International Journal of Systems Science
Vol. 43, No. 11, November 2012, 2107–2113

Using zero-norm constraint for sparse probability density function estimation

X. Hong[a]*, S. Chen[b] and C.J. Harris[b]

[a]*School of Systems Engineering University of Reading, Reading RG6 6AY, UK;* [b]*School of Electronics and Computer Science University of Southampton, Southampton SO17 1BJ, UK*

(*Received 16 July 2009; final version received 28 February 2010*)

A new sparse kernel probability density function (pdf) estimator based on zero-norm constraint is constructed using the classical Parzen window (PW) estimate as the target function. The so-called zero-norm of the parameters is used in order to achieve enhanced model sparsity, and it is suggested to minimize an approximate function of the zero-norm. It is shown that under certain condition, the kernel weights of the proposed pdf estimator based on the zero-norm approximation can be updated using the multiplicative nonnegative quadratic programming algorithm. Numerical examples are employed to demonstrate the efficacy of the proposed approach.

Keywords: cross-validation; Parzen window; probability density function; sparse modelling

1. Introduction

One of the fundamental problems in data-based nonlinear system modelling is to find the minimal model with the best model generalization performance from observational data. In linear-in-the-parameters modelling and kernel methods, the number of terms in the model is referred as the zero-norm of the parameters. Minimizing such quantity is related to variable and feature selection, ensuring model sparsity and generalization (Bradley and Mangasarian 1998; Weston, Ellisseeff, Scholkopf, and Tipping 2003). Because of the intractability in the minimization of the zero-norm, there is considerable research on the approximation schemes on the zero-norm (Bradley and Mangasarian 1998; Weston et al. 2003) and the associated computational complexities.

The estimation of the probability density function (pdf) from observed data samples is a fundamental problem in many machine learning and pattern recognition applications (Duda and Hart 1973; Silverman 1986; Bishop 1995). The Parzen window (PW) estimate is a simple yet remarkably accurate nonparametric density estimation technique (Parzen 1962; Duda and Hart 1973; Bishop 1995). A general and powerful approach to the problem of pdf estimation is the finite mixture model (McLachlan and Peel 2000). The finite mixture model includes the PW estimate as a special case in that equal weights are adopted in the PW, with the number of mixtures equal to the number of training data samples. A disadvantage associated with the PW estimate is its high computational cost of the point density estimate for a future data sample in the cases where the training data set is very large. Clearly by taking a much smaller number of mixture components, the finite mixture model can be regarded as a condensed representation of data (McLachlan and Peel 2000). Note that the mixing weights in the finite mixture model need to be determined through parametric optimization, unlike just adopting equal weights in the PW. It is desirable to develop methods of fitting a finite mixture model with the capability to infer a minimal number of mixtures from the data efficiently.

Motivated by this, there is a considerable interest in the research into the sparse pdf estimate, including support vector machine (SVM) density estimation technique (Weston et al. 1999; Vapnik and Mukherjee 2000), the reduced set density estimator (RSDE) (Girolami and He 2003). Alternatively a novel regression-based probability density estimation method has been introduced (Choudhury 2002), in which the empirical cumulative distribution function was constructed as the desired response (Weston et al. 1999). The regression-based idea of Choudhury (2002) and the approach in Hong, Sharkey, and Warwick (2003) and Chen, Hong, Harris, and Sharkey (2004b) have been extended to yield sparse density estimation algorithm based on an orthogonal forward regression (OFR) algorithm (Chen, Hong, and Harris 2004a) which is capable of automatically constructing very sparse kernel density estimate, with comparable performance to that of PW estimate. Alternatively, a simple and viable alternative approach has been

*Corresponding author. Email: x.hong@reading.ac.uk

ISSN 0020–7721 print/ISSN 1464–5319 online

http://dx.doi.org/10.1080/00207721.2011.564673
http://www.tandfonline.com

proposed to use the kernels directly as regressors and the target response as PW estimate (Chen, Hong, and Harris 2008).

Following the idea in Chen et al. (2008) of using PW estimate as the target function and based on zero-norm constraint, this article introduces a new sparse kernel pdf estimator. It is suggested to minimize an approximate function of the zero-norm of the kernel weights vector. It is analyzed that under certain condition, 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.

2. The kernel density estimator

Given a finite data set consisting of N data samples, $D=\{\mathbf{x}_1,\ldots,\mathbf{x}_j,\ldots,\mathbf{x}_N\}$, where the feature vector variable $\mathbf{x}_j\in\Re^m$ follows an unknown pdf $p(\mathbf{x})$, the problem under study is to find a sparse approximation of $p(\mathbf{x})$ based on D.

A general kernel-based density estimate of $p(\mathbf{x})$ is given by

$$\hat{p}(\mathbf{x};\boldsymbol{\beta},\rho)=\sum_{j=1}^{N}\beta_j K_\rho(\mathbf{x},\mathbf{x}_j),\quad \text{subject to}\quad \beta_j\geq 0,\ j=1,\ldots,N,\ \boldsymbol{\beta}^T\mathbf{1}=1, \tag{1}$$

where $\boldsymbol{\beta}=[\beta_1,\beta_2,\ldots,\beta_N]^T$. β_j's are the kernels weights. $\mathbf{1}$ is a vector with an appropriate dimension and all elements as ones. $K_\rho(\mathbf{x},\mathbf{x}_j)$ is a chosen kernel function with kernel width ρ. In this study,

$$K_\rho(\mathbf{x},\mathbf{x}_j)=\frac{1}{(2\pi\rho^2)^{m/2}}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_j\|^2}{2\rho^2}\right) \tag{2}$$

is used. Let the well-known PW estimator be denoted by $\hat{p}(\mathbf{x};\boldsymbol{\beta}^{\mathrm{Par}},\rho^{\mathrm{Par}})$, where $\boldsymbol{\beta}^{\mathrm{Par}}=[\beta_1^{\mathrm{Par}},\ldots,\beta_N^{\mathrm{Par}}]^T$, $\beta_j^{\mathrm{Par}}=\frac{1}{N}\ \forall j$. The log-likelihood for $\boldsymbol{\beta}$ can be formed using observed data D as $\log L$ as

$$\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}(\mathbf{x}_i;\boldsymbol{\beta},\rho)=\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{j=1}^{N}\beta_j K_\rho(\mathbf{x}_i,\mathbf{x}_j)\right). \tag{3}$$

Note that by the law of large numbers the log-likelihood of (3) tends to

$$\int_{\Re^m} p(\mathbf{x})\log\hat{p}(\mathbf{x};\boldsymbol{\beta},\rho)\mathrm{d}\mathbf{x}, \tag{4}$$

as $N\to\infty$ with probability one. Equation (4) is simply the negative cross-entropy or divergence between the true density $p(\mathbf{x})$ and the estimate $\hat{p}(\mathbf{x};\boldsymbol{\beta},\rho)$. It can be shown that the PW estimator $\beta_j^{\mathrm{Par}}=\frac{1}{N}\ \forall j$ can be obtained as an optimal estimator via the maximization of (3) with respective to $\boldsymbol{\beta}$ subject to the constraints $\beta_j\geq 0$, $j=1,\ldots,N$, $\boldsymbol{\beta}^T\ \mathbf{1}=1$. Note that the choice of ρ^{Par} is crucial in density estimation using PW (Silverman 1986). Based on the principle of minimizing the mean integrated square error (MISE) (Silverman 1986), ρ^{Par} can be found so as to minimize the least squares cross validation criterion $M(\rho)$ given by Silverman (1986)

$$\begin{aligned}&\frac{1}{N^2}\sum_{i,j=1}^{N}K_{\sqrt{2}\rho}(\mathbf{x}_i,\mathbf{x}_j)-\frac{2}{N(N-1)}\sum_{i,j=1,j\neq i}^{N}K_\rho(\mathbf{x}_i,\mathbf{x}_j)\\&\quad\approx\frac{1}{N^2}\sum_{i,j=1}^{N}K_\rho^*(\mathbf{x}_i,\mathbf{x}_j)+\frac{2}{N(2\pi\rho^2)^{m/2}},\end{aligned} \tag{5}$$

where $K_\rho^*(\mathbf{x}_i,\mathbf{x}_j)=K_{\sqrt{2}\rho}(\mathbf{x}_i,\mathbf{x}_j)-2K_\rho(\mathbf{x}_i,\mathbf{x}_j)$. The computational cost of finding ρ^{Par} is $O(N^2)$, this is scaled by the number of grid search set by the user.

With the PW estimator, the associated computational cost for evaluating the probability density estimate for a future sample scales directly with the sample size N. Therefore it is desirable to devise a sparse representation of $\hat{p}(\mathbf{x};\boldsymbol{\beta},\rho)$, in which the terms are composed of a small subset of data samples.

Clearly any good sparse kernel density estimator $\hat{p}(\mathbf{x};\boldsymbol{\beta},\rho)$ should be devised as close as possible to the unknown true density $p(\mathbf{x})$. Because the PW estimators have the property of optimality, it was suggested (Chen et al. 2008) that it is possible to use the PW estimator as the target of the proposed sparse kernel density estimator. Specifically we can write a regression equation linking $\hat{p}(\mathbf{x};\boldsymbol{\beta},\rho)$ and $\hat{p}(\mathbf{x};\boldsymbol{\beta}^{\mathrm{Par}},\rho^{\mathrm{Par}})$ as

$$\hat{p}(\mathbf{x};\boldsymbol{\beta}^{\mathrm{Par}},\sigma^{\mathrm{Par}})=\sum_{j=1}^{N}\beta_j K_\rho(\mathbf{x},\mathbf{x}_j)+\varepsilon(\mathbf{x}), \tag{6}$$

where $\varepsilon(\mathbf{x})$ is the modelling error at $\mathbf{x}$ between the sparse kernel density estimator $\hat{p}(\mathbf{x};\boldsymbol{\beta},\rho)$ and the PW density estimator $\hat{p}(\mathbf{x};\boldsymbol{\beta}^{\mathrm{Par}},\rho^{\mathrm{Par}})$ that is initially constructed based on D. The aims are to obtain β_j via minimizing some modelling error criterion, e.g. $E[\varepsilon^2(\mathbf{x})]$, and simultaneously to achieve a sparse representation of $\hat{p}(\mathbf{x};\boldsymbol{\beta},\sigma)$ (with most elements in $\boldsymbol{\beta}$ being zeros in (6)) subject to the constraints $\beta_j\geq 0$, $j=1,\ldots,N$, $\boldsymbol{\beta}^T\ \mathbf{1}=1$.

Define $y_k=\hat{p}(\mathbf{x}_k;\boldsymbol{\beta}^{\mathrm{Par}},\sigma^{\mathrm{Par}})$, $\boldsymbol{\phi}(k)=[K_{k,1}\ \ K_{k,2}\cdots K_{k,N}]^T$ with $K_{k,i}=K_\rho(\mathbf{x}_k,\ \mathbf{x}_i)$ and $\varepsilon(k)=\varepsilon(\mathbf{x})$, then model (6) at data point $\mathbf{x}_k\in D$ can be expressed as

$$y_k=\hat{y}_k+\varepsilon(k)=\phi^T(k)\boldsymbol{\beta}+\varepsilon(k). \tag{7}$$

Over the training data set D, model (6) can be written in the matrix form

$$\mathbf{y}=\Phi\boldsymbol{\beta}+\boldsymbol{\varepsilon}, \tag{8}$$

with the following additional notations $\Phi = [K_{k,i}] \in \Re^{N\times N}$, $1 \le i,\ k \le N$, $\boldsymbol{\varepsilon} = [\varepsilon(1)\ \ \varepsilon(2)\ \cdots\ \varepsilon(N)]^T$ and $\mathbf{y} = [y(1)\ \ y(2)\ \cdots\ y(N)]^T$. The kernel weights vector $\boldsymbol{\beta}$ can be obtained by solving the following constrained nonnegative quadratic programming

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\boldsymbol{\beta}^T \mathbf{B}\boldsymbol{\beta} - \mathbf{v}^T\boldsymbol{\beta} \right\}, \tag{9}$$

$$\text{s.t. } \boldsymbol{\beta}^T\mathbf{1} = 1 \quad \text{and} \quad \beta_j \ge 0,\ j = 1, \ldots, N, \tag{10}$$

where $\mathbf{B} = \Phi^T\Phi$ is the related design matrix and $\mathbf{v} = \Phi^T\mathbf{y}$. The solution can be readily solved using an iterative MNQP algorithm (Sha, Saul, and Lee 2002; Girolami and He 2003; Chen et al. 2008). Denote $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_j, \ldots, \boldsymbol{\phi}_N]$, where $\boldsymbol{\phi}_j$ denotes column vectors of $\boldsymbol{\Phi}$. Note the row vectors of $\boldsymbol{\Phi}$ is denoted by $\boldsymbol{\phi}(k)$.

3. Sparse pdf estimation using zero-norm constraint

In order to improve the sparsity of model (6), the quantity $\|\boldsymbol{\beta}\|_0$, that counts the number of nonzero entries in $\boldsymbol{\beta}$ and is referred to as zero-norm of $\boldsymbol{\beta}$, can be utilised as an additional constraint (Bradley and Mangasarian 1998; Weston et al. 2003). It is a very hard problem to directly minimize the zero-norm of $\boldsymbol{\beta}$ (Amaldi and Kann 1998; Weston et al. 2003), so the work of Bradley and Mangasarian (1998) proposed an approximate approach with

$$\|\boldsymbol{\beta}\|_0 \approx \sum_{i=1}^{N}\left(1 - e^{-\alpha|\beta_i|}\right), \tag{11}$$

in which $\alpha > 0$ is a properly chosen parameter. ($\|\boldsymbol{\beta}\|_0 \to 0$ if $\alpha \to 0$, and $\|\boldsymbol{\beta}\|_0 \to N$ if $\alpha \to +\infty$.) Following the idea in Bradley and Mangasarian (1998), the objective function (9) can be modified to yield

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\boldsymbol{\beta}^T \mathbf{B}\boldsymbol{\beta} - \mathbf{v}^T\boldsymbol{\beta} \right\} + \lambda \sum_{i=1}^{N}\left(1 - e^{-\alpha|\beta_i|}\right),$$
$$\text{s.t. } \boldsymbol{\beta}^T\mathbf{1} = 1 \quad \text{and} \quad \beta_j \ge 0,\ j = 1, \ldots, N, \tag{12}$$

where $\lambda > 0$ is a small parameter that regulates the tradeoff between the two objectives. Here we propose a further approximation by using the Taylor series expansion up to the second order, such that

$$e^{-\alpha|\beta_i|} \approx 1 - \alpha|\beta_i| + \frac{\alpha^2\beta_i^2}{2}, \tag{13}$$

and

$$\sum_{i=1}^{N}\left(1 - e^{-\alpha|\beta_i|}\right) \approx \alpha\sum_{i=1}^{N}|\beta_i| - \frac{\alpha^2}{2}\sum_{i=1}^{N}\beta_i^2. \tag{14}$$

Applying the constraint $\boldsymbol{\beta}^T\mathbf{1} = 1$ and $\beta_j \ge 0, j = 1, \ldots, N$ to (14), we obtain

$$\sum_{i=1}^{N}\left(1 - e^{-\alpha|\beta_i|}\right) \approx \alpha - \frac{\alpha^2}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}. \tag{15}$$

Based on (15), (12) can be approximately reformulated

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\boldsymbol{\beta}^T \mathbf{A}\boldsymbol{\beta} - \mathbf{v}^T\boldsymbol{\beta} \right\}, \tag{16}$$

$$\text{s.t. } \boldsymbol{\beta}^T\mathbf{1} = 1 \quad \text{and} \quad \beta_j \ge 0,\ j = 1, \ldots, N, \tag{17}$$

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

Provided that δ is set in a manner that $\mathbf{A}$ is a positive-definite matrix, the solution to the above can be readily solved using an iterative MNQP algorithm (Sha et al. 2002; Girolami and He 2003; Chen et al. 2008) as that of (9).

Lemma 1: *Assuming that* $\mathbf{B}$ *is full rank, the* condition *for* $\mathbf{A}$ *to be positive definite matrix is* $\delta < \sigma_N = \sigma_{\min}$, *where* $\sigma_{\min}$ *is the smallest eigenvalue of* $\mathbf{B}$.

Proof: Consider the singular value decomposition (SVD) of matrix $\mathbf{B}$ with orthonormal matrix $\mathbf{Q} \in \Re^{N\times N}$, such that

$$\mathbf{Q}^T\mathbf{B}\mathbf{Q} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N) \in \Re^{N\times N}, \tag{18}$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_N$ are N nonnegative singular values of $\mathbf{B}$. Applying $\mathbf{A} = \mathbf{B} - \delta\mathbf{I}$ and $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, we obtain

$$\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathrm{diag}(\sigma_1', \sigma_2', \ldots, \sigma_N') \in \Re^{N\times N}, \tag{19}$$

where $\sigma_i' = \sigma_i - \delta > 0$, $i = 1, 2\ldots, N$ are eigenvalues of $\mathbf{A}$. This concludes the proof.

For completeness, the MNQP algorithm for solving (16) is described below (Girolami and He 2003; Chen et al. 2008). For convenience, denote $\mathbf{A} = [a_{i,j}]$, $\mathbf{v} = [v_1 \cdots v_N]^T$. Since the elements of $\mathbf{A}$ and $\mathbf{v}$ are strictly positive, the Lagrangian for the above problem can be formed as Girolami and He (2003)

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}\frac{\beta_j^{(t)}(\beta_i^{(t+1)})^2}{\beta_i^{(t)}} - \sum_{i=1}^{N} v_i\beta_i^{(t+1)} - h^{(t)}\left(\sum_{i=1}^{N}\beta_i^{(t+1)} - 1\right), \tag{20}$$

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

$$\frac{\partial\mathcal{L}}{\partial\beta_i^{(t+1)}} = 0 \quad \text{and} \quad \frac{\partial\mathcal{L}}{\partial h^{(t)}} = 0 \tag{21}$$

yields the following updating equations:

$$c_i^{(t)} = \beta_i^{(t)} \left(\sum_{j=1}^{N} a_{i,j} \beta_j^{(t)} \right), \quad 1 \le i \le N, \tag{22}$$

$$h^{(t)} = \left(\sum_{j=1}^{N} c_i^{(t)} \right)^{-1} \left(1 - \sum_{i=1}^{N} c_i^{(t)} v_i \right), \tag{23}$$

$$\beta_i^{(t+1)} = c_i^{(t)} (v_i + h^{(t)}). \tag{24}$$

The initial condition can be set as $\beta_i^{(0)} = \frac{1}{N}$, $1 \le i \le N$. It is easy to verify that (17) is maintained over the iterations. Over the iterations, some of the kernel weights are driven to near zero, and the corresponding kernels can be removed from the model (6).

Remarks:

(1) The general problem of minimising $\|\boldsymbol{\beta}\|_q$ for $0 < q < 1$ is nonconvex (hard). With the two successive approximation step of (11) and (13), the computational difficulty is alleviated by changing the problem to a convex optimisation problem.

(2) From (15), it is seen that the minimisation of the proposed zero-norm approximation, combined with the convexity constraint of the kernel parameter vector, 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 condition (17), the model sparsity is equivalent to the unevenness in the distribution of the parameters magnitude, e.g. the two-norm of the parameters is maximised as one when $\beta_k = 1$, and $\beta_j = 0$ for $\forall j \neq k$, $j = 1, \ldots, N$, corresponding to the smallest zero-norm of 1 when the parameters are the most unevenly distributed. The two-norm of the parameters is minimised as $\frac{1}{N}$ when $\beta_j = \frac{1}{N}$, $j = 1, \ldots, N$, which corresponds to the largest zero-norm of N when the parameters are uniformly distributed. This leads to nonsparse estimate.

(3) Remark 2 is interesting as it shows that the maximal, not the minimal, of the two-norm of the parameters leads to model sparsity. It is worth noting that whenever other constraints to the parameters are present, only the minimal of zero-norm can be considered as the natural measure of model sparsity without the need of further mathematical proof.

(4) The strength of the zero norm constraint is represented by a high value of δ which is upper bounded by the smallest eigenvalue of the design matrix according to Lemma 1. This implies that the proposed algorithm may be most effective when it is applied following some model subset selection preprocessing procedures. This is because it is common for the design matrix of a large data set to be ill-conditioned. Alternatively the proposed algorithm can be applied by gradually increasing δ from 0 to a higher value, while the kernels with close to zero parameters are removed over the iterations to improve the condition of the updated design matrix using only a subset of the kernels. In this article we used the forward D-optimality criteria subset selective algorithm (Appendix and Chen, Hong, and Harris 2010) followed by applying the proposed algorithm.

4. Illustrative examples

In the following examples, a data set of N points was randomly drawn from a given distribution described below ($N = 500$ in Example 1 and $N = 600$ in Example 2). This was used to construct the sparse pdf $\hat{p}(\mathbf{x}; \mathbf{g}, \sigma)$ using the proposed MNQP approach based on zero-norm constraint, following the preprocessing using the forward D-optimality criteria subset selective algorithm (Appendix). For each example, the experiment was repeated for 100 different random runs. For each random run, a separate test data set of $N_{\text{test}} = 10{,}000$ points was used for evaluation according to

$$L_1 = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} |p(\mathbf{x}_k) - \hat{p}(\mathbf{x}_k; \mathbf{g}, \sigma)|. \tag{25}$$

In both examples, ρ^{Par} was found using a coarse grid search based on (5) for each of 100 training data sets. Then $\rho = \gamma \rho^{\text{Par}}$, where γ was empirically set based on the first data set to save computational cost. For the forward D-optimality criteria subset selective algorithm (Appendix), a predetermined $n_s = 30$ was set for all runs to obtain subset models, then for each of the subset models the smallest eigenvalue was found and set as δ. The other four methods used for comparison are (a) the PW estimate; (b) the sparse density construction (SDC) algorithm (Chen et al. 2004a); (c) the sparse Kernel density construction (SKD) algorithm (Chen et al. 2008) and (d) the RSDE-MNQP (Sha et al. 2002; Girolami and He 2003). The results of the proposed method in comparison with

Table 1. Performance of kernel density estimates for Examples 1 and 2.

Method	L_1 test error (mean ± STD) (mean ± STD)	Kernel numbers (mean ± STD)
Example 1		
PW	$(4.2 \pm 0.8) \times 10^{-3}$	500 ± 0
SDC (Chen et al., 2004a)	$\mathbf{(3.8 \pm 0.8) \times 10^{-3}}$	$\mathbf{11.9 \pm 2.6}$
SKD (Chen et al., 2008)	$(3.8 \pm 0.8) \times 10^{-3}$	15.3 ± 3.9
RSDE-MNQP (Hong, Chen, and Harris 2008)	$(4.2 \pm 0.8) \times 10^{-3}$	129.4 ± 35.7
Proposed method	$(3.9 \pm 0.9) \times 10^{-3}$	21.7 ± 2.1
Example 2		
PW	$(3.2 \pm 0.1) \times 10^{-5}$	600 ± 0
SDC (Chen et al., 2004a)	$(4.5 \pm 1.2) \times 10^{-5}$	14.9 ± 2.1
SKD (Chen et al., 2008)	$(3.1 \pm 0.5) \times 10^{-5}$	$\mathbf{9.4 \pm 1.9}$
RSDE-MNQP (Hong et al., 2008)	$(3.7 \pm 0.7) \times 10^{-5}$	29.4 ± 10.1
Proposed method	$\mathbf{(2.9 \pm 0.2) \times 10^{-5}}$	11.8 ± 2.1

Note: Bold values denote the best results.

other approaches are shown in Table 1, where the results of the SDC, SKD, RSDE-MNQP are quoted from Chen et al. (2004a), Chen et al. (2008) and Hong et al. (2008).

Example 1: The density to be estimated for this 2-D example was given by the mixture of two densities of a Gaussian and a Laplacian, as defined by

$$p(\mathbf{x}) = \frac{1}{4\pi}\exp\left(-\frac{(x_1-2)^2}{2}\right)\exp\left(-\frac{(x_2-2)^2}{2}\right) + \frac{0.35}{8}\exp(-0.7|x_1+2|)\exp(-0.5|x_2+2|). \tag{26}$$

Example 2: The density to be estimated for this 6-D example was defined by

$$p(\mathbf{x}) = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{(2\pi)^3\sqrt{\det(\Gamma_i)}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Gamma}_i^{-1}(\mathbf{x}-\mu_i)\right), \tag{27}$$

with $\boldsymbol{\mu}_1 = [1.0, 1.0, 1.0, 1.0, 1.0, 1.0]^T$, $\boldsymbol{\mu}_2 = [-1.0, -1.0, -1.0, -1.0, -1.0, -1.0]^T$, $\boldsymbol{\mu}_3 = [0, 0, 0, 0, 0, 0]^T$, $\boldsymbol{\Gamma}_1 = \mathrm{diag}\{1.0, 2.0, 1.0, 2.0, 1.0, 2.0\}$ $\boldsymbol{\Gamma}_2 = \mathrm{diag}\{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}$ and $\boldsymbol{\Gamma}_3 = \mathrm{diag}\{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}$.

Note that the involved computational cost is mainly from the forward D-optimality criteria subset selective algorithm (Appendix) of $O(N^2)$. The computational cost in the MNQP is negligible in comparison to $O(N^2)$. From the results in Table 1, it is shown that the proposed method has comparable accuracy to PW other sparse pdf estimators, and is effective in building sparse pdf models.

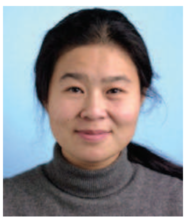
5. Conclusions

In this article, we proposed the idea of integrating the zero-norm constraint into the construction of a sparse kernel pdf estimator that uses the classical PW estimate as the target function. By making use of the convexity constraint for the kernel parameters and the proposed approximation function of the zero-norm, this hard problem is alleviated and approximately formulated as a simple quadratic programming problem, which lends itself to the utilisation of the MNQP algorithm. Some analysis are provided to suggest that the proposed approach can be benefited from preprocessing procedures to improve the condition of the kernel matrix. The simulation experiments are based on the proposed algorithm, following a preprocessing stage using the forward D-optimality criteria subset selective algorithm as described in the appendix.

Acknowledgements

The authors would like to thank the reviewers for the helpful comments.

Notes on contributors



Xia Hong received her BSc and MSc degrees in 1984 and 1987, respectively, at National University of Defense Technology, P.R. China, and PhD in 1998 from University of Sheffield, UK, all in Automatic Control. She worked as a research assistant in Beijing Institute of Systems Engineering, Beijing, China from 1987 to 1993. She worked as a research fellow in the Department of Electronics and Computer Science at University of Southampton from 1997 to 2001. She is currently a Reader at School of Systems Engineering, University of Reading. She is actively engaged in research into nonlinear systems identification, data modelling, estimation and intelligent control, neural networks, pattern recognition, learning theory and their applications. She has published over 100 research papers, and co-authored a research book. She was awarded a Donald Julius Groen Prize by IMechE in 1999.

Sheng Chen obtained his BEng degree in Control Engineering from the East China Petroleum Institute in 1982, and a PhD degree in control engineering from the City University at London in 1986. He joined the University of Southampton in September 1999. He previously held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth. Prof Chen is Fellow of IEEE in the USA. His recent research works include adaptive nonlinear signal processing, modelling and identification of nonlinear systems, neural network research, finite-precision digital controller design, evolutionary computation methods and optimisation. He has published over 300 research papers.

Chris Harris received his BSc degree at Leicester, Oxford (MA) and PhD at Southampton. He previously held appointments at the Universities of Hull, UMIST, Oxford and Cranfield, as well as being employed by the UK Ministry of Defence. His research interest areas are intelligent and adaptive systems theory and its application to intelligent autonomous systems, management infrastructures, intelligent control and estimation of dynamic processes, multi-sensor data fusion and systems integration. He has authored or co-authored 12 books and over 400 research papers, and he was the associate editor of numerous international journals including *Automatica*, *Engineering Applications of AI*, *International Journal of General Systems Engineering*, *International Journal of System Science* and the *International Journal on Mathematical Control and Information Theory*. He was elected to the Royal Academy of Engineering in 1996, was awarded the IEE Senior Achievement medal in 1998 for his work on autonomous systems, and the highest international award in IEE, the IEE Faraday medal in 2001 for his work in Intelligent Control and Neurofuzzy System.

References

Amaldi, E., and Kann, V. (1998), 'On the Approximability of Minimising Non-zero Variables or Unsatisfied Relations in Linear Systems', *Theoretical Computer Science*, 209, 237–260.

Atkinson, A.C., and Donev, A.N. (1992), *Optimum Experimental Designs*, Oxford: Clarendon Press.

Bishop, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford, UK: Oxford University Press.

Bradley, P.S., and Mangasarian, O.L. (1998), 'Feature Selelction via Concave Minimisation and Support Vector Machines', in *Proceedings of the 13th ICML*, San Francisco, CA, pp. 82–90.

Chen, S., Billings, S.A., and Luo, W. (1989), 'Orthogonal Least Squares Methods and Their Applications to Non-linear System Identification', *International Journal of Control*, 50, 1873–1896.

Chen, S., Hong, X., and Harris, C.J. (2004a), 'Sparse Kernel Density Construction Using Orthogonal Forward Regression with Leave-One-Out Test Score and Local Regularization', *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 34, 1708–1717.

Chen, S., Hong, X., and Harris, C.J. (2008), 'An Orthogonal Forward Regression Technique for Sparse Kernel Density Estimation', *Neurocomputing*, 71, 931–943.

Chen, S., Hong, X., and Harris, C.J. (2010), 'Regression Based D-optimality Experimental Design for Sparse Kernel Density Estimation', *Neurocomputing*, 73, 727–739.

Chen, S., Hong, X., Harris, C.J., and Sharkey, P.M. (2004b), 'Sparse Modelling Using Orthogonal Forward Regression with PRESS Statistic and Regularization', *IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics*, 34, 898–911.

Choudhury, A. (2002), 'Fast Machine Learning Algorithms for Large Data', Ph.D. Thesis, School of Engineering Sciences, University of Southampton.

Duda, R.O., and Hart, P.E. (1973), *Pattern Classification and Scene Analysis*, New York: John Wiley & Sons.

Girolami, M., and He, C. (2003), 'Probability Density Estimation from Optimally Condensed Data Samples', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1253–1264.

Hong, X., Chen, S., and Harris, C.J. (2008), 'A Forward-constrained Regression Algorithm for Sparse Kernel Density Estimation', *IEEE Transactions on Neural Networks*, 19, 193–198.

Hong, X., Sharkey, P.M., and Warwick, K. (2003), 'Automatic Nonlinear Predictive Model Construction Using Forward Regression and the PRESS Statistic', *IEE Proceedings Control Theory Applications'*, 150, 245–254.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley & Sons.

Parzen, E. (1962), 'On Estimation of a Probability Density Function and Mode', *The Annals of Mathematical Statistics*, 33, 1065–1076.

Sha, F., Saul, L.K., and Lee, D.D. (2002), 'Multiplicative Updates for Non-negative Quadratic Programming in Support Vector Machines', Technical Report, MS-CIS-02-09, University of Pennsylvania, USA.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Vapnik, V., and Mukherjee, S. (2000), 'Support Vector Machine for Multivariate Density Estimation', in *Advances in Neural Information Processing Systems*, eds. T. Leen, S. Solla, and K.R. Müller, Cambridge, MA, USA: MIT Press, pp. 659–665.

Weston, J., Ellisseeff, A., Scholkopf, B., and Tipping, M. (2003), 'Use of the Zero-norm with Linear Models and Kernel Methods', *Journal of Machine Learning Research*, 3, 1439–1461.

Weston, J., Gammerman, A., Stitson, M.O., Vapnik, V., Vovk, V., and Watkins, C. (1999), 'Support Vector Density Estimation', in *Advances in Kernel Methods*, eds. C. Burges, B. Schölkopf, and A.J. Smola, Cambridge, MA, USA: MIT Press, pp. 293–306.

Appendix

The forward D-optimality criteria subset selective algorithm

The D-optimality design criterion (Atkinson and Donev 1992) can be applied as a model selection criterion that maximizes the determinant of the design matrix defined as $\mathbf{B}_k = \mathbf{\Phi}_k^T \mathbf{\Phi}_k$, where $\mathbf{\Phi}_k \in \Re^{N \times n_s}$ denotes the resultant regression matrix, consisting of n_s regressors selected from N regressors in $\mathbf{\Phi}$. Specifically

$$\max \left\{ J_D = \det(\mathbf{B}_k) = \prod_{k=1}^{n_s} \sigma_k \right\}, \tag{28}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n_s}$ are n_s nonnegative singular values of $\mathbf{B}_k$. The same notation as (18) is used for simplicity.

An orthogonal decomposition of $\mathbf{\Phi}$ is

$$\mathbf{\Phi} = \mathbf{WR}, \tag{29}$$

where

$$\mathbf{R} = \begin{bmatrix} 1 & r_{1,2} & \cdots & r_{1,N} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{N-1,N} \\ 0 & \cdots & 0 & 1 \end{bmatrix}, \tag{30}$$

and $\mathbf{W}$ is an $N \times N$ matrix with orthogonal columns that satisfy

$$\mathbf{W}^T \mathbf{W} = \mathrm{diag}\{\kappa_1, \ldots, \kappa_j, \ldots, \kappa_N\}, \tag{31}$$

with

$$\kappa_j = \mathbf{w}_j^T \mathbf{w}_j, \quad j = 1, \ldots, N. \tag{32}$$

Denote $\mathbf{\Phi}_k = \mathbf{W}_k \mathbf{R}_k$, where $\mathbf{R}_k$ is a unit upper triangular matrix based on the orthogonal triangularisation of $\mathbf{\Phi}_k$. Because

$$\begin{aligned} \det(\mathbf{B}_k) &= \det(\mathbf{R}_k^T) \det(\mathbf{W}_k^T \mathbf{W}_k) \det(\mathbf{R}_k) \\ &= \det(\mathbf{W}_k^T \mathbf{W}_k) = \prod_{j=1}^{k} \kappa_k \end{aligned} \tag{33}$$

due to $\det(\mathbf{R}_k) = 1$, the selection of a subset of $\mathbf{\Phi}_k$ from $\mathbf{\Phi}$ for higher value of J_D can be achieved via the maximization of $\prod_{j=1}^{k} \kappa_k$.

The forward D-optimality criteria subset selective algorithm outlined below involves selecting a set of n_s kernels (regressors) $\mathbf{\Phi}_k = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_k]$, $k = 1, \ldots, n_s$, from N kernels to form a set of orthogonal basis $\mathbf{w}_k$, $k = 1, \ldots, n_s$ (e.g. via using the modified Gram–Schmidt orthogonalisation procedure (Chen, Billings, and Luo 1989) in a forward selective manner. At the kth selection, a candidate regressor is selected as the kth basis of the subset if it produces the largest value of κ_k from the remaining $(N - k + 1)$ candidates. The variable selection is terminated when

$$\kappa_{n_s} < \rho, \tag{34}$$

where ρ is a preset small positive number. Because this algorithm is used at the preprocessing stage, the choice of ρ can be coarse, e.g. it can be set so that n_s is sufficiently large.

Note that if the above algorithm is used as preprocessing, then in (9) and (16), the associated matrices $\mathbf{B}$ is replaced by $\mathbf{B}_k$, and $\mathbf{v}$ is replaced by $\Phi_k^T \mathbf{y}$ and the matrix dimensions and entries in (22)–(24) are adjusted accordingly.