Classifiers in Almost Empty Spaces

Robert P.W.Duin

Pattern Recognition Group, Department of Applied Physics

Delft University of Technology, The Netherlands

Barcelona, September 2000

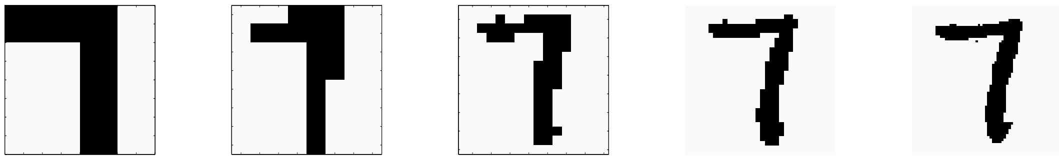
P.O. Box 5046, 2600GA Delft, The Netherlands.
Phone: +(31) 15 2786143, FAX: +(31) 15 2786740,
E-mail: duin@ph.tn.tudelft.nl

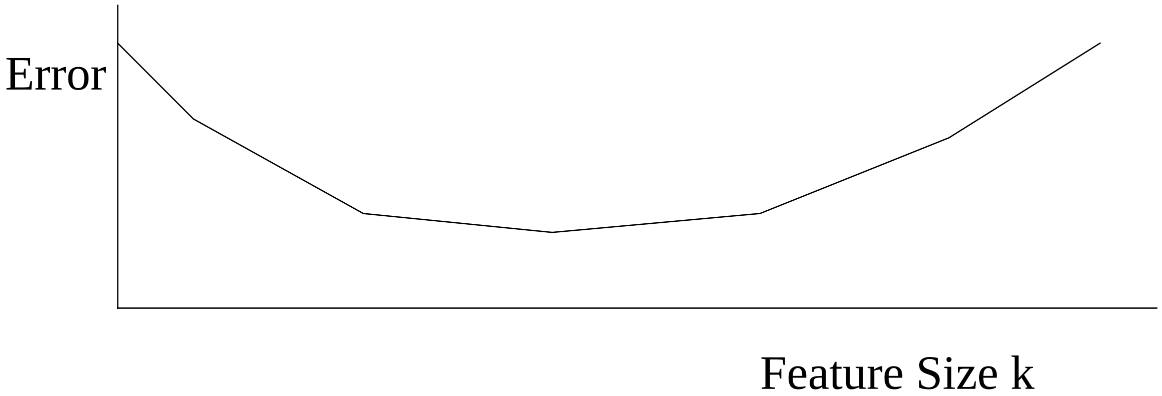
Statistical PR Paradox

$\mathbf{x} = (x^1, x^2, \ldots, x^k)$ - k dimensional feature space

$$\left.\begin{array}{l}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} \text{ - training set} \\ \{\lambda_1, \lambda_2, \ldots, \lambda_m\} \text{ - class labels}\end{array}\right\} D(\mathbf{x}) \text{ - classifier }, \varepsilon = \text{Prob} (D(\mathbf{x}) \neq \lambda(\mathbf{x}))$$

$\varepsilon(m)$: monotonically decreasing , $\varepsilon(k)$: peaks !





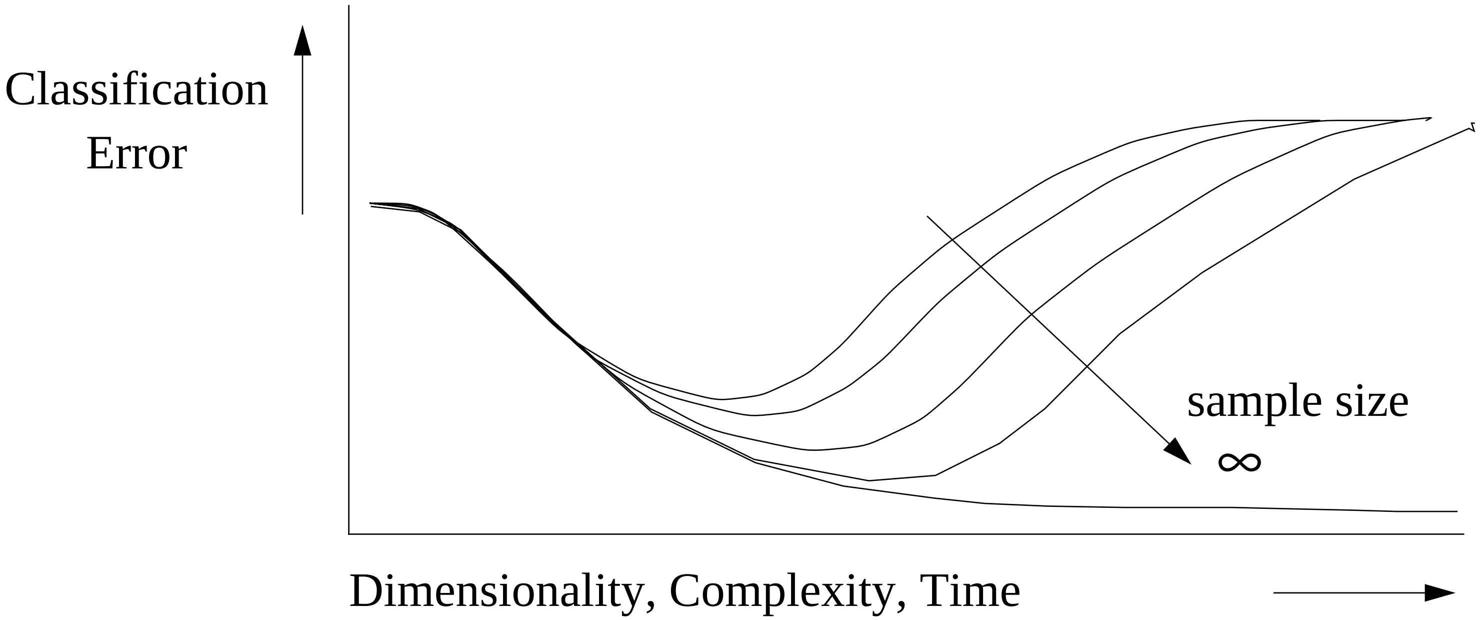
Contents

- Peaking, Curse of Dimensionality, Overtraining
- Pseudo Fisher Linear Discriminant Experiments
- Representation Sets and Kernel Mapping
- Support Vector Classifier
- Dissimilarity Based Classifier
- Subspace Classifier
- Conclusions

Old Peaking Examples

- De Dombal, 1971
- Ullman, 1969
- Raudys, 1976
- Jain and Waller, 1978

Peaking, Curse of Dimensionality, Overtraining

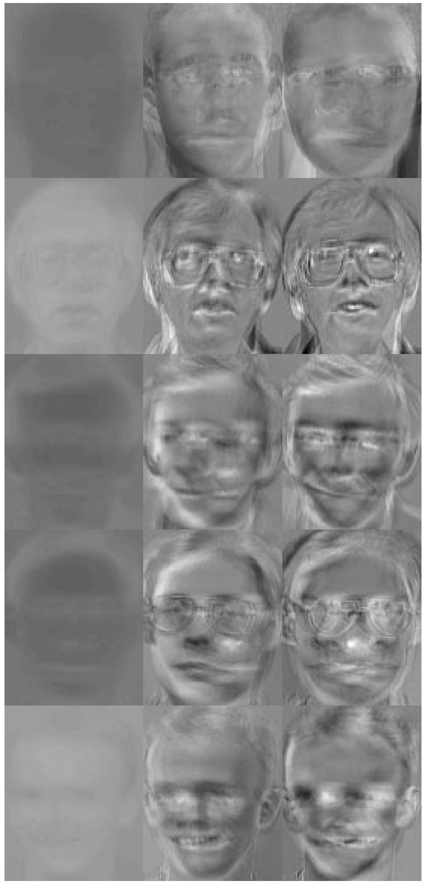


Asymptotically increasing classification error due to:

- Increasing Dimensionality — *Curse of Dimensionality*
- Increasing Complexity — *Peaking Phenomenon*
- Decreasing Regularization
- Increasing Computational Effort

} *Overtraining*

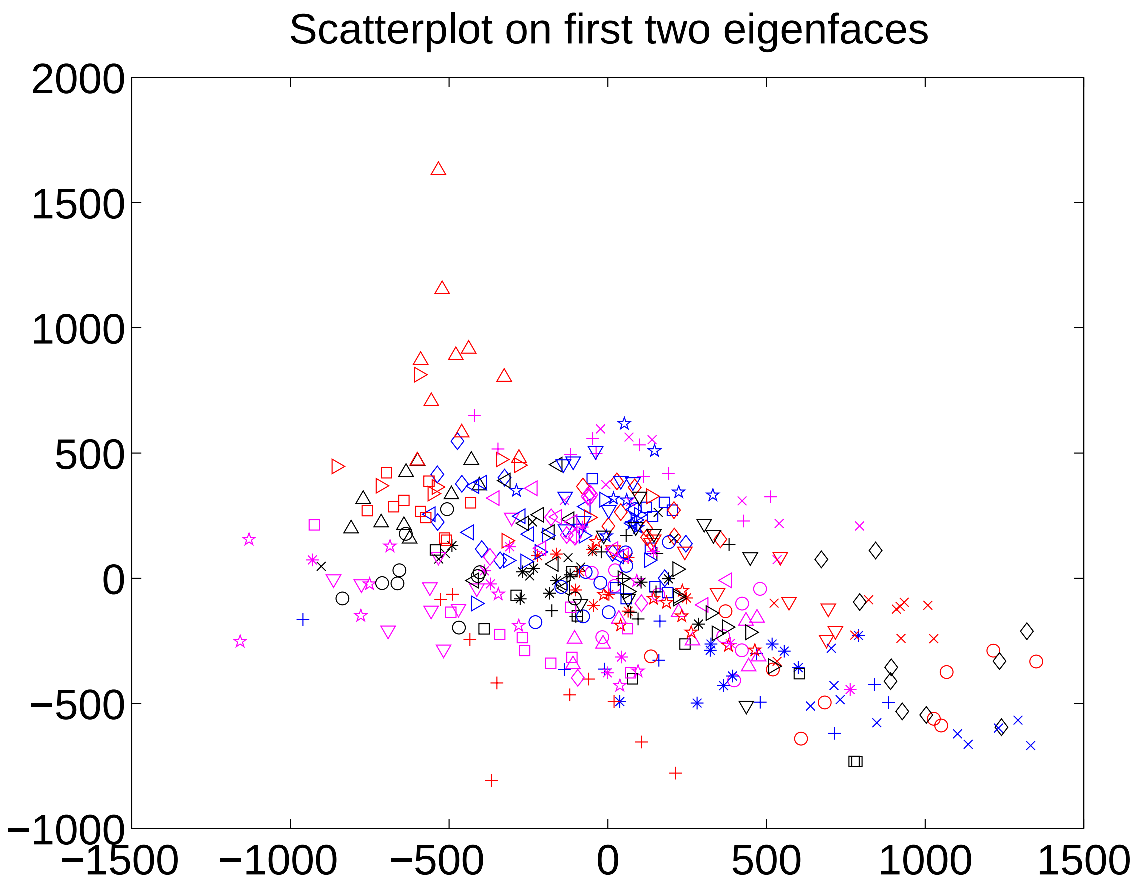
Eigenfaces

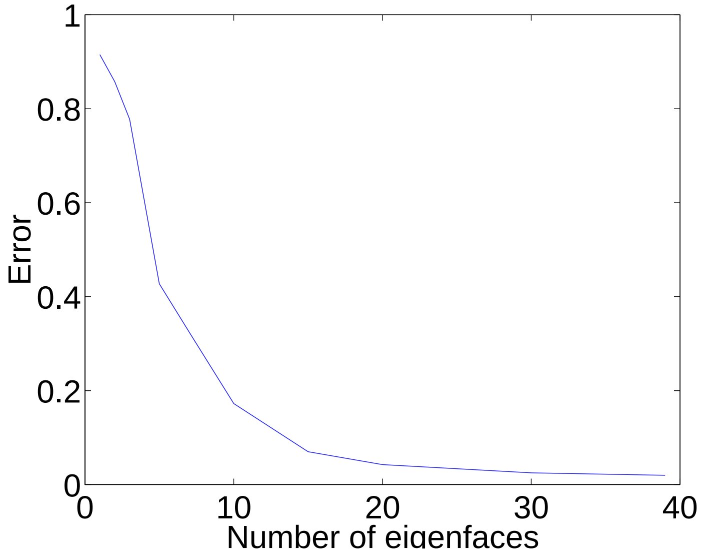


10 pictures of 5 subjects

eigenfaces 1 - 3

PCA Classification of Faces



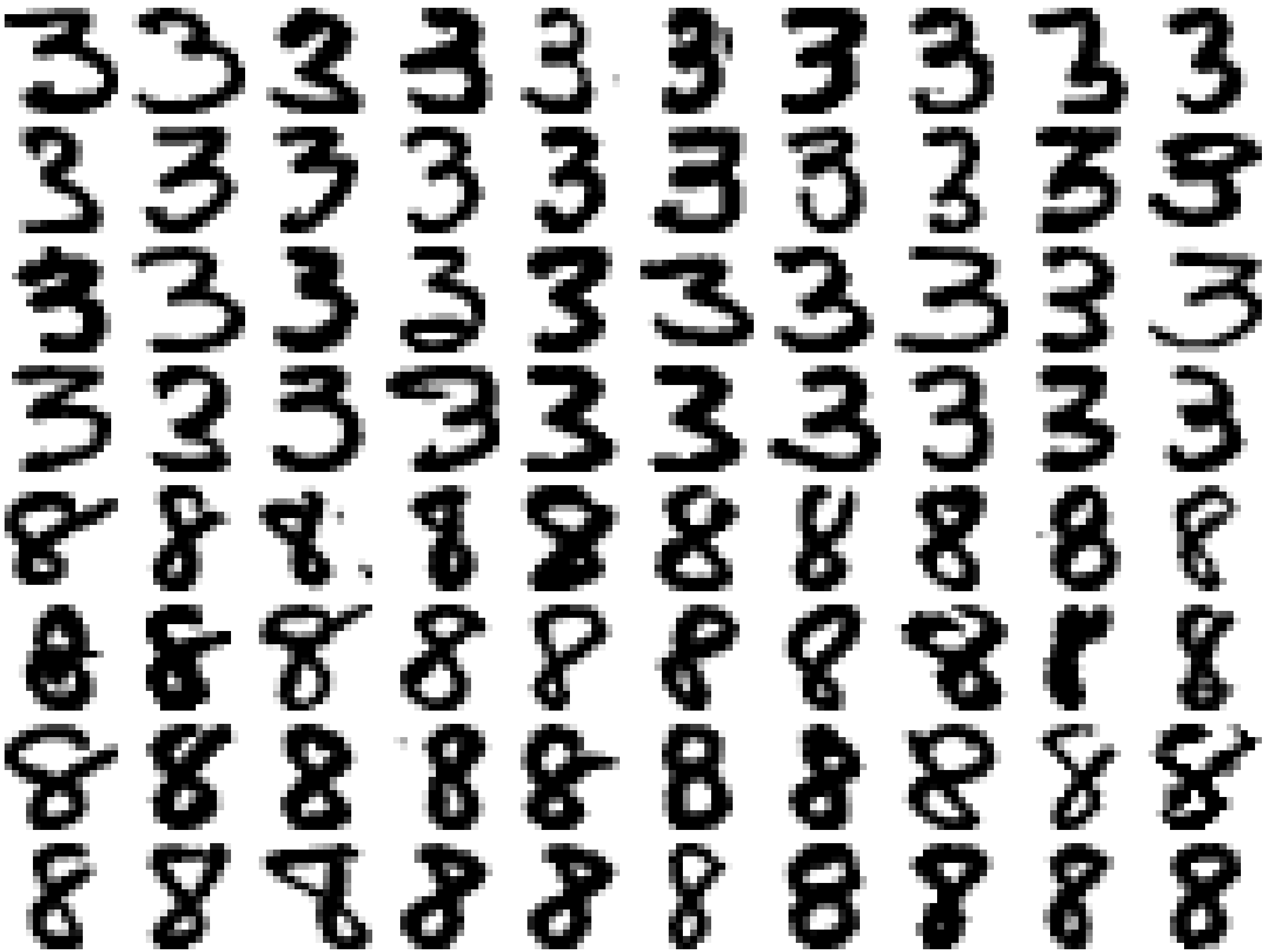


Training Set: 1 image, 40 persons

Feature Size: 92 x 112 = 10304

Test Set: 9 * 40 = 360;

Normalized NIST Data



2 x 2000 Characters

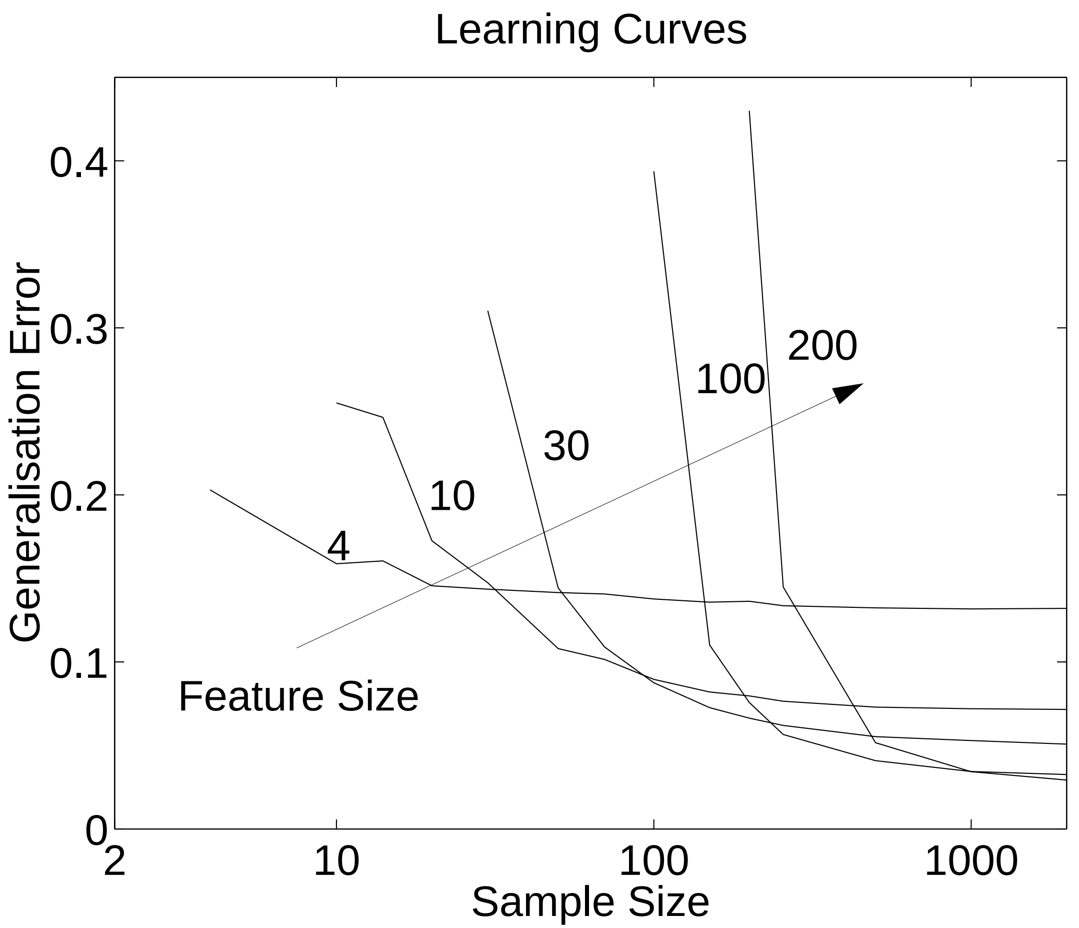
Random Subsets:

- 2 x 1000 Training
- 2 x 1000 Testing

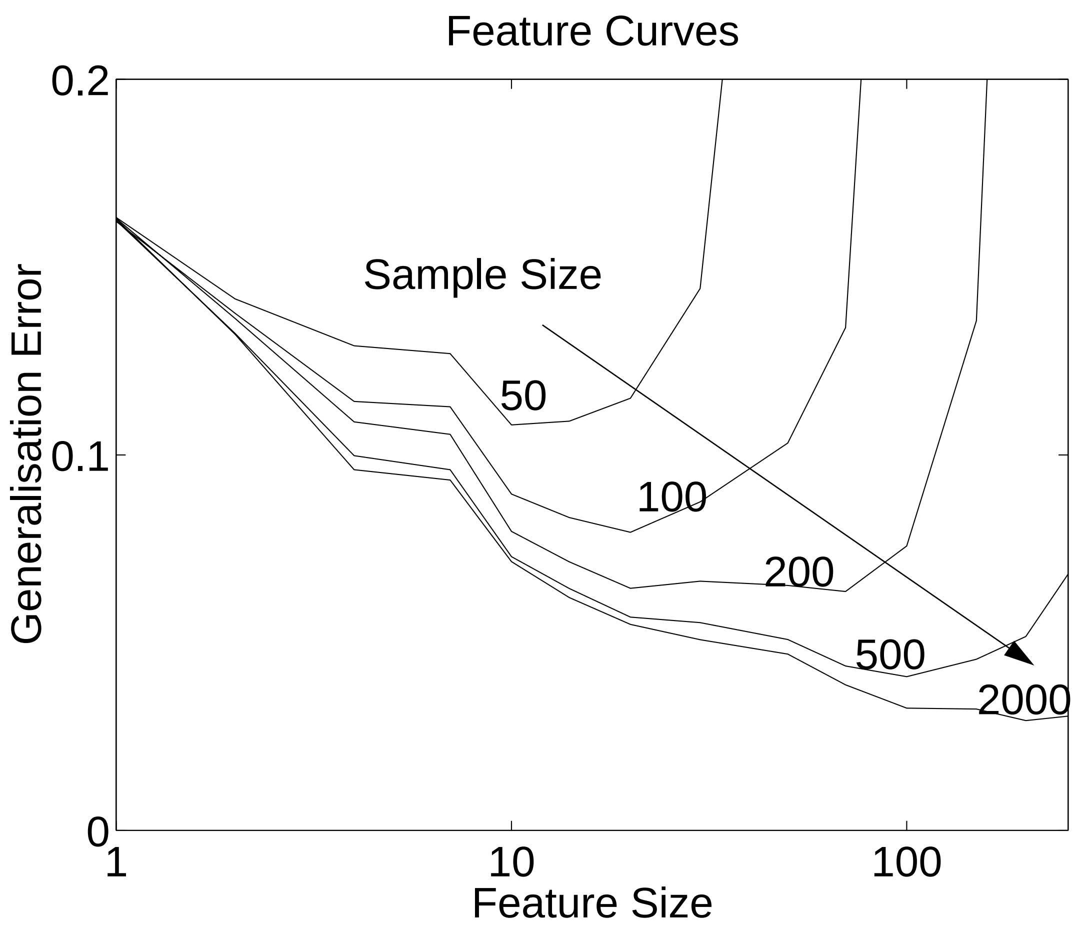
Errors averaged over

50 experiments

Fisher Results



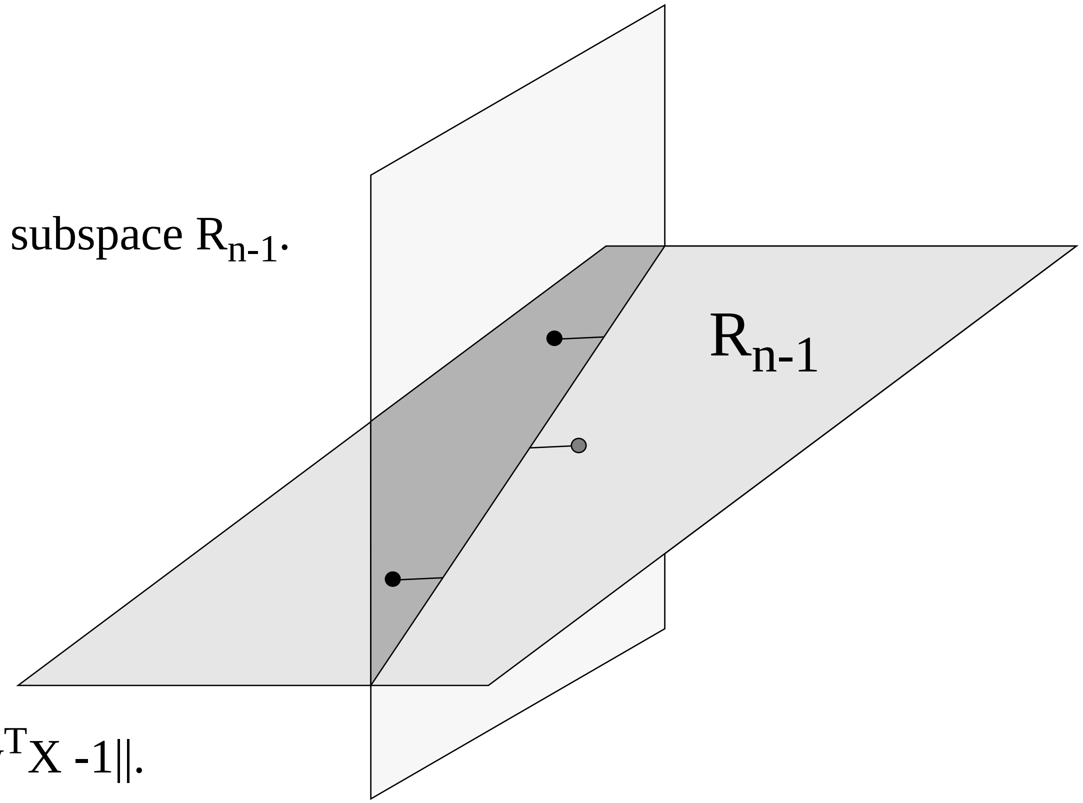
Peaking of the generalization error of FLD as a function of the feature size.



Learning curves of the FLD .

Pseudo Fisher Linear Discriminant

n points in R_k are in a (n-1) dimensional subspace R_{n-1}.



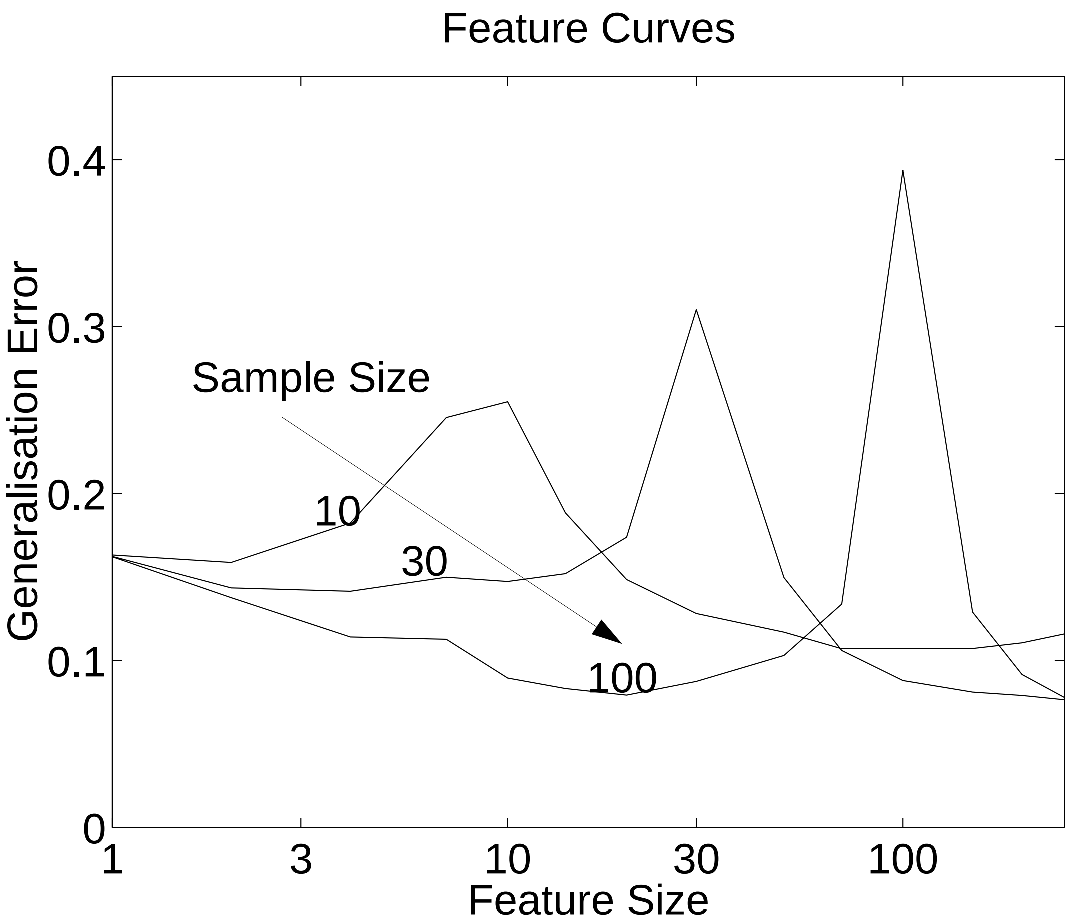
$\rightarrow$ w is the minimum norm solution of $\|w^T X - 1\|$.

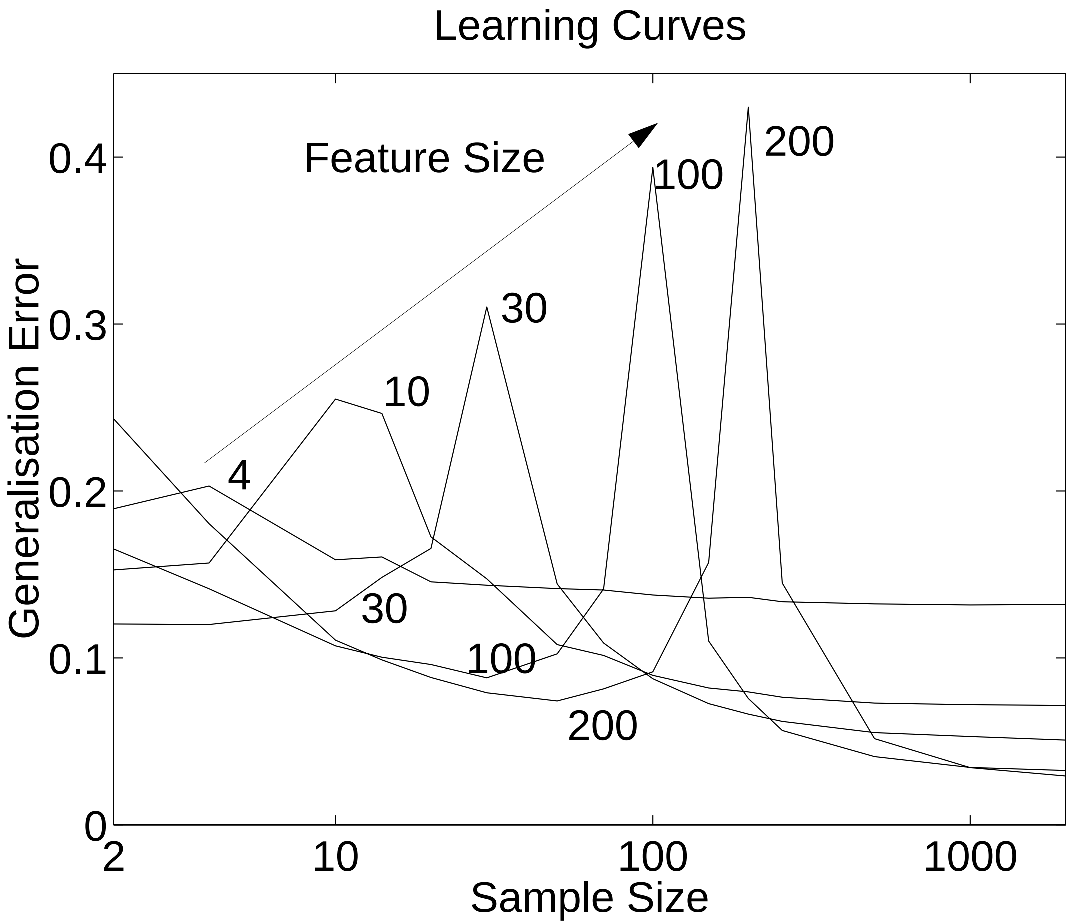
$\rightarrow$ Use Moore-Penrose pseudo-inverse: $w^T = \mathrm{pinv}(X)$.

For n > k the same formula defines Fisher's Discriminant.

$X = \{ (x_i,1),\ x_i \in A,\ (-x_j,-1),\ x_j \in B \}$

Feature Curves and Learning Curves

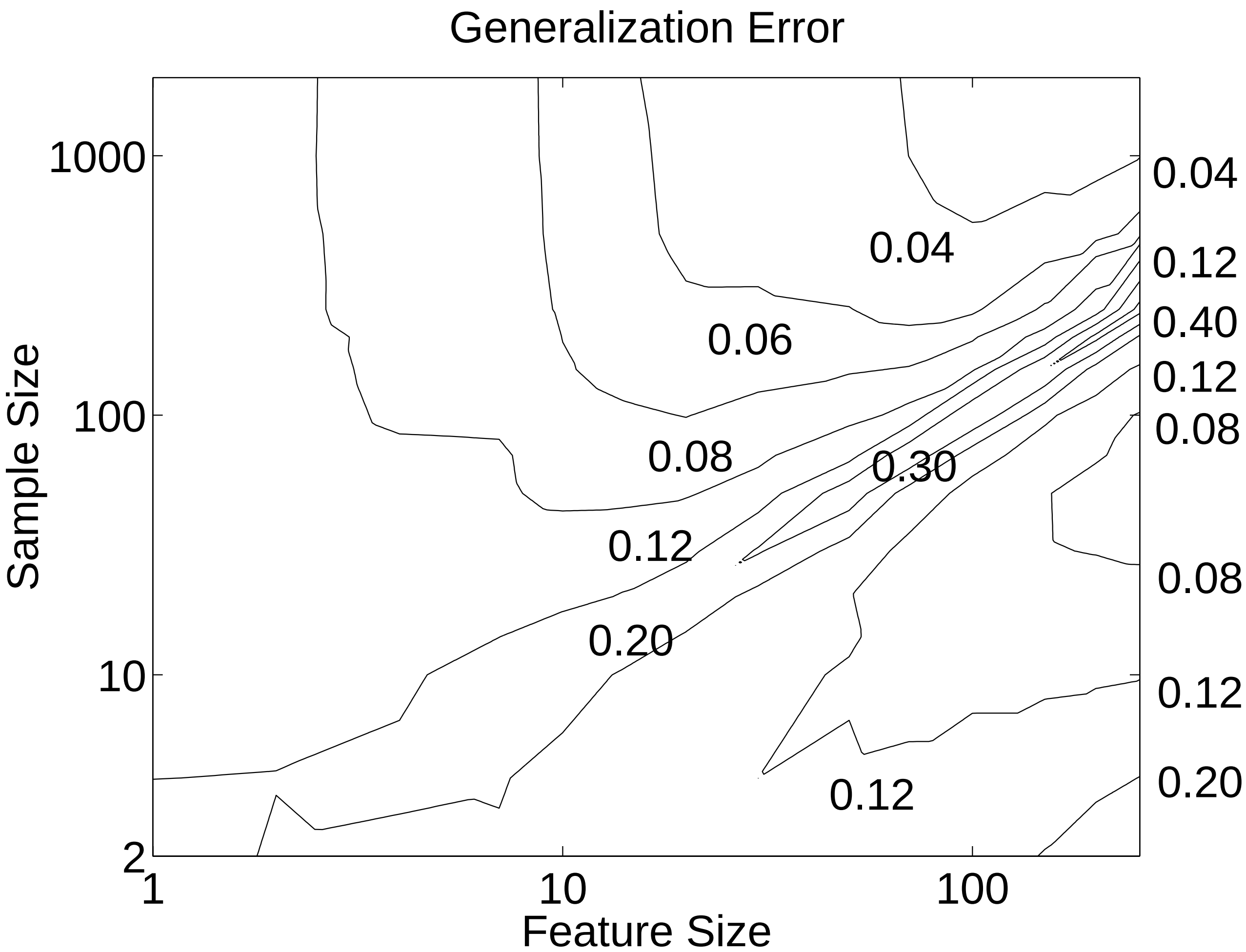




The generalization error of the PFLD as a function of the feature size.

Learning curves of the PFLD.

Feature Size <--> Sample Size Error



The generalization error of the PFLD as a function of feature size and sample size.

Improving Pseudo Fisher Linear Discriminant (PFLD)

PFLD is for dimensionalities > sample size fully overtrained.

Still good results are possible (ε = 0.08, 30 objects in 256 D)

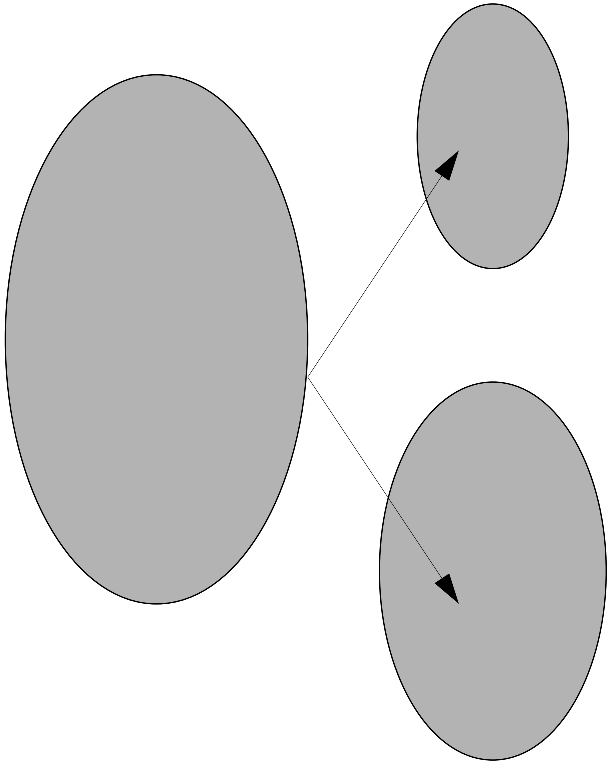
--> Better results are possible

- Regularization e.g. by $D(x) = (\hat{\mu}_A - \hat{\mu}_B)^T (\hat{G} + \lambda I)^{-1} x$

- Change of representation to lower dimensional spaces.

Representation Sets and Kernel Mapping

Representation Set
$Y = \{\mathbf{y}_i\}, \|Y\| = n$



Training Set $X = \{\mathbf{x}_i\}$
Possibly $Y \subset X$

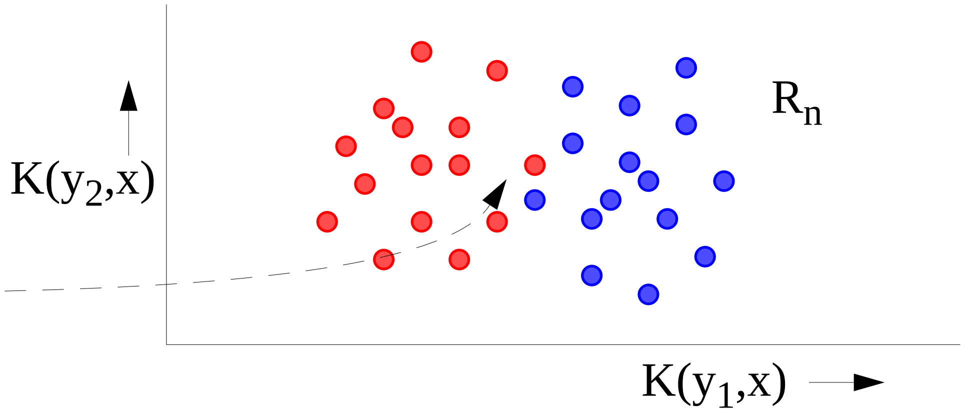
$K = K(Y,\mathbf{x}) = (K(\mathbf{y}_i,\mathbf{x}),\ i=1, \ldots, n),\ \ K \in R_n$

maps an arbitrary object $\mathbf{x}$ into R_n

Polynomials: $K(\mathbf{y}_i, \mathbf{x}) = (\mathbf{x} \bullet \mathbf{y}_i + 1)^p$

Gaussians: $K(\mathbf{y}_i, \mathbf{x}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{y}_i\|^2}{2\sigma^2}\right)$

--> Nonlinear Mapping



Representation Sets

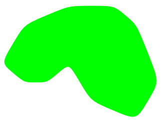
- Dimensionality is controlled by the size of the Representation Set Y.

- Original objects may have arbitrary representation (feature size), just K(**y,x**) has to be defined.

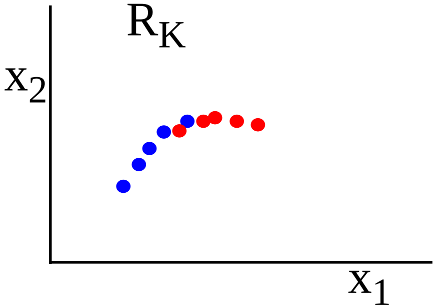
- In the feature space defined by the Representation Set, traditional classifiers may be used.

- Problems: choice of Y, choice of K

Feature Approach and Representation Sets



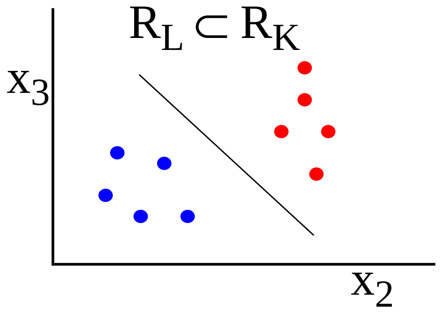
object



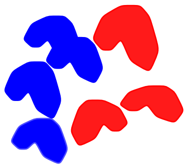
feature space

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ x_{31} & x_{32} & \cdots & x_{3k} \\ x_{41} & x_{42} & \cdots & x_{4k} \end{pmatrix}$$

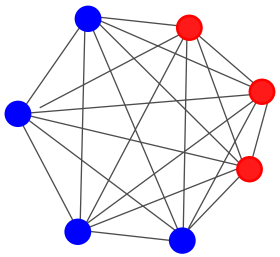
subspace selection



classifier



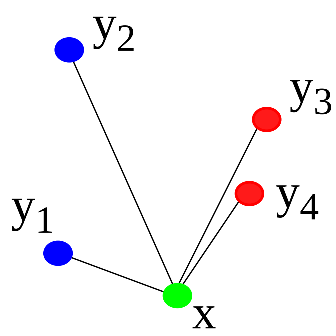
set of objects



relation matrix (kernel K(x,x))

$$\begin{array}{ccc} & \qquad\qquad Y & \\ K = & \begin{pmatrix} k_{11} & k_{12} & k_{13} & k_{14} \\ k_{21} & k_{22} & k_{23} & k_{24} \\ k_{31} & k_{32} & k_{33} & k_{34} \\ k_{41} & k_{42} & k_{43} & k_{44} \end{pmatrix} & X \end{array}$$

selection of Representation Set Y



kernel based classification

Support Vector Classifier

$$\begin{array}{c} Y \\ K = \begin{pmatrix} k_{11}k_{12}k_{13}k_{14}k_{15}k_{16} \\ k_{21}k_{22}k_{23}k_{24}k_{25}k_{26} \\ k_{31}k_{32}\boxed{k_{33}k_{34}}k_{35}k_{36} \\ k_{41}k_{42}\boxed{k_{43}k_{44}}k_{45}k_{46} \\ k_{51}k_{52}k_{53}k_{54}k_{55}k_{56} \\ k_{61}k_{62}k_{63}k_{64}k_{65}k_{66} \end{pmatrix} \end{array} X$$

Reduce training set X to

minimum size 'support set' Y

such that

if used as Representation Set,

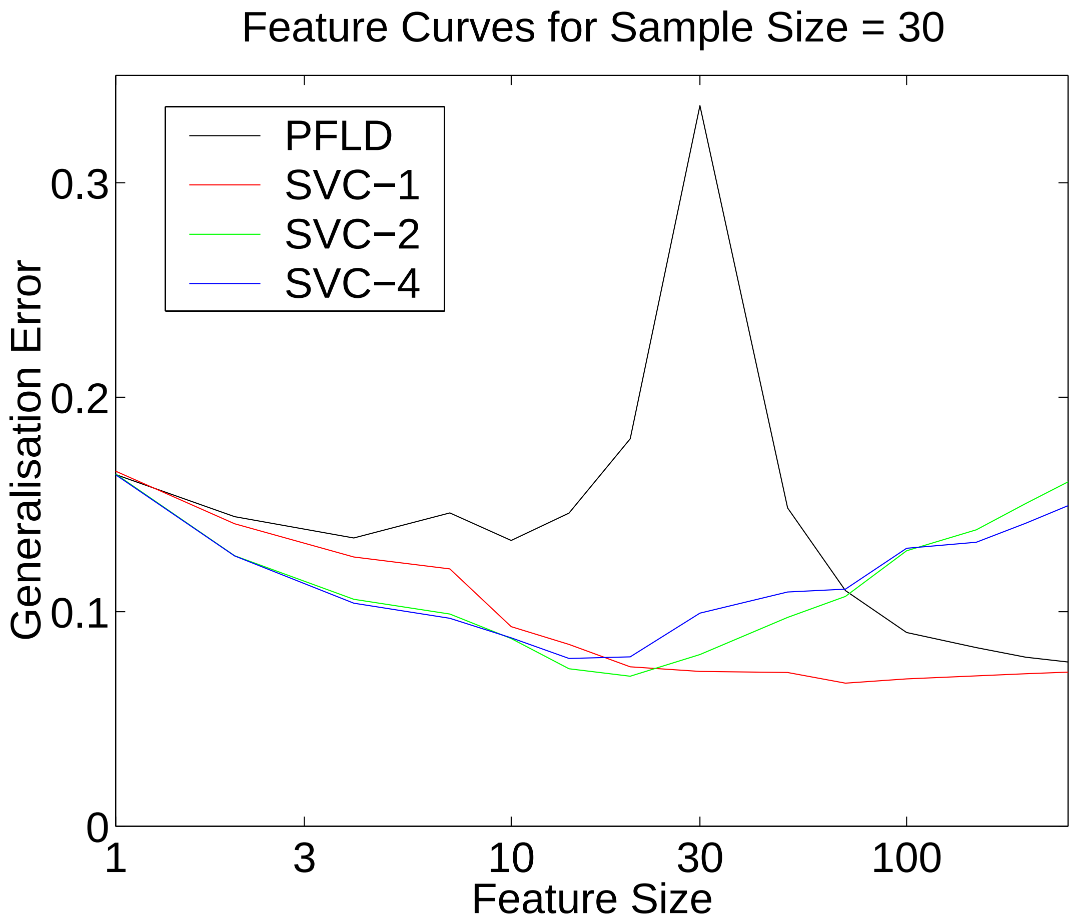
X is error free classified:

$$\min_{\|Y\|}[\text{classf\_error}(K(Y,X))]$$

Notes:

- Classifier is written as a function of n points in R_n
- Not all kernels allowed (Mercer's theorem)

Support Vector Classifier Results



The generalization errors of the PFLD and the SVC as a function of the feature size for a sample size of 30. In the SVC polynomial kernels are used of the orders 1,2 and 4. Number of support vectors: 10 - 30.

Dissimilarity Based Classification

$$K = \begin{array}{c} Y \\ \begin{pmatrix} k_{11} & k_{12} & k_{13} & k_{14} & k_{15} & k_{16} \\ k_{21} & k_{22} & k_{23} & k_{24} & k_{25} & k_{26} \\ k_{31} & k_{32} & k_{33} & k_{34} & k_{35} & k_{36} \\ k_{41} & k_{42} & k_{43} & k_{44} & k_{45} & k_{46} \\ k_{51} & k_{52} & k_{53} & k_{54} & k_{55} & k_{56} \\ k_{61} & k_{62} & k_{63} & k_{64} & k_{65} & k_{66} \end{pmatrix} \end{array} \quad X$$

(Random) selection of Representation Set Y.

All objects X are used for training.

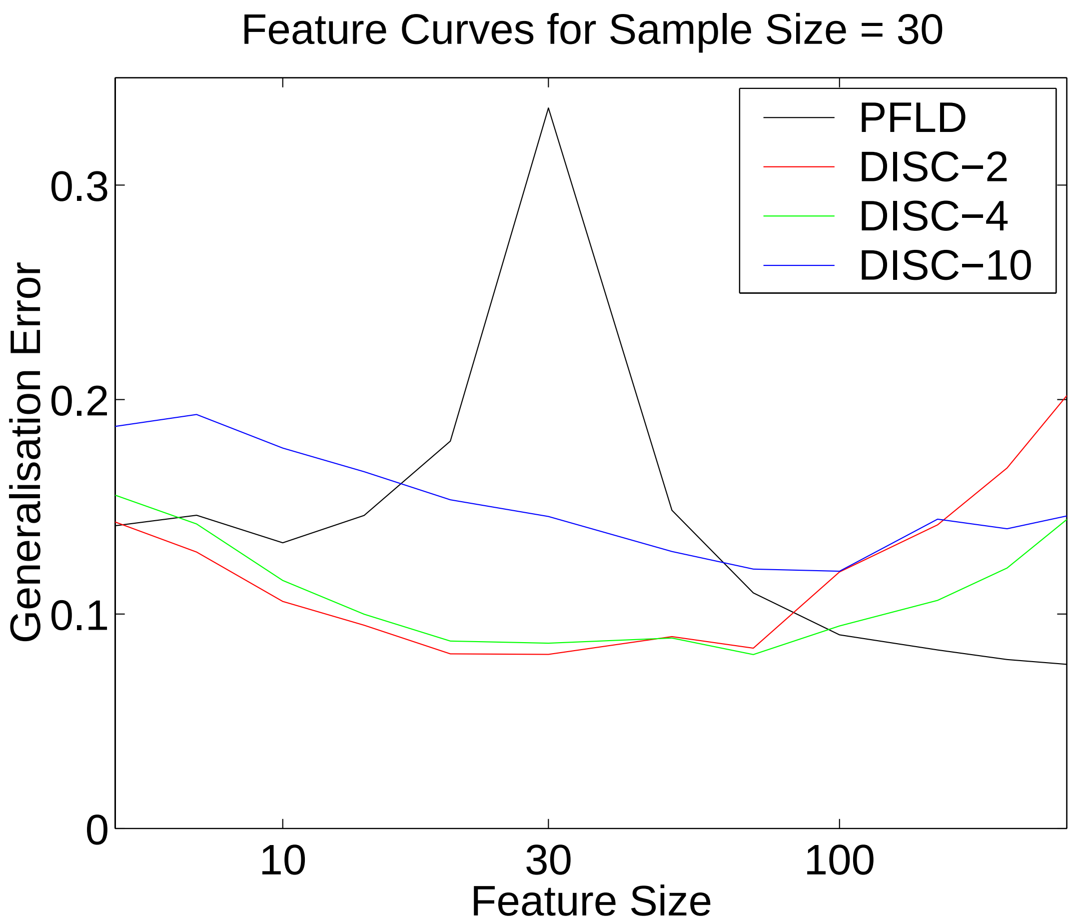
Any kernel $K(\mathbf{y},\mathbf{x})$ is allowed (e.g. $\|\mathbf{x} - \mathbf{y}\|$)

Fast training (simple selection of Y).

Possibly fast testing (choose small Y).

E. Pekalska et al., Classifiers for dissimilarity-based pattern recognition, ICPR15

Dissimilarity Based Classification Results



The generalization errors of the PFLD and a linear dissimilarity based classifier (DISC) as a function of the feature size, using a sample size of 30. For DISC three sizes of the representation set are used: 2, 4 and 10.

Subspace Classifier

$$
\begin{array}{cc}
 & Y \\
K = & \begin{pmatrix} k_{11} k_{12} k_{13} k_{14} k_{15} k_{16} \\ k_{21} k_{22} k_{23} k_{24} k_{25} k_{26} \\ k_{31} k_{32} k_{33} k_{34} k_{35} k_{36} \\ k_{41} k_{42} k_{43} k_{44} k_{45} k_{46} \\ k_{51} k_{52} k_{53} k_{54} k_{55} k_{56} \\ k_{61} k_{62} k_{63} k_{64} k_{65} k_{66} \end{pmatrix} X
\end{array}
$$

$K' = PCA(K)$

Training Set X equals Representation Set Y.

Dimension reduction per class by PCA.

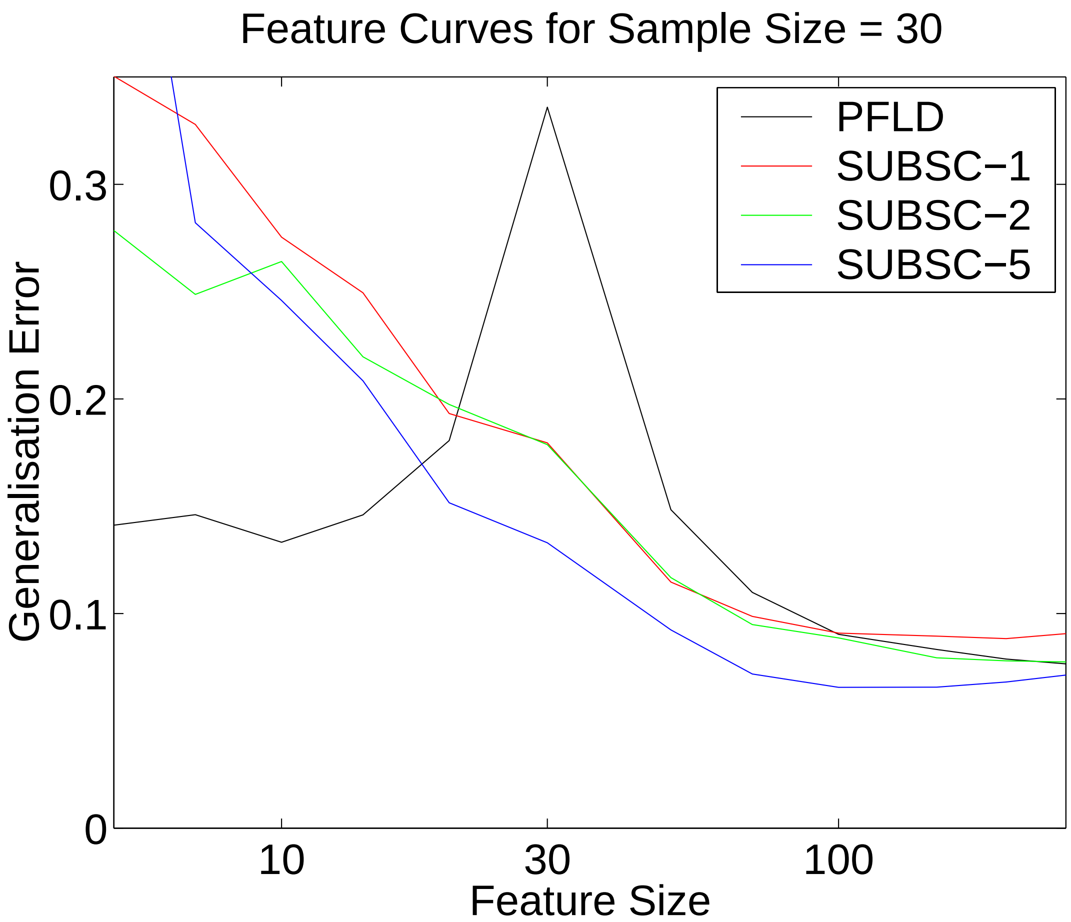
Classification by nearest subspace.

Compare Eigenface method (linear subspace).

Compare feature extraction (no selection).

Test objects have to be compared with entire training set (not true for linear inner product kernel).

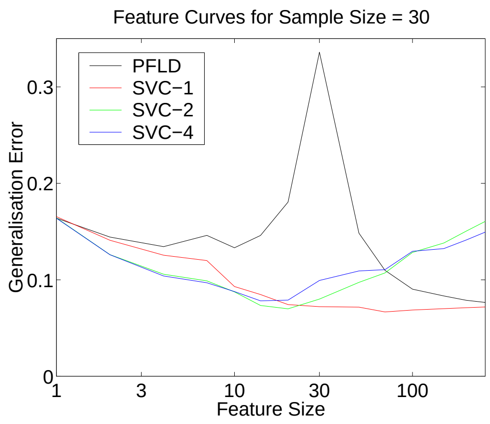
Subspace Classifier Results

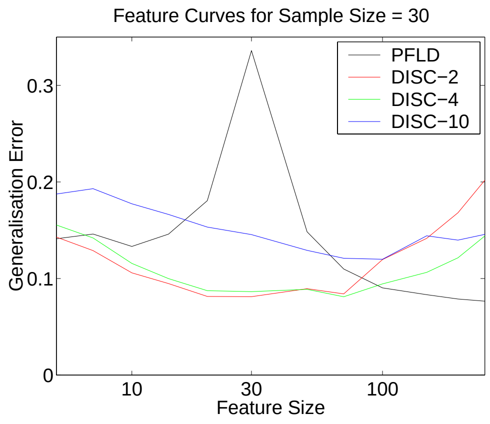


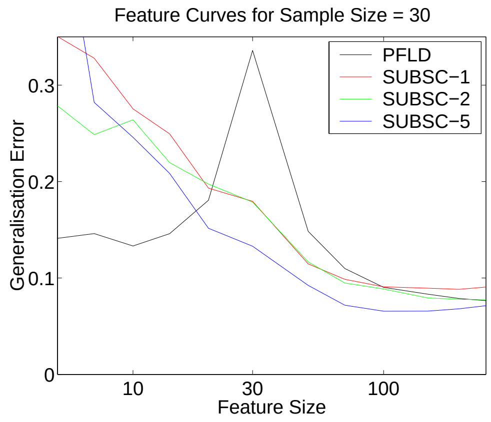
The generalization errors of the PFLD and the subspace classifier (SUBSC) as a function of the feature size for a sample size of 30.

For SUBSC three subspace dimensionalities per class are used: 1, 2 and 5.

Summary







	Support Vector Classifier	Dissimilarity based Classification	Subspace Classifier
Representation Set Selection	Optimized on Minimum Error	Heuristics Free Choice of n	None
Dimension of Representation	n	n	(n=) k
Size of Final Training Set	n	k	k
Training Effort	high	low	moderate
Test Effort	O(n)	O(n)	O(k)

Training Set X (size k); Representation Set Y (size n); $n < k$ ($n << k$), $Y \subset X$

Conclusion

The use of Kernel based Representation Sets allows for the construction of generalizable, nonlinear classifiers in very high-dimensional feature spaces based on relatively small training sets (i.e. size lower than the dimensionality.