STAT 215A HW 2

Jonathan Fischer SID: 24962996

October 7, 2014

1 Kernel Density Estimation

The fact that bias increases when variance decreases (and vice versa) as functions of the bandwith h is known as the bias-variance tradeoff for histograms and kernel density estimates. The tradeoff tells us that there is no value of h that minimizes both the point-wise bias and variance of our estimates simultaneously. Furthermore, it says that the cost of minimizing one of these quantities is to increase the other. Let's consider the qualitative effect of varying h on the bias and variance. Small h should give us a good local picture of the density, decreasing bias. However, this means that our estimates will be less robust since the kernel is more sensitive to individual data points, so variance increases; we get an undersmoothed estimate. Conversely, larger h correspond to a kernel that incorporates data in a more global sense, dropping the variance. Unfortunately, this causes nearby points to have potentially excessive influence on our estimates and increases the bias. This can result in oversmoothing. As the number of sampled points increases, we should anticipate the variance to drop while the bias will be unaffected. This is because as we obtain more data, the proportion of points falling within an arbitrary bin will approach what is predicted by the true density. The kernel operations will then give values approaching those arising from applying the same kernel operations to the density itself. Hence our estimate will converge to some value, indicating decreasing variance. Evaluation of the kernel function introduces non-local effects according to the bandwith, biasing the estimator. These effects have no dependence on n and thus will not dissipate as the sample size increases.

We define our estimate and impose conditions on our kernel as follows, under the assumption that $\{x_i\}_0^n$ are i.i.d.:

$$\hat{f}_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x), \tag{1}$$

$$K_h(t) = \frac{1}{h}K(\frac{t}{h}), \tag{2}$$

$$\int K_h(t)dt = 1, \tag{3}$$

$$\int tK_h(t)dt = 0. \tag{4}$$

The calculate the bias we must take the expectation of our estimate $\hat{f}_{n,h}(x)$:

$$E[\hat{f}_{n,h}(x)] = E[\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)] = E[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x_i - x}{h})].$$

Fubini's thereom allows us to interchange the integration and summation. A change of variables with $\frac{x_i - x}{h} \to t$ combined with the i.i.d. assumption results in

$$= \frac{1}{nh}\sum_{i=1}^{n} E[K(\frac{x_i - x}{h})] = \frac{1}{nh}\sum_{i=1}^{n}\int K(\frac{x_i - x}{h})f(x_i)dx_i = \frac{1}{n}\sum_{i=1}^{n}\int K(t)f(x+ht)dt = \int K(t)f(x+ht)dt.$$

In the general case, this is the best we can do to express the bias. Representing $f(x+ht)$ as a Taylor series (when f is nice enough to do so) elucidates the h-dependence of the bias. Using

$$f(x+ht) = \sum_{j=0}^{\infty} \frac{(ht)^j f^{(j)}(x)}{j!}$$

we see that bias increases with h. For small h, we can get a cleaner form by dropping all terms of order greater than 2. This relaxes the differentiability constraints on f by only requiring the existence of derivates through second order. Again applying Fubini's theorem, the expression becomes

$$\int K(t)f(x+ht)dt \approx f(x)\int K(t)dt + hf'(x)\int tK(t)dt + h^2 f''(x)\int t^2 K(t)dt.$$

Applying (3) and (4) simplifies this to

$$f(x) + h^2 f''(x)\int t^2 K(t)dt.$$

To get the bias, we subtract $f(x)$ and see, for small h,

$$bias \approx h^2 f''(x)\int t^2 K(t)dt.$$

If we think of $K(t)$ as a probability distribution on t, we can replace the integral with σ_t^2, the variance of t. The bias is then approximately $h^2 f''(x)\sigma_t^2$. While this change doesn't yield any new information about the effect of h, it agrees with our intuition that increased globality induces larger biases since σ_t^2 represents the spread of the kernel. For all h, the term is

$$bias = \int K(t)f(x+ht)dt.$$

We can now conclude that small bandwiths give small biases and large bandwiths lead to large biases. As we posited earlier, there is no dependence on the number of samples. Interestingly to note, the curvature of the density f at point x appears in the bias term via $f''(x)$. As kernel density estimation produces a smoothed, or averaged-over picture of the density, it makes sense that the magnitude of the bias increases with the magnitude of the curvature since smoothing deletes this information.

2 Multidimensional Scaling

1. By definition, $a_{rs} = -\frac{1}{2}d_{rs}$. The matrix $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}' = \mathbf{I} - n^{-1}\mathbf{J}$ where $\mathbf{J}$ is the Hadamard identity matrix (all ones). Define $\mathbf{B} = \mathbf{HAH}$. If we have $\mathbf{D}$ Euclidean, $d_{rs}^2 = (\mathbf{z}_r - \mathbf{z}_s)'(\mathbf{z}_r - \mathbf{z}_s)$. We can thus compute the entries of $\mathbf{A}$.

$$\mathbf{B} = \mathbf{HAH} = (\mathbf{I} - n^{-1}\mathbf{J})\mathbf{A}(\mathbf{I} - n^{-1}\mathbf{J}) = \mathbf{A} - n^{-1}\mathbf{JA} - n^{-1}\mathbf{AJ} + n^{-2}\mathbf{JAJ}.$$

Consider b_{rs}, where $\mathbf{B} = (b_{rs})$. Then

$$b_{rs} = a_{rs} - \frac{1}{n}\sum_{v=1}^{n} a_{rv} - \frac{1}{n}\sum_{u=1}^{n} a_{us} + \frac{1}{n^2}\sum_{v=1}^{n}\sum_{u=1}^{n} a_{uv}.$$

Inserting the expression for a_{rs} yields

$$b_{rs} = (\mathbf{z}_r - \mathbf{z}_s)'(\mathbf{z}_r - \mathbf{z}_s) - \frac{1}{n}\sum_{v=1}^{n}(\mathbf{z}_r - \mathbf{z}_v)'(\mathbf{z}_r - \mathbf{z}_v) - \frac{1}{n}\sum_{u=1}^{n}(\mathbf{z}_u - \mathbf{z}_s)'(\mathbf{z}_u - \mathbf{z}_s) + \frac{1}{n^2}\sum_{v=1}^{n}\sum_{u=1}^{n}(\mathbf{z}_u - \mathbf{z}_v)'(\mathbf{z}_u - \mathbf{z}_v).$$

We expand the products to obtain

$$b_{rs} = -\frac{1}{2}(\mathbf{z}_r'\mathbf{z}_r - \mathbf{z}_r'\mathbf{z}_s - \mathbf{z}_s'\mathbf{z}_r + \mathbf{z}_s'\mathbf{z}_s) + \frac{1}{2n}\sum_{v=1}^{n}(\mathbf{z}_r'\mathbf{z}_r - \mathbf{z}_r'\mathbf{z}_v - \mathbf{z}_v'\mathbf{z}_r + \mathbf{z}_v'\mathbf{z}_v) + \frac{1}{2n}\sum_{u=1}^{n}(\mathbf{z}_u'\mathbf{z}_u - \mathbf{z}_u'\mathbf{z}_s - \mathbf{z}_s'\mathbf{z}_u + \mathbf{z}_s'\mathbf{z}_s)$$

$$-\frac{1}{2n^2}\sum_{u=1}^{n}\sum_{v=1}^{n}(\mathbf{z}_u'\mathbf{z}_u - \mathbf{z}_u'\mathbf{z}_v - \mathbf{z}_v'\mathbf{z}_u + \mathbf{z}_v'\mathbf{z}_v).$$

Note that $\frac{1}{n}\sum_{u=1}^{n}\mathbf{z}_u'\mathbf{z}_s = \bar{\mathbf{z}}'\mathbf{z}_s$. Making this substitution, cancelling like terms and using the general fact that $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x}$ for column vectors $\mathbf{x}$ and $\mathbf{y}$, simplifies the equation to

$$b_{rs} = \mathbf{z}_r'\mathbf{z}_s - \mathbf{z}_r'\bar{\mathbf{z}} - \bar{\mathbf{z}}'\mathbf{z}_s + \frac{1}{n}\sum_{v=1}^{n}\mathbf{z}_v'\mathbf{z}_v - \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n}(\mathbf{z}_u'\mathbf{z}_u - \mathbf{z}_u'\mathbf{z}_v)$$

$$= \mathbf{z}_r'\mathbf{z}_s - \mathbf{z}_r'\bar{\mathbf{z}} - \bar{\mathbf{z}}'\mathbf{z}_s + \frac{1}{n}\sum_{v=1}^{n}\mathbf{z}_v'\mathbf{z}_v - \frac{1}{n}\sum_{u=1}^{n}\mathbf{z}_u'\mathbf{z}_u + \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n}\mathbf{z}_u'\mathbf{z}_v.$$

The fourth and fifth terms are equal since they are sums of the same thing ranging over all indices, and therefore cancel. The evaluate the double sum, we sum over each index separately, giving $\bar{\mathbf{z}}'\bar{\mathbf{z}}$. The whole expression for b_{rs} is now

$$b_{rs} = \mathbf{z}_r'\mathbf{z}_s - \mathbf{z}_r'\bar{\mathbf{z}} - \bar{\mathbf{z}}'\mathbf{z}_s + \bar{\mathbf{z}}'\bar{\mathbf{z}} = (\mathbf{z}_r - \bar{\mathbf{z}})'(\mathbf{z}_s - \bar{\mathbf{z}}),$$

as we have sought to show.

Define a matrix $\mathbf{Z}$ whose columns are the points $\mathbf{z}_1, ..., \mathbf{z}_n$. Consider the matrix $\mathbf{HZZ'H} = (\mathbf{HZ})(\mathbf{HZ})'$. This matrix must be positive semidefinite, and we show that it is equivalent to $\mathbf{B}$, making $\mathbf{B}$ PSD. The entry at position (r, s) of the matrix $\mathbf{ZZ'}$ is $\mathbf{z}_r\mathbf{z}_s' = \mathbf{z}_r'\mathbf{z}_s$. We have

$$\mathbf{HZZ'H} = \mathbf{ZZ'} - n^{-1}\mathbf{JZZ'} - n^{-1}\mathbf{ZZ'J} + n^{-2}\mathbf{JZZ'J}.$$

The matrix $\mathbf{J}$ has the effect of summing of the column (or row) depending on whether it is right or left mulitplied. With the n^{-1} factor, this means we get $\bar{\mathbf{z}}$ or $\bar{\mathbf{z}}'$ multiplied by $\mathbf{z}_r'$ or $\mathbf{z}_s$, respectively. Then the entry at (r, s) of $\mathbf{HZZ'H}$ is

$$\mathbf{z}_r'\mathbf{z}_s - \mathbf{z}_r'\bar{\mathbf{z}} - \bar{\mathbf{z}}'\mathbf{z}_s + \bar{\mathbf{z}}'\bar{\mathbf{z}} = b_{rs}.$$

Thus $\mathbf{B} = \mathbf{HZZ'H} = (\mathbf{HZ})(\mathbf{HZ})'$, and it is PSD.

2. We begin by showing that $\mathbf{B}$ is the inner product matrix of the given configuration. Our first step is to standardize the eigenvectors so that $\mathbf{x}_{(i)}'\mathbf{x}_{(i)} = 1$. This is done by taking $\mathbf{X\Lambda}^{-1/2}$ for $\mathbf{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$. Since B is real and symmetric, we can diagonalize it as $\mathbf{B} = \mathbf{\Gamma\Lambda\Gamma'}$ where $\mathbf{\Gamma} = \mathbf{X\Lambda}^{-1/2}$. Inserting this for $\mathbf{\Gamma}$ gives

$$B = \mathbf{\Gamma\Lambda\Gamma'} = \mathbf{X\Lambda}^{-1/2}\mathbf{\Lambda}[\mathbf{X\Lambda}^{-1/2}]' = \mathbf{X\Lambda}^{-1/2}\mathbf{\Lambda\Lambda}^{-1/2}\mathbf{X'} = \mathbf{XX'} = \mathbf{X'X}.$$

This proves $\mathbf{B}$ is the inner product matrix with $b_{rs} = \mathbf{x}_r'\mathbf{x}_s$.

To check that $\mathbf{D}$ is the interpoint distance matrix, we see

$$(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s) = \mathbf{x}_r'\mathbf{x}_r - \mathbf{x}_r'\mathbf{x}_s - \mathbf{x}_s'\mathbf{x}_r + \mathbf{x}_s'\mathbf{x}_s = \mathbf{x}_r'\mathbf{x}_r - 2\mathbf{x}_r'\mathbf{x}_s + \mathbf{x}_s'\mathbf{x}_s = b_{rr} - 2b_{rs} + b_{ss}.$$

From part 1, we have

$$b_{rs} = a_{rs} - \frac{1}{n}\sum_{v=1}^{n}a_{rv} - \frac{1}{n}\sum_{u=1}^{n}a_{us} + \frac{1}{n^2}\sum_{v=1}^{n}\sum_{u=1}^{n}a_{uv}.$$

Plugging this in and cancelling terms leads to

$$(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s) = a_{rr} - 2a_{rs} + a_{ss} = -2a_{rs} = d_{rs}^2$$

since $a_{ii} = 0$. Thus $\mathbf{D}$ is the interpoint distance matrix for the given configuration.

Finally, we verify the the center of gravity of $\mathbf{X}$ is the $\mathbf{0}$ vector. Consider $\mathbf{B1} = \mathbf{HAH1}$. Since $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{J}$, $\mathbf{H1} = \mathbf{1} - n^{-1}n\mathbf{1} = \mathbf{0}$. Then $\mathbf{B1} = \mathbf{0}$ and $\mathbf{1}$ is an eigenvector of $\mathbf{B}$ associated with eigenvalue 0. Hence $\mathbf{1}$ is orthogonal to the eigenvectors $\mathbf{x}_{(i)}$, $i = 1, ..., p$. This is equivalent to

$$\bar{\mathbf{x}}_i = \frac{1}{n}\sum_{j=1}^{n}x_{(i)j} = \frac{1}{n}\mathbf{x}_{(i)}'\mathbf{1} = 0.$$

Thus $\bar{\mathbf{x}} = \mathbf{0}$.