Lamda-phage DNA tracking data analysis using bootstrap method

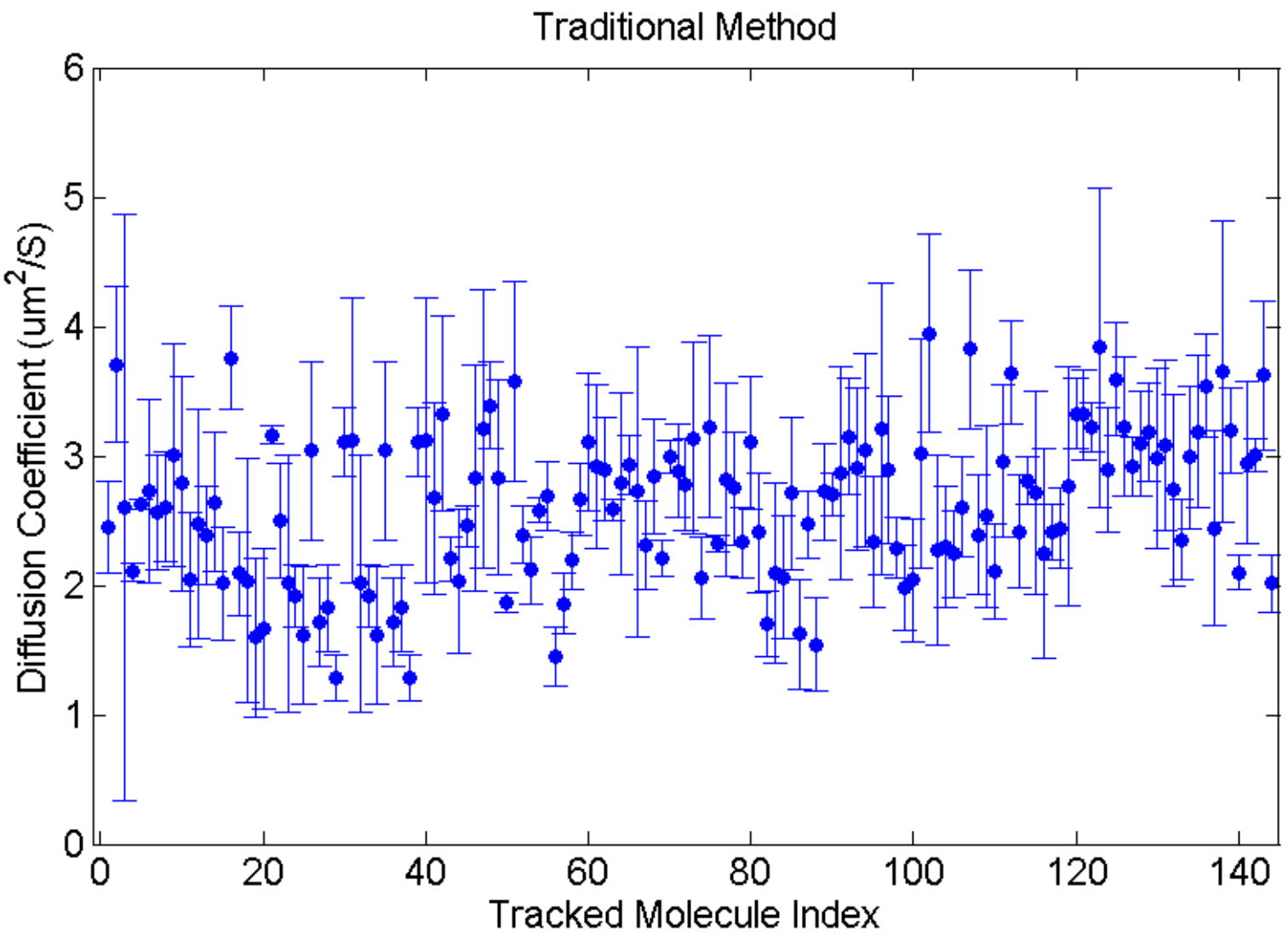
Motivation

- a large distribution of diffusion coefficients in our second lambda phage DNA measurement (3.8kbp ones)
- Wide diffusion coefficients distribution is due to that there are several species of DNA molecules with different polymer length.

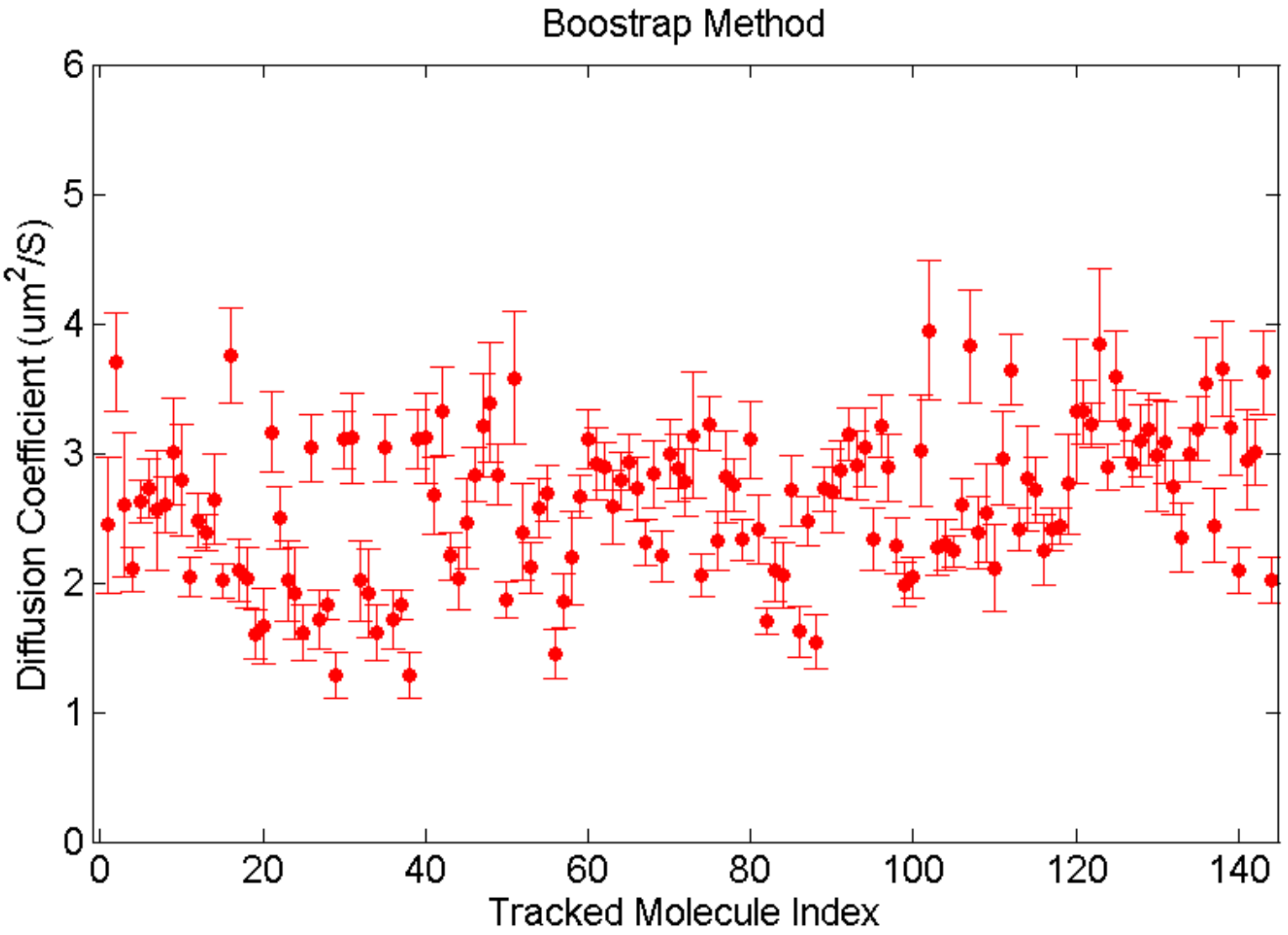
- Can we can identify subpopulation with their diffusion coefficient?

- Traditional way to determine diffusion coefficient and their confidence intervals is:
  - Compute MSD(tau)
  - Fit to the model and obtain x,y,z diffusion coe
  - Use xyz average and std as mean and measurement error.

- This method gives large measurement error
  - Fits subjects to many other parameters such as fluorescence background, gain settings and model used.
  - It converges slower (statistically bootstrap is an order better than central limit theorem.).
  - Our x,y,z do behave differently.



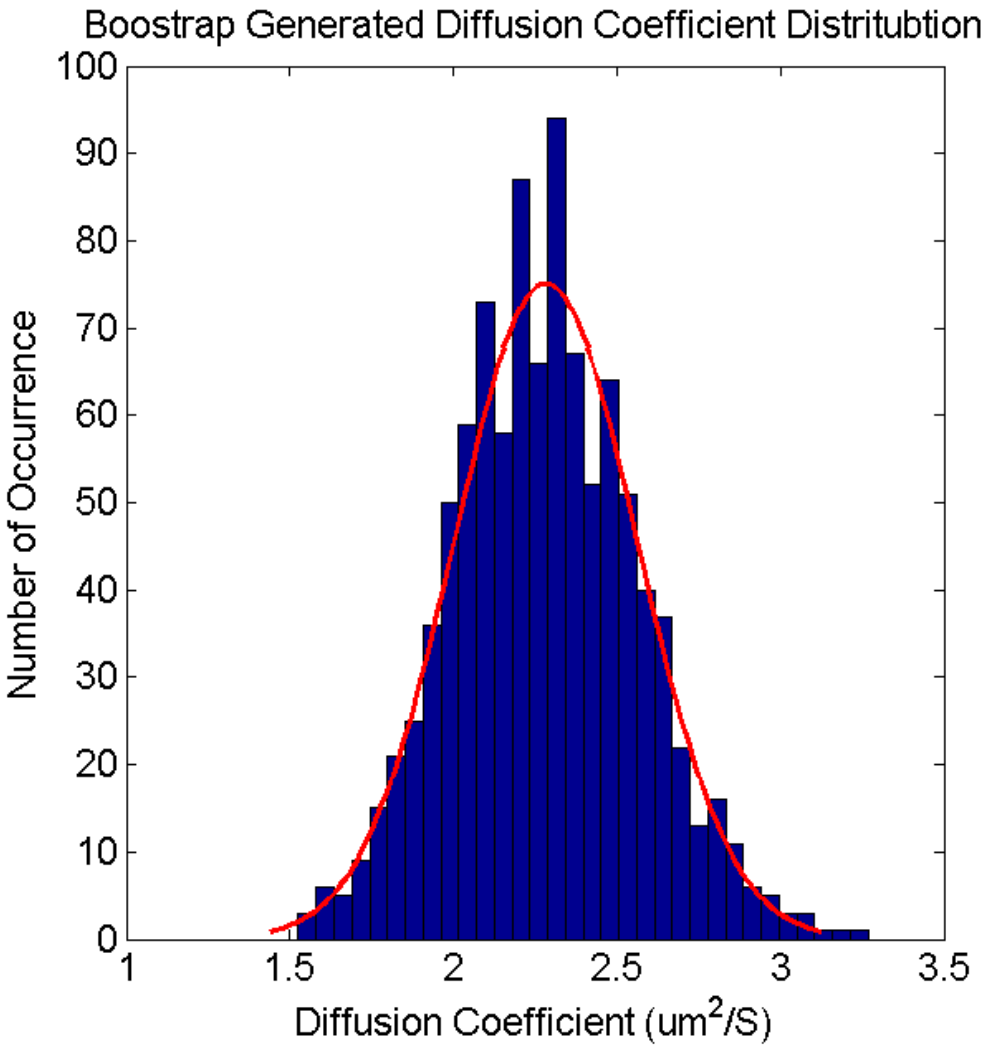
Our diffusion coe errors are too large to allow identifying any subpopulations.



Std from traditional method: 0.50 um^2/S.
Std from bootstrap method: 0.27 um^2/S.

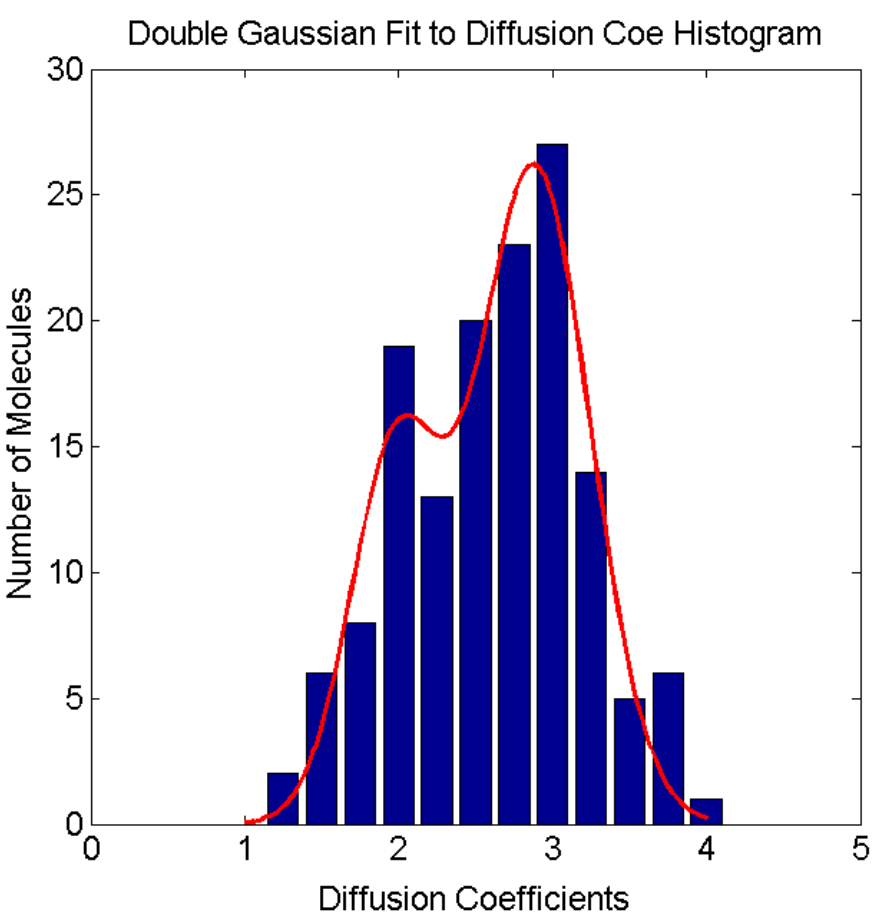
Std determines bin size in histograms…

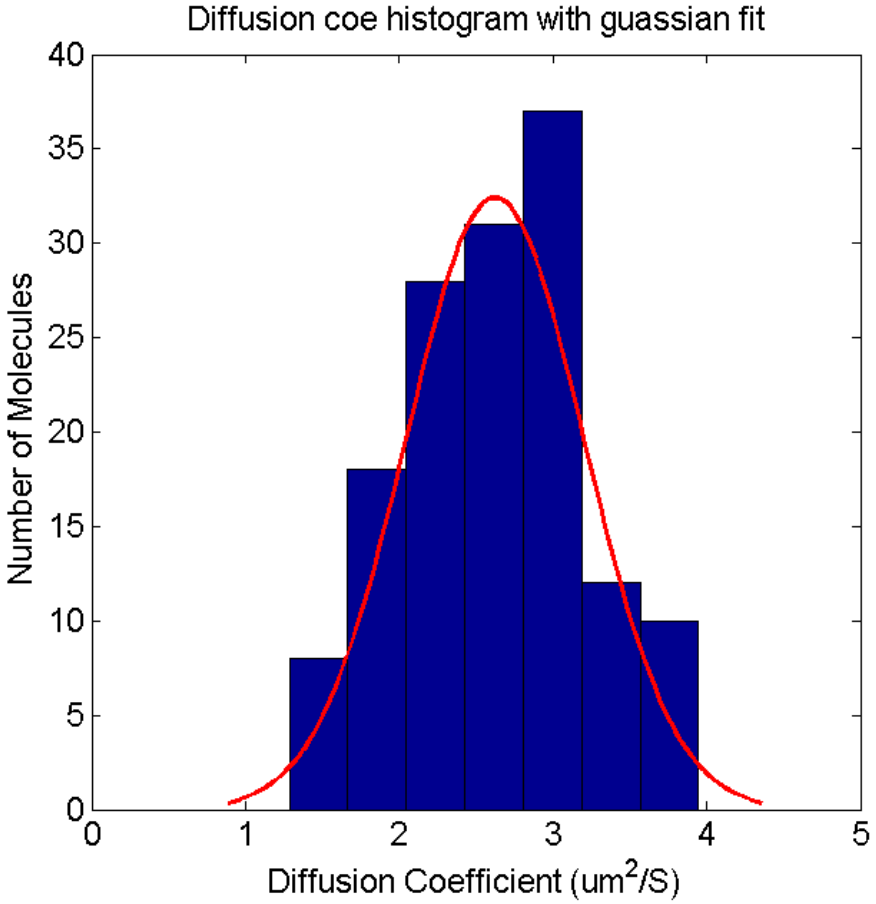
How does bootstrap do this?



- Take a trajectory. Get DeltaX = X(deltaT*(k+1))-X(deltaT*k).
- We know from diffusion model DeltaX's are iid, so we can generate bootstrap states using DeltaX, with some rules…
- Calculate diffusion coe from each generated DeltaX set.
- Use diffusion coe distribution to obtain new c.i.
- For example, figure on the left is a generated diffusion coe distribution.

A good reference book: An introduction to the Bootstrap by Efron.





Now let's look at 3.8kbp DNA diffusion coe histogram.

Can only plot histogram with 0.5um^2/S bin size by traditional way. Single guassian fit tells that a single species with diffusion coe of 2.6 um^2/S.

With bootstrap, 0.25 um^2/S bin size. Refined histogram.
At least two population, with diffusion coefficients 1.98 and 2.88 um^2/S. Roughly 36% is 1.98 um^2/S with the rest 2.88 um^2/S. If there is more species we need to take more data.

Conclusion

- Bootstrap method is very useful in our diffusion data analysis by:
  - Converging faster and thus providing tight error bounds on diffusion coe
  - Fast and robust analysis than our traditional fitting methods. Fits now and then won't work due to non-converging issues.
  - You can even do more analysis with bootstrap analysis as it is transformation invariant.