Kernel Density Construction Using Orthogonal Forward Regression

S. Chen[†], X. Hong[‡] and C.J. Harris[†]

[†] School of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, U.K.
E-mails: sqc@ecs.soton.ac.uk cjh@ecs.soton.ac.uk

[‡] Department of Cybernetics
University of Reading, Reading, RG6 6AY, U.K.
E-mail: x.hong@reading.ac.uk

Presented at 5th International Conference on Intelligent Data Engineering and Automated Learning, Exeter, U.K., August 25-27, 2004

Overview

◯ Density estimation is a recurrent theme in machine learning and many fields of engineering — It is a hard, ill-posed and unsupervise "learning problem"

◯ Non-parametric techniques

Parzen window estimate: remarkably simple and accurate but non-sparse

SVM based sparse kernel density estimation technique

Related reduced data density estimation technique

◯ This contribution proposes a sparse kernel density construction based on orthogonal forward regression — an efficient technique widely used in parsimonious data modelling

Kernel Density Estimation as Regression

◯ Estimate unknown PDF $p(\mathbf{x})$ from finite sample set $\mathcal{D} = \{\mathbf{x}_k\}_{k=1}^N$ using kernel model

$$\hat{p}(\mathbf{x}) = \sum_{k=1}^{N} \beta_k K(\mathbf{x}, \mathbf{x}_k)$$

where $\mathbf{x}_k = [x_{1,k} \cdots x_{m,k}]^T \in \mathcal{R}^m$, with constraints

$$\beta_k \geq 0, \ 1 \leq k \leq N; \quad \sum_{k=1}^{N} \beta_k = 1$$

◯ Define empirical distribution function

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

where $\theta(x) = 1$ if $x > 0$ and $\theta(x) = 0$ if $x \leq 0$, and “regressor”

$$q(\mathbf{x}, \mathbf{x}_k) = \int_{-\infty}^{\mathbf{x}} K(\mathbf{u}, \mathbf{x}_k) \, d\mathbf{u}$$

Regression Modelling (continue)

◯ This leads to regression model

$$\mathbf{f} = \mathbf{\Phi}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{f} = [f_1 \cdots f_N]^T$ with $f_k = f(\mathbf{x}_k; N)$, $\boldsymbol{\beta} = [\beta_1 \cdots \beta_N]^T$
$\mathbf{\Phi} = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_N]$ with $\boldsymbol{\phi}_k = [q_{1,k} \cdots q_{N,k}]^T$ and $q_{i,k} = q(\mathbf{x}_i, \mathbf{x}_k)$

◯ Let orthogonal decomposition

$$\mathbf{\Phi} = \mathbf{W}\,\mathbf{A}$$

where orthogonal matrix $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_N]$ has orthogonal columns

◯ Orthogonal regression model

$$\mathbf{f} = \mathbf{W}\,\mathbf{g} + \boldsymbol{\epsilon}$$

with $\mathbf{g} = \mathbf{A}\,\boldsymbol{\beta}$

Sparse Density Construction

○ Effectively becomes a sparse regression modeling

○ Efficient orthogonal forward selection algorithm to select a subset model:

Incrementally minimize leave-one-out test error, a direct measure of model generalization ability

Multiple-regularizer or local regularization further enforce model sparsity

Automatically construct a sparse subset model (user does not need to specify any algorithmic parameters)

○ Details in: 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, August 2004.

A Two-Dimensional Example

- Density to be estimated:

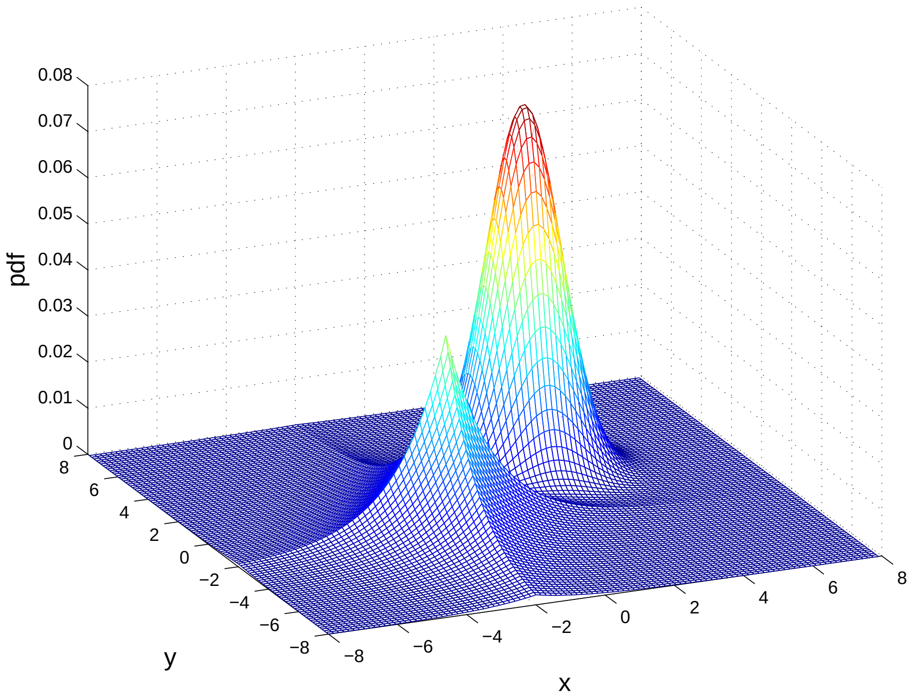
$$p(x,y) = 0.5\frac{1}{2\pi}e^{-\frac{(x-2)^2}{2}}\, e^{-\frac{(y-2)^2}{2}} + 0.5\frac{0.35}{4}e^{-0.7|x+2|}\, e^{-0.5|y+2|}$$

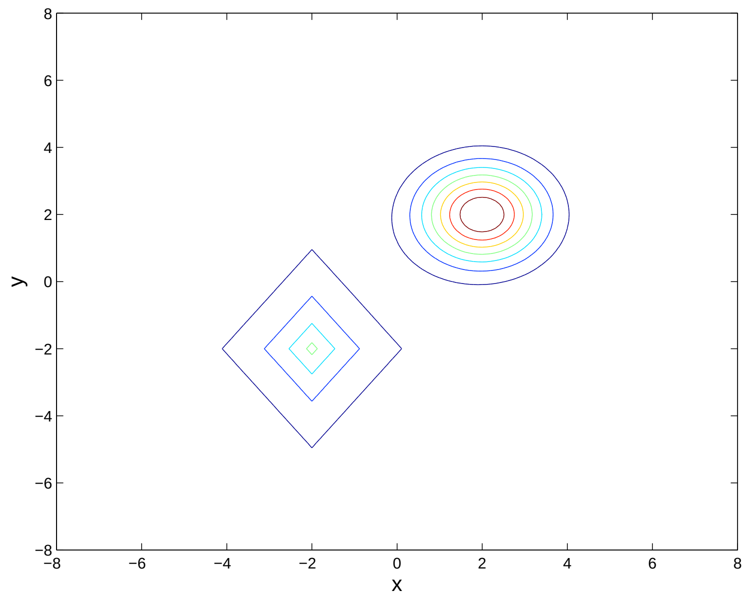
- Estimation set: 500 samples, Test set for calculate L_1 error: 10000 samples, Gaussian kernel used

- Mean and standard deviation for 100 experiments

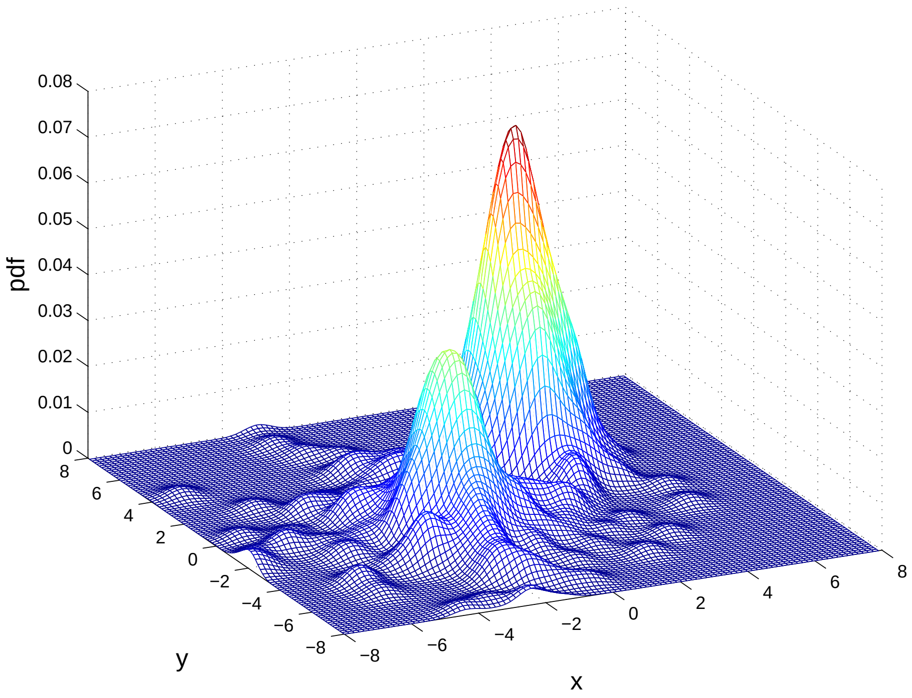
method	L_1 test error	kernel number
PW	$(4.084 \pm 0.779) \times 10^{-3}$	500 ± 0
SDC	$(3.628 \pm 0.826) \times 10^{-3}$	11.9 ± 2.6

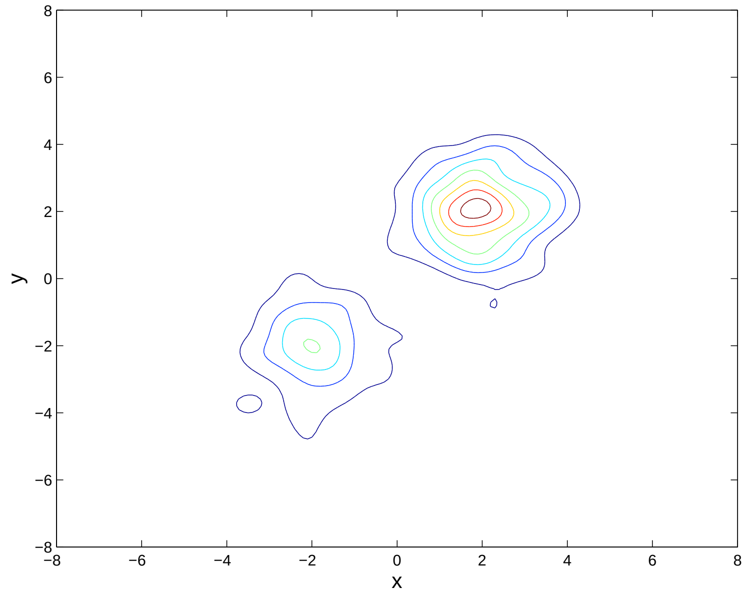
Result of SDC also compares favorably with known result of SVM for this example



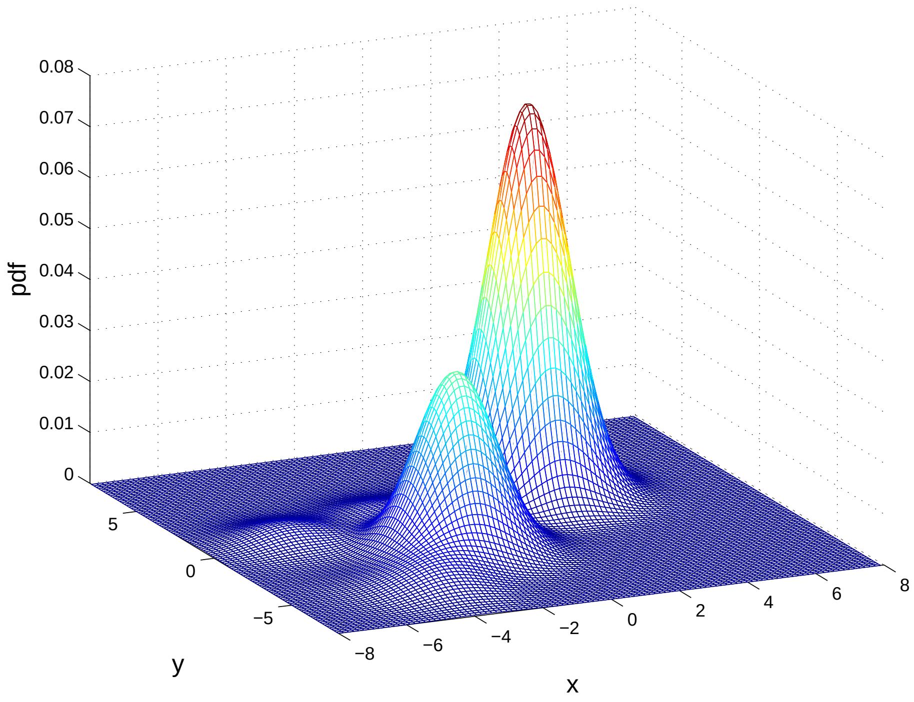


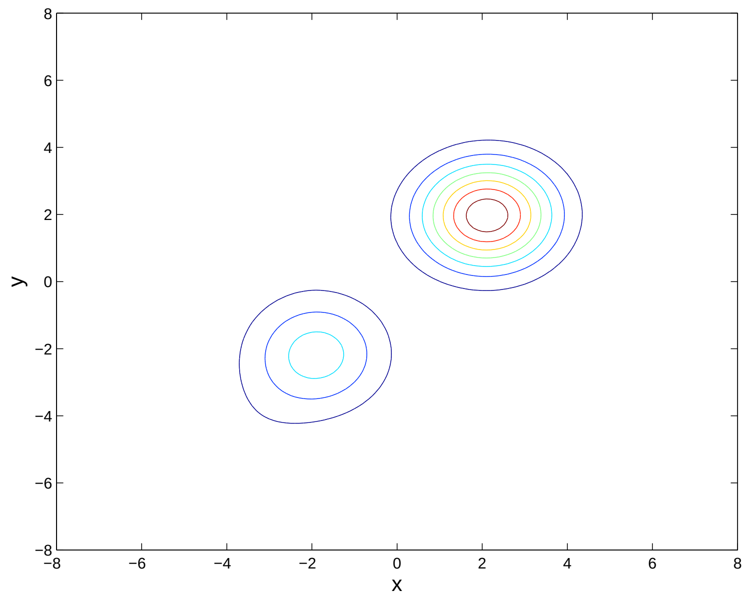
2-D Example: True Density





2-D Example: A Parzen Window Estimate





2-D Example: A Sparse Density Construction Estimate

A Classification Example

○ Synthetic 2-class classification in 2-D feature space from:

`http://www.stats.ox.ac.uk/PRNN/`

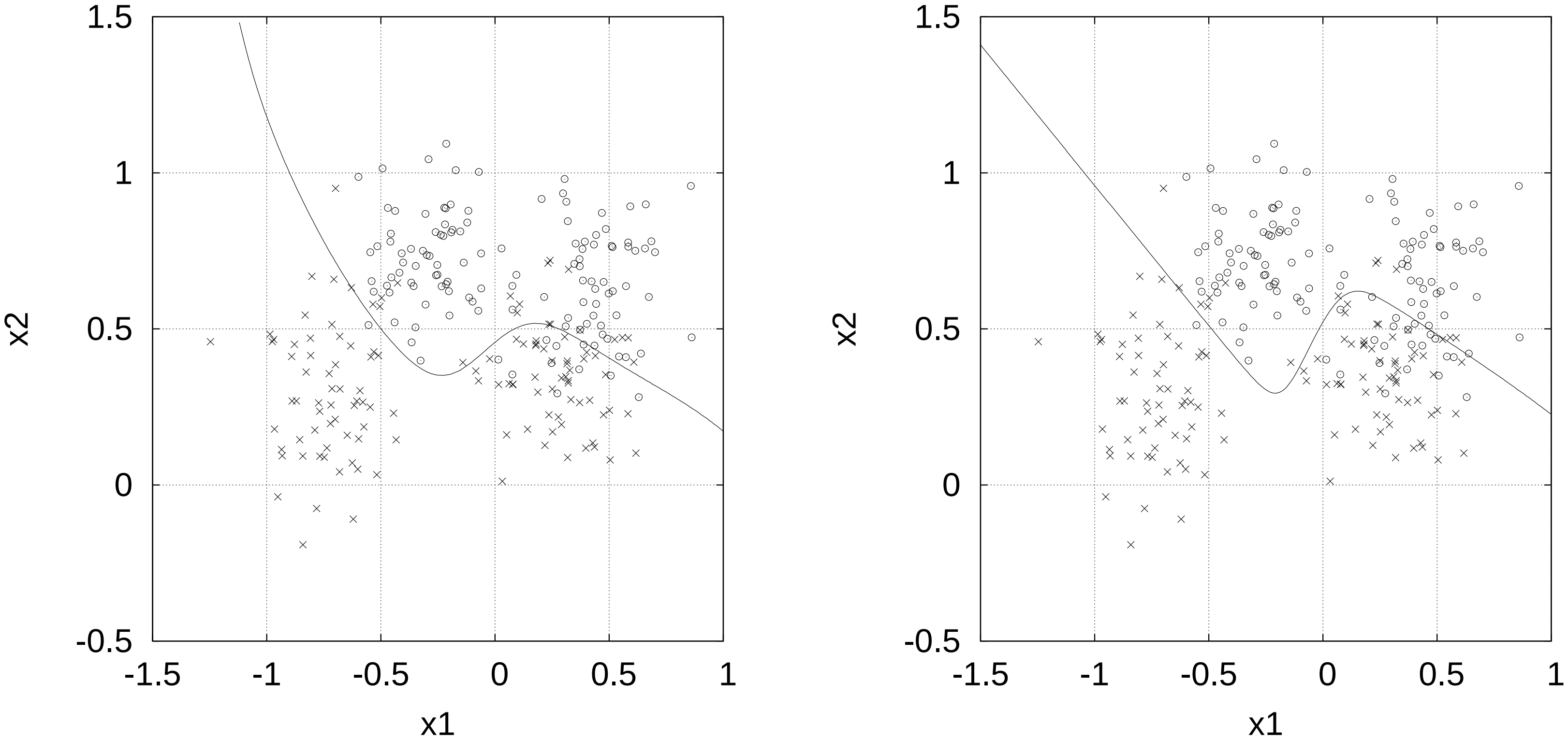
○ Training set: 250 samples and 125 points for each class, Test set: 1000 samples and 500 points for each class, optimal Bayes error rate for test set $\approx 8\%$

○ With Gaussian kernel, construct two class-conditional PDFs, then use them to form Bayes classifier

method	$\hat{p}(\bullet\|C0)$	$\hat{p}(\bullet\|C1)$	test error rate
PW	125 kernels	125 kernels	8.1%
SDC	5 kernels	4 kernels	8.3%

Result of SDC also compares favorably with known result of SVM classification for this example (38-kernel classifier with test error rate 10.6%)

Classification Example: Decision Boundary



(a) (b)

(a) Parzen window estimate and (b) sparse density construction estimation, where circles represent class-1 training data and crosses class-0 training data

Conclusions

- Efficient construction algorithm has been presented for obtaining kernel density estimates based on orthogonal forward regression that incrementally minimizes leave-one-out test score, coupled with local regularization to further enforce sparsity
- Proposed method is simple to implement and computationally efficient, and except for kernel width the algorithm contains no other free parameters that require tuning
- It offers a state-of-art technique for sparse kernel density estimation in practical applications