Probability Density Estimation from Optimally Condensed Data Samples

Mark Girolami and Chao He

Authors are with the Applied Computational Intelligence Research Unit, School of Information and Communication Technologies, University of Paisley, High Street, Paisley, PA1 2BE, Scotland, UK. E-mail:{mark.girolami, chao.he}@paisley.ac.uk.

This work is supported by SHEFC RDG grant 'INCITE' `http://www.incite.org.uk`. Matlab implementation and example data sets available at `http://cis.paisley.ac.uk/giro-ci0/reddens`.

Abstract

The requirement to reduce the computational cost of evaluating a point probability density estimate when employing a Parzen window estimator is a well known problem. This paper presents the Reduced Set Density Estimator that provides a kernel based density estimator which employs a small percentage of the available data sample and is optimal in the L_2 sense. Whilst only requiring $\mathcal{O}(N^2)$ optimisation routines to estimate the required kernel weighting coefficients, the proposed method provides similar levels of performance accuracy and sparseness of representation as Support Vector Machine density estimation, which requires $\mathcal{O}(N^3)$ optimisation routines, and which has previously been shown to consistently outperform Gaussian Mixture Models. It is also demonstrated that the proposed density estimator consistently provides superior density estimates for similar levels of data reduction to that provided by the recently proposed Density Based Multiscale Data Condensation algorithm and in addition has comparable computational scaling. The additional advantage of the proposed method is that no extra free parameters are introduced such as regularisation, bin width or condensation ratios making this method a very simple and straightforward approach to providing a reduced set density estimator with comparable accuracy to that of the full sample Parzen density estimator.

I. INTRODUCTION

The estimation of the probability density function (PDF) of a continuous distribution from a representative sample drawn from the underlying density is a problem of fundamental importance to all aspects and applications of machine learning and pattern recognition, see for example [33], [29], [5].

When it is reasonable to assume a particular functional form for the PDF, perhaps due to *a priori* knowledge of the data generating process, then the problem reduces to the estimation of the parameters defining the density function and is often referred to as *Parametric* density estimation. As an example the Gaussian PDF is one of the two-parameter members of the exponential family of distributions [4], and so estimation of the sufficient statistics is all that is required to fully define the Gaussian density. In other words the full characteristics of the data distribution can be summarized by condensing the data sample into the estimated sufficient statistics, in the case of the multi-variate Gaussian the mean vector and covariance matrix [2].

There are very many cases where a single parametric form for the PDF to be estimated is inappropriate, for instance in the situation where there are a number of sub-populations within the population being characterized [18], [34]. In such a case the PDF may be comprised of a finite number of simple parametric forms which define a constrained mixture of parametric

PDF's, the constraint being that the mixture also defines and satisfies the conditions of a density function. Finite mixture models [18], also known as *Semi-Parametric* density estimators, are a very powerful approach to estimating arbitrary density functions and the specific case of the Mixture-of-Gaussians [5] is employed in many practical applications, for example in defining the emission probabilities of a hidden Markov model for speech recognition [23], or in devising an outlier detector [34], [24] amongst many other applications. *Semi-Parametric* density estimation [18] therefore also provides a condensed representation of the data sample in terms of the sufficient statistics of each of the mixture components and their respective mixing weights.

The *Semi-Parametric* approach to density estimation relaxes the number of explicit assumptions required on the form of the underlying density, however it is the *Non-Parametric* approach to density estimation that has the least number of assumptions imposed[1]. *Non-Parametric* density estimators make no assumption on the structural form of the PDF and examples are Histograms, K-nearest Neighbour, Orthogonal Series Basis Expansions, and kernel estimators, see for example [29] and [13] for an extensive review. The kernel density estimators, also known as Parzen windows [22], have been extensively studied and provide a useful *Non-Parametric* alternative to mixture-models. Indeed a Parzen window estimator can be viewed as the limiting form of a mixture model where the number of mixture components will equal the number of points in the data sample [18]. Unlike the other approaches to density estimation considered where only sufficient statistics are required in estimation, Parzen density estimates employ the full data sample in defining subsequent density estimates and whilst large sample sizes ensure reliable density estimates they also ensure a computational cost for testing which scales directly with the sample size. Herein lies the main practical difficulty with employing certain *Non-Parametric* methods such as Parzen window density estimators.

This paper considers the case where data scarcity is not an application constraint and that the continuous distributional characteristics of the data suggest the existence of a well formed density function which requires to be estimated. Such situations are quite the norm in the majority of practical applications such as continuous monitoring of the condition of a machine or biomedical process, computer vision e.g. [24], [6] - indeed the reverse 'problem' is often experienced

[1] Although the term *Non-Parametric* implies that there are no free parameters which require to be set this is not the case. The bin width in histograms, the value of K in K-nearest neighbours, the window width in Parzen windows, the series length in Orthogonal Series estimators all have to be selected according to some optimality criterion.

in many situations where there is an overwhelming amount of data logged [19]. In situations where the volume of data to be processed is large a semi-parametric mixture model can provide a condensed representation of the reference data sample, in the form of the estimated mixing coefficients and component sufficient statistics, for estimating the value of the density of further observed data. On the other hand the Parzen window density estimator requires the full reference set for testing [13] which in such practical circumstances can be prohibitively expensive for online testing purposes.

This paper addresses the above problem by providing a Parzen window density estimator which employs a reduced set of the available data sample. The proposed *Reduced Set Density Estimator* (RSDE) is optimal in the L_2 sense in that the integrated squared error between the true density and the reduced set density estimator is minimised in devising the estimator. The required optimisation turns out to be a straightforward quadratic optimisation with simple equality constraints and thus suitable forms of Multiplicative Updating [27] or Sequential Minimal Optimisation as introduced in [30] can be employed which ensures at most quadratic scaling in the original sample size. This is an improvement over the cubic scaling optimisation required of the Support Vector Method of density estimation proposed in [20]. The additional advantage of the proposed method is that, apart from the weighting coefficients, no additional free parameters are introduced into the representation such as regularisation terms [35], bin widths [26], [11], or number of nearest neighbours [19]. The RSDE is shown to have similar convergence rates as the Parzen window estimator and performs, in terms of accuracy, similarly to the SVM density estimator [20] whilst requiring a much less costly optimisation, and consistently outperforms the multiscale data condensation method [19] at specified data reduction rates when used for density estimation.

The following section provides a brief review of methods which have been proposed in reducing the computational cost of density estimation using a kernel (Parzen window) density estimator.

II. Computation Reduction Methods for Kernel Density Estimation

The Parzen window form of non-parametric probability density estimation [22] is particularly attractive when no *a priori* information is available to guide the choice of the form of parametric density function with which to fit the data. A probability density estimate $\hat{p}(\mathbf{x};\boldsymbol{\theta})$ can be obtained

from the finite data sample $\mathcal{S} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \in \mathcal{R}^d$ drawn from the density $p(\mathbf{x})$ by employing the isotropic product form of the univariate Parzen window density estimator [28], [13]

$$\hat{p}(\mathbf{x}; h) = \frac{1}{Nh^d} \sum_{n=1}^{N} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \tag{1}$$

where the well-known constraints on the window (also referred to as the weighting or kernel) function hold i.e. it should also be a density function, see [13] for a comprehensive review.

However, as already stated, the main disadvantage of such an approach is the high computational requirements when large data samples are available as the estimation of the density at one point is an order-N type problem.

Two distinct approaches to resolving this practical problem of computational load have been adopted. The first concentrates on providing an approximation to the kernel function which decouples the point under consideration from the points of the sample in such a way that the summation over the sample can be performed separately in a manner akin to orthogonal series density estimators [13]. The second approach focuses on reducing the required number of computations by reducing the effective size of the sample.

A. Approximate Kernel Decompositions

The notion of multi-pole expansions of potential functions is exploited in [16] to provide a reduced cost kernel[2] density estimator. In [16] it is noted that the summation in (1) can be decomposed and approximated by truncating the inner-product summation defining the kernel at M terms such that

$$\sum_{n=1}^{N} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) = \Phi(\mathbf{x}) \cdot \sum_{n=1}^{N} \Phi(\mathbf{x}_n) \approx \sum_{m=1}^{M} \Phi_m(\mathbf{x}) \sum_{n=1}^{N} \Phi_m(\mathbf{x}_n) \tag{2}$$

The terms $\sum_{n=1}^{N} \Phi_m(\mathbf{x}_n)$ can be pre-computed and stored so that a point density estimate will scale as $\mathcal{O}(M)$ rather than $\mathcal{O}(N)$ which clearly denotes a computational saving when $M << N$. However there is no longer any guarantee that point estimates will necessarily be positive using this approach, [13] discusses such truncated orthogonal series estimators in detail and [9] points out the relationship between such estimators and kernel principal component analysis [31].

[2]Both the terms 'kernel' and 'Parzen' will be used interchangeably in the text and will refer to the same form of non-parametric density estimator.

B. Data Reduction Methods

A number of approaches have been taken in reducing the effective number of computations required in giving a point estimate of the density. In [28] the Fourier transform is used to reduce the effective number of computations required, whilst in [26] the data sample is pre-binned and the kernel density estimator employs the bin centres as the 'sample' points which are each weighted by the normalised bin-counts. Somewhat recently the multivariate form of the binned kernel density estimator has been analysed in [11]. However, now the bin width and also possible binning strategies (equal width bins or variable spacing) have to be selected for each dimension in the multivariate case.

Rather than binning the sample data an alternative strategy is to cluster the sample and employ the cluster centres as the reduced data set. In [14] a clustering-based branch and bound approach is adopted, whilst in [3] clustering is employed in identifying a set of reference vectors to be employed in a Parzen-window classifier. In [12] the Self-Organising Map [15] is used to provide the reference vectors for the density estimators. The main detractor of employing clustering based data reduction methods is that a nonlinear optimization is required for the data partitioning and as such the solution is dependent on initial conditions, so the relative simplicity of the non-parametric density estimator is lost.

In [19] a data reduction method is proposed which employs hyper-discs of varying radii which are dependent on the density of the data in the region being considered. This provides a very elegant density dependent data reduction method, in other words a multi-scale approach to data reduction is employed so that larger numbers of points will be removed from regions of high density. This has the additional benefit that the algorithm is deterministic based on the value of the free parameter k the number of 'nearest neighbours' which determines the rate of data reduction. The value of k can of course be selected to minimize an error criterion between the estimate based on the reduced sample and the full sample, the algorithm has at most $\mathcal{O}(kN^2)$ scaling where N is the number of points in the full sample.

C. Data Reduction via Sparse Functional Approximations

In [8], [7] a computationally costly search based approach is adopted in approximating an entropic distance between the density estimate based on a subset of the available data sample

and that based on the full sample. Support vector regression [33] was originally proposed in [35] as a means of providing a sparse Parzen density estimator, i.e. many of the points in the sample are not used in the density estimate. The trade-off between sparsity and accuracy is controlled by the regularization term which requires to be selected in addition to the width of the kernel.

In [35] and [20], [21] the support vector approach to density estimation has been proposed as a means of solving the ill-posed linear operator problem $\int_{-\infty}^{x} p(t)dt = F(x)$ where $p(t)$ denotes the PDF and the distribution function at the point x is given as $F(x)$. The support vector density estimator $\hat{p}(\mathbf{x}) = \sum_{i=1}^{N} \beta_i \mathcal{K}_h(\mathbf{x}, \mathbf{x}_i)$ where $\mathcal{K}_h(\mathbf{x}, \mathbf{x}_i) \equiv \frac{1}{h^d}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$, can be considered as a generalisation of the Parzen density estimator where now each β_i act as the non-uniform weighting coefficients. The following constrained quadratic optimization is required to define the weighting coefficients [20].

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{K}\boldsymbol{\beta} \\ &\text{subject to } \ |\mathbf{f} - \mathbf{E}\boldsymbol{\beta}| \leq \boldsymbol{\epsilon}, \quad \text{and} \quad \boldsymbol{\beta}^{\mathrm{T}}\mathbf{1} = 1 \ , \ \beta_i \geq 0 \ \forall \ i \end{aligned} \tag{3}$$

Where $\mathbf{K}$ is the $N \times N$ matrix whose elements are all $\mathcal{K}_h(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{1}$ is the $N \times 1$ vector of ones, $\mathbf{f}$ is the $N \times 1$ vector whose i^{th} element $\hat{F}_N(\mathbf{x}_i)$ is the empirical distribution function of the random vector $\mathbf{x}_i$ computed as the product of the empirical distribution of each vector element. The $N \times N$ matrix $\mathbf{E}$ whose i, j^{th} element corresponds to $\prod_{k=1}^{d} \int_{-\infty}^{x_j^k} \mathcal{K}_h(x_i^k, t)dt$ and $\boldsymbol{\epsilon}$ the $N \times 1$ vector whose elements are all ϵ_N completes the definitions required for the above optimisation. The ϵ_N denotes the accuracy value of the Kolmogorv-Smirnov statistic (the absolute deviation between the empirical distribution function and the distribution function derived from the model) [20] which the solution is desired to achieve and this is used in selecting the bandwidth of the kernel [20]. The constraints required for this optimisation are dense and there is no dual form [33] which reduces the complexity of the constraints, as such the solution of (3) requires generic quadratic optimisation packages which typically scale as $\mathcal{O}(N^3)$.

The support vector approach to density estimation provides a sparse and therefore reduced computational cost when testing, it has also been shown to provide excellent results in testing [20], [21]. However, for large sample sizes, it is essential to obtain an optimisation which will have scaling better than $\mathcal{O}(N^3)$ as in [20], and does not require the setting of any additional free parameters which control the regularisation of the solution as in [35], [26]. The following section

presents the RSDE which enjoys at most $\mathcal{O}(N^2)$ scaling to estimate the weighting coefficients and only has one free parameter to set, the width of the kernel as in a standard Parzen estimator.

III. Reduced Set Density Estimator

Based on a reduced or condensed set of the reference data sample say $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_M\} \subset \mathcal{S}$ where $M << N$ a Parzen window density estimate can be obtained as

$$\begin{aligned} \hat{p}_R(\mathbf{x}; h) &= \frac{1}{M}\sum_{n=1}^{M} \mathcal{K}_h(\mathbf{x}, \mathbf{x}_n) = \sum_{n=1}^{N} \frac{1}{M} I_{\{\mathbf{x}_n \in \mathcal{D}\}} \mathcal{K}_h(\mathbf{x}, \mathbf{x}_n) \\ &\equiv \sum_{n=1}^{N} \gamma_n \mathcal{K}_h(\mathbf{x}, \mathbf{x}_n) \end{aligned}$$

such that $\sum_n \gamma_n = 1$, $\gamma_n = \frac{1}{M}\ \forall\ \mathbf{x}_n \in \mathcal{D}$ and $\gamma_n = 0\ \forall\ \mathbf{x}_n \notin \mathcal{D}$, where in this case each γ_n acts both as an indicator function and weighting term. Once the reduced data set has been selected (see [19] for a review of data condensation methods) a density estimator can be devised and for the Parzen method typically the bandwidth h is selected to minimise a distance measure between $\hat{p}_R(\mathbf{x}; h)$ and $\hat{p}(\mathbf{x}; h)$ for example the mean squared error. In [19] a multi-scale approach to data condensation is employed after which the reduced data set can be used in devising a non-parametric density estimator. If the ultimate goal of the data condensation is to provide a non-parametric density estimator with reduced testing cost, as opposed to a classifier for example, then we note that the weighting coefficient constraints required to ensure the satisfaction of the requirements for $\hat{p}_R(\mathbf{x}; h)$ to be a bona fide density function are $\sum_n \gamma_n = 1$ and $\gamma_n \geq 0\ \ \forall\ \ n$. However, it is not a necessary condition for every non-zero γ_n to give equal weighting $\frac{1}{M}$ to each selected point in the reduced set for $\hat{p}_R(\mathbf{x}; h)$ to be a density function. It is clear then that each of the γ_n will be a data dependent variable and as such the selector and weighting coefficients can be estimated based on the optimisation of an appropriate criterion. The maximum-likelihood criterion yields values for the weighting coefficients which will each be $\frac{1}{M}$, however penalised likelihood methods have been proposed to provide a smoother density estimator to alleviate the problem of overfitting to the sample [13], [29]. Rather than consider explicit penalised (regularised) likelihood this paper considers the L_2 criterion based on the Integrated Squared Error (ISE) between the true density and the estimate.

Consider then a sample of data $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ drawn from a distribution which has a density $p(\mathbf{x})$. The ISE is a measure of the global accuracy of a density estimate [13], [29] which con-

verges to the mean squared error asymptotically. For a reduced set density estimate from the data sample denoted as $\hat{p}_R(\mathbf{x};\boldsymbol{\theta}) = \sum_{n=1}^{N} \gamma_n \mathcal{K}_h(\mathbf{x}_n, \mathbf{x})$ where $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, h\}$ and $\mathcal{K}_h(\mathbf{x}_n, \mathbf{x})$ denotes a window function satisfying the requirements for a density [13], [29] with width h, the minimum of the ISE is as follows.

$$\begin{aligned} \min_{\boldsymbol{\theta}} I(\boldsymbol{\theta}) &= \min_{\boldsymbol{\theta}} \int_{\mathcal{R}^d} |p(\mathbf{x}) - \hat{p}_R(\mathbf{x};\boldsymbol{\theta})|^2 d\mathbf{x} \\ &= \min_{\boldsymbol{\theta}} \int_{\mathcal{R}^d} \hat{p}_R^2(\mathbf{x};\boldsymbol{\theta}) d\mathbf{x} - 2E_{p(\mathbf{x})}\{\hat{p}_R(\mathbf{x};\boldsymbol{\theta})\} \end{aligned} \tag{4}$$

Where the term $\int_{\mathcal{R}^d} p^2(\mathbf{x}) d\mathbf{x}$ has been dropped from the above due to its independence of the $\boldsymbol{\theta}$ parameters and $E_{p(\mathbf{x})}\{\cdot\}$ denotes expectation with respect to $p(\mathbf{x})$. Now note that the minimum of the Kullback-Leibler (KL) divergence [5] between the true density and the estimate based on the reduced set is defined as

$$\begin{aligned} \min_{\boldsymbol{\theta}} E_{p(\mathbf{x})}\left\{ log\left(\frac{p(\mathbf{x})}{\hat{p}_R(\mathbf{x};\boldsymbol{\theta})}\right)\right\} &= \min_{\boldsymbol{\theta}} \left\{ -E_{p(\mathbf{x})}\{log(\hat{p}_R(\mathbf{x};\boldsymbol{\theta})\}\right\} \\ &\leq \min_{\boldsymbol{\theta}} \left\{ -log\left(E_{p(\mathbf{x})}\{\hat{p}_R(\mathbf{x};\boldsymbol{\theta})\}\right)\right\} \end{aligned}$$

So we see that minimisation of the ISE (4) is equivalent to minimising an upper-bound on the KL divergence between the reduced set density model and the true density whilst minimising the compactness (norm) of the functional form of the density estimate. The issue of determining the selected points in the reduced set along with their weightings is considered in the following section.

A. Estimation of Weighting Coefficients

As already mentioned equality of weighting of each window function is not a necessary condition for $\hat{p}_R(\mathbf{x};\boldsymbol{\theta})$ to be a valid density estimator provided the positivity and summation constraints on $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_N]^{\mathrm{T}}$ are satisfied. The empirical estimate of the expectation in the above expression $E_{p(\mathbf{x})}\{\hat{p}_R(\mathbf{x};\boldsymbol{\theta})\} \simeq N^{-1}\sum_{n=1}^{N} \hat{p}_R(\mathbf{x}_n;\boldsymbol{\theta})$ is employed[3] and so for a fixed

[3]Note that the expectation is taken over the true density and as such the empirical estimate will be conservative. Standard cross-validation is therefore employed in providing the estimate of $I_{cv}(h) = \frac{1}{N}\sum_{n=1}^{N} I^{-n}(h)$ to assess the appropriate kernel width, where $I^{-n}(h)$ denotes the estimated n^{th} component of ISE obtained from the solution found with the n^{th} sample point removed [25].

bandwidth window the optimisation of (4) over $\boldsymbol{\gamma}$ satisfying the requirements of a density function vis. $\sum_{n=1}^{N}\gamma_n = 1$ and $\gamma_n \geq 0 \ \ \forall \ \ n$, is

$$\begin{aligned} & \min_{\boldsymbol{\gamma}} \ \sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j \int_{\mathcal{R}^d} \mathcal{K}_h(\mathbf{x},\mathbf{x}_i)\mathcal{K}_h(\mathbf{x},\mathbf{x}_j)d\mathbf{x} - \frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\mathcal{K}_h(\mathbf{x}_i,\mathbf{x}_j) \\ = & \min_{\boldsymbol{\gamma}} \ \sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j\mathcal{C}(\mathbf{x}_i,\mathbf{x}_j) - 2\sum_{i=1}^{N}\gamma_i\frac{1}{N}\sum_{j=1}^{N}\mathcal{K}_h(\mathbf{x}_i,\mathbf{x}_j) \\ = & \min_{\boldsymbol{\gamma}} \ \sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j\mathcal{C}(\mathbf{x}_i,\mathbf{x}_j) - 2\sum_{i=1}^{N}\gamma_i\hat{p}_h(\mathbf{x}_i) \end{aligned}$$

Due to the summation and positivity constraints on the weighting coefficients many of the γ terms associated with points having low density estimate $\hat{p}_h(\mathbf{x})$ will be set to zero in the above optimisation, thus effectively selecting a reduced set from high density regions in the data sample. So the minimisation of the ISE of the reduced set density estimator can be written as a constrained quadratic optimisation which in familiar matrix form is

$$\begin{aligned} & \min_{\boldsymbol{\gamma}} \ \frac{1}{2}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{C}\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{p} \\ & \texttt{subject to } \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{1} = 1 \ \texttt{ and } \ \gamma_i \geq 0 \ \ \forall \ \ i \end{aligned} \tag{5}$$

Where the $N \times N$ matrices with elements $\mathcal{C}(\mathbf{x}_i,\mathbf{x}_j) = \int_{\mathcal{R}^d}\mathcal{K}_h(\mathbf{x},\mathbf{x}_i)\mathcal{K}_h(\mathbf{x},\mathbf{x}_j)d\mathbf{x}$ and $\mathcal{K}_h(\mathbf{x}_i,\mathbf{x}_j)$ are defined as $\mathbf{C}$ and $\mathbf{K}$ respectively. The $N \times 1$ vector of Parzen density estimates of each point in the sample $\hat{p}_h(\mathbf{x}_i) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{K}_h(\mathbf{x}_i,\mathbf{x}_j)$ is defined as $\mathbf{p} = \mathbf{K}\mathbf{1}_N$, where $\mathbf{1}_N$ is the $N \times 1$ vector whose elements are all $\frac{1}{N}$.

As one specific example[4] we can employ an isotropic Gaussian window at a point $\mathbf{x}$ with common width (variance) h and centre $\mathbf{x}_i$ denoted as $\mathcal{G}_h(\mathbf{x},\mathbf{x}_i)$ then the individual terms of the matrices $\mathbf{K}$ and $\mathbf{C}$ have the specific form of $\mathcal{K}_h(\mathbf{x}_i,\mathbf{x}_j) = \mathcal{G}_h(\mathbf{x}_i,\mathbf{x}_j)$ and $\mathcal{C}(\mathbf{x}_i,\mathbf{x}_j) = \int_{\mathcal{R}^d}\mathcal{G}_h(\mathbf{x},\mathbf{x}_i)\mathcal{G}_h(\mathbf{x},\mathbf{x}_j)d\mathbf{x} = \mathcal{G}_{2h}(\mathbf{x}_i,\mathbf{x}_j)$ and so (5) can be written simply as

$$\min_{\boldsymbol{\gamma}} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j\mathcal{G}_{2h}(\mathbf{x}_i,\mathbf{x}_j) - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\mathcal{G}_h(\mathbf{x}_i,\mathbf{x}_j) \tag{6}$$

Note that the only free parameter (apart from the weighting coefficients) which requires to be set is the window width, there are no regularisation or additional parameters which require to

[4] Other kernels, such as the finite-support Bartlett-Epanenchnikov kernel, can easily be integrated over the range of the sample to obtain the $\mathcal{C}(\cdot,\cdot)$ terms.

be determined. In addition the constraints on the optimisation are simpler than those required for the SVM density estimator (3) thus enabling a possibly faster means of optimisation. Unlike the binned Parzen density estimator [26] or the data condensation approach [19] the problematic choice of bin width (binning strategy), or effective disc width selection is not required. Examining the form of (6) an intuitive insight into how the data reduction mechanism operates can be obtained. The minimum value of ISE will be penalised by contributions of large inter-point distances in the window function $\mathcal{G}_h(\cdot,\cdot)$ so the empirical expected value of the right-hand term will be maximised by selecting a small number of points (due to the summation constraint) in regions of high-density (low average inter-point distance). The left-hand term alone will cause the selection of points with high inter-point distances, as defined by the metric associated with the left-hand convolution operator, therefore the overall effect will be that points in regions of high-density (as defined by the specific width of the window function) will be selected to provide a smoothed density estimate.

B. Optimisation

As the quadratic programme specified by (5) only has simple positivity and equality constraints then a number of alternative optimisation strategies are now available. A standard trick of introducing a dummy variable and applying the soft-max [5] function such that $\gamma_i = \frac{exp(\alpha_i)}{\sum_{n=1}^{N} exp(\alpha_n)}$ converts the required constrained quadratic optimisation (5) to an unconstrained nonlinear optimisation over the dummy variables and conjugate gradients [5] provide a linear $\mathcal{O}(N)$ scaling optimisation. However, moving from a linear to nonlinear optimisation is not particularly appealing due to the inherent initialisation dependent variability of the solutions. Somewhat recently multiplicative updating methods for non-negative quadratic programming have been proposed in [27] and it is straightforward to employ these for the optimisation of (5). However, in terms of speed of convergence it has been found in our experiments that a form of the Sequential Minimal Optimisation (SMO) as presented in [30] suitable for solving (5) is superior to multiplicative updating. As detailed [30] this can achieve $\mathcal{O}(N^2)$ scaling as opposed to $\mathcal{O}(N^3)$ scaling achievable for the standard quadratic optimisation packages. In the following experiments the variant of SMO for (5) is employed and this is detailed in the appendix. A `MATLAB` implementation of RSDE as well as the data sets employed in the reported experiments is available at the following website `http://cis.paisley.ac.uk/giro-`

`ci0/reddens`. So the above optimisation (5), in the case of a Gaussian window, will provide a non-parametric estimate of the data density based on a subset of the original data sample defined as $\hat{p}_R(\mathbf{x}) = \sum_{\gamma_n \neq 0} \gamma_n \mathcal{G}_h(\mathbf{x}, \mathbf{x}_n)$. A number of experiments are now provided to demonstrate the proposed RSDE method.

IV. EXPERIMENTS

A. One and Two Dimensional Examples

The first demonstration of the RSDE employs a 1-D data sample which is drawn from a heavily skewed distribution defined as $p(x) = \frac{1}{8}\sum_{i=1}^{8} \mathcal{G}_{h_i}(\mu_i, x)$ where $h_i = \left(\frac{2}{3}\right)^i$ and $\mu_i = 3\left(h_i - 1\right)$ [25]. A sample of 200 points is drawn from the distribution and a Parzen window density estimator employing a Gaussian kernel is devised using the data sample. The width of the kernel is found by leave-one-out cross validation. A further sample of 10,000 data points are then drawn from the density and the L_2 error between the Parzen estimate and true density is computed, this procedure is repeated 200 times. The error was found to be (median value & interquartile range) 0.0033 & 0.0033. Figure 1. shows the true density and the estimated density for a particular sample realisation along with the individual kernel functions placed at the sample points[5].

The RSDE is applied to this data using, as above, a Gaussian kernel and the width of the kernel is also set by cross-validation. However it was noted in the reported experiments that measuring the cross-entropy[5] between the RSDE and the existing Parzen estimator then selecting the width value which returns the minimal cross-entropy is found to give similar results to cross-validation whilst reducing the effective number of optimisation runs (time taken) required for width selection. From the two hundred samples the median value for the number of non-zero weighting coefficients was 13 - amounting to less than 8% of the original sample - the minimum and maximum values of non-zero weighting coefficient was 5 and 42 respectively. The corresponding L_2 error based on 10,000 data points for 200 sample realisations was measured to be 0.0035 & 0.0030. Due to the highly asymmetric nature of the distribution of errors a Rank sum Wilcoxon test [17] is applied and shows that both error distributions for the full Parzen and RSDE estimators, at the 5% significance level, are identical. This is a somewhat satisfying result

[5]Every fifth data point is used in the figure for the purposes of clarity.

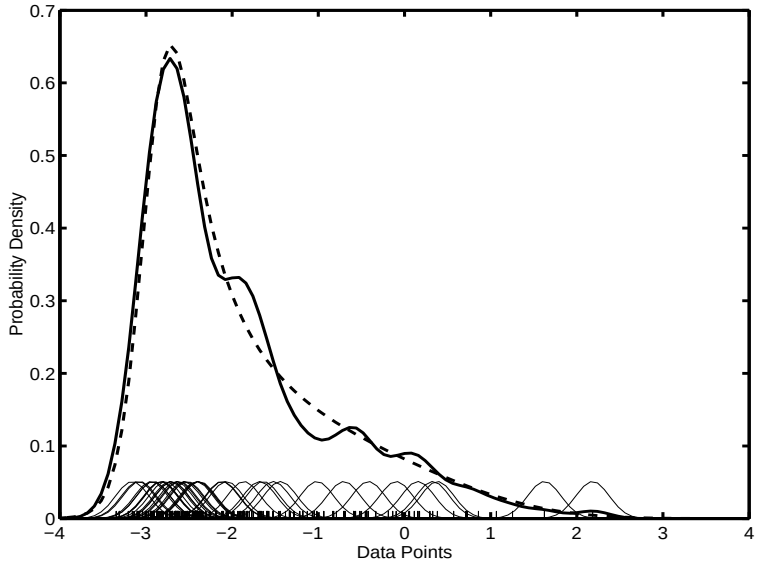

Fig. 1. The true density (dashed line) and the Parzen window estimate (solid line), each of the kernel functions is placed at the appropriate sample data point.

in that the accuracy of the RSDE is shown to be the same as the Parzen for this particular density function. The resulting estimate is shown in Figure 2. Notice that both methods estimate the mode well and the ripples in the tail, which are characteristic of finite sample Parzen estimates of long tailed behaviour, can be seen to be somewhat smoothed by the RSDE.

The second demonstration is primarily illustrative and employs a sample (200 points) of 2-D data which is generated with equal probability from an isotropic Gaussian and two Gaussians with both positive and negative correlation structure. The probability density is estimated using a Parzen window employing a Gaussian kernel and leave-one-out cross-validation was employed in selecting the kernel bandwidth. The probability density iso-contours, along with the data sample, is shown in the left-hand plot of Figure 3. By way of a comparison the multi-scale density based data condensation method of [19] is applied to this toy example and the results are shown in the right-hand plot of Figure 3. A similar level of data reduction to that of RSDE is achieved, where large circles denote identified regions of low density with smaller ones defining regions of high density. The selected data points are encircled. As a means of data condensation with the specific aim of non-parametric density estimation the multi-scale approach [19] has been shown to consistently outperform the data reduction methods proposed by Fukunaga and Mantock [8] and Astrahan [1].

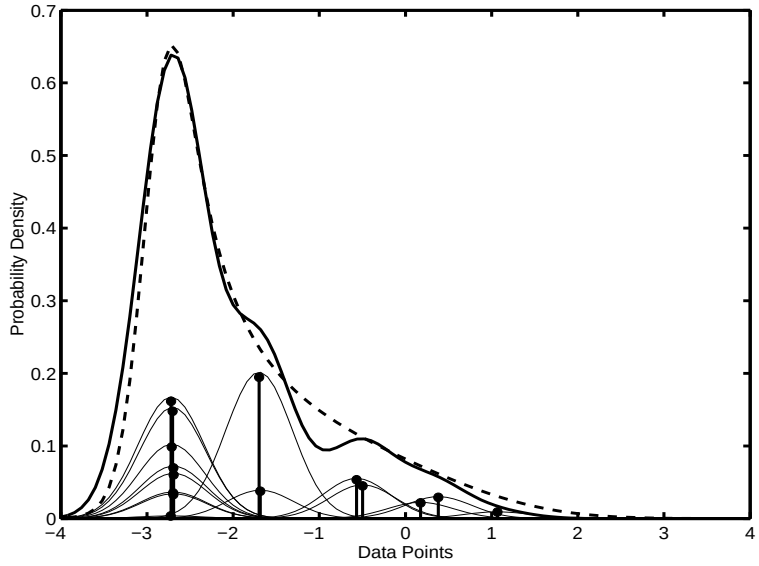

Fig. 2. The true density (dashed line) and the RSDE (solid line), each of the non-zero kernel functions is placed at the appropriate sample data point and the length of the vertical line denotes the value of the corresponding weighting coefficient.

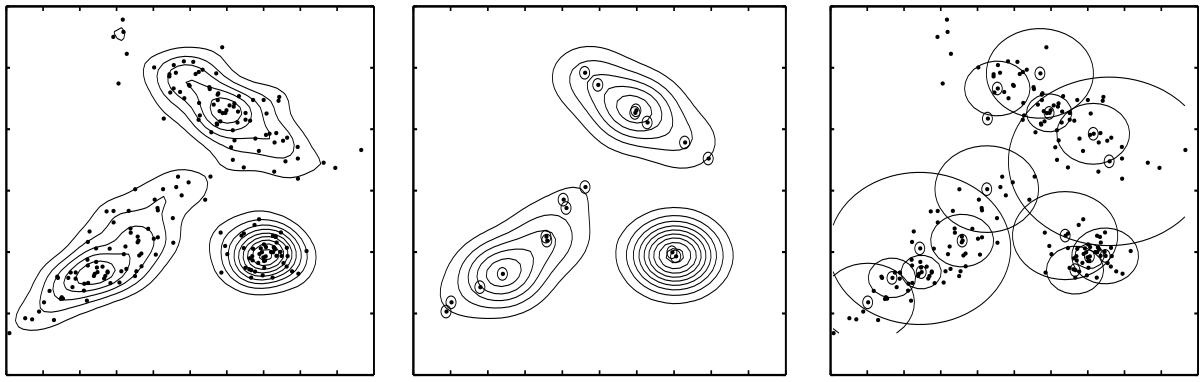

Fig. 3. The left hand plot shows the Parzen window density estimate. The middle plot shows the RSDE with the retained points circled. The right hand plot shows the results of the multiscale data condensation method where the selected points are encircled and the corresponding discs are shown.

The RSDE is obtained by optimising equation (5) and employing a Gaussian kernel in this case. As before the kernel bandwidth is selected by minimising the cross-entropy between the Parzen window estimate and the RSDE. The central plot in Figure 3. shows the corresponding iso-contours along with the reduced data set, denoted by the encircled points, which amounts to a 91% reduction in the number of points required to estimate the density of further data points. It is interesting to note that the selected points (non-zero weighting) occur in the regions of highest density of the sample, and indeed lie approximately on the principal axis of the two elongated

Gaussians. To illustrate this further 3000 data points from the 2-D *S*-shaped distribution[6] are used to estimate the associated PDF. The left hand plot in Figure 4. shows the data sample and the iso-contours of the Parzen density estimate. The right hand plot shows the density iso-contours obtained using RSDE and the selected points (12% of the original sample) are encircled as in the previous example. The selected points lie in the centre of the distribution and the shape they form is somewhat reminiscent of that obtained by Principal Curves [10]. This similarity may form an interesting area of future investigation. This observation is in contrast to the support vector data description methods [32], [30] where the boundary points of the sample tend to be selected.

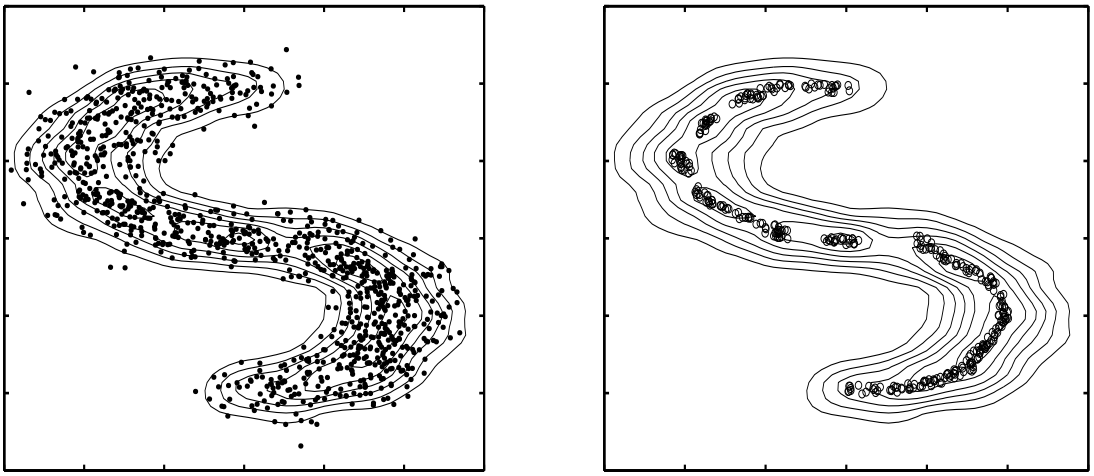

Fig. 4. The left hand plot shows the data sample and the Parzen window density estimate contours. The right hand plot shows the RSDE with the retained points circled.

B. Comparative Experiments

The first experiment in this section compares the RSDE with the SVM approach to density estimation [20]. The 1-D density function employed in [20] is used in this experiment i.e. $p(x) = \frac{1}{2\sqrt{2\pi}} exp(-0.5|x-2|^2) + \frac{0.7}{4} exp(-|x+2|)$. This density is a particularly useful test as it possesses both bi-modality and long tailed behaviour in one of the modes. As in [20] samples of 100 points are drawn from the density and then the SVM, RSDE, and Parzen density estimators are devised, a further 10,000 samples are then drawn from the PDF and used to compute, in

[6]This data set is used to demonstrate the use of Principal Curves [10] and is available at `http://www.iro.umontreal.ca/ kegl/research/pcurves/`

this case as in [20], the L_1 error, the integrated absolute deviation of the estimate from the true density value. This procedure was then repeated 1000 times to assess the bias and variance associated with each of the estimators. The free parameter (kernel width and ϵ) values reported in [20] for the SVM estimator were employed throughout, whilst leave-one-out cross-validation was used to set the Gaussian width for the Parzen window, and minimum cross-entropy between the Parzen and RSDE was used to set the kernel width for the RSDE. The results are shown in Figure 5.

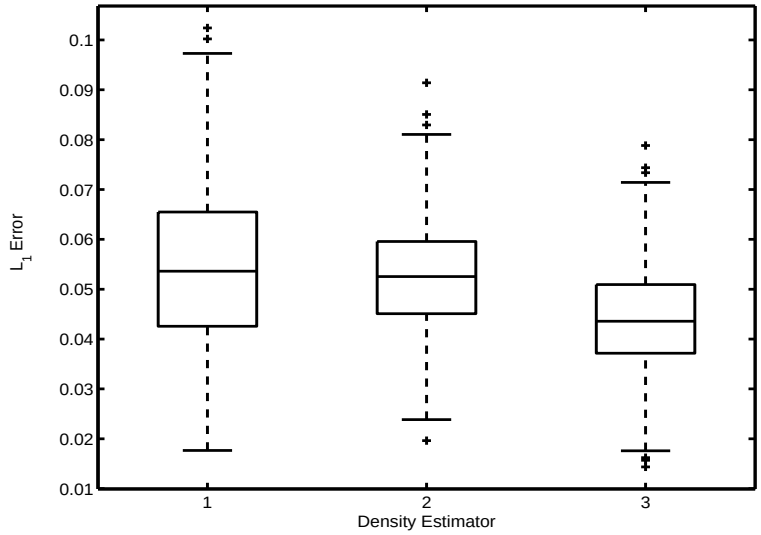

Fig. 5. Boxplot of the L_1 error for the SVM (1), RSDE (2), and Parzen (3) density estimators.

The box shows the quartiles of the distribution of error values whilst the whiskers show the range of the error values and the points beyond the whiskers fall outwith 1.5 times the inter-quantile range. It is interesting to note that both the RSDE and SVM estimators introduce an equally small amount of difference from the Parzen estimator, though the variability of the SVM estimator is slightly larger in this case. However, the RSDE can take advantage of the less computationally costly SMO routine in estimating the weighting coefficients. Figure 6. shows the number of non-zero values for both the SVM and RSDE estimators. Both have the same median value of 4 whilst the RSDE shows greater variability in the number of non-zero coefficients, primarily due to the kernel width value varying based on each sample in RSDE whilst the value of ϵ in the SVM approach stayed fixed for each sample.

To further test RSDE varying sizes of sample are drawn from both uni-modal and bi-modal distributions at different dimensionalities and the accuracy of the RSDE is compared with the

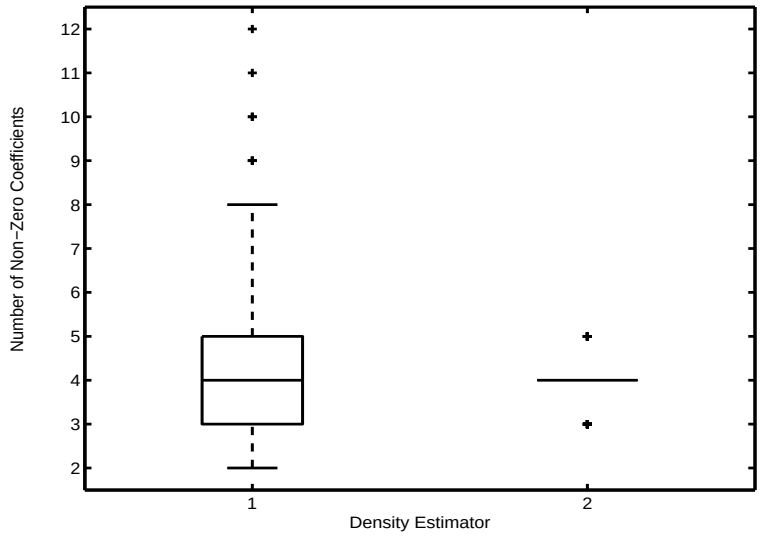

Fig. 6. Boxplot of the number of non-zero weighting coefficients for RSDE (1) and the SVM (2) density estimators.

Parzen density estimator. By way of further comparison with an alternative data reduction method the density based multiscale data condensation method [19] is employed to obtain a reduced size data sample from which to obtain a Parzen estimator with reduced computational complexity. This was chosen primarily due to the excellent results obtained in [19] with this method.

C. Multidimensional Unimodal Distribution

A multivariate (2-D & 5-D) Gaussian which is centered at the origin and has a covariance matrix such that $C_{ij} = 1$ where $i = j$ and $C_{ij} = 0.5$ where $i \neq j$ is used in this experiment. Samples of size 30 to 700 data points are drawn from the distribution and both a Parzen estimator and RSDE are fit to the data. A test sample of 10,000 points is then drawn and the L_2 error is computed this is then repeated 200 times for each sample size. For each sample size the average value of the level of sample size reduction achieved by the RSDE is then used to set the value of the number of nearest neighbours - k - the free parameter value in the multiscale condensation method of [19] which defines the associated condensation level. This reduced set is then used to devise a Parzen density estimate which is then tested alongside the full Parzen and the proposed RSDE.

The results are summarised in Figure 7. and Figure 8. where the L_2 error for the Parzen, RSDE and multiscale method is plotted against sample size. The points to note from Figure 7.

are that the rate of convergence of both the full Parzen estimators and RSDE are similar and that the variance of the estimators both decrease at the same rate with sample size. The levels of data reduction for the various sample sizes are given in Table I.

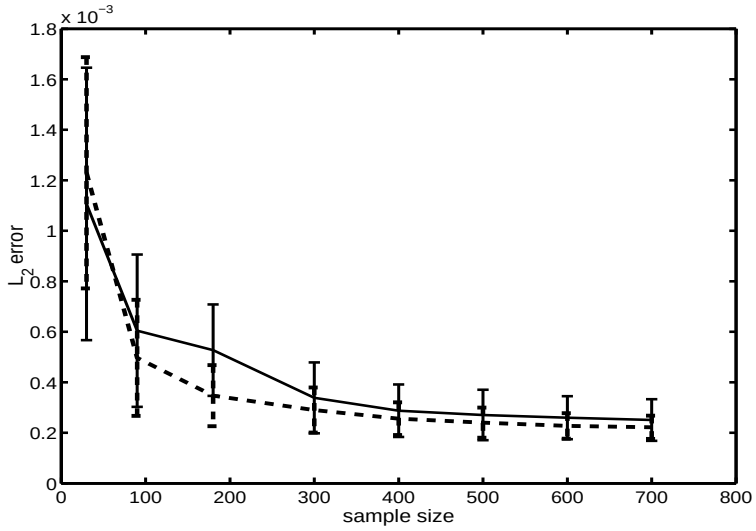

Fig. 7. L_2 error between the true density of a 2-D Gaussian and density estimators against sample size. The bars denote one standard deviation. Parzen window denoted by a solid line; and the RSDE denoted by the dashed line.

TABLE I

SS - SAMPLE SIZE, RD - REMAINING DATA (PERCENTAGE). RD DENOTES THE SIZE OF THE REDUCED SET AS A PERCENTAGE OF THE ORIGINAL SAMPLE (SS) DRAWN FROM A 2-D GAUSSIAN.

SS	30	90	180	300	400	500	600	700
RD	3.33%	4.44%	3.33%	2.33%	1.50%	1.40%	1.00%	1.29%

The levels of data reduction remain relatively constant at sample sizes of 400 and beyond with only on average 1.5% of the sample being used. These data reduction rates are then used to select the appropriate parameter value for the data condensation method of [19] in order to yield a similar level of data reduction. The accuracy of the Parzen estimator obtained by the multiscale data condensation method [19] is measured as above and is shown in Figure 8. It is clear that for similar levels of data reduction the RSDE provides an order of magnitude improvement in accuracy in terms of L_2 metric for this type of data.

The same experiment is conducted for data samples drawn from a similar 5-D Gaussian. The

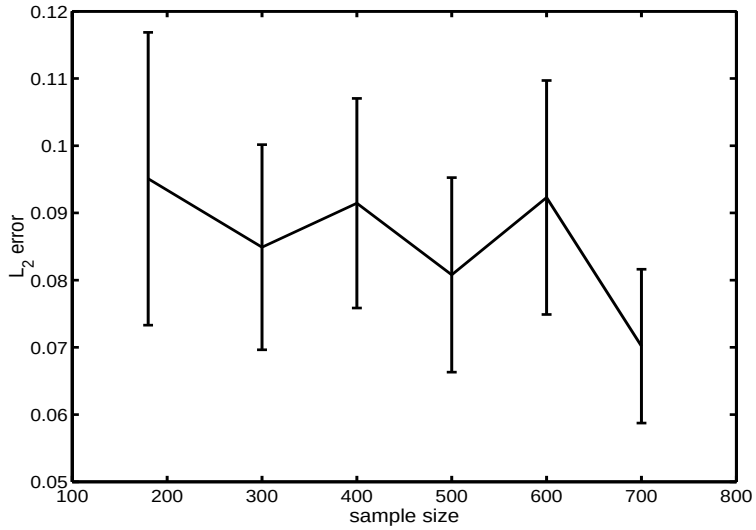

Fig. 8. L_2 error between true density of a 2-D Gaussian and Parzen density estimator based on condensed data set obtained from the multiscale data condensation method.

results are tabulated in Tables II and III.

TABLE II

L_2 ERROR (COMPUTED OVER 200 TRIALS) BETWEEN TRUE DENSITY OF 5-D GAUSSIAN AND RESPECTIVE DENSITY ESTIMATORS AGAINST SAMPLE SIZE.

	L_2 Error (Mean $\pm$ standard deviation)$\times 10^{-5}$		
Sample size	Parzen window	RSDE	Density condensation
30	(1.67$\pm$0.30)	(1.33$\pm$0.45)	0.12$\pm$0.67
90	(1.27$\pm$0.199)	(0.99$\pm$0.21)	0.12$\pm$0.47
180	(0.90$\pm$0.15)	(0.38$\pm$0.13)	(1.31$\pm$0.15)
300	(0.84$\pm$0.13)	(0.26$\pm$0.09)	(1.54$\pm$0.11)
400	(0.83$\pm$0.11)	(0.68$\pm$0.08)	(2.10$\pm$0.07)
500	(0.82$\pm$0.10)	(0.66$\pm$0.087)	(1.44$\pm$0.11)
600	(0.82$\pm$0.08)	(0.64$\pm$0.06)	(1.45$\pm$0.09)
700	(0.82$\pm$0.08)	(0.63$\pm$0.06)	(1.41$\pm$0.10)

From the table of results a similar trend in the accuracy of the estimate is observed as for the 2-D case. However, it can be seen that the level of data reduction is not so aggressive at the small sample sizes with 63% of the sample being retained for the small 30 point sample.

TABLE III

SS - SAMPLE SIZE, RD - REMAINING DATA (PERCENTAGE). RD DENOTES THE SIZE OF THE REDUCED SET AS A PERCENTAGE OF THE ORIGINAL SAMPLE (SS) DRAWN FROM A 5-D GAUSSIAN.

SS	30	90	180	300	400	500	600	700
RD	63.33%	8.89%	13.33%	15.33%	2.50%	3.00%	2.83%	2.00%

This is a nice example showing that the data reduction obtained is driven by the reduction of ISE. Clearly excessive reduction of the small sample size would result in large residual ISE due to the higher dimensionality of the data in this case. This is in contrast to the data reduction method of [19] where the data reduction is governed by the chosen value of k as such there is no automatic or implicit means of controlling the ensuing error in density estimate by the adoption of the method of [19].

D. Multidimensional Bimodal Distribution

The experiments in the previous section are now repeated for a bimodal distribution composed of two Gaussians centered at (1,1) and (-1,-1) with common covariance (1 0.5;0.5 1) in the 2-D case. In the 5-D case each Gaussian is centered at (1,1,1,1,1) and (-1,-1,-1,-1,-1) with common covariance as defined for the Gaussian of the previous section. The accuracy results for the 2-D case are given in Figure 9. and as with the unimodal density the RSDE has similar bias and variance to the Parzen density estimator, whilst the data condensation approach has a higher bias level for the same amount of data reduction.

The associated data reduction rates for the RSDE are given in Table IV

TABLE IV

SS - SAMPLE SIZE, RD - REMAINING DATA (PERCENTAGE). RD DENOTES THE SIZE OF THE REDUCED SET AS A PERCENTAGE OF THE ORIGINAL SAMPLE (SS) DRAWN FROM A 2-D MIXTURE OF TWO GAUSSIANS.

SS	30	90	180	300	400	500	600	700
RD	23.33%	6.67%	11.67%	8.67%	13.50%	13.00%	7.33%	8.57%

As with the previous section the same experiment is run on a 5-D example of the mixture of

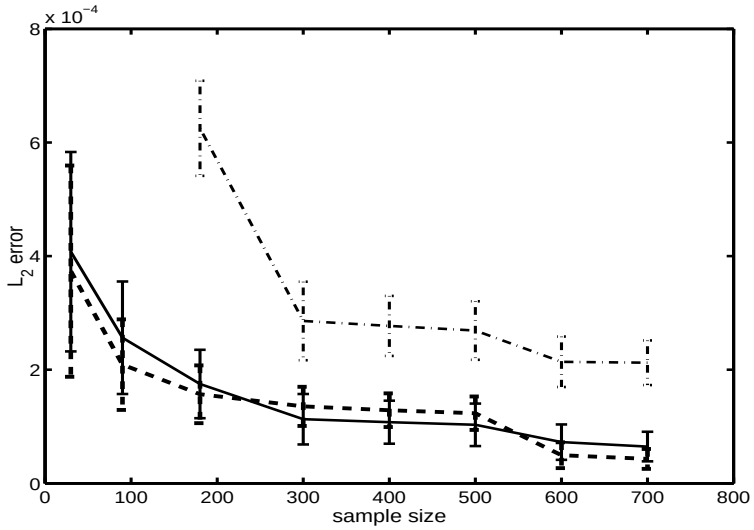

Fig. 9. L_2 error between true density of 2-D mixture of two Gaussians and the following density estimators - Solid line: Parzen window; Dashed line: RSDE; Dash-dot line: Density condensation, charted against sample size.

two Gaussians and the results are given in Tables V and VI.

TABLE V

L_2 ERROR (COMPUTED OVER 200 TRIALS) BETWEEN TRUE DENSITY OF 5-D MIXTURE OF TWO GAUSSIANS AND RESPECTIVE DENSITY ESTIMATORS AGAINST SAMPLE SIZE.

Sample size	L_2 Error (Mean $\pm$ standard deviation)$\times 10^{-6}$		
	Parzen window	RSDE	Density condensation
30	(2.87±0.50)	(2.56±0.84)	0.039±0.28
90	(2.23±0.32)	(1.34±0.40)	(3.44±0.33)
180	(2.22±0.25)	(1.86±0.26)	(3.63±0.23)
300	(2.22±0.20)	(1.75±0.20)	(3.83±0.15)
400	(2.21±0.17)	(1.71±0.18)	(3.54±0.16)
500	(2.19±0.15)	(1.66±0.15)	(3.85±0.13)
600	(1.63±0.18)	(1.64±0.15)	(4.13±0.09)
700	(1.63±0.15)	(1.63±0.13)	(3.49±0.14)

As in the case of unimodal data the bias of the RSDE follows that of the full sample Parzen estimator with the Parzen estimator based on the data condensation method showing a consistently larger bias. Again it can be seen that the reduction rate of the RSDE is governed by the

TABLE VI

SS - SAMPLE SIZE, RD - REMAINING DATA (PERCENTAGE). RD DENOTES THE SIZE OF THE REDUCED SET AS A PERCENTAGE OF THE ORIGINAL SAMPLE (SS) DRAWN FROM A 5-D MIXTURE OF TWO GAUSSIANS.

SS	30	90	180	300	400	500	600	700
RR	80.00%	22.22%	7.22%	4.33%	4.25%	4.20%	3.33%	2.29%

ISE and as such very small reductions are obtained for the 30 and 90 sample sizes (Table VI).

V. CONCLUSIONS AND DISCUSSION

The experiments reported have demonstrated that the RSDE provides very similar estimation accuracy as the Parzen window estimator whilst employing greatly reduced (as little as 1%) numbers of points from the available sample. The results also suggest that the RSDE can have a beneficial effect in terms of reducing the amount of overfitting experienced by the full Parzen estimator. The SVM approach to density estimation [20], [21], [35] sets out to solve the inverse linear operator problem and so estimates the empirical distribution function from the sample. The ϵ-insensitive loss employed [20], [21] provides the sparse representation of the density, however, from the perspective of practical implementation the dense nature of the constraints requires generic quadratic optimisation routines. One of the alternate SVM approaches proposed in [35] was to minimise the L_2 error between the SVM density estimate and a Parzen estimate whilst enforcing sparsity of representation by a suitable regularising term, which then introduces the added complexity of selecting the appropriate trade-off between sparsity and accuracy. The approach taken herein is fundamentally different in that the ISE between the true (unknown) density and the reduced set estimator is minimised. The sparsity of representation (data condensation) emerges naturally from direct minimisation of ISE due to the required constraints on the functional form of $\hat{p}_R(\mathbf{x})$ without the requirement to resort to additional sparsity inducing regularisation terms or employing L_1 or ϵ-insensitive losses [33], [21], [35]. As indicated in the text the minimisation of ISE based on an available sample can also be viewed as the minimisation of a regularised form of an upper-bound on the estimated KL-divergence between the RSDE and true density.

The Density Based Multiscale Data Condensation method [19] offers a straightforward means

of providing a sparse representation of a kernel density estimator once the parameter (k - number of nearest neighbours) which controls the rate of data condensation is set. It should be noted that this method returns a subset of the original data sample, the representation is multi-scale as regions of estimated high density have more points removed than regions of low density. This sample can then be employed, amongst other uses, in devising a density estimator. When a pre-defined data reduction ratio (that obtained by RSDE in the reported experiments) is employed to define the free parameter k the accuracy of the resulting Parzen density estimators have more bias than that obtained by RSDE.

One final point to note is that the reduced sample set returned by RSDE has a prototypical nature and this has been demonstrated on multivariate Gaussians where the selected points tend to lie on the principal axis of the distribution, and with isotropic Gaussians the points selected lie close to the distribution mean. Further, for an arbitrary non-Gaussian distribution the selected points tend to lie on what could be considered to be the principal curve of the distribution.

In summary this paper has presented a method that provides a kernel (Parzen) density estimator which employs a small subset of the available data sample based on the minimisation of the integrated square error between the estimator and the true density. Other than the weighting coefficients which can be obtained through straightforward quadratic optimisation, no additional free parameters e.g regularisation term, bin width or condensation ratio are introduced into the proposed estimator. Due to the simple constraints on the error criterion optimisation methods which have scaling of the order of $\mathcal{O}(N) \sim \mathcal{O}(N^2)$ can be employed. In testing it has been shown that the proposed density estimation method has similar convergence rates to the Parzen window estimator which employs the full data sample and has been shown to have comparable performance to the SVM density estimation method [20], [21].

It has also been shown to have improved performance over the density based multiscale data condensation method at predefined condensation rates. The proposed RSDE will find application in the many instances where a high-accuracy estimate of a PDF with low computational cost is required.

ACKNOWLEDGMENTS

The authors would like to acknowledge the helpful discussions with Dr Ewan MacArthur regarding this work.

REFERENCES

[1] Astrahan, M.M. (1970) *Speech Analysis by Clustering or the Hyperplane Method*, Stanford A.I. Project Memo, Stanford University, CA.

[2] Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*, Wiley.

[3] Babich, G.A. and Camps, O.I (1996) *Weighted Parzen Windows for Pattern Classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(5), pp 567 - 570.

[4] Barndorff-Nielson, O. (1978) *Information and Exponential Families in Statistical Theory*, Chichister: Wiley.

[5] Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Oxford University Press.

[6] Elgammal, E. Harwood, D. Davis, L. (2000) *Non-parametric Model for Background Subtraction*, 6th European Conference on Computer Vision, Lecture Notes in Computer Science 1843 Springer, ISBN 3-540-67686-4, pp 751 - 761.

[7] Fukunaga, K. and Hayes, R.R. (1989) *The Reduced Parzen Classifier*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(4), pp 423-425.

[8] Fukunaga, K. and Mantock, J.M. (1984) *Nonparametric Data Reduction*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6: pp 115-118.

[9] Girolami, M. (2002) *Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem*, Neural Computation, MIT Press, 14(3), pp 669 - 688.

[10] Hastie, T. and Stuetzle, W. (1989) *Principal Curves*, Journal of the American Statistical Association vol. 84, no. 406, pp. 502-516.

[11] Holmström, L. (2000) *The Error and the Computational Complexity of a Multivariate Binned Kernel Density Estimator*, Journal of Multivariate Analysis, 72(2):264-309.

[12] Holmström L. and Hämäläinen, A. (1993) *The Self-Organising Reduced Kernel Density Estimator*, IEEE International Conference on Neural Networks, Vol. 1, pp 417 - 421.

[13] Izenman, A.J. (1991) *Recent Developments in Nonparametric Density Estimation.* Journal of the American Statistical Association, **86**:205-224.

[14] Jeon, B. and Landgrebe, D.A. (1994) *Fast Parzen Density Estimation Using Clustering-Based Branch and Bound*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(9), pp 950-954.

[15] Kohonen, T. *Self-Organising Maps*, Springer-Verlag, 1995.

[16] Lambert, C. Harrington, S. Harvey, C. and Glodjo, A. (1999) *Efficient Online Nonparametric Kernel Density Estimation*, Algorithmica, 25, pp 37 - 57.

[17] Lehmann, E.L. (1975) *Nonparametric Statistical Methods Based on Ranks*, New York: McGraw-Hill.

[18] McLachlan, G. and Peel, D. (2000) *Finite Mixture Models*, Wiley.

[19] Mitra, P. Murthy, C.A. and Pal, S.K. (2002) *Density Based Multiscale Data Condensation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**: 6.

[20] Mukherjee, S. and Vapnik, V. (1999) *Support Vector Method for Multivariate Density Estimation*, CBCL Paper #170, AI Memo #1653.

[21] Mukherjee, S. and Vapnik, V. (2000) *Support Vector Method for Multivariate Density Estimation*, In Sara Solla, Todd Leen, Klaus-Robert Muller (eds), Advances in Neural Information Processing Systems, pp 659 - 665, MIT Press.

[22] Parzen, E. (1962) *On Estimation of a Probability Density Function and Mode*, Annals of Mathematical Statistics, 33:1065-1076.

[23] Rabiner, L.R. (1989) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the I.E.E.E, 77,(2), 257 - 286.

[24] Roberts, S. (2000) *Extreme Value Statistics for Novelty Detection in Biomedical Signal Processing*, IEE Proceedings Science, Technology & Measurement, 47:6, pp 363-367.

[25] Sain, S. (1994) *Adaptive Kernel Density Estimation*, PhD Thesis, Rice University.

[26] Scott, D.W. and Sheather, S.J. (1985) *Kernel density Estimation with Binned Data*, Communications in Statistics - Theory and Methods, Vol. 14, pp 1353 - 1359.

[27] Sha, F. Saul, L. and Lee, D.D. (2002) *Multiplicative Updates for Non-Negative Quadratic Programming in Support Vector Machines*, Technical Report MS-CIS-02-19, University of Pennsylvania.

[28] Silverman, B.W. (1982) *Kernel Density Estimation using the Fast Fourier Transform*, Applied Statistics, 31: pp 93-99.

[29] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

[30] Schölkopf, B. Platt, J. Shawe-Taylor, J. Smola, A and Williamson, R. (2001) *Estimating the Support of a High-Dimensional Distribution*, Neural Computation, 13, pp 1443-1471.

[31] Schölkopf, B. Smola. A. and Müller, K.R. (1998) Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, 10(5):1299-1219.

[32] Tax, D.M.J. and Duin, R.P.W. (1999) *Support Vector Data Description*, Pattern Recognition Letters, 20:(11-13), pp 1191-1199.

[33] Vapnik, V. N. (1998) *Statistical Learning Theory*, New York, John Wiley & Sons.

[34] Wang, S. Woodward, W.A. Gray, H.L. Wiechecki, S. and Sain, S.R. (1997) *A New Test for Outlier Detection from a Multivariate Mixture Distribution*, Journal of Computational and Graphical Statistics 6, pp 285-299.

[35] Weston, J. Gammerman, A. Stitson, M.O. Vapnick, V. Vovk, V. and Watkins, C. (1999) *Support Vector Density Estimation*, Advances in Kernel Methods, MIT Press.

VI. APPENDIX

A modified version of SMO (Sequential Minimal Optimisation) as developed in [30] is proposed specifically for the RSDE method. The updates for (5) are almost identical to those of [30] apart from the one additional term in (5) which requires to be incorporated. For completeness the derivation is included here and follows [30]. A MATLAB implementation is available at the website given in Section. III-B.

To fulfil the summation constraint, we resort to optimising over pairs of variables as in [30]. The SMO elementary optimisation step for optimising γ_1 and γ_2 with all other variables fixed follows. The quadratic optimisation problem

$$\min \ \frac{1}{2}\sum_{ij}\gamma_i\gamma_j Q_{ij} - \frac{1}{N}\sum_{ij}\gamma_i P_{ij}$$

can be written as

$$\min_{\gamma_1,\gamma_2} \frac{1}{2}\sum_{i,j=1}^{2}\gamma_i\gamma_j Q_{ij} + \sum_{i=1}^{2}\gamma_i C_i + C - \sum_{i=1}^{2}\gamma_i D_i - D \tag{7}$$

by defining

$$C_i = \sum_{j=3}^{N} \gamma_j Q_{ij};\ C = \frac{1}{2}\sum_{i,j=3}^{N} \gamma_i \gamma_j Q_{ij};\ D_i = \frac{1}{N}\sum_j P_{ij};\ D = \frac{1}{N}\sum_{i=3}^{N}\sum_j \gamma_i P_{ij}$$

subject to $\sum_{i=1}^{2} \gamma_i = \Delta, \quad \gamma_1, \gamma_2 \geq 0$, where $\Delta = 1 - \sum_{i=3}^{N} \gamma_i$. Following [30] we discard C and D in (7), which are independent of γ_1 and γ_2, and eliminate γ_1 to obtain

$$\begin{aligned} L &= \min_{\gamma_1,\gamma_2}\{\frac{1}{2}[(\Delta-\gamma_2)^2 Q_{11} + 2(\Delta-\gamma_2)\gamma_2 Q_{12} + {\gamma_2}^2 Q_{22}] \\ &\quad +(\Delta-\gamma_2)C_1 + \gamma_2 C_2 - (\Delta-\gamma_2)D_1 - \gamma_2 D_2\}. \end{aligned} \tag{8}$$

Setting the derivative of the above to zero and solving for γ_2 then

$$\gamma_2 = \frac{\Delta(Q_{11}-Q_{12}) + (C_1 - C_2) - (D_1 - D_2)}{Q_{11} - 2Q_{12} + Q_{22}}. \tag{9}$$

is obtained. γ_1 can be recovered from $\gamma_1 = \Delta - \gamma_2$.

Let γ_1^*, γ_2^* denote the parameter values before the step, and

$$K_i = Q_{1i}\gamma_1^* + Q_{2i}\gamma_2^* + C_i - D_i, \tag{10}$$

we can give the update equation for γ_2 as

$$\gamma_2 = \gamma_2^* + \frac{K_1 - K_2}{Q_{11} - 2Q_{12} + Q_{22}} \tag{11}$$

which does not explicitly depend on γ_1^*.

A. Overall optimisation procedure

1) Initialisation: The variables are initialized as $\gamma_i = p_i / \sum_i p_i$, where $p_i = \frac{1}{N}\sum_j P_{ij}$ is the Parzen window density estimate. This results in the points with higher density initially having larger γ values.

2) Optimisation algorithm: 1. Searching γ_2 and γ_1:

After initialisation the points with higher density will have larger γ values, we select the largest γ value in turn as the first variable γ_2 for the elementary optimisation step, and search for the second variable γ_1 which can generate the largest value of K_i. When γ_2 is less than a preset tolerance, stop the current searching loop, and go to check the terminating criterion.

2. Updating γ_2 and γ_1:

If γ_1 is greater than the preset tolerance, update γ_2. If the updated $\gamma_2 < 0$, set $\gamma_2 = 0$. Then, update γ_1 by $\Delta - \gamma_2$, if the updated $\gamma_1 < 0$, set $\gamma_1 = 0$ and $\gamma_2 = \Delta$.

3. Terminating criterion:

There are two criteria to terminate the algorithm.

(1)Comparing the value of the objective function with the same value obtained in the previous searching loop: if it decreases and the difference is greater than the preset error tolerance, then restart another searching loop, otherwise, recover the previous γ value and terminate the algorithm.

(2)If no variables are updated during a loop, terminate the algorithm.