Lab/Homework 7

ORIE 4740

TA: Jian Wu

Vincent Nguyen

**What characteristics of the data do you notice in this plot and why might they be there?**

The distribution of data points is not uniform: in particular, there seem to be clusters where the density is higher, and regions where the density of events is lower (near (x= 0,y= -5) ). This might be the case because that region is dangerous (crime, environment), which would cause a higher incidence of ambulance visits/drive-throughs.

**What bandwidths did kde2d use in each dimension?**

A bandwidth of 4.940 was used for V1

A bandwidth of 3.425137 was used for V2

The new density estimate (c(0.4,0.4)) may be worse than the density estimate that used the default specification because it does not interpolate nearly as well (undersmoothing): which is to say that it may be underestimating the true density between points (there are many "red" regions between many "hot" regions).

The new density estimate (c(14,14)) may be worse than the default specification because it is oversmoothing.

However, the new density estimate may be better than the default in the sense that it captures much greater detail: it better repesents the finer "hot" regions and allows the user to see the small "cool" regions even within the middle of a large "hot" region.

**What does your plot tell you about the eﬀect of bandwidth on the predictive**

**accuracy for the ambulance data? What value of the bandwidth gives the best**

**predictive accuracy?**

The plot shows that the bandwidth worsens the predictive accuracy for very low bandwidths and very high bandwidths. In other words, there are bandwidths that oversmooth and bandwidths that undersmooth. The predictive accuracy is best when the bandwidth is approximately 0.6 (logscore of -6.4).