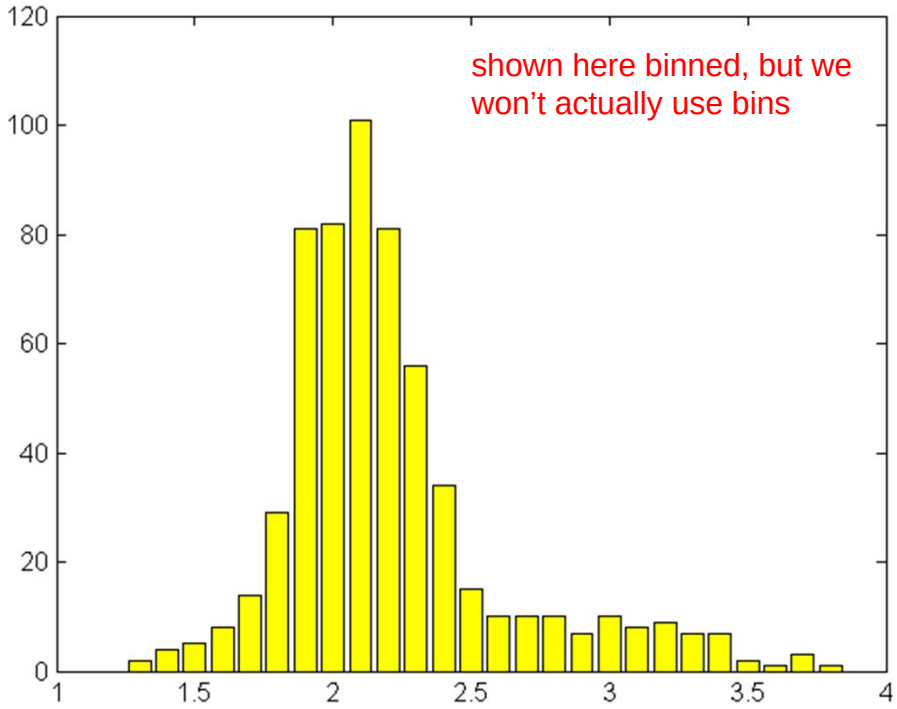
Opinionated Lessons in Statistics

by Bill Press

#40 MCMC: Example 1

Let's try MCMC on our two-Student-t model, assuming (as in a previous lesson) 600 data points (actually drawn as a random sample from the full data set)

The data is the (log) length of 600 random exons in the human genome. If it is the sum of two peaked components, what is the ratio of their areas?



6 parameters: two centers, two widths, ratio of peak heights, and Student t index.

multivariate Normal proposal distribution based on the covariance matrix of the maximum likelihood parameters

```
cstart = [2.1 0.19 0.09 3.1 0.26 4.2]
loglikefn = @(cc) twostudloglikenu(cc,data);
covar = 0.1 * inv(hessian(loglikefn,cstart,.001))
cstart =
    2.1000    0.1900    0.0900    3.1000    0.2600    4.2000
covar =
    0.0000    0.0000   -0.0000    0.0000   -0.0000   -0.0001
    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0005
   -0.0000    0.0000    0.0000   -0.0000   -0.0000    0.0003
    0.0000    0.0000   -0.0000    0.0003   -0.0001   -0.0010
   -0.0000   -0.0000   -0.0000   -0.0001    0.0002    0.0013
   -0.0001    0.0005    0.0003   -0.0010    0.0013    0.0904
```

(they're not really zeros, they just print that way)

Here's the Metropolis-Hastings step function:

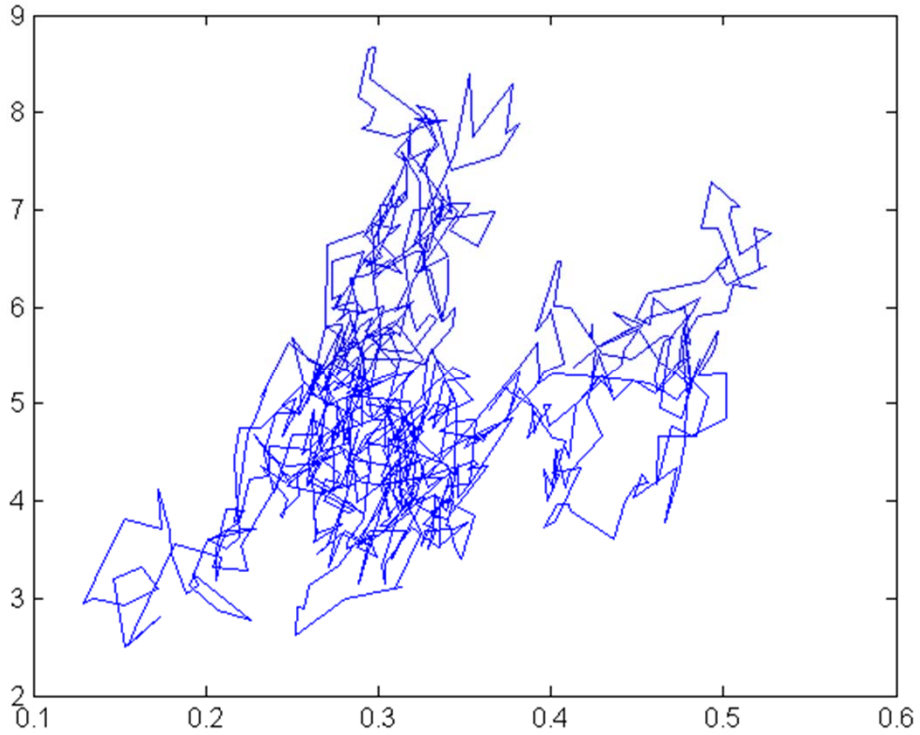
```
function cnew = mcmcstep(cold, covar)
cprop = mvnrnd(cold,covar);
alpha = min(1,exp(loglikefn(cold)-loglikefn(cprop)));
if (rand < alpha)
    cnew = cprop;
else
    cnew = cold;
end
```

Let's see the first 1000 steps:

```
chain = zeros(1000,6);
chain(1,:) = cstart;
for i=2:1000
  chain(i,:) = mcmcstep(chain(i-1,:),covar);
end
plot(chain(:,5),chain(:,6))
```

width of 2nd component

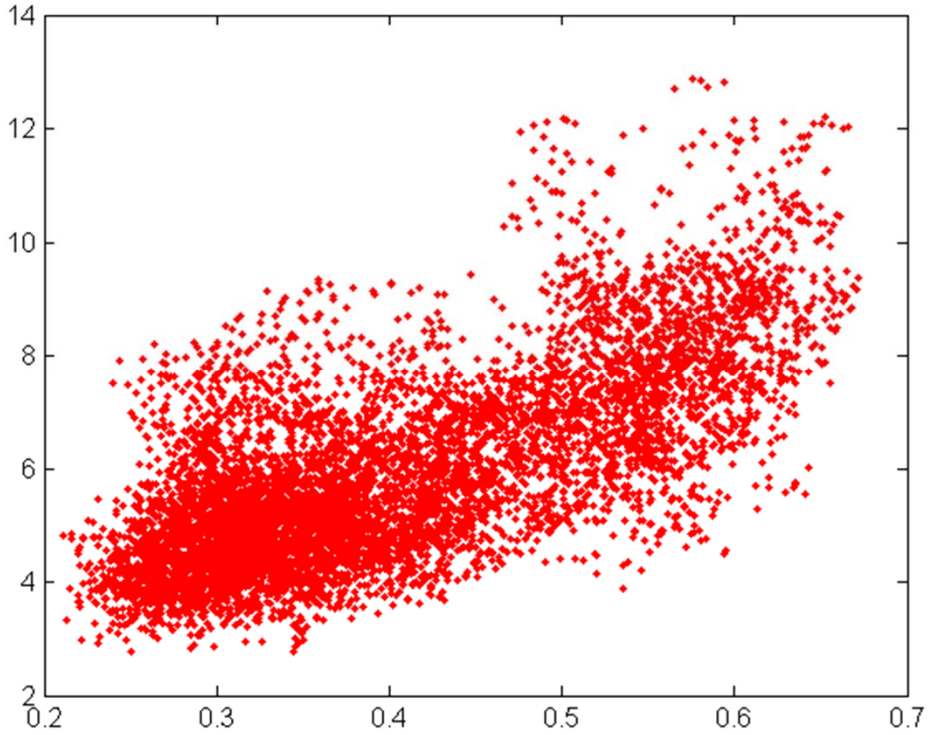
Student-t index



we're only plotting 2 components of chain, but it of course actually is a sample of the joint distribution of all the parameters

Try 10000 steps:

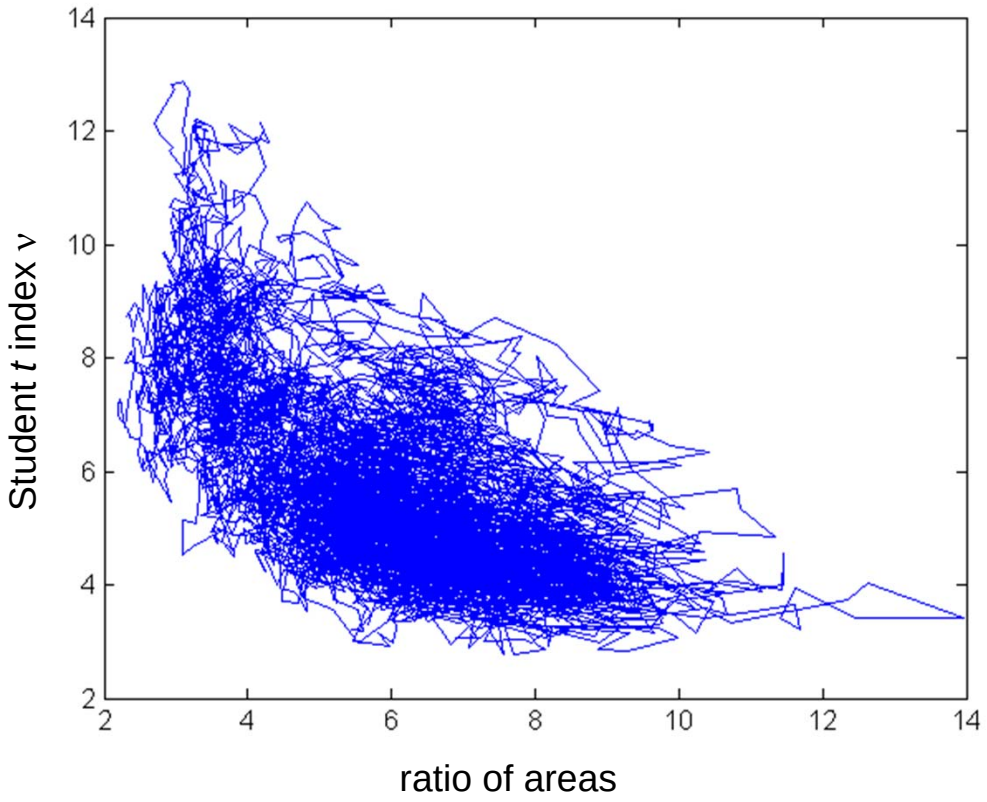
```
chain = zeros(10000,6);
chain(1,:) = cstart;
for i=2:10000, chain(i,:) = mcmcstep(chain(i-1,:),covar); end
plot(chain(:,5),chain(:,6),'.r')
```



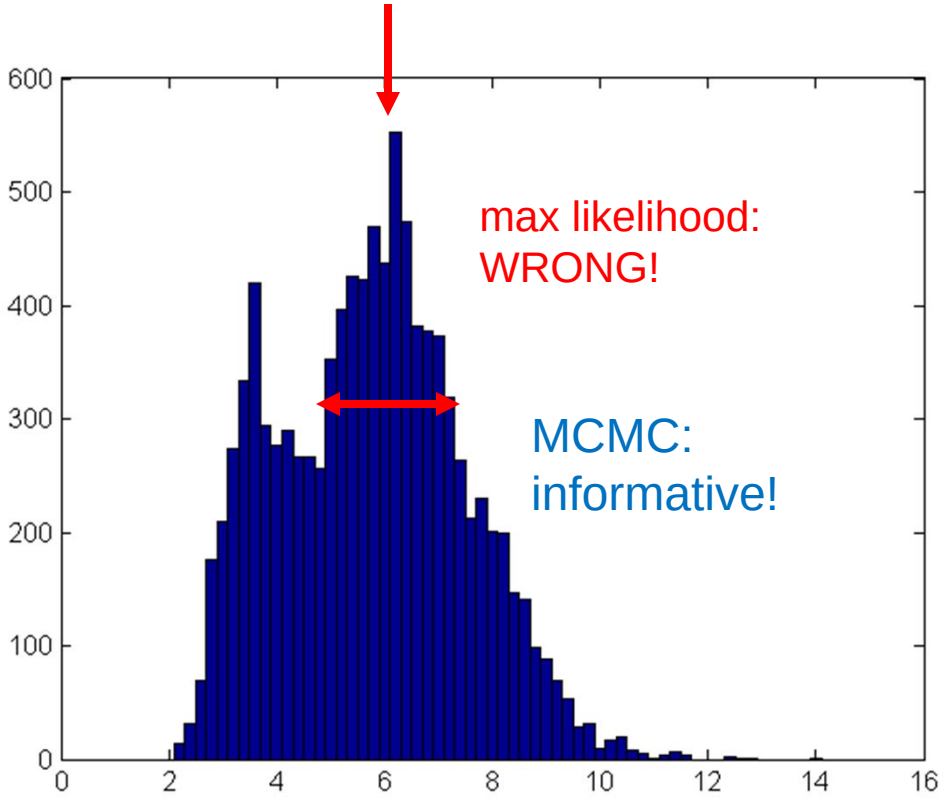
OK, plausibly ergodic. I should probably do 100000 steps, but Matlab is too slow and I'm too lazy to program it in C right now. (Don't you be!)
There are various ways of checking for convergence more rigorously, none of them foolproof; see NR3.

The payoff now is that we can look at the posterior distribution of any quantity, or derived quantity, or joint distribution of quantities, etc., etc.

```
areas = chain(:,2)./(chain(:,3).*chain(:,5));
plot(areas,chain(:,6))
```



```
hist(areas,1:.2:15)
```



We can now see why the ratio of areas is so hard to determine:
It is rather degenerate with the Student t index (i.e., sensitive to the tails) and, with only 600 data points, it can be bimodal. Maximum likelihood and derivatives of the log-likelihood (Fisher information matrix) don't capture this. MCMC and Bayes posteriors do.