




kdensity: An R package for kernel density estimation with parametric starts and asymmetric kernels

Jonas Moss¹ and Martin Tveten¹

¹ University of Oslo

DOI: [10.21105/joss.01566](https://doi.org/10.21105/joss.01566)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Submitted: 11 July 2019

Published: 03 October 2019

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

It is often necessary to estimate a probability density non-parametrically, that is, without making strong parametric assumptions such as normality. This R (R Core Team, 2019) package provides a non-parametric density estimator that can take advantage of some of the knowledge the user has about the probability density.

Kernel density estimation (Silverman, 2018) is a popular method for non-parametric density estimation based on placing kernels on each data point. Hjort & Glad (1995) extended kernel density estimation with *parametric starts*. The parametric start is a parametric density that is multiplied with the kernel estimate. When the data-generating density is reasonably close to the parametric start density, kernel density estimation with that parametric start will outperform ordinary kernel density estimation.

Moreover, when estimating densities on the half-open interval $[0, \infty)$ and bounded intervals, such as $[0, 1]$, symmetric kernels are prone to serious boundary bias that should be corrected (Marron & Ruppert, 1994). Asymmetric kernels have been designed to avoid boundary bias and many of them are implemented in *kdensity* in addition to the classical symmetric kernels. For the unit interval, the Gaussian copula kernel of M. Jones & Henderson (2007) and the beta kernels of Chen (1999) are supported. The gamma kernel of Chen (2000) is available for the half-open interval.

The supported non-parametric starts include the normal, Laplace, Gumbel, exponential, gamma, log-normal, inverse Gaussian, Weibull, Beta, and Kumaraswamy densities. The parameters of all parametric starts are estimated using maximum likelihood. The implemented bandwidth selectors are the classical bandwidth selectors from *stats*, unbiased cross-validation, the Hermite polynomial method from Hjort & Glad (1995), and the tailored bandwidth selector for the Gaussian copula method of M. Jones & Henderson (2007). User-defined parametric starts, kernels and bandwidth selectors can also be set.

Several R packages deal with kernel estimation, see Deng & Wickham (2011) for an overview. While no other R package handles density estimation with parametric starts, several packages supports methods that handle boundary bias. Hu & Scarrott (2018) provides a variety of boundary bias correction methods in the *bckden* functions. Nagler & Vatter (2019) corrects for boundary bias using probit or logarithmically transformed local polynomial kernel density estimation. A. T. Jones, Nguyen, & McLachlan (2018) corrects for boundary bias on the half line using a logarithmic transform. Duong (2019) supports boundary correction through the *kde.boundary* function, while Wansouwé, Somé, & Kokonendji (2015) corrects for boundary bias using asymmetric kernels.

The following example uses the *airquality* data set from the built-in R package data sets. Since the data is positive we use Chen's gamma kernel. As the data is likely to be better approximated by a gamma distribution than a uniform distribution, we use the gamma parametric start. The plotted density is in figure 1, where the gamma distribution

with parameters estimated by maximum likelihood is in red and the ordinary kernel density estimate in blue. Notice the boundary bias of the ordinary kernel density estimator.

```
# install.packages("kdensity")
library("kdensity")
kde = kdensity(airquality$Wind, start = "gamma", kernel = "gamma")
plot(kde, main = "Wind speed (mph)")
lines(kde, plot_start = TRUE, col = "red")
lines(density(airquality$Wind, adjust = 2), col = "blue")
rug(airquality$Wind)
```

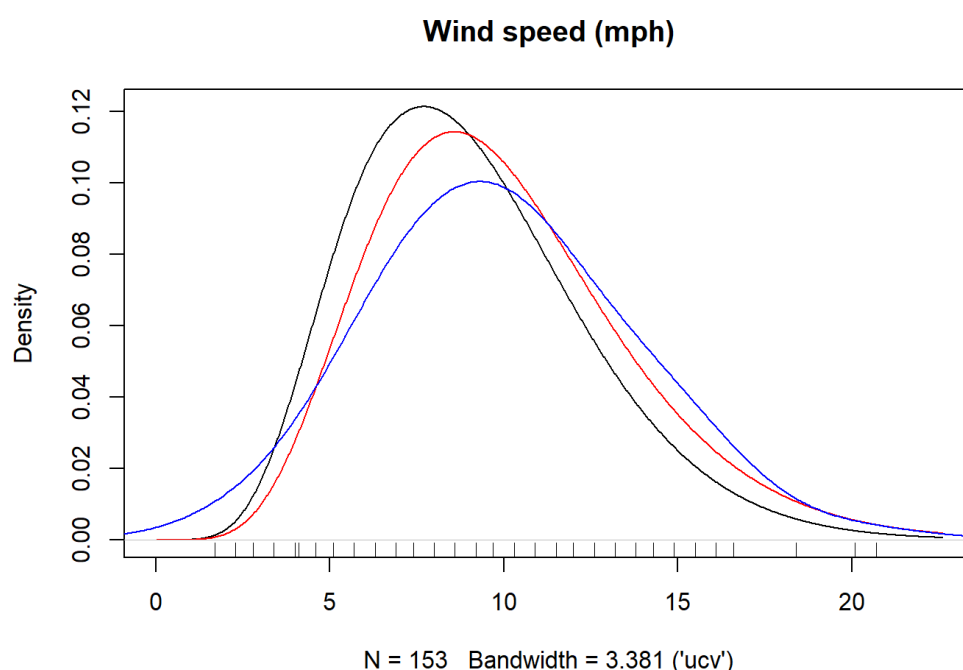


Figure 1: The *airquality* data set. Kernel density estimate in black and estimated gamma distribution in red.

References

- Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2), 131–145. doi:[10.1016/S0167-9473\(99\)00010-9](https://doi.org/10.1016/S0167-9473(99)00010-9)
- Chen, S. X. (2000). Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3), 471–480. doi:[10.1023/A:1004165218295](https://doi.org/10.1023/A:1004165218295)
- Deng, H., & Wickham, H. (2011). Density estimation in r. Retrieved from <http://vita.had.co.nz/papers/density-estimation.pdf>
- Duong, T. (2019). *Ks: Kernel smoothing*. Retrieved from <https://CRAN.R-project.org/package=ks>
- Hjort, N. L., & Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 882–904. doi:[10.1214/aos/1176324627](https://doi.org/10.1214/aos/1176324627)

- Hu, Y., & Scarrott, C. (2018). *evmix*: An R package for extreme value mixture modeling, threshold estimation and boundary corrected kernel density estimation. *Journal of Statistical Software*, 84(5), 1–27. doi:[10.18637/jss.v084.i05](https://doi.org/10.18637/jss.v084.i05)
- Jones, A. T., Nguyen, H. D., & McLachlan, G. J. (2018). LogKDE: Log-transformed kernel density estimation. *Journal of Open Source Software*, 3(28), 870. doi:[10.21105/joss.00870](https://doi.org/10.21105/joss.00870)
- Jones, M., & Henderson, D. (2007). Kernel-type density estimation on the unit interval. *Biometrika*, 94(4), 977–984. doi:[10.1093/biomet/asm068](https://doi.org/10.1093/biomet/asm068)
- Marron, J. S., & Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4), 653–671. doi:[10.1111/j.2517-6161.1994.tb02006.x](https://doi.org/10.1111/j.2517-6161.1994.tb02006.x)
- Nagler, T., & Vatter, T. (2019). *Kde1d: Univariate kernel density estimation*. Retrieved from <https://CRAN.R-project.org/package=kde1d>
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge. doi:[10.1201/9781315140919](https://doi.org/10.1201/9781315140919)
- Wansouwé, W. E., Somé, S. M., & Kokonendji, C. C. (2015). *Ake: Associated kernel estimations*. Retrieved from <https://CRAN.R-project.org/package=Ake>