CANONICAL HIGHER-ORDER KERNELS FOR DENSITY DERIVATIVE ESTIMATION

DANIEL J. HENDERSON AND CHRISTOPHER F. PARMETER*

Binghamton University and University of Miami

Abstract. In this note we present rth order kernel density derivative estimators using canonical higher-order kernels. These canonical rescalings uncouple the choice of kernel and scale factor. This approach is useful for selection of the order of the kernel in a data-driven procedure as well as for visual comparison of kernel estimates.

Journal of Economic Literature Subject Classification: C14.

1. Introduction

The issue of kernel selection on the performance of density estimation is widely believed to have little overall impact relative to that of the bandwidth. Yet, given that the bandwidth controls the smoothing of the density estimator, using different kernels (e.g., Epanechnikov vs. Gaussian) can produce different results. Here we generalize the canonical second-order kernel equivalence class developed in Marron & Nolan (1989) to higher-order kernels.[1] The development of canonical higher order and derivative kernels may be useful for data-driven selection of the bandwidth and the kernel order simultaneously (Hall & Marron 1988).

The remainder of the paper is laid out as follows. Section 2 discusses the construction of the higher-order equivalence class and provides the canonical kernel scalings. Section 3 gives a simple set of illustrations to emphasize our results.

E-mail address: djhender@binghamton.edu,cparmeter@miami.edu.

Date: July 13, 2010.

Key words and phrases. Derivative Estimation, AMISE.

* Daniel J. Henderson (djhender@binghamton.edu) Department of Economics, State University of New York at Binghamton and Christopher F. Parmeter (Corresponding Author, cparmeter@miami.edu) Department of Economics, University of Miami.

[1] See Abadir & Lawford (2004) for a similar exercise using asymmetric kernels.

2. Estimators

The ν^{th}-order, r^{th} derivative kernel density estimator for $f^{(r)}(\cdot)$, based on an *iid* random sample $\{X_1, \ldots, X_n\}$ from $f(\cdot)$, the probability density, is defined as

$$\hat{f}^{(r)}(x) = n^{-1} h^{-(1+r)} \sum_{i=1}^{n} k_{\nu,s}^{(r)}\left(\frac{x - X_i}{h}\right), \tag{1}$$

where h is the bandwidth and $k_{\nu,s}^{(r)}$ is the r^{th} derivative of the ν^{th}-order s-kernel (see (2) below). The most common setup is $\nu = 2$ and $r = 0$ which constitutes a second-order kernel used to estimate the density itself. The use of a higher-order kernel is typically undertaken to reduce the bias of the density estimator.

There are also many cases where interest may lie in the r^{th} derivative. For example, one may be interested in looking for the location of modes ($r = 1$). Indirect interest in derivatives estimates also exists. For example, estimation of the roughness of the second derivative ($r = 2$) of the density, $R(f^{(2)}(\cdot)) = \int f^{(2)}(x)^2 dx$, is required for plug-in bandwidth selection.

Marron & Nolan (1989) discuss an equivalence class of kernels for the case $\nu = 2$ and $r = 0$. Here we generalize their results for the arbitrary ν, r^{th} case. We discuss estimation for the class of kernels defined as

$$k_s(u) = \frac{(2s+1)!!}{2^{s+1} s!}(1 - u^2)^s \mathbf{1}\{|u| \leq 1\}, \tag{2}$$

where the double factorial is defined as $(2s+1)!! = (2s+1) \cdot (2s-1) \cdots 5 \cdot 3 \cdot 1$ (commonly known as the odd factorial). As $s \to \infty$, $k_s(u) \to e^{-u^2/2}$. Rescaling this particular case by $1/\sqrt{2\pi}$ delivers the common Gaussian kernel, which we denote as $k_\phi(u)$. The Epanechnikov ($s = 1$), biweight ($s = 2$) and triweight ($s = 3$) are also popular kernels from this class. Notice that as s increases, the kernel possesses more derivatives and thus is 'smoother'.

For the class of polynomial kernels of order s, a ν^{th}-order s-kernel can be constructed as (Hansen 2005, Theorem 1)

$$k_{\nu,s}(u) = B_{\nu/2,s}(u) k_s(u),$$

where

$$B_{\nu/2,s}(u) = \frac{\left(\frac{3}{2}\right)_{\nu/2-1}\left(\frac{3}{2}+s\right)_{\nu/2-1}}{(s+1)_{\nu/2-1}} \sum_{j=0}^{\nu/2-1} \frac{(-1)^j\left(\frac{1}{2}+s+\nu/2\right)_j x^{2j}}{j!(\nu/2-1-j)!\left(\frac{3}{2}\right)_j}.$$

The notation $(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}$ is Pochhamer's symbol, where $\Gamma(a) = \int\limits_0^\infty t^{a-1}e^{-t}dt$. See Wand & Schucany (1990, Theorem 2.1) for a similar expression for higher order Gaussian kernels.

We first impose several generic properties on our kernel function involving the 'moments' of the kernel. Letting $\kappa_j(k) = \int\limits_{-\infty}^{\infty} u^j k(u)du$, we say a kernel is of ν^{th}-order if $\kappa_0(k) = 1$, $\kappa_j(k) = 0$ for $1 \leq j \leq \nu - 1$ and $\kappa_\nu(k) < \infty$. With symmetric kernels we have that $\kappa_\ell(k) = 0$ for $\ell = 2j+1$, i.e. all odd moments of our kernel are zero so that we may treat only the cases where ν is even. Requiring $\kappa_\nu(\cdot)$ to be finite is necessary to obtain meaningful expressions for the bias of our kernel density estimator.

As pointed out by Marron & Nolan (1989, pg. 197), a representative kernel has a best element.[2] This kernel is known as the canonical kernel and it is such that it has exactly the same effect on the squared bias and variance components which make up the asymptotic mean integrated squared error ($AMISE$). Here we extend the results of Marron & Nolan (1989) to both higher-order and derivative kernels. To establish our basis, we quantify the requisite amount of smoothing via $AMISE$ which one can easily show for the density derivative estimator in (1) to be

$$AMISE(\hat{f}^{(r)}(x)) \approx \frac{R\left(f^{(r+\nu)}(\cdot)\right)h^{2\nu}\kappa_\nu^2(k_{\nu,s})}{(\nu!)^2} + \frac{R\left(k_{\nu,s}^{(r)}\right)}{nh^{1+2r}}, \tag{3}$$

provided $h \to 0$ and $nh^{1+2r} \to \infty$ and assuming that $f(\cdot)$ is $r+\nu$ times continuously differentiable.

We immediately notice that (3) quantifies the smoothing trade-off since decreasing h results in the first term (the squared bias) collapsing towards zero with the second term (the variance) increasing. This produces an estimated curve with too much variation. Alternatively, allowing h to increase clearly raises the first term while the second term shrinks, producing a curve that is uninformative as it possesses almost no local variation since it averages over too large of a neighborhood surrounding x. Moreover, regardless of the order of the derivative of interest, the

[2] For a given kernel one may scale it by any positive constant, thus, each scaled kernel is an element from the class of kernels defined by s and ν in our language.

bias always depends on the kernel, not its derivative. Alternatively, the variance depends directly on the r^{th} derivative of the kernel being used in estimation.

2.1. **Optimal scaling.** We now discuss a rescaled version of the kernel which decouples the impact that the choice of kernel has on each component in (3). Following the setup of Marron & Nolan (1989), we seek to rescale $k_{\nu,s}^{(r)}$ so that k and h are separate in (3). This can be done by noting that the kernel impacts the bias through $\kappa_{\nu,s}^2\left(k_{\nu,s}\right)$, while it impacts the variance through $R\left(k_{\nu,s}^{(r)}\right)$. By selecting the scale of the kernel so that these are equivalent, we see that we can equalize their individual contributions to $AMISE$.[3]

Rescaling our kernel by δ,

$$k_{\nu,s,\delta}^{(r)}(\cdot) = k_{\nu,s}^{(r)}(\cdot/\delta)/\delta^{r+1},$$

our optimal scale (δ) is found by solving

$$\left[\int x^{\nu} k_{\nu,s,\delta}(x)dx\right]^2 = \int \left(k_{\nu,s,\delta}^{(r)}(x)\right)^2 dx.$$

Integration by substitution yields

$$\delta_0 = \left[R\left(k_{\nu,s}^{(r)}\right)\right]^{1/(1+2r+2\nu)} \left[\kappa_{\nu}\left(k_{\nu,s}\right)\right]^{-2/(1+2r+2\nu)}. \tag{4}$$

When $r = 0$ and $\nu = 2$, (4) gives the optimal scaling found in Marron & Nolan (1989).

With the choice $\delta = \delta_0$, using the kernel $k_{\nu,s,\delta_0}^{(r)}(\cdot)$, delivers

$$AMISE\left(\hat{f}^{(r)}(x)\right) \approx C\left(k_{\nu,s,\delta_0}^{(r)}(\cdot)\right)\left[R\left(f^{(r+\nu)}(\cdot)\right)h^{2\nu}(\nu!)^{-2} + (nh^{1+2r})^{-1}\right], \tag{5}$$

where

$$C\left(k_{\nu,s,\delta_0}^{(r)}(\cdot)\right) = \left[R\left(k_{\nu,s}^{(r)}\right)\right]^{2\nu/(1+2r+2\nu)} \left[\kappa_{\nu}\left(k_{\nu,s}\right)\right]^{(2+4r)/(1+2r+2\nu)}.$$

These scalings are produced for various values of s, ν, and r in Table 1. Here we first note the obvious; the table shows fewer kernel functions (s) when the order of the derivative (r) increases as some of our kernel functions only possess derivatives up to a given order. The relative difference between estimates for a given kernel function can be seen by the relative values for δ. For example, when $r = 0$, the effective smoothness of the Epanechnikov and Gaussian kernels differ by a factor in

[3] An interesting extension would be to develop an equivalence class for the bias reducing kernels proposed in Mynbaev & Martins-Filho (2010).

excess of 2 for higher-order kernels and this difference increases with ν. Note that this is when the same bandwidth is used for smoothing. In contrast, we can surmise that comparisons between the biweight and triweight kernels based on the same bandwidth would produce more closely related density estimates for any order kernel.

3. Illustration

For both illustrative and comparison purposes, we consider the same data generating process as Marron & Nolan (1989). Specifically, we simulate 500 data points from a mixture density: $(0.7)\text{Beta}(4, 8) + (0.3)\text{Beta}(40, 20)$. We plot estimates using fourth-order kernels ($\nu = 4$).

Figure 1, panel (a) uses Gaussian ($s = \phi$) and triweight ($s = 3$) kernels to estimate the underlying density ($r = 0$), while panel (b) uses their respective canonical fourth-order kernels. Both panels also provide vertically rescaled kernels centered at 0.5 to provide an equivalence (dotted lines). Notice that in panel (a) the kernels do not appear to look similar and as such, even with the same amount of smoothing, $h = 0.11$, the resulting curves are different (due to the difference in the kernels). In panel (b), we see that differences in the density estimates are harder to detect visually given the use of canonical kernels. This is further buttressed by the vertically rescaled canonical kernels at 0.5. The differences in these canonical kernels are difficult to detect relative to panel (a).

In Figure 2 panel (a), we look at the estimates of the second-order derivatives of our density ($r = 2$) using standard and canonical fourth-order kernels ($\nu = 4$), respectively. To make the panels visually appealing we plot the vertically rescaled kernels using -400 as the $x-$axis (again dotted lines represent the kernels). We use the bandwidth $h = 0.11$ to construct our density derivative estimates. We note that the 4th order, 2nd derivative density estimate using the triwieght kernel is highly variable, suggesting that this bandwidth is too small (increased variance) whereas this same bandwidth produces a viable estimate using the Gaussian kernel. However, the peaks and the troughs are all underestimated suggesting perhaps that this bandwidth is too large (increased bias). The differences in these estimated curves is easily gleaned by noting the differences in the corresponding kernels at the bottom of each plot. Panel (b) provides the canonical forms of these density estimates, where as expected, the differences in the estimated curves are difficult to detect visually.

In sum, the use of canonical higher order derivative kernels may prove useful in a data driven procedure which simultaneously selects the kernel order and the bandwidth since the use of canonical kernels decouples the problem of kernel and bandwidth selection. This is an area of research that has received little attention in the applied nonparametric literature.

References

Abadir, K. M. & Lawford, S. (2004), 'Optimal asymmetric kernels', *Economics Letters* **83**, 61–68.

Hall, P. & Marron, J. S. (1988), 'Choice of kernel order in density estimation', *Annals of Statistics* **16**(1), 161–173.

Hansen, B. E. (2005), 'Exact mean integrated squared error of higher order kernel estimators', *Econometric Theory* **21**, 1031–1057.

Marron, J. S. & Nolan, D. (1989), 'Canonical kernels for density estimation', *Statistics & Probability Letters* **7**(3), 195–199.

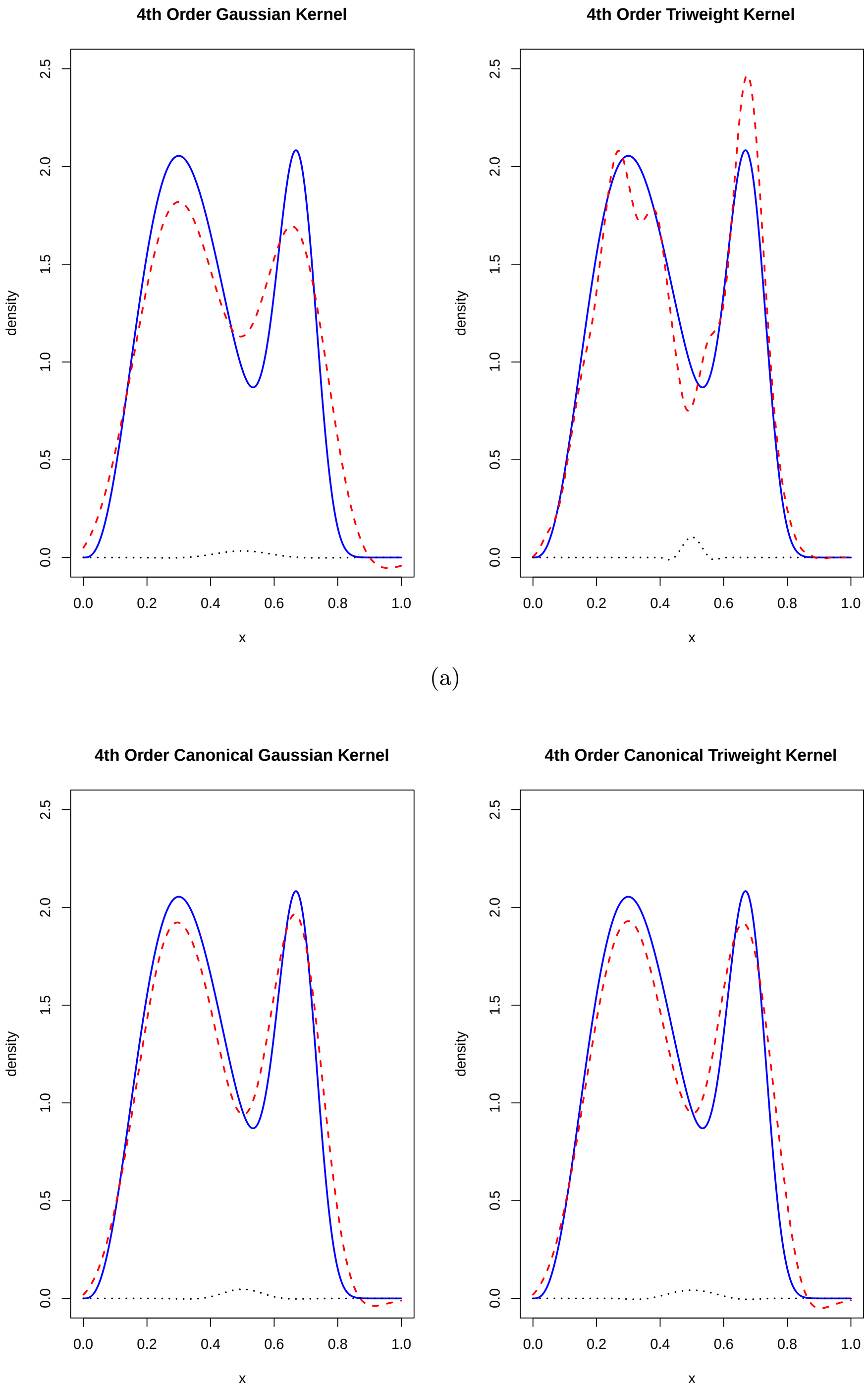
Mynbaev, K. & Martins-Filho, C. (2010), 'Bias reduction in kernel density estimation via Lipschitz condition', *Journal of Nonparametric Statistics* **22**, 219–235.

Wand, M. P. & Schucany, W. R. (1990), 'Gaussian-based kernels', *Canadian Journal of Statistics* **18**, 197–204.

Table 1. Canonical kernel scalings (δ): s refers to the kernel type. $s = 1, 2, 3, \phi$ representing Epanechnikov, biweight, triweight and Gaussian kernels, respectively. $v = 2, 4, 6, 8, 10$ refers to the order of the kernel funciton. r refers to the order of the derivative of the function. $r = 0, 1, 2$ represents the density itself, and its first and second order derivatives, respectively. The remaining values in the table represent δ, the canonical kernel scalings.

$s \setminus \nu$	2	4	6	8	10
			$r = 0$		
1	1.7188	2.0165	2.0834	2.1021	2.1062
2	2.0362	2.2591	2.2694	2.2513	2.2302
3	2.3122	2.4788	2.4416	2.3913	2.3478
ϕ	0.7764	0.7214	0.6358	0.5686	0.5169
			$r = 1$		
2	1.9442	2.3658	2.4539	2.4607	2.4443
3	2.2103	2.5973	2.6410	2.6143	2.5736
ϕ	0.7559	0.7668	0.6959	0.6280	0.5717
			$r = 2$		
3	2.4189	2.8627	2.9335	2.9083	2.8585
ϕ	0.8415	0.8566	0.7818	0.7054	0.6405

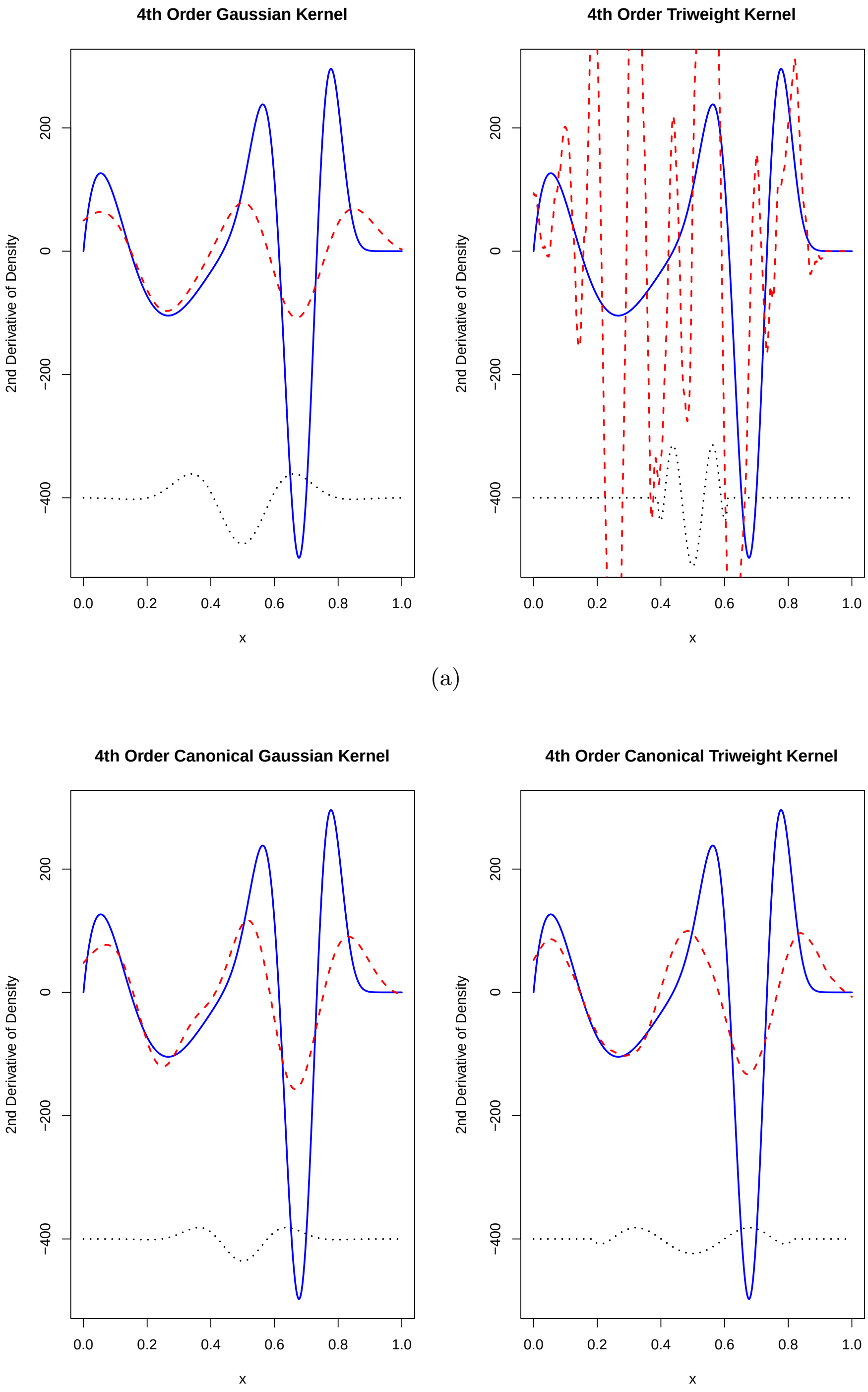
FIGURE 1. The solid curve is the true underlying density function while the dashed curve is the corresponding kernel density estimate. Both density estimates use 4th order (canonical) kernels with bandwidth, $h = 0.11$. Vertically rescaled (canonical) kernels appear as a dotted line in each panel.



(a)

(b)

Figure 2. The solid curve is the true underlying 2nd derivative of the density function while the dashed curve is the corresponding 2nd derivative kernel density estimate. Both derivative density estimates use 4th order (canonical) kernels with bandwidth, $h = 0.11$. Vertically rescaled (canonical) kernels appear as a dotted line in each panel.



(a)

(b)