Bandwidths for Univariate Kernel Density Estimation

Brad Stieber

Introduction

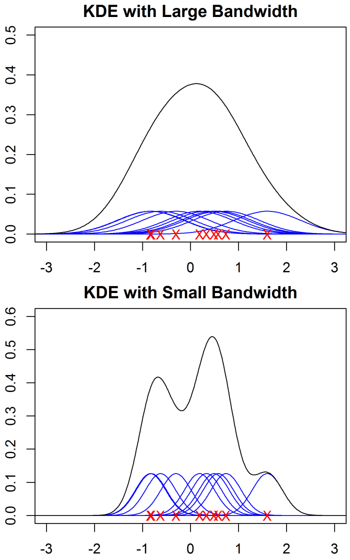
KDE (choose kernel K and bandwidth h):

$$\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Optimal Bandwidth:

$$h_{AMISE} = \left(\frac{R(K)}{n\sigma_K^4 R(f'')}\right)^{\frac{1}{5}}$$

- $R(g) = \int g^2$: roughness of a function
- Don't know $R(f'') \to$ bandwidth selections rely on getting around this unknown



Candidate Bandwidths (1/2)

Unbiased Cross Validation (E[UCV + R(f)] = MISE)

Minimize

$$UCV(h) = R\left(\widehat{f}\right) - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(x_i),$$

instead.

$$\widehat{f}_{-i}(x_i) = \frac{1}{h(n-1)}\sum_{j\neq i} K\left(\frac{x_i - x_j}{h}\right)$$

is the LOO estimator. Used to estimate the second term in $ISE(h) = \int \widehat{f}^2 - 2\int \widehat{f}_h f + \int f^2$.

Issue: excessive variation

Candidate Bandwidths (2/2)

Terrell's Maximal Smoothing

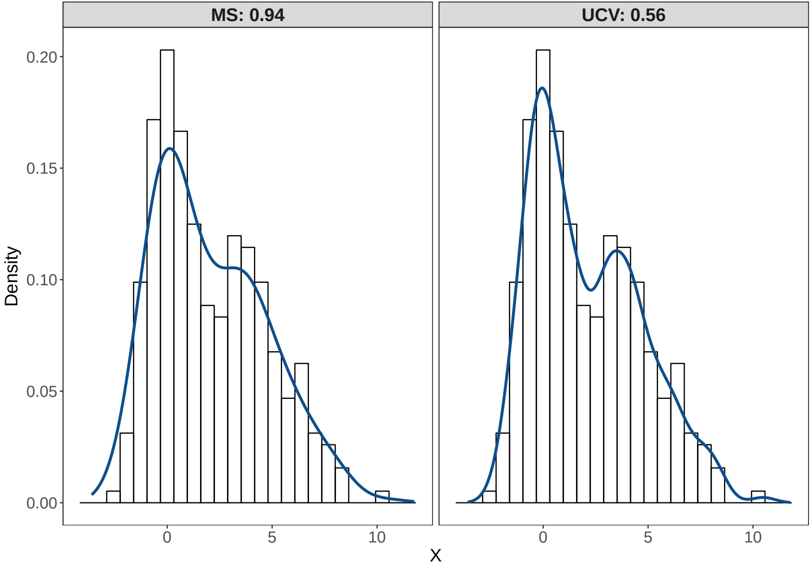
Instead of estimating $R(f^{''})$, what if we tried to minimize it? Built on the result that the $beta(k+2, k+2)$ family minimizes $\int (f^{(k)})^2$ for a given standard deviation.

$$h_{MS} = 3\hat{\sigma}\left(\frac{R(K)}{35n}\right)^{\frac{1}{5}}.$$

Issue: upper bound on $h_{opt} \rightarrow$ oversmooths interesting features of the data

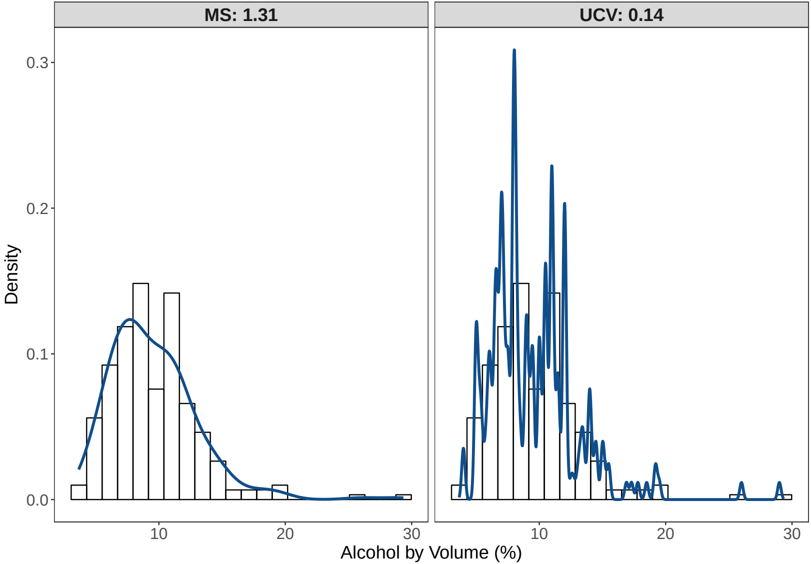
Example 1

50/50 mixture of $N(0,1)$ and $N(2,2^2)$



Example 2

Alcohol by volume of Beer Advocate's top 250 beers



Conclusion

- Choosing a bandwidth should be an iterative process
- Bias - variance tradeoff
  - Too smooth: low variance, high bias
    - what *is not there* might be
  - Too wiggly: high variance, low bias
    - what *is there* might be too hard to see