Excercise sheet 1 (problems 2, 3)

Khodosevich Leonid

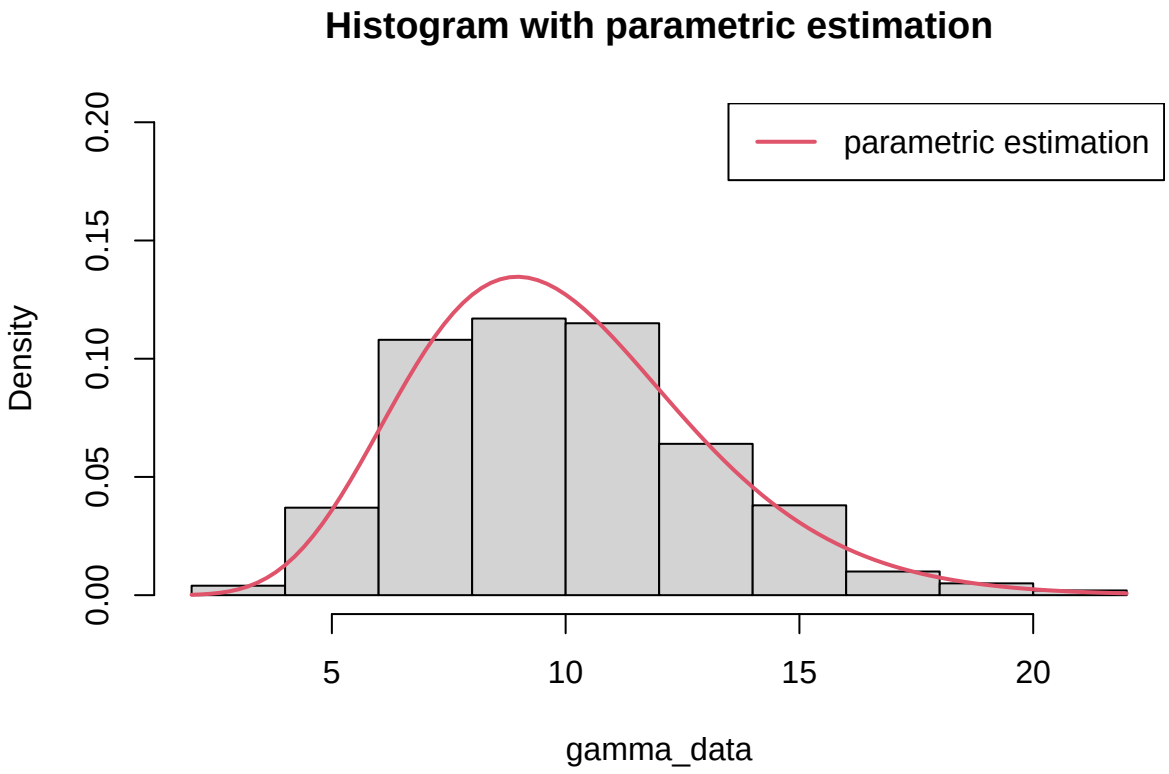
Problem 2

Subtask 1.

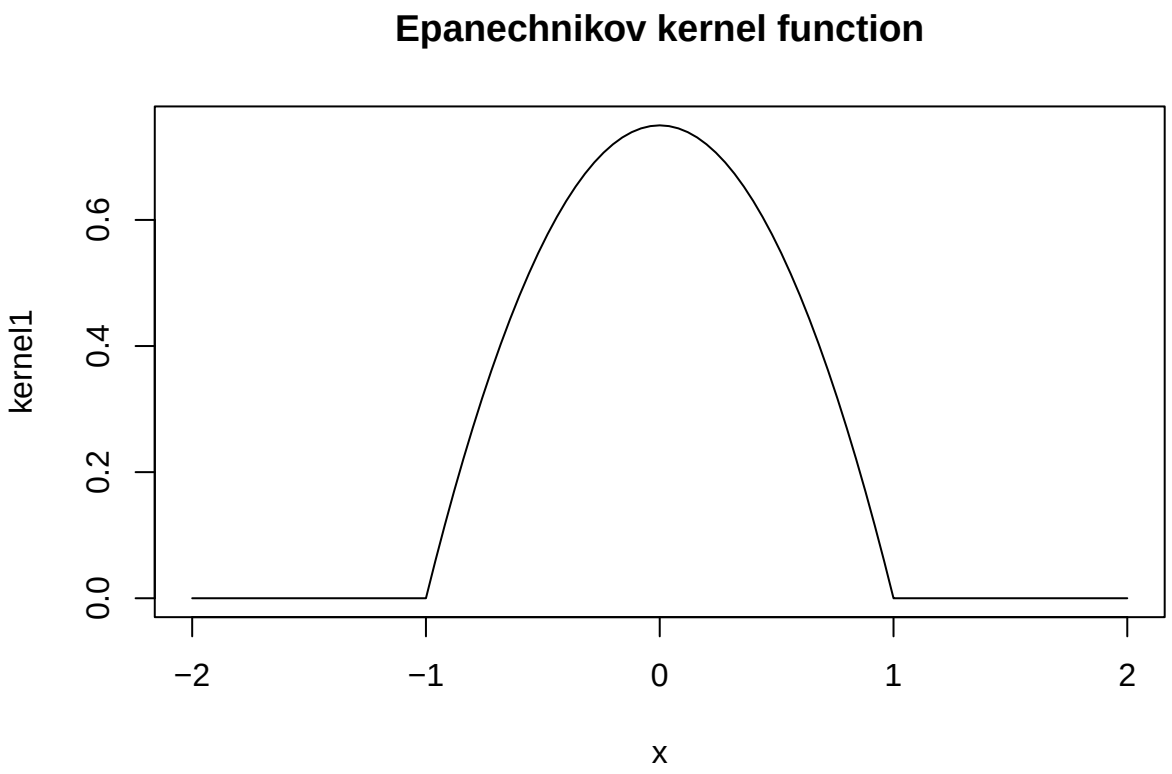
Simulate a dataset of 500 observations from the Gamma distribution with the shape pa- rameter 10 and the rate parameter 1. Set seed to 2425 to ensure comparability.

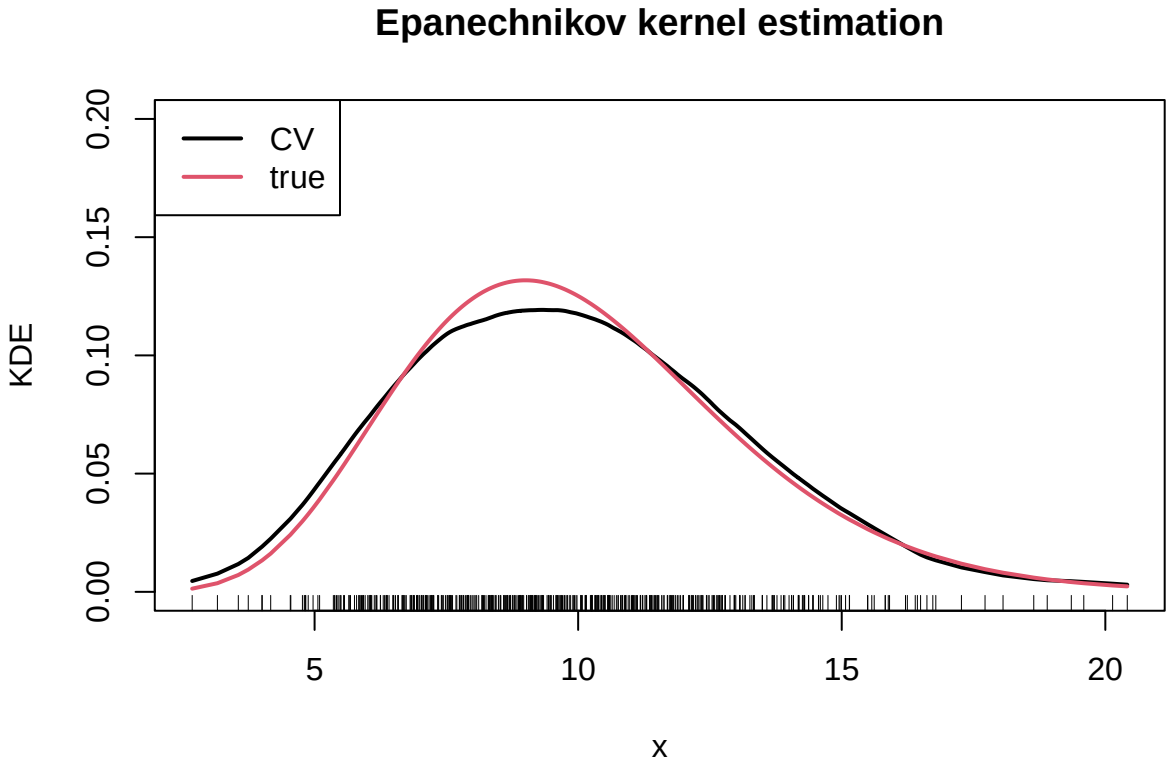
(a) Obtain a parametric estimator for the density of these data, employing moment estimators for the parameters. (1 point)
Estimators

$$alpha := \frac{Mean(X)^2}{Var(X)}, beta := \frac{Var(X)}{Mean(X)}$$

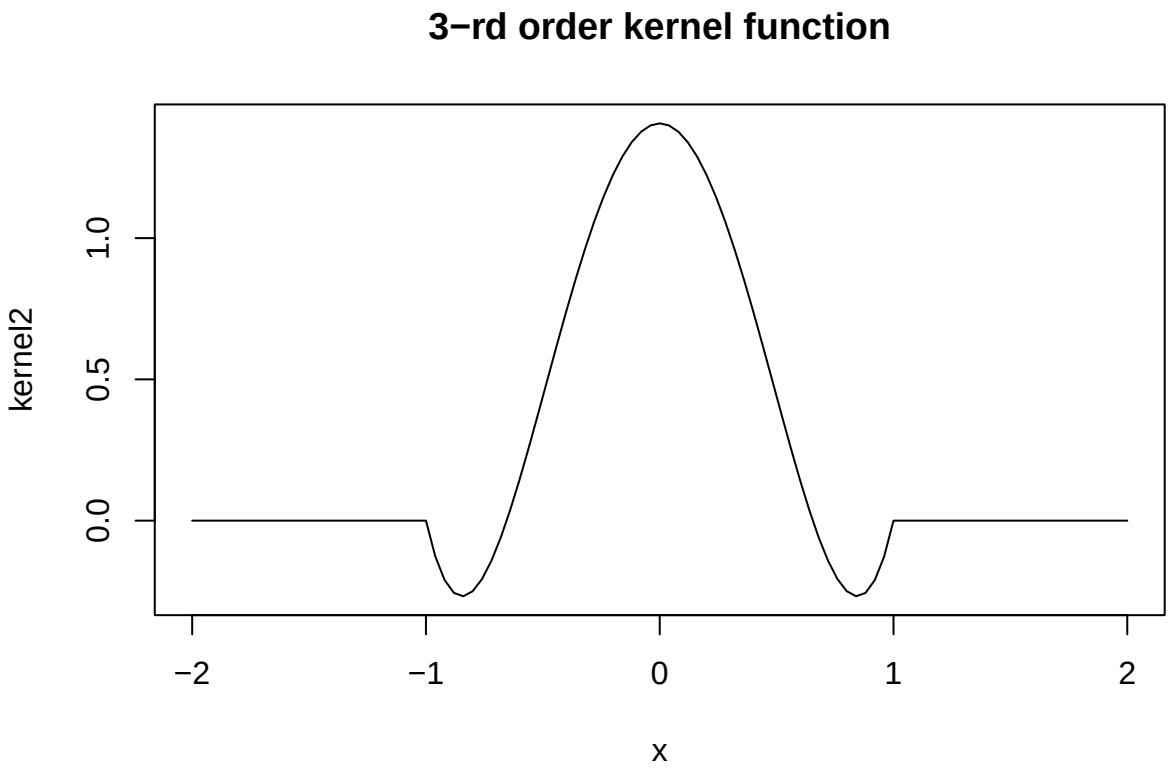


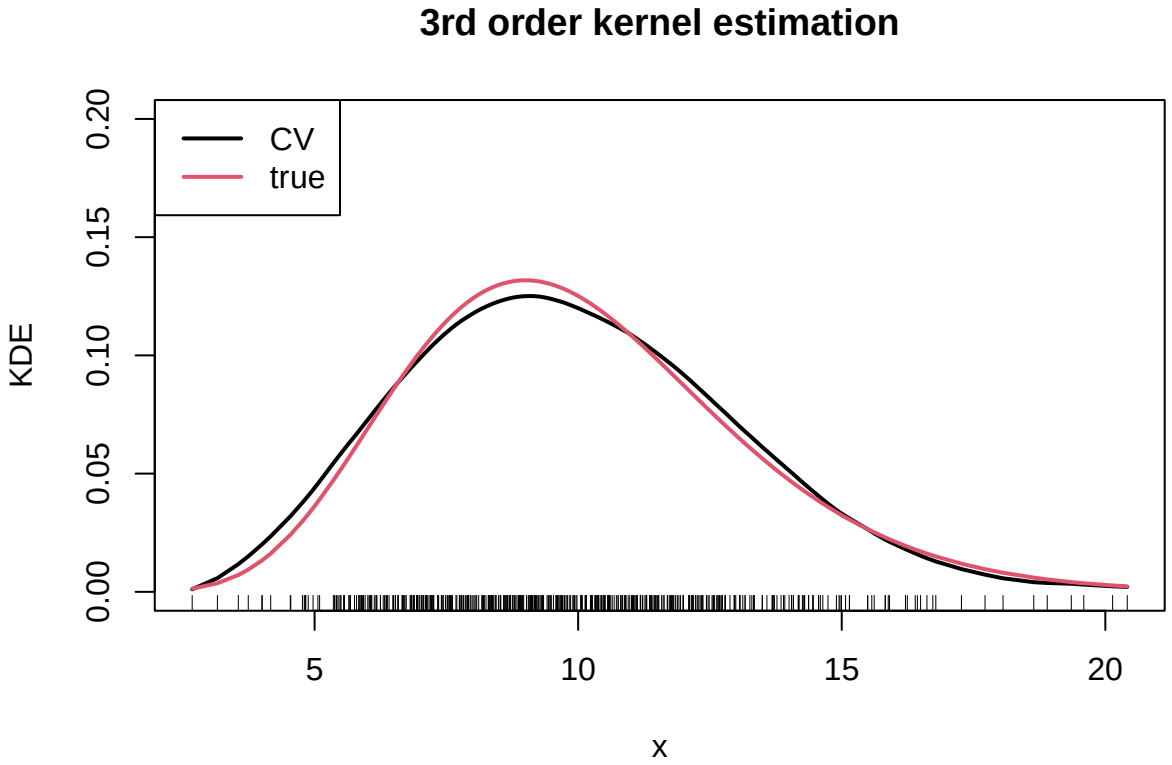
(b) Obtain a kernel density estimator with Epanechnikov kernel and a bandwidth chosen by cross-validation. Ensure that the cross-validation criterion has a global minimum by plotting it on a suitable range of bandwidths. (2 points)



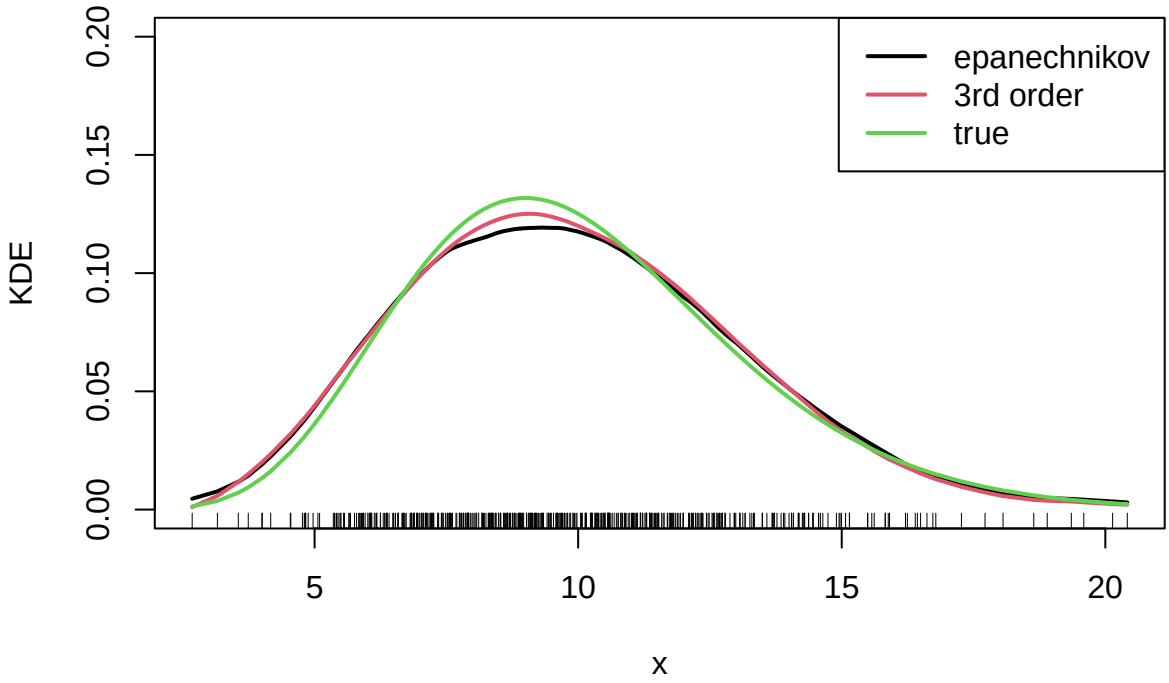


(c) Obtain a kernel density estimator as in (b), replacing Epanechnikov kernel by a kernel of order 3 (1 point)





(d) Plot estimators obtained in (a) – (c) together with the true density on one plot, putting a legend. Comment on the results. Compare also the bandwidths obtained in (b) and (c): which one is larger and why (give theoretical justification)? (2 points)




```
h1.cv
```

```
## [1] 2.307834
```

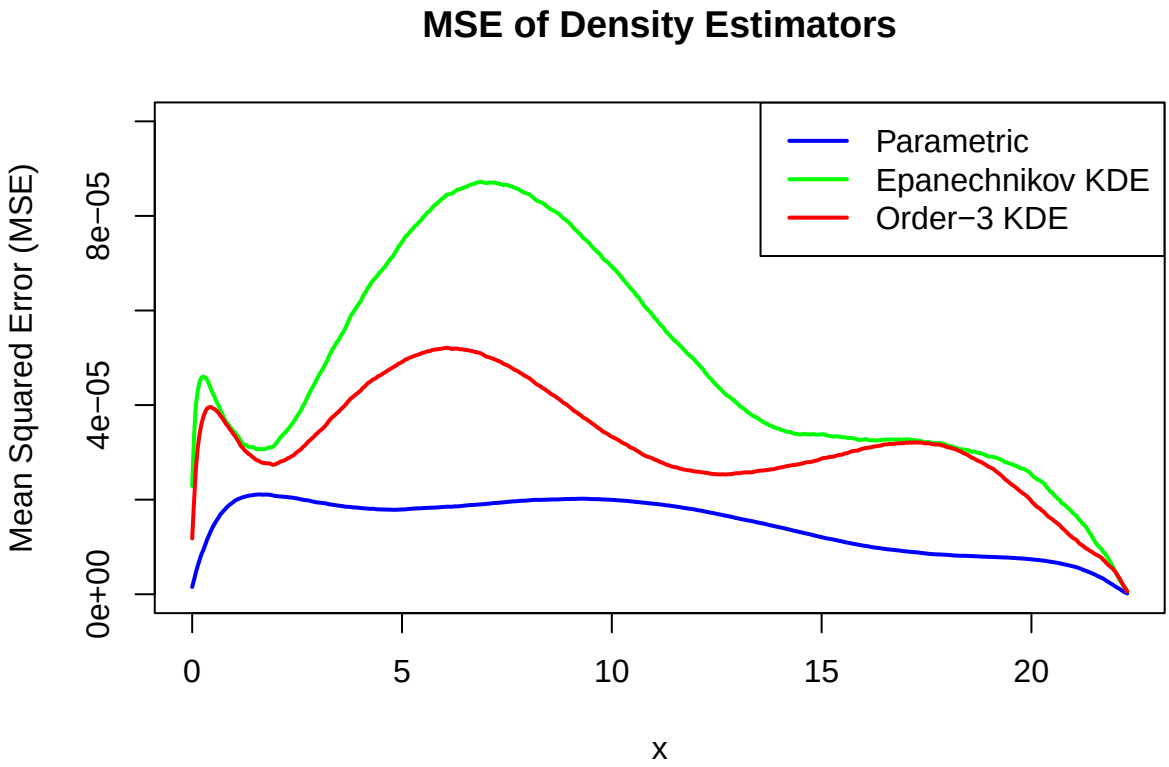
```
h2.cv
```

```
## [1] 4.931752
```

- 3rd order kernel has bigger smoothness, compared to 1st order epanechnikov kernel. For smoother kerels we need to have bigger bandwidth to prevent them from overfitting to data. If we would decrease brandwidth it would lead to decrease of bias term and increase of varience. kde would oscillate around true pdf.

Subtask 2

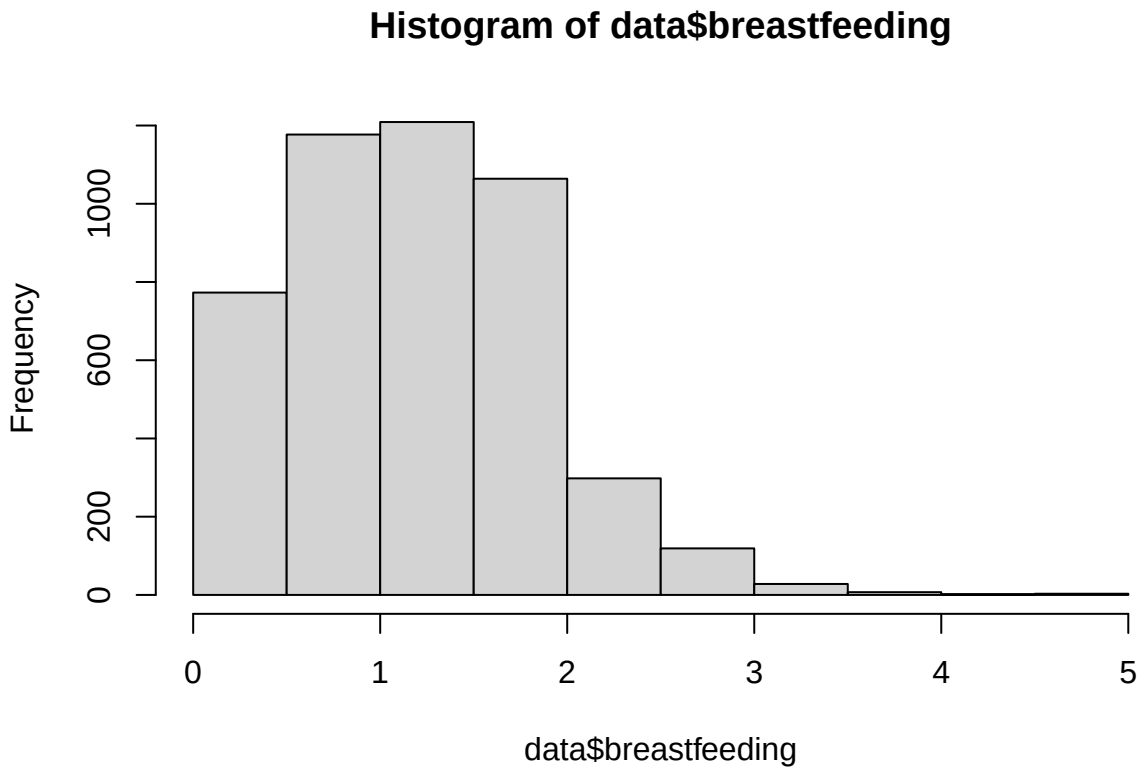
2. Simulate 300 samples as in 1, setting seed again to 2425. Estimate each of these samples by methods from 1(a)-1(c). With this, obtain a Monte Carlo estimator of the mean squared error at each point x of all three estimators (parametric, kernel density estimator with a first order kernel and kernel density estimator with a third order kernel). Plot the (estimated) mean squared errors of all three density estimators as a function of x and comment on the results (provide theoretical justification). (2 points)



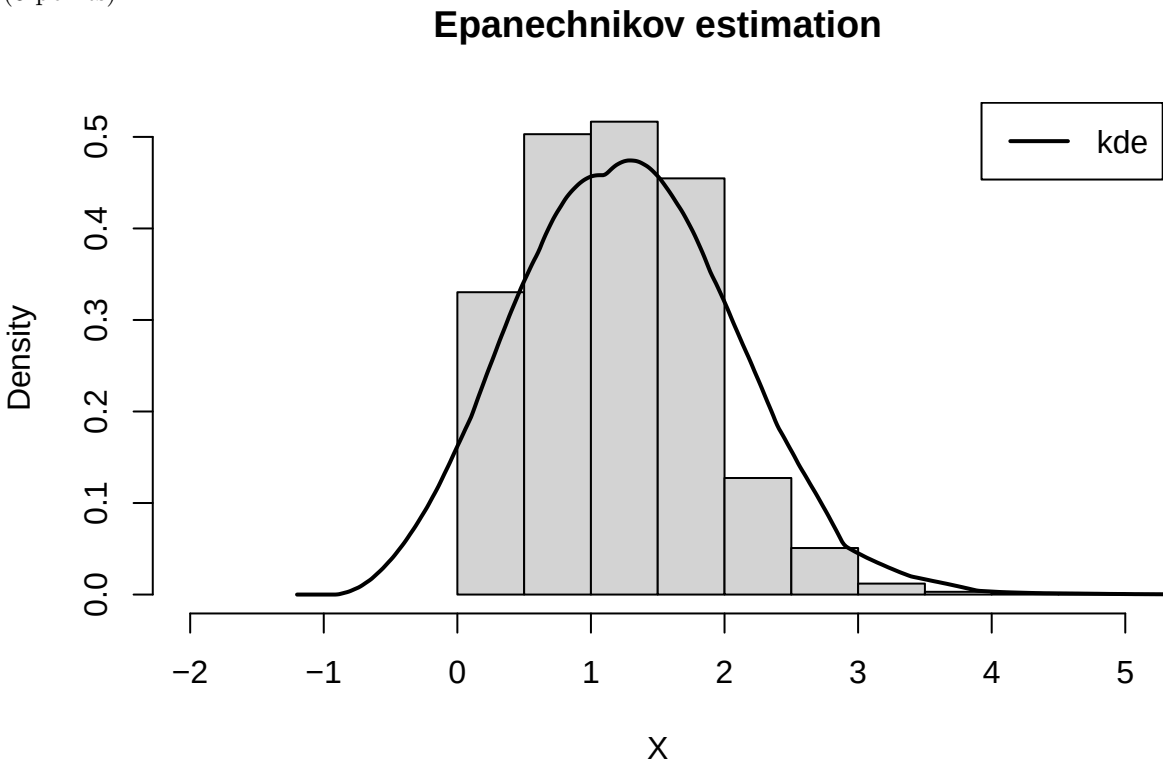
- There is a theorem that shows, that max likelihood estimator, if exists, is an efficient estimator of distribution parameters. Which means, that no estimators can be better. In this case, when sampling from known distribution (gamma), which has max lkelihood estimators for parameters, the best MSE-wise estimator is the parametric one. Order 3 estimator with properly chosen bandwidth also has better estimation then 1st order one, which is shown on graph.

Problem 3

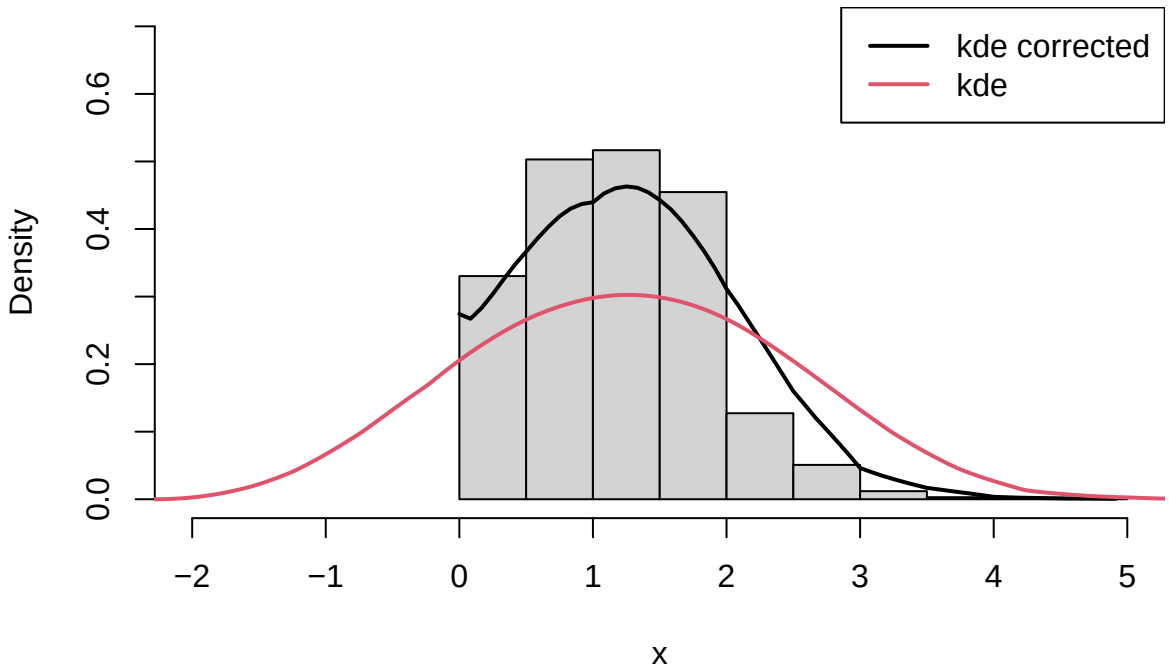
Read the dataset "Kenya DHS" into R and consider the variable breastfeeding, which gives the duration of breastfeeding in months. First, redefine this variable to give the duration in years and plot its histogram. Argue why and at which boundary a correction necessary. Next, estimate the density of this variable employing a kernel density estimator with Epanechnikov kernel and an appropriate boundary correction suggested by Gasser, Mueller. (1979), as given in Lecture 4. Set the bandwidth to h = 0.8. Compare this boundary corrected estimator with a usual kernel density estimator, that uses the same kernel and the same bandwidth, putting both on the histogram. Comment on the results.



(6 points)



Correction is required in x = 0.



- Corrected estimator correctly deals with 0 point. With this bandwidth corrected estimator deas better not only with 0, but also with density itself.