A Fast Non-Parametric Density Estimation Algorithm

ÖMER EĞECIOĞLU and ASHOK SRINIVASAN

Department of Computer Science, University of California Santa Barbara, CA 93106

SUMMARY

Non-parametric density estimation is the problem of approximating the values of a probability density function, given samples from the associated distribution. Non-parametric estimation finds applications in discriminant analysis, cluster analysis, and flow calculations based on Smoothed Particle Hydrodynamics. Usual estimators make use of kernel functions, and require on the order of n^2 arithmetic operations to evaluate the density at n sample points. We describe a sequence of special weight functions which requires almost linear number of operations in n for the same computation.

KEY WORDS: Non-parametric estimation, probability density, kernel method.

1 INTRODUCTION

The use of probability density estimation in data analysis is well established [15, 12, 11, 2]. In the non-parametric case no assumption is made about the type of the distribution from which the samples are drawn. This is in contrast to parametric estimation in which the density is assumed come to from a given family, and the parameters are then estimated by various statistical methods. A comprehensive bibliography on non-parametric estimation can be found in Silverman [16], and Nadaraya [10]. Results of experimental comparison of some widely used methods appear in [5].

An important application of non-parametric density estimation is in computational fluid mechanics. When flow calculations are performed in a Lagrangian framework, a set of points in space are evolved using the governing equations. However in time this usually leads to mesh distortion and numerical difficulties. Problems with mesh distortion can be eliminated to a certain extent by the use of Smoothed Particle Hydrodynamics (SPH) [9, 4]. SPH treats the points being tracked as samples drawn from an unknown probability distribution. Because of the iterative nature of SPH, density is estimated at every time step, and the efficiency of the computational aspects becomes even more important.

We propose a cosine based estimator which is similar to kernel estimators in convergence properties, but computationally more efficient. Conditions for the convergence of the method and experimental verification of its accuracy as compared to kernel based estimators for practical test cases are also presented. The experiments indicate how to choose the optimal estimator as a function of the number of points while keeping the error small.

2 METHODS OF ESTIMATION

Various methods have been proposed for non-parametric density estimation, such as the kernel [11, 1] and the orthogonal series methods [13]. In the kernel method, the value of the density at x is estimated as

$$f_n(x) = \frac{1}{nA_h} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

where X_i are the positions of the samples drawn from a probability distribution with an unknown density function f, K is a kernel function, h is the window width, and A_h is a normalization factor to make f_n into a probability density. One of the drawbacks of the kernel method is the computational cost involved. Even though it is possible to reduce the cost using a series expansion of the kernel, this approach becomes problematic for narrow window widths.

Our estimator for non-parametric density uses properties of the cosine function, has convergence properties which are similar to those of kernel based estimators, but at the same time has the ease of evaluation of a series expansion. The role of the window size parameter h of the kernel method is replaced by a parameter m in our method, and f_n is now of the form

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} c_m(x - X_i) \ . \tag{1}$$

The definition of the function c_m allows for the efficient computation of the n values

$$f_n(X_1), f_n(X_2), \ldots, f_n(X_n) \tag{2}$$

using only $O(m^2 n)$[1] arithmetic operations, where m need not be large as long as it increases without bound with n. This is in contrast to the $O(n^2)$ operations required by the kernel method. After the computation of $f_n(X_1), f_n(X_2), \ldots, f_n(X_n)$, our method requires only $O(m^2)$ operations for each subsequent evaluation $f_n(x)$ for an arbitrary x. This is in contrast to the $O(n)$ operations required by the kernel method. The factor m^2 in this operation count is replaced by m^{1+d} when the sample space is the d-dimensional unit sphere S^d. Such sample spaces arise naturally in problems involving directional data.

3 CONVERGENCE OF THE METHOD

Let $X_j,\ j = 1, 2, \ldots$ be a sequence of i.i.d. random variables (observations) with values on the real line $\mathbb{R}$. Suppose their common distribution function has density density $f(x)$. As an estimator to $f(x)$, we consider a non-parametric estimator of the form (1), where m is to be determined as a function of n. Let

$$c_m(x) \ = \ \begin{cases} \frac{1}{A_m} \cos^{2m}(\frac{x}{2}) \ , & x \in [-\pi, \pi] \\ 0 & otherwise, \end{cases} \tag{3}$$

where the factor A_m normalizes c_m to a probability density function. Two popular criteria for convergence that are usually considered are given below. The estimate converges to the true density under the *Mean Integrated Square Error* (MISE) criterion if

$$\int E(f_n(x) - f(x))^2 dx \to 0$$

[1] $f(n) = O(g(n))$ means $\lim_{n\to\infty} f(n)/g(n) < \infty$, and $f(n) = o\,(g(n))$ means $\lim_{n\to\infty} f(n)/g(n) = 0$.

as $n \to \infty$. We have *almost sure* (a.s.) convergence to the true density if

$$\sup |f_n(x) - f(x)| \to 0$$

almost surely as $n \to \infty$. First, we state conditions for the almost sure convergence of the estimates of the density and its derivatives on $\mathbb{R}$, and the convergence in the MISE sense on S^1 and S^2. The conditions are similar to those required by kernel based methods. The proofs of these theorems appear elsewhere [3].

Theorem 1 *Suppose $f(x)$ is uniformly continuous on $\mathbb{R}$, and $f_n(x)$ is as given in (1). If $m \to \infty$ as $n \to \infty$, and $m = o\,(\frac{n}{\log n})$, then*

$$\sup_{x \in \mathbb{R}} |f_n(x) - f(x)| \to 0 \qquad a.s.$$

Flow calculations often require the computation of the values of not only the unknown density f, but also its derivatives. Let $C^r(\mathbb{R})$ denote r-fold continuously differentiable functions on $\mathbb{R}$. Assume that $f \in C^r(\mathbb{R})$. Denote the r-th derivative of $f(x)$ by $f^{(r)}(x)$. As an estimator of $f^{(r)}(x)$ we take the r-th derivative $f_n^{(r)}(x)$ of $f_n(x)$. The conditions for almost sure convergence of the estimates of the derivatives of the density on $\mathbb{R}$ are given by the following theorem.

Theorem 2 *Suppose $f(x) \in C^r(\mathbb{R})$, $f^{(r)}(x)$ is uniformly continuous, and $f_n^{(r)}(x)$ is as obtained by differentiating (1). If $m \to \infty$ as $n \to \infty$, and $m^{2r+1} = o\,(\frac{n}{\log n})$, then*

$$\sup_{x \in \mathbb{R}} |f_n^{(r)}(x) - f^{(r)}(x)| \to 0 \qquad a.s.$$

Consider the estimation of a density function defined on S^d. Define the weight function by

$$c_m(x) \;=\; \frac{1}{A_m} \cos^{2m}(\frac{\alpha_x}{2}) \tag{4}$$

where $x \in S^d$ and α_x is the length of the shortest arc between x and a fixed point on S^d (which we take to be the point $\vec{0}$). As before A_m is the normalization constant. Conditions for convergence of the MISE for this case are given by

Theorem 3 *Suppose $f \in C^1(S^d)$, $d = 1, 2$, and let $f_n(x)$ be as given in* (1) *and* (4). *If $m \to \infty$ as $n \to \infty$ and $m = o\,(n^{2/d})$, then*

$$\text{MISE} \;=\; \int E(f_n(x) - f(x))^2 dx \to 0$$

4 FAST COMPUTATION OF THE DENSITY ESTIMATOR

In this section, we describe an efficient algorithm for the computation of the values of the density estimator $f_n(x)$ at a set of n observed points $X_1, X_2, \ldots, X_n$ on the unit circle S^1 (1-D case) and on the unit sphere S^2 (2-D case). The idea of the 1-D case easily extends to $\mathbb{R}$. We also show that if the value of f_n at some arbitrary x is desired, then this is also easily accomplished once $f_n(X_1), f_n(X_2), \ldots, f_n(X_n)$ have been computed.

We show that for any x, $f_n(x)$ can be expressed as a polynomial of total degree m in the coordinates of x, and that the coefficients of this polynomial can be determined in turn from the coordinates of the X_i. Using the half angle formula for cosine, we obtain from (1) and (3)

$$f_n(x) = \frac{1}{nA_m} \sum_{i=1}^{n} \left(\frac{1 + \cos(x - X_i)}{2} \right)^m . \tag{5}$$

Let $x = (x_1, x_2)$ and $X_i = (X_{i1}, X_{i2})$ be the Cartesian coordinates of x and X_i, $i = 1, 2, \ldots, n$ on S^1 and let $<,>$ represent the standard inner product on $\mathbb{R}^2$. Then $\cos(x - X_i) =< x, X_i >$. Substituting this into (5) and using the multinomial theorem we get

$$f_n(x) = \frac{1}{2^m A_m} \sum_{r+s \leq m} \binom{m}{r, s} M(r, s)\, x_1^r x_2^s \tag{6}$$

where

$$\binom{m}{r, s} = \frac{m!}{r!s!(m - r - s)!}$$

and

$$M(r, s) = \frac{1}{n} \sum_{i=1}^{n} X_{i1}^r X_{i2}^s .$$

Now consider the number of operations required for the evaluation of $f_n(x)$ given the observations $X_1, X_2, \ldots, X_n$ on S^1: The powers X_{i1}^r and X_{i2}^r for a fixed i and $r = 1, 2, \ldots, m$ can be computed with $O(m)$ multiplications. Doing this for $i = 1, 2, \ldots, n$ requires $O(mn)$ multiplications. After the conclusion of this step, each of the $O(m^2)$ averages $M(r, s)$ for a given r and s can be computed with an additional $O(n)$ operations. Since there are a total of $O(m^2)$ terms in $f_n(x)$ corresponding to the pairs r, s with $0 \leq r, s \leq m$, this means that the coefficients of the polynomial in (6) can be computed with a total of $O(m^2 n)$ arithmetic operations. Once the coefficients of $f_n(x)$ have been computed, to evaluate $f_n(x)$ with $x = (x_1, x_2)$ we calculate the powers x_1^r and x_2^r for $r = 1, 2, \ldots, m$ in $O(m)$ operations. Since the coefficients are already available, the remaining computation in (6) requires only an additional $O(m^2)$ multiplications and additions. Note that kernel based non-parametric estimators to the density require $O(n^2)$ arithmetic operations for the computation of the values in (2), since each $K(\frac{X_j - X_i}{h})$ needs to be evaluated for $1 \leq i, j \leq n$.

A fast algorithm for the evaluation of $f_n(x)$ defined on S^2 is constructed similarly. In this case the values (2) of the approximate density can be evaluated in $O(m^3 n)$ arithmetic operations. After this preprocessing step, $f_n(x)$ can be computed for any $x \in S^2$ with only $O(m^3)$ additional operations.

5 EXPERIMENTAL RESULTS

For our convergence conditions to be satisfied, m must increase without bound with n, but not too fast. Theoretically, we can take m to be as slowly increasing as we like. The problem with this is that the magnitude of m controls the error terms of the convergence proofs. This is a crucial point which makes experimental results invaluable in determining good values of m in practice. With this in mind, numerical experiments were carried out to determine the error as a function of m and n using various distributions. The results for two tested distributions are presented below. Comparisons were also made with the following kernel estimator [8]:

$$K(x) = \begin{cases} \frac{1}{A}\left(1 - 1.5x^2 + 0.75x^3\right) , & x \in [0,1] \\ \frac{1}{A}\, 0.25(2-x)^3 , & x \in [1,2] \\ 0 & otherwise \end{cases}$$

where A is a normalization constant.

Figure 1 compares the accuracy of the cosine and the kernel estimates for a normal distribution with mean π and standard deviation 1, generated by using the function *randn* of *MATLAB*. The results given are the mean of 50 trials. Note that the two curves are virtually identical, and they both underestimate the real density near the mean.

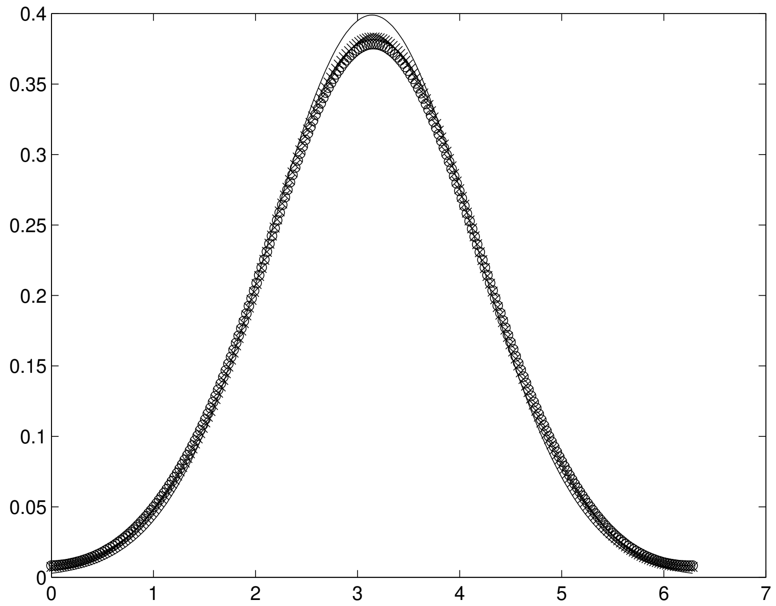

Figure 1: *Comparison of the cosine and kernel estimators for a normal density with mean π and standard deviation 1. The solid line represents the normal density. The points marked by x represent the kernel estimate with $h = 0.55$. The points marked by o represent the cosine estimate with $m = 15$. The estimates are based on the mean of 50 trials with samples of size $n = 250$ each.*

In the second experiment we compare the MISE versus different values of m and n for the cosine method itself, where the distribution is again normal with mean π and standard deviation 1. The resulting curves are given in Figure 2. In Figure 3 the same comparison is made for the derivative of the density function. In each case the MISE was determined by numerical integration using 250 points. In the determination of the MISE, trials were performed with 7 sets of samples each time, and the mean is reported in the figures given.

As with most estimators, the estimation of bimodal distributions is more difficult. We performed

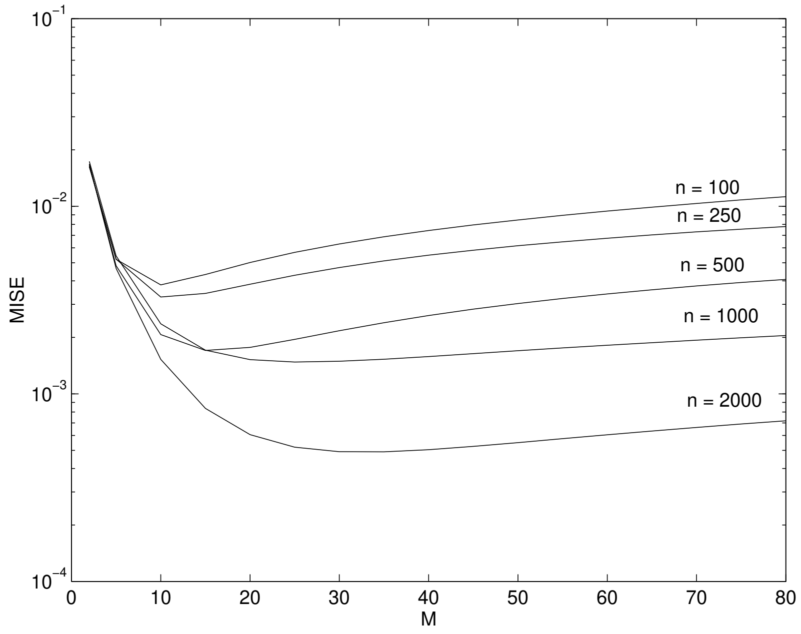

Figure 2: *MISE versus m and n for the normal density with mean π and standard deviation* 1.

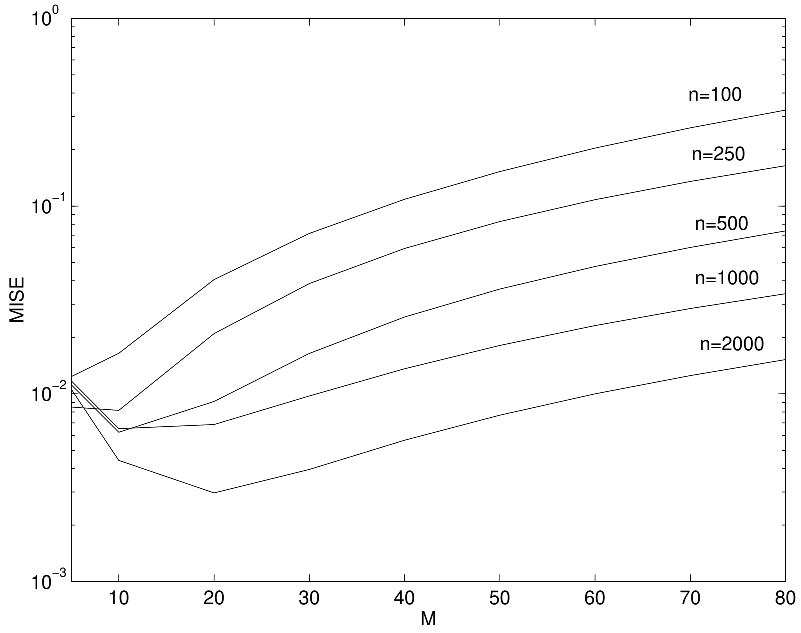

Figure 3: *MISE versus m and n for the derivative of the normal density with mean π and standard deviation* 1.

experiments with a sum of β distributions in the form

$$0.35\,\beta(6,2) + 0.65\,\beta(2,6)$$

generated by the rejection-acceptance method. The β function was scaled to take values in $[0, 2\pi]$. The results are given in Figure 4. It is evident that compared with a unimodal distribution, more samples are required to get accurate results.

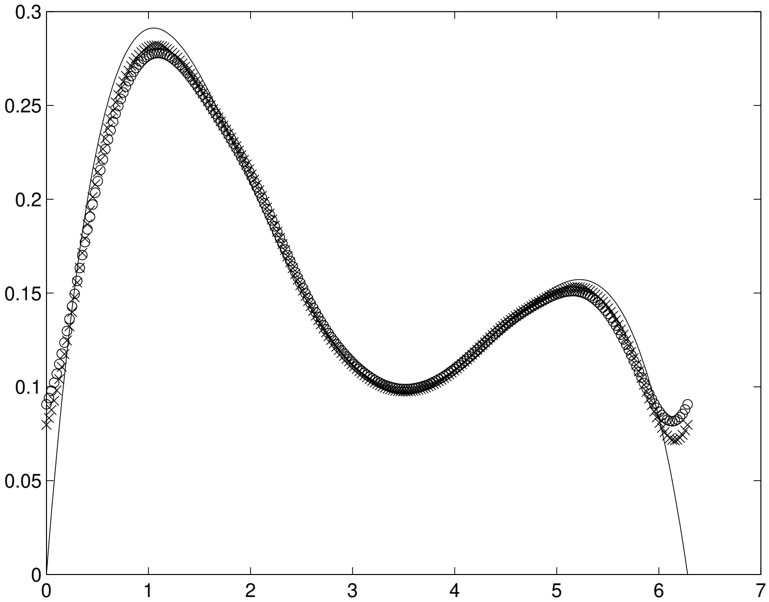

Figure 4: *Comparison of cosine and kernel estimators for a sum of* β *distributions on* $[0, 2\pi]$. *The solid line represents the true density. The points marked by* × *represent the kernel estimate with* $h = 0.4$. *The points marked by* o *represent the cosine estimate with* $m = 25$. *The estimates are based on the mean of* 50 *trials with samples of size* 500 *each.*

In iterative calculations, one can choose the sample size to obtain the desired accuracy in the results. In order to compare the computational efficiency of the two methods, a reasonable criterion appears to be to compare the time requirements of the two methods for the same error. The sample sizes may not necessarily be the same in the two cases to obtain the same accuracy. Figures 5 and 6 compare the computational effort required for the cosine weight function estimator and the kernel estimator for the normal density in this sense. The data was obtained by the following procedure: Estimates were computed for various values of n, m, and h. The time required (in seconds) for the computation, together with the resulting MISE were determined. The lower envelope of the data was chosen as the representative curve for that particular estimator. Note that the cosine estimator performs substantially better than the kernel estimator for the cases tested.

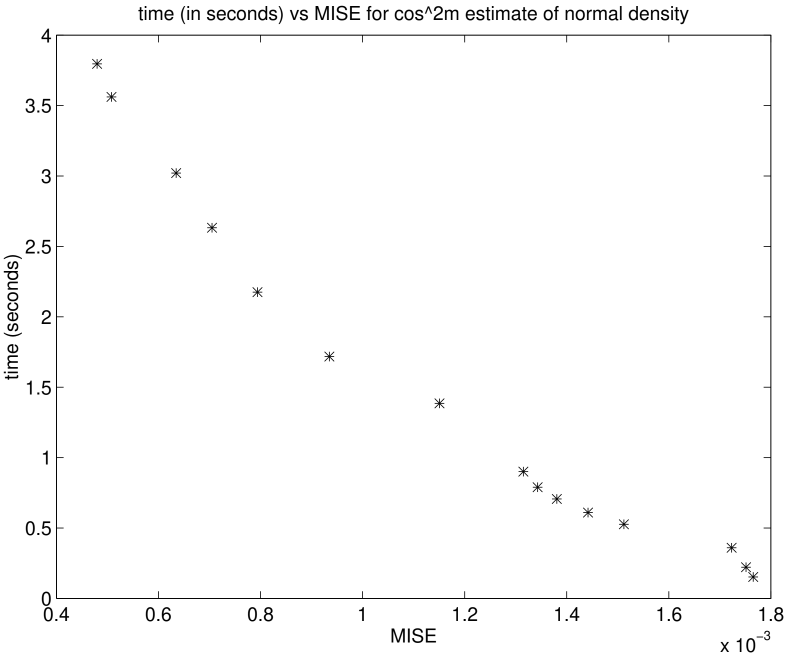

Figure 5: *Plot of time (in seconds) versus the MISE for the cosine estimation of the normal density.*

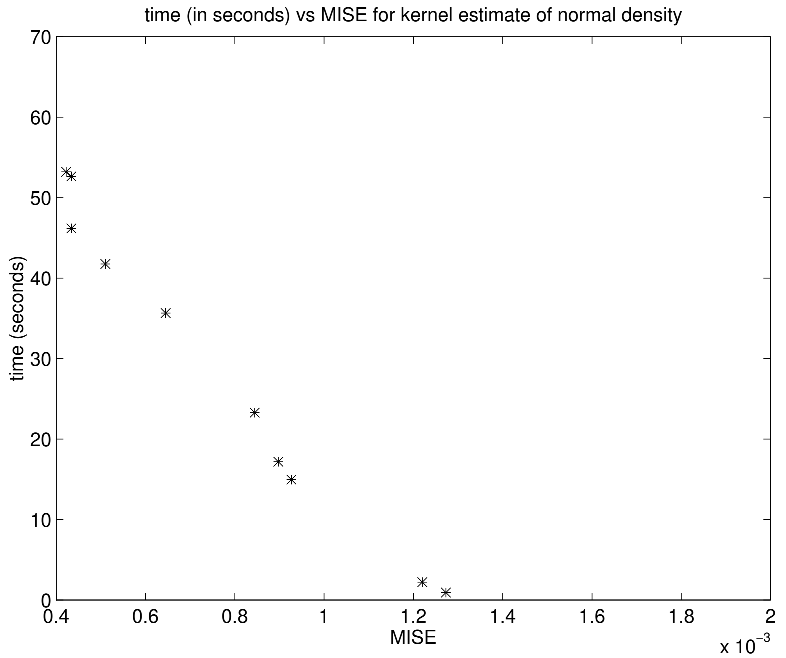

Figure 6: *Plot of time (in seconds) versus the MISE for the kernel estimation of the normal density.*

6 REMARKS

The experiments show that the optimal values of m are sufficiently small for the distributions tested to make the proposed scheme faster than the kernel estimate. However, as the distribution function gets more complicated (multimodal, for instance), we need higher values of m to get good estimates. Therefore such cases would require more computational effort than those with a more uniform distribution. If derivatives of the density are also required, then we are forced to increase m at a slower rate with respect to n to avoid sharp peaks, but then the cosine method becomes even more advantageous.

Plots of MISE versus m and n follow the expected trends. As the sample size increases, the error at the optimal value of m decreases. Besides, the optimal value of m increases as the sample size increases, and it is smaller for the derivative than for the density estimation itself. It can also be seen that as the number of points increases, the range of m over which the estimate performs well also increases.

Based on the experiments, we can perform a least squares analysis and approximate the magnitude of m as $kn^{1/2.5}$ for 1-D density estimation, and as $kn^{1/(3+\delta)}$ for the derivative, for some constant k. δ can be any small number and is introduced in order to satisfy the upper bound for derivatives required for convergence. Choosing $m = kn^{1/3}$ appears to give reasonable estimates for density on S^2. The magnitude of k depends on the complexity of the density function itself, and it varies between 1 and 10 for the distributions considered here. If a more automatic choice of m is desired, it should be possible to make k a function of the variation. This can be done by choosing an initial value for m and and estimating the variation using the approximate density currently available. Next density estimation computations can then be performed using the new value of m. We are currently studying schemes for the automatic determination of the best exponent.

We remark that computations for the algorithm presented here parallelize easily by a straightforward distribution of the samples across the processors. Each processor computes the coefficients $M(r,s)$ based on only the samples locally present. Subsequently the corresponding coefficients in each processor are summed. This can be accomplished by a global reduction operation, for which efficient library functions are normally provided in parallel computers. In order to reduce inter-processor communication while using kernel estimators, it is necessary to redistribute the samples so that points that are close to each other reside in the same processor. This overhead is avoided in the cosine based estimator.

References

[1] Bickel P.J. and M. Rosenblatt *On some global measures of the deviations of density function estimates*, Annals of Statistics, 1 (1973) pp. 1071–1095.

[2] Chentsov N.N. *Estimation of unknown probability density based on observations*, Dokl. Akad. Nauk SSSR, 147 (1962) pp. 45–48 (in Russian).

[3] Eğecioğlu Ö. and Srinivasan A. *Efficient Non-parametric Estimation of Probability Density Functions*, SIAM J. of Scientific Computing, (in review).

[4] Hernquist L. and Katz N. *TREESPH: A unification of SPH with the hierarchical tree method*, Astrophys. J. suppl., 70 (1989) pp. 419–446.

[5] Hwang J. *Non-parametric multivariate density estimation: A comparative study*, IEEE Trans. Signal Processing, 42 (1994) pp. 2795–2810.

[6] Knuth D.E. *The Art of Computer Programming, Vol. 2, Seminumerical Algorithms*, Addison-Wesley, Menlo Park, CA, 1968. pp. 41–52.

[7] Markushevich A.I. *Theory of Functions of a Complex Variable*, Chelsea Publishing Company, New York, 1985, Vol 2. pp. 26–34.

[8] Monaghan J.J. and Lattanzio J.C. *A refined particle method for astrophysical problems*, Astron. Astrophys. 149 (1985) pp. 135–143.

[9] Monaghan J.J. *Smoothed particle hydrodynamics*, Annu. Rev. Astron. Astrophys. 30 (1992) pp. 543–574.

[10] Nadaraya E.A. *Non-parametric Estimation of Probability Densities and Regression Curves*, Mathematics and Applications (Soviet Series)1, Kluwer Academic Publishers, Boston, 1989.

[11] Parzen E. *On estimation of a probability density function and mode*, Ann. Math. Statist., 33 (1962) pp. 1065–1076.

[12] Rosenblatt M. *Remarks on some non-parametric estimates of a density function*, Ann. Math. Statist., 27 (1956) pp. 832–837.

[13] Schwartz S.C. *Estimation of probability density by an orthogonal series*, Ann. Math. Statist., 38 (1967) pp. 1261–1265.

[14] Silverman B.W. *Kernel density estimation using the fast Fourier transform*, Appl. Statist., 31 (1982) pp. 93–99.

[15] Smirnov M.V *On the approximation of probability densities of random variables*, Scholarly Notes of Moscow State Polytechnical Institute, 16 (1951) pp. 69–96 (in Russian).

[16] Silverman B.W. *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1993.

[17] Whittle P. *On the smoothing of probability density functions*, J. Roy. Statist. Soc. B, 20 (1958) pp. 334–343.