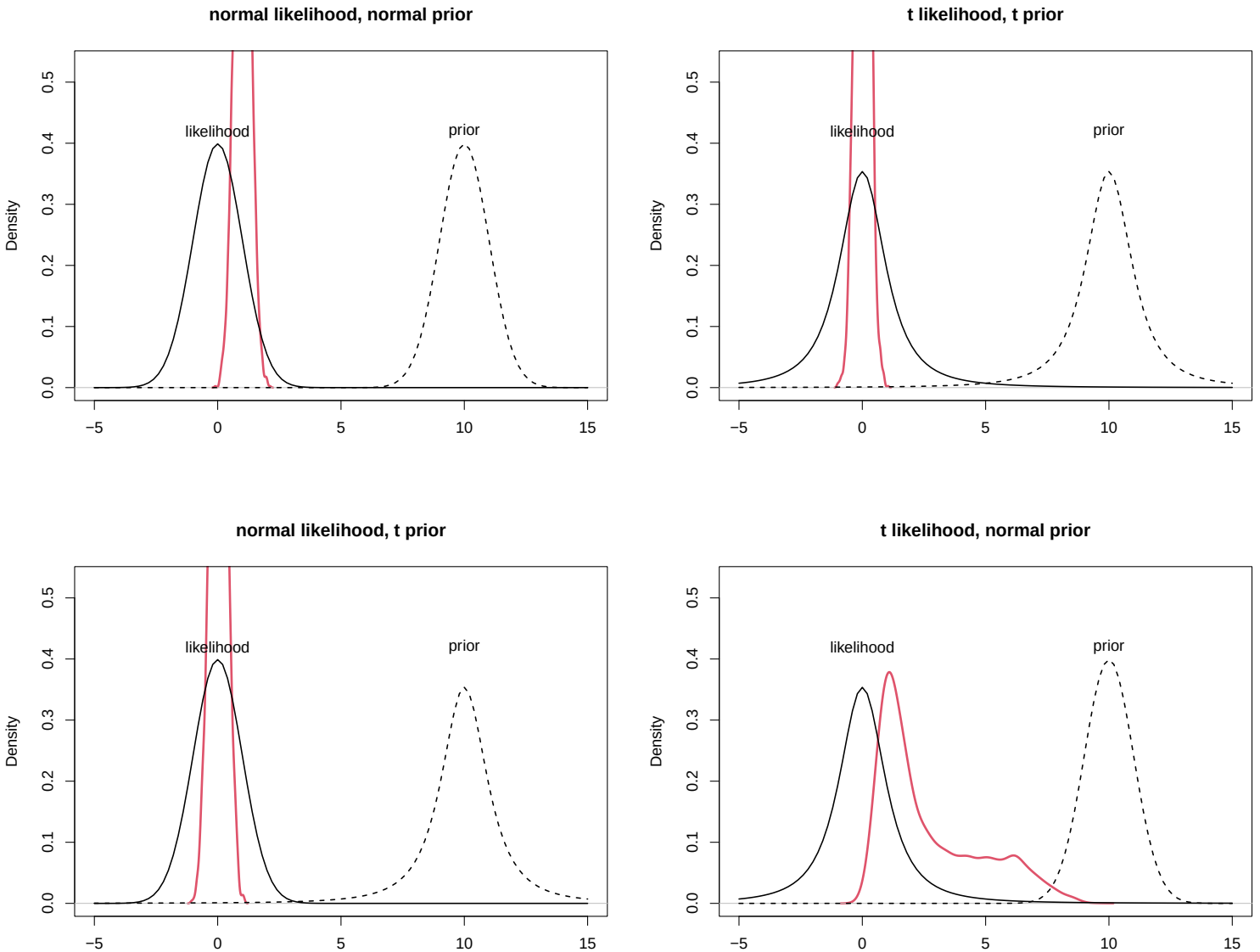
What if we have two data points $y = (0, 0, ..., 0)$



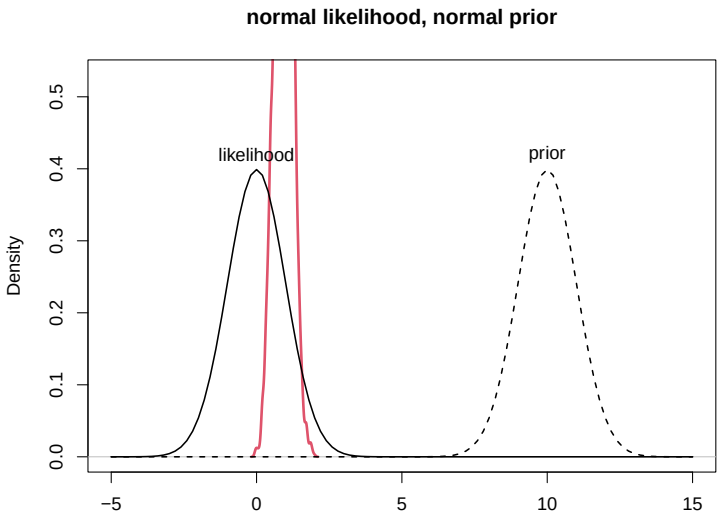
What if we have five data points $y = (0, 0, ..., 0)$

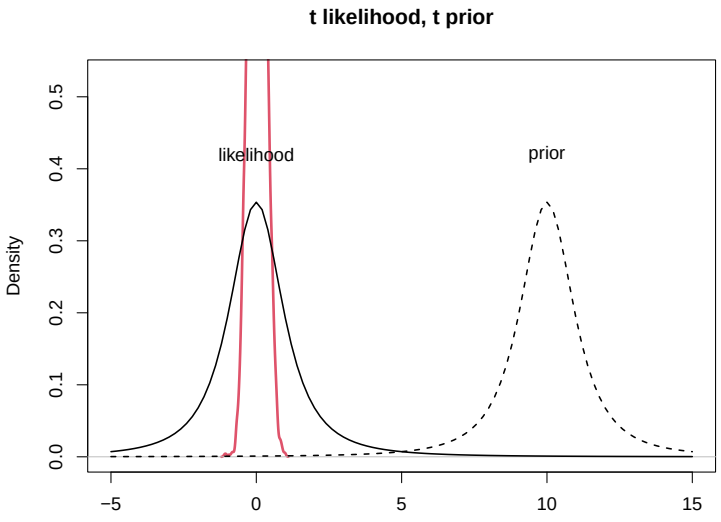


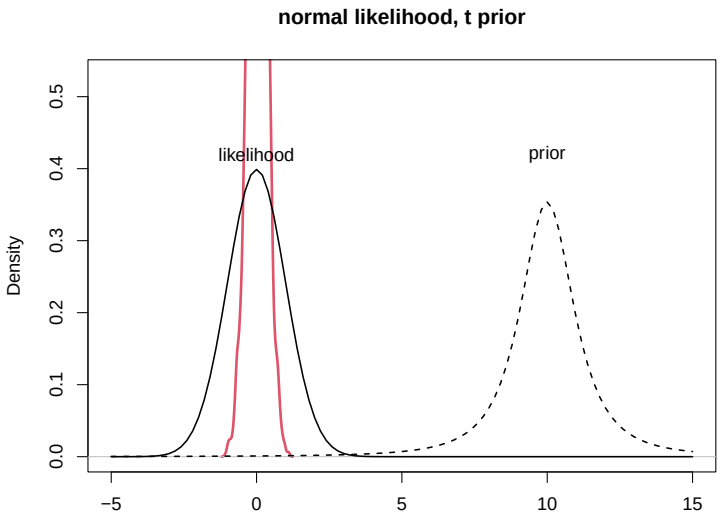
What if we have nine data points $y = (0, 0, ..., 0)$

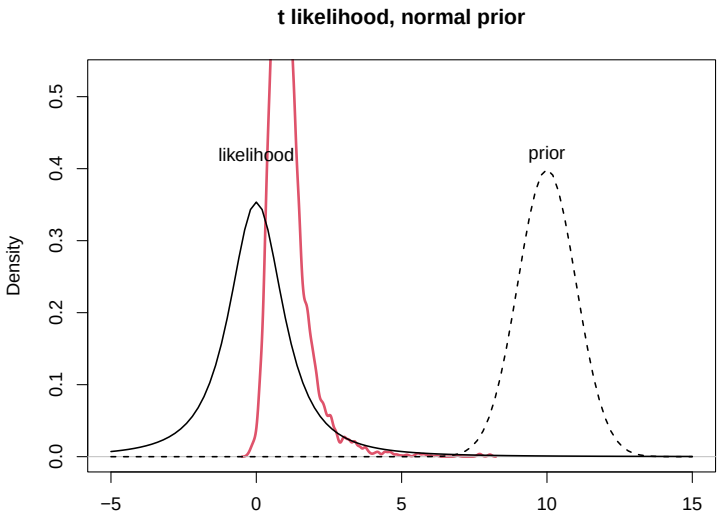


What if we have nine data points $y = (0, 0, ..., 0)$





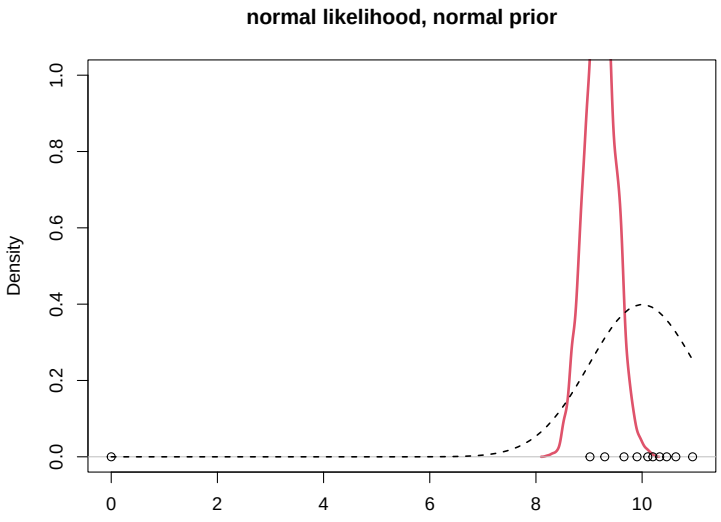


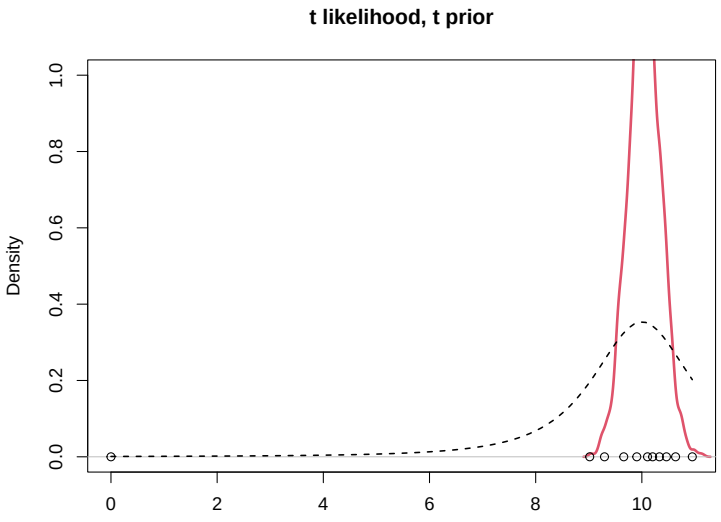


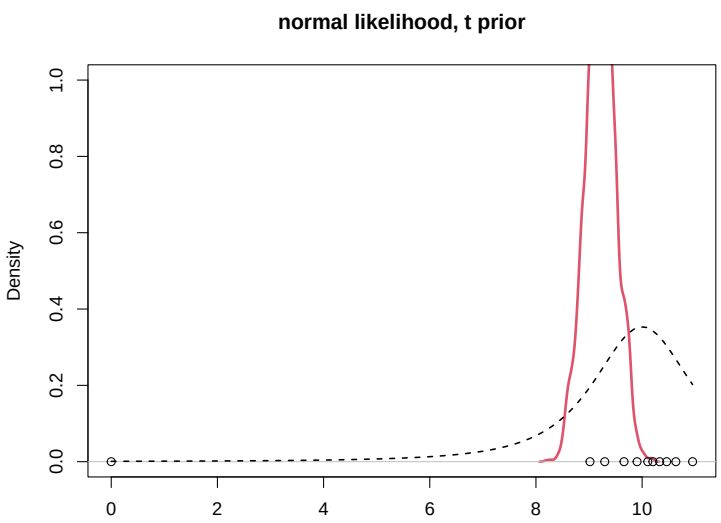
A less toy example

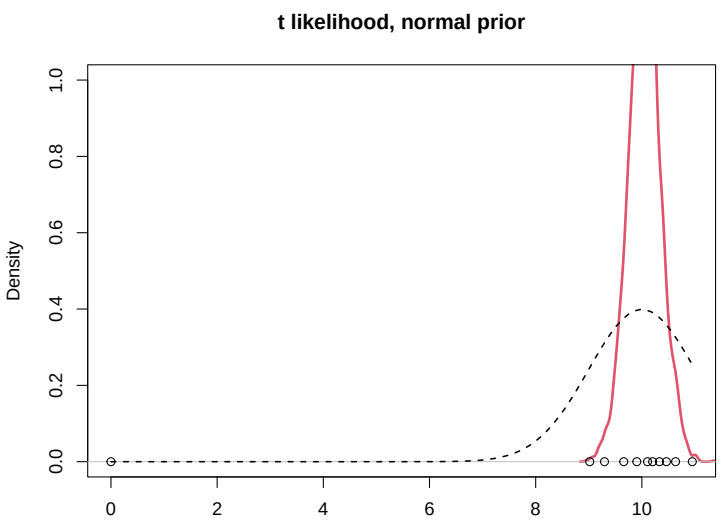
- The previous example considers two types of beliefs encoded in the model:
  - Can the data be a long way from the mean?
  - Can the mean be a long way from where we expect it to be (10)?
- It also illustrates that given a single surprising observation, Bayesian inference can sensibly 'reject' the information from either prior or likelihood, if one of them have a heavy tail.
- Now consider a more realistic situation: $y_i \sim N(10, 1)$ for $i = 1, ..., 10$, and $y_{11} = m \ll 10$ is an outlier.
- How would the posterior distribution of μ look like under the previous four cases?

$m = 0$

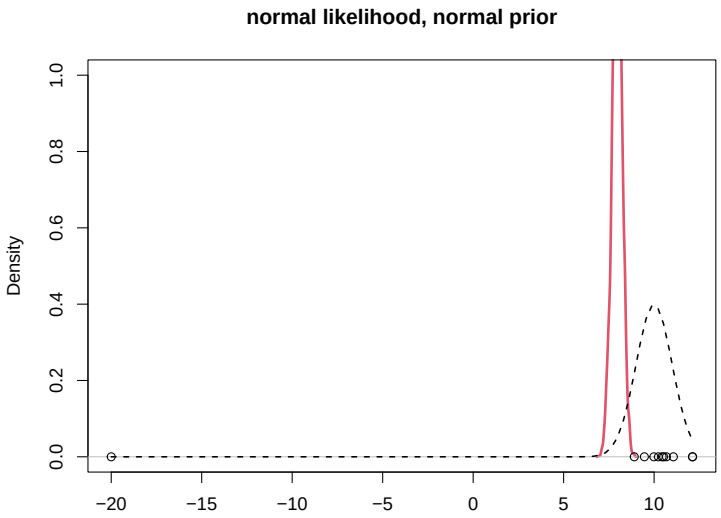

normal likelihood, normal prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

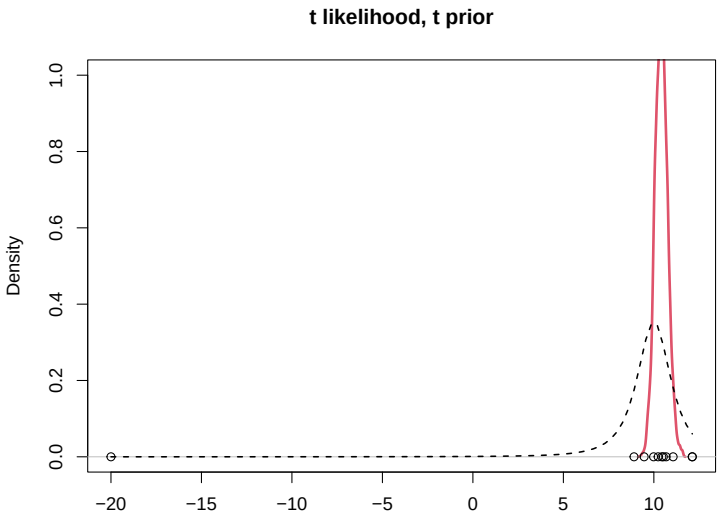

t likelihood, t prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

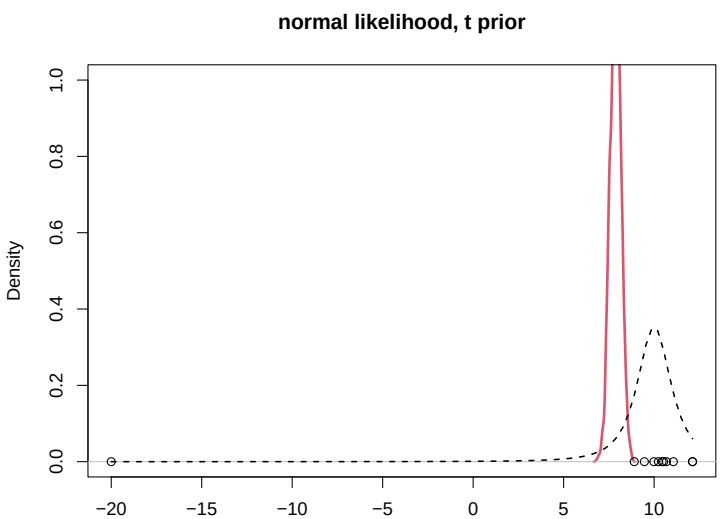

normal likelihood, t prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

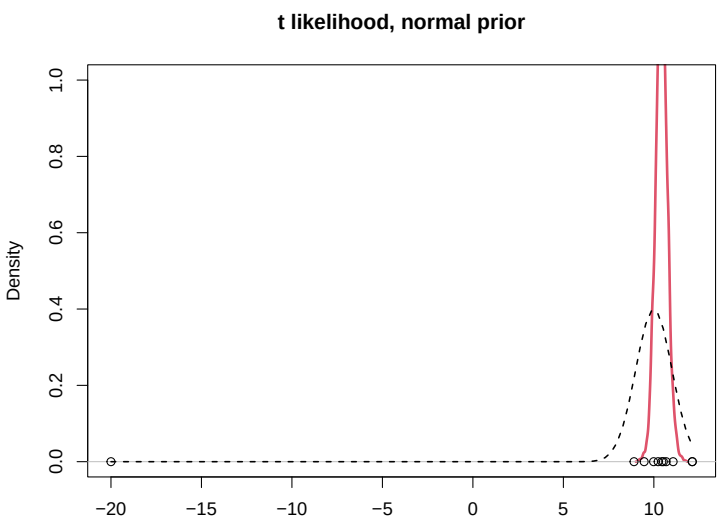

t likelihood, normal prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

$m = -20$

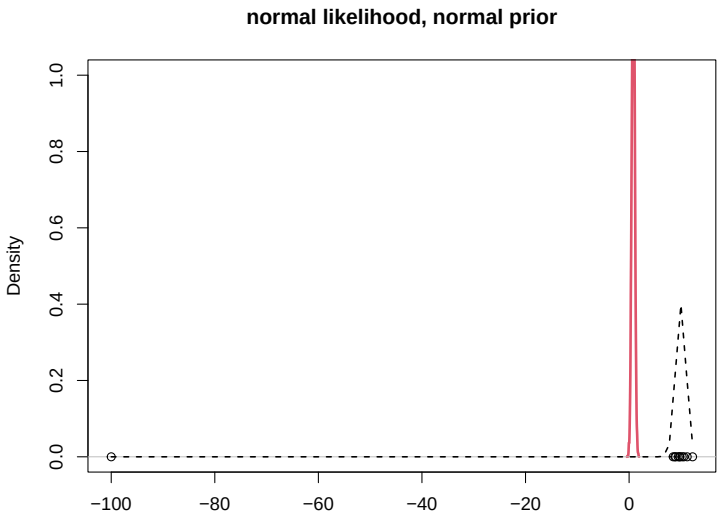

normal likelihood, normal prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−20
−15
−10
−5
0
5
10

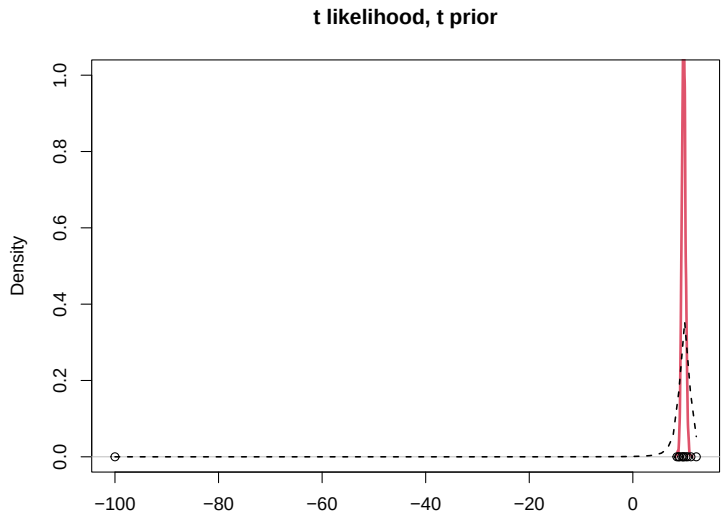

t likelihood, t prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−20
−15
−10
−5
0
5
10

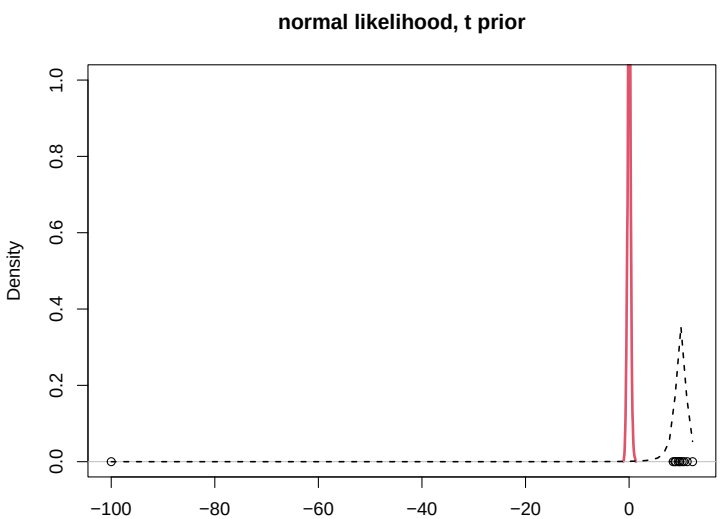

normal likelihood, t prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−20
−15
−10
−5
0
5
10

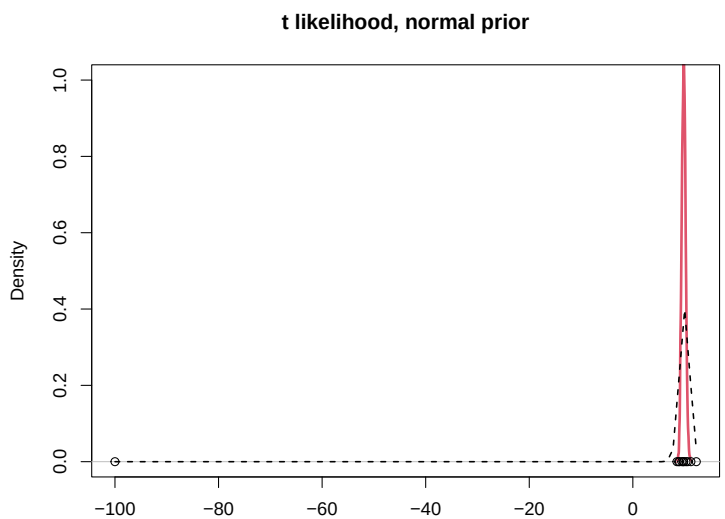

t likelihood, normal prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−20
−15
−10
−5
0
5
10

$m = -100$


normal likelihood, normal prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−100
−80
−60
−40
−20
0


t likelihood, t prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−100
−80
−60
−40
−20
0


normal likelihood, t prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−100
−80
−60
−40
−20
0


t likelihood, normal prior
Density
0.0
0.2
0.4
0.6
0.8
1.0
−100
−80
−60
−40
−20
0

A regression example

- Now consider something more realistic, for a regression model

$$y_i = \sum_j x_{ij}\beta_j + \epsilon_i$$

  Previously we usually let $\beta_j \sim N(\mu, \tau^2)$ and $\epsilon_i \sim N(0, \sigma^2)$.
- For fixed effects, it is usually fine to use a normal prior, provided that x_{ij} has been rescaled so that the prior on β_j is reasonable.
- What about the normality assumption on the noise term?

Simulation

```
N <- 50
beta0 <- 0.5
x1 <- rnorm(N)
y <- x1 * beta0 + rnorm(N, sd = 1)
standata <- list(N = N, x1 = x1, y = y)
y.out1 <- x1 * beta0 + rt(N, df = 3)
standata.out1 <- list(N = N, x1 = x1, y = y.out1)
y.out2 <- x1 * beta0 + rt(N, df = 2)
standata.out2 <- list(N = N, x1 = x1, y = y.out2)
y.out3 <- x1 * beta0 + rt(N, df = 1)
standata.out3 <- list(N = N, x1 = x1, y = y.out3)
```

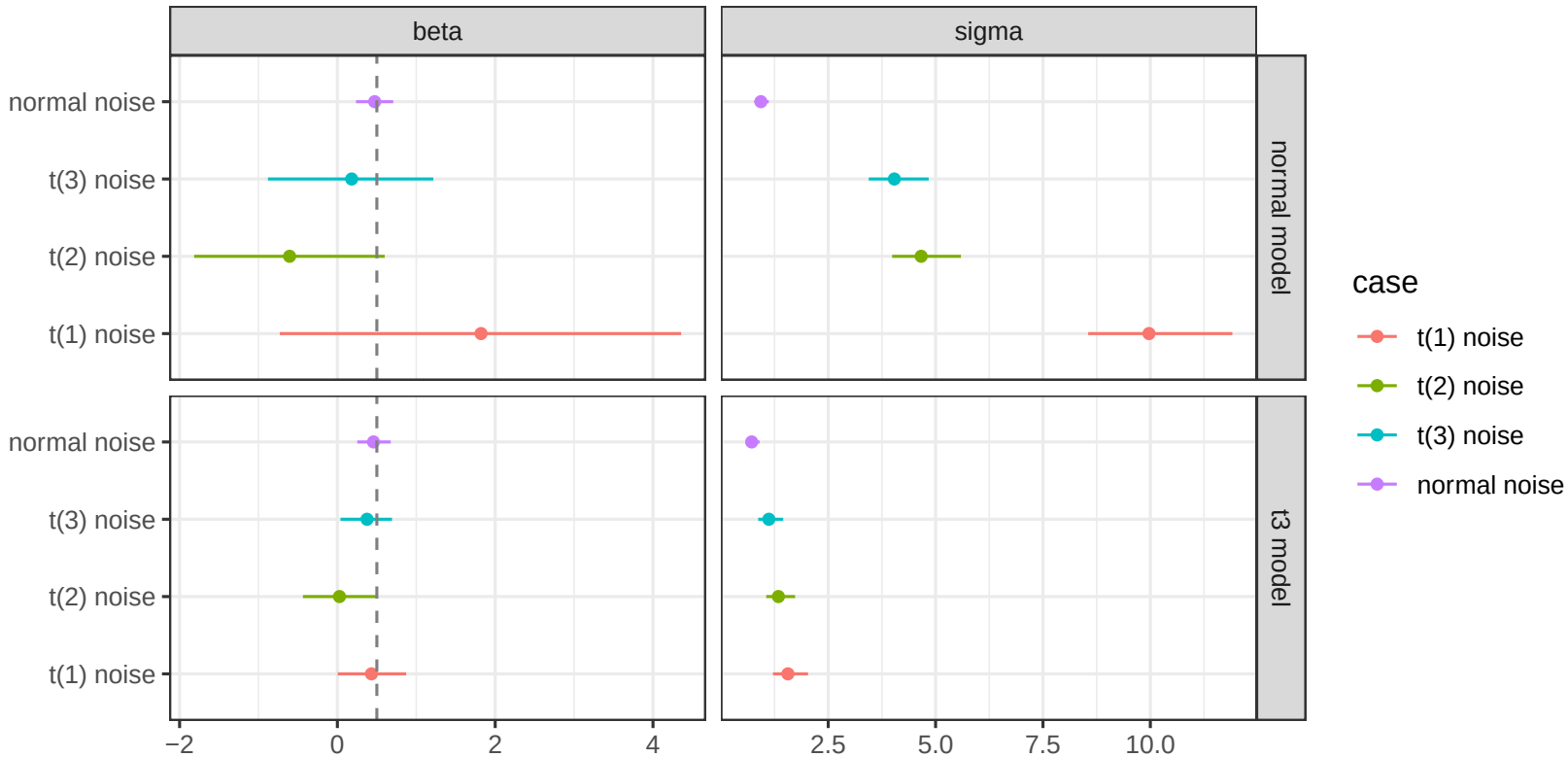
Simulation: normal error model

```
normal_model <- " data {
      int N;            // number of observations
      vector[N] x1;      // covariate
      vector[N] y;       // outcome
    }
    parameters {
      real beta;  // fixed effects
      real<lower=0> sigma;  // sd of y
    }
    model {
      real yhat;
      beta ~ normal(0, 1000);
      for(i in 1:N){
        y[i] ~ normal(beta * x1[i], sigma);
      }
    }"
 fit.stan <- stan(model_code = normal_model,
             data = standata,
             iter=4000, chains = 4)
```

Simulation: t_2 error model

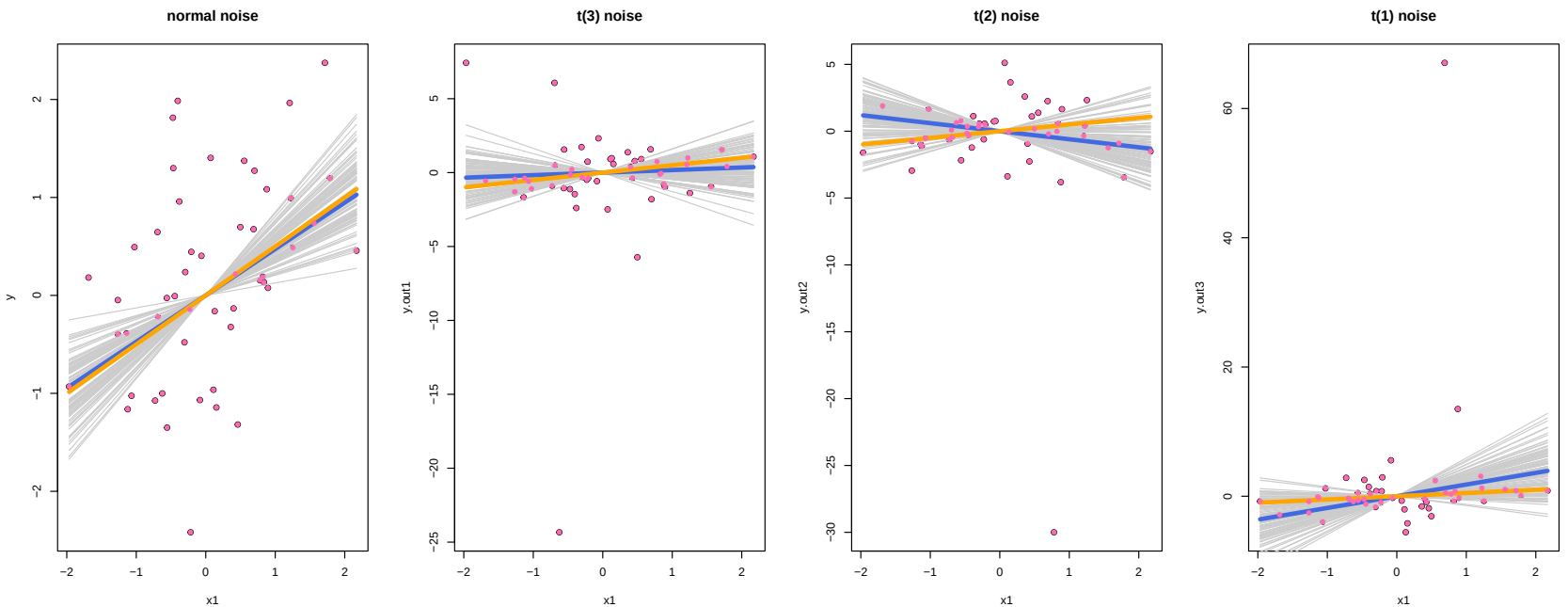
```
t2_model <- " data {
          int N;            // number of observations
          vector[N] x1;      // covariate
          vector[N] y;       // outcome
        }
        parameters {
          real beta;  // fixed effects
          real<lower=0> sigma;  // sd of y
        }
        model {
          real yhat;
          beta ~ normal(0, 1000);
          for(i in 1:N){
            y[i] ~ student_t(3, beta * x1[i], sigma);
          }
        }"

fit.stan <- stan(model_code = t2_model,
                 data = standata,
                 iter=4000, chains = 4)
```

Parameter estimation



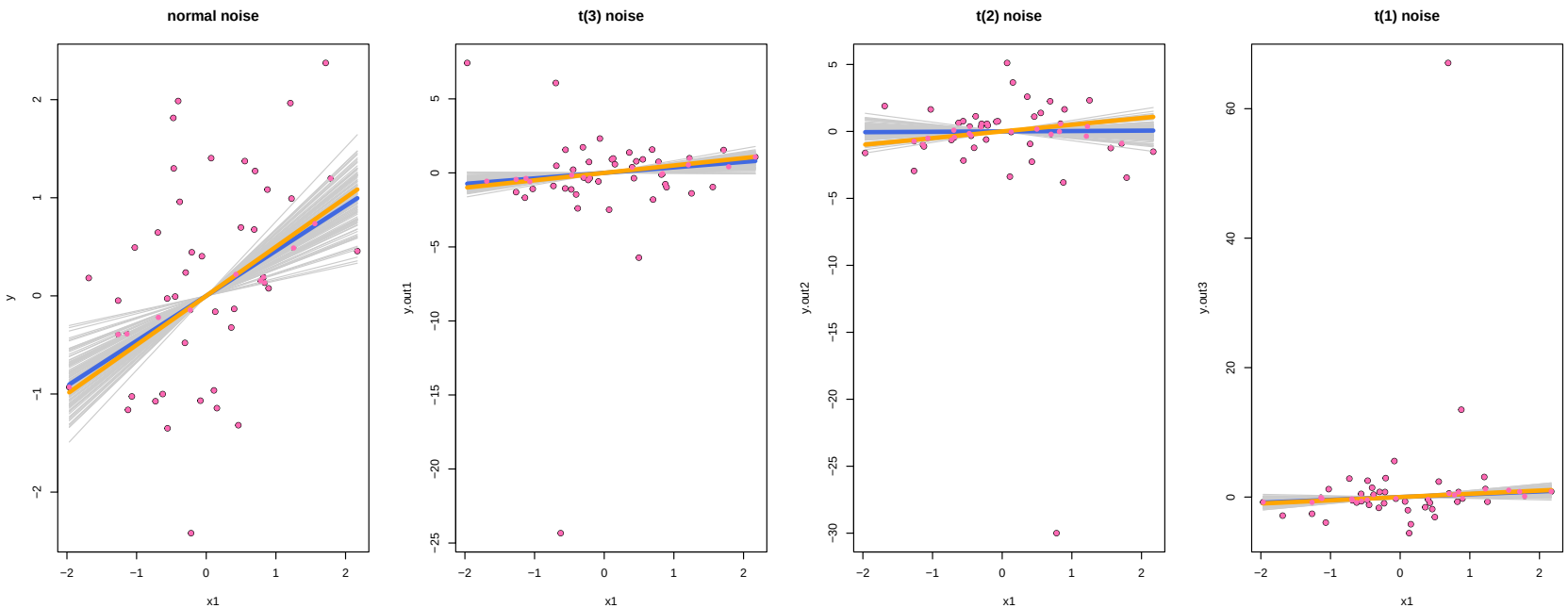
Posterior draws and posterior mean from the normal model

Blue is estimated. Orange is truth



Posterior draws and posterior mean from the t model

Blue is estimated. Orange is truth



Posterior inference

- Parameter inference under the t-distribution model can be easily implemented through data augmentation. $x \sim t(\nu, \mu, \sigma)$ can be equivalently expressed as

$$x \mid v \sim N(\mu, s^2)$$

$$s^2 \sim \text{InvGamma}(\frac{\nu}{2}, \frac{\nu\sigma^2}{2})$$

or

$$x \mid v \sim N(\mu, s^2\sigma^2)$$

$$s^2 \sim \text{InvGamma}(\frac{\nu}{2}, \frac{\nu}{2})$$

- So the regression model with a t-distribution error, $\epsilon_i \sim_{iid} t(v, 0, \sigma)$, can be expressed as the hierarchical form, with prior choices for β and σ^2,

$$y_i \sim N(\sum_j x_{ij}\beta_j, s_i^2\sigma^2)$$

$$s_i^2 \sim \text{InvGamma}(\frac{\nu}{2}, \frac{\nu}{2}), \quad i = 1, ..., n$$

$$\beta_j \sim N(0, \tau^2), \quad j = 1, ..., p$$

$$p(\sigma^2) \propto 1$$

Posterior inference

- The full posterior conditionals are

$$\boldsymbol{\beta} \mid \cdot \sim N((\frac{1}{\tau^2}\boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{y}, (\frac{1}{\tau^2}\boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X})^{-1})$$

$$s_i^2 \mid \cdot \sim \text{InvGamma}(\frac{\nu+1}{2}, \frac{\nu + (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2/\sigma^2}{2})$$

$$\sigma^2 \mid \cdot \sim \text{InvGamma}(\frac{n}{2} + 1, \frac{1}{2}\sum_i \frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{s_i^2})$$

$p(v \mid -)$ has no closed form, and need M-H step if treating as random

A random effect regression example

- The previous example makes the error distribution robust to outliers.
- Now consider random effects that share information across the β's. For easier notation, let's focus on the random intercept only model

$$y_{ij} = \beta_i + \epsilon_{ij}$$

  where $\beta_i \sim N(\mu, \tau^2)$.
- We may also want to robustify the prior for β_i if we believe there are intercepts that are far away from others.
- That is, let $\beta_i \sim t(\nu, \mu, \tau^2)$ instead of the normal.
- Data augmentation can be carried out in the same way as before.
- We will omit details here, but see the textbook Chapter 17.4 for this model applied to the eight schools example.

Case Study 2: overdispersion and model expansion

- Using the t-distribution as alternative to normal model can also be considered as introducing overdispersion to the model.
- Overdispersion means the variability in the data is larger than expected by the model.
- Another example as we have seen before is the binomial distribution, $y_i \mid p \sim \text{Bin}(n_i, p)$, which leads to $\mathbb{E}(y_i \mid p) = n_i p$ and $\text{var}(y_i \mid p) = n_i p(1-p)$.
- The mean-variance relationship is determined by the assumed binomial likelihood.
- The binomial model assumes individual outcomes are independent Bernoulli. When there are clustering of units within the data, e.g., consider C_i clusters of size K_i within group i,

$$y_{ic} \mid p_{ic} \sim \text{Bin}(k_i, p_{ic})$$

  where $y_i = \sum_c y_{ic}$, and $\mathbb{E}(p_{ic}) = p$.
- We can show that the count of successes has a larger variance than $n_i p(1-p)$ using the law of total variances.
- Similarly, one can show that overdispersion can arise in other likelihood models such as Poisson and exponential.

Overdispersed binomial

- One straightforward way to introduce overdispersion in binomial data is the beta-binomial model

$$y_i \mid q_i \sim \text{Bin}(n_i, q_i), \quad q_i \sim \text{Beta}(a, b)$$

  or equivalently

$$y_i \mid p, d \sim \text{BetaBin}(n_i, p, d)$$

  where $p = a/(a+b)$ and $d = \frac{1}{a+b+1}$ as we have seen before. We then have $\text{var}(y_i) = np(1-p)(1+(n-1)d)$.

- Similarly other hierarchical models can introduce overdispersion too, e.g., in a logit model,

$$y_i \mid q_i \sim \text{Bin}(n_i, q_i), \quad \text{logit}(q_i) = \mu_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Overdispersed Poisson

- For a Poisson model $y_i \sim \text{Poi}(\lambda_i)$, a common overdispersion model is

$$y_i \mid \theta_i \sim \text{Poi}(\theta_i), \quad \theta_i \sim \text{Gamma}(a_i, b_i)$$

or equivalently

$$y_i \mid a_i, b_i \sim \text{NegBin}(a_i, b_i)$$

where $\mathbb{E}(y_i) = a_i/b_i = \lambda_i$ and $\text{var}(y_i) = \frac{b_i+1}{b_i}\frac{a_i}{b_i} > \mathbb{E}(y_i)$.

Simulate overdispersed binomial counts

```
N <- 100
n <- sample(10:100, N, replace=TRUE)
p <- 0.5
d <- 0.7 # overdispersion parameter
a <- p * (1-d)/d
b <- (p * d - p - d + 1)/d
q <- rbeta(N, a, b)
y <- rbinom(N, n, q)
```

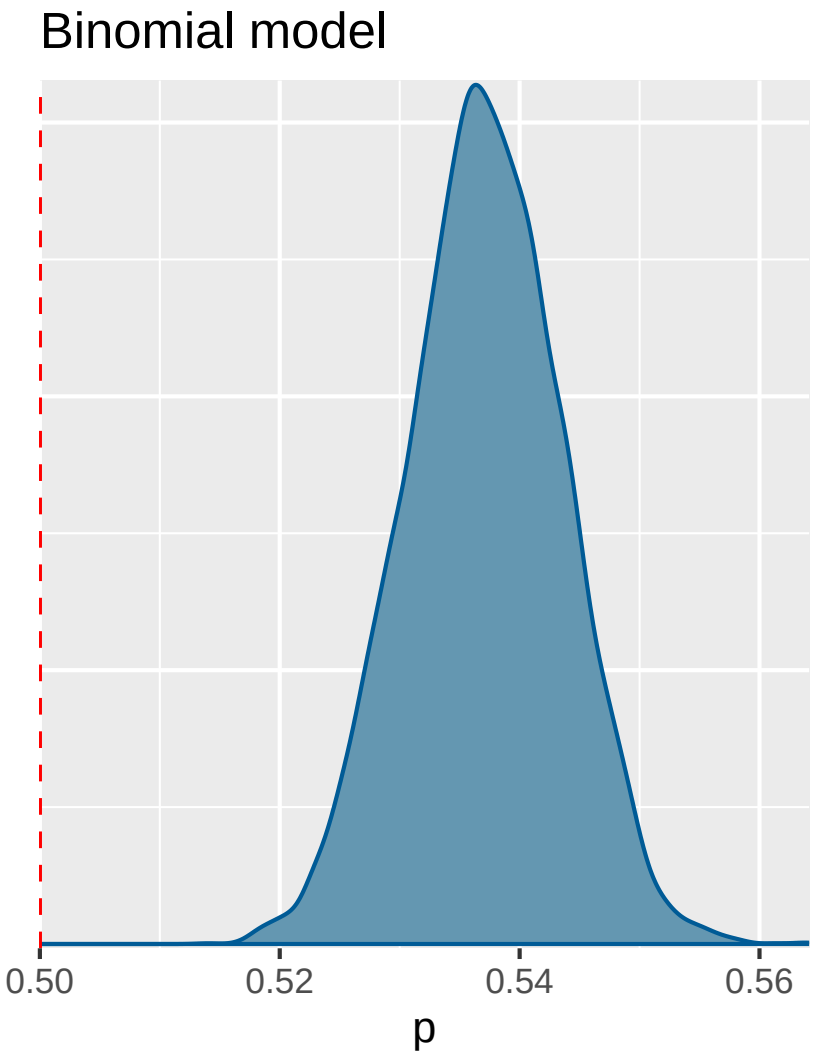
Binomial model

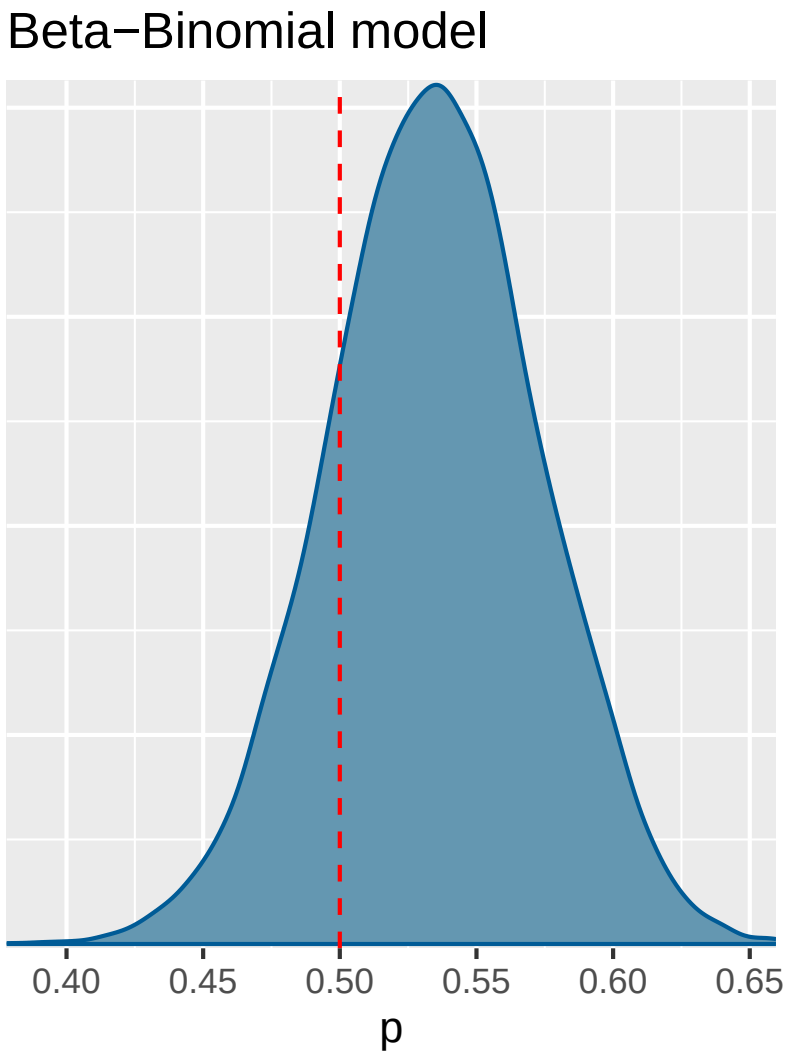
```
bin_model <- " data {
   int N;
   int y[N];
   int n[N];
 }
 parameters {
   real<lower=0, upper=1> p;
 }
 model {
   p ~ beta(1, 1);
   for(i in 1:N){
     y[i] ~ binomial(n[i], p);
   }
 }"
```

Beta Binomial model

```
betabin_model <- " data {
   int N;
   int y[N];
   int n[N];
 }
 parameters {
   real logit_d;
   real<lower=0, upper=1> p;
 }
 transformed parameters {
     real<lower=0, upper=1> d = exp(logit_d) / (1 + exp(logit_d));
 }
 model {
   p ~ beta(1, 1);
   logit_d ~ normal(0, 0.5);
   for(i in 1:N){
     y[i] ~ beta_binomial(n[i], p * (1-d)/d, (p * d - p - d + 1)/d);
   }
 }"
```

Parameter estimation





Model misspecification

- Another major class of methods that we did not touch upon is the robustness to model misspecification.
- For example, the MLE theory tells us that when the true model is not in the model class we consider, we can define the pseudo-true parameter $\theta^\star$ such that $P_{\theta^\star}$ is the closet distribution in our model space to the true model. Then the MLE $\hat{\theta}$ can be consistent for $\theta^\star$.
- And the limiting distribution of MLE is normal with a sandwich variance, derived from estimating equations.
- The Bayesian analogy for such general model misspecification is not very straightforward, and has been a major area of research. More references on Canvas homepage.
- Lastly, sometimes we may challenge whether we should handle "robustness" within the model. For example, outliers may not always be a concern for the model: Maybe it means we need to clean the data better!