Available online at www.sciencedirect.com
SCIENCE DIRECT
Economics Letters 83 (2004) 61–68
economics letters
www.elsevier.com/locate/econbase

Optimal asymmetric kernels

Karim M. Abadir[a,b], Steve Lawford[c,*]

[a] Department of Mathematics, University of York, Heslington, York YO10 5DD, UK
[b] Department of Economics, University of York, Heslington, York YO10 5DD, UK
[c] European Centre for Advanced Research in Economics and Statistics, Université Libre de Bruxelles, 50 Avenue F. D. Roosevelt, CP 114, B-1050 Brussels, Belgium

Received 9 January 2003; accepted 28 July 2003

Abstract

Kernels are typically justified as a smoothing device in nonparametric analysis. We provide alternative interpretations, which could lead to the use of asymmetric kernels. We thus derive the class of optimal asymmetric kernels, and analyse its main properties. We illustrate numerically its optimality by showing how well it fares in locating the mode and tail quantiles of some common asymmetric densities.

Keywords: Nonparametric analysis; Optimal asymmetric kernels; Integrated mean squared-error (IMSE); Tail quantiles; Income distribution

JEL classification: C14; C15

1. Introduction

Kernels are used in nonparametric analysis for smoothing purposes. For example, suppose data X_j, $j=1, 2, \ldots, n$, are available from a random sampling of a continuous density $f(x)$. A smoothed estimate $\hat{f}(x)$ of $f(x)$ may be obtained by using a kernel. In the simplest setting, $\hat{f}(x)=(nh)^{-1}\sum_{j=1}^{n} K((x-X_j)h^{-1})$, where where h is the smoothing parameter and $K(\cdot)$ is a chosen continuous kernel. The interpretation of each of

$$\kappa(x-X_j;h)\equiv h^{-1}K((x-X_j)h^{-1}), \qquad j=1,2,\ldots,n, \tag{1}$$

* Corresponding author. Fax: +32-2-650-4475.
E-mail address: steve_lawford@yahoo.co.uk (S. Lawford).

0165-1765/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2003.07.017

is one whereby the kernel is a weighting device, say a density function, and h^{-1} indicates the concentration of the assigned weights around the points X_j. The estimate $\hat{f}(x)$ is then the average of these $\kappa(\cdot)$. It has been shown that, in large samples, the optimal (in a mean squared-error sense) symmetric kernel is a quadratic one, the Epanechnikov kernel (Epanechnikov, 1969), and that its optimum window width h_e falls as the sample size n increases; e.g. see Silverman (1986) or Härdle and Linton (1994) for an introduction. In a sense, there is less need for smoothing as $n \to \infty$ because the empirical distribution becomes almost continuous. We now give additional different interpretations to kernels, then re-examine the optimality of the quadratic kernel.

In some statistical applications, one may assign some uncertainty as to the accuracy of the available data on X. For example, there may be errors in reporting, recording, etc., which mean that an observation X_j could have come from somewhere else in the neighbourhood of X_j, say x, with probability $\kappa(x - X_j; h)$. Kernels can then be given an errors-in-variables interpretation[1]. To take a simple example, it is likely that economic agents will misreport their taxable incomes, whether by malice or by ignorance of tax-rebate entitlements. If there is a lack of explicit quantifiable evidence, it may not be known how to model this tendency explicitly. If one believes that a hypothetically fully efficient version of an existing tax system could raise more tax receipts, it could be of interest to let $\kappa(\cdot)$ be asymmetric with positive skewness. An important question arises: is the quadratic (symmetric) Epanechnikov kernel still optimal and, if not, what is the optimal one?

One need not subscribe to the income distribution example to consider asymmetric kernels. An alternative justification for asymmetric kernels would be that density estimates tend to inherit the salient properties of their kernels in moderately-sized samples. If a density is suspected to be highly skewed, it may be sensible to contemplate asymmetric kernels. Some benefits will ensue, for example, when estimating the mode and tail quantiles of such a density. It is important to be able to locate more precisely the mode of possibly asymmetric densities when one wishes to maximise a smoothed profile likelihood with respect to a parameter of interest; e.g. see Silverman et al. (1990). Also, smoothed quantiles are important in the literature on robust statistical methods such as the regression quantiles of Koenker and Bassett (1978) or Koenker et al. (1994), and they can also be used to estimate probabilities for extreme-value events for problems such as in Tawn (1992).

The list of potential applications of and/or motivations for asymmetric kernels is large, and we have highlighted only a few. For example, spectral analysis in time series is another huge area of activity where asymmetric kernels are the rule rather than the exception, though the prime concern there is consistency rather than smoothing. The piecewise cubic kernel of Parzen is one example, based on the mean of an underlying uniform distribution. See Anderson (1971) (ch. 9) or Granger and Newbold (1986) (ch. 2) for an introduction, and Taniguchi and Kakizawa (2000) (ch. 6) for more recent applications.

Other successful applications of asymmetric kernels include Samiuddin and El-Sayyad (1990), Holiday (1995), Chen (1999, 2000), Scaillet (2001) and Fernandes and Scaillet (2002); where such kernels are shown to lead to desirable properties. Since asymmetric kernels are already in use, and there are good reasons for continuing to employ them, the question we address is the following: if one were to

[1] This is not to say that errors-in-variables problems should necessarily be dealt with by means of kernels. Such deconvolution problems have been treated extensively elsewhere, and are only used here for interpretation/illustration purposes.

use asymmetric kernels, which class would be optimal and what would its features be? In this paper, we derive an explicit answer to this question.

A potentially worrying aspect of considering asymmetric kernels is the paper by Cline (1988). He shows that asymmetric kernels are asymptotically ($n \to \infty$ in his proof on p. 1424) inadmissible, though not differing by an order of magnitude from their symmetrised counterpart (op. cit., p. 1425). Our context differs in that there is available qualitative information in the form of an a priori skewness in one direction. It is, therefore, more efficient to incorporate this qualitative feature into the estimation procedure.

It is also possible to use transformations in order to reduce the problems caused by the asymmetry of the density; e.g. see Wand et al. (1991) and the subsequent literature. See also Breiman and Friedman (1985), Tibshirani (1988) and Nychka and Ruppert (1995) for the different framework of regression models. Alternatively, a class of hypergeometric transformations of flexible form (including as special cases Box-Cox, Hermite polynomials and many others) can be applied as suggested in Abadir (1999, 2002). Our approach avoids the difficulty of identifying optimal transformations in samples, which are not large, but would be less effective if one had a priori reasons to believe that a class of transformations should symmetrise the variate of interest.

The next section derives a class of optimal asymmetric kernels, which turns out to be cubic with zero mean, and we solve explicitly for the parameters of this cubic. Its main numerical features are analysed in the subsequent section. Further numerical results are in Lawford (2001) who shows that the optimal asymmetric kernel is better than Epanechnikov's at estimating modes and extreme-value quantiles of asymmetric densities. Some general comments about our approach conclude the paper.

2. Asymmetric kernels

Suppose that possibly asymmetric kernels are contemplated, what functional form should they optimally take? One may argue from the point of view of minimising the distance between $f(x)$ and $\hat{f}(x)$, say the integrated mean squared-error (IMSE) or some other measure of distance.[2] For a kernel with skewness c, we have the following result based on minimising IMSE and solving the calculus-of-variations problem.

Theorem 1. *The standardised (unit-variance) kernel with skewness c, which minimises the leading term in the IMSE expansion is*

$$K_c(t) \equiv \alpha(t-\lambda_1)(t-\lambda_2)(t-\lambda_3)1_{t\in(\lambda_1,\lambda_2)}, \tag{2}$$

where $\alpha = 12(p-1)(p^2-3p+1)^2/[5(p+1)^5]$,

$$\lambda_1 = -\sqrt{\frac{5}{-p^2+3p-1}}, \quad \lambda_2 = p\sqrt{\frac{5}{-p^2+3p-1}}, \quad \lambda_3 = -\frac{3-4p+3p^2}{(p-1)\sqrt{5(-p^2+3p-1)}},$$

[2] Some information-theoretic measures necessitate further care in the treatment of supports for $\hat{f}(\cdot)$ (hence $K(\cdot)$) and/or $f(\cdot)$ if their logarithms are required. Others need not be formulated in terms of logarithms; e.g. $(\int f(x)^{1+\delta}\mathrm{d}x - 1)/\delta$ for some $\delta \neq 0$, the case $\delta = 0$ giving the usual Shannon-entropy $\int f(x)\ln(f(x))\mathrm{d}x$.

and $p \equiv -\lambda_2/\lambda_1 \in (1, \ 3/2]$ *solves* $\sqrt{5}\,(8-27p+27p^2-8p^3)/[7(-p^2+3p-1)^{3/2}]=c$ *uniquely for* $c \in (0, \ 2/7]$.

Proof. In order to reduce the bias in the IMSE, ceteris paribus, we should choose a second-order kernel, i.e. one with zero mean. Then, the leading (small-h, large-n) term in the IMSE expansion is

$$\frac{h^4}{4}\left(\int t^2 K(t)\mathrm{d}t\right)^2 \int\left(\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\right)^2 \mathrm{d}x + \frac{1}{nh}\int K(t)^2 \mathrm{d}t, \tag{3}$$

which achieves its minimum with respect to h at

$$\frac{5}{4n^{4/5}}\left(\left(\int K(t)^2 \mathrm{d}t\right)^2 \Big/ \int t^2 K(t)\mathrm{d}t\right)^{2/5}\left(\int\left(\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\right)^2 \mathrm{d}x\right)^{1/5};$$

e.g. see Silverman (1986) (pp. 39–41) for details. When the latter integral is finite, the problem of finding the optimal second-order $K(t)$ is equivalent to minimising $\int K(t)^2 \mathrm{d}t$ subject to $K(t)$ being nonnegative, $\int K(t)\mathrm{d}t = 1 = \int t^2 K(t)\mathrm{d}t$ and $\int t^3 K(t)\mathrm{d}t = c$. Modifying the derivations in Hodges and Lehmann (1956) for our extra constraint (the last one), we get the kernel in Eq. (2) where the parameters

$$\lambda_1 < 0 < \lambda_2 \leq \lambda_3 \quad \text{and} \quad 0 < \alpha \tag{4}$$

are all real, and to be determined. Then, the conditions

$$\int_{\lambda_1}^{\lambda_2} K_c(t)\mathrm{d}t = 1, \quad \int_{\lambda_1}^{\lambda_2} tK_c(t)\mathrm{d}t = 0, \quad \int_{\lambda_1}^{\lambda_2} t^2 K_c(t)\mathrm{d}t = 1, \quad \int_{\lambda_1}^{\lambda_2} t^3 K_c(t)\mathrm{d}t = c \tag{5}$$

determine the parameters. Solving the first constraint by definite integration of $K_c(t)$ from Eq. (2), $\lambda_3 = 6/[\alpha(\lambda_2-\lambda_1)^3] + (\lambda_1+\lambda_2)/2$, and upon substitution into the second constraint,

$$\alpha = 60(\lambda_1+\lambda_2)/(\lambda_2-\lambda_1)^5. \tag{6}$$

The formula for α can be used to simplify λ_3 to

$$\lambda_3 = (3\lambda_1^2 + 4\lambda_2\lambda_1 + 3\lambda_2^2)/[5(\lambda_1+\lambda_2)]. \tag{7}$$

Substituting Eq. (6) and Eq. (7) into the remaining two constraints of Eq. (5) gives

$$1 = -(\lambda_2^2 + 3\lambda_1\lambda_2 + \lambda_1^2)/5 \quad \text{and} \quad c = -(\lambda_1+\lambda_2)(8\lambda_1^2 + 19\lambda_1\lambda_2 + 8\lambda_2^2)/35. \tag{8}$$

The right-hand sides of all the equations in Eqs. (6)–(8) are homogeneous functions, and it facilitates the solution if we let

$$\lambda_2 \equiv -p\lambda_1, \tag{9}$$

where $p \in \mathbb{R}_+$ to satisfy Eq. (4). The first equation of Eq. (8) determines $\lambda_1 = -\sqrt{5/(-p^2+3p-1)}$, which can be used in Eq. (6), Eq. (7), Eq. (9), and the second equation of (8) to give

$$\alpha = \frac{12(p-1)(p^2-3p+1)^2}{5(p+1)^5}, \quad \lambda_3 = \frac{3-4p+3p^2}{(p-1)\sqrt{5(-p^2+3p-1)}},$$
$$\lambda_2 = p\sqrt{\frac{5}{-p^2+3p-1}}, \quad c = \frac{\sqrt{5}(8-27p+27p^2-8p^3)}{7(-p^2+3p-1)^{3/2}}, \tag{10}$$

respectively. Notice that the polynomial in the numerator of c has the three roots $p=1$, $(19/16) \pm ((1/16)\sqrt{105})$, which may seem to indicate that c can take any value in $\mathbb{R}$ for $p \in ((3/2)-((1/2)\sqrt{5}), (3/2)+((1/2)\sqrt{5}))$, and that multiple solutions for p may exist for any given c; see Fig. 1. This is not the case since p is further constrained to $p \in (1, 3/2]$ and bounded by the two vertical lines on the graph, as we now show.. The first of these further restrictions on p arises from the expression for α in Eq. (10). Since Eq. (4) requires $\alpha>0$, we must have $p>1$. The second restriction arises from the requirement in Eq. (4) that $\lambda_3 \geq \lambda_2$, with both given in Eq. (10). Therefore, p cannot lie outside the interval $[-1, 3/2]$. Combining both restrictions on p, we have $p \in [1, 3/2]$, with maximum c (and λ_1) at $p=3/2$. □

Remark 1. It was assumed that $c>0$ in the statement of the theorem. For $c<0$, one needs the mirror image of the kernel given there.

Remark 2. If there are no a priori grounds for believing that the underlying density $f(x)$ is skewed, then the derivations in Hodges and Lehmann (1956) or Epanechnikov (1969) apply. However, if the density is known to be qualitatively asymmetric (see the introductory example on income distributions), then $c \neq 0$ should be considered. In practice, c could be selected according to a cross-validation approach, and is going to be positively related with the standardised empirical skewness of X.

Remark 3. In spite of being asymmetric, the optimal kernel has zero mean. This reduces the bias of the estimated density, ceteris paribus. A zero mean, coupled with $c>0$, implies that the median and mode of

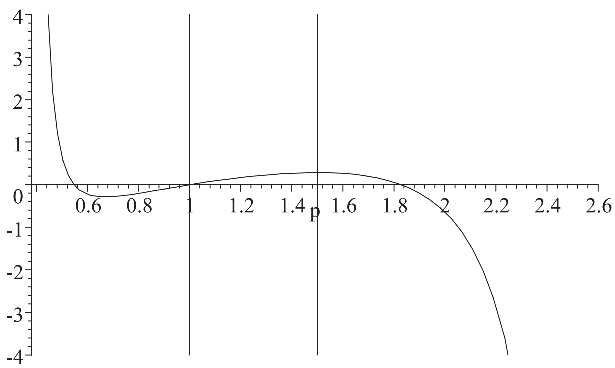

Fig. 1. Plot of c against p.

$K_c(t)$ occur at $t<0$. This may seem to violate the discussion in the income distribution example of Introduction, but this is not the case. An increasing marginal tax rate would guarantee higher tax receipts (in spite of the zero mean of $K_c(\cdot)$) if the fully-efficient tax collection system were implemented.

Remark 4. It follows from minimizing Eq. (3) that, for large samples, $h_c \simeq \text{constant}/n^{1/5}$.

3. Main intrinsic numerical features of the cubic $K_c(\cdot)$

We investigate the numerical features of the cubic $K_c(\cdot)$. We compare it mainly to its close relative, the Epanechnikov kernel $K_e(\cdot)$, that arises as the special case of $K_c(\cdot)$ with $c=0$. We, therefore, take the other extreme allowable value of c, namely $c=2/7$, giving rise to

$$K_c(t) = \frac{12}{625}(t+2)(t-3)^2 1_{t\in(-2,3)}.$$

We present the main results that are intrinsic to $K_c(\cdot)$. In Lawford (2001), Monte-Carlo experiments are given to show how $K_c(\cdot)$ performs better than the Epanechnikov kernel at estimating modes and extreme-value quantiles of asymmetric densities.

Recall Eq. (1) which gives a class of functions (kernels) integrating to 1, all from the same family as $K(\cdot)$ but with different scales. When comparing kernels' IMSE performances, we wish to uncouple the effect of the choices of $K(\cdot)$ and h. To this end, we need to find the $\kappa(\cdot)$ from each family K (= Epanechnikov, Cubic, etc.) such that its asymptotic IMSE separates into two multiplicative terms: one in $\kappa(\cdot)$ and another in h. This $\kappa(t;\, s_K)$ is called the (asymptotic) canonical kernel for the class K, with (asymptotic) canonical scale s_K. Marron and Nolan (1989) (p. 197) show that

$$s_K = \left(\int K(t)^2 \mathrm{d}t \bigg/ \left(\int t^2 K(t)\mathrm{d}t\right)^2\right)^{1/5}.$$

Using $\kappa(t;\, s_K) = s_K^{-1} K(t/s_K)$ in Eq. (3) instead of $K(t)$, then changing the variable $t/s_K \mapsto t$, we obtain the asymptotic IMSE (after correcting for a typo in (2.4) of Marron and Nolan, 1989 (p. 197)

$$\left[\left(\int K(t)^2 \mathrm{d}t\right)^2 \int t^2 K(t)\mathrm{d}t\right]^{2/5} \left(\frac{h^4}{4}\int\left(\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\right)^2 \mathrm{d}x + \frac{1}{nh}\right),$$

where the term in the (first) square bracket is now the only function of $K(t)$, and it is invariant to changes of scale, i.e. invariant to h if $K(t)$ were to be replaced by any $\kappa(t;\, h)$. The importance of such a separation is also highlighted by the derivations in Samiuddin and El-Sayyad (1990) (especially pp. 865–867).

In order to be able to translate a window width of $K_c(\cdot)$ into those of commonly-used kernels, we need to provide an exchange-rate table based on s_K/s_c. Such exchange rates are not transitive, and Table 1 lists them for a variety of commonly-referenced kernels, namely K = Uniform, Triangular, Epanechnikov, Quartic and Gaussian. Because of the asymmetric support of the cubic, we have normalised all kernels to have unit variances, instead of normalising them to be on the $(-1,\,1)$ support. It is seen that the exchange rates are all approximately equal to 1, once variance-standardisation is performed. This reveals

Table 1
Exchange rates for kernel K relative to the cubic

K	$K_K(t)$	s_K/s_c
Uniform	$(1/2\sqrt{3})1_{t\in(-\sqrt{3},\sqrt{3})}$	1.0103
Triangular	$(1/\sqrt{6})(1-(1/\sqrt{6})\|t\|)1_{t\in(-\sqrt{6},\sqrt{6})}$	0.9984
Epanechnikov	$(3/4\sqrt{5})(1-(1/5)t^2)1_{t\in(-\sqrt{5},\sqrt{5})}$	0.9956
Quartic	$(15/16\sqrt{7})(1-(1/7)t^2)^2 1_{t\in(-\sqrt{7},\sqrt{7})}$	0.9968
Gaussian	$(1/\sqrt{2\pi})\exp(-(1/2)t^2)$	1.0056

that (asymptotic) exchange rates for kernels are essentially equivalent to a transformation of kernels to have unit variances.

4. Concluding comments

We have derived the optimal functional form that asymmetric kernels should take. Even though we have focused on density estimation in our discussion of the results, kernels are also of use in other areas of nonparametric analysis, such as the ones mentioned in Introduction. It is also worth pointing out that a special case of the calculus-of-variations problem which we have solved in the proof of our theorem arises in the seemingly unrelated area of assessing the efficiency of nonparametric tests; e.g. see Hodges and Lehmann (1956) and Hallin and Tribel (2000). Our new kernel could have some applicability there too. Finally, Ruppert and Cline (1994) suggest an iterated empirical transformation, based on the smoothed empirical distribution, for the sake of asymptotic bias reduction. Our kernel could be used there too (except for the boundaries of the target uniform distribution) in order to achieve finite-sample improvements.

Acknowledgements

We would like to thank Peter Robinson and Oliver Linton for useful comments. The first author acknowledges support from the ESRC (UK) grant R000239538, and the second acknowledges the support of the Marie Curie Fellowships under the European Commission's Improving Human Potential programme. The authors are solely responsible for information communicated and the European Community is not responsible for any views or results expressed.

References

Abadir, K.M., 1999. An introduction to hypergeometric functions for economists. Econometric Reviews 18, 287–330.
Abadir, K.M., 2002. Comment on "Box-Cox transformations in linear models: large sample theory and tests of normality" by Chen, G., Lockhart, R.A., Stephens, M.A. Canadian Journal of Statistics 30 (210), 233–234.
Anderson, T.W., 1971. The Statistical Analysis of Time Series Wiley, New York.
Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association 80, 580–619 (with discussion).

Chen, S.X., 1999. Beta kernel estimators for density functions. Computational Statistics and Data Analysis 31, 131–145.

Chen, S.X., 2000. Probability density function estimation using gamma kernels. Annals of the Institute of Statistical Mathematics 52, 471–480.

Cline, D.B.H., 1988. Admissible kernel estimators of a multivariate density. Annals of Statistics 16, 1421–1427.

Epanechnikov, V.A., 1969. Non-parametric estimation of a multivariate probability density. Theory of Probability and its Applications 14, 153–158.

Fernandes, M., Scaillet, O., 2002. Testing for symmetry and conditional symmetry using asymmetric kernels, mimeo, HEC, Université de Genève.

Granger, C.W.J., Newbold, P., 1986. Forecasting Economic Time Series. Academic Press, San Diego.

Hallin, M., Tribel, O., 2000. The efficiency of some nonparametric competitors to correlogram methods. In: Bruss, F.T., Le Cam, L. (Eds.), Game Theory, Optimal Stopping, Probability, and Statistics, Papers in honor of T.S. Ferguson on the occasion of his 70th birthday (I.M.S. Lecture Notes-Monograph Series, Hayward, California), 249–262 http://www.imstat.org/publications/lecnotes.htm.

Härdle, W., Linton, O., 1994. Applied nonparametric methods. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. Elsevier, Amsterdam, pp. 2295–2339.

Hodges, J.L., Lehmann, E.L., 1956. The efficiency of some non-parametric competitors of the t-test. Annals of Mathematical Statistics 27, 324–335.

Holiday, D., 1995. Near optimal weights in nonparametric regression under some common restrictions. Statistics and Probability Letters 22, 33–42.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.

Lawford, S.D.C., 2001 Improved modelling in finite-sample and nonlinear frameworks, D.Phil. in Economics, University of York.

Marron, J.S., Nolan, D., 1989. Canonical kernels for density estimation. Statistics and Probability Letters 7, 195–199.

Nychka, D., Ruppert, D., 1995. Nonparametric transformations for both sides of a regression model. Journal of the Royal Statistical Society, Series B 57, 519–532.

Ruppert, D., Cline, D.B.H., 1994. Bias reduction in kernel density estimation by smoothed empirical transformations. Annals of Statistics 22, 185–210.

Samiuddin, M., El-Sayyad, G.M., 1990. On nonparametric kernel density estimates. Biometrika 77, 865–874.

Scaillet, O., 2001. Density estimation using inverse and reciprocal inverse Gaussian kernels, IRES Discussion Paper 2001/17, Université Catholique de Louvain.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Silverman, B.W., Jones, M.C., Wilson, J.D., Nychka, D.W., 1990. A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. Journal of the Royal Statistical Society, Series B 52, 271–324 (with discussion).

Taniguchi, M., Kakizawa, Y., 2000. Asymptotic Theory of Statistical Inference for Time Series. Springer, New York.

Tawn, J.A., 1992. Estimating probabilities of extreme sea-levels. Applied Statistics 41, 77–93.

Tibshirani, R., 1988. Estimating transformations for regression via additivity and variance stabilisation. Journal of the American Statistical Association 83, 394–405.

Wand, M.P., Marron, J.S., Ruppert, D., 1991. Transformations in density estimation. Journal of the American Statistical Association 86, 343–361 (with discussion).