Statistics & Probability Letters 7 (1989) 195–199
North-Holland

December 1988

CANONICAL KERNELS FOR DENSITY ESTIMATION

J.S. MARRON *

Department of Statistics, University of North Carolina, Chapel Hill, NC, USA

D. NOLAN **

Department of Statistics, University of California, Berkeley, CA, USA

Received May 1987
Revised April 1988

Abstract: The kernel function in density estimation is uniquely determined up to a scale factor. In this paper, we advocate one particular rescaling of a kernel function, called the canonical kernel, because it is the only version which uncouples the problems of choice of kernel and choice of scale factor. This approach is useful for both pictoral comparison of kernel density estimators and for optimal kernel theory.

AMS 1980 Subject Classification: 62G05.

Keywords: canonical kernels, density estimation, optimal kernels, smoothing

1. Introduction and motivation

For practical kernel density estimation, an obvious way to choose among kernel functions is to compare plots of estimates for the data set of interest. A major problem with this approach occurs when the standard representations of kernel functions are coupled with identical bandwidths. For then the estimates are based on different amounts of smoothing and the comparison becomes meaningless. To illustrate this point, Figures 1a and b show kernel estimates using the standard normal and triweight kernels, for a simulated data set of size 200 from a mixture density (0.7)Beta(4,8) + (0.3)Beta(40,20), where standard normal and triweight kernels are used with the same bandwidth $h = 0.066$.

Notice that the triweight kernel estimate is much less smooth than the normal. Although the bandwidth is the same for both estimates, the "effective bandwidth", observed in the spread of the kernel function at the bottom of each picture, is not the same. Local averaging drives the density estimator, and it is very different for these two estimates. This makes visual comparison of kernel estimates nearly impossible because the critical problem of bandwidth selection is confounded with the problem of comparison of kernel functions.

A well known way of overcoming this problem is to readjust the bandwidth for each of the different estimators, by making the variance-squared bias trade off in Mean Square Error the same (see Scott, 1976). Our approach is to rescale the kernel functions, allowing the same bandwidth to represent the same amount of smoothing across kernels. In Section 2 it is seen that each kernel has exactly one rescaling, called the canonical representation of the kernel, that allows this sensible and convenient comparison.

* Research partially supported by NSF Grant DMS-84-00602 and NSF Grant DMS-8701201.
** Research partially supported by NSF Grant DMS-850-3347, NSF Grant DMS-840-3230 and AFSOR Grant No. F49620 85C 0144.

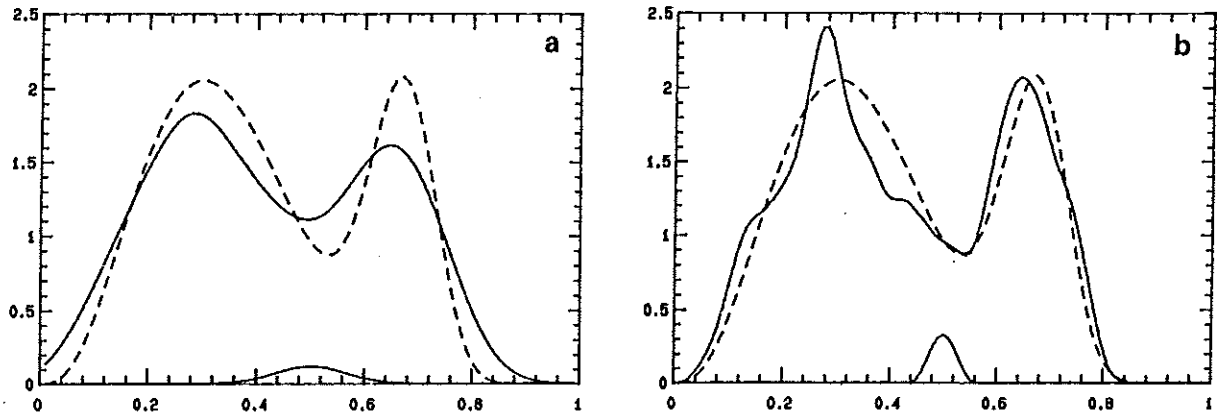

Fig. 1. Dashed curve is a mixture probability density function, (0.7)Beta(4.8)+(0.3)Beta(40,20). Solid curves are kernel density estimates with bandwidth 0.066 together with one vertically rescaled kernel function, for (a) the standard normal kernel, (b) the triweight kernel.

Since the canonical rescaling of a kernel separates the problems of kernel and bandwidth selection, it provides a fresh approach to the problem of optimal kernel selection. Section 3 shows how this idea can be used to considerably strengthen the conventional notion of optimal kernel.

Section 4 gives the form of the canonical kernel for a rich family of kernel functions that includes essentially all kernels used in practice. Also given are specific values of the appropriate constants for the examples in Figure 1.

It is important to note that the ideas of this paper carry over entirely to kernel estimators in other nonparametric curve estimation settings, such as regression, spectral density and hazard function estimation. We restrict explicit statement of our results to the density estimation context, because that is the simplest and most widely treated in the literature.

2. Comparison of kernel estimators

The kernel estimator, of a probability density f, based on a random sample $X_1, \ldots, X_n$ from f, is usually defined as

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{2.1}$$

where

$$K_h(\cdot) = K(\cdot/h)/h. \tag{2.2}$$

The function K is called the kernel and is usually taken to be a probability density. The constant h is called the bandwidth or smoothing parameter; it controls the amount of smoothing or local averaging. To see this, consider the plots in Figure 1 where the function $K_h(\cdot)$ is shown at the bottom of each plot. The estimator centers a kernel about each of the observations, then averages them together.

The choice of h is crucial to the performance of the estimator. However, in this paper we choose to concentrate on the choice of K. See Devroye and Györfi (1984) and Silverman (1986) for discussion of the classical theory and of subjective methods for choosing h. For an access to the literature on data based methods for choosing h see Marron (1986) and Marron (1988).

Note that since h controls the scale of K_h in (2.1), the function K may be rescaled without changing the estimator at all (provided the rescaling is absorbed by suitably changing h as well). Hence it makes sense to express the choice between kernel functions as a choice among equivalence classes of kernel functions, where two kernels are considered equivalent when they are rescalings of each other. The main point of this paper is that there is a best way of producing a

representative element from each equivalence class that facilitates comparison of kernels or classes.

An approach to finding a representative member of each equivalence class may be found in Epanechnikov (1969), where the kernel which satisfies

$$\int x^2 K(x)\,dx = 1$$

is used. Another approach, taken by Gasser, Müller and Mammitzsch (1985) insists the support of the kernel be $[-1, 1]$. A drawback to both of these methods is that they are rather arbitrary. Neither makes an attempt to allow a single choice of bandwidth to give the same amount of smoothing for different kernels, although the former comes fairly close because the variance of the kernel function (thinking of it as a probability density) does provide a rough quantification of the type of "scale" that is pertinent here. While Epanechnikov's variance adjustment is not far from suitable, it is still rather ad hoc, and the main point of this paper is to provide a firm mathematical basis for choosing a particular kernel rescaling, and for consideration of kernel choice problems.

To establish this firm basis, we need to somehow quantify the amount of smoothing. A standard way for doing this is to use the Mean Square Error,

$$\mathrm{MSE} = E\left[\hat{f}_h(x) - f(x)\right]^2,$$

or the Mean Integrated Square Error,

$$\mathrm{MISE} = \int \mathrm{MSE}\,dx.$$

We explicitly use the MISE here, although the same theory, with exactly the same conclusion, is very simply adapted to the MSE. MISE (and also MSE) is convenient for quantifying the smoothing problem because it allows a variance-squared bias decomposition that admits the asymptotic representation:

$$\mathrm{MISE} \cong n^{-1}h^{-1}\int K^2 + h^4\left[\int x^2 K\right]^2\left[\int (f''/2)^2\right], \tag{2.3}$$

as $n \to \infty$, $h \to 0$ with f twice continuously differentiable. See, for example, (3.20) of Silverman (1986).

To see that (2.3) quantifies the smoothing trade-off, note that the first term on the right side of (2.3) gets large when h is too small, i.e. when the curve is too wiggly, because there is too much variance caused by averaging over too few points. On the other hand, the second term gets large when h is too big, i.e. when features of the actual density are smoothed away, because there is too much bias introduced by averaging over too large a neighborhood.

We seek a representative of the possible rescalings of the form

$$K_\delta(\cdot) = K(\cdot/\delta)/\delta,$$

that separates h and K in

$$n^{-1}h^{-1}\int K_\delta^2 + h^4\left[\int x^2 K_\delta\right]^2\left[\int (f''/2)^2\right].$$

To do this, solve for δ such that the contribution made by K_δ to both terms equal each other:

$$\int K_\delta(x)^2\,dx = \left[\int x^2 K_\delta(x)\,dx\right]^2.$$

Integration by substitution gives

$$\delta_0 = \left[\int K^2\right]^{1/5}\left[\int x^2 K\right]^{-2/5}.$$

Observe that the kernel K_{δ_0} gives

$$\mathrm{MISE} \cong C(K_{\delta_0})[n^{-1}h^{-1} + h^4]\left[\int (f''/2)^2\right], \tag{2.4}$$

where

$$C(K) = \left[\int K^2\right]^{4/5}\left[\int x^2 K\right]^{2/5}. \tag{2.5}$$

In the proof of their Lemma 18 in Chapter 5, Devroye and Györfi point out the interesting fact that $C(K)$ is invariant within each equivalence class in the sense that for any δ_1 and δ_2,

$$C(K_{\delta_1}) = C(K_{\delta_2}).$$

Because δ_0 does not depend on the particular scaling of K and because $C(K)$ is invariant, any

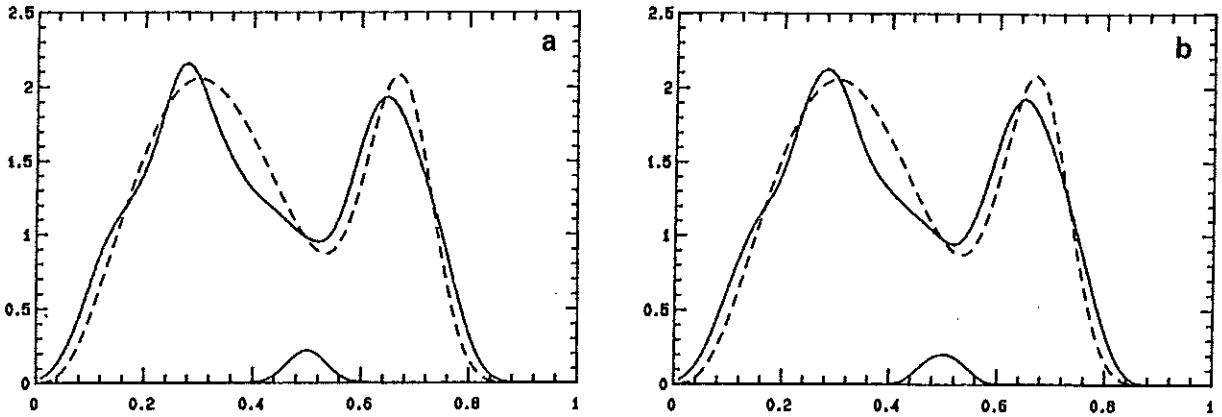

Fig. 2. Dashed curve is a mixture probability density function, (0.7)Beta(4.8)+(0.3)Beta(40,20). Solid curves are kernel density estimates with bandwidth 0.046 together with one vertically rescaled kernel function, for (a) the canonical normal kernel, (b) the canonical triweight kernel.

equivalence class of rescalings of K has a uniquely defined representative of that class, K_{δ_0}.

Equation (2.4) reinforces K_{δ_0} as a useful representative for visual comparison of kernel functions. In particular, if and only if K_{δ_0} is used, the kernel function no longer plays a role in the variance-squared bias trade off which quantifies the smoothing problem. Hence we call K_{δ_0} the canonical kernel for the equivalence class of rescalings of K.

To see how use of the canonical kernel makes it easy to compare kernel functions, consider Figure 2. This has the same setup as Figure 1, except now the canonical versions of the kernels and the L^2 optimal bandwidths, $h = 0.046$ are used. Note that it takes a very careful inspection before any difference at all can be observed between the two estimates.

3. Optimal choice of kernel

An added benefit of canonical kernels is that they provide a more rigorous basis for the standard optimal kernel theory. The problem of the confounding of h and K is typically handled by evaluating the right hand side of (2.3) at the value of h which minimizes it, and then solving the calculus of variation problem for the function K which minimizes the resulting expression. This approach is unsatisfactory because it essentially assumes that the minimizing value of h is known. Of course in practice the minimizer is not known, and the results of Hall and Marron (1987a, b) show that in a strong sense, estimates of the optimal h will typically be subject to a good deal of error.

If instead we reformulate the problem as a choice among equivalence classes of kernels, then we recognize the kernel that uncouples the problems of choosing h and K is the sensible approach. Note that while the calculus of variation problem for minimizing $C(K)$ in (2.4) looks different from that of Epanechnikov (1969) (discussed in Section 3.3.2 of Silverman), the invariance of $C(K)$ shows that they are the same. Thus when the canonical kernel is used to derive the optimal kernel the answer is the same as that of Epanechnikov.

Our approach makes it clear that the optimal choice of kernel does not depend on knowledge of the bandwidth. The Epanechnikov kernel is optimal for any h, i.e. any amount of smoothing. Note that the ideas of this section are related in spirit to some of those in Chapter 5 of Devroye and Györfi (1984).

Table 1

Kernel	α	C_α	δ_0
Uniform	0	$\frac{1}{2}$	$(\frac{9}{2})^{1/5}$ $\cong 1.3510$
Epanechnikov	1	$\frac{3}{4}$	$15^{1/5}$ $\cong 1.7188$
Biweight	2	$\frac{15}{16}$	$35^{1/5}$ $\cong 2.0362$
Triweight	3	$\frac{35}{32}$	$(\frac{9450}{143})^{1/5}$ $\cong 2.3122$
Normal	∞	–	$(\frac{1}{4\pi})^{1/10} \cong 0.7764$

4. Examples

One interesting family of kernels, which contains many of the kernels used in practice is

$$K^\alpha(x) = C_\alpha(1-x^2)^\alpha 1_{[-1,1]}(x), \tag{4.1}$$

where 1 denotes the indicator function and the constant C_α makes K^α a probability density:

$$C_\alpha = \Gamma(2\alpha+2)\Gamma(\alpha+1)^{-2}2^{-2\alpha-1}.$$

The first three columns of Table 1 show the values of α and C_α for the most common special cases. The normal kernel is not explicitly of the form (4.1); it is the degenerate case obtained by taking the limit as $\alpha \to \infty$.

It is simple to check that the rescaling factor δ_0, for each K^α, is

$$\delta_0 = 2^{-1/5}\Gamma(\alpha+1)^{-4/5}\Gamma(2\alpha+3)^{2/5}\Gamma(2\alpha+2)^{2/5} \times \Gamma(2\alpha+1)^{2/5}\Gamma(4\alpha+2)^{-1/5}.$$

The fourth column of Table 1 gives the value of δ_0 for these special cases.

The very large differences between the pictures of Figure 1 and those of Figure 2 are explained by the differences between the numbers appearing in the fourth column of Table 1.

References

Devroye, L. and L. Györfi (1984), *Nonparametric Density Estimation: The L_1 View* (Wiley, New York).

Epanechnikov, V.A. (1969), Nonparametric estimation of a multivariate probability density, *Theory of Probability and its Applications* **14**, 153–158.

Gasser, T., H.G. Müller and V. Mammitzsch (1985), Kernels for nonparametric curve estimation, *Journal of the Royal Statistical Society Series B* **47**, 238–252.

Hall, P. and J.S. Marron (1987a), Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation, *Probability Theory and Related Fields*, **74**, 567–581.

Hall, P. and J.S. Marron (1987b), The amount of noise inherent in bandwidth selection for a kernel density estimator, *Annals of Statistics* **15**, 163–181.

Marron, J.S. (1986), Will the art of smoothing ever become a science?, in: J.S. Marron, ed. *Function Estimates* (American Mathematical Society Series: Contemporary Mathematics 9) pp. 169–178.

Marron, J.S. (1988), Automatic smoothing parameter selection: A survey, *Empirical Economics*, to appear.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York).

Scott, D.W. (1976), Nonparametric probability density estimation by optimization theoretic techniques, Doctoral dissertation, Rice University, Houston, Texas.