Sparse Kernel Density Estimator Using Orthogonal Regression Based on D-Optimality Experimental Design

S. Chen, X. Hong and C.J. Harris

Abstract— **A novel sparse kernel density estimator is derived based on a regression approach, which selects a very small subset of significant kernels by means of the D-optimality experimental design criterion using an orthogonal forward selection procedure. The weights of the resulting sparse kernel model are calculated using the multiplicative nonnegative quadratic programming algorithm. The proposed method is computationally attractive, in comparison with many existing kernel density estimation algorithms. Our numerical results also show that the proposed method compares favourably with other existing methods, in terms of both test accuracy and model sparsity, for constructing kernel density estimates.**

I. Introduction

The problem of estimating probability density functions (PDFs) is of fundamental importance to machine learning and all fields of engineering, see for example [1], [2], [3], [4], [5], [6]. The non-parametric approach for estimating the PDF based on a realisation sample drawn from the underlying density [1], [2], [3] has attracted considerable interests, because it does not require to specify the functional form for the unknown underlying density distribution. The best-known non-parametric density estimation technique is perphase the classical Parzen window (PW) estimate [1], which is remarkably simple and accurate. However, the PW estimate, also known as the kernel density estimate, employs the full data sample set in defining density estimate for subsequent observation. Thus its computational cost for testing scales directly with the sample size. In today's data rich environment, this may become a practical difficulty in employing the PW estimator. It also motivates the research on the sparse kernel density (SKD) estimation techniques.

The support vector machine (SVM) method was applied to SKD estimation in [7], [8] and an interesting SKD estimation technique was proposed in [9]. These techniques employ the full data set as the kernel set and obtain a sparse representation by making as many kernel weights to (near) zero as possible based on some chosen criteria. A regression-based estimation method was reported in [10], which selects SKD estimates based on an orthogonal forward regression (OFR) algorithm that incrementally minimises the training mean square error (MSE). We proposed an OFR algorithm for SKD estimation based on the leave-one-out test MSE and regularisation [11]. Similar to the SVM-based density estimation [7], [8], the algorithms of [10], [11] select SKD estimates in the cumulative distribution function (CDF) space by converting the kernels into the associated CDFs and using the empirical distribution function (EDF) calculated on the training set as the desired respone. As a PDF estimate, the kernel weights must satisfy the nonnegative and unity constraints. In [10], [11], the unity constraint is met by normalising the kernel weight vector and the nonnegative constraint is ensured by adding a test to the OFR selection procedure at the cost of an increased complexity. Recently, we have developed an OFR algorithm [12] that selects SKD estimates in the original PDF space by adopting the PW estimate as the desired response and calculating the kernel weights of the selected kernel model using the multiplicative nonnegative quadratic programming (MNQP) algorithm [13].

Optimal experimental designs [14] have been used for data analysis to construct smooth model response surface based on the setting of the experimental variables under well controlled experimental conditions, where model adequacy is evaluated by design criteria that are statistical measures of goodness of experimental designs by virtue of design efficiency and experimental effort. For regression models, quantitatively model adequacy is measured as function of the eigenvalues of the design matrix, as these eigenvalues are linked to the covariance matrix of the least squares (LS) parameter estimate. There exist a variety of optimal experimental design criteria based on different aspects of experimental design [14], and the D-optimality criterion is most effective in optimising the parameter efficiency and model robustness via maximisation of the determinant of the design matrix. Optimal experimental designs have been adopted to construct sparse regression models based on an OFR procedure [15], [16], [17], [18]. These previous works have demonstrated the effectiveness of optimal experimental design methods in obtaining a robust and parsimonious model structure with unbiased model parameter estimate.

Motivated by the effectiveness of optimal experimental designs in constructing robust and sparse regression models, we propose a simple yet effective regression-based method for SKD estimation using the D-optimality criterion. Our proposed method first selects a very small subset of significant kernels from the full kernel set generated from the training data set. Note that the problem of kernel density estimation is enssentially an unsupervised learning problem and typically an ill-conditioned one. Our proposed OFR procedure based on the D-optimality is a computationally efficient unsupervised learning method and it is capable of yielding robust and accurate as well as sparse kernel model structure. Thus, this D-optimality based OFR method is well-

S. Chen and C.J. Harris are with School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. E-mails: {sqc, cjh}@ecs.soton.ac.uk

X. Hong is with School of Systems Engineering, University of Reading, Reading RG6 6AY, U.K. E-mail: x.hong@reading.ac.uk

978-1-4244-1821-3/08/$25.00

suited to the problem of kernel density estimation. After obtaining a very sparse kernel model structure, the associated kernel weights can readily be calculated using a modified version of the MNQP algorithm of [13]. Because the size of kernel model is very small, this MNQP algorithm requires little extra computational effort. Moreover, it further sets some kernel weights to zero, yielding an even sparser kernel density estimate. The proposed SKD estimation approach based on the combined D-optimality design and MNQP algorithm is computationally much more efficient than other existing regression-based methods. Several examples demonstrate that this proposed method compares favourably with other existing methods for constructing SKD estimates, both in terms of test accuracy and model sparsity.

II. Kernel Density Estimation via Regression

Let a finite data sample set $D_N = \{\mathbf{x}_k\}_{k=1}^N$ be drawn from a density $p(\mathbf{x})$, where the data sample $\mathbf{x}_k = [x_1 \ x_2 \cdots x_m]^T \in \mathcal{R}^m$. The task is to infer the unknown density $p(\mathbf{x})$ using the kernel density estimate of the form

$$\hat{p}(\mathbf{x}; \boldsymbol{\beta}_N, \rho) = \sum_{k=1}^{N} \beta_k K_\rho(\mathbf{x}, \mathbf{x}_k) \tag{1}$$

with the constraints

$$\beta_k \geq 0, \ 1 \leq k \leq N, \tag{2}$$

and

$$\boldsymbol{\beta}_N^T \mathbf{1}_N = 1, \tag{3}$$

where $\boldsymbol{\beta}_N = [\beta_1 \ \beta_2 \cdots \beta_N]^T$ is the kernel weight vector, $\mathbf{1}_N$ denotes the vector of ones with dimension N, and $K_\rho(\bullet, \bullet)$ is a chosen kernel function with the kernel width ρ. In this study, we use the Gaussian kernel of the form

$$K_\rho(\mathbf{x}, \mathbf{x}_k) = \frac{1}{(2\pi\rho^2)^{m/2}} e^{-\frac{\|\mathbf{x}-\mathbf{x}_k\|^2}{2\rho^2}}, \tag{4}$$

but many other types of kernel functions can also be used in the density estimate (1). The well-known PW estimate is obtained by setting all the elements of $\boldsymbol{\beta}_N$ to $\frac{1}{N}$. The optimal kernel width ρ is typically determined via cross validation.

The PW estimate in fact can be derived as the maximum likelihood estimator using the divergence-based criterion [19]. The negative cross-entropy or divergence between the true density $p(\mathbf{x})$ and the estimate $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_N, \rho)$ is defined as

$$\int_{\mathcal{R}^m} p(\mathbf{u}) \log \hat{p}(\mathbf{u}; \boldsymbol{\beta}_N, \rho) \, d\mathbf{u} \approx \frac{1}{N} \sum_{k=1}^{N} \log \hat{p}(\mathbf{x}_k; \boldsymbol{\beta}_N, \rho)$$

$$= \frac{1}{N} \sum_{k=1}^{N} \log \left(\sum_{n=1}^{N} \beta_n K_\rho(\mathbf{x}_k, \mathbf{x}_n) \right). \tag{5}$$

Minimising this divergence subject to the constraints (2) and (3) leads to $\beta_n = \frac{1}{N}$ for $1 \leq n \leq N$, i.e. the PW estimate. Because of this "optimality" property, we may view the PW estimate as the "observation" of the true density contaminated by some "observation noise", namely

$$\hat{p}(\mathbf{x}; \mathbf{1}_N/N, \rho_{\mathrm{Par}}) = p(\mathbf{x}) + \tilde{\epsilon}(\mathbf{x}), \tag{6}$$

where ρ_{Par} denotes the kernel width used for the PW estimate. Thus the generic kernel density estimation problem (1) can be viewed as the following regression problem with the PW estimate as the "desired response"

$$\hat{p}(\mathbf{x}; \mathbf{1}_N/N, \rho_{\mathrm{Par}}) = \sum_{k=1}^{N} \beta_k K_\rho(\mathbf{x}, \mathbf{x}_k) + \epsilon(\mathbf{x}) \tag{7}$$

subject to the constraints (2) and (3), where $\epsilon(\mathbf{x})$ is the modelling error at $\mathbf{x}$.

Most SKD estimation techniques [7], [8], [10], [11] reformulate the density estimation problem (1) into a regression one by using the EDF as the desired response and converting the kernels into CDFs, that is,

$$F_N(\mathbf{x}) = \sum_{k=1}^{N} \beta_k q_\rho(\mathbf{x}, \mathbf{x}_k) + \hat{\epsilon}(\mathbf{x}) \tag{8}$$

where the regressor $q_\rho(\mathbf{x}, \mathbf{x}_k)$ is defined by

$$q_\rho(\mathbf{x}, \mathbf{x}_k) = \int_{-\infty}^{\mathbf{x}} K_\rho(\mathbf{u}, \mathbf{x}_k) \, d\mathbf{u} \tag{9}$$

and the EDF $F_N(\mathbf{x})$ is defined by

$$F_N(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^{N} \prod_{j=1}^{m} \theta(x_j - x_{j,k}) \tag{10}$$

with

$$\theta(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0, \end{cases} \tag{11}$$

where $\mathbf{x}_k = [x_{1,k} \ x_{2,k} \cdots x_{m,k}]^T \in D_N$. Our regression-based approach is computationally simpler and it can use any type of kernel function. Moreover, our empirical results will demonstrate that the proposed method yields sparser kernel density estimates without sacrificing the accuracy.

Define $\boldsymbol{\phi}^T(k) = [K_{k,1} \ K_{k,2} \cdots K_{k,N}]$ with $K_{k,i} = K_\rho(\mathbf{x}_k, \mathbf{x}_i)$, $y_k = \hat{p}(\mathbf{x}_k; \mathbf{1}_N/N, \rho_{\mathrm{Par}})$, and $\epsilon_k = \epsilon(\mathbf{x}_k)$. Then the model (7) at the data point $\mathbf{x}_k \in D_N$ is expressed as

$$y_k = \hat{y}_k + \epsilon_k = \boldsymbol{\phi}^T(k) \boldsymbol{\beta}_N + \epsilon_k. \tag{12}$$

Introduce the regression matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \ \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_N]$ with $\boldsymbol{\phi}_k = [K_{1,k} \ K_{2,k} \cdots K_{N,k}]^T$. Note that $\boldsymbol{\phi}_k$ is the kth column of $\boldsymbol{\Phi}$, while $\boldsymbol{\phi}^T(k)$ is the kth row of $\boldsymbol{\Phi}$. With the additional notations $\boldsymbol{\epsilon} = [\epsilon_1 \ \epsilon_2 \cdots \epsilon_N]^T$ and $\mathbf{y} = [y_1 \ y_2 \cdots y_N]^T$, the model (12) over D_N can be written in the matrix form

$$\mathbf{y} = \boldsymbol{\Phi} \boldsymbol{\beta}_N + \boldsymbol{\epsilon}. \tag{13}$$

III. Proposed Sparse Density Estimation Method

Our aim is to seek a sparse representation for $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_N, \rho)$ with most elements of $\boldsymbol{\beta}_N$ being zero and yet maintaining a comparable test performance or generalisation capability to that of the PW estimate. One approach is to work on the full regression matrix $\boldsymbol{\Phi}$ and to make as many kernel weights to (near) zero as possible based on some appropriate criteria, thus yielding a sparse representation, as in [7], [8], [9]. Alternatively, the efficient OFR procedure can be used to select a small subset of significant kernels based on some relevant criteria, thus constructing a sparse kernel model, as in [10], [11], [12]. We adopt the second approach here.

A. Subset Kernel Selection Using D-Optimality Criterion

Consider the model (13) in the generic data modelling context. The LS estimate of $\boldsymbol{\beta}_N$ is given by $\hat{\boldsymbol{\beta}}_N = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{y}$. Assume that (13) represents the true data generating process and the design matrix $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is nonsingular. The estimate $\hat{\boldsymbol{\beta}}_N$ is unbiased and the covariance matrix of the estimate is determined by the design matrix, namely

$$\begin{cases} E\left[\hat{\boldsymbol{\beta}}_N\right] = \boldsymbol{\beta}_N, \\ \text{Cov}\left[\hat{\boldsymbol{\beta}}_N\right] \propto \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}. \end{cases} \tag{14}$$

The condition number of the design matrix is given by

$$C = \frac{\max\{\lambda_i, 1 \le i \le N\}}{\min\{\lambda_i, 1 \le i \le N\}} \tag{15}$$

with λ_i, $1 \le i \le N$, being the eigenvalues of $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$. Too large a condition number will result in unstable LS parameter estimate while a small C improves model robustness. The D-optimality design criterion [14] maximises the determinant of the design matrix for the constructed model. Specifically, let $\boldsymbol{\Phi}_{N_s}$ be a column subset of $\boldsymbol{\Phi}$ representing a constructed N_s-term subset model. According to the D-optimality criterion, the selected subset model is the one that maximises $\det\left(\boldsymbol{\Phi}_{N_s}^T\boldsymbol{\Phi}_{N_s}\right)$. This helps to prevent the selection of an oversized ill-posed model and the problem of high parameter estimate variances. Moreover, the design matrix does not depend on $\mathbf{y}$ explicitly. Hence, the D-optimality design is an unsupervised learning, making it particularly suitable for determining the structure of kernel density estimate.

Let an orthogonal decomposition of the regression matrix $\boldsymbol{\Phi}$ be $\boldsymbol{\Phi} = \mathbf{WA}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & a_{1,2} & \cdots & a_{1,N} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{N-1,N} \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{16}$$

and $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2 \cdots \mathbf{w}_N]$ with orthogonal columns satisfying $\mathbf{w}_i^T\mathbf{w}_j = 0$, if $i \ne j$. Similarly, the orthogonal matrix corresponding to $\boldsymbol{\Phi}_{N_s}$ is denoted as $\mathbf{W}_{N_s}$. Maximising $\det\left(\boldsymbol{\Phi}_{N_s}^T\boldsymbol{\Phi}_{N_s}\right)$ is identical to maximising $\det\left(\mathbf{W}_{N_s}^T\mathbf{W}_{N_s}\right)$ or, equivalently, minimising $-\log\det\left(\mathbf{W}_{N_s}^T\mathbf{W}_{N_s}\right)$. In fact,

$$\begin{aligned} \det\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right) &= \det\left(\mathbf{A}^T\right)\det\left(\mathbf{W}^T\mathbf{W}\right)\det\left(\mathbf{A}\right) \\ &= \det\left(\mathbf{W}^T\mathbf{W}\right) = \prod_{i=1}^{N}\lambda_i, \end{aligned} \tag{17}$$

and

$$-\log\det\left(\mathbf{W}^T\mathbf{W}\right) = \sum_{i=1}^{N} -\log\left(\mathbf{w}_i^T\mathbf{w}_i\right). \tag{18}$$

Denote $\mathbf{B} = \boldsymbol{\Phi}^T\boldsymbol{\Phi} = [b_{i,j}] \in \mathcal{R}^{N\times N}$. The fast algorithm for the modified Gram-Schmidt orthogonalisation procedure [20] can readily be used to orthogonalise $\mathbf{B}$ and to calculate $\mathbf{A}$. For convenience, the same notation $\mathbf{B} = [b_{i,j}]$ is used to denote the design matrix after its first $n \times n$ block has been orthogonalised. The n-th stage of the D-optimality based OFR selection procedure is given as follows.

D-optimality based OFR

Begin: For $n \le j \le N$, calculate $J_n^{(j)} = -\log\left(b_{j,j}\right)$ and find $J_n = J_n^{(j_n)} = \min\{J_n^{(j)},\ n \le j \le N\}$

- If
$$J_n > \xi \tag{19}$$
where ξ is a threshold value that determines the size of the subset model, goto *Stop*.
- Otherwise, the j_n-th column of $\mathbf{B}$ is interchanged from the n-th row upwards with the n-th column of $\mathbf{B}$, and then the j_n-th row of $\mathbf{B}$ is interchanged from the n-th column upwards with the n-th row of $\mathbf{B}$. The j_n-th column of $\mathbf{A}$ is interchanged up to the $(n-1)$-th row with the n-th column of $\mathbf{A}$.
This effectively selects the j_n-th candidate as the n-th regressor in the subset model.
- For $n+1 \le j \le N$, compute $\alpha_{n,j} = b_{n,j}/b_{n,n}$, and for $n+1 \le j \le N$ and $j \le l \le N$, compute
$$\begin{cases} b_{j,l} = b_{j,l} - \alpha_{n,j}\alpha_{n,l}b_{n,n}, \\ b_{l,j} = b_{j,l}. \end{cases}$$
Set $n = n+1$ and go to *Begin*.

Stop: This selects $n-1$ most significant kernels according to the D-optimality criterion to form the selected subset model.

The desired threshold value ξ is problem dependent, and it is typically determined by simply observing the values of $-\log\left(\mathbf{w}_i^T\mathbf{w}_i\right) = -\log\left(b_{i,i}\right)$ for $i = 1, 2, \cdots$, and terminating the selection when it is appropriate. Alternatively, one can simply set a maximum number N_s for the selected kernels, where $N_s \ll N$. It does not matter if N_s is set too large, as the MNQP algorithm [13] used will automatically make some of the kernel weights to (near) zero, and thus reduces the model size to an appropriate level. It can be seen that the computational complexity of this D-optimality based OFR algorithm is no more than $\mathcal{O}(N^2)$, which is much simpler than other existing regression-based algorithms for SKD estimation [10], [11], [12].

B. Calculating Kernel Weights

After the structure determination using the D-optimality based OFR, we obtain a N_s-term subset kernel model, where $N_s \ll N$. The resulting regression modelling problem is

$$\mathbf{y} = \boldsymbol{\Phi}_{N_s}\boldsymbol{\beta}_{N_s} + \boldsymbol{\epsilon} \tag{20}$$

subject to the constraints

$$\boldsymbol{\beta}_{N_s}^T\mathbf{1}_{N_s} = 1 \text{ and } \beta_i \ge 0,\ 1 \le i \le N_s. \tag{21}$$

where $\boldsymbol{\beta}_{N_s}^T = [\beta_1\ \beta_2 \cdots \beta_{N_s}]$. The kernel weight vector can be obtained by solving the following constrained nonnegative quadratic programming

$$\begin{aligned} &\min_{\boldsymbol{\beta}_{N_s}}\{\tfrac{1}{2}\boldsymbol{\beta}_{N_s}^T\mathbf{B}_{N_s}\boldsymbol{\beta}_{N_s} - \mathbf{v}_{N_s}^T\boldsymbol{\beta}_{N_s}\} \\ &\text{s.t. } \boldsymbol{\beta}_{N_s}^T\mathbf{1}_{N_s} = 1 \text{ and } \beta_i \ge 0,\ 1 \le i \le N_s, \end{aligned} \tag{22}$$

where $\mathbf{B}_{N_s} = \mathbf{\Phi}_{N_s}^T \mathbf{\Phi}_{N_s} = [b_{i,j}] \in \mathcal{R}^{N_s \times N_s}$ and $\mathbf{v}_{N_s} = \mathbf{\Phi}_{N_s}^T \mathbf{y} = [v_1 \ v_2 \cdots v_{N_s}]^T$. The solution for $\boldsymbol{\beta}_{N_s}$ can be obtained iteratively using a modified MNQP algorithm [9].

As the elements of $\mathbf{B}_{N_s}$ and $\mathbf{v}_{N_s}$ are strictly positive, the Lagrangian for the above problem can be formed as [9]

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} b_{i,j} \frac{\beta_j^{(t)} \left(\beta_i^{(t+1)} \right)^2}{\beta_i^{(t)}} - \sum_{i=1}^{N_s} v_i \beta_i^{(t+1)} \\ &\quad - \eta^{(t)} \left(\sum_{i=1}^{N_s} \beta_i^{(t+1)} - 1 \right) \end{aligned} \tag{23}$$

where the superindex $^{(t)}$ denotes the iteration index and η is the lagrangian multiplier. Setting the gradients of $\mathcal{L}$ with respect to $\beta_i^{(t+1)}$ and $\eta^{(t)}$ to zeros, leads to the following updating equations

$$c_i^{(t)} = \beta_i^{(t)} \left(\sum_{j=1}^{N_s} b_{i,j} \beta_j^{(t)} \right)^{-1}, \ 1 \le i \le N_s, \tag{24}$$

$$\eta^{(t)} = \left(\sum_{i=1}^{N_s} c_i^{(t)} \right)^{-1} \left(1 - \sum_{i=1}^{N_s} c_i^{(t)} v_i \right), \tag{25}$$

$$\beta_i^{(t+1)} = c_i^{(t)} \left(v_i + \eta^{(t)} \right). \tag{26}$$

The initial condition can be set as $\beta_i^{(0)} = \frac{1}{N_s}$, $1 \le i \le N_s$.

During the iterative procedure, some of the kernel weights may be driven to (near) zero, particularly when the subset model size N_s is chosen to be larger than necessary. The corresponding kernels can then be removed from the kernel model, leading to a reduction in the subset model size. Because N_s is typically very small, this MNQP algorithm imposes only a small amount of computational requirements.

IV. Numerical Experiments

Several examples were used to test the proposed SKD estimator using the combined D-optimality OFR and MNQP algorithm and to compare its performance with the PW estimator as well as other existing SKD estimators. The value of the kernel width ρ was determined by test performance via cross validation. For each example, a data set of N randomly drawn samples was used to construct kernel density estimates, and a separate test data set of $N_{\text{test}} = 10,000$ samples was used to calculate either the L_2 or the L_1 test errors for the resulting estimate according to

$$L_2 = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} \left| p(\mathbf{x}_k) - \hat{p}(\mathbf{x}_k; \boldsymbol{\beta}_{N_s}, \rho) \right|^2, \tag{27}$$

and

$$L_1 = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} \left| p(\mathbf{x}_k) - \hat{p}(\mathbf{x}_k; \boldsymbol{\beta}_{N_s}, \rho) \right|, \tag{28}$$

respectively. The experiment was repeated by N_{run} different random runs for each example.

TABLE I

Performance comparison in terms of L_2 test error and number of kernels required for the one-dimensional example of eight-Gaussian mixture, quoted as mean $\pm$ standard deviation over 200 runs.

method	L_2 test error	kernel number
PW estimate	$(2.9311 \pm 2.0601) \times 10^{-3}$	200 ± 0
SKD estimate of [12]	$(3.0181 \pm 2.0991) \times 10^{-3}$	10.2 ± 1.6
proposed SKD estimate	$(2.8762 \pm 2.0775) \times 10^{-3}$	8.7 ± 0.9

A. One-Dimensional Examples

Example 1. The density to be estimated was the mixture of eight Gaussian distributions given by

$$p(x) = \frac{1}{8} \sum_{i=0}^{7} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \tag{29}$$

with

$$\sigma_i = \sqrt{\left(\frac{2}{3}\right)^i}, \ \mu_i = 3\left(\left(\frac{2}{3}\right)^i - 1\right), \ 0 \le i \le 7. \tag{30}$$

Our previous SKD estimator [12] was shown to be favourable, compared with the estimator of [9], in terms of test performance and model sparsity. Here we compared the proposed SKD estimator with the PW estimator and our previous estimator [12]. The number of data points for density estimation was $N = 200$. The experiment was repeated $N_{\text{run}} = 200$ times. The optimal kernel widths were found to be $\rho = 0.17$ and $\rho = 0.31$ empirically for the PW and proposed SKD estimators, respectively. We observed that the significant kernel terms according to the D-optimality criterion were in the range of 8 to 10 and the threshold value could be set to $\xi = -1.0$. However, we simply set the maximum number of selected kernels by the D-optimality OFR to be $N_s = 16$ and let the MNQP algorithm to decide the final model size. Table I compares the performance of the three density estimates. It can be seen that the accuracy of the proposed SKD estimator was comparable to that of the PW estimator, but it realised very sparse estimates with an average kernel number less than 5% of the data samples.

Example 2. The density to be estimated for this one-dimensional example was the mixture of Gaussian and Laplacian distributions given by

$$p(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} + \frac{0.7}{4} e^{-0.7|x+2|}. \tag{31}$$

TABLE II

Performance comparison in terms of L_1 test error and number of kernels required for the one-dimensional example of Gaussian and Laplacian mixture, quoted as mean $\pm$ standard deviation over 200 runs.

method	L_1 test error	kernel number
PW estimator	$(1.9503 \pm 0.5881) \times 10^{-2}$	100 ± 0
SKD estimate of [11]	$(2.1785 \pm 0.7468) \times 10^{-2}$	4.8 ± 0.9
SKD estimate of [12]	$(1.9436 \pm 0.6208) \times 10^{-2}$	5.1 ± 1.3
proposed SKD estimate	$(1.8333 \pm 0.6144) \times 10^{-2}$	3.3 ± 0.7

The number of data points for density estimation was $N = 100$. The optimal kernel widths were found to be $\rho = 0.54$ and $\rho = 1.1$ for the PW and SKD estimators, respectively. The experiment was repeated $N_{\rm run} = 200$ times. According to the D-optimality criterion, only three kernel terms were significant and the threshold value could be set to $\xi = 0.0$. But we simply set the maximum number of selected kernels by the D-optimality based OFR to be $N_s = 10$ and let the MNQP algorithm to further reduce the model size. Table II compares the performance of the four kernel density estimates, in terms of the L_1 test error and the number of kernels required. Compared with the results given in [9], it can be seen that our proposed SKD estimator also had better performance than the SKD estimator of [9] and the SVM-based density estimator, both in terms of test performance and model sparsity, for this one-dimensional example.

B. Two-Dimensional Examples

Example 3. The density to be estimated for this two-dimensional example was defined by the mixture of Gaussian and Laplacian distributions given as follows

$$\begin{aligned} p(x_1, x_2) &= \frac{1}{4\pi} e^{-\frac{(x_1-2)^2}{2}} e^{-\frac{(x_2-2)^2}{2}} \\ &\quad + \frac{0.35}{8} e^{-0.7|x_1+2|} e^{-0.5|x_2+2|}. \end{aligned} \tag{32}$$

The estimation data set contained $N = 500$ samples, and the empirically found optimal kernel widths were $\rho = 0.42$ for the PW estimate and $\rho = 1.1$ for the SKD estimate. The experiment was repeated $N_{\rm run} = 100$ times. We simply set the maximum selected kernels by the D-optimality OFR procedure to be $N_s = 16$ and let the MNQP algorithm to decide the final model size. Table III lists the L_1 test errors and the numbers of kernels required for the four density estimators, where it can be seen that our proposed SKD estimator produced the best result. Note that the proposed SKD estimator is also computationally much simpler than our previous density estimators [11], [12].

Example 4. The true density to be estimated for this two-dimensional example was defined by the mixture of five Gaussian distributions given as follows

$$p(x_1, x_2) = \sum_{i=1}^{5} \frac{1}{10\pi} e^{-\frac{(x_1-\mu_{i,1})^2}{2}} e^{-\frac{(x_2-\mu_{i,2})^2}{2}} \tag{33}$$

and the means of the five Gaussian distributions, $[\mu_{i,1}\ \mu_{i,2}]$, $1 \le i \le 5$, were $[0.0\ -4.0]$, $[0.0\ -2.0]$, $[0.0\ 0.0]$, $[-2.0\ 0.0]$, and $[-4.0\ 0.0]$, respectively. The number of data points for density estimation was $N = 500$. The optimal kernel widths were found to be $\rho = 0.5$ and $\rho = 1.0$ for the PW and SKD estimators, respectively. The experiment was repeated $N_{\rm run} = 100$ times. The maximum number of selected kernels by the D-optimality based OFR was set to $N_s = 16$, and the final model size was determined by the MNQP algorithm automatically. Table IV compares the performance of the three density estimators.

TABLE III

PERFORMANCE COMPARISON IN TERMS OF L_1 TEST ERROR AND NUMBER OF KERNELS REQUIRED FOR THE TWO-DIMENSIONAL EXAMPLE OF GAUSSIAN AND LAPLACIAN MIXTURE, QUOTED AS MEAN $\pm$ STANDARD DEVIATION OVER 100 RUNS.

method	L_1 test error	kernel number
PW estimate	$(4.2453 \pm 0.8242) \times 10^{-3}$	500 ± 0
SKD estimate of [11]	$(3.8381 \pm 0.8263) \times 10^{-3}$	11.9 ± 2.6
SKD estimate of [12]	$(3.8379 \pm 0.7797) \times 10^{-3}$	15.3 ± 3.9
proposed SKD estimate	$(3.7672 \pm 0.6937) \times 10^{-3}$	8.6 ± 1.0

TABLE IV

PERFORMANCE COMPARISON IN TERMS OF L_1 TEST ERROR AND NUMBER OF KERNELS REQUIRED FOR THE TWO-DIMENSIONAL EXAMPLE OF FIVE-GAUSSIAN MIXTURE, QUOTED AS MEAN $\pm$ STANDARD DEVIATION OVER 100 RUNS.

method	L_1 test error	kernel number
PW estimate	$(3.6204 \pm 0.4394) \times 10^{-3}$	500 ± 0
SKD estimate of [12]	$(3.6100 \pm 0.5025) \times 10^{-3}$	13.2 ± 2.9
proposed SKD estimate	$(3.2355 \pm 0.5575) \times 10^{-3}$	7.9 ± 0.8

Example 5. This was a two-class classification problem in a two-dimensional feature space [21] and we obtained the data from [22]. The training set contained 250 samples with 125 points for each class, and the test set had 1000 points with 500 samples for each class. The optimal Bayes error rate based on the true underlying probability distribution was known to be 8%. We first estimated the two conditional PDFs $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C0})$ and $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C1})$ from the training data, and then applied the Bayes decision rule

$$\left.\begin{array}{ll} \text{if } \hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C0}) \ge \hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C1}), & \mathbf{x} \in \text{C0} \\ \text{else}, & \mathbf{x} \in \text{C1} \end{array}\right\} \tag{34}$$

to the test data set. Table V lists the results obtained by the four density estimators, where the values of the kernel width ρ were found by cross validation. It can be seen that the proposed SKD estimator required only two kernels for each conditional PDF estimate, and the resulting test error rate was identical to the optimal Bayes classification error rate.

C. Multi-Dimensional Examples

Example 6. The underlying density to be estimated was

$$p(\mathbf{x}) = \frac{1}{3} \sum_{i=1}^{3} \frac{1}{(2\pi)^{6/2}} \frac{1}{\det^{1/2} |\boldsymbol{\Gamma}_i|} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Gamma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)} \tag{35}$$

with

$$\begin{aligned} \boldsymbol{\mu}_1 &= [1.0\ 1.0\ 1.0\ 1.0\ 1.0\ 1.0]^T, \\ \boldsymbol{\Gamma}_1 &= \text{diag}\{1.0, 2.0, 1.0, 2.0, 1.0, 2.0\}, \end{aligned} \tag{36}$$

TABLE V

PERFORMANCE COMPARISON FOR THE TWO-CLASS TWO-DIMENSIONAL CLASSIFICATION EXAMPLE.

method	$\hat{p}(\bullet|\text{C0})$	$\hat{p}(\bullet|\text{C1})$	test error rate
PW estimate	125 kernels	125 kernels	8.0%
SKD estimate of [11]	5 kernels	4 kernels	8.3%
SKD estimate of [12]	6 kernels	5 kernels	8.0%
proposed SKD estimate	2 kernels	2 kernels	8.0%

TABLE VI

PERFORMANCE COMPARISON IN TERMS OF L_1 TEST ERROR AND NUMBER OF KERNELS REQUIRED FOR THE SIX-DIMENSIONAL EXAMPLE OF THREE-GAUSSIAN MIXTURE, QUOTED AS MEAN $\pm$ STANDARD DEVIATION OVER 100 RUNS.

method	L_1 test error	kernel number
PW estimate	$(3.5195 \pm 0.1616) \times 10^{-5}$	600 ± 0
SKD estimate of [11]	$(4.4781 \pm 1.2292) \times 10^{-5}$	14.9 ± 2.1
SKD estimate of [12]	$(3.1134 \pm 0.5335) \times 10^{-5}$	9.4 ± 1.9
proposed SKD estimate	$(2.7822 \pm 0.2271) \times 10^{-5}$	8.4 ± 0.9

$$\begin{aligned} \boldsymbol{\mu}_2 &= [-1.0 \ -1.0 \ -1.0 \ -1.0 \ -1.0 \ -1.0]^T, \\ \boldsymbol{\Gamma}_2 &= \text{diag}\{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}, \end{aligned} \tag{37}$$

$$\begin{aligned} \boldsymbol{\mu}_3 &= [0.0 \ 0.0 \ 0.0 \ 0.0 \ 0.0 \ 0.0]^T, \\ \boldsymbol{\Gamma}_3 &= \text{diag}\{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}. \end{aligned} \tag{38}$$

The estimation data set contained $N = 600$ samples. The optimal kernel width was found to be $\rho = 0.65$ for the PW estimate and $\rho = 1.2$ for the SKD estimate, via cross validation. The experiment was repeated $N_{\text{run}} = 100$ times. The number of kernels selected by the D-optimality OFR was set to $N_s = 16$, and the final model size was left to the MNQP algorithm to determine. The results obtained by the four density estimators are summarised in Table VI.

Example 7. This was a two-class classification data set, Titanic [23]. The feature space dimension was $m = 3$. There were 100 realisations of the data set, each containing 150 training samples and 2051 test samples. Note that the two-class data samples were imbalanced, with the class-0 training samples approximately twice of the class-1 training samples. In [24], a range of classifiers were applied to this data set, and the best classification test error rate in %, obtained by the SVM classifier, averaged over the 100 realisations was 22.42 ± 1.02. We estimated the two conditional PDFs $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C0})$ and $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C1})$ from the training data, and then applied the Bayes decision rule (34) to the test data and calculated the corresponding error rate. The kernel width was set to $\rho = 1.0$ for both the PW and SKD estimates. The results obtained in terms of total kernel number required for $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C0})$ and $\hat{p}(\mathbf{x}; \boldsymbol{\beta}_{N_s}, \rho|\text{C1})$ and test performance are listed in Table VII.

V. CONCLUSIONS

An efficient construction algorithm has been proposed for obtaining SKD estimates. A very small subset of significant kernels was first selected using the OFR procedure based on the D-optimality criterion. The associated kernel weights are calculated using a modified MNQP algorithm, which can further reduce the kernel model size by making some of the kernel weights to zero. The proposed method is simple to implement and computationally efficient. Several examples have demonstrated that the proposed method compares favourably with other existing SKD estimation methods, both in terms of test accuracy and sparsity of the estimate.

TABLE VII

PERFORMANCE COMPARISON FOR THE TITANIC DATA SET IN TERMS OF TOTAL KERNEL NUMBER REQUIRED FOR TWO CONDITIONAL PDF ESTIMATES AND TEST ERROR RATE OVER 100 REALISATIONS.

method	kernel number	test error rate in %
PW estimate	150 ± 0	22.48 ± 0.43
proposed SKD estimate	7.8 ± 4.4	22.34 ± 0.34

REFERENCES

[1] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol.33, pp.1066–1076, 1962.

[2] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman Hall, 1986.

[3] A.W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*. Oxford, U.K.: Oxford University Press, 1997.

[4] C.M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford University Press, 1995.

[5] H. Wang, "Robust control of the output probability density functions for multivariable stochastic systems with guaranteed stability," *IEEE Trans. Automatic Control*, vol.44, no.11, pp.2103–2107, 1999.

[6] S. Chen, A.K. Samingan, B. Mulgrew and L. Hanzo, "Adaptive minimum-BER linear multiuser detection for DS-CDMA signals in multipath channels," *IEEE Trans. Signal Processing*, vol.49, no.6, pp.1240–1247, 2001.

[7] J. Weston, A. Gammerman. M.O. Stitson, V. Vapnik, V. Vovk and C. Watkins, "Support vector density estimation," in: B. Schölkopf, C. Burges and A.J. Smola, eds., *Advances in Kernel Methods — Support Vector Learning*, Cambridge MA: MIT Press, 1999, pp.293–306.

[8] V. Vapnik and S. Mukherjee, "Support vector method for multivariate density estimation," in: S. Solla, T. Leen and K.R. Müller, eds., *Advances in Neural Information Processing Systems*, MIT Press, 2000, pp.659–665.

[9] M. Girolami and C. He, "Probability density estimation from optimally condensed data samples," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.25, no.10, pp.1253–1264, 2003.

[10] A. Choudhury, *Fast Machine Learning Algorithms for Large Data*. PhD Thesis, Computational Engineering and Design Center, School of Engineering Sciences, University of Southampton, 2002.

[11] S. Chen, X. Hong and C.J. Harris, "Sparse kernel density construction using orthogonal forward regression with leave-one-out test score and local regularization," *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol.34, no.4, pp.1708–1717, 2004.

[12] S. Chen, X. Hong and C.J. Harris, "An orthogonal forward regression technique for sparse kernel density estimation," *Neurocomputing*, vol.71, no.4-6, pp.931–943, 2008.

[13] F. Sha, L.K. Saul and D.D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," *Technical Report*. MS-CIS-02-19, University of Pennsylvania, USA, 2002.

[14] A.C. Atkinson and A.N. Donev, *Optimum Experimental Designs*. Oxford, U.K.: Clarendon Press, 1992.

[15] X. Hong and C.J. Harris, "Neurofuzzy design and model construction of nonlinear dynamical processes from data," *IEE Proc. Control Theory and Applications*, vol.148, no.6, pp.530–538, 2001.

[16] X. Hong and C.J. Harris, "Nonlinear model structure detection using optimum design and orthogonal least squares," *IEEE Trans. Neural Networks*, vol.12, no.2, pp.435–439, 2001.

[17] X. Hong and C.J. Harris, "Nonlinear model structure design and construction using orthogonal least squares and D-optimality design," *IEEE Trans. Neural Networks*, vol.13, no.5, pp.1245–1250, 2002.

[18] S. Chen, X. Hong and C.J. Harris, "Sparse kernel regression modeling using combined locally regularized orthogonal least squares and D-optimality experimental design," *IEEE Trans. Automatic Control*, vol.48, no.6, pp.1029–1036, 2003.

[19] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley, 2000.

[20] S. Chen and J. Wigger, "Fast orthogonal least squares algorithm for efficient subset model selection," *IEEE Trans. Signal Processing*, vol.43, no.7, pp.1713–1715, 1995.

[21] B.D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge University Press, 1996.

[22] `http://www.stats.ox.ac.uk/PRNN/`

[23] `ida.first.fhg.de/projects/bench/benchmarks.htm`

[24] G. Rätsch, T. Onoda, and K.R. Müller, "Soft margins for AdaBoost," *Machine Learning*, vol.42, no.3, pp.287–320, 2001.