Empirical Methods for Applied Micro

Problem Set 1

Alberto Cappello

2/1/2021

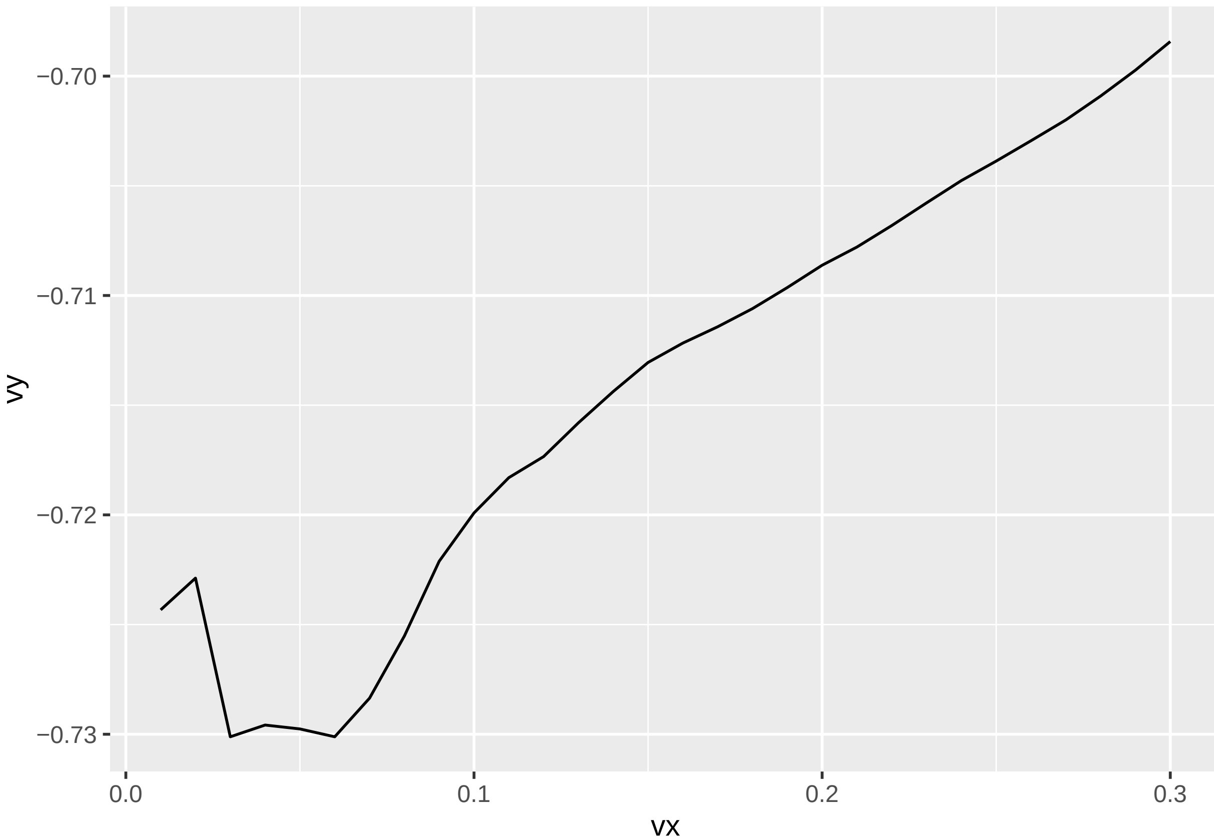
Load data

1. Density estimation with optimal h

```
# Under assumed normal distribution
n = length(bids)
sd = sqrt((sd(bids) ^ 2) * ((n - 1) / n))
# optimal bandwidth
h_star = 1.06*sd*(n)^(-0.2)
# Plug in rule for gaussian and epanechnikov kernel
h_plug = dpik(X, scalest = "minim")
kde = Vectorize(function(kernel,h) density(X,bw=h,kernel = kernel)$y)
kernels = c("gaussian","epanechnikov")
estimates = cbind(kde(kernels[1],h_star),kde(kernels,h_plug))
colnames(estimates) = c("normal",kernels)
```

2. LSCV for Epanechnikov kernel

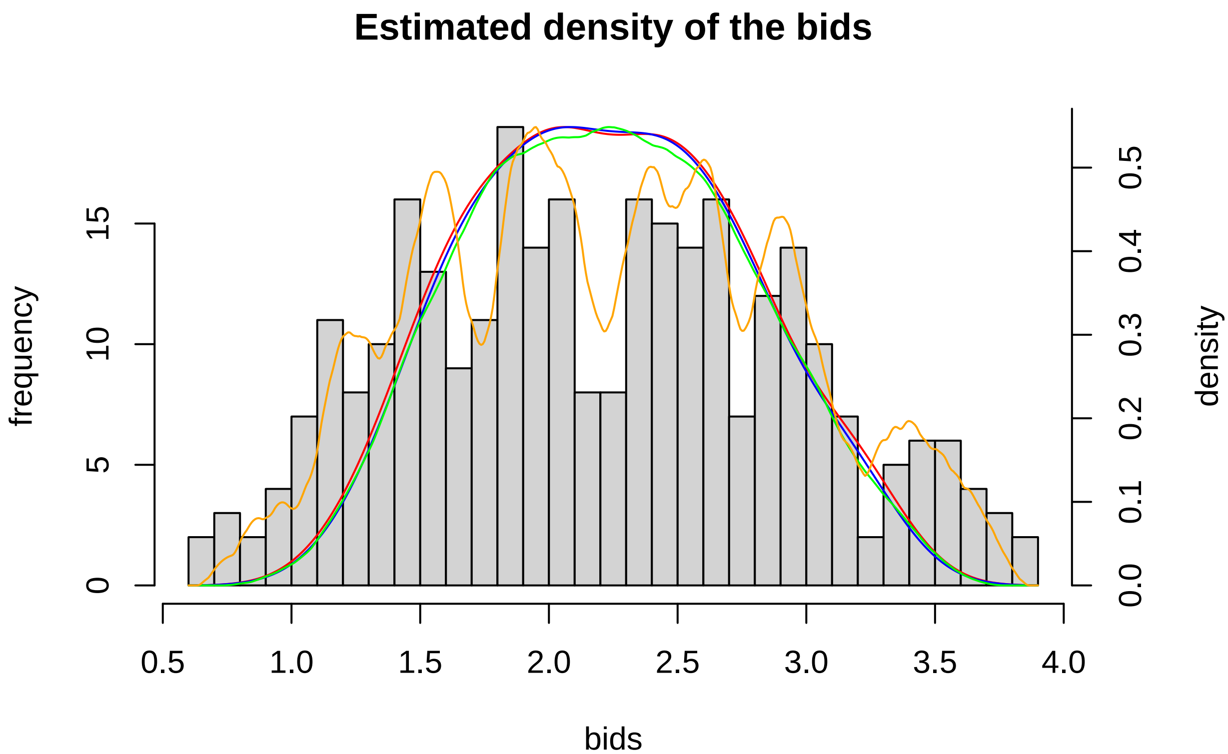
```
# find h that minimize the Risk function
kernel = kernels[2]
J<- function(h){
  fhat=Vectorize(function(x) density(X,from=x,to=x,n=1,bw=h,kernel = kernel)$y)
  fhati=Vectorize(function(i) density(X[-i],from=X[i],to=X[i],n=1,bw=h,kernel = kernel)$y)
  F=fhati(1:length(X))
  return(integrate(function(x) fhat(x)^2,-10^(-5),10^(5))$value-2*mean(F))
}
vx=seq(.01,0.3,by=.01)
vy=Vectorize(J)(vx)
df=data.frame(vx,vy)
ggplot(data=df, aes(x=vx, y=vy)) + geom_line()
```




```
# estimated density via CV
h_CV = vx[which(vy==min(vy))]
estimates = cbind(estimates, kde(kernels[2],h_CV))
colnames(estimates) = c("normal",kernels,"epanec_CV")
y_CV = density(X,bw=h_CV,kernel = kernel)$y
x_CV = density(X,bw=h_CV,kernel = kernel)$x
```

The objective function as a function of h is plotted above and is minimized at h=0.06. This is a fairly low value for the bandwidth so we expect to have an estimated density that is not very smooth.

3. Estimated densities Plot



The legend og the plot above is as follows:

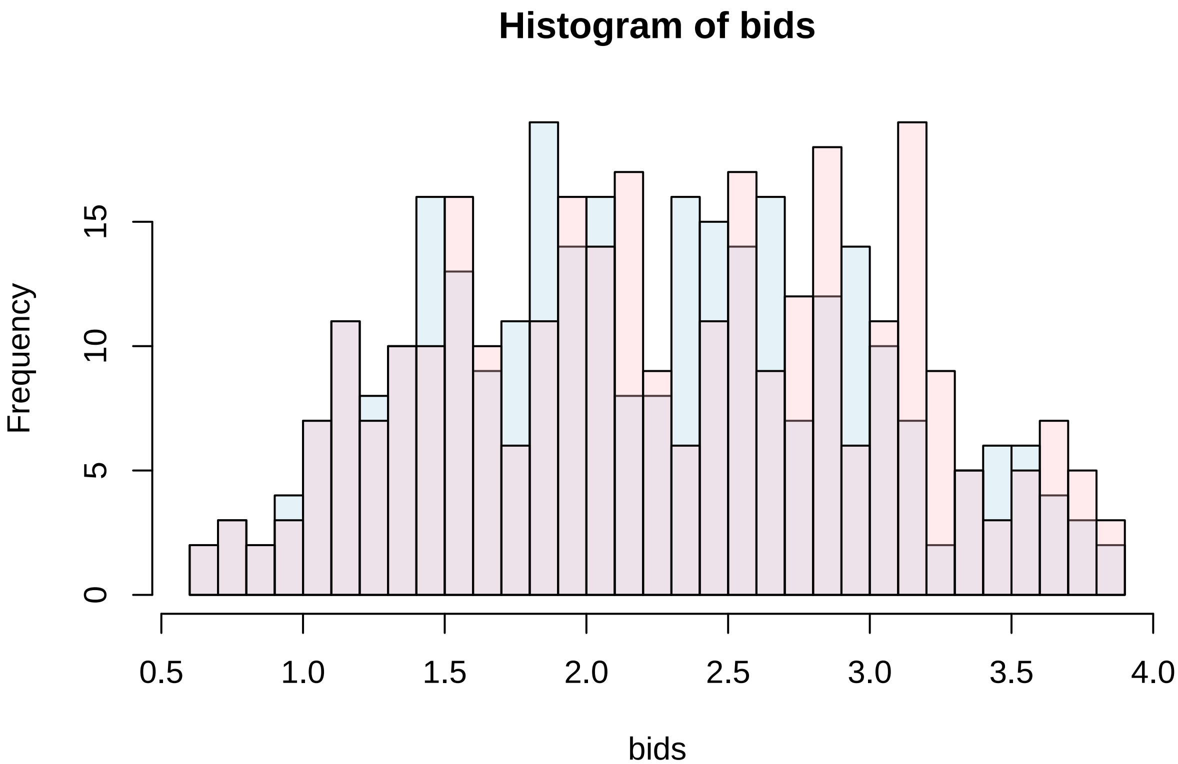
- red line = assumed normal distribution with optimal h
- blue line = Gaussian kernel with plug-in rule
- green line = Epanechnikov kernel with plug-in rule
- orange line = LSCV with Epanechnikov kernel

Densities estimated with the plug-in rule appear to attribute to be too high around the center of the distribution where the observations have low frequencies.This is probably due to the large value of the optimal bandwidth $h^* = 0.27$ Instead, density estimated with CV for the Epanechnikov kernel has very low bandwidth $h^* = 0.06$ and therefore it captures the apparent underlying multimodal distribution, but there's reasons to think that this low value of h leads to overfitting.

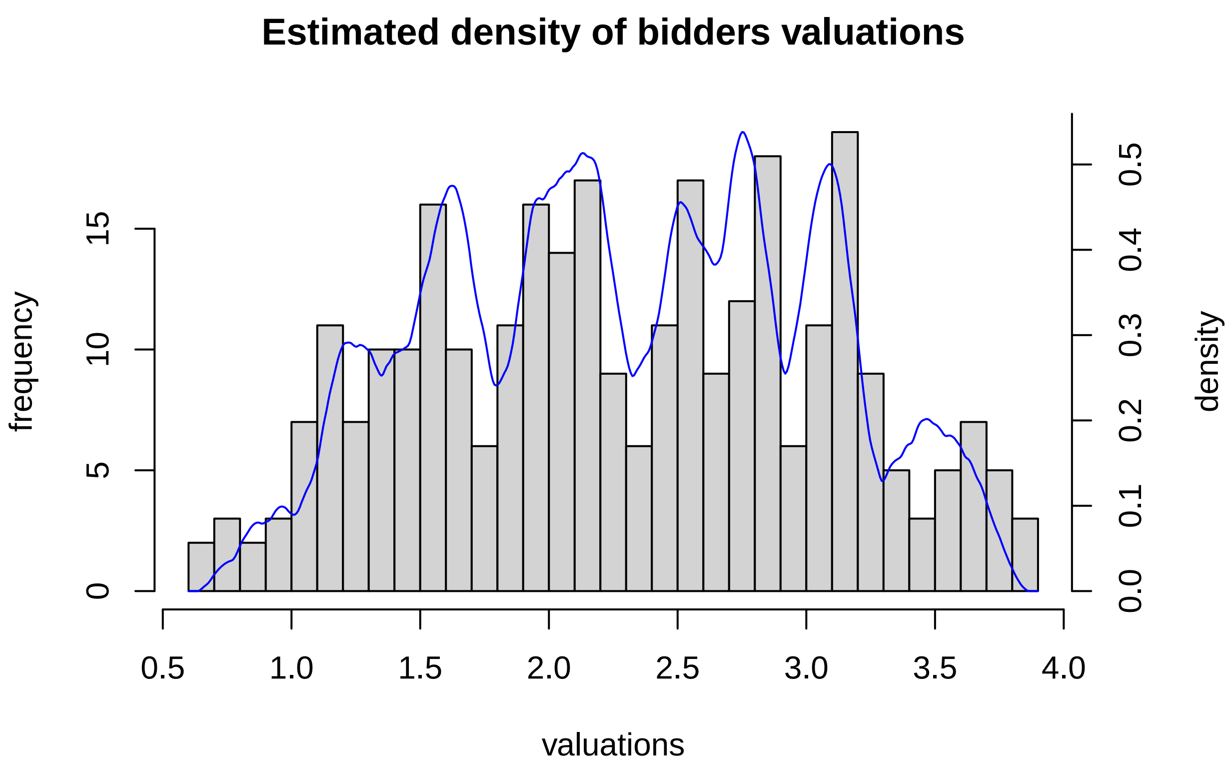
4. Evaluations implied the bids

```
#GPV to estimate the bids density
G = ecdf(X)
# interpolate to get pdf and cdf at the actual bids
g = approx(x_CV,y_CV,X)$y
N=3#of bidders

# Valuations
valuations = X + G(X)/(N-1)*g
v_pdf = density(valuations, bw = h_CV,kernel = kernels[2])
```

The plot below are the histograms of bids (light blue) and valuation (light red). We see strong overlap between bids and valuations meaning that the estimated valuations are reasonable.



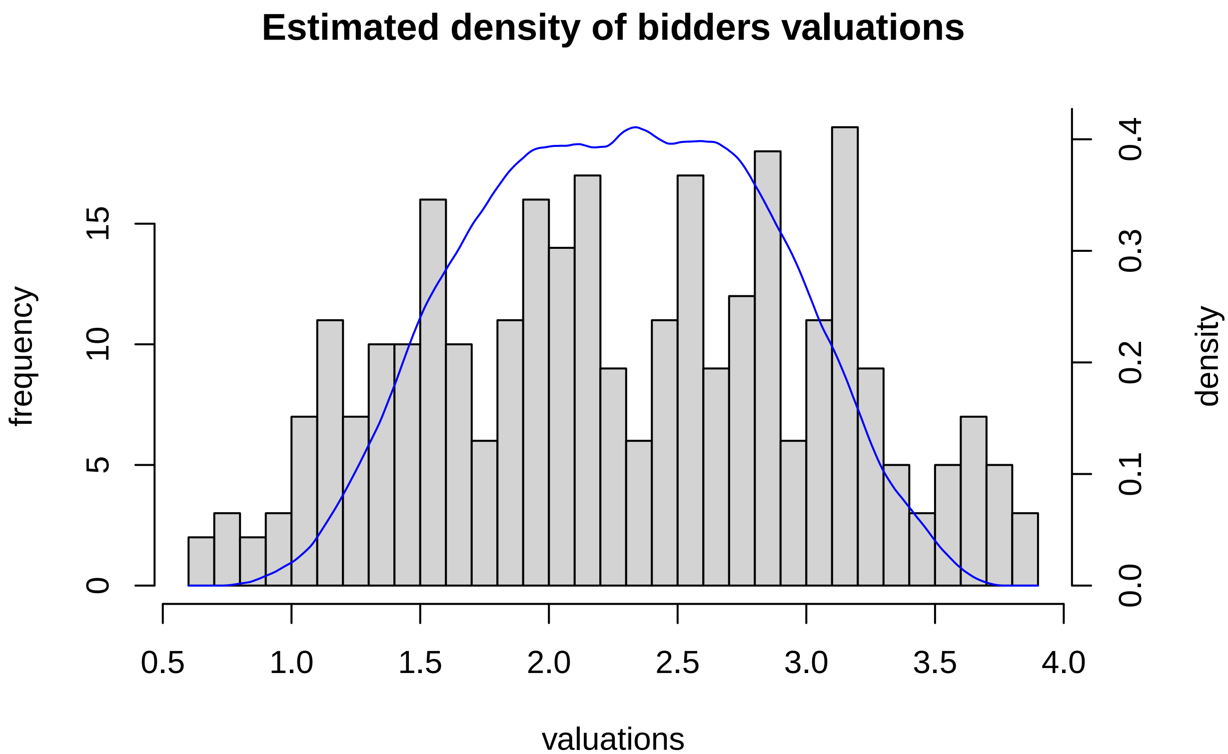
The plot below represent the histogram and the estimated density of the bidders evaluations implied by the bids estimated using the cross-validated Epanechnikov kernel. Also in this case we are probably overfitting the data.




```
#estimate the bids cdf with h from the plug in rule
G = ecdf(X)
```

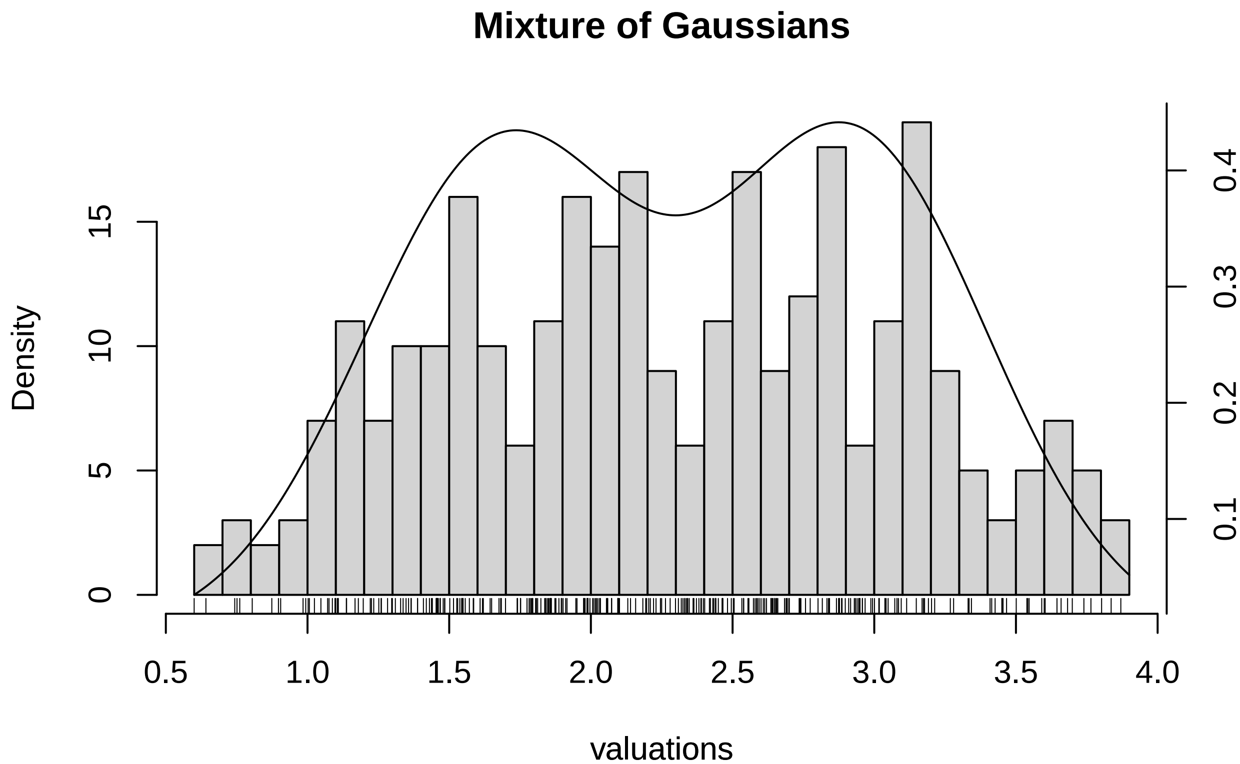
```r
# interpolate to get pdf and cdf at the actual bids
g = approx(x_CV,y_CV,X)$y
N=3#of bidders

# Valuations
valuations = X + G(X)/(N-1)*g

h_v = dpik(valuations, scalest = "minim")
v_pdf = density(valuations, bw = h_v,kernel = kernels[2])
```

The plot below represent the histogram and the estimated density of the bidders evaluations implied by the bids estimated using the Epanechnikov kernel with optimal plug-in bandwidth. Also in this case we are probably underfitting the data cause the estimated deensity does not capture the modes of the underlying distribution.



Guess the valuation distribution

```r
fit = Mclust(valuations, G=2, model="V")
hist(valuations,breaks = 30,main = "Mixture of Gaussians",ylab="")
par(new = TRUE)
plot(fit, what="density", main="Mixture of Gaussians",axes = FALSE,ylab="")
rug(X)
axis(side=4, at = pretty(range(0,0.5)))
```



My guess is that the true density of the bidders valuations is a mixture of Gaussians with means around 1.5 and 3 respectively and same variance 1.