Sparse kernel density estimation technique based on zero-norm constraint

Sparse Kernel Density Estimation Technique Based on Zero-Norm Constraint

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

***Abstract*— A sparse kernel density estimator is derived based on the zero-norm constraint, in which the zero-norm of the kernel weights is incorporated to enhance model sparsity. The classical Parzen window estimate is adopted as the desired response for density estimation, and an approximate function of the zero-norm is used for achieving mathemtical tractability and algorithmic efficiency. Under the mild condition of the positive definite design matrix, the kernel weights of the proposed density estimator based on the zero-norm approximation can be obtained using the multiplicative nonnegative quadratic programming algorithm. Using the D-optimality based selection algorithm as the preprocessing to select a small significant subset design matrix, the proposed zero-norm based approach offers an effective means for constructing very sparse kernel density estimates with excellent generalisation performance.**

I. Introduction

A fundamental principle in data modelling is the parsimonious principle of ensuring the smallest possible model with the best model generalisation performance from observational data. In linear-in-the-parameters modelling, the number of terms in the model is referred to as the zero-norm of the parameters. Minimising this zero norm is related to variable and feature selection, which ensures model sparsity and enhances model generalisation [1], [2]. Because of the intractability in the minimisation of the zero-norm, considerable research efforts have been focused on the approximation schemes on the zero-norm [1], [2] and their associated computational complexities.

The estimation of probability density function (PDF) from observed data samples is a fundamental problem in all the fields of engineering [3]–[7]. The Parzen window (PW) estimate is a simple yet remarkably accurate nonparametric density estimation technique [8]. A general and powerful approach to the problem of PDF estimation is the finite mixture model (FMM) [9]. The FMM includes the PW estimate as a special case in that the number of mixtures is set to the number of training data samples and equal weights are adopted in the PW estimator. A disadvantage associated with the PW estimate is its high computational cost of the point density estimate for a future data sample in the case where the training data set is very large. Clearly by taking a much smaller number of mixture components, the FMM can be regarded as a condensed representation of data [9]. Note that the mixing weights and the parameters of mixing components in the FMM need to be determined through parametric optimisation using for example the expectation-maximisation (EM) algorithm, which is an ill-posed and costly nonlinear optimisation problem. For the Gaussian mixture model (GMM), the EM algorithm can be derived in an explicit and simple iterative form [10].

The high test cost of the PW estimator has motivated a considerable interest in the research into the sparse PDF estimate, including th support vector machine (SVM) density estimation technique [11], [12] and the reduced set density estimator (RSDE) [13]. Alternatively, a regression-based PDF estimation method was introduced [14], in which the empirical cumulative distribution function is constructed as the desired response, similar to the SVM method of [11]. With the aid of the sparse modelling technique [15], [16], the regression-based idea of [14] has been extended to yield sparse density estimation algorithm based on an orthogonal forward regression (OFR) algorithm [17], which is capable of automatically constructing very sparse kernel density estimate with excellent generalisation performance. Alternatively, a simple and viable alternative approach has been proposed to use the kernels directly as regressors and the PW estimate as the target response [18].

Following the idea of using PW estimate as the target function [18], we introduce a new sparse kernel PDF estimator based on the zero-norm constraint. For mathemtical tractability and algorithmic efficiency, an approximate function of the zero-norm of the kernel weights is minimised. We show that, within the constrained kernel density estimation, it is the maximisation, not minimisation, of the two-norm of the kernel weights which leads to model sparsity. We further show that, when the zero-norm constrained design matrix is positive definite, the kernel weights of the proposed PDF estimator based on the zero-norm approximation can be updated using the multiplicative nonnegative quadratic programming (MNQP) algorithm of [19]. Thus it is highly desirable to apply a preprocessing for selecting a small significant subset design matrix, and we propose to use the efficient D-optimality based method [20] for this preprocessing. Numerical examples included demonstrate that the proposed estimator based on the zero-norm constraint offers a highly efficient means for selecting very sparse kernel density estimates with excellent generalisation performance.

II. The kernel density estimation

Let a finite data set consisting of N data samples, $D_N = \{\mathbf{x}_k\}_{k=1}^{N}$, be drawn from a density $p(\mathbf{x})$, where $\mathbf{x}_k \in \mathcal{R}^m$. The problem under study is to infer the unknown $p(\mathbf{x})$ based

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

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

on D_N using the kernel density estimate of the form

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

subject to

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

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

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 the dimension N, and $K_\rho(\bullet,\bullet)$ is the chosen kernel function with kernel width ρ. In this study, we use the Gaussian kernel

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

where $\mathbf{c}\in\mathcal{R}^m$ is the kernel centre vector. But any kernel function, satisfying $K_\rho(\mathbf{x},\mathbf{c})\geq 0\ \forall\mathbf{x}\in\mathcal{R}^m$ and

$$\int_{\mathcal{R}^m}K_\rho(\mathbf{x},\mathbf{c})\,d\mathbf{x}=1 \tag{5}$$

can also be is used in the density estimate (1). We point out that the kernel width ρ is assumed to be provided, for example, via cross validation.

Let the PW estimator be denoted by $\hat{p}(\mathbf{x};\boldsymbol{\beta}_N^{\text{Par}},\rho^{\text{Par}})$, where the elements of $\boldsymbol{\beta}_N^{\text{Par}}$ are all equal, namely, $\beta_i^{\text{Par}}=\frac{1}{N}$, $1\leq i\leq N$. The log-likelihood for $\boldsymbol{\beta}_N$ can be formed using the observed data D_N as

$$\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}(\mathbf{x}_i;\boldsymbol{\beta}_N,\rho)=\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{j=1}^{N}\beta_j K_\rho(\mathbf{x}_i,\mathbf{x}_j)\right). \tag{6}$$

By the law of large numbers the log-likelihood (6) tends to

$$\int_{\mathcal{R}^m}p(\mathbf{x})\log\hat{p}(\mathbf{x};\boldsymbol{\beta}_N,\rho)\,d\mathbf{x} \tag{7}$$

as $N\rightarrow\infty$ with probability one. The measure (7) is simply the negative cross-entropy or divergence between the true density $p(\mathbf{x})$ and the estimate $\hat{p}(\mathbf{x};\boldsymbol{\beta}_N,\rho)$. It can be shown that the PW estimator $\beta_i^{\text{Par}}=\frac{1}{N}$, $1\leq i\leq N$, is obtained by maximising the log-likelihood (6) with respective to $\boldsymbol{\beta}_N$ subject to the constraints (2) and (3). Note that the choice of ρ^{Par} is crucial in density estimation using the PW estimate [5]. Based on the principle of minimising the mean integrated square error [5], ρ^{Par} can be found by minimising the following the least squares criterion evaluated on D_N

$$\begin{aligned}&\frac{1}{N^2}\sum_{i,j=1}^{N}K_{\sqrt{2}\rho}(\mathbf{x}_i,\mathbf{x}_j)-\frac{2}{N(N-1)}\sum_{i,j=1,j\neq i}^{N}K_\rho(\mathbf{x}_i,\mathbf{x}_j)\\ &\approx\ \frac{1}{N^2}\sum_{i,j=1}^{N}K_\rho^*(\mathbf{x}_i,\mathbf{x}_j)+\frac{2}{N(2\pi\rho^2)^{m/2}},\end{aligned} \tag{8}$$

where $K_\rho^*(\mathbf{x}_i,\mathbf{x}_j)=K_{\sqrt{2}\rho}(\mathbf{x}_i,\mathbf{x}_j)-2K_\rho(\mathbf{x}_i,\mathbf{x}_j)$. Typically, ρ^{Par} is found by a grid search.

With the PW estimator, the associated computational cost for evaluating the PDF estimate for a future sample scales directly with the sample size N. Therefore it is desirable to devise a sparse representation of $\hat{p}(\mathbf{x};\boldsymbol{\beta}_N,\rho)$, in which most of the elements in $\boldsymbol{\beta}_N$ are zero. Because the PW estimator has the above-mentioned "optimal" property, it was suggested [18] that the PW estimator can be used as the desired response for the sparse kernel density estimator. Specifically, a regression model linking $\hat{p}(\mathbf{x};\boldsymbol{\beta}_N,\rho)$ and $\hat{p}(\mathbf{x};\boldsymbol{\beta}_N^{\text{Par}},\rho^{\text{Par}})$ can be written as

$$\hat{p}(\mathbf{x};\boldsymbol{\beta}_N^{\text{Par}},\rho^{\text{Par}})=\sum_{i=1}^{N}\beta_i K_\rho(\mathbf{x},\mathbf{x}_i)+\varepsilon(\mathbf{x}) \tag{9}$$

where $\varepsilon(\mathbf{x})$ denotes the modelling error at $\mathbf{x}$. Define $y_k=\hat{p}(\mathbf{x}_k;\boldsymbol{\beta}_N^{\text{Par}},\rho^{\text{Par}})$, $\boldsymbol{\phi}_N(k)=[K_{k,1}\ K_{k,2}\cdots K_{k,N}]^T$ with $K_{k,i}=K_\rho(\mathbf{x}_k,\mathbf{x}_i)$ and $\varepsilon(k)=\varepsilon(\mathbf{x}_k)$. Then the model (9) at the data point $\mathbf{x}_k\in D_N$ can be expressed as

$$y_k=\hat{y}_k+\varepsilon(k)=\boldsymbol{\phi}_N^T(k)\boldsymbol{\beta}_N+\varepsilon(k). \tag{10}$$

The objective is to obtain $\boldsymbol{\beta}_N$ via minimising some modelling error criterion, such as the mean square error $E[\varepsilon^2(k)]$, and simultaneously to achieve a sparse representation of $\hat{p}(\mathbf{x};\boldsymbol{\beta}_N,\boldsymbol{\sigma})$ with most of the elements in $\boldsymbol{\beta}_N$ being zeros, subject to the constraints (2) and (3).

Over the training data set D_N, the model (9) can be written in the matrix form

$$\mathbf{y}=\boldsymbol{\Phi}_N\boldsymbol{\beta}_N+\boldsymbol{\varepsilon} \tag{11}$$

with the additional notations $\boldsymbol{\Phi}_N=[\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\cdots\boldsymbol{\phi}_N]=[K_{k,i}]\in\mathcal{R}^{N\times N}$, $1\leq i,k\leq N$, $\boldsymbol{\varepsilon}=[\varepsilon(1)\ \varepsilon(2)\cdots\varepsilon(N)]^T$ and $\mathbf{y}=[y(1)\ y(2)\cdots y(N)]^T$. Note that $\boldsymbol{\phi}_k$ denotes the kth column of $\boldsymbol{\Phi}_N$ while $\boldsymbol{\phi}_N^T(k)$. is the kth row of $\boldsymbol{\Phi}_N$. The kernel weight vector $\boldsymbol{\beta}_N$ can be obtained by solving the following constrained nonnegative quadratic programming

$$\begin{aligned}&\min_{\boldsymbol{\beta}_N}\left\{\tfrac{1}{2}\boldsymbol{\beta}_N^T\mathbf{B}_N\boldsymbol{\beta}_N-\mathbf{v}_N^T\boldsymbol{\beta}_N\right\},\\ &\text{s.t. }\boldsymbol{\beta}_N^T\mathbf{1}_N=1\text{ and }\beta_i\geq 0,1\leq i\leq N,\end{aligned} \tag{12}$$

where $\mathbf{B}_N=\boldsymbol{\Phi}_N^T\boldsymbol{\Phi}_N$ is the design matrix and $\mathbf{v}_N=\boldsymbol{\Phi}_N^T\mathbf{y}$. Provided that $\mathbf{B}_N$ is positive definite, the solution can readily be obtained using the MNQP algorithm [13], [18], [19].

III. Sparse estimation with zero-norm constraint

The quantity $\|\boldsymbol{\beta}_N\|_0$ that counts the number of non-zero entries in $\boldsymbol{\beta}_N$ is referred to as the zero-norm of $\boldsymbol{\beta}_N$. In order to improve the sparsity of the model (9), $\|\boldsymbol{\beta}_N\|_0$ can be utilised as an additional constraint [1], [2]. It is a very hard problem to directly minimise the zero-norm of $\boldsymbol{\beta}_N$ [1], [21], and the work of [2] proposed an approximation with

$$\|\boldsymbol{\beta}_N\|_0\approx\sum_{i=1}^{N}\left(1-e^{-\alpha|\beta_i|}\right), \tag{13}$$

in which $\alpha>0$ is a chosen parameter. Following the idea in [2], the objective function in (12) can be modified to yield

$$\begin{aligned}&\min_{\boldsymbol{\beta}_N}\Big\{\tfrac{1}{2}\boldsymbol{\beta}_N^T\mathbf{B}_N\boldsymbol{\beta}_N-\mathbf{v}_N^T\boldsymbol{\beta}_N\\ &\qquad+\lambda\textstyle\sum_{i=1}^{N}\left(1-e^{-\alpha|\beta_i|}\right)\Big\},\\ &\text{s.t. }\boldsymbol{\beta}_N^T\mathbf{1}_N=1\text{ and }\beta_i\geq 0,1\leq i\leq N,\end{aligned} \tag{14}$$

where $\lambda > 0$ is a small parameter that regulates the tradeoff between the two objectives. We propose a further approximation by using the Taylor series expansion up to the second order for $e^{-\alpha|\beta_i|}$

$$e^{-\alpha|\beta_i|} \approx 1 - \alpha|\beta_i| + \frac{\alpha^2\beta_i^2}{2} \tag{15}$$

such that

$$\sum_{i=1}^{N}\left(1 - e^{-\alpha|\beta_i|}\right) \approx \alpha\sum_{i=1}^{N}|\beta_i| - \frac{\alpha^2}{2}\sum_{i=1}^{N}\beta_i^2. \tag{16}$$

Applying the constraints (2) and (3) to (16), we obtain

$$\sum_{i=1}^{N}\left(1 - e^{-\alpha|\beta_i|}\right) \approx \alpha - \frac{\alpha^2}{2}\boldsymbol{\beta}_N^T\boldsymbol{\beta}_N. \tag{17}$$

Based upon (17), the constrained optimisation (14) can be approximately reformulated as

$$\begin{array}{l}\min\limits_{\boldsymbol{\beta}_N}\left\{\frac{1}{2}\boldsymbol{\beta}_N^T\mathbf{A}_N\boldsymbol{\beta}_N - \mathbf{v}_N^T\boldsymbol{\beta}_N\right\}, \\ \text{s.t. } \boldsymbol{\beta}_N^T\mathbf{1}_N = 1 \text{ and } \beta_i \geq 0, 1 \leq i \leq N,\end{array} \tag{18}$$

where $\mathbf{A}_N = \mathbf{B}_N - \delta\mathbf{I}_N$, $\mathbf{I}_N$ is the $N \times N$ identity matrix and $\delta = \lambda\alpha^2$ is a predetermined small parameter.

Again assume that $\mathbf{B}_N$ is full rank. Provided that δ is set in a manner such that $\mathbf{A}_N$ is a positive definite matrix, the problem (18) is a constrained nonnegative quadratic programming whose solution can readily be solved using the MNQP algorithm [13], [18], [19], as for (12). In fact, let the N eigenvalues of $\mathbf{B}_N$ be arranged in the order $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_N = \sigma_{\min} > 0$. Then the condition for $\mathbf{A}_N$ to be a positive definite matrix is obvious: $\delta < \sigma_{\min}$. For completeness, the MNQP algorithm [13], [18] for solving (18) is described below. Denote $\mathbf{A}_N = [a_{i,j}] \in \mathcal{R}^{N\times N}$, $\mathbf{v}_N = [v_1\ v_2 \cdots v_N]^T$. Since the elements of $\mathbf{A}_N$ and $\mathbf{v}_N$ are strictly positive, the Lagrangian for the problem (18) can be formed as [13]

$$\begin{aligned}\mathcal{L} \;=\; & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_{i,j}\frac{\beta_j^{(t)}(\beta_i^{(t+1)})^2}{\beta_i^{(t)}} - \sum_{i=1}^{N}v_i\beta_i^{(t+1)} \\ & -h^{(t)}(\sum_{i=1}^{N}\beta_i^{(t+1)} - 1),\end{aligned} \tag{19}$$

where the superscript $^{(t)}$ denotes the iteration index and h is the Lagrangian multiplier. Setting

$$\frac{\partial\mathcal{L}}{\partial\beta_i^{(t+1)}} = 0 \text{ and } \frac{\partial\mathcal{L}}{\partial h^{(t)}} = 0 \tag{20}$$

yields the following updating equations

$$c_i^{(t)} = \beta_i^{(t)}\left(\sum_{j=1}^{N}a_{i,j}\beta_j^{(t)}\right), 1 \leq i \leq N, \tag{21}$$

$$h^{(t)} = \left(\sum_{i=1}^{N}c_i^{(t)}\right)^{-1}\left(1 - \sum_{i=1}^{N}c_i^{(t)}v_i\right), \tag{22}$$

$$\beta_i^{(t+1)} = c_i^{(t)}(v_i + h^{(t)}). \tag{23}$$

The initial condition can be set as $\beta_i^{(0)} = \frac{1}{N}$, $1 \leq i \leq N$. It is easy to verify that the constraints (2) and (3) are maintained during the iterative procedure. Over the iterations, some of the kernel weights are driven to near zero, and the corresponding kernels can be removed from the model (9).

Remark 1: From (17), it is seen that the minimisation of the proposed zero-norm approximation, combined with the convexity constraint of the kernel parameter vector, (2) and (3), is equivalent to the maximisation of the two-norm of the parameters. The fact that the maximisation of the two-norm of the parameters, subject to the convexity constraint of the parameters, encourages model sparsity is explained as follows. Under the convexity condition (2) and (3), the model sparsity is equivalent to the unevenness in the distribution of the parameter magnitudes. For example, the two-norm of the parameters is maximised as 1 when $\beta_k = 1$ and $\beta_j = 0$ for $\forall j \neq k$, which corresponds to the smallest zero-norm of 1. In this case, the parameters are the most unevenly distributed. On the other hand, the two-norm of the parameters is minimised as $\frac{1}{N}$ when $\beta_i = \frac{1}{N}$, $1 \leq i \leq N$, which corresponds to the largest zero-norm of N. In this case, the parameters are uniformly distributed, leading to a non-sparse estimate.

Remark 1 is interesting as it shows that the maximisation, not the minimisation, of the two-norm of the parameters leads to model sparsity. The strength of the zero norm constraint is represented by the value of δ which is upper bounded by the smallest eigenvalue of the design matrix. This implies that the proposed algorithm is most effective when it is applied following some model subset selection preprocessing. This is because it is common for the design matrix of a large data set to be ill-conditioned. We use the D-optimality based OFR algorithm [20] for this preprocessing.

IV. D-OPTIMALITY BASED SUBSET SELECTION

Consider the model (11) in the generic data modelling context. The least squares estimate of $\boldsymbol{\beta}_N$ is given by $\hat{\boldsymbol{\beta}}_N = \mathbf{B}_N^{-1}\boldsymbol{\Phi}_N^T\mathbf{y}$. Assume that (11) represents the true data generating process and the design matrix $\mathbf{B}_N$ is nonsingular. The estimate $\hat{\boldsymbol{\beta}}_N$ is unbiased and the covariance matrix of the estimate is proportional to the inverse of the design matrix

$$\text{Cov}\left[\hat{\boldsymbol{\beta}}_N\right] \propto \mathbf{B}_N^{-1}. \tag{24}$$

The condition number of the design matrix is given by

$$C = \frac{\max\{\sigma_i, 1 \leq i \leq N\}}{\min\{\sigma_i, 1 \leq i \leq N\}} \tag{25}$$

with σ_i, $1 \leq i \leq N$, being the eigenvalues of $\mathbf{B}_N$. Too large a condition number will result in unstable parameter estimate while a small C improves model robustness. The D-optimality design criterion [22] maximises the determinant of the design matrix for the constructed model. Specifically, let $\boldsymbol{\Phi}_{N_s}$ be a column subset of $\boldsymbol{\Phi}_N$ representing a constructed N_s-term subset model. According to the D-optimality criterion, the selected subset model is the one that maximises

$$\det\left(\boldsymbol{\Phi}_{N_s}^T\boldsymbol{\Phi}_{N_s}\right) = \det\left(\mathbf{B}_{N_s}\right). \tag{26}$$

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 $\mathbf{\Phi}_N$ be $\mathbf{\Phi}_N = \mathbf{W}_N \mathbf{R}_N$, where

$$\mathbf{R}_N = \begin{bmatrix} 1 & r_{1,2} & \cdots & r_{1,N} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{N-1,N} \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{27}$$

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

$$\begin{aligned} \det\left(\mathbf{B}_N\right) &= \det\left(\mathbf{R}_N^T\right)\det\left(\mathbf{W}_N^T \mathbf{W}_N\right)\det\left(\mathbf{R}_N\right) \\ &= \det\left(\mathbf{W}_N^T \mathbf{W}_N\right) = \prod_{i=1}^{N} \sigma_i, \end{aligned} \tag{28}$$

and

$$-\log\det\left(\mathbf{W}_N^T \mathbf{W}_N\right) = \sum_{i=1}^{N} -\log\left(\mathbf{w}_i^T \mathbf{w}_i\right). \tag{29}$$

Recall the notation $\mathbf{B}_N = [b_{i,j}] \in \mathcal{R}^{N\times N}$. The fast algorithm for the modified Gram-Schmidt orthogonalisation procedure [23] can readily be used to orthogonalise $\mathbf{B}_N$ and to calculate $\mathbf{R}_N$. For convenience, the same notation $\mathbf{B}_N$ 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.

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

- If
$$J_n > \xi \tag{30}$$
where ξ is a threshold value that determines the size of the subset model, goto *Stop*.
- Otherwise, the j_n-th column of $\mathbf{B}_N$ is interchanged from the n-th row upwards with the n-th column of $\mathbf{B}_N$, and then the j_n-th row of $\mathbf{B}_N$ is interchanged from the n-th column upwards with the n-th row of $\mathbf{B}_N$. The j_n-th column of $\mathbf{R}_N$ is interchanged up to the $(n-1)$-th row with the n-th column of $\mathbf{R}_N$.
This effectively selects the j_n-th candidate as the n-th regressor in the subset model.
- For $n+1 \leq j \leq N$, compute $r_{n,j} = b_{n,j}/b_{n,n}$, and for $n+1 \leq j \leq N$ and $j \leq l \leq N$, compute
$$\left\{ \begin{array}{l} b_{j,l} = b_{j,l} - r_{n,j} r_{n,l} b_{n,n}, \\ b_{l,j} = b_{j,l}. \end{array} \right.$$
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 subset model.

As the D-optimality based OFR algorithm is only used for preprocessing, the termination test (30) can be replaced by simply setting 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 will automatically make some of the kernel weights to (near) zero, and thus reduces the model size to an appropriate level. It can be shown that the computational complexity of this D-optimality based OFR algorithm is no more than $\mathcal{O}(N^2)$ [23].

After this preprocessing, the kernel weights are determined by solving the resulting subset nonnegative quadratic programming

$$\begin{aligned} &\min_{\boldsymbol{\beta}_{N_s}} \left\{ \tfrac{1}{2} \boldsymbol{\beta}_{N_s}^T \mathbf{A}_{N_s} \boldsymbol{\beta}_{N_s} - \mathbf{v}_{N_s}^T \boldsymbol{\beta}_{N_s} \right\}, \\ &\text{s.t. } \boldsymbol{\beta}_{N_s}^T \mathbf{1}_{N_s} = 1 \text{ and } \beta_i \geq 0, 1 \leq i \leq N_s, \end{aligned} \tag{31}$$

using the MNQP algorithm, where $\mathbf{v}_{N_s} = \mathbf{\Phi}_{N_s}^T \mathbf{y}$, $\mathbf{A}_{N_s} = \mathbf{B}_{N_s} - \delta \mathbf{I}_{N_s}$, and $\mathbf{B}_{N_s} = \mathbf{\Phi}_{N_s}^T \mathbf{\Phi}_{N_s}$ is the related subset design matrix. Furthermore, the value of δ can be set according to $\delta < \mathbf{w}_{N_s}^T \mathbf{w}_{N_s}$. Since N_s is typically very small, the computational complexity of the iterative MNQP algorithm is negligible, in comparison with $\mathcal{O}(N^2)$ of the D-optimality based OFR preprocessing.

V. Numerical examples

Two two-dimensional (2-D) and one 6-D PDF examples were used to test the proposed zero-norm constraint enhanced sparse kernel density (SKD) estimator and to compare its performance with the three kernel estimators, namely, the nonsparse PW estimator, our previous SKD estimator [18] and the RSDE estimator of [13], as well as the GMM estimator. For each case, a data set of N randomly drawn samples was used to construct density estimate, and a separate test data set of $N_{\text{test}} = 10,000$ samples was used to calculate the L_1 test error between the true density $p(\mathbf{x})$ and the resulting estimate $\hat{p}(\mathbf{x}_k; \boldsymbol{\beta}_N, \rho)$ according to

$$L_1 = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} \left| p(\mathbf{x}_k) - \hat{p}(\mathbf{x}_k; \boldsymbol{\beta}_N, \rho) \right|. \tag{32}$$

For the two 2-D examples, the Kullback-Leibler divergence (KLD), defined as

$$D_{\text{KL}}(p|\hat{p}) = \int_{\mathcal{R}^m} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x}; \boldsymbol{\beta}_N, \rho)} \, d\mathbf{x}, \tag{33}$$

was also used to validate the resulting estimates. Specifically, the KLD was approximated by partitioning the integration range $[x_{1,\min}, \ x_{1,\max}] \times [x_{2,\min}, \ x_{2,\max}]$ into the $N_{\text{par}} \times N_{\text{par}}$ small equal-area intervals and calculated

$$D_{\text{KL}}(p|\hat{p}) \approx \sum_{k=1}^{N_{\text{par}}} \sum_{l=1}^{N_{\text{par}}} p(k,l) \log \frac{p(k,l)}{\hat{p}(k,l)} \left(\Delta x\right)^2, \tag{34}$$

where $\Delta x = (x_{1,\max} - x_{1,\min})/N_{\text{par}} = (x_{2,\max} - x_{2,\min})/N_{\text{par}}$, $p(k,l) = p(x_{1,\min} + k\Delta x, x_{2,\min} + l\Delta x)$ and

TABLE I

PERFORMANCE COMPARISON OF THE PW ESTIMATOR, PREVIOUS SKD ESTIMATOR [18], RSDE ESTIMATOR [13], GMM ESTIMATOR AND PROPOSED SKD ESTIMATOR FOR THE TWO-DIMENSIONAL EXAMPLE OF GAUSSIAN AND LAPLACIAN MIXTURE, OVER 100 RUNS.

estimator	PW	previous SKD [18]	RSDE [13]	GMM	**proposed SKD**
kernel type	fixed, $\rho^{\text{Par}} = 0.42$	fixed, $\rho = 1.1$	fixed, $\rho = 1.2$	tunable	fixed, $\rho = 1.1$
L_1 test error $\times 10^3$	4.036 ± 0.693	3.838 ± 0.780	4.053 ± 0.446	3.474 ± 0.990	3.562 ± 0.692
KLC $\times 10$	1.466 ± 0.228	1.403 ± 0.534	0.896 ± 0.411	0.608 ± 0.172	1.303 ± 0.310
kernel no.	500	15.3 ± 3.9	16.2 ± 3.4	11	11.0 ± 1.5
maximum	500	25	24	11	14
minimum	500	8	9	11	8

TABLE II

PERFORMANCE COMPARISON OF THE PW ESTIMATOR, PREVIOUS SKD ESTIMATOR [18], RSDE ESTIMATOR [13], GMM ESTIMATOR AND PROPOSED SKD ESTIMATOR FOR THE TWO-DIMENSIONAL EXAMPLE OF FIVE-GAUSSIAN MIXTURE, OVER 100 RUNS.

estimator	PW	previous SKD [18]	RSDE [13]	GMM	**proposed SKD**
kernel type	fixed, $\rho^{\text{Par}} = 0.5$	fixed, $\rho = 1.1$	fixed, $\rho = 1.2$	tunable	fixed, $\rho = 1.0$
L_1 test error $\times 10^3$	3.620 ± 0.439	3.610 ± 0.503	3.631 ± 0.362	3.675 ± 0.672	3.322 ± 0.634
KLC $\times 10^2$	3.422 ± 0.548	3.665 ± 0.920	3.537 ± 0.485	3.392 ± 0.870	2.899 ± 1.087
kernel no.	500	13.2 ± 2.9	13.2 ± 3.0	8	7.8 ± 1.3
maximum	500	22	21	8	11
minimum	500	8	6	8	5

$\hat{p}(k,l) = \hat{p}(x_{1,\min} + k\Delta x, x_{2,\min} + l\Delta x; \boldsymbol{\beta}_N, \rho)$. To ensure the accuracy of the approximation, $N_{\text{par}} > 100$ was chosen. The experiment was repeated by N_{run} different random runs for each example.

The optimal values of the kernel widths, ρ^{Par} for the PW estimator and ρ for the other three SKD estimators, were found empirically via cross validation. For the GMM, the number of mixing Gaussian components, N_m, must be determined. Instead of exhaustedly trying different values for the number of mixing components based on cross validation, we simply set N_m to the average model size obtained by the proposed zero-norm constraint enhanced SKD estimator. The parameters of the GMM were determined using the EM algorithm of [10].

Example 1. The density to be estimated for this 2-D example was defined by the mixture of Gaussian and Laplacian distributions given as follows

$$p(x_1, x_2) = \frac{1}{4\pi} e^{-\frac{(x_1-2)^2}{2}} e^{-\frac{(x_2-2)^2}{2}} + \frac{0.35}{8} e^{-0.7|x_1+2|} e^{-0.5|x_2+2|}. \quad (35)$$

The estimation data set contained $N = 500$ samples, and the experiment was repeated $N_{\text{run}} = 100$ times. For the propose SKD estimator, we simply set $N_s = 16$ for the D-optimality based OFR preprocessing. Because we had an average model size of 11.0 for the proposed SKD estimate, $N_m = 11$ was used for the GMM. Table I lists the L_1 test errors and the KLD values as well as the numbers of kernels required for the five density estimates compared. For this example, the GMM estimator achieved the best test performance. The proposed SKD estimator and the RSDE estimator also did well, but the former achieved a smaller average model size.

Example 2. The true density to be estimated for this 2-D example was defined by the mixture of five Gaussian distributions given as

$$p(x, y) = \sum_{i=1}^{5} \frac{1}{10\pi} e^{-\frac{(x-\mu_{i,1})^2}{2}} e^{-\frac{(y-\mu_{i,2})^2}{2}} \quad (36)$$

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 $N = 500$, and the experiment was repeated $N_{\text{run}} = 100$ times. For the propose SKD estimator, we simply set $N_s = 14$ for the D-optimality based OFR preprocessing. Since the average model size obtained by the proposed SKD estimate was 7.8, we set $N_m = 8$ for the GMM estimator. Table II compares the performance of the five density estimators studied, where it can be seen that the proposed SKD estimator achieved the best test performance and the smallest model size, in comparison with the other two benchmark SKD estimators.

Example 3. In this 6-D example, the underlying density to be estimated was given by

$$p(\mathbf{x}) = \frac{1}{3} \sum_{i=1}^{3} \frac{1}{(2\pi)^{6/2}} \frac{1}{\det^{1/2} |\boldsymbol{\Gamma}_i|} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Gamma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)} \quad (37)$$

with

$$\begin{aligned} \boldsymbol{\mu}_1 &= [1.0\ 1.0\ 1.0\ 1.0\ 1.0\ 1.0]^T, \\ \boldsymbol{\Gamma}_1 &= \text{diag}\{1.0, 2.0, 1.0, 2.0, 1.0, 2.0\}, \end{aligned} \quad (38)$$

$$\begin{aligned} \boldsymbol{\mu}_2 &= [-1.0\ -1.0\ -1.0\ -1.0\ -1.0\ -1.0]^T, \\ \boldsymbol{\Gamma}_2 &= \text{diag}\{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}, \end{aligned} \quad (39)$$

$$\begin{aligned} \boldsymbol{\mu}_3 &= [0.0\ 0.0\ 0.0\ 0.0\ 0.0\ 0.0]^T, \\ \boldsymbol{\Gamma}_3 &= \text{diag}\{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}. \end{aligned} \quad (40)$$

TABLE III

PERFORMANCE COMPARISON OF THE PW ESTIMATOR, PREVIOUS SKD ESTIMATOR [18], RSDE ESTIMATOR [13], GMM ESTIMATOR AND PROPOSED SKD ESTIMATOR FOR THE SIX-DIMENSIONAL EXAMPLE OF THREE-GAUSSIAN MIXTURE, OVER 100 RUNS.

estimator	PW	previous SKD [18]	RSDE [13]	GMM	**proposed SKD**
kernel type	fixed, $\rho^{\mathrm{Par}} = 0.65$	fixed, $\rho = 1.2$	fixed, $\rho = 1.2$	tunable	fixed, $\rho = 1.2$
L_1 test error $\times 10^5$	3.520 ± 0.162	3.113 ± 0.534	2.739 ± 0.500	1.743 ± 0.285	2.767 ± 0.242
kernel no.	600	9.4 ± 1.9	14.2 ± 3.6	8	7.9 ± 1.3
maximum	600	16	25	8	12
minimum	600	7	8	8	5

The estimation data set was set to $N = 600$, and the experiment was repeated $N_{\mathrm{run}} = 100$ times. The results obtained by the five density estimators are summarised in Table III. For this example, it can be seen that the GMM estimator achieved the best test performance. The proposed SKD estimator and the RSDE estimator also did well, in terms of test performance. The proposed zero-norm constraint aided estimator was seen to achieve a much sparser PDF estimate than the RSDE estimator.

VI. CONCLUSIONS

We have proposed the idea of integrating the zero-norm constraint into the construction of a sparse kernel density estimator that uses the classical Parzen window estimate as the desired response. By making use of the convexity constraint for the kernel parameters and the proposed approximation function of the zero-norm, this hard problem becomes mathematically tractable and can be solved effectively using the multiplicative nonnegative quadratic programming algorithm. It is interesting to see that within the convexity constraint of kernel density estimation, the maximisation, not minimisation, of the two-norm of the kernel weights leads to model sparsity. It has been shown that the proposed approach can be benefited from preprocessing procedures to improve the condition of the kernel design matrix, and we have proposed to apply the efficient D-optimality based method for selecting a small significant subset kernel matrix. Computational complexity of the proposed sparse kernel density estimator compares favourably with other existing sparse kernel density estimators. Numerical results obtained have demonstrated that the proposed zero-norm constraint aided estimator offers an efficient means for selecting very sparse kernel density estimates with excellent generalisation performance.

REFERENCES

[1] J. Weston, A. Ellisseeff, B. Schölkopf and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *J. Machine Learning Research*, vol.3, pp.1439–1461, 2003.

[2] P.S. Bradley and O.L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 13th ICML* (San Francisco, CA, USA), 1998, pp.82–90.

[3] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

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

[5] B.W. Silverman, *Density Estimation*. London: Chapman and Hall, 1996.

[6] 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.

[7] 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.

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

[9] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: John Wiley, 2000.

[10] J.A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and hidden Markov models," *Technical Report*, ICSI-TR-97-021, University of Berkeley, 1997.

[11] 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*, MIT Press, Cambridge MA, 1999, pp.293–306.

[12] 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.

[13] 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.

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

[15] X. Hong, P.M. Sharkey and K. Warwick, "Automatic nonlinear predictive model construction algorithm using forward regression and the PRESS statistic," *IEE Proc. Control Theory and Applications*, vol.150, no.3, pp.245–254, 2003.

[16] S. Chen, X. Hong, C.J. Harris and P.M. Sharkey, "Sparse modeling using orthogonal forward regression with PRESS statistic and regularization," *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol.34, no.2, pp.898–911, 2004.

[17] 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.

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

[19] 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.

[20] S. Chen, X. Hong and C.J. Harris, "Sparse kernel density estimator using orthogonal regression based on D-optimality experimental design," in *Proc. IJCNN 2008* (Hong Kong, China), June 1-6, 2008, pp.1–6.

[21] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol.209, pp.237–260, 1998.

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

[23] 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.