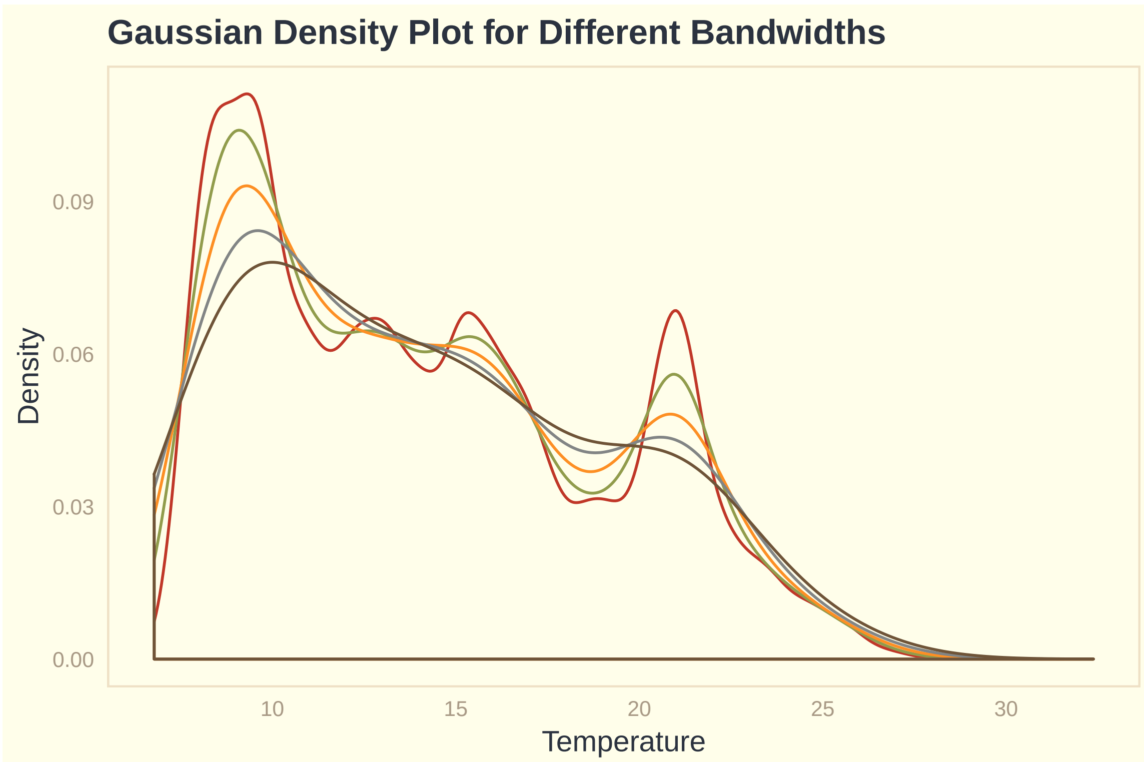
Lab 1 Part 2 - Stat 215A

Todd Faulkenberry

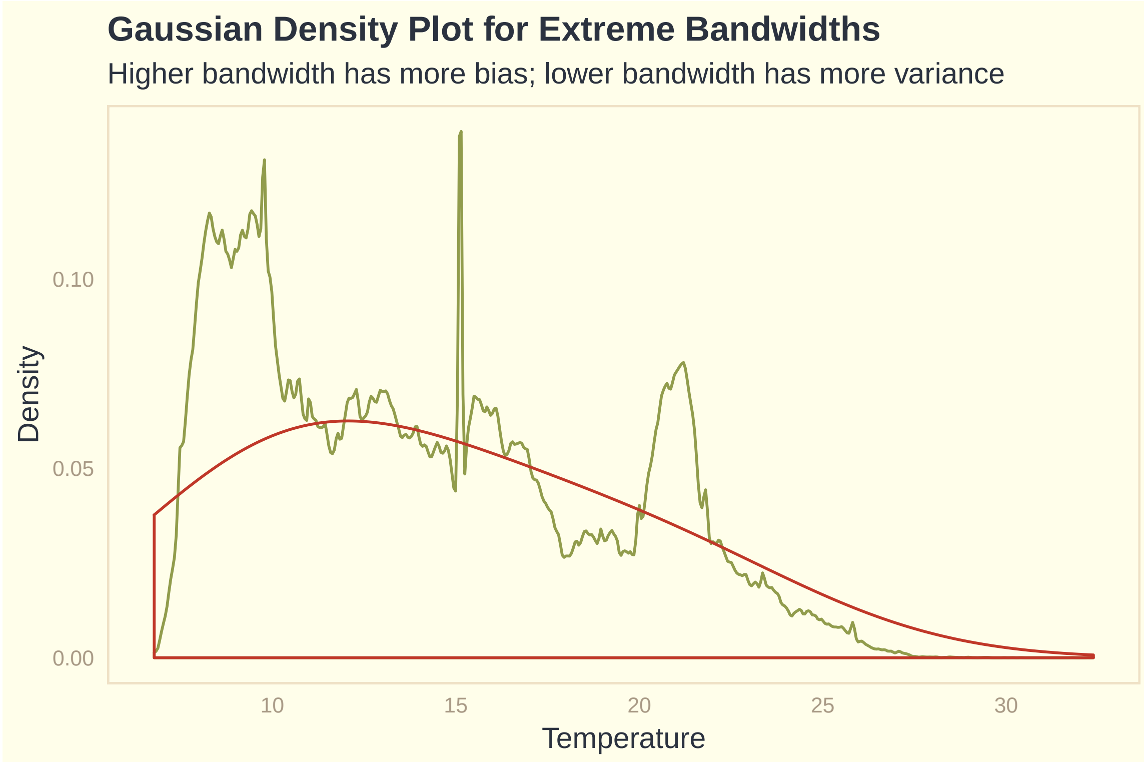
10/5/2018

Density Estimate

Below is a density estimate of temperature with bandwidths 1-5.



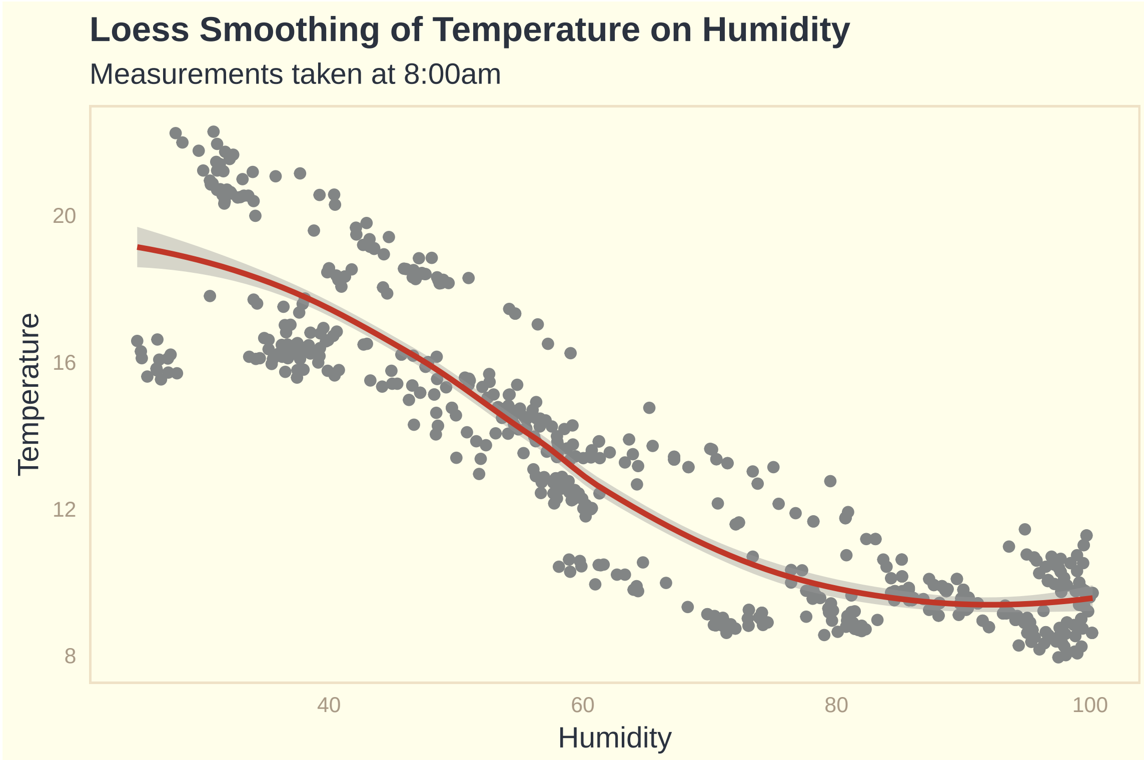
In the graph above, the kernels with the more extreme values have lower bandwidth. This demonstrates the trade off between bias and variance that we experience with bandwidth. If we choose a bandwidth that is too high, our estimate will systemically differ from the true value, creating bias. If we choose one too low, our estimate becomes jittery, changing dramatically in ways we don't expect. The graph below demonstrates this problem by showing two densities: One at one-tenth of the default bandwidth, and one ten times the default bandwidth:



The lower bandwidth overfits the model, giving us a variance that very jagged and not reflective of the underlying data. Meanwhile, the higher the bandwidth, the smoother data. This creates a similarly non-representative estimate, but in the opposite direction.

Loess Smoothing

Next, I plotted temperature against humidity for all data that was collected at 8:00am from the tree(s):



Here, I used bandwidth and polynomial (quadratic) for the loess smoother. I stuck with the default because higher degrees did not improve the fit substantially. Even though this smoother has higher variance at its boundaries (especially its left one), it has lower bias than, say, a liner polynomial while not overfitting the data. We see a clear negative relationship between temperature and humidity – an increase in temperature is associated with a decrease in humidity.