Things that can go wrong

MATH 66356

November 27, 2016 at 22:26

Contents

Summary of asymptotics **1**

Common examples where things fail **1**
Underidentified models and nonidentified parameters 2
Number of parameters increasing with sample size 2
Aliasing 2
Unbounded likelihoods 2

Improper posterior distribution **9**

Prior distributions that exclude the point of convergence **9**

Convergence to the edge of the parameter space **9**

Tails of the distribution **9**

To run these examples it is assumed that you have already successfully installed STAN and R packages associated with it.
See RStan Getting Started.

Summary of asymptotics

As $n \to \infty$, the posterior distribution, $p(\theta|y)$ is well behaved under general conditions:

1. θ_{true} is in the interior of
$$\Omega \subset \mathbb{R}^p$$
$ where Ω is the set of possible values of θ.
2. The model for an individual observation, $p(y_i|\theta)$ is the same for each i. Although this is sufficient, it is rarely exactly satisfied in practice. There are better statements of this condition.
3. $p(y|\theta)$ is continuous in θ

If the model is valid, i.e. the true distribution of y is in the model given by $p(y|\theta_{true})$ and if the proper prior has positive density at θ_{true} then the posterior mode $\hat{\theta}$ converges in distribution to θ_{true} and is asymptotically normal with θ_{true} and variance $I(\hat{\theta})^{-1}$ which itself converges to $I(\theta_{true})^{-1}$.

So, for large n, Bayesian and frequentist approaches coincide and the posterior distribution can be summarized by the posterior mean and variance.

Common examples where things fail

Largely taken from pp 89–91 of BDA3.

Underidentified models and nonidentified parameters

Consider estimating the distribution of treatment effects with a randomized experiment.

Suppose the response of subject i to treatment $t = 1, 2$ is modeled as:

$$y_{it} = \phi_{it} + \epsilon_{it}$$

where ϕ_{it} is the expected response of subject i to treatment t. The expected effect of treatment on subject i is $\delta_i = \phi_{i2} - \phi_{i1}$ and $\epsilon_{it} \sim N(0, \sigma_\epsilon)$ is the variability in response under fixed conditions.

In a randomized experiment in which each subject is observed under only one of treatments 1 or 2, we can estimate the expected value of δ_i, the mean treatment effect, but we are very limited in the estimation of its distribution. Thus, although we can tell a subject the 'expected benefit' of a treament, we cannot readily estimate the probability that one treatment will be superior to the other.

Number of parameters increasing with sample size

For example if, in hierarchical data, the number of clusters increases – thus increasing n – but the cluster size has a maximum, then the larger n will not result is an asymptotically vanishing standard deviation for the estimate of a cluster mean if the number of observations within that cluster is fixed.

Aliasing

Lack of identifiability can happen in situation that may be difficult to see. For example in a regression, three predictors could be collinear although any pair are not collinear. The regression as a whole will not be estimable although, pairwise, the predictors are not collinear.

Unbounded likelihoods

This is not uncommon in hierarchical models if the variability between clusters is small relative to what it would be expected to be compared with the within cluster variability.

Consider the eight school example from BDA3.

```
# Original from BDA3
schools_dat <- list(J = 8,
                    y = c(28,  8, -3,  7, -1,  1, 18, 12),
                    sigma = c(15, 10, 16, 11,  9, 11, 10, 18))
```

The marginal variance of the y's obeys:

$$var(y) = E(var(y_i|i)) + var(E(y_i|i)) = \frac{\sum \sigma_i^2}{J} + \tau^2$$

In this case the observed variance of y is 109.07 and the mean variance of the individual y's is 166.

If the variance of y's were less that the mean of the sigma squared, the y's would be said to be *underdispersed*. An extreme case would suggest that the y's could not have been generated by the process that is modeled.

Using mixed models to fit highly underdispersed data usually results in convergence problems because the likelihood is unbounded as the between cluster variance, τ^2 approaches 0.

We can use the 'eight schools' data to explore what happens with HMC by increasing 'sigma' in the data to make the data underdispersed. Surprisingly, the posterior, in this case, does not produce a large number of

divergent transitions associated with small estimated values of τ. Curiously, with considerable underdispersion, the posterior for τ seems to suggest larger values.

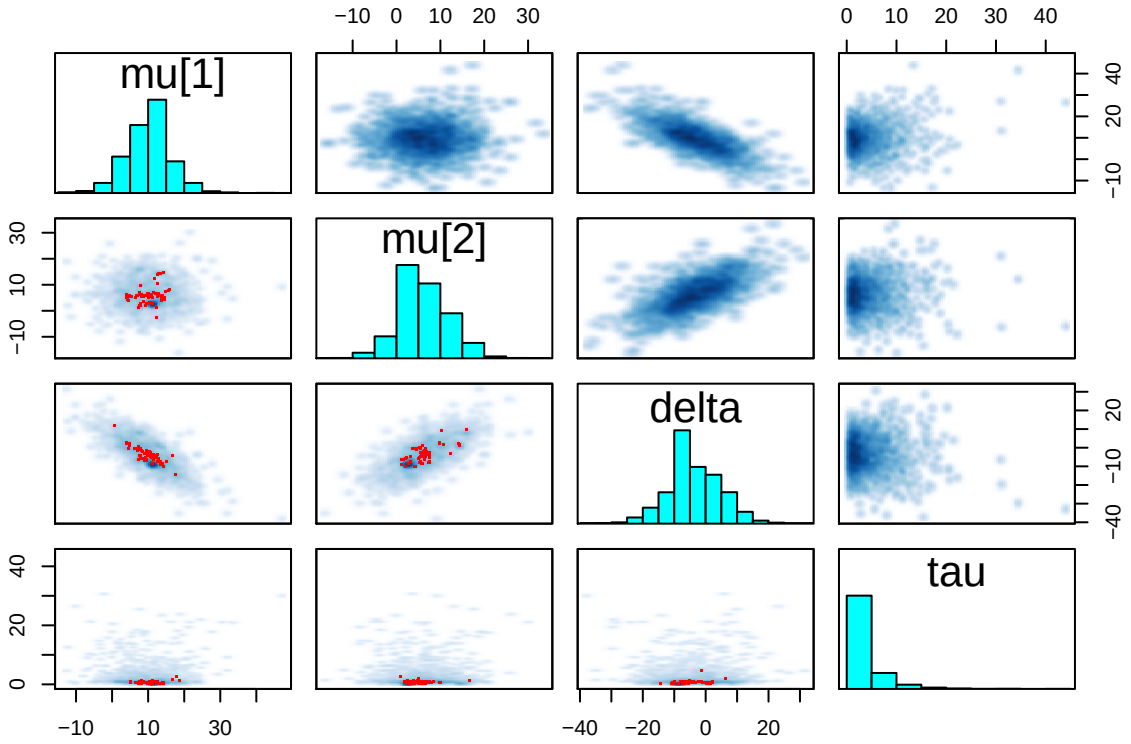
This seems to suggest that fitting models that 'don't make sense' can produce results that make even less sense. The lesson is that one must try to understand the data and the processes generating it well and not rely on advanced modeling methods to correct, or even provide a warning, when the model and data don't accord.

Let's add two types so we can estimate the difference between school types.

```
schools_dat$type <- c(1,1,1,1,2,2,2,2)

code <- ('
  data{
    int J;
    real y[J];
    real sigma[J];
    int type[J];
  }
  parameters{
    real mu[2];
    real u[J];
    real log_tau;
  }
  transformed parameters{
    real delta;
    real tau;
    delta = mu[2] - mu[1];
    tau = exp(log_tau);
  }
  model{
    for(j in 1:J) u[j] ~ normal(0,tau);
    for(j in 1:J) y[j] ~ normal(mu[type[j]]+u[j],sigma[j]);
  }
')
system.time(
  fit <- stan(model_code = code,
              data = schools_dat,
              iter = 2012,
              chains = 4,
              verbose = verbose)
)
```

```
    Warning: There were 72 divergent transitions after warmup. Increasing adapt_delta above 0.8 may hel
    http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup

    Warning: Examine the pairs() plot to diagnose sampling problems

       user  system elapsed
       1.81    0.58   82.12
```

```
pairs(fit,pars=c('mu','delta','tau'))
```




```
data_small_sigma <- schools_dat
data_small_sigma$sigma <- schools_dat$sigma*0.2
system.time(
  fit2 <- stan(model_code = code,
               data = data_small_sigma,
               iter = 2012,
               chains = 4,
               verbose = verbose)
)
```

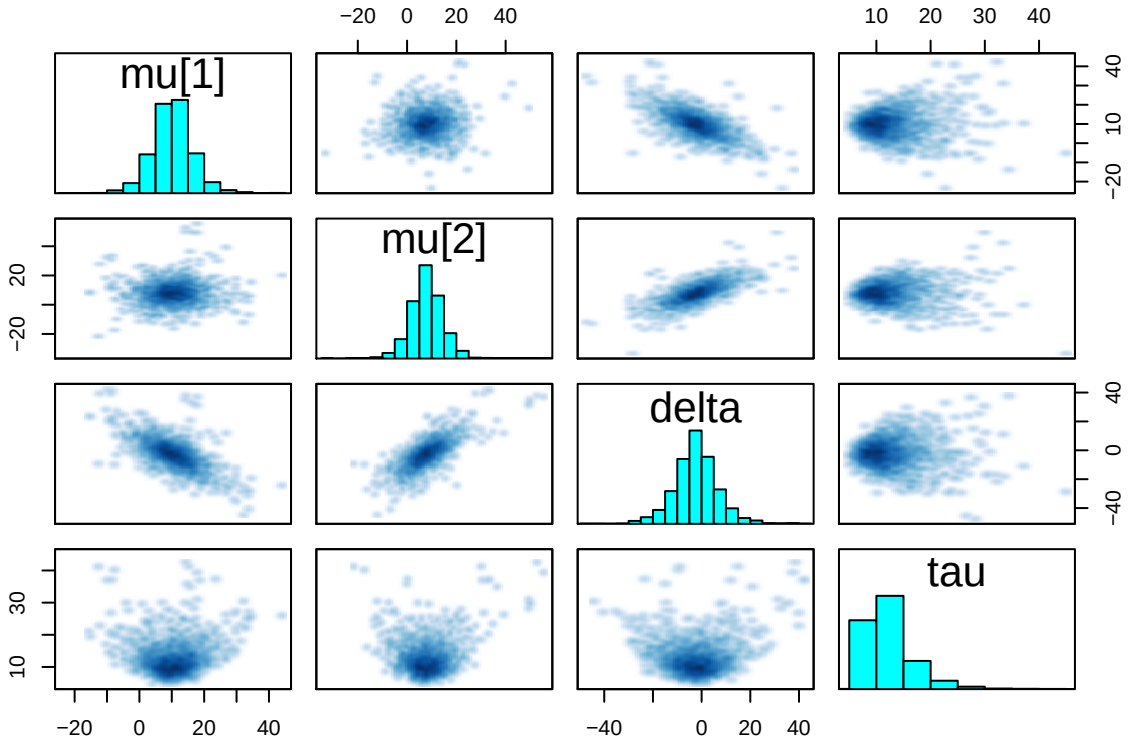
```
   user  system elapsed
   0.92    0.59   68.47
```

```
fit2
```

```
Inference for Stan model: 912c2165fec2071ed3d7a74be48deb36.
4 chains, each with iter=2012; warmup=1006; thin=1;
post-warmup draws per chain=1006, total post-warmup draws=4024.

          mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
mu[1]    10.19    0.21 6.28  -2.65   6.48  10.17  13.78  23.05   878    1
mu[2]     7.96    0.22 6.83  -5.43   4.06   7.80  11.83  21.45   935    1
u[1]     16.46    0.21 6.59   3.81  12.48  16.30  20.24  30.22   983    1
u[2]     -2.12    0.21 6.46 -15.26  -5.79  -2.09   1.65  10.75   931    1
u[3]    -12.11    0.22 6.70 -25.89 -15.94 -11.89  -7.98   1.01   946    1
u[4]     -3.12    0.21 6.46 -16.67  -6.86  -3.16   0.79   9.60   924    1
u[5]     -8.72    0.22 6.92 -22.39 -12.69  -8.53  -4.57   4.70   958    1
u[6]     -6.71    0.22 6.98 -20.35 -10.67  -6.54  -2.56   7.04   992    1
```

```
u[7]      9.72    0.22 6.92  -3.95   5.66   9.70  13.77  23.39   991    1
u[8]      3.55    0.23 7.17  -9.88  -0.77   3.64   7.64  17.73  1004    1
log_tau   2.45    0.01 0.33   1.88   2.23   2.42   2.64   3.17  1000    1
delta    -2.23    0.31 9.03 -21.04  -7.34  -2.21   2.77  15.96   850    1
tau      12.25    0.15 4.46   6.55   9.29  11.24  14.07  23.87   944    1
lp__    -27.46    0.07 2.47 -33.19 -28.87 -27.06 -25.68 -23.72  1144    1

Samples were drawn using NUTS(diag_e) at Sun Nov 27 22:29:32 2016.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

```
pairs(fit2,pars=c('mu','delta','tau'))
```



with a large sigma:

```
data_large_sigma <- schools_dat
data_large_sigma$sigma <- schools_dat$sigma*5
system.time(
  fit3 <- stan(model_code = code,
               data = data_large_sigma,
               iter = 2012,
               chains = 4,
               verbose = verbose)
)
```

```
Warning: There were 73 divergent transitions after warmup. Increasing adapt_delta above 0.8 may hel
http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
```

```
Warning: Examine the pairs() plot to diagnose sampling problems
```

```
   user  system elapsed
   0.47    0.33    6.02
```

```
fit3
```

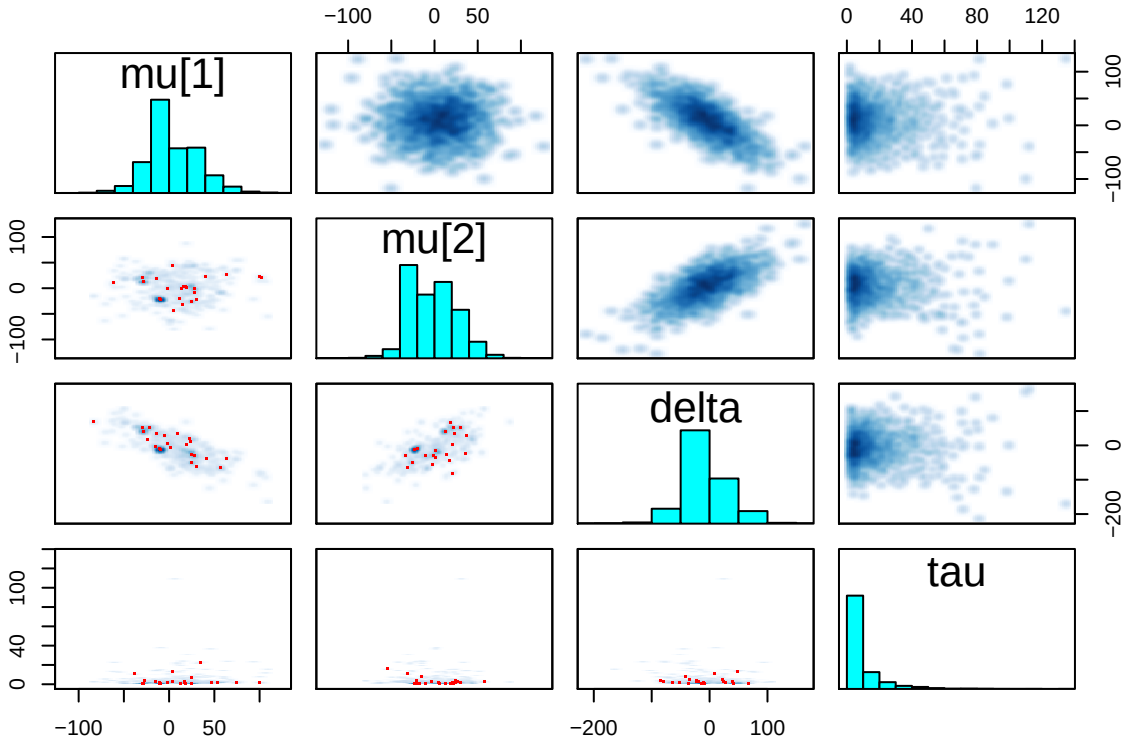
```
Inference for Stan model: 912c2165fec2071ed3d7a74be48deb36.
4 chains, each with iter=2012; warmup=1006; thin=1;
post-warmup draws per chain=1006, total post-warmup draws=4024.

          mean se_mean    sd   2.5%    25%    50%   75% 97.5% n_eff Rhat
mu[1]     4.00    7.39 28.25 -48.39 -10.43  -4.41 24.07 64.87    15 1.09
mu[2]    -1.24    9.36 26.29 -47.07 -22.67  -2.79 17.99 52.57     8 1.14
u[1]      0.61    0.21 11.73 -21.86  -1.12  -0.35  1.78 28.85  3075 1.00
u[2]     -0.25    0.19 11.25 -25.18  -1.58  -0.14  1.64 23.37  3462 1.00
u[3]     -0.17    0.26 13.37 -29.19  -1.53   0.63  1.21 27.21  2675 1.00
u[4]     -0.11    0.21 13.05 -27.51  -1.46  -0.22  1.52 25.46  4024 1.00
u[5]     -0.05    0.24 12.01 -26.36  -1.74   0.05  2.08 25.29  2602 1.00
u[6]     -0.22    0.19 11.93 -26.98  -1.92   0.09  1.38 25.12  4024 1.00
u[7]      0.78    0.21 13.28 -23.43  -1.05  -0.54  1.73 31.00  4024 1.01
u[8]      0.29    0.20 12.68 -25.83  -1.73   0.47  1.71 27.76  4024 1.00
log_tau   1.10    0.73  1.42  -0.69  -0.21   1.06  2.28  3.74     4 1.62
delta    -5.24    5.51 36.73 -79.45 -21.35 -13.40 16.97 68.77    44 1.05
tau       7.79    3.21 11.99   0.50   0.81   2.90  9.80 42.10    14 1.11
lp__    -13.78    6.04 11.70 -35.45 -23.29 -13.47 -2.62  2.00     4 1.61

Samples were drawn using NUTS(diag_e) at Sun Nov 27 22:29:40 2016.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

```
pairs(fit3,pars=c('mu','delta','tau'))
```

```
Warning in KernSmooth::bkde2D(x, bandwidth = bandwidth, gridsize = nbin, :
Binning grid too coarse for current (small) bandwidth: consider increasing
'gridsize'
```

```
Warning in KernSmooth::bkde2D(x, bandwidth = bandwidth, gridsize = nbin, :
Binning grid too coarse for current (small) bandwidth: consider increasing
'gridsize'
```

```
Warning in KernSmooth::bkde2D(x, bandwidth = bandwidth, gridsize = nbin, :
Binning grid too coarse for current (small) bandwidth: consider increasing
'gridsize'
```



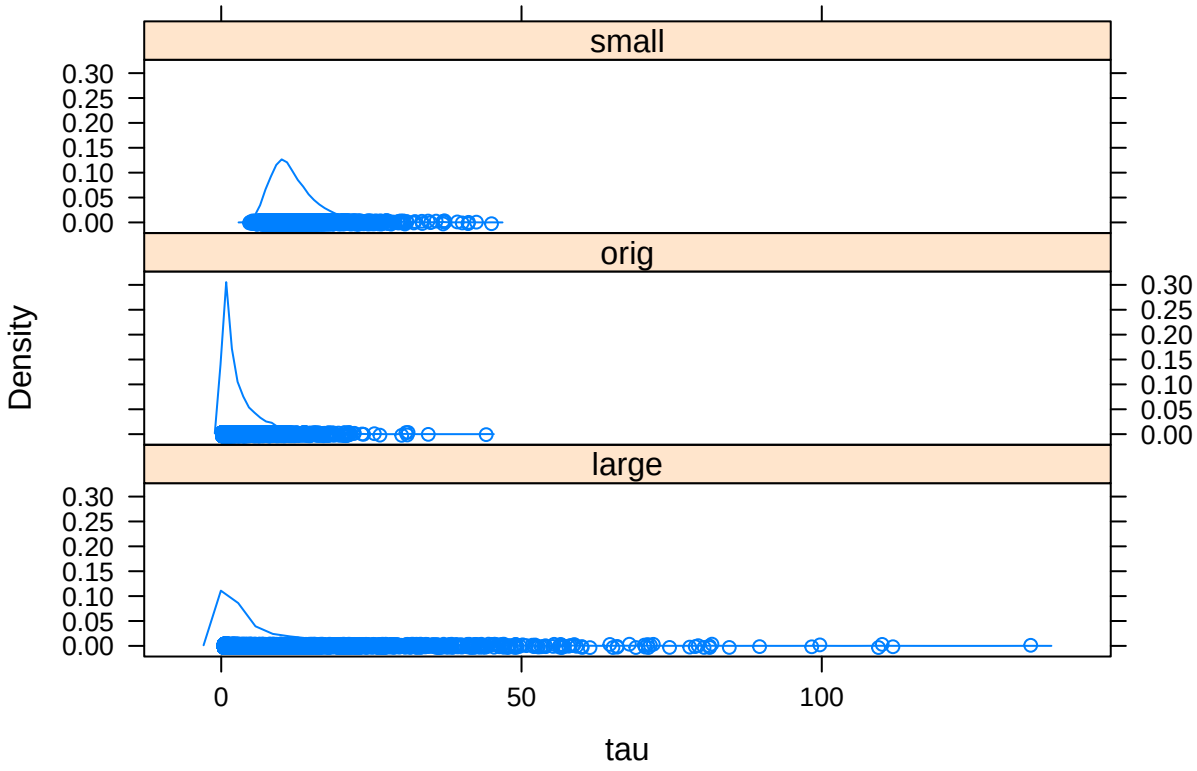

```
dd <- lapply(list(fit,fit2,fit3), rstan::extract) %>% lapply(as.data.frame)
taus <- sapply(dd, function(d) d$tau)
length(taus)/3
```

```
[1] 4024
```

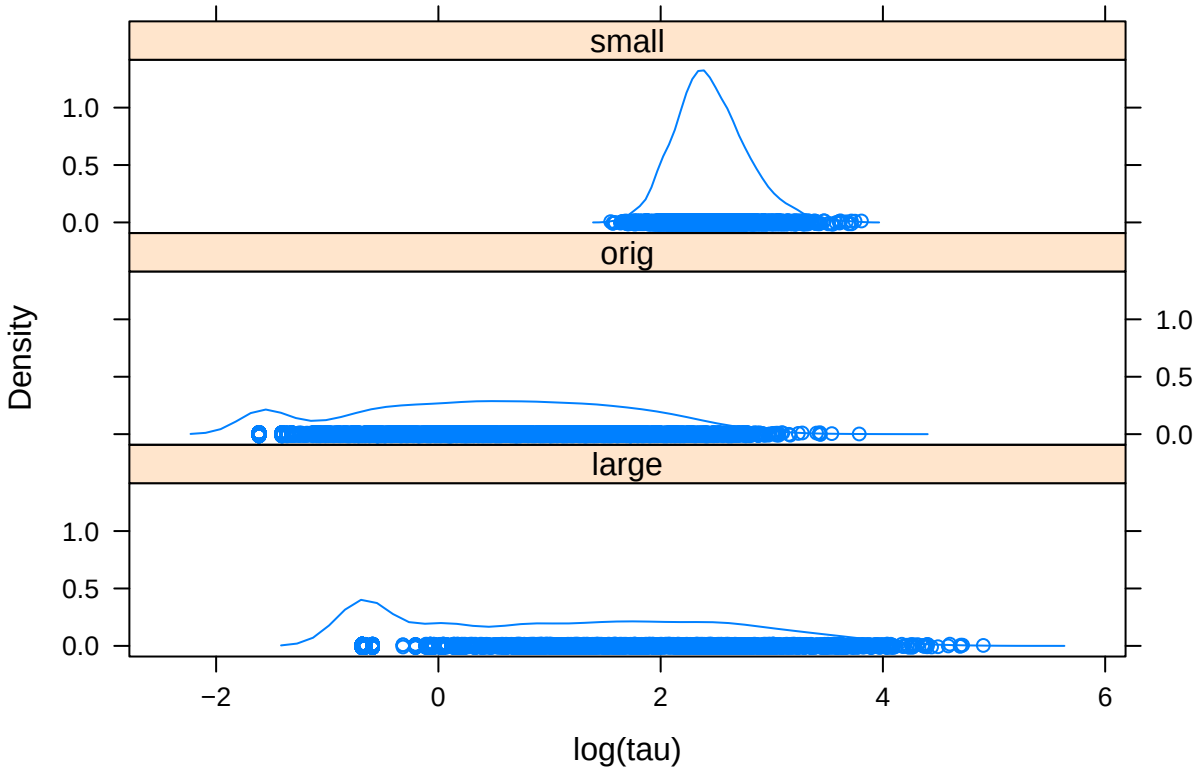
```
dp <- data.frame(tau=c(taus), model = rep(c('orig','small','large'),
                                          each = length(taus)/3))
head(dp)
```

```
          tau model
1  5.6481067  orig
2 10.5437002  orig
3  0.6859223  orig
4  1.2425951  orig
5 10.5162430  orig
6  5.0593960  orig
```

```
densityplot(~tau | model, dp, layout = c(1,3))
```




```
densityplot(~log(tau) | model, dp, layout = c(1,3))
```



Improper posterior distribution

In this case, MCMC will just wander around the entire space without giving consistent results from run to run.

An improper posterior can only happen with an improper prior. A proper prior will never result in an improper posterior.

Prior distributions that exclude the point of convergence

Convergence to the edge of the parameter space

Tails of the distribution

A common example would be a ratio of normals, or a ratio of variance where the denominator has one degree of freedom. Any ratio where the denominator does not have zero density at the origin will produce a thick-tailed distribution like the Cauchy distribution. The distribution has neither mean nor variance and attempting to summarize it with a mean and standard deviation will invariably lead to an underestimate of the probability of outcomes in the extreme tails, a serious issue if a parameter is indicator of risk.