International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:5, No:7, 2011

Orthogonal Polynomial Density Estimates: Alternative Representation and Degree Selection

Serge B. Provost and Min Jiang

Abstract—The density estimates considered in this paper comprise a base density and an adjustment component consisting of a linear combination of orthogonal polynomials. It is shown that, in the context of density approximation, the coefficients of the linear combination can be determined either from a moment-matching technique or a weighted least-squares approach. A kernel representation of the corresponding density estimates is obtained. Additionally, two refinements of the Kronmal-Tarter stopping criterion are proposed for determining the degree of the polynomial adjustment. By way of illustration, the density estimation methodology advocated herein is applied to two data sets.

Keywords—kernel density estimation, orthogonal polynomials, moment-based methodologies, density approximation.

I. Introduction

A wide array of parametric, nonparametric and hybrid techniques are available for estimating a density function on the basis of a sample of observations. An informative account of the main nonparametric density estimation methodologies available can be found for instance in [1]. The density estimates being considered in this paper can be expressed as the product of an initial estimate referred to as *base density* and an *adjustment component* consisting a linear combination of orthogonal polynomials.

Various aspects of several density estimation and approximation techniques that involve orthogonal series have been studied in numerous papers, including [2]–[8]. The concept of making use of a base density and adjusting it has been previously discussed by [9]–[12], among others.

Although attention is focused on density *estimation*, some preliminary results on a density *approximation* methodology that relies on orthogonal polynomials are required. Accordingly, orthogonal polynomial density approximants are defined in Section II. It is also explained therein that a sequence of orthogonal polynomials can be generated from a given weight function and that the coefficients of the linear combination of the orthogonal polynomials constituting the adjustment component of the estimates can be determined either from a moment-matching technique or by minimizing the integrated squared error with respect to a certain weighting function.

As pointed out in Section III, orthogonal polynomial density estimates can be viewed as counterparts of the density approximants discussed in [6]; such density estimates are shown to possess a kernel representation in addition to their prime representation which is given in terms of *sample* moments. Two refinements of the Kronmal-Tarter criterion whereby the number of terms to be included in the polynomial adjustment component of the estimates can be determined, are also proposed. The step-by-step description of the orthogonal polynomial density estimation methodology contained in Section III-C should render it more easily implementable and more widely accessible. Two illustrative examples are presented in Section IV.

Explicit representations of the kernels associated with the Legendre, Jacobi, Laguerre and Hermite orthogonal polynomials are provided in Appendices $A-D$. In view of the fact that their associated base densities are respectively the uniform, beta, gamma and Gaussian density functions (up to certain affine transformations), such orthogonal polynomials are likely to be frequently utilized in conjunction with the density estimation approach advocated in this paper.

Serge Provost is Professor of Statistics in the Department of Statistical & Actuarial Sciences at The University of Western Ontario, London, Canada, N6A 5B7. Min Jiang is with Statistics Canada, Ottawa, Ontario, K1A 0T6. Corresponding author's e-mail address: provost@stats.uwo.ca.

II. Orthogonal Polynomials and Density Approximants

A. Introduction

Let

$$\varphi_k(x) = \sum_{\ell=0}^{k} \delta_{k,\ell}\ x^{\ell},\ k = 0, \ldots, m\,, \tag{1}$$

be polynomials defined on the interval (a, b), which satisfy the orthogonality property,

$$\int_a^b w(x)\,\varphi_i(x)\,\varphi_j(x)\,\mathrm{d}x = \begin{cases} \theta_i & \text{for } \ i = j \\ 0 & \text{for } \ i \neq j, \end{cases} \tag{2}$$

where $w(x)$ denotes a certain nonnegative weight function whose 'moments' given by $\int_a^b x^k\ w(x)\ \mathrm{d}x$, exist for $k = 0, 1, \ldots$, and θ_i will be referred to as the i^{th} degree orthogonality factor. Then, $\{\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)\}$ is said to form a set of orthogonal polynomials with respect to $w(x)$.

As was explained in [6], in most instances, it is possible to approximate a continuous probability density function, $f(x)$, defined on the interval (a, b), by means of an approximant of the form

$$f_m(x) = c\ w(x) \sum_{j=0}^{m} a_j\,\varphi_j(x) \tag{3}$$

where the normalizing constant c is such that $\int_a^b c\ w(x)\ \mathrm{d}x = 1$. The weight function $w(x)$ is chosen so that $c\,w(x)$ provides a suitable initial density approximation to $f(x)$. The function $f_m(x)$ will be referred to as an m^{th} degree orthogonal polynomial density approximant. It was demonstrated in [7] that

such approximants, which are based on the exact moments of a continuous distribution, can yield very accurate percentiles.

It is shown in Section II-B that, given a weight function $w(x)$ whose moments exist, one can generate a specific sequence of orthogonal polynomials. If $w(x)$ depends on p parameters, such parameters can be determined for instance by solving the equations, $c\int_a^b x^h\, w(x)\mathrm{d}x = \mu_X(h), \quad h = 1, \ldots, p$, where $\mu_X(h)$ denotes the h^{th} moment of the random variable whose density function, $f(x)$, is being approximated. It is assumed in the sequel that the distributions under consideration are uniquely determined from their moments. We note that this is always the case for random variables whose support is compact; conditions ensuring uniqueness in the case of infinite or semi-infinite ranges are specified for instance in [13]. As shown in Sections II-C and II-D, the coefficients a_j in (3) can be equivalently obtained from two distinct approaches.

B. Determination of the Orthogonal Polynomials

In certain instances, such as in the case of the classical orthogonal polynomials discussed in the Appendices , the coefficients $\delta_{k,\ell}$ appearing in (1), as well as the associated weight functions, are known. In general, one can generate a sequence of orthogonal polynomials from any proper weight function by making use of the Gram-Schmidt orthogonalization process. This procedure constructs an orthogonal basis over an interval (a, b) with respect to an arbitrary weight function $w(x)$ from a nonorthogonal set of linearly independent functions.

Using the notation

$$< p_i(x), p_j(x) > \; = \int_a^b w(x)\, p_i(x)\, p_j(x)\, \mathrm{d}x\,, \tag{4}$$

where $w(x)$ is a weight function, and defining the first two polynomials as

$$\varphi_0(x) = 1 \;\equiv\; \delta_{0,0} \tag{5}$$

and

$$\varphi_1(x) = x - \frac{< x, 1 >}{< 1, 1 >} \;\equiv\; \delta_{1,0} + \delta_{1,1}\, x\,, \tag{6}$$

one can construct all higher order orthogonal polynomials from the recurrence relation,

$$\begin{aligned}\varphi_{i+1}(x) &= \left(x - \frac{< x\,\varphi_i(x), \varphi_i(x) >}{< \varphi_i(x), \varphi_i(x) >}\right)\varphi_i(x) \\ &\qquad - \left(\frac{< \varphi_i(x), \varphi_i(x) >}{< \varphi_{i-1}(x), \varphi_{i-1}(x) >}\right)\varphi_{i-1}(x)\,, \\ &\equiv \sum_{\ell=0}^{i+1} \delta_{i+1,\ell}\; x^\ell\;, \quad i = 1, 2, \ldots,\end{aligned} \tag{7}$$

see *e.g.* [14].

C. Moment-Based Density Approximants

It is shown in this section that the coefficients $a_j,\ j = 0, 1 \ldots, m$, appearing in the approximant $f_m(x)$ defined by (3) can be determined by matching the first m moments of $f_m(x)$ to those of $f(x)$, the density function being approximated. First, one can easily establish that the equalities

$$\int_a^b x^j f_m(x)\,\mathrm{d}x = \int_a^b x^j f(x)\,\mathrm{d}x, \quad j = 0, 1, \ldots, m, \tag{8}$$

are mathematically equivalent to

$$\int_a^b \varphi_j(x) f_m(x)\,\mathrm{d}x = \int_a^b \varphi_j(x) f(x)\,\mathrm{d}x, \quad j = 0, 1, \ldots, m. \tag{9}$$

Accordingly, if (9) holds, which amounts to assuming that the first m moments of the approximate distribution are equal to those associated with the density function being approximated, one has

$$\int_a^b c\; w(x) \sum_{i=0}^{m} a_i\, \varphi_i(x)\, \varphi_j(x)\,\mathrm{d}x = \int_a^b \varphi_j(x) f(x)\,\mathrm{d}x,$$

that is,

$$\sum_{i=0}^{m} c\; a_i \int_a^b w(x)\, \varphi_i(x)\, \varphi_j(x)\,\mathrm{d}x = \int_a^b \varphi_j(x) f(x)\,\mathrm{d}x \tag{10}$$

or

$$c\; a_j\; \theta_j = \int_a^b \varphi_j(x) f(x)\,\mathrm{d}x\,,$$

so that

$$a_j = \frac{\int_a^b \varphi_j(x) f(x)\,\mathrm{d}x}{c\; \theta_j}\,, \tag{11}$$

where

$$\begin{aligned}\int_a^b \varphi_j(x) f(x)\,\mathrm{d}x &= \int_a^b \sum_{\ell=0}^{j} \delta_{j,\ell}\; x^\ell f(x)\,\mathrm{d}x \\ &= \sum_{\ell=0}^{j} \delta_{j,\ell}\; \mu_X(\ell)\,,\end{aligned} \tag{12}$$

$\mu_X(\ell)$ denoting the ℓ^{th} moment of the distribution specified by $f(x)$. Thus,

$$a_j = \sum_{\ell=0}^{j} \frac{\delta_{j,\ell}\; \mu_X(\ell)}{c\; \theta_j}\,, \quad j = 0, 1, \ldots, m\,, \tag{13}$$

and the m^{th} degree density approximant can be expressed as follows:

$$f_m(x) = w(x) \sum_{j=0}^{m} \sum_{\ell=0}^{j} \frac{\delta_{j,\ell}\; \mu_X(\ell)}{\theta_j}\; \varphi_j(x)\,, \tag{14}$$

where θ_j can be determined from (2) and $\delta_{j,\ell}$ denotes the coefficient of x^ℓ in $\varphi_j(x)$.

D. Weighted Least-Squares Density Approximants

Alternatively, the coefficients a_j appearing in (3) can be obtained from a weighted least-squares approach. On denoting the integrated weighted squared error by $W(a_0, \ldots, a_m)$ while making use of the reciprocal of the weight function as the weighting function as was done for instance in [15], one has

$$\begin{aligned}&W(a_0, \ldots, a_m) \\ &\quad = \int_a^b \frac{1}{w(x)}\big(f(x) - f_m(x)\big)^2\,\mathrm{d}x \\ &\quad = \int_a^b \frac{f^2(x)}{w(x)}\,\mathrm{d}x - 2\int_a^b c\; \sum_{i=0}^{m} a_i\, \varphi_i(x)\, f(x)\,\mathrm{d}x \\ &\qquad\qquad + \int_a^b w(x)\Big(c\; \sum_{i=0}^{m} a_i \varphi_i(x)\Big)^2\,\mathrm{d}x\,.\end{aligned} \tag{15}$$

Then, equating

$$\begin{aligned}
&\tfrac{\partial W(a_0,\ldots,a_m)}{\partial a_j} \\
&= -2\int_a^b c\;\varphi_j(x)\,f(x)\,\mathrm{d}x \\
&\quad +2\int_a^b w(x)\Big(c^2\;\sum_{i=0}^{m} a_i\,\varphi_i(x)\Big)\,\varphi_j(x)\,\mathrm{d}x \\
&= -2\;c\;\Big(\int_a^b \varphi_j(x)f(x)\;\mathrm{d}x - c\;a_j\theta_j\Big),\;\; j=0,1,\ldots,m\,,
\end{aligned}$$

to zero, it is seen that the coefficients a_j, $j=0,1,\ldots,m$, that minimize $W(a_0,\ldots,a_m)$ are given by

$$a_j = \frac{\int_a^b \varphi_j(x)f(x)\,\mathrm{d}x}{c\;\theta_j} \tag{16}$$

or, in light of (12),

$$a_j = \sum_{\ell=0}^{j} \frac{\delta_{j,\ell}\;\mu_X(\ell)}{c\;\theta_j}\,, \tag{17}$$

which coincides with the representation of a_j given in (13). Interestingly, Equation (15) can be rewritten as

$$W(a_0,\ldots,a_m) = \int_a^b w(x)\Big(g(x)-p_m(x)\Big)^2\,\mathrm{d}x, \tag{18}$$

where $g(x)=f(x)/w(x)$ and $p_m(x)=f_m(x)/w(x)$. Thus, the a_j's also minimize the integrated weighted squared error for $f(x)/w(x)$ with $w(x)$ as the weighting function where $f_m(x)/w(x)$ is the polynomial adjustment component of the density estimate. Moreover, as stated in [16], a necessary condition for an m^{th} degree polynomial approximant $p_m(x)$ to converge to a function $g(x)$, is that $g(x)$ be $L^2_{w(x)}-$integrable. Thus, $g(x)$ must satisfy the conditions, $\int_a^b w(x)\,g(x)\,\mathrm{d}x<\infty$ and $\int_a^b w(x)\,g^2(x)\,\mathrm{d}x\;<\;\infty$. Clearly, the first condition is always satisfied when $g(x)=f(x)/w(x)$ and $f(x)$ is a density function, and in terms of $f(x)$, the second one can be re-expressed as $\int_a^b f^2(x)/w(x)\,\mathrm{d}x<\infty$.

III. Density Estimation

A. Kernel Representation of the Density Estimates

Density estimates that are the counterparts of the orthogonal polynomial density approximants discussed in the previous section are shown to admit a certain kernel representation.

Let $\{x_1,x_2,\ldots,x_n\}$ be a simple random sample from a population whose distribution is specified by the random variable X. On replacing the exact raw moments, $\mu_X(\ell)$, by the sample moments, $\hat{\mu}_X(\ell)=\frac{1}{n}\sum_{i=1}^n x_i^\ell,\;\ell=0,1,\ldots,m$, in (14), one obtains the m^{th} degree orthogonal polynomial density estimate,

$$\begin{aligned}
\hat{f}_m(x) &= w(x)\sum_{j=0}^{m}\frac{\varphi_j(x)}{\theta_j}\sum_{\ell=0}^{j}\delta_{j,\ell}\,\hat{\mu}_X(\ell) &(19)\\
&= \frac{w(x)}{n}\sum_{j=0}^{m}\frac{1}{\theta_j}\,\varphi_j(x)\sum_{\ell=0}^{j}\delta_{j,\ell}\sum_{i=1}^{n}x_i^\ell \\
&= \frac{w(x)}{n}\sum_{i=1}^{n}\sum_{j=0}^{m}\frac{1}{\theta_j}\,\varphi_j(x_i)\,\varphi_j(x)\,. &(20)
\end{aligned}$$

On making use of the Christoffel–Darboux formula, that is,

$$\begin{aligned}
&\textstyle\sum_{k=0}^{m}\frac{\varphi_k(x)\;\varphi_k(y)}{\theta_k} \\
&= \frac{\delta_{m,m}}{\delta_{m+1,m+1}}\,\frac{\varphi_{m+1}(x)\;\varphi_m(y)-\varphi_m(x)\;\varphi_{m+1}(y)}{\theta_m\,(x-y)}\,, \quad (21)
\end{aligned}$$

see, for instance, [17], $\delta_{k,k}$ being the coefficient of x^k in $\varphi_k(x)$, and letting

$$\mathcal{K}_m(x,x_i) = w(x)\sum_{j=0}^{m}\frac{1}{\theta_j}\,\varphi_j(x_i)\,\varphi_j(x) \tag{22}$$

or equivalently,

$$\begin{aligned}
&\mathcal{K}_m(x,x_i) \\
&= \frac{w(x)\,\delta_{m,m}}{\delta_{m+1,m+1}}\Big(\frac{\varphi_{m+1}(x)\,\varphi_m(x_i)-\varphi_m(x)\,\varphi_{m+1}(x_i)}{\theta_m\;(x-x_i)}\Big)\,,
\end{aligned} \tag{23}$$

one has the following kernel representation of the orthogonal polynomial density estimate:

$$\hat{f}_m(x) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_m(x,x_i), \tag{24}$$

which is mathematically equivalent to the representation given in (19) in terms of the first m sample moments. It can easily be shown that such kernels integrate to one.

In some instances and, in particular, if one wishes to make use of available results in connection with certain classical orthogonal polynomials, it may be indicated or even necessary to transform the data prior to resorting to the representations the density estimates given in (19), (20) or (24). For example, on letting $y_1,y_2,\ldots,y_n$ be a simple random sample from a distribution specified by the density function, $f(y)$, and making the change of variables $x=g(y)$, where $g(y)$ is a differentiable function of y, the density estimate corresponding to (24) becomes

$$\hat{f}_m(y) = \frac{|g'(y)|}{n}\sum_{i=1}^{n}\mathcal{K}_m(g(y),g(y_i))\,, \tag{25}$$

where $\mathcal{K}_m(\cdot,\cdot)$ is as defined in Equation (23). Oftentimes, it suffices to apply an affine transformation such as

$$g(y) = \frac{y-\tau}{\nu}\,, \tag{26}$$

so that support or the first moment or perhaps the first two moments of the transformed variable coincide(s) with that (those) of the normalized weight function associated with a given type of orthogonal polynomials. This will be illustrated in the Appendices in connection with four classical orthogonal polynomials. However, it should be noted that, as explained in Section II-A, one can always generate a set of orthogonal polynomials from a suitable weight function, in which case there is no need to apply any transformation to the data. For illustrative purposes, two kernels are plotted in Section IV-B; these kernels were obtained from the orthogonal polynomials generated from a mixture of three Gaussian density functions.

When one applies the linear transformation specified by Equation (26), the density estimates (19), (20) and (24), respectively become

$$\hat{f}_m(y) = \frac{1}{\nu}\, w\Big(\frac{y-\tau}{\nu}\Big) \sum_{j=0}^{m} \frac{\varphi_j\Big(\frac{y-\tau}{\nu}\Big)}{\theta_j} \sum_{\ell=0}^{j} \delta_{j,\ell}\ \hat{\mu}_X(\ell) \quad (27)$$

with

$$\hat{\mu}_X(\ell) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{y_i-\tau}{\nu}\Big)^{\ell}$$

$$= \frac{1}{\nu^{\ell}} \sum_{k=0}^{\ell} \binom{\ell}{k} \hat{\mu}_Y(k)\,(-1)^{\ell-k}\,\tau^{\ell-k}\,,$$

where

$$\hat{\mu}_Y(k) = \frac{1}{n}\sum_{i=1}^{n} y_i^k\,, \quad k = 0, 1, \ldots, \ell\,,$$

$$\hat{f}_m(y) = \frac{1}{n\,\nu}\, w\Big(\frac{y-\tau}{\nu}\Big) \sum_{i=1}^{n}\sum_{j=0}^{m} \frac{1}{\theta_j}\, \varphi_j\Big(\frac{y_i-\tau}{\nu}\Big)\ \varphi_j\Big(\frac{y-\tau}{\nu}\Big) \quad (28)$$

and

$$\hat{f}_m(y) = \frac{1}{n\,\nu}\sum_{i=1}^{n} \mathcal{K}_m\Big(\frac{y-\tau}{\nu}, \frac{y_i-\tau}{\nu}\Big)\,. \quad (29)$$

In the case of a density approximant, which is based on $\mu_Y(k) = E(Y^k),\ k = 0, 1, \ldots, m$, the exact moments of the random variable Y, the linear transformation yields the following density approximant corresponding to (14):

$$f_m(y) = w\Big(\frac{y-\tau}{\nu}\Big) \sum_{j=0}^{m}\sum_{\ell=0}^{j} \frac{\delta_{i,\ell}\ \mu_X(\ell)}{\nu\,\theta_j}\ \varphi_j\Big(\frac{y-\tau}{\nu}\Big)\,, \quad (30)$$

where

$$\mu_X(\ell) = \frac{1}{\nu^{\ell}} \sum_{k=0}^{\ell} \binom{\ell}{k} \mu_Y(k)(-1)^{\ell-k}\tau^{\ell-k}\,.$$

The kernels associated with the Legendre, Jacobi, Laguerre and Hermite classical orthogonal polynomials are explicitly given in Appendices $A - D$. Additionally, when the weight functions involve parameters, convenient estimates thereof are provided in terms of the sample moments.

The main results derived in this section, that is, the connection between approximants and estimates and the dual representation of the density estimates, ought to provide valuable insights into the orthogonal polynomial density estimation methodology advocated herein and lead to a heightened appreciation of this approach as a viable alternative to other density estimation techniques.

Since the kernels specified by Equation (23) become more concentrated as the number of moments being used increases, caution needs be exercised when selecting the number of terms to be included in the polynomial adjustment component.

B. Degree Selection Criterion

This section addresses the question of determining the degree of the polynomial adjustment in the orthogonal polynomial density estimates. A convenient criterion was proposed by [18] in connection with Fourier series. Some preliminary considerations are in order before discussing its extension to orthogonal polynomial density estimates and proposing some refinements.

We first note that $\hat{f}_m(x)$, as given in (20), can be rewritten as

$$\hat{f}_m(x) = c\, w(x) \sum_{j=0}^{m} \hat{a}_j\, \varphi_j(x) \quad (31)$$

where

$$\hat{a}_j = \frac{1}{c\,n}\sum_{i=1}^{n}\frac{1}{\theta_j}\,\varphi_j(x_i)\,. \quad (32)$$

It can be assumed without any loss of generality that $c = 1$ and $\theta_j = 1$ (by renormalizing the orthogonal polynomials) for $j = 0, 1, \ldots, m$. The degree of the adjustment component acts in fact as a smoothing parameter, smaller values of m leading to smoother estimates. The number of terms to be included in the adjustment component will be determined from an estimate of $J(m)$, the mean integrated weighted squared error between the true density and an m^{th} degree orthogonal polynomial density estimate, that is,

$$\begin{aligned} J(m) &= E\Big(\int_a^b \frac{1}{w(x)}\,(\hat{f}_m(x) - f(x))^2\,\mathrm{d}x\Big) \\ &= E\Big(\int_a^b \frac{1}{w(x)}\,\Big(w(x)\sum_{i=0}^{m}(\hat{a}_i - a_i)\,\varphi_i(x) \\ &\quad - w(x)\sum_{i=m+1}^{\infty} a_i\,\varphi_i(x)\Big)^2\,\mathrm{d}x\Big) \end{aligned} \quad (33)$$

where $\hat{f}_m(x)$ is as specified in (31) or (24) and a_j is as specified in (13) in terms of the moments of the random variable X. Making use of the orthogonality property specified by (2) with $\theta_j = 1$, for $j = 0, 1, \ldots$, one can re-express Equation (33) as follows:

$$\begin{aligned} J(m) &= E\Big(\sum_{i=0}^{m}(\hat{a}_i - a_i)^2 + \sum_{i=m+1}^{\infty} a_i^2\Big) \\ &= \sum_{i=0}^{m} n^{-1}(d_i^2 - a_i^2) + \sum_{i=m+1}^{\infty} a_i^2, \\ &= \sum_{i=0}^{m}\big(n^{-1}(d_i^2 - a_i^2) - a_i^2\big) + \sum_{i=0}^{\infty} a_i^2\,, \end{aligned} \quad (34)$$

where $d_i^2 = E\left(\varphi_i^2(X)\right)$, as explained for instance in [15] and [19].

It is seen from Equation (34) that the j^{th} term should be included whenever $J(j) < J(j-1)$, which is equivalent to $n^{-1}(d_j^2 - a_j^2) < a_j^2$. Clearly,

$$\widehat{d_i^2} = \frac{1}{n}\sum_{k=1}^{n}\varphi_i^2(x_k), \quad (35)$$

is an unbiased estimator of d_i^2, while

$$\widehat{a_i^2} = \frac{n}{n-1}(\hat{a}_i^2 - \frac{1}{n}\widehat{d_i^2}) \tag{36}$$

is an unbiased estimator of a_i^2, see [19]. Thus, in terms of the unbiased estimates of d_i^2 and a_i^2, the condition $n^{-1}(d_j^2 - a_j^2) < a_j^2$ becomes

$$\mathrm{KTS}(j) \;\equiv\; \hat{a}_j^2 - 2\,\widehat{d_j^2}/(n+1) > 0\,, \tag{37}$$

where KTS(j) denotes the j^{th} degree *Kronmal-Tarter statistic*, the estimates $\hat{a}_j$ and $\widehat{d_j^2}$ being as given in (32) and (35), respectively.

The *Kronmal-Tarter criterion* consists of including all the terms until t successive terms fail the test specified by inequality (37). Such a rule was suggested in [18] and [20] for density estimators expressed in terms of Fourier series, and later generalized to orthogonal series density estimators in [19] which pointed out some of its drawbacks.

An empirical simulation study in which we generated various types of distributions, suggests that the Kronmal-Tarter criterion often produces unsatisfactory density estimates when applied to skewed or multimodal distributions. This is due to the fact that this criterion can select a degree m that is smaller than m_0, the degree corresponding to the maximum value of the the Kronmal-Tarter statistic.

Two refinements are being proposed. The first will be referred to as the *MKT criterion* as it chooses the degree m_0 corresponding to the maximum value of the Kronmal-Tarter statistic. Referring to (34), it is seen that m_0 is the degree associated with the largest of the negative terms in the first sum, in absolute value, so that additional terms will not contribute as significantly to reducing the estimated mean integrated weighted squared error. It was also observed that for strongly skewed or multimodal distributions, the *MKT* criterion produces estimates that tend to be too smooth. Thus, in order to obtain estimates that duplicate more closely the features of the underlying distribution in such instances, we propose to select the degree, m_1 that corresponds to the first occurrence of a negative value of the Kronmal-Tarter statistic *past* m_0—at which point the estimated mean integrated squared error starts to increase. We shall refer to this rule as the *RSKT criterion* for *right-sided Kronmal-Tarter criterion*. Intermediate values of m could also be considered.

Before making use of either of the proposed stopping rules, the minimum value of m could generally be taken to be four. However, if it is known a priori that the underlying distribution is symmetric and unimodal and that it closely matches the selected base distribution, then the minimum degree could be set to two. On the other hand, if the distribution is multimodal and its modes are separated, the minimum degree could be set equal to twice the number of modes plus two. For illustration and comparison purposes, both the *MKT* and *RSKT* criteria are utilized in the examples presented in Section IV.

C. The Density Estimation Methodology

The orthogonal polynomial density estimation methodology comprises the following steps:

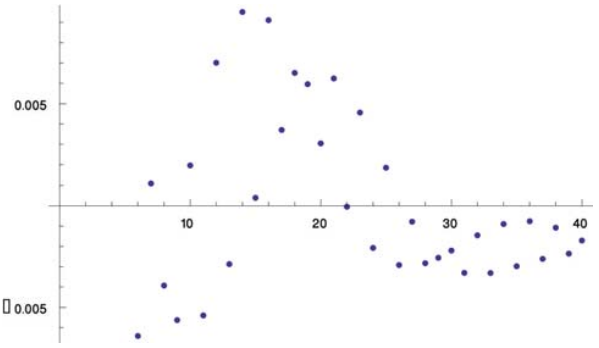

Fig. 1. The Kronmal-Tarter statistic for the certificates of deposit rates data

1) An initial density estimate, referred to as base density and denoted $c\,w(x)$, is selected on the basis of some preliminary estimate such as a histogram of the data or some prior knowledge of the underlying distribution. A mixture of densities is indicated in the case of a multimodal distribution whose modes are separated.
2) A sequence of orthogonal polynomials $\{\varphi_0(x), \varphi_1(x), \ldots, \}$ is generated from the chosen base density as explained in Section II-B, and the associated kernels are determined from (23). For the uniform, beta, gamma and normal base densities, the corresponding orthogonal kernels (up to certain affine transformations) are the Legendre, Jacobi, Laguerre and Hermite kernels, which are explicitly provided in Appendices $A - D$.
3) The density is estimated by means of the kernel formula given in (24) or equivalently from (19) in terms of the sample moments, the degree m being determined from the MKT and/or the RSKT criteria introduced in Section III-B.
4) The end points the resulting density estimate are taken to be the points of intersection with the abscissa. The resulting function, which is non-negative, is then renormalized to produce a *bona fide* density function.

IV. Application to Two Data Sets

The density estimation technique described in Section III-C is applied to the certificates of deposit rates and the galaxy velocities data sets. All the calculations were carried out with the symbolic computational software, *Mathematica*. The code is available from the authors upon request.

A. The Certificates of Deposit Rates Data Set

Consider the data set analyzed in [21], which represents the three-month certificates of deposit rates for 69 Long Island banks and thrift institutions, as given in the August 23, 1989 issue of *Newsday*. The 13-bin histogram of this data set appears to be mainly bimodal, with a primary mode around 8.5 and a secondary mode around 7.9. Not surprisingly, the data contains two subgroups, namely 29 commercial banks and 40 thrift (Savings and Loans) institutions.

In order to make use of the estimate given in Equation (29) in conjunction with the Hermite polynomial kernel specified

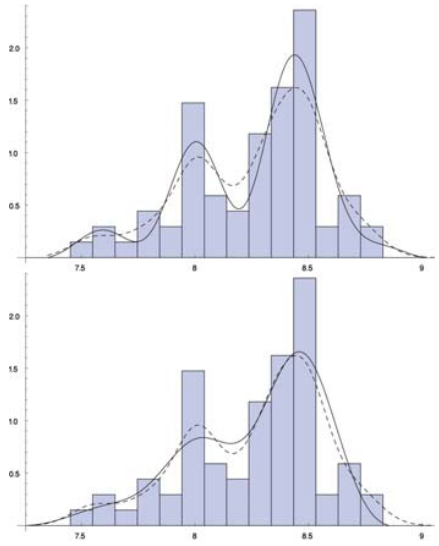

Fig. 2. Histogram, kernel density estimate (dashed line) and Hermite polynomial density estimates (upper panel: degree 21; lower panel: degree 14) for the the certificates of deposit rates data

by Equation (A.4.1), one needs to apply to the original data the linear transformation, $g(y) = (y - \hat{\mu})/(\sqrt{2}\hat{\sigma})$, with $\hat{\mu} = 8.26$ and $\hat{\sigma} = 0.298$, as explained in Appendix D. Equivalently, one may make use of a Gaussian density function with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ as initial estimate, generate a sequence of orthogonal polynomial by means of the Gram-Schmidt orthogonalization process and apply formula (24), in which case no transformation is required. Referring to the plot of the Kronmal-Tarter statistic shown in Figure 1, one would select degrees $m_0 = 14$ and $m_1 = 21$ in accordance with the MKT and RSKT criteria introduced in Section III-B. The Hermite polynomial density estimates of degrees 21 and 14, which can be evaluated from formula (29) conjunction with the Hermite kernels specified in (A.4.1), are respectively plotted (as solid lines) in the upper and lower panels of Figure 2 along with a two-stage plug-in kernel density estimate with bandwidth 0.0913129 (dashed line) and a 13-bin histogram of the data. As expected, the density estimates exhibit two main modes. It should be observed that the higher-degree density estimate captures a possibly spurious third mode of lesser importance which is present in the left tail of the histogram.

B. The Galaxy Velocities Data Set

In this case, the proposed density estimation methodology is applied to a data set consisting of 82 galaxies velocities in km/sec from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. Multimodality in such surveys is evidence for voids and superclusters in the far universe. This data set has been previously analyzed in [22]. It is often used as a benchmark example in mixture analysis. Given that the modes are clearly separated, a mixture of three Gaussian density functions—weighted in proportion to the number of points included in each of the three subintervals—was used as base density. The associated orthogonal polynomials were generated from Equation (7). For illustrative

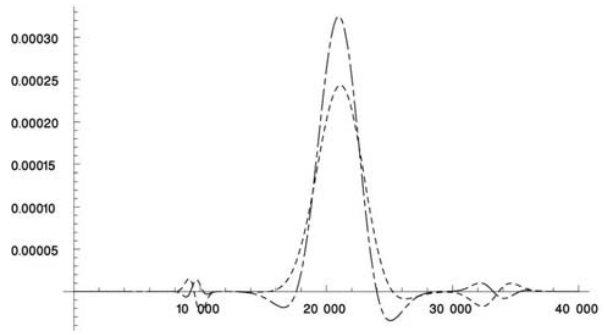

Fig. 3. Two kernels for the galaxy velocities data. $\mathcal{K}_4(x, 20828.2)$: short dashes; $\mathcal{K}_8(x, 20828.2)$: long dashes

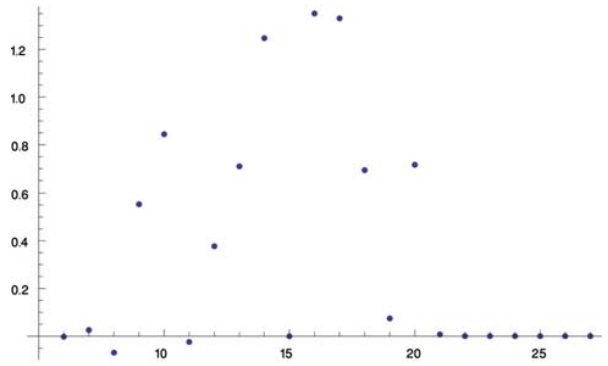

Fig. 4. The Kronmal-Tarter statistic for the galaxy velocities data

purposes, two kernels are plotted in Figure 3, one of degree four and the other of degree eight. The number of terms in the polynomial component of the density estimates were determined by applying the MKT and the RSKT stopping criteria (both described in Section III-B). It is seen from Figure 4 that $m_0 = 16$ and $m_1 = 23$. The resulting density estimates of degrees 23 and 16, which can be obtained from (27) in terms of the sample moments or from (29) in terms of kernels, are respectively plotted in the upper and lower panels of Figure 5 along with a two-stage plug-in kernel density estimate with bandwidth 1155.33 (the dashed line in both panels) and a 30-bin histogram of the data.

APPENDIX A
LEGENDRE POLYNOMIAL KERNELS

Let $\varphi_k^{\mathrm{L}}(x) = \sum_{\ell=0}^{k} \delta_{k,\ell}^{\mathrm{L}}\, x^\ell$ denote a k^{th} degree Legendre polynomial; then the coefficient of x^ℓ as given explicitly in [8] in a compact form is

$$\delta_{k,\ell}^{\mathrm{L}} = \frac{(-1)^k + (-1)^\ell}{2^{k+1}} \frac{(-1)^{\frac{3k-\ell}{2}}(k+\ell)!}{\Gamma(\frac{k-\ell}{2}+1)\,\Gamma(\frac{k+\ell}{2}+1)\;\ell!}, \quad \ell = 0, 1, \ldots, k,$$

so that $\delta_{n,n}^{\mathrm{L}} = 2n!/\left(2^n(n!)^2\right)$. In this case, the support is $(-1, 1)$, the weight function is $w(x) = 1/2$, and the orthogonality factor is $\theta_m = 2/(2m+1)$. Thus, according

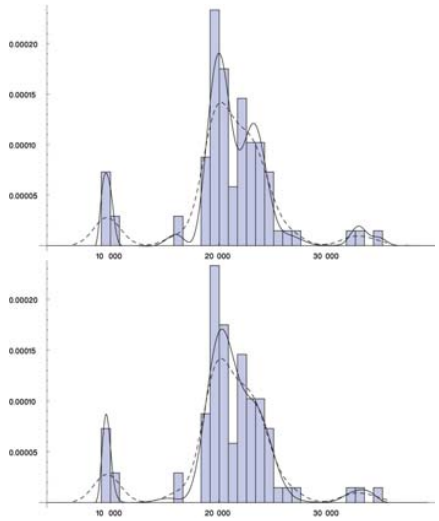

Fig. 5. Histogram, kernel density estimate (dashed line) and mixture of three normals polynomial density estimates (upper panel: degree 23; lower panel: degree 16) for the galaxy velocities data

to (23), the m^{th} degree kernel associated with the Legendre polynomials is

$$\mathcal{K}_m(x,x_i)=\frac{(m+1)}{2}\,\frac{\left(\varphi^{\mathrm{L}}_{m+1}(x)\,\varphi^{\mathrm{L}}_{m}(x_i)-\varphi^{\mathrm{L}}_{m}(x)\,\varphi^{\mathrm{L}}_{m+1}(x_i)\right)}{x-x_i} \tag{A.1.1}$$

where $\varphi^{\mathrm{L}}_k(x)=\sum_{i=0}^{k}\delta^{\mathrm{L}}_{k,i}\,x^i$ denotes a Legendre polynomial of degree k.

In order to estimate a density function defined on the interval (a,b) in terms of Legendre polynomials, one must first apply the linear transformation,

$$g(y)=\frac{2y-(a+b)}{b-a}, \tag{A.1.2}$$

which maps the interval (a,b) onto the interval $(-1,1)$. The density estimates are then obtained either from Equation (27) or Equation (29) in conjunction with the kernel representation given in (A.1.1) after letting $\tau=\frac{a+b}{2}$ and $\nu=\frac{b-a}{2}$, referring to (26).

Appendix B
Jacobi Polynomial Kernels

The following explicit representation of the coefficients of a k^{th} degree Jacobi polynomial with parameters α and β whose support is the interval $(-1,1)$, is given in [8]:

$$\delta^{\alpha,\beta}_{k,\ell}=\sum_{h=0}^{k}\sum_{j=0}^{\ell}\frac{(k+\alpha)!\,(k+\beta)!\,(-1)^{k-h-\ell+j}}{2^k(k+\alpha-h)!\,j!\,(\beta+h)!\,(h-j)!\,(\ell-j)!\,(k-h-\ell+j)!} \tag{A.2.1}$$

and as shown in [17], $\delta^{\alpha,\beta}_{n,n}=\frac{\Gamma(2n+\alpha+\beta+1)}{2^n\,n!\;\Gamma(n+\alpha+\beta+1)}$. In this case, the weight function and m^{th} degree orthogonality factor are respectively $w(x)=(1-x)^{\alpha}(1+x)^{\beta},\quad -1<x<1$, and

$$\theta_m=\frac{2^{\alpha+\beta+1}\Gamma(m+\alpha+1)\,\Gamma(m+\beta+1)}{m!\,(2m+\alpha+\beta+1)\,\Gamma(m+\alpha+\beta+1)}, \tag{A.2.2}$$

the normalizing constant c being equal to $\Gamma(\alpha+\beta+2)/\left(2^{\alpha+\beta+1}\,\Gamma(\alpha+1)\,\Gamma(\beta+1)\right)$. Thus, the m^{th} degree Jacobi polynomial kernel is

$$\mathcal{K}_m(x,x_i)=(1-x)^{\alpha}(1+x)^{\beta}\;\frac{(m+1)(m+\alpha+\beta+1)}{(2m+\alpha+\beta+2)}\times\frac{m!\,\Gamma(m+\alpha+\beta+1)}{2^{\alpha+\beta}\,\Gamma(m+\alpha+1)\,\Gamma(m+\beta+1)}\times\frac{\left(\varphi^{\alpha,\beta}_{m+1}(x)\varphi^{\alpha,\beta}_{m}(x_i)-\varphi^{\alpha,\beta}_{m}(x)\varphi^{\alpha,\beta}_{m+1}(x_i)\right)}{(x-x_i)},$$

where $\varphi^{\alpha,\beta}_k(x)=\sum_{i=0}^{k}\delta^{\alpha,\beta}_{k,i}\,x^i$ denotes a Jacobi polynomial of degree k with parameters α and β. The Jacobi polynomial kernels behave similarly to the the Legendre polynomial kernels.

The transformation specified by Equation (A.1.2) for distributions whose support is the interval (a,b) also applies in this case. Letting $\hat{\mu}_Y(j)$ denote the j^{th} sample raw moment of the original observations, the parameters α and β are estimated as follows by matching the moments of Y to those of the normalized weight function:

$$\hat{\alpha}=\frac{(b-\hat{\mu}_Y(1))(\hat{\mu}_Y(2)+ab-(a+b)\hat{\mu}_Y(1))}{(b-a)(\hat{\mu}_Y(1)^2-\hat{\mu}_Y(2))}-1 \tag{A.2.3}$$

and

$$\hat{\beta}=\frac{(\hat{\mu}_Y(1)-a)(\hat{\mu}_Y(2)+ab-(a+b)\hat{\mu}_Y(1))}{(b-a)(\hat{\mu}_Y(1)^2-\hat{\mu}_Y(2))}-1. \tag{A.2.4}$$

Appendix C
Laguerre Polynomial Kernels

In the case of the Laguerre polynomials with parameter ϕ, the support is the interval $(0,\infty)$,

$$\delta^{\phi}_{k,\ell}=\frac{(-1)^{\ell}\Gamma(k+\phi+1)}{(k-\ell)!\,\Gamma(\phi+\ell+1)\,\ell!}$$

and $\delta^{\phi}_{n,n}=(-1)^n/n!$. Moreover, $w(x)=x^{\phi}e^{-x}$, $c=1/\Gamma(\phi+1)$ and $\theta_m=\Gamma(\phi+m+1)/m!$. Thus, according to (23), the m^{th} degree Laguerre polynomial kernel is given by

$$\mathcal{K}_m(x,x_i)=-x^{\phi}e^{-x}\frac{(m+1)!\left(\varphi^{\phi}_{m+1}(x)\,\varphi^{\phi}_{m}(x_i)-\varphi^{\phi}_{m}(x)\,\varphi^{\phi}_{m+1}(x_i)\right)}{\Gamma(\phi+m+1)\,(x-x_i)},$$

where $\varphi^{\phi}_k(x)=\sum_{i=0}^{k}\delta^{\phi}_{k,i}x^i$ denotes a k^{th} degree Laguerre polynomial with parameter ϕ.

Let $\hat{\mu}_Y(j),\ j=0,1,2,\ldots$, denote the j^{th} sample moment associated with a random variable Y for which a gamma distribution is a suitable initial approximation, and let $\hat{\nu}=\left(\hat{\mu}_Y(2)-\hat{\mu}_Y(1)^2\right)/\hat{\mu}_Y(1)$, and $\hat{\phi}=(\hat{\mu}_Y(1)/\hat{\nu})-1$. Then, the transformation, $g(y)=y/\hat{\nu}$, is required since the scaling factor in the weight function is one. The resulting density estimates are obtained either from Equations (27), (28) or (29) with $\tau=0$ and ν and ϕ respectively estimated by $\hat{\nu}$ and $\hat{\phi}$.

APPENDIX D
HERMITE POLYNOMIALS KERNELS

As for Hermite polynomials,

$$\delta^{\mathrm{H}}_{k,\ell} = \frac{((-1)^k + (-1)^\ell)(-1)^{\frac{3k-\ell+2}{2}} 2^{\frac{(k+\ell-2)}{2}} k!}{(k-\ell)!\,\ell!} \prod_{j=0}^{\frac{k-\ell}{2}} (2j-1),$$

and $\delta^{\mathrm{H}}_{n,n} = 2^n$. Their associated weight function and orthogonality factor are respectively $w(x) = e^{-x^2}$ and $\theta_m = \sqrt{\pi}\ 2^m\ m!$, the normalizing constant c being equal to $1/\sqrt{\pi}$. Thus, the m^{th} degree kernel associated with Hermite polynomials is

$$\mathcal{K}_m(x, x_i) = e^{-x^2} \frac{1}{\sqrt{\pi}\ 2^{m+1}\ m!} \frac{\varphi^{\mathrm{H}}_{m+1}(x)\varphi^{\mathrm{H}}_m(x_i) - \varphi^{\mathrm{H}}_m(x)\varphi^{\mathrm{H}}_{m+1}(x_i)}{(x - x_i)}, \tag{A.4.1}$$

where $\varphi^{\mathrm{H}}_k(x) = \sum_{i=0}^{k} \delta^{\mathrm{H}}_{k,i} x^i$ denotes a k^{th} degree Hermite polynomial.

In this case, the appropriate transformation is $g(y) = \frac{y-\hat{\mu}}{\sqrt{2}\hat{\sigma}}$ where $\hat{\mu}$ and $\hat{\sigma}$ are respectively the sample mean and sample standard deviation of the original observations, so that τ is $\hat{\mu}$ and ν is $\sqrt{2}\,\hat{\sigma}$ in Equations (27), (28) and (29). We note that this transformation produces a distribution having mean zero and variance $1/2$, as is the case for the normalized weight function $\pi^{-1/2}\, e^{-x^2}$.

ACKNOWLEDGMENT

The financial support of the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

REFERENCES

[1] Izenman A J. Recent development in nonparametric density density estimation. Journal of the American Statistical Association, 1991, 86:205–224.
[2] Watson G S. Density estimation by orthogonal series. Annals of Mathematical Statistics, 1969, 40:1496–1498.
[3] Rosenblatt M. Curve estimates. The Annals of Mathematical Statistics, 1971, 42:1815–1842.
[4] Hall P. On the rate of convergence of orthogonal series density estimators. Journal of Royal the Royal Statistical Society, Ser. B, 1986, 48:115–122.
[5] Johnstone I M, Silverman B W. Speed of estimation in positron emission tomography and related inverse problems. The Annals of Statistics, 1990, 18:251–280.
[6] Provost S B. Moment-based density approximants. The Mathematica Journal, 2005, 9:727–756.
[7] Ha H-T, Provost S B. A viable alternative to resorting to statistical tables. Communications in Statistics–Simulation and Computation, 2007, 36:1135–1151.
[8] Provost S B, Ha H-T. On the inversion of certain moment matrices. Linear Algebra and Its Applications, 2009, 430:2650–2658.
[9] Brunk H B. Univariate density estimation by orthogonal series. Biometrika, 1978, 65:521–528.
[10] Parzen E. Nonparametric statistical data modeling. Journal of the American Statistical Association, 1979, 74:105–121.
[11] Ruppert D, Cline D B H. Bias reduction in kernel density estimation by smoothed empirical transformations. Institute of Mathematical Statistics, 1994, 22:185–210.
[12] Hjort N L, Jones M C. Locally parametric nonparametric density estimation. The Annals of Statistics, 1996, 24:1619–1647.
[13] Rao C R. Linear Statistical Inference and Its Applications. New York: Wiley, 1973.
[14] Arfken G. Gram-Schmidt Orthogonalization. Orlando: Academic Press, 1985.
[15] Anderson G L, De Figueiredo R J P. An adaptive orthogonal-series estimator for probability density functions. The Annals of Statistics, 1980, 8:347–376.
[16] Alexits G. Convergence Problems of Orthogonal Series. New York: Pergamon Press, 1961.
[17] Hildebrand F B. Introduction to Numerical Analysis. New York: McGraw-Hill, 1956.
[18] Kronmal R, Tarter M. The estimation of probability densities and cumulatives by Fourier series methods. Journal of the American Statistical Association, 1968, 63:925–952.
[19] Diggle P J, Hall P. The selection of terms in an orthogonal series density estimator. Journal of American Statistical Association, 1986, 81:230–233.
[20] Tarter M, Kronmal R. An introduction to the implementation and theory of nonparametric density estimation. The American Statistician, 1976, 30:105–112.
[21] Simonoff J S. Smoothing Methods in Statistics. New York: Springer, 1996.
[22] Roeder K. Density estimation with confidence sets exemplified by superclusters and voids in galaxies. Journal of the American Statistical Association, 1990, 85:617–624.