Fisher Discriminant Analysis with Kernels

- by Mika, Rätsch, Weston, Schölkpof, Müller
- presented by Boley.

Discriminate between two classes

- Need to identify good set of features
- PCA: unsupervised algorithm to reduce reconstruction error
- Better to take advantage of label info
- Classical approaches: bayes classifier - requires assumptions on data distribution within each class
- Often: assume Gaussian distribution within each class

  $\rightarrow$ leads to quadratic or linear discriminants, like Fisher

This work

- Authors propose kernel idea used in SVMs, K-PCA.
- Use in supervised Fisher's Discriminant
- Result often competitive with K SVMs.
- Dot-product in kernel space $\rightarrow$ closed form solution

Classical Fisher Linear Discriminant

- samples from two classes: $X_1 = [\mathbf{x}_1, \ldots, \mathbf{x}_{\ell_1}]$, $X_2 = [\mathbf{x}_{\ell_1+1}, \ldots, \mathbf{x}_{\ell_1+\ell_2}]$, with $\ell = \ell_1 + \ell_2$.
- Fisher's discriminant projects all the data onto a direction $\mathbf{w}$ maximizing the separation of the means along the projection while minimizing the scatter with each class

$$\max J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

where

S_B	$=$	$(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$	between cluster scatter
S_W	$=$	$\sum_{i=1,2} \sum_{\mathbf{x} \in X_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$	within class scatter
$\mathbf{m}_i$	$=$	$\frac{1}{\ell_i} \sum_{\mathbf{x} \in X_i} \mathbf{x}$	class mean
$\mathbf{m}$	$=$	$\frac{1}{\ell} \sum_{\mathbf{x}} \mathbf{x} = \frac{\ell_1}{\ell} \mathbf{m}_1 + \frac{\ell_2}{\ell} \mathbf{m}_2$	global mean

Statistical Motivation - Bayes

- Optimal Bayes assigns class based on maximum a-posteriori probability
- Simplifying assumption: each class has a normal distribution
- Measures Mahalanobis distance of a sample to class center
- Result is a quadratic separator
- With a single common Covariance matrix $\rightarrow$ linear separator
- linear separator advantage: robust against noise
- Direction of separator aligned with direction of maximal variance within each class
- Linear separator $\leftrightarrow$ Fisher's $\mathbf{w}$.
- Crucial: have enough samples to get good estimate of Covariance.

Fisher's discriminant in feature space

- Linear discriminant is not rich enough
- Want to keep robustness and statistical foundation while allowing richer separators
- Answer: use high-dimensional feature space $\mathcal{F}$
- Map $\mathbf{x} \mapsto \hat{\mathbf{x}} = \phi(\mathbf{x}) \in \mathcal{F}$.
- Fisher's Disc. is now:

$$\max J(\mathbf{w}) = \frac{\mathbf{w}^T \hat{S}_B \mathbf{w}}{\mathbf{w}^T \hat{S}_W \mathbf{w}}$$

where

$\hat{S}_B$	$=$	$(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2)(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2)^T$	between cluster scatter
$\hat{S}_W$	$=$	$\sum_{i=1,2} \sum_{\hat{\mathbf{x}} \in \hat{X}_i} (\hat{\mathbf{x}} - \hat{\mathbf{m}}_i)(\hat{\mathbf{x}} - \hat{\mathbf{m}}_i)^T$	within class scatter
$\hat{\mathbf{m}}_i$	$=$	$\frac{1}{\ell_i} \sum_{\hat{\mathbf{x}} \in \hat{X}_i} \hat{\mathbf{x}}$	class mean
$\hat{\mathbf{m}}$	$=$	$\frac{1}{\ell} \sum_{\hat{\mathbf{x}}} \mathbf{x} = \frac{\ell_1}{\ell} \hat{\mathbf{m}}_1 + \frac{\ell_2}{\ell} \hat{\mathbf{m}}_2$	global mean

Kernel Function

- Need to formulate problem in terms of dot-products of input patterns
- Any solution $\mathbf{w}$ must lie in span of training samples $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_\ell$ in $\mathcal{F}$.
- $\mathbf{w} = \sum_1^\ell \alpha_j \hat{x}_j = \sum_1^\ell \alpha_j \phi(\mathbf{x}_j)$.
- Inner Product with mean: $\mathbf{w}^T \hat{\mathbf{m}}_i = \sum_{j=1}^{\ell} \alpha_j \underbrace{\frac{1}{\ell_i} \sum_{\mathbf{x} \in X_i} \mathrm{k}(\mathbf{x}_j, \mathbf{x})}_{(\boldsymbol{M}_i)_j}$.
- Wish to optimize $\max J(\mathbf{w}) = \mathbf{w}^T \hat{S}_B \mathbf{w} \;/\; \mathbf{w}^T \hat{S}_W \mathbf{w}$
- Numerator: $\mathbf{w}^T \hat{S}_B \mathbf{w} = \boldsymbol{\alpha}^T \underbrace{(\boldsymbol{M}_1 - \boldsymbol{M}_2)(\boldsymbol{M}_1 - \boldsymbol{M}_2)^T}_{M} \boldsymbol{\alpha}$
- Here $\boldsymbol{M}_i$ is the ℓ-vector of weighted row sums of the kernel matrix $K = \{K_{ij}\} = \{\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1,\ldots,\ell}$.

Kernel Function 2

- Wish to optimize $\max J(\mathbf{w}) = \mathbf{w}^T \hat{S}_B \mathbf{w} \ / \ \mathbf{w}^T \hat{S}_W \mathbf{w}$
- Denominator: $\mathbf{w}^T \hat{S}_W \mathbf{w} = \boldsymbol{\alpha}^T \underbrace{(K_1(I - \mathbf{1}_{\ell_1})K_1^T) + (K_2(I - \mathbf{1}_{\ell_2})K_2^T)}_{N} \boldsymbol{\alpha}$

$$\begin{aligned} \text{where } K_1 &= \{(K_1)_{ij}\} = \{\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j)\}_{i=1,\ldots,\ell}^{j=1,\ldots,\ell_1} \ (\ell \times \ell_1 \text{ matrix}) \\ K_2 &= \{(K_2)_{ij}\} = \{\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j)\}_{i=1,\ldots,\ell}^{j=1,\ldots,\ell_2} \ (\ell \times \ell_2 \text{ matrix}) \\ K &= (K_1 \ , \ K_2). \end{aligned}$$

Kernel Fisher Discriminant

- KFD is now solved by optimizing

$$\max J(\mathbf{w}) = \frac{\mathbf{w}^T N \mathbf{w}}{\mathbf{w}^T M \mathbf{w}}.$$

- Solve by finding leading eigenvector of $N^{-1}M$ [or better, solve generalized eigenproblem $M\mathbf{w} = \lambda N \mathbf{w}$].

- Project new pattern $\hat{\mathbf{x}} = \phi(\mathbf{x})$ onto $\mathbf{w}$ by

$$\langle \mathbf{w}, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{\ell} \alpha_i \mathrm{k}(\mathbf{x}_i, \mathbf{x})$$

Numerical Issues

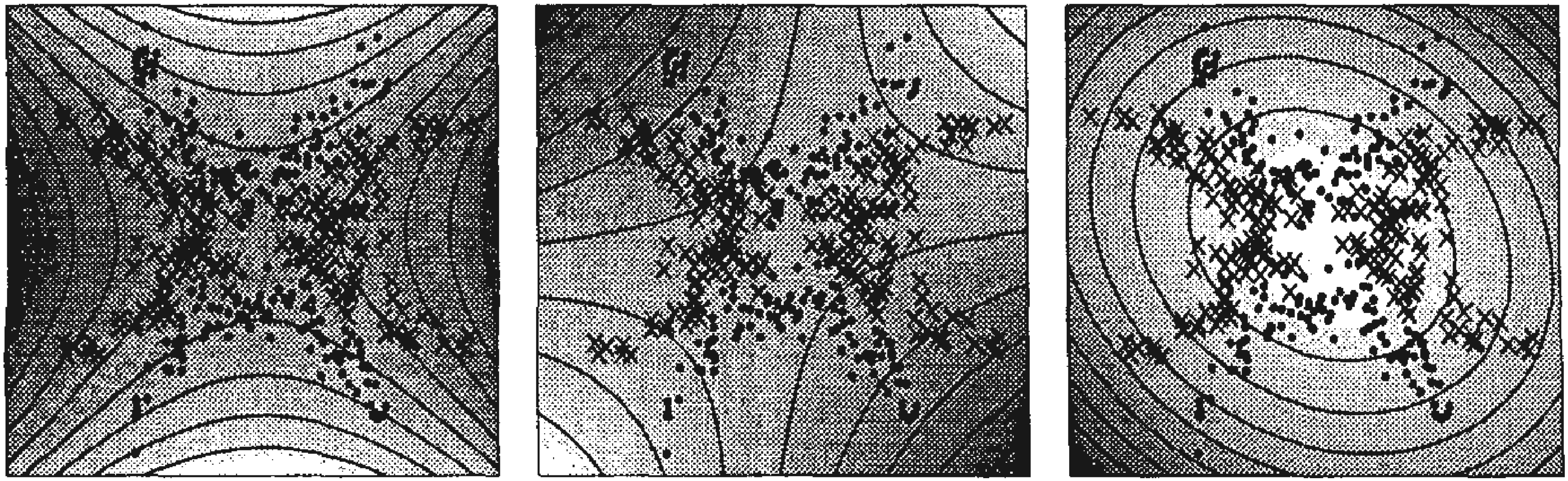
- Estimating ℓ covariance structures from ℓ samples $\rightarrow$ ill-posed.
- N could be singular or badly conditioned
- Need capacity control in $\mathcal{F}$

Solution

- Replace N with $N_\mu = N + \mu I$.
- Effect: Makes N better conditioned
- Decreases bias in sample-based eigenvalue estimates
- Imposes regularization on $\|\boldsymbol{\alpha}\|^2$, favoring solutions with small expansion coefficients.
- Regularization effect not fully understood.
- Other forms of regularization possible.

Illustration

Figure 1: Comparison of feature found by KFD (left) and those found by Kernel PCA: first (middle) and second (right); details see text.



- KFD: polynomial kernel degree two, regularized with $\mu = 10^{-3}$.
- Two classes ($\times$'s & $\bullet$'s), parabolic mirrored around axes.
- Contour lines = level sets
- KFD level sets discriminate classes well
- KPCA less so.

Experiments

- Compare to other state-of-the-art classifiers
- KFD: Kernel Fisher Discrminant with Gaussian kernel
  - Once $\mathbf{w}$ obtained, used 1-d linear SVM to classify
- Adaboost
- Regularized Adaboost
- SVM: Support Vector Machine with Gaussian kernel

Data Sets

- Sources: ICI DELVE STATLOG Benchmark data sets
- Treated all as two-class problems
- 100 partitions into training/test sets (about 60%:40%)
- Hyperparameters estimated using 5-fold cross-validation over first 5 realizations
- Table shows average test error & standard deviation over 100 runs

Results

Preliminary Experiment with USPS Digit Data

- Used 3000 training samples
- Compared KFD with KSVM, both with Gaussian kernels
- 10 class error: KFD: 3.7%, KSVM: 4.2%

In General

- Noticed: both KFD & SVM yield optimal hyperplane in $\mathcal{F}$: often former is better.
- Complexity of SVM classifier is $O(support vectors)$.
- Complexity of KFD classifier is $O(all training vectors)$.
- Dependence on all training vectors $\rightarrow$ maybe more robust.
- KFD: closed form solution.
  Other methods involve a search or an optimization problem.
- Table on next page: **1st place in bold**, *2nd place in italic* (lower is better)

Experiments

Table 1: Comparison between KFD, a single RBF classifier, AdaBoost (AB), regularized AdaBoost (AB_R) and Support Vector Machine (SVM) (see text). Best method in bold face, second best emphasized.

	RBF	AB	AB_R	SVM	KFD
Banana	**10.8±0.6**	12.3±0.7	*10.9±0.4*	11.5±0.7	**10.8±0.5**
B.Cancer	27.6±4.7	30.4±4.7	26.5±4.5	*26.0±4.7*	**25.8±4.6**
Diabetes	24.3±1.9	26.5±2.3	23.8±1.8	*23.5±1.7*	**23.2±1.6**
German	24.7±2.4	27.5±2.5	24.3±2.1	**23.6±2.1**	*23.7±2.2*
Heart	17.6±3.3	20.3±3.4	16.5±3.5	**16.0±3.3**	*16.1±3.4*
Image	3.3±0.6	**2.7±0.7**	**2.7±0.6**	*3.0±0.6*	4.8±0.6
Ringnorm	1.7±0.2	1.9±0.3	*1.6±0.1*	1.7±0.1	**1.5±0.1**
F.Sonar	34.4±2.0	35.7±1.8	34.2±2.2	**32.4±1.8**	*33.2±1.7*
Splice	*10.0±1.0*	10.1±0.5	**9.5±0.7**	10.9±0.7	10.5±0.6
Thyroid	4.5±2.1	*4.4±2.2*	4.6±2.2	4.8±2.2	**4.2±2.1**
Titanic	23.3±1.3	*22.6±1.2*	*22.6±1.2*	**22.4±1.0**	23.2±2.0
Twonorm	2.9±0.3	3.0±0.3	*2.7±0.2*	3.0±0.2	**2.6±0.2**
Waveform	10.7±1.1	10.8±0.6	**9.8±0.8**	*9.9±0.4*	*9.9±0.4*

Conclusions and Discussion

- Fisher's discriminant: standard linear statistical technique, but too limited.
- This is one of many approcahes to obtain more general class separability.
- Advantage: closed form solution.
- Flexibility: wide choice of kernels.
- Experimental results: competitive with many other methods.
- Complexity scales with all training samples (not just the difficult ones)

Future Work

- Suitable approximation schemes
- Numerical methods to find a few leading eigenvectors
- Multi-class discriminants
- Generalization bounds.

Novelty Detection

Kernel PCA for Novelty Detection by Heiko Hoffman

- Novelty Detection is a one-class classification problem.
- Use training data to see typical acceptable data.
- Called One-Class because training data contains only acceptable data.
- Test data may be similar to training data or not: objective is to distinguish those that are different.
- Abnormal examples are generally rare.
- Alternate algorithm: One-class SVM: find tightest separator from origin in $\mathcal{F}$.
- Alternate algorithm: SVDD: Find smallest enclosing sphere in kernel space $\mathcal{F}$. RBF kernel leads to same as one-class SVM.
- Here we try to generate a simplified model.
- Alternate approaches: • Gaussian Mixture models, • auto-associative multilayer perceptror
  • principal curves and surfaces,
  All these lead to non-linear (often non-convex) optimization problems.
- Here we use PCA in kernel space to reduce dimensionality.

Method

- Training data are mapped into an infinite-dimensional feature space.

- In this space, kernel PCA extracts the principal components of the data distribution. Eigenvectors $\{\mathbf{v}_\ell\}_{\ell=1}^{q}$ of $\bar{K}$ with $\bar{K}_{ij} = K_{ij} - \frac{1}{n}\sum_r K_{ir} - \frac{1}{n}\sum_r K_{rj} + \frac{1}{n^2}\sum_{r,s} K_{rs}$ where $K_{ij} = \mathrm{k}(\mathbf{x}_i, \mathbf{x}_j)$.

- Potential: $p_S(\mathbf{z}) = \|\phi(\mathbf{z}) - \bar{\phi}\|_2^2 = \mathrm{k}(\mathbf{z},\mathbf{z}) - \dfrac{2}{n}\sum_{i=1}^{n}\mathrm{k}(\mathbf{z},\mathbf{x}_i) + \dfrac{1}{n^2}\sum_{i,j}^{n}\mathrm{k}(\mathbf{x}_i,\mathbf{x}_j)$

- Projection: $f_\ell(\mathbf{z}) = \left\langle \left[\phi(\mathbf{z}) - \dfrac{1}{n}\sum_{r=1}^{n}\phi(\mathbf{x}_r)\right], \left[\mathbf{v}_l - \bar{\phi}(\mathbf{x})\right] \right\rangle$

  where $\mathbf{v}_l = \ell$-th eigenvector & $\bar{\phi}(\mathbf{x})$ is center in $\mathcal{F}$(both linear comb's of $\phi(\mathbf{x}_i)$'s).

- The squared distance to the corresponding principal subspace is the measure for novelty:

  $$p(\mathbf{z}) = p_s(\mathbf{z}) - \sum_{i=1}^{q} f_\ell(\mathbf{z})^2$$

Diagram

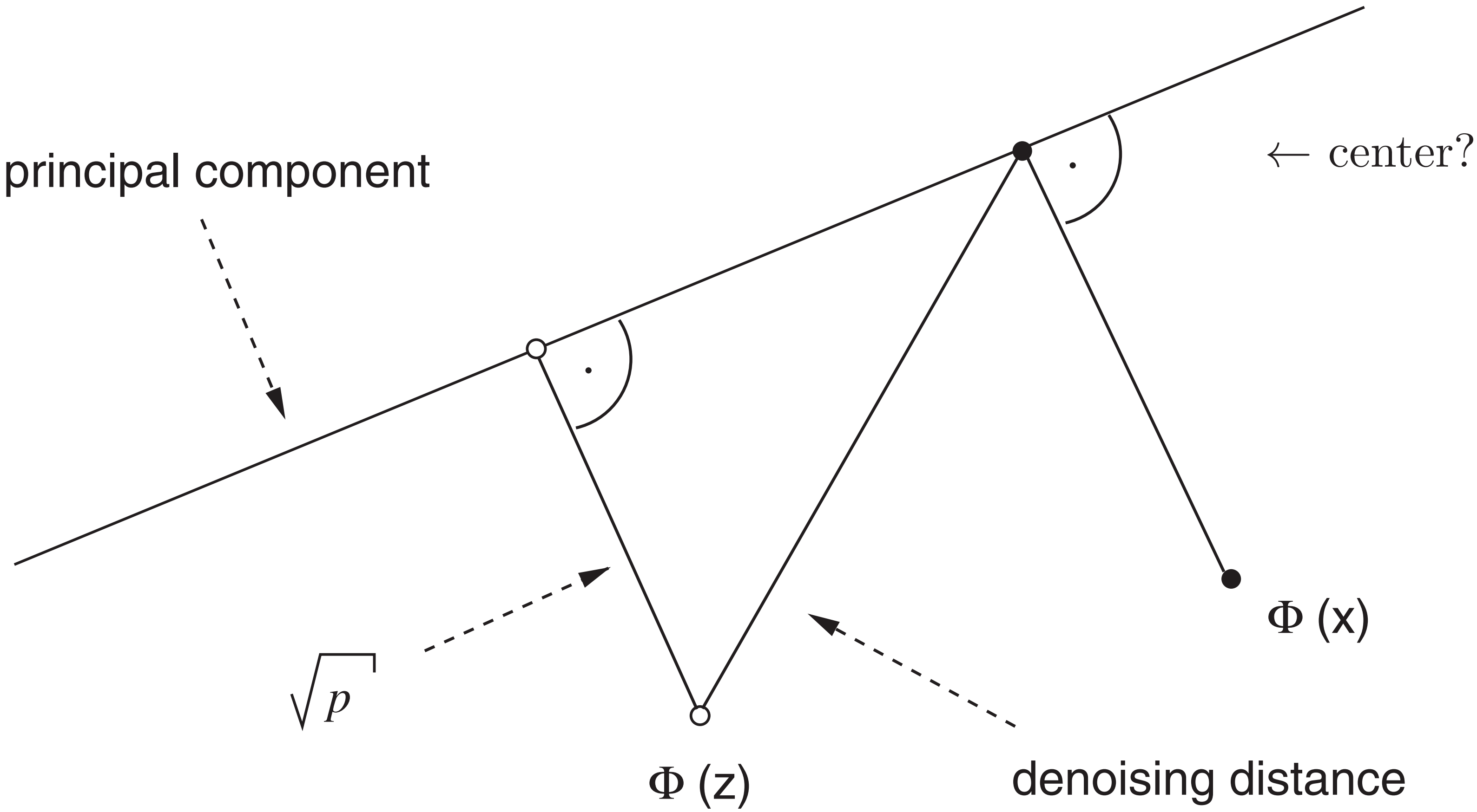
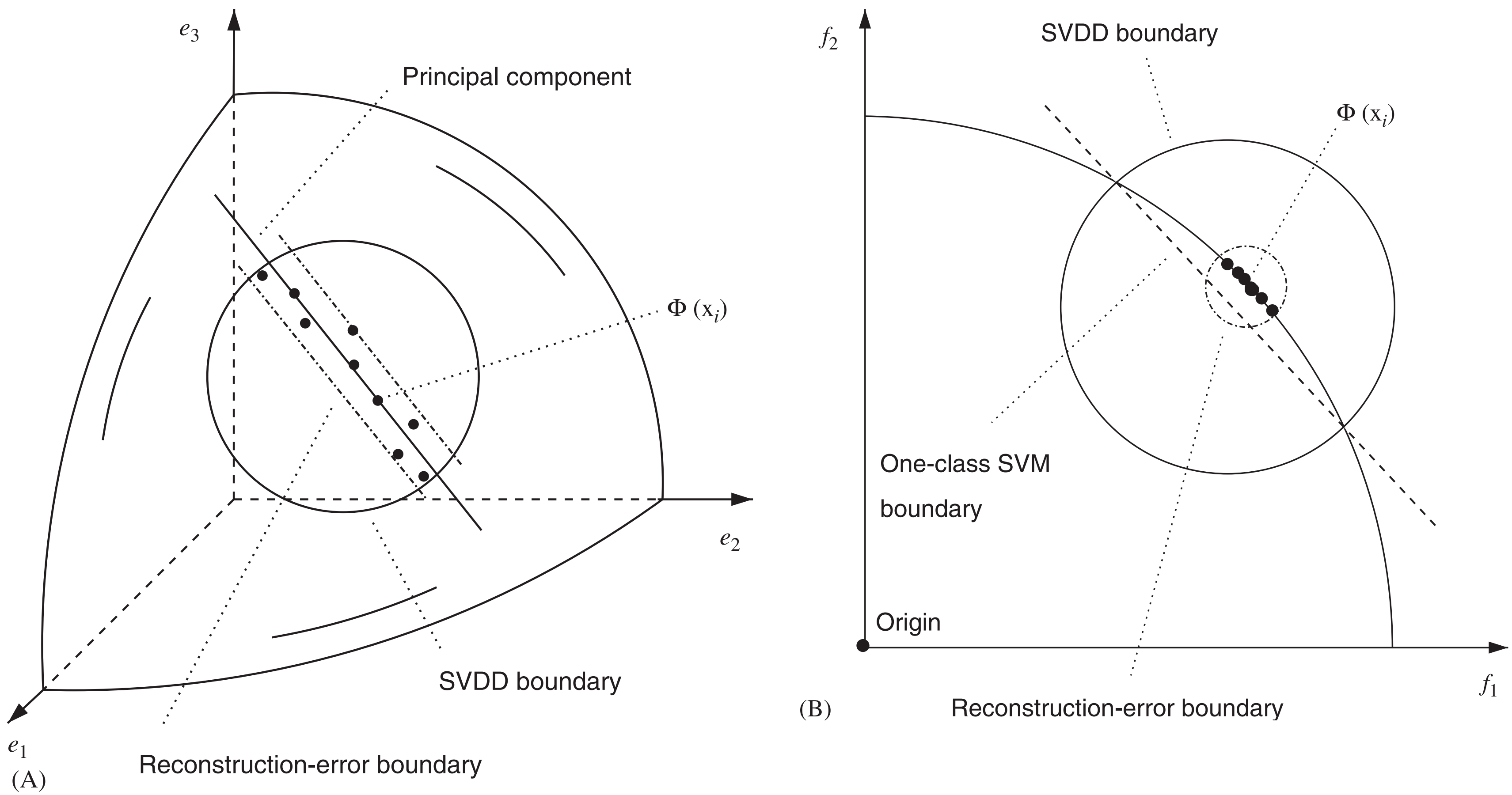


Fig. 12. The difference between the distance to be optimized in denoising and the reconstruction error p.

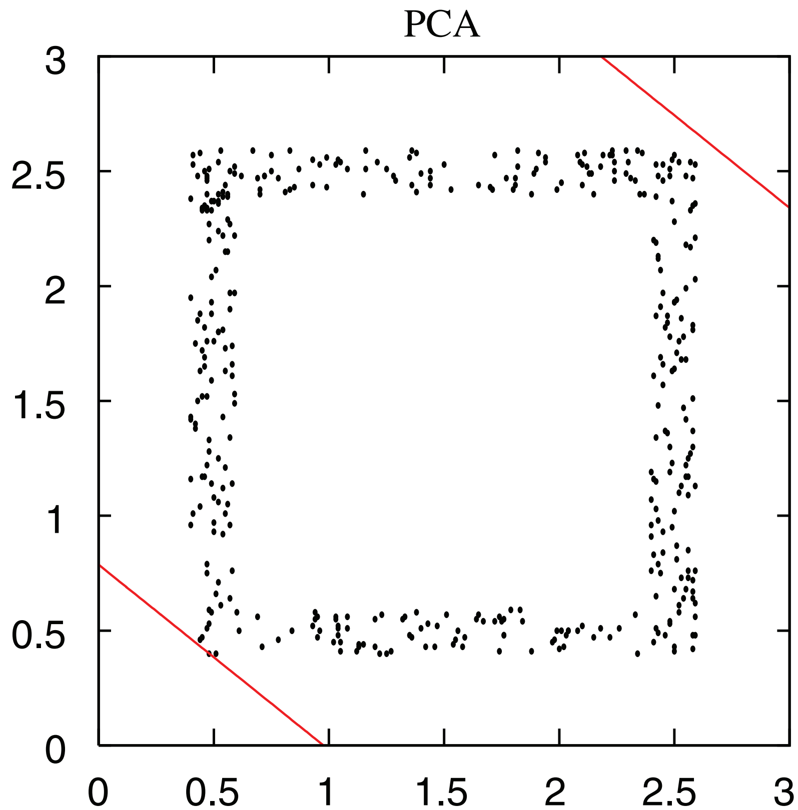
Decision Boundary Sketch

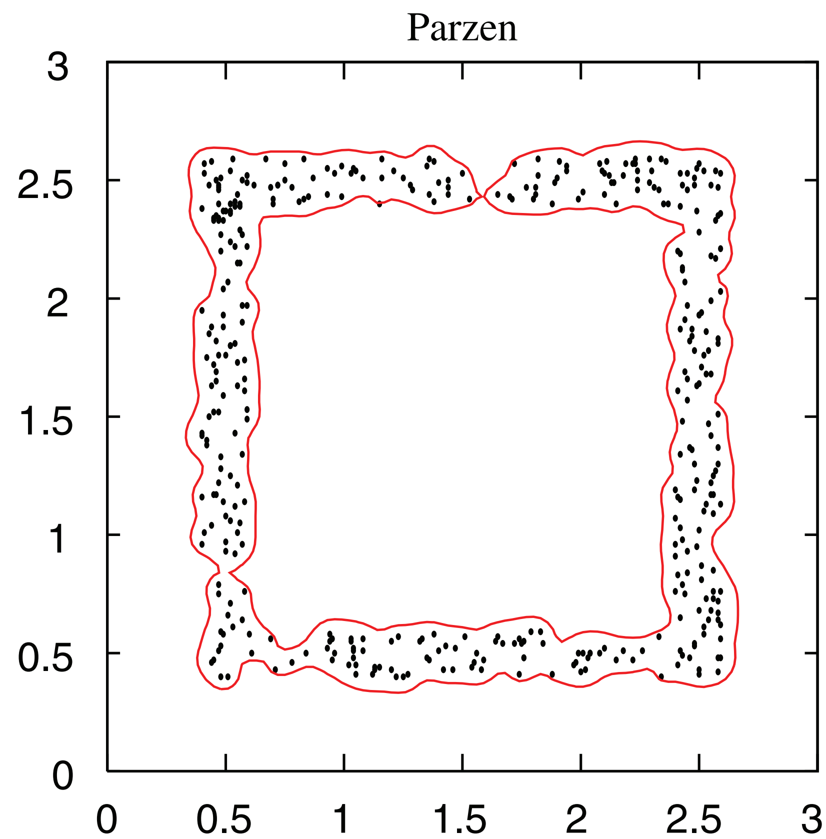
Fig. 1. Decision boundaries in the feature space of an RBF kernel, comparing one-class SVM, SVDD, and the reconstruction error: (A) The boundaries are illustrated in a three-dimensional feature space. All data points $\mathbf{\Phi}(\mathbf{x}_i)$ lie on a sphere. (B) Cross-section through the center of the SVDD sphere and orthogonal to the principal component for the situation in A.

Illustration

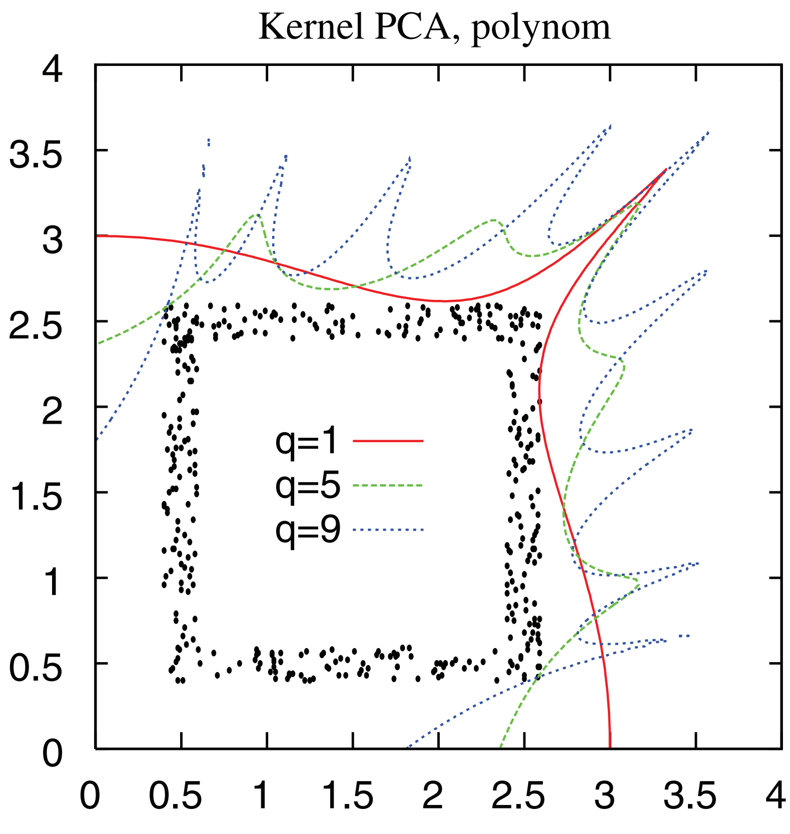


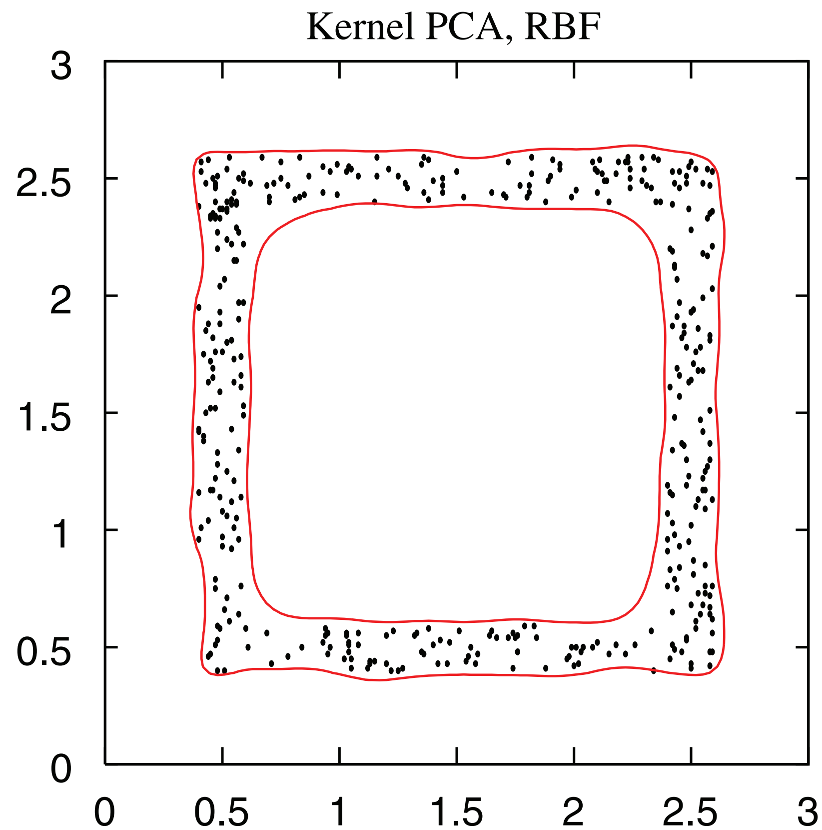
Example - classical methods





Example - kernel methods





Ring Square Boundary

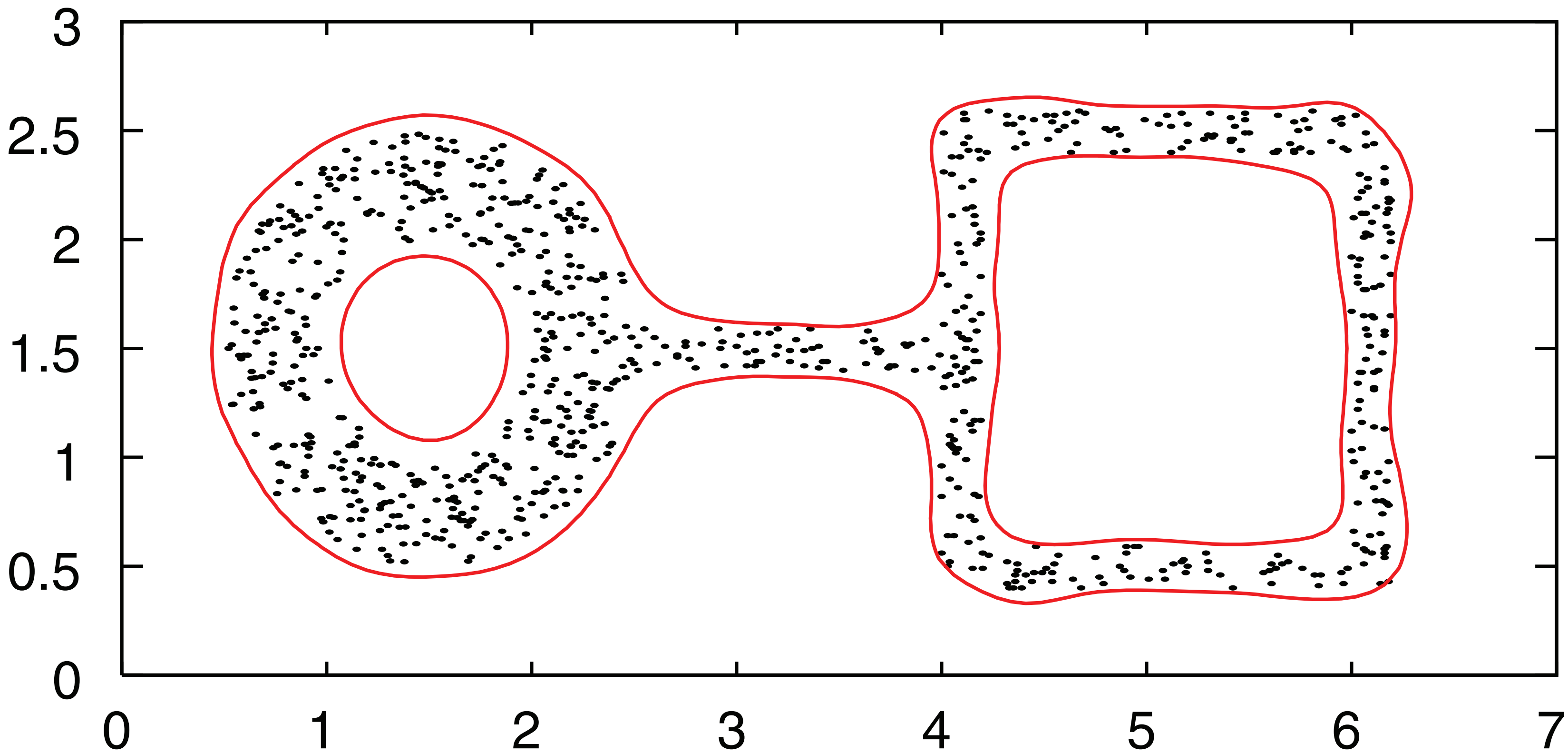


Fig. 3. Decision boundary for the ring-line-square distribution using the reconstruction error in $\mathscr{F}$ with $\sigma = 0.4$ and $q = 40$.

spiral

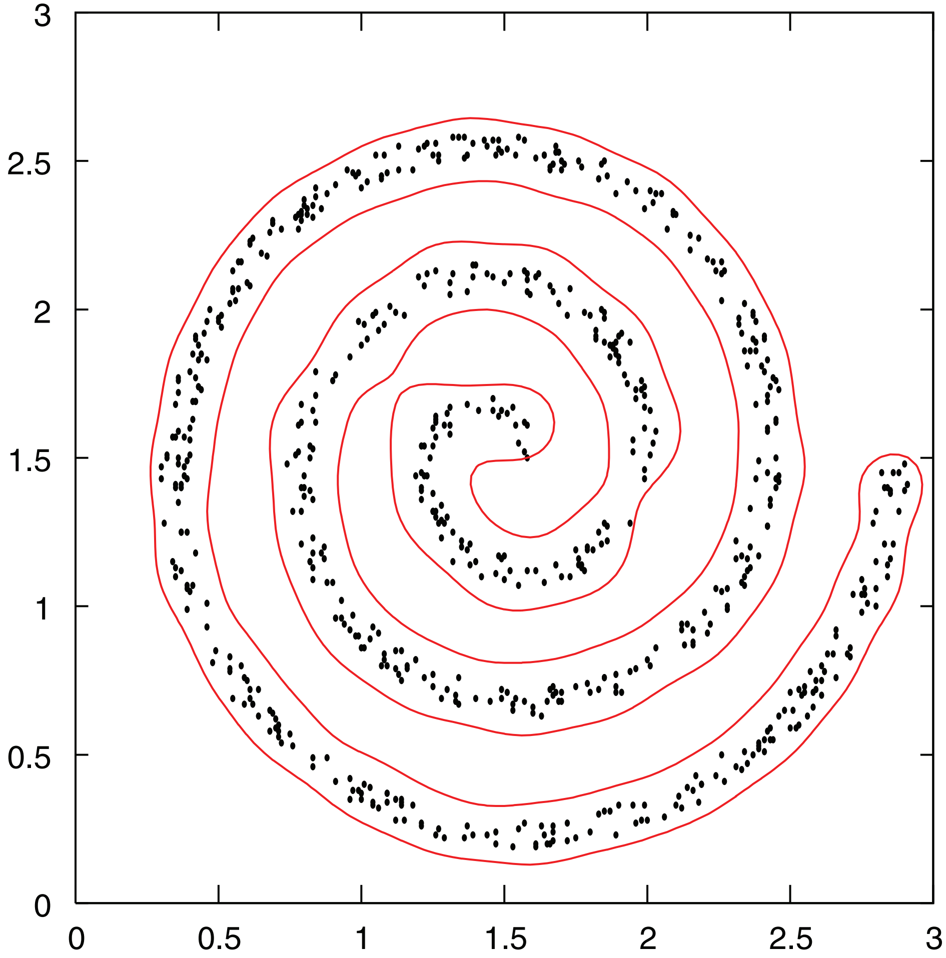
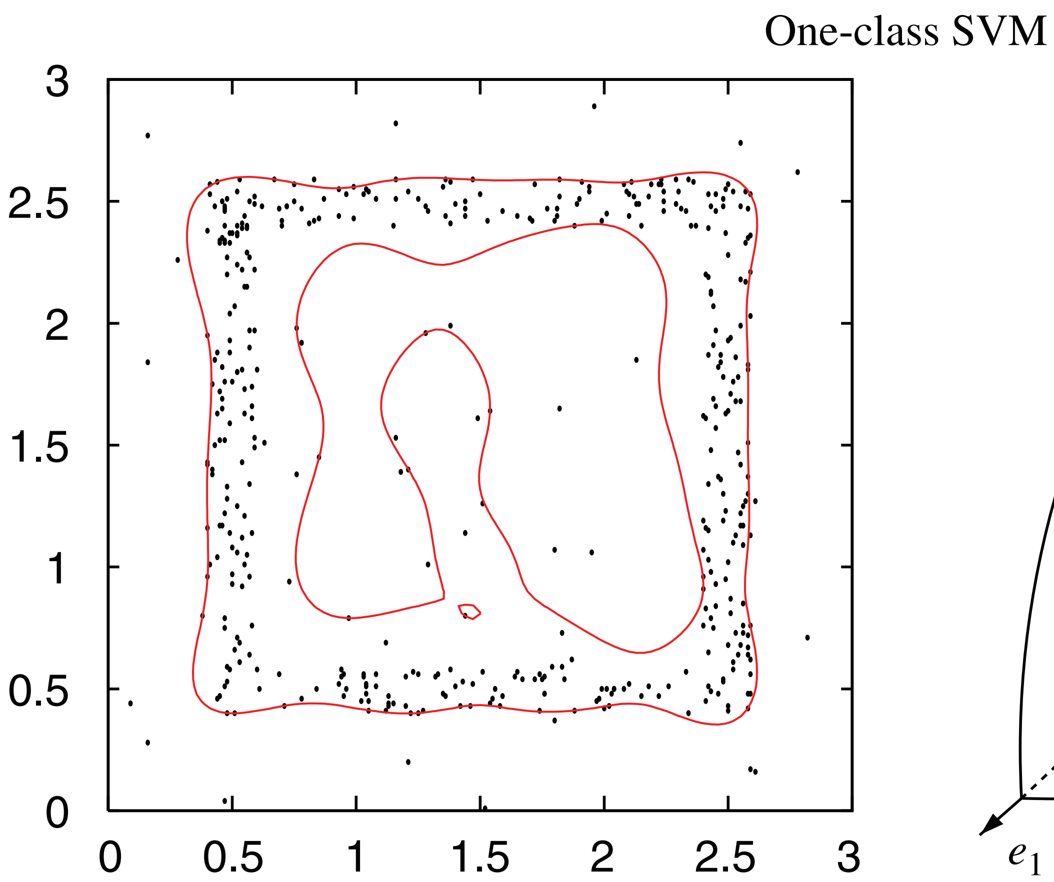
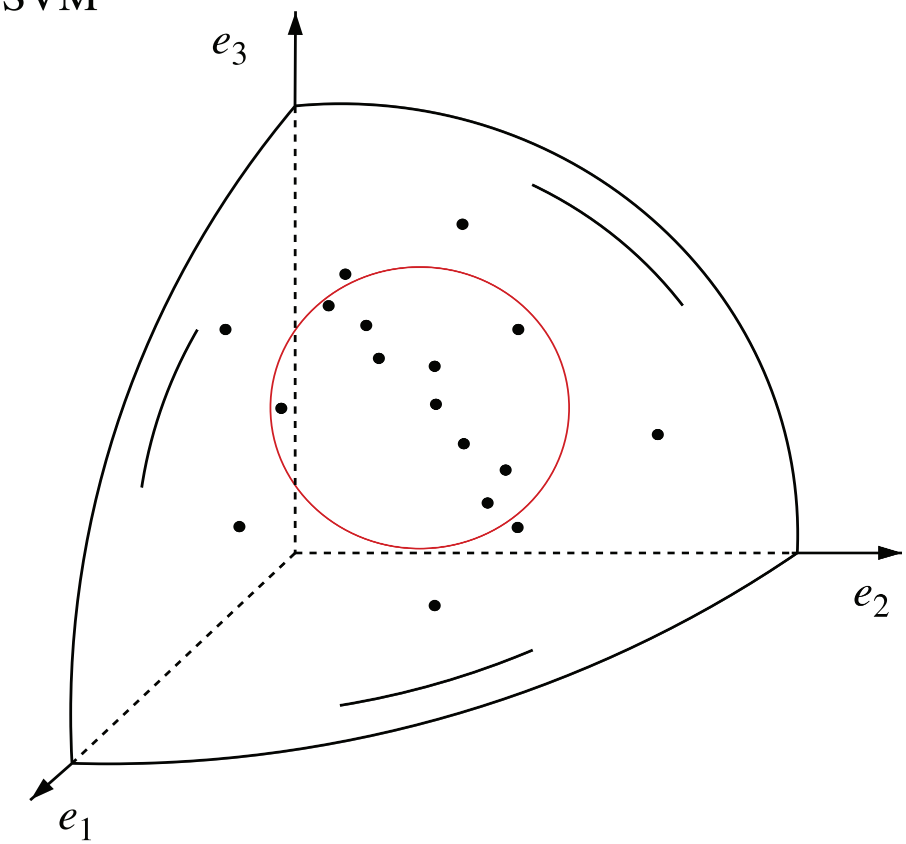


Fig. 4. Decision boundary for the spiral distribution using the reconstruction error in $\mathscr{F}$ with $\sigma = 0.25$ and $q = 40$.

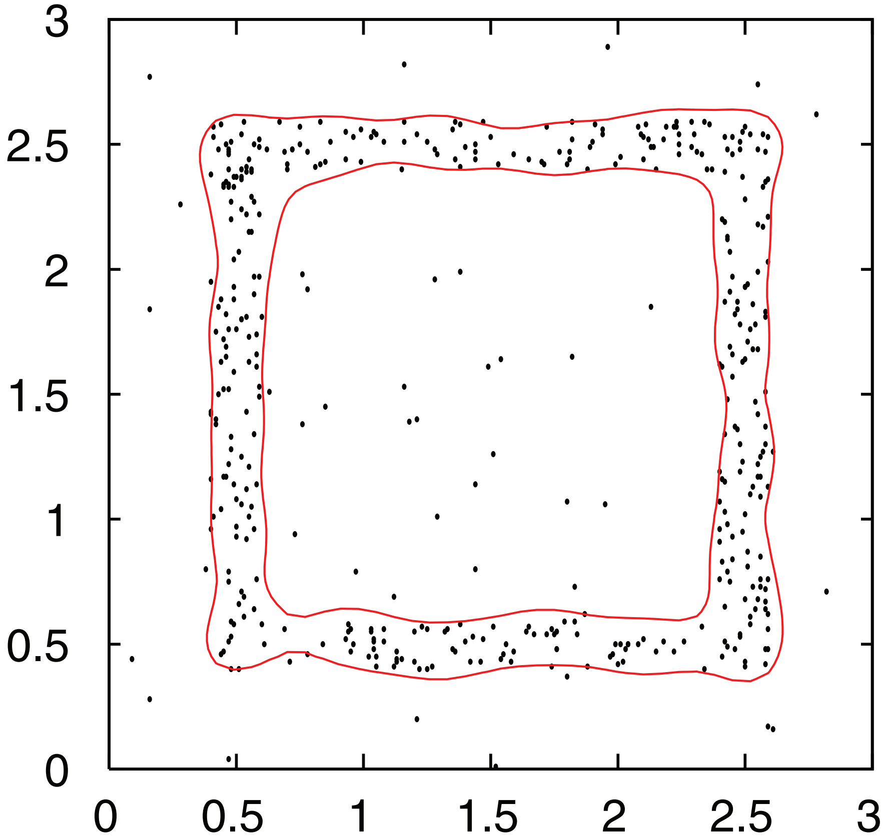
Noisy Data - One-class SVM

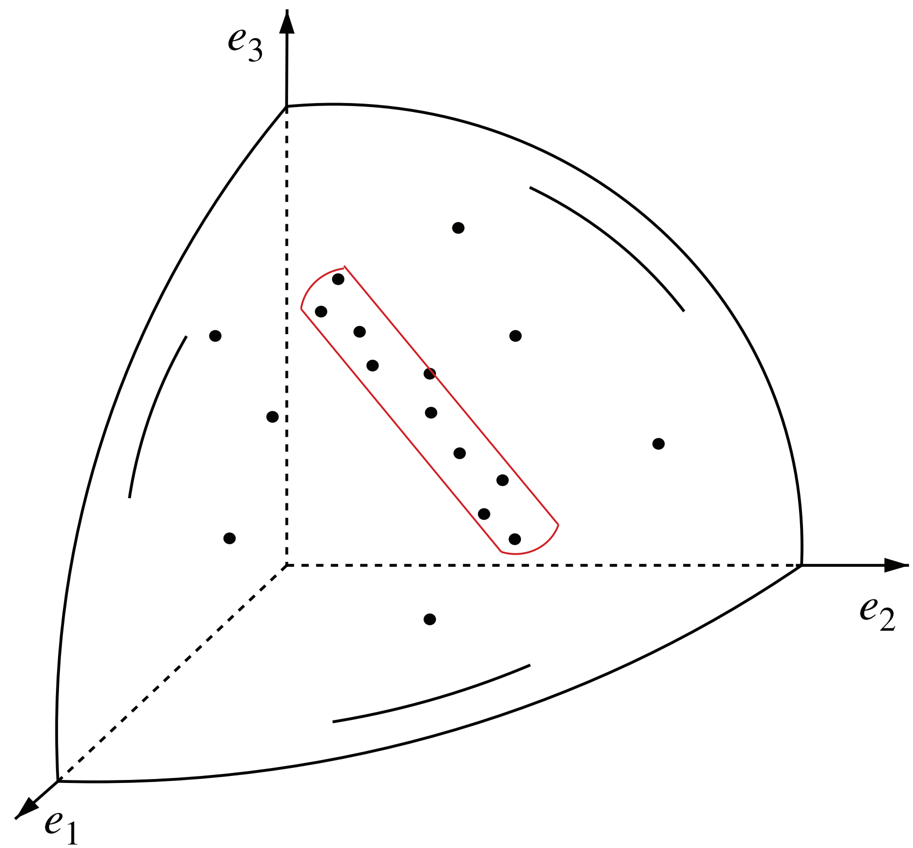




Noisy Data - K-PCA

Kernel PCA





Samples from sine curve plus uniform noise

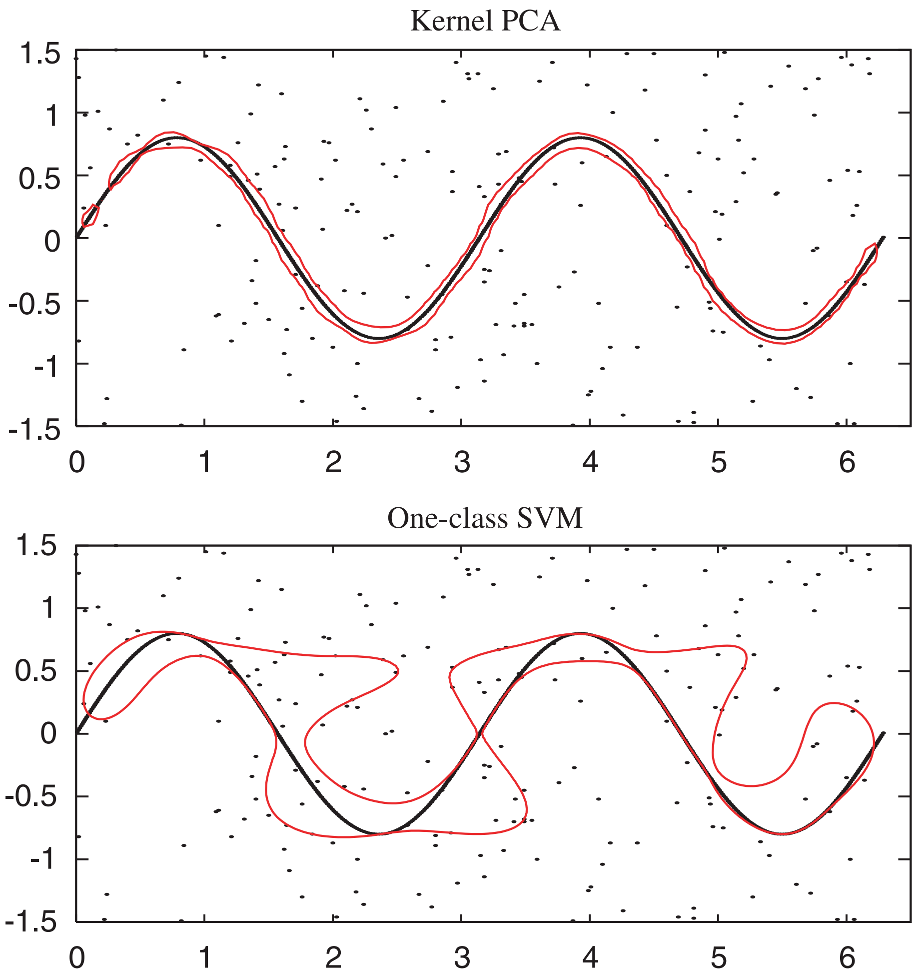
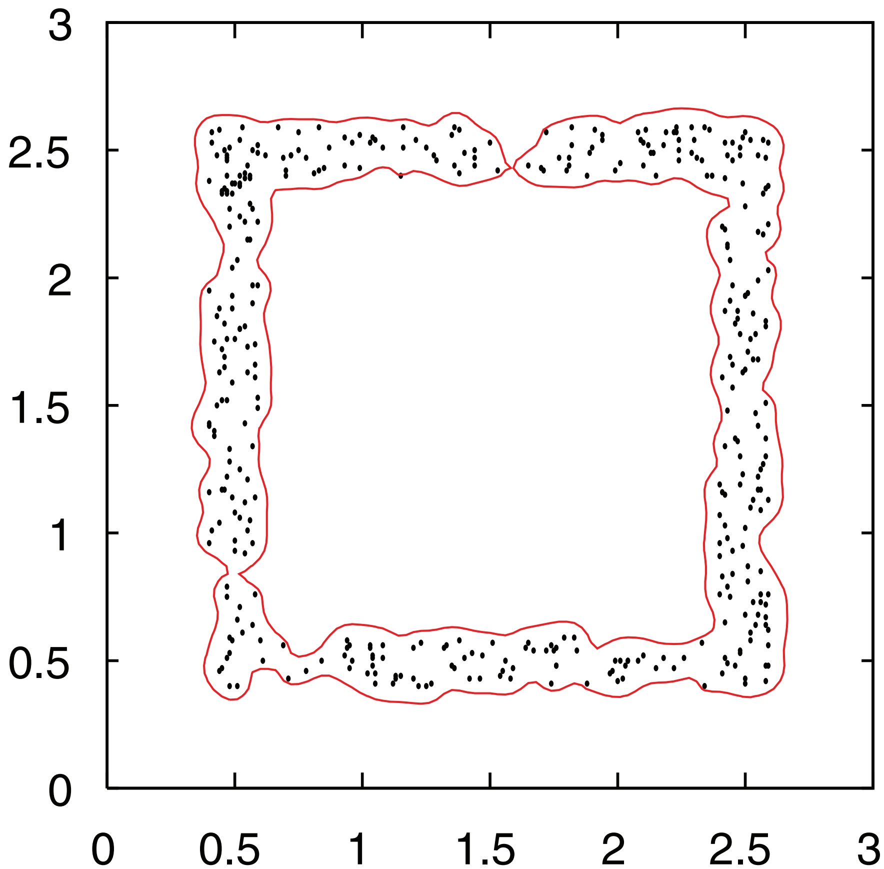
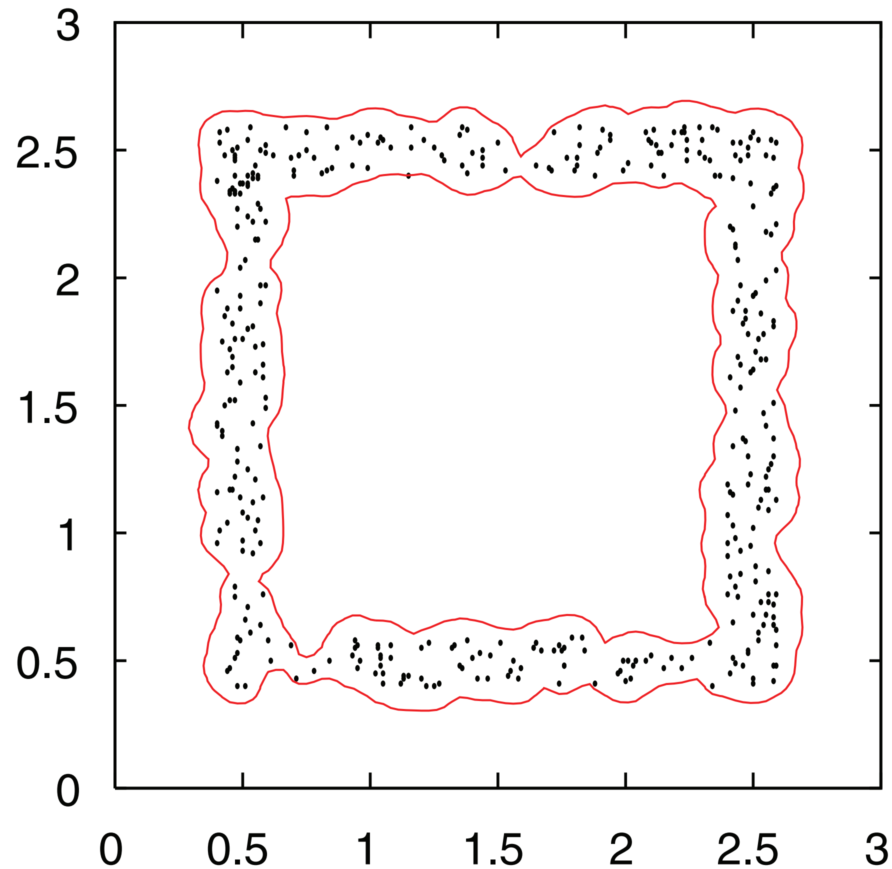


Fig. 6. Decision boundaries for the sine-noise distribution comparing kernel PCA ($\sigma = 0.4$, $q = 40$) with the one-class SVM ($\sigma = 0.489$, $\nu = \frac{2}{7}$).

Vary Parameters: $\sigma = .05$

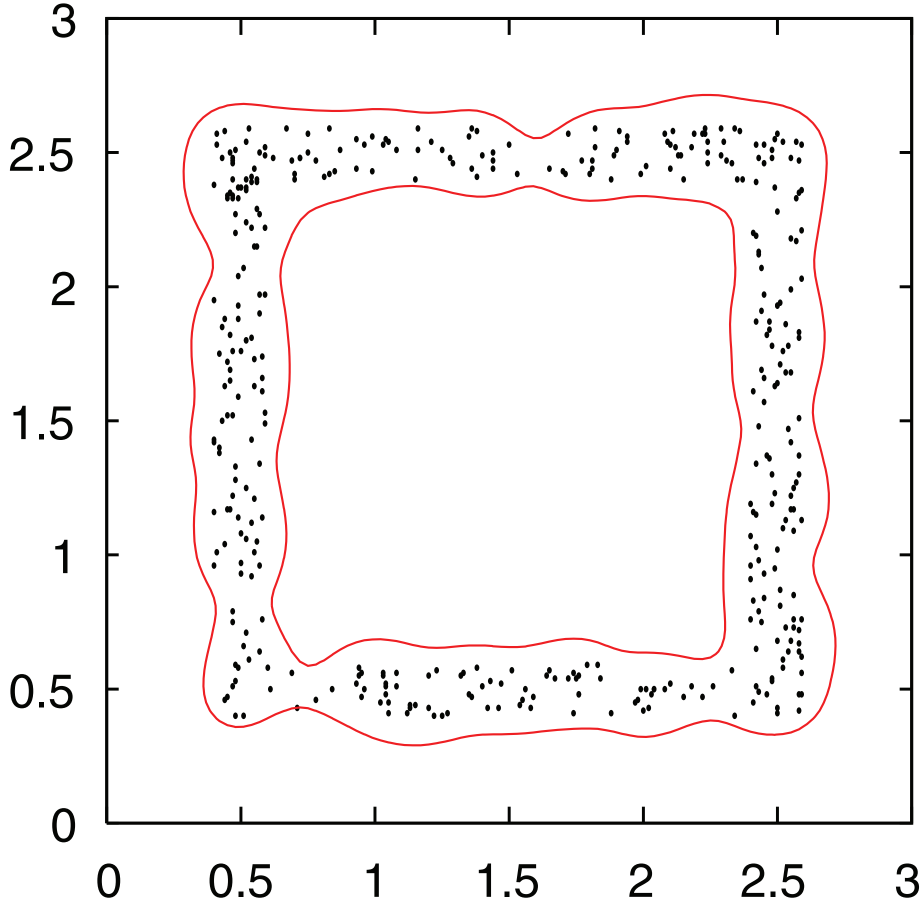


$q = 0$

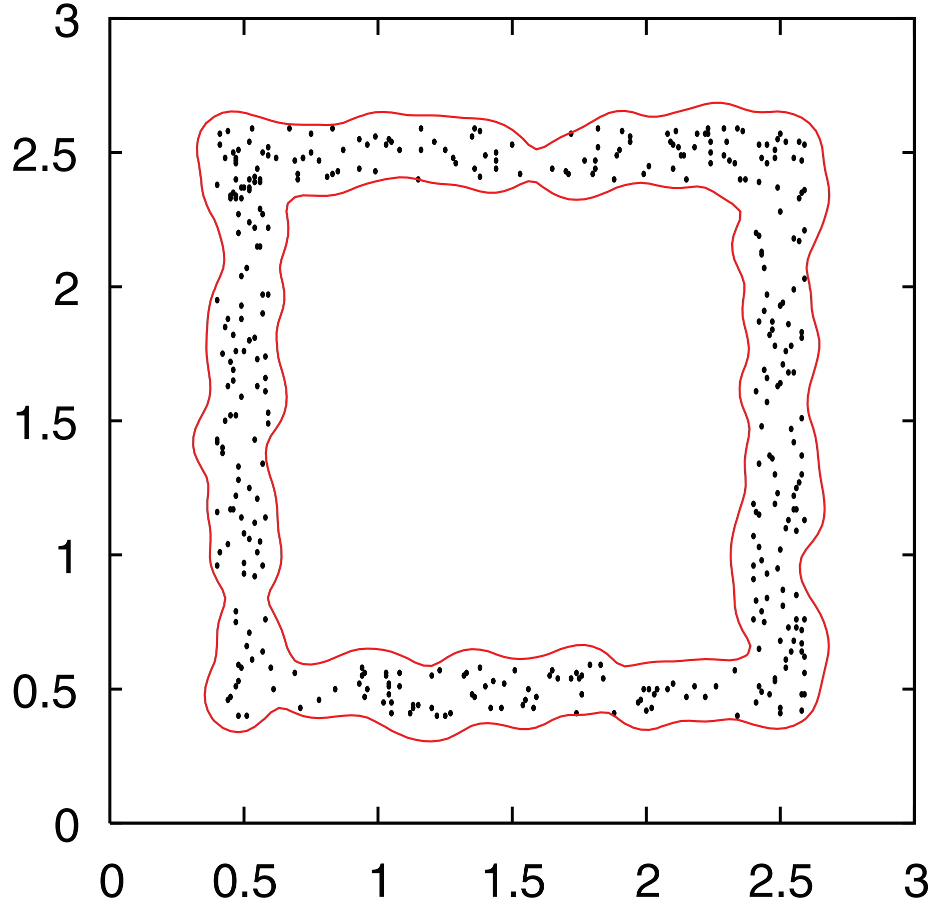


$q = 40$

Vary Parameters: $\sigma = .10$

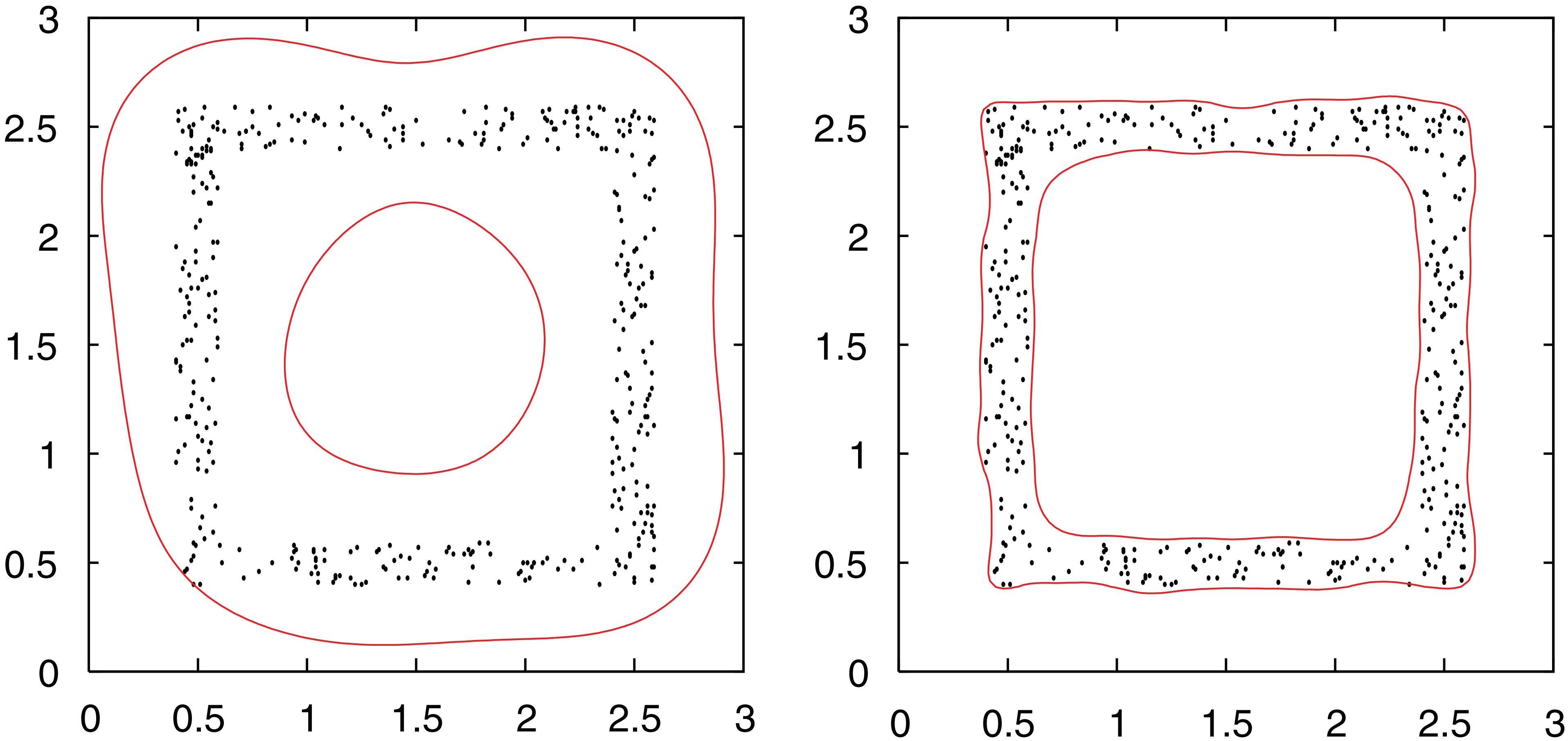


$q = 0$



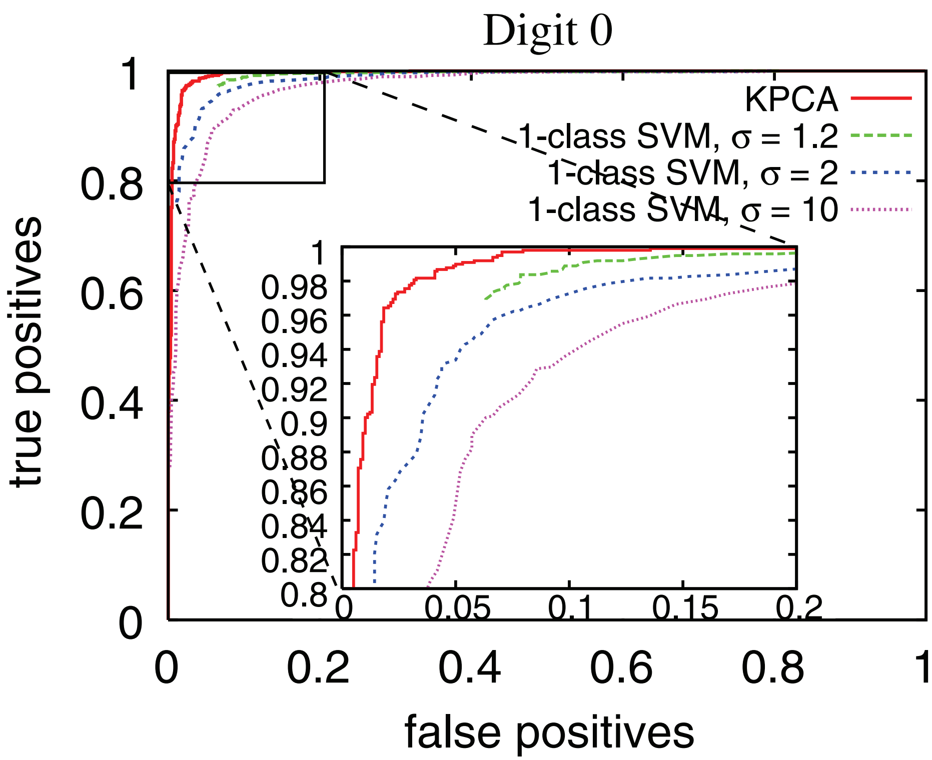
$q = 40$

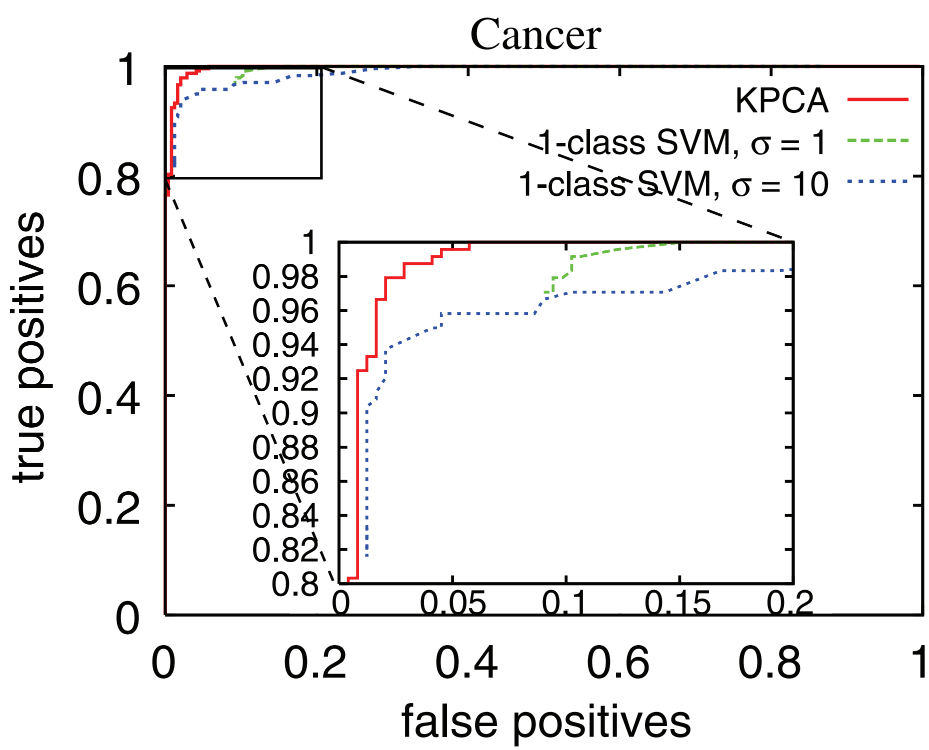
Vary Parameters: $\sigma = .40$



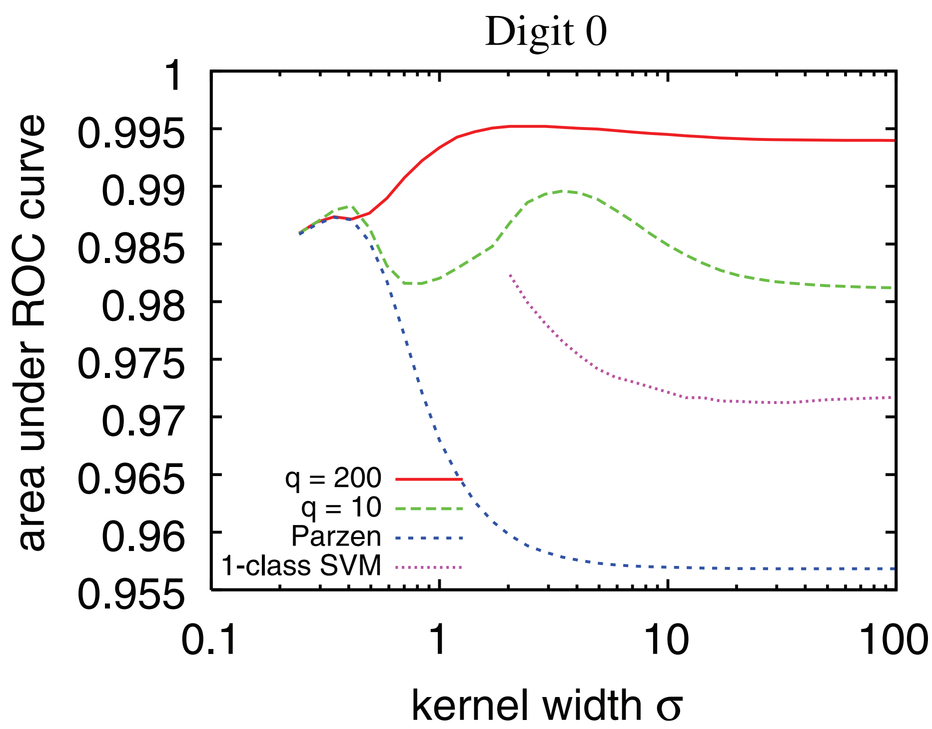
$q = 0$ $\quad$ $q = 40$

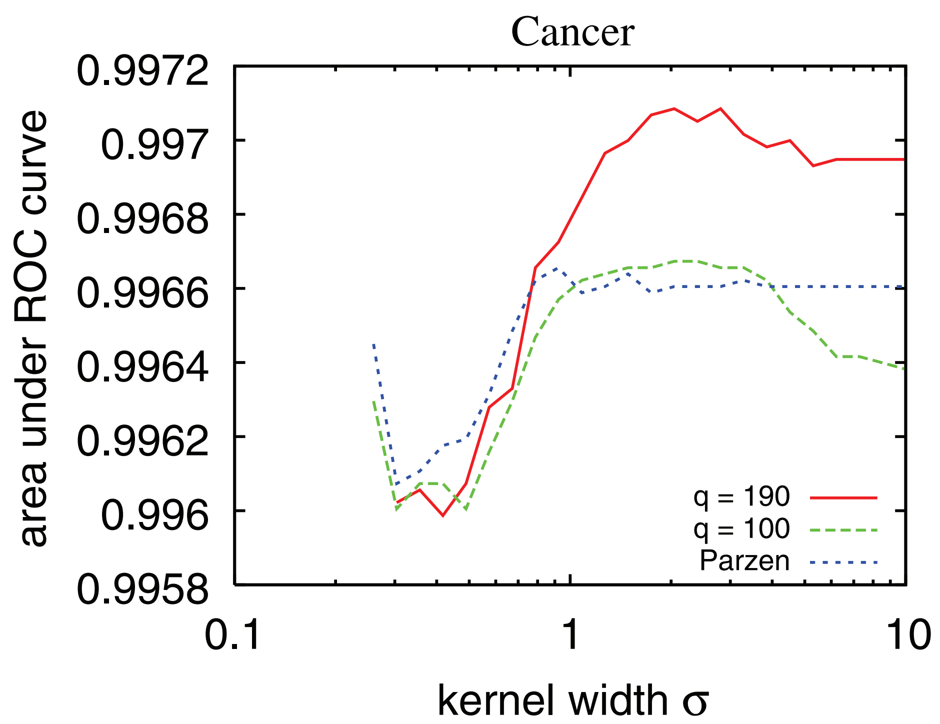
Real Data ROC curves : Classifier



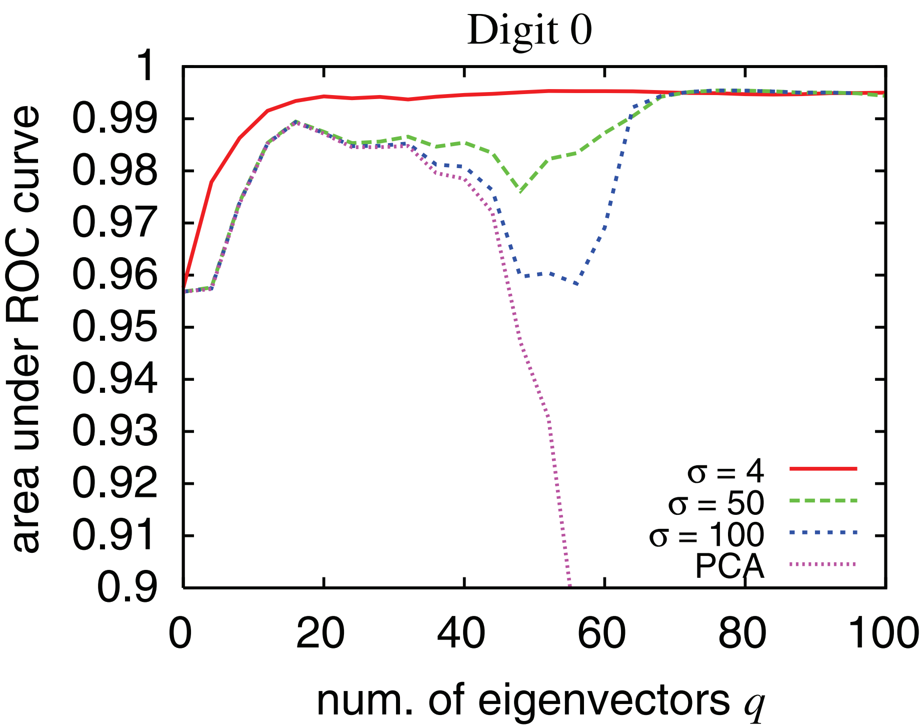


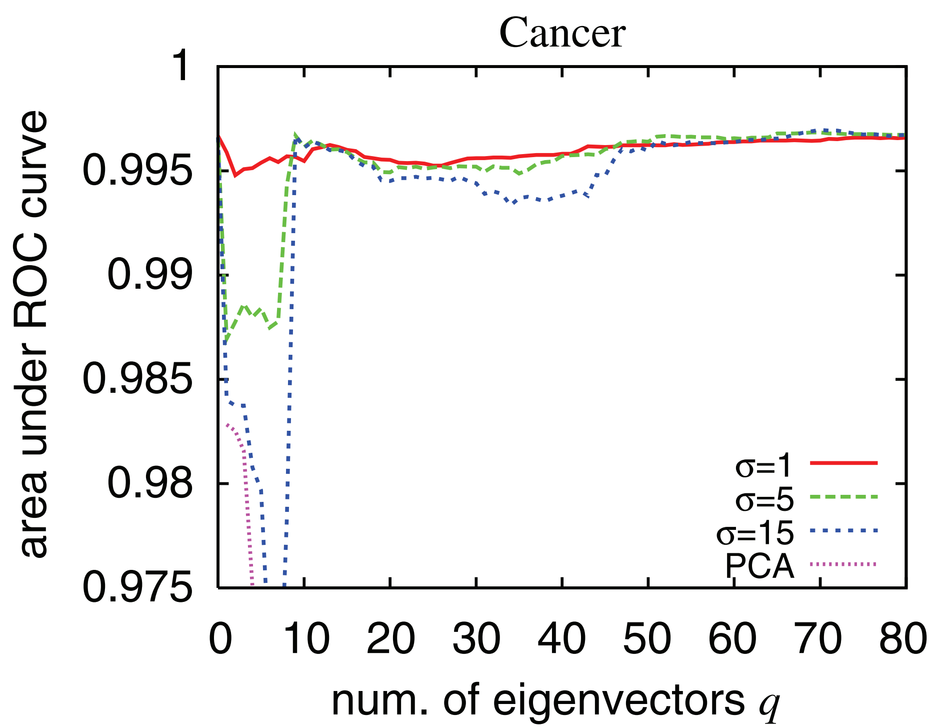
Real Data: vary kernel width





Real Data: vary # eigenvectors





Most Unusual Zero Digits

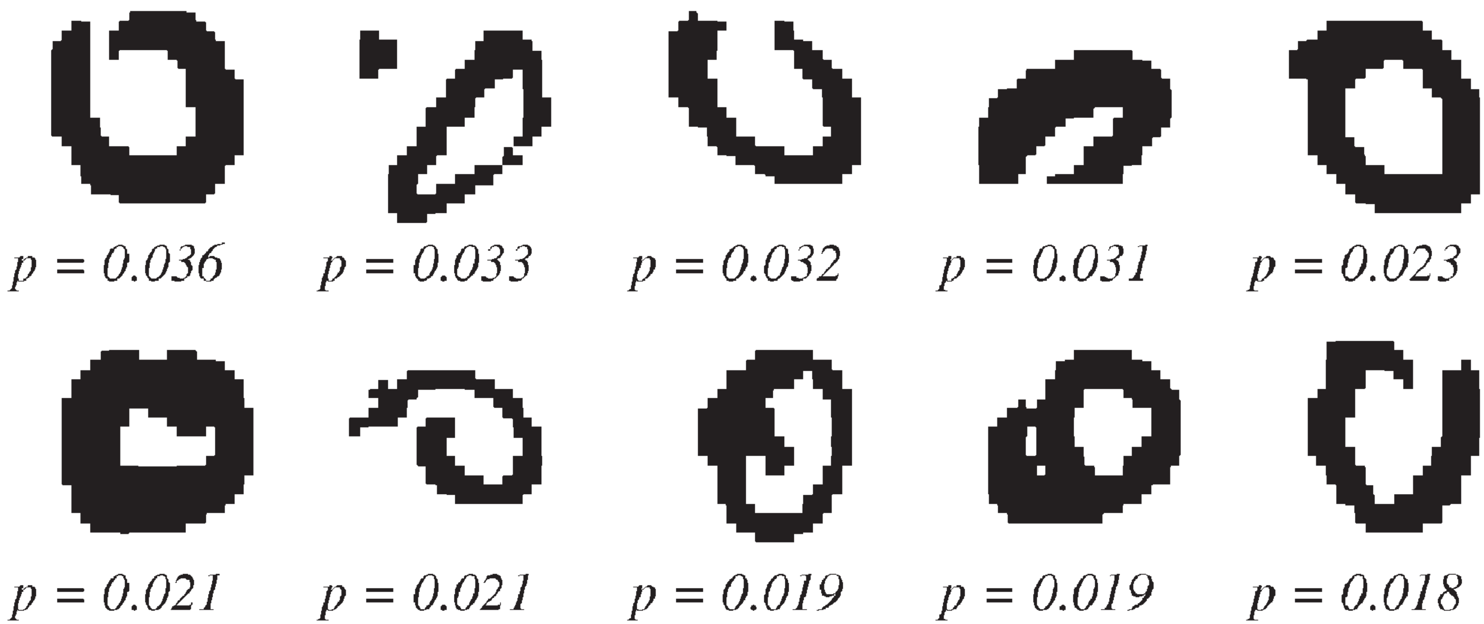


Fig. 11. The 10 most unusual '0' digits from the MNIST test set. The digits are arranged in descending order of their reconstruction error p ($\sigma = 4$, $q = 100$). The figure shows the unprocessed digits of size 28×28 pixels; for novelty detection, however, the processed digits (8×8 pixels) were used.

Diagram

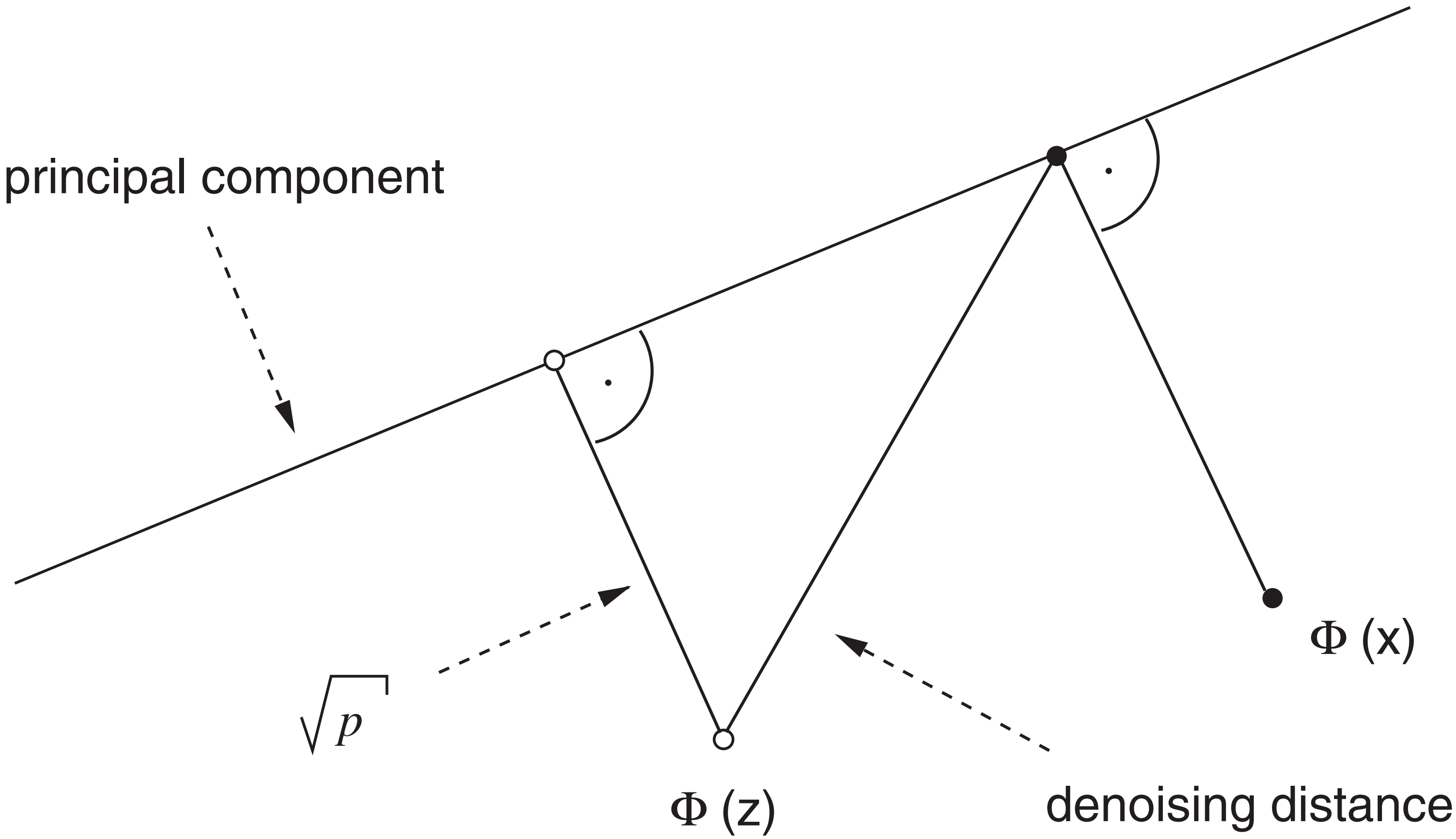


Fig. 12. The difference between the distance to be optimized in denoising and the reconstruction error p.