Article type: Focus Article

Filtered Kernel Density Estimation **Article ID**

David Marchette

Naval Surface Warfare Center

Keywords

nonparametric, probability density, kernel estimator, mixture model, multiple bandwidths

Abstract

This article describes a multiple-bandwidth version of the kernel estimator for nonparametric probability density estimation, in which the bandwidths are chosen using a set of functions, called filter functions, which determine the support of the density appropriate to the different bandwidths. These filter functions are usually defined using a normal mixture fit to the data. Thus the estimator uses different bandwidths in different regions of the support of the distribution, as controlled by the filter functions.

Except for the histogram, the kernel estimator is probably the most popular nonparametric density estimator in use today. Given iid $\{x_1, \ldots, x_n\} \in \mathbb{R}$, a kernel function K and a bandwidth h, the kernel estimator is defined as

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{1}$$

One of the problems associated with the kernel estimator is the reliance on a single bandwidth to define the amount of smoothing of the data. Often, data appear to be drawn from distributions which have very different scales in different regions (see Figure 1 below). Several solutions have been proposed to solve this problem. For example, Abramson [1982] uses a pilot estimate of the density to provide local bandwidths at each observation x_i. Other researchers suggest providing a collection of kernel estimators using a range of values for h. This gives a kind of multi-resolution view of the density. See for example Marron and Chaudhuri [1998] and Zhang and Marron [2005] for an approach along these lines, called SiZer.

We will describe a compromise between the single bandwidth kernel estimator and the adaptive kernel estimator of Abramson [1982]. This uses a user-supplied set of "filters" to split the data into regions in which different bandwidths are appropriate. As will be seen, the technique is more general than this, allowing the overlap of different regions to have a "mixture" of bandwidths.

The Estimator

The filtered kernel estimator first appeared in Marchette et al. [1996]. The idea is to use knowledge about local scale to improve the fit of a kernel estimator. The scale information is coded in a set of "filter functions" which indicate the regions in which the different scales are active.

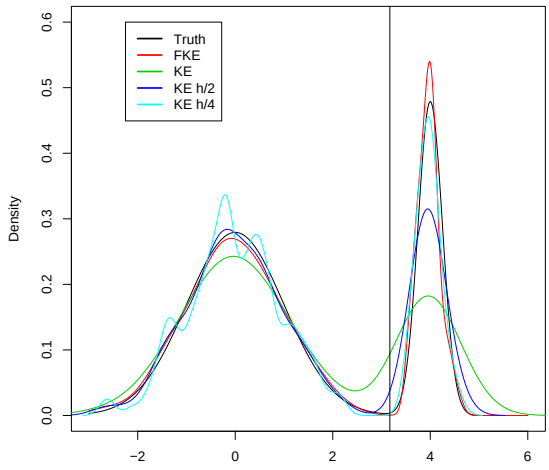

Figure 1: A probability density consisting of a mixture of two normals. We illustrate four estimators computed on 200 random variates drawn from this model: the filtered kernel estimator, using a two component mixture of normals fit to the data to define the filtering functions; a standard kernel estimator; the kernel estimator with the bandwidth inflated or reduced in order to match the two components.

Figure 1 shows the basic motivation for the filtered kernel estimator. In this, a simple mixture of two normals has two different scales: a standard deviation of 1 on the left, and of 0.25 on the right. If we wanted to estimate the probability density for data drawn from this distribution, the standard kernel estimator would leave us with a dilemma: do we under-smooth the left, over-smooth the right, or both? Knowing the situation, the solution is clear: we should use a relatively large bandwidth for points to the left of the vertical line, and a smaller bandwidth for points to the right, and produce a "mixture" of kernel estimators. We will now make this explicit.

Let $X = \{x_1, \ldots, x_n\}$ be iid random real numbers. Let K be a kernel as in the standard kernel estimator. Given a collection of filter functions $\{\rho_1, \ldots, \rho_m\}$, with $0 \leq \rho_j(x) \leq 1$ and $\sum_{j=1}^{m} \rho_j(x) = 1$ for all $x \in \mathbb{R}$, and m bandwidths $h_j > 0$, we define

the filtered kernel estimator to be

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\rho_j(x_i)}{h_j}K\left(\frac{x-x_i}{h_j}\right) \tag{2}$$

This general formula is useful if one has prior information of where the density can be expected to have different scales. The user can then design filter functions to appropriately model the density in these different regions. For example, in Figure 1, the user may choose the filter functions to be the indicator functions for the regions on either side of the vertical line. In this case, the filtered kernel estimator is simply a "cobbling together" of two kernel estimators, each one best suited to its distinct region.

A less general, but perhaps more usual application of the filtered kernel estimator, is to use a mixture model fit to the data to define the filter functions. Let

$$\widetilde{f}(x) = \sum_{j=1}^{m}\pi_j g_j(x). \tag{3}$$

Let σ_j be the standard deviation of the jth component g_j. We then define the single bandwidth version of the filter kernel estimator with filter functions defined by the posteriors,

$$\rho_j(x) = \frac{\pi_j g_j(x)}{\widetilde{f}(x)},$$

as

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\rho_j(x_i)}{h\sigma_j}K\left(\frac{x-x_i}{h\sigma_j}\right) \tag{4}$$

We refer to $\widetilde{f}$ as the *filtering mixture*.

Note that, just like the standard kernel estimator, this can easily be extended to multivariate data. One advantage, is that since we effectively have different kernels in different regions of the support (guided by the filter functions), the kernels can now have different covariances in these different regions, better fitting the structure of the data.

This formulation provides guidance for selecting h. Suppose we believed that the true density was in fact the mixture $\widetilde{f}(x)$. Then the optimal (in mean integrated squared error) bandwidth h_{opt} is:

$$h_{\text{opt}} = \left(\frac{\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{B_{jk}}{\sqrt{\sigma_j^2+\sigma_k^2}}}{\sqrt{2\pi}n\sum_{j=1}^{m}\sum_{k=1}^{m}A_{jk}\sigma_j^2\sigma_k^2}\right)^{\frac{1}{5}}, \tag{5}$$

where

$$A_{jk} = \pi_j\pi_k\int g_j''(x)g_k''(x)dx,$$

and

$$B_{jk} = \pi_j \pi_k \int \frac{g_j(x) g_k(x)}{\widetilde{f}(x)} d(x).$$

Thus, we can use the mixture as both the set of filter functions and to find the optimal bandwidth. Since we use the variances of the terms locally, we end up with a locally adaptive kernel estimator, with only one parameter (given the filtering mixture model). In the examples that follow, we will always use a Gaussian mixture model for the filtering mixture.

Examples

In Figure 1 we illustrate the filtered kernel estimator for the motivating example. Given 200 observations drawn from the true distribution, we fit a two component normal mixture model and use this to define the filter functions as above. The resultant estimate is shown in red. Notice that it simultaneously does a very good job of fitting both components, and also the tails and between-component region. The standard kernel estimator, illustrated with three versions in the other colors, cannot simultaneously fit both components. In this case, the standard method for bandwidth selection using a normal fit to the data (green curve) does not work. This is because the spread of the two components increases the variance. Instead, we show two kernel estimators giving good fits to the two components. Fitting the left hand component over-smooths the right, and fitting the right under-smooths the left. The filtered kernel estimator allows appropriate smoothing for both components.

Figures 2 and 3 demonstrate a different situation, one in which a density with a long tail and a skewed mode is the target. The data are drawn from the log-normal distribution (the figures represent different numbers of observations, but otherwise the details are the same). We first select a two component normal mixture by looking at the histogram and determining by eye a rough idea of where the mode and tail of the distribution are. The mixture selected is:

$$\frac{8}{10} N(\frac{1}{2}, \frac{4}{10}) + \frac{2}{10} N(3, 5).$$

Note that this mixture was deliberately chosen to be only a very crude fit to the data. In addition, a mixture was fit to the data using this model as a starting point in an EM algorithm. These two models are then used to define filtered kernel estimators, which are plotted in the figure, compared to a standard kernel density estimate.

In Figure 2 a case can be made that the filtered kernel estimator is doing a better job of estimation near the mode of the distribution then the standard kernel estimator, whichever mixture we use. For these data, the estimators are pretty much equivalent in the right tail. In Figure 3 the filtered kernel estimators are superior to the standard kernel estimator pretty much throughout. Further, even the crude initial mixture, having only a tenuous relationship to the actual data, produces a better estimator. The point of these examples is that unlike the situation in Figure 1, the true distribution is not a two component mixture model, and yet such a model can be easily used to impart local

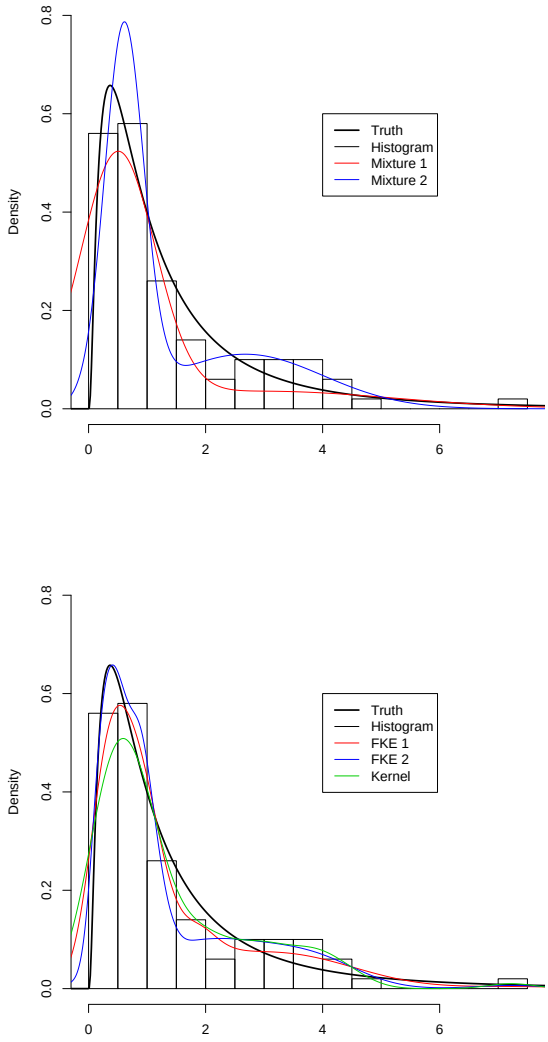

Figure 2: Log-normal example: 100 observations drawn from a log-normal distribution. The top figure shows the data compared to two mixture models. The first, in red, was selected purely by eye as compared to the histogram, intended to indicate the mode and tail of the distribution. The second, in blue, was fit using the EM algorithm with the first mixture as a starting point. The bottom figure shows the two filtered kernel estimators using the two mixture models compared to a standard kernel estimator.

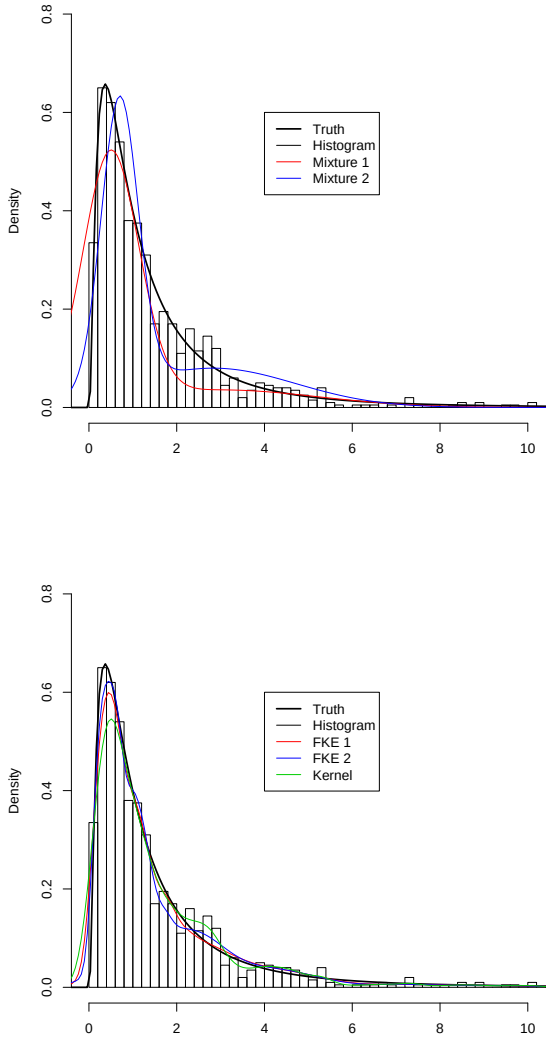

Figure 3: Log-normal example: 100 observations drawn from a log-normal distribution. The figures are as in Figure 2.

scale information to the filtered kernel estimator, providing an improvement over the standard kernel estimator.

Conclusion

The filtered kernel estimator is a relatively simple, easy to use method for multiscale probability density estimation. It allows the user to provide information about different scales in different regions of the data through the filtering functions. The most natural way to do this, in the author's opinion, is through the use of a mixture model fit to the data. In this case, only a single bandwidth need to be fit (once the mixture has been selected), and the variances (or covariances, in the case of multidimensional data) are used to provide the local scales.

When used with a mixture model or other density estimate providing the filtering functions, the filtered kernel estimator can be used to provide a nonparametric evaluation of the parametric fit. Essentially, by using the nonparametric estimator "most related to" the parametric one, the user can use the comparison to the nonparametric fit for model selection and validation. These ideas are discussed in more detail in Priebe and Marchette [2000].

References

I. S. Abramson. On bandwidth variation in kernel estimates – a square root law. *The Annals of Statistics*, 10:1217–1223, 1982.

D. J. Marchette, C. E. Priebe, G. W. Rogers, and J. L. Solka. Filtered kernel density estimation. *Computational Statistics*, 11(2):95–112, 1996.

J. S. Marron and P. Chaudhuri. Significance of features via SiZer. In Brian Marxh and Herwig Friedl, editors, *Proceedings of the 13th International Workshop on Statistical Modelling*, pages 65–75, 1998.

C. E. Priebe and D. J. Marchette. Alternating kernel and mixture density estimates. *Computational Statistics and Data Analysis*, 35:43–65, 2000.

J. T. Zhang and J. S. Marron. SiZer for smoothing splines. *Computational Statistics*, 20:481–502, 2005.

Cross-References

Kernel density estimation Finite mixture distributions