Primal and Dual Model Representations in Supervised and Unsupervised Kernel-based Learning

Johan Suykens

K.U. Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.be
http://www.esat.kuleuven.be/scd/

Learning Theory and Approximation
MFO July 2008
Contents

• Context: support vector machines and kernel based learning

• Primal and dual model representations in supervised and unsupervised learning

• Core models, constructive approximation and learning

• Examples

• Sparse approximation

• Robustness and learning theory
Support vector machines and kernel methods: context

- New technologies (e.g. microarrays, proteomics): massive data sets, high dimensional.

- Tasks and objectives: predictive modelling, knowledge discovery, data fusion, classification, feature selection, prior knowledge incorporation, correlation analysis, ranking, ...

- Supervised, unsupervised or semi-supervised

- Models able to handle different data types (numerical, categorical, sequences, graphs ...)

- Linear as well as nonlinear models

- Reliable methods: numerically, computationally, statistically
Kernel based learning: interdisciplinary challenges

neural networks

data mining

linear algebra

pattern recognition

mathematics

machine learning

statistics

optimization

signal processing

SVM & Kernel Methods

systems and control theory

MFO 2008 ◊ Johan Suykens
Different views on kernel based models

- SVM
- LS–SVM
- Kriging
- RKHS
- Gaussian Processes

Some early history on RKHS:

1910-1920: Moore
1940: Aronszajn
1951: Krige
1970: Parzen
1971: Kimeldorf & Wahba

Obtaining complementary insights from different perspectives:

kernels are used in different methodologies

Support vector machines (SVM):
optimization approach (primal/dual)
Reproducing kernel Hilbert spaces (RKHS):
variational problem, functional analysis
Gaussian processes (GP):
probabilistic/Bayesian approach
Estimation in Reproducing Kernel Hilbert Spaces (RKHS)

- **Variational problem:** [Wahba, 1990; Poggio & Girosi, 1990; Evgeniou et al., 2000]
  find function $f$ such that

\[
\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \|f\|_K^2
\]

with $L(\cdot, \cdot)$ the loss function. $\|f\|_K$ is norm in RKHS $\mathcal{H}$ defined by $K$.

- **Representer theorem:** for convex loss function, solution of the form

\[
f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)
\]

Reproducing property $f(x) = \langle f, K_x \rangle_K$ with $K_x(\cdot) = K(x, \cdot)$

- **Some special cases:**
  
  \[L(y, f(x)) = (y - f(x))^2: \text{regularization network}\]
  
  \[L(y, f(x)) = |y - f(x)|_\epsilon: \text{SVM regression with } \epsilon\text{-insensitive loss function}\]
**SVM classifier: primal and dual problem**

- **Primal problem**: [Vapnik, 1995]

\[
\min_{w, b, \xi} \mathcal{J}(w, \xi) = \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i \quad \text{s.t.} \begin{cases} y_i [w^T \varphi(x_i) + b] \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1, \ldots, N \end{cases}
\]

Trade-off between margin maximization and tolerating misclassifications (training set \( \{(x_i, y_i)\}_{i=1}^{N}, x_i \in \mathbb{R}^p, y_i \in \{-1, 1\} \), feature map \( \varphi(\cdot): \mathbb{R}^p \rightarrow \mathbb{R}^{nh} \))

- Conditions for optimality from Lagrangian. Express the solution in the Lagrange multipliers.

- **Dual problem**: QP problem (convex problem)

\[
\max_{\alpha} Q(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{N} y_i y_j K(x_i, x_j) \alpha_i \alpha_j + \sum_{j=1}^{N} \alpha_j \quad \text{s.t.} \begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq c, \quad \forall i \end{cases}
\]

MFO 2008 ◇ Johan Suykens
SVM classifier model representations

• Classifier: primal representation

\[ y(x) = \text{sign}[w^T \varphi(x) + b] \]

Positive definite kernel: \( K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \) (Mercer Theorem)

• Dual representation:

\[ y(x) = \text{sign}\left[ \sum_i \alpha_i y_i K(x, x_i) + b \right] \]

Some possible kernels:
- \( K(x, x_i) = x_i^T x \) (linear)
- \( K(x, x_i) = (x_i^T x + \tau)^d \) (polynomial)
- \( K(x, x_i) = \exp(-\|x - x_i\|_2^2/\sigma^2) \) (RBF)
- \( K(x, x_i) = \tanh(\kappa x_i^T x + \theta) \) (MLP)

• Sparseness property (many \( \alpha_i = 0 \))
SVMs: living in two worlds ...

Primal space:

\[ y(x) = \text{sign}(w^T \varphi(x) + b) \]

Dual space:

\[ y(x) = \text{sign}(\sum_{i=1}^{\#sv} \alpha_i y_i K(x, x_i) + b) \]

"Kernel trick"

\[ K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \]
Wider use of the “kernel trick”

- **Angle between vectors:** (e.g. correlation analysis)
  
  **Input space:**
  \[
  \cos \theta_{xz} = \frac{x^T z}{\|x\|_2 \|z\|_2}
  \]
  
  **Feature space:**
  \[
  \cos \theta_{\varphi(x),\varphi(z)} = \frac{\varphi(x)^T \varphi(z)}{\|\varphi(x)\|_2 \|\varphi(z)\|_2} = \frac{K(x, z)}{\sqrt{K(x, x)\sqrt{K(z, z)}}}
  \]

- **Distance between vectors:** (e.g. for “kernelized” clustering methods)
  
  **Input space:**
  \[
  \|x - z\|_2^2 = (x - z)^T(x - z) = x^T x + z^T z - 2x^T z
  \]
  
  **Feature space:**
  \[
  \|\varphi(x) - \varphi(z)\|_2^2 = K(x, x) + K(z, z) - 2K(x, z)
  \]
Least Squares Support Vector Machines: “core problems”

- **Regression**

  \[
  \min_{w,b,e} w^T w + \gamma \sum_i e_i^2 \quad \text{s.t.} \quad y_i = w^T \varphi(x_i) + b + e_i, \ \forall i
  \]

- **Classification**

  \[
  \min_{w,b,e} w^T w + \gamma \sum_i e_i^2 \quad \text{s.t.} \quad y_i(w^T \varphi(x_i) + b) = 1 - e_i, \ \forall i
  \]

- **Principal component analysis**

  \[
  \min_{w,b,e} -w^T w + \gamma \sum_i e_i^2 \quad \text{s.t.} \quad e_i = w^T \varphi(x_i) + b, \ \forall i
  \]

- **Canonical correlation analysis/partial least squares**

  \[
  \min_{w,v,b,d,e,r} w^T w + v^T v + \nu_1 \sum_i e_i^2 + \nu_2 \sum_i r_i^2 - \gamma \sum_i e_i r_i \quad \text{s.t.} \quad \left\{ \begin{array}{l} e_i = w^T \varphi_1(x_i) + b \\ r_i = v^T \varphi_2(y_i) + d \end{array} \right. \]

  - partially linear models, spectral clustering, subspace algorithms, ...
Core models + additional constraints

• Monotonicity constraints: [Pelckmans et al., 2005]

\[
\min_{w,b,e} w^T w + \gamma \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad \begin{cases}
y_i = w^T \varphi(x_i) + b + e_i, & (i = 1, \ldots, N) \\
w^T \varphi(x_i) \leq w^T \varphi(x_{i+1}), & (i = 1, \ldots, N - 1)
\end{cases}
\]

• Structure detection: [Pelckmans et al., 2005; Tibshirani, 1996]

\[
\min_{w,e,t} \rho \sum_{p=1}^{P} t_p + \sum_{p=1}^{P} w^{(p)^T} w^{(p)} + \gamma \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad \begin{cases}
y_i = \sum_{p=1}^{P} w^{(p)^T} \varphi^{(p)}(x^{(p)}_i) + e_i, & (\forall i) \\
-t_p \leq w^{(p)^T} \varphi^{(p)}(x^{(p)}_i) \leq t_p, & (\forall i, \forall p)
\end{cases}
\]

• Autocorrelated errors: [Espinoza et al., 2006]

\[
\min_{w,b,r,e} w^T w + \gamma \sum_{i=1}^{N} r_i^2 \quad \text{s.t.} \quad \begin{cases}
y_i = w^T \varphi(x_i) + b + e_i, & (i = 1, \ldots, N) \\
e_i = \rho e_{i-1} + r_i, & (i = 2, \ldots, N)
\end{cases}
\]

• Spectral clustering: [Alzate & Suykens, 2006; Chung, 1997; Shi & Malik, 2000]

\[
\min_{w,b,e} -w^T w + \gamma e^T D^{-1} e \quad \text{s.t.} \quad e_i = w^T \varphi(x_i) + b, \quad (i = 1, \ldots, N)
\]

MFO 2008 ◇ Johan Suykens
**LS-SVM classifier**

- Preserve support vector machine [Vapnik, 1995] methodology, but simplify via least squares and equality constraints [Suykens, 1999]

- **Primal problem:**

  \[
  \min_{w,b,e} \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad y_i [w^T \varphi(x_i) + b] = 1 - e_i, \ i = 1, \ldots, N
  \]

- **Dual problem:**

  \[
  \begin{bmatrix}
  0 & y^T \\
  y & \Omega + I/\gamma
  \end{bmatrix}
  \begin{bmatrix}
  b \\
  \alpha
  \end{bmatrix}
  =
  \begin{bmatrix}
  0 \\
  1_N
  \end{bmatrix}
  \]

  where \( \Omega_{ij} = y_i y_j \varphi(x_i)^T \varphi(x_j) = y_i y_j K(x_i, x_j) \) and \( y = [y_1; \ldots; y_N] \).

- LS-SVM classifiers perform very well on 20 UCI data sets [Van Gestel et al., ML 2004] Winning results in competition WCCI 2006 by [Cawley, 2006]
Kernel principal component analysis (KPCA)

Kernel PCA [Schölkopf et al., 1998]:
take eigenvalue decomposition of the kernel matrix

\[
\begin{bmatrix}
K(x_1, x_1) & \ldots & K(x_1, x_N) \\
\vdots & \ddots & \vdots \\
K(x_N, x_1) & \ldots & K(x_N, x_N)
\end{bmatrix}
\]

(applications in dimensionality reduction and denoising)

Where is the regularization?
Kernel PCA: primal and dual problem

- Underlying primal problem with regularization term [Suykens et al., 2003]
- Primal problem:

$$\min_{w,b,e} -\frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad e_i = w^T \varphi(x_i) + b, \ i = 1, ..., N.$$  

(or alternatively $\min \frac{1}{2} w^T w - \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2$)
- Dual problem = kernel PCA :

$$\Omega c \alpha = \lambda \alpha \quad \text{with} \quad \lambda = 1/\gamma$$

with $\Omega_{c,i,j} = (\varphi(x_i) - \hat{\mu}_\varphi)^T (\varphi(x_j) - \hat{\mu}_\varphi)$ the centered kernel matrix.

- Score variables (allowing also out-of-sample extensions):

$$z(x) = w^T (\varphi(x) - \hat{\mu}_\varphi) = \sum_{j} \alpha_j (K(x_j, x) - \frac{1}{N} \sum_r K(x_r, x) - \frac{1}{N} \sum_r \sum_s K(x_r, x_s))$$
Spectral clustering: weighted KPCA

• Spectral graph clustering [Chung, 1997; Shi & Malik, 2000; Ng et al., 2002]

• Normalized cut problem

\[ Lq = \lambda Dq \]

with \( L = D - W \) the Laplacian of the graph. Cluster membership indicators are given by \( q \).

• Weighted KPCA - LS-SVM formulation to normalized cut:

\[
\min_{w,b,e} -\frac{1}{2} w^T w + \gamma \frac{1}{2} e^T V e \quad \text{such that} \quad e_i = w^T \varphi(x_i) + b, \forall i = 1, ..., N
\]

with \( V = D^{-1} \) the inverse degree matrix [Alzate & Suykens, 2006]. Allows for **out-of-sample extensions** on test data.
Application to image segmentation

Given image \((240 \times 160)\)

Image segmentation

subsampling + training + out-of-sample extension

[Alzate & Suykens, 2006]
• **Primal problem:** [Suykens et al. 2002]

\[
\begin{align*}
\min_{w,v,e,r} & \quad \frac{1}{2} w^T w + \frac{1}{2} v^T v + \nu_1 \frac{1}{2} \sum_{i=1}^{N} e_i^2 + \nu_2 \frac{1}{2} \sum_{i=1}^{N} r_i^2 - \gamma \sum_{i=1}^{N} e_i r_i \\
\text{such that} & \quad e_i = w^T (\varphi_1(x_i) - \hat{\mu}_{\varphi_1}), \; (i = 1, \ldots, N) \\
& \quad r_i = v^T (\varphi_2(y_i) - \hat{\mu}_{\varphi_2}), \; (i = 1, \ldots, N) \\
\end{align*}
\]

with \( \hat{\mu}_{\varphi_1} = (1/N) \sum_{i=1}^{N} \varphi_1(x_i) \), \( \hat{\mu}_{\varphi_2} = (1/N) \sum_{i=1}^{N} \varphi_2(y_i) \).

• **Dual problem:** generalized eigenvalue problem

\[
\begin{bmatrix}
0 & \Omega_{c,2} \\
\Omega_{c,1} & 0
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}
= \lambda
\begin{bmatrix}
\nu_1 \Omega_{c,1} + I & 0 \\
0 & \nu_2 \Omega_{c,2} + I
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}, \; \lambda = 1/\gamma
\]

with \( \Omega_{c,1_{ij}} = (\varphi_1(x_i) - \hat{\mu}_{\varphi_1})^T (\varphi_1(x_j) - \hat{\mu}_{\varphi_1}) \), \( \Omega_{c,2_{ij}} = (\varphi_2(y_i) - \hat{\mu}_{\varphi_2})^T (\varphi_2(y_j) - \hat{\mu}_{\varphi_2}) \)

Kernels \( K_1(x_i, x_j) = \varphi_1(x_i)^T \varphi_1(x_j) \), \( K_2(y_i, y_j) = \varphi_2(y_i)^T \varphi_2(y_j) \)

Related work [Bach & Jordan, 2002; Vert et al., 2003; Shawe-Taylor et al., 2004]
Independent Component Analysis

Measure of independence based on multivariate kernel CCA:

$$\max_{w^{(l)}, e^{(l)}} \gamma \sum_l \sum_{k \neq l} e^{(l)^T} e^{(k)} - \sum_l \nu_l e^{(l)^T} e^{(l)} - \sum_l w^{(l)^T} w^{(l)} \quad \text{s.t.} \quad e^{(l)} = \Phi^{(l)} w^{(l)}, \; l = 1, \ldots, m$$

original IC

observed mixtures
(top-left: train, bottom-right: validation, other: test)

estimated IC

[Alzate & Suykens, NN 2008], related work: [Bach & Jordan, 2002; Gretton et al., 2005]
**System identification of Hammerstein systems**

- **Hammerstein system:**
  \[
  \begin{align*}
  x_{t+1} &= Ax_t + B f(u_t) + \nu_t \\
  y_t &= C x_t + D f(u_t) + \nu_t
  \end{align*}
  \]
  with
  \[
  E\left\{ \begin{bmatrix} \nu_p \\ v_p \end{bmatrix} \begin{bmatrix} \nu_q^T \\ v_q^T \end{bmatrix} \right\} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \delta_{pq}.
  \]

- **System identification problem:**
  given \( \{(u_t, y_t)\}_{t=0}^{N-1} \), estimate \( A, B, C, D, f \).

- **Subspace algorithms** [Goethals et al., IEEE-AC 2005]:
  first estimate state vector sequence (can be done by kernel CCA)
  (for linear systems equivalent to Kalman filtering)

- **Related problems:** linear non-Gaussian models, links with ICA
  Kernels for linear systems and gait recognition [Bissacco et al., 2007]
Semi-supervised learning in RKHS

- Semi-supervised learning in RKHS [Belkin & Niyogi, 2004]:

\[
\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} V(y_i, f(x_i)) + \lambda \|f\|_K^2 + \eta f^T L f
\]

with \(V(\cdot, \cdot)\) loss function, \(L\) Laplacian matrix, \(\|f\|_K\) norm in RKHS \(\mathcal{H}\), \(f = [f(x_1); \ldots; f(x_{N_l+N_u})]\) \((N_l, N_u\) number of labeled and unlabeled data)

- Laplacian term: discretization of the Laplace-Beltrami operator

- Representer theorem: \(f(x) = \sum_{i=1}^{N_l+N_u} \alpha_i K(x, x_i)\)

- Least squares solution case: Laplacian acts on kernel matrix
Formulation by adding constraints

- Semi-supervised LS-SVM model [Luts et al., 2007]:

\[
\begin{align*}
\min_{w, \epsilon, b, \hat{y}} & \quad \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2 + \frac{1}{2} \eta \sum_{i,j=1}^{N} v_{ij}(\hat{y}_i - \hat{y}_j)^2 \\
\text{s.t.} & \quad \hat{y}_i = w^T \varphi(x_i) + b, \quad i = 1, ..., N \\
& \quad \hat{y}_i = \nu_i y_i - e_i, \quad \nu_i \in \{0, 1\}, \quad i = 1, ..., N \\
\end{align*}
\]

where \( \nu_i = 0 \) for unlabeled data, \( \nu_i = 1 \) for labeled data.

- MRI image: healthy tissue versus tumor classification [Luts et al., 2007]
Dimensionality reduction and data visualization

- Traditionally:
  commonly used techniques are e.g. principal component analysis, multi-dimensional scaling, self-organizing maps

- More recently:
  isomap, locally linear embedding, Hessian locally linear embedding, diffusion maps, Laplacian eigenmaps
  (“kernel eigenmap methods and manifold learning”)
  [Roweis & Saul, 2000; Coifman et al., 2005; Belkin et al., 2006]

- Relevant issues:
  - learning and generalization [Cucker & Smale, 2002; Poggio et al., 2004]
  - model representations and out-of-sample extensions
  - convex/non-convex problems, computational complexity [Smale, 1997]

- Kernel maps with reference point (KMref) [Suykens, 2008]:
  data visualization and dimensionality reduction by solving linear system
Kernel maps with reference point: problem statement

- Kernel maps with reference point [Suykens, IEEE-TNN 2008]:
  - **LS-SVM core part**: realize dimensionality reduction $x \mapsto z$
  - **reference point** $q$ (e.g. first point; sacrificed in the visualization)

- Example: $d = 2$

$$\min_{z,w_1,w_2,b_1,b_2,e_{i,1},e_{i,2}} \frac{1}{2}(z - P_D z)^T(z - P_D z) + \frac{\nu}{2}(w_1^T w_1 + w_2^T w_2) + \frac{\eta}{2} \sum_{i=1}^{N}(e_{i,1}^2 + e_{i,2}^2)$$

such that

- $c_{1,1}^T z = q_1 + e_{1,1}$
- $c_{1,2}^T z = q_2 + e_{1,2}$
- $c_{i,1}^T z = w_1^T \varphi_1(x_i) + b_1 + e_{i,1}$, $\forall i = 2, \ldots, N$
- $c_{i,2}^T z = w_2^T \varphi_2(x_i) + b_2 + e_{i,2}$, $\forall i = 2, \ldots, N$

Coordinates in low dimensional space: $z = [z_1; z_2; \ldots; z_N] \in \mathbb{R}^{dN}$

Regularization term: $(z - P_D z)^T(z - P_D z) = \sum_{i=1}^{N} \|z_i - \sum_{j=1}^{N} s_{ij} D z_j\|_2^2$

with $D$ diagonal matrix and $s_{ij} = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$. 

MFO 2008 ◈ Johan Suykens 23
Kernel maps with reference point: solution

- The unique solution to the problem is given by the linear system

\[
\begin{bmatrix}
U & -V_1 M_1^{-1} 1 & -V_2 M_2^{-1} 1 \\
-1^T M_1^{-1} V_1^T & 1^T M_1^{-1} 1 & 0 \\
-1^T M_2^{-1} V_2^T & 0 & 1^T M_2^{-1} 1
\end{bmatrix}
\begin{bmatrix}
z \\b_1 \\ b_2
\end{bmatrix}
=
\begin{bmatrix}
\eta (q_1 c_{1,1} + q_2 c_{1,2}) \\
0 \\
0
\end{bmatrix}
\]

with matrices

\[
U = (I - P_D)^T (I - P_D) - \gamma I + V_1 M_1^{-1} V_1^T + V_2 M_2^{-1} V_2^T + \eta c_{1,1} c_{1,1}^T + \eta c_{1,2} c_{1,2}^T
\]

\[
M_1 = \frac{1}{\nu} \Omega_1 + \frac{1}{\eta} I, \quad M_2 = \frac{1}{\nu} \Omega_2 + \frac{1}{\eta} I
\]

\[
V_1 = [c_{2,1} ... c_{N,1}], \quad V_2 = [c_{2,2} ... c_{N,2}]
\]

kernel matrices \( \Omega_1, \Omega_2 \in \mathbb{R}^{(N-1) \times (N-1)}: \)

\[
\Omega_{1,ij} = K_1(x_i, x_j) = \varphi_1(x_i)^T \varphi_1(x_j), \quad \Omega_{2,ij} = K_2(x_i, x_j) = \varphi_2(x_i)^T \varphi_2(x_j)
\]

positive definite kernel functions \( K_1(\cdot, \cdot), K_2(\cdot, \cdot). \)
Kernel maps with reference point: model representations

• The **primal and dual model representations** allow making **out-of-sample extensions**. Evaluation at point \( x^* \in \mathbb{R}^p \):

\[
\hat{z}_{*,1} = w_1^T \varphi_1(x^*) + b_1 = \frac{1}{\nu} \sum_{i=2}^{N} \alpha_{i,1} K_1(x_i, x^*) + b_1 \\
\hat{z}_{*,2} = w_2^T \varphi_2(x^*) + b_2 = \frac{1}{\nu} \sum_{i=2}^{N} \alpha_{i,2} K_2(x_i, x^*) + b_2
\]

Estimated coordinates for visualization: \( \hat{z}_* = [\hat{z}_{*,1}; \hat{z}_{*,2}] \).

• \( \alpha_1, \alpha_2 \in \mathbb{R}^{N-1} \) are the unique solutions to the **linear systems**

\[
M_1 \alpha_1 = V_1^T z - b_1 1_{N-1} \quad \text{and} \quad M_2 \alpha_2 = V_2^T z - b_2 1_{N-1}
\]

and \( \alpha_1 = [\alpha_{2,1}; \ldots; \alpha_{N,1}], \alpha_2 = [\alpha_{2,2}; \ldots; \alpha_{N,2}], 1_{N-1} = [1; 1; \ldots; 1] \).
Model selection: \[
\min \sum_{i,j} \left( \frac{\hat{z}_i^T \hat{z}_j}{\|\hat{z}_i\|_2 \|\hat{z}_j\|_2} - \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2} \right)^2
\]
Alon colon cancer microarray data set: 3D projections

Dimension input space: 62
Number of genes: 1500 (training: 500, validation: 500, test: 500)

Model selection: $\sigma^2 = 10^4$, $\sigma_1^2 = 10^3$, $\sigma_2^2 = 0.5\sigma_1^2$, $\sigma_3^2 = 0.1\sigma_1^2$, $\eta = 1$, $\nu = 100$, $D = \text{diag}\{10, 5, 1\}$, $q = [+1; -1; -1]$. 
Solving in primal or dual?

Example of microarray data (10,000 genes & 50 training data)

Classifier model:

\[ \text{sign}(w^T x + b) \quad \text{(primal)} \]
\[ \text{sign}(\sum_i \alpha_i y_i x_i^T x + b) \quad \text{(dual)} \]

primal: \( w \in \mathbb{R}^{10,000} \)
dual: \( \alpha \in \mathbb{R}^{50} \) (kernel matrix: 50 × 50)

Example of datamining problem (20 inputs & 1,000,000 training data)

primal: \( w \in \mathbb{R}^{20} \)
dual: \( \alpha \in \mathbb{R}^{1,000,000} \) (kernel matrix: 1,000,000 × 1,000,000)
Fixed-size kernel methods: estimation in primal space

Primal space

Regression

Dual space

Nyström method
Kernel PCA
Density estimate
Entropy criteria

Eigenfunctions
SV selection

Link Nyström approximation (GP) - kernel PCA - density estimation
[Girolami, 2002; Williams & Seeger, 2001]

Modelling in view of primal-dual representations
[Suykens et al., 2002]: primal space estimation, sparse, large scale
Sparse representations with estimation in primal space [Suykens et al., 2002]
Application: electric load forecasting

Short-term load forecasting (1-24 hours)
Important for power generation decisions
Hourly load values from substations in Belgian grid
Seasonal/weekly/intra-daily patterns

1-hour ahead

24-hours ahead

Fixed-size LS-SVM → Linear ARX model

[Espinoza et al., IEEE CSM 2007]
Weighted versions and robustness

Convex cost function

Convex optimiz.

SVM solution

Weighted version with modified cost function

robust statistics

LS-SVM solution

Weighted LS-SVM

- Weighted LS-SVM: \( \min_{w, b, e} \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} v_i e_i^2 \) s.t. \( y_i = w^T \varphi(x_i) + b + e_i, \forall i \)

  with \( v_i \) determined from \( \{e_i\}_{i=1}^{N} \) of unweighted LS-SVM [Suykens et al., 2002].

  Robustness and stability of reweighted kernel based regression [Debruyne et al., 2006].

- SVM solution by applying iteratively weighted LS [Perez-Cruz et al., 2005]
Generalizations to Kernel PCA: other loss functions

• Consider general loss function $L$ ($L_2$ case = KPCA):

$$\min_{w,b,e} \left[ -\frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} L(e_i) \right] \text{ s.t. } e_i = w^T \varphi(x_i) + b, \ i = 1, \ldots, N.$$ 

Generalizations of KPCA that lead to robustness and sparseness, e.g. Vapnik $\epsilon$-insensitive loss, Huber loss function [Alzate & Suykens, 2006].

• Weighted least squares versions and incorporation of constraints:

$$\min_{w,b,e} \left[ -\frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} v_i e_i^2 \right] \text{ s.t. } \begin{cases} 
 e_i = w^T \varphi(x_i) + b, \ i = 1, \ldots, N \\
 \sum_{i=1}^{N} e_i e_i^{(1)} = 0 \\
 \vdots \\
 \sum_{i=1}^{N} e_i e_i^{(i-1)} = 0 
\end{cases}$$

Find $i$-th PC w.r.t. $i-1$ orthogonality constraints (previous PC $e_i^{(j)}$). The solution is given by a generalized eigenvalue problem.
Kernel Component Analysis: robust denoising

Images corrupted with Gaussian noise and outliers

Classical Kernel PCA
Robust method

Bottom rows: application of different pre-image algorithms

Robust method: improved results and fewer components needed
Kernel Component Analysis: sparseness

Sparseness from choice of $\epsilon$-insensitive loss [Alzate & Suykens, 2006]
(top figure: denoising; bottom figures: different support vectors (in black) per principal component vector)
Learning Theory

• Typical setting: fixed but unknown distribution $P$, i.i.d. data

• Generalization error:

$$\min_{f} \int L(Y - f(X))dP(X,Y)$$

Statistical learning theory [Vapnik, 1998]
Approximation and sample error, bias-variance trade-off [Cucker & Smale, 2002; Cucker & Zhou, 2007]
Stability and generalization [Bousquet & Elisseeff, 2001]
Rademacher complexity, ...

• Leave-one-out error plays a vital role:
  Uniform stability [Bousquet & Elisseeff, 2001]
  Cross-validation stability [Kutin & Niyogi, 2002]
  $CV_{1oo}$ stability [Poggio et al., Nature, 2004]
Robustness and Learning Theory (1)

• What happens if $P$ is contaminated? [Debruyne et al., 2008]

• Consider distributions

$$P_{\epsilon,z} = (1 - \epsilon)P + \epsilon \Delta_z$$

($\Delta_z$ Dirac distribution at $z \in \mathcal{X} \times \mathcal{Y}$) [Hampel et al., 1986; Huber, 1981]

• For a functional $T$ it is desirable that

$$T(P_{\epsilon,z}) \text{ is not too far away from } T(P) \ (\forall z \text{ and } \forall \epsilon)$$

• **Robustness measure** for $T$: influence function

$$IF(z; T, P) = \lim_{\epsilon \to 0} \frac{T(P_{\epsilon,z}) - T(P)}{\epsilon}$$
• Higher order influence functions: \( IF_k(z; T, P) = \frac{\partial}{\partial \epsilon} T(P_{\epsilon, z}) |_{\epsilon=0} \)

• **Taylor expansion:**

\[
T(P_{\epsilon, z}) = T(P) + \epsilon IF(z; T, P) + \frac{\epsilon^2}{2!} IF_2(z; T, P) + \ldots
\]

• Approximation to the leave-one-out error:

\[
T(P_{n-i}) = T(P_n) + \sum_{j=1}^{\infty} \left(\frac{-1}{n-1}\right)^j IF_j(z_i; T, P_n) \frac{1}{j!}
\]

by taking \( P_{\epsilon, z} = P_{n-i} \), \( \epsilon = -1/(n - 1) \), \( P = P_n \) where \( P_{n-i} \) empirical distribution of sample without \( z_i = (x_i, y_i) \)

(other approximation: generalized cross-validation [Wahba, 1990])
Application to function estimation in RKHS

- Define

\[ f_{\lambda,K,P} := \arg\min_{f \in \mathcal{H}} \mathbb{E}_P L(Y - f(X)) + \lambda \| f \|^2_{\mathcal{H}} \]

\[ f_{\lambda,K,P_n} := \arg\min_{f \in \mathcal{H}} \frac{1}{n} L(y_i - f(x_i)) + \lambda \| f \|^2_{\mathcal{H}} \]

Matrix \([IFM_k]\) with \(ij\)-th entry \(IF_k(z_j; f_{\lambda,K,P_n})(x_i)\).

- Robust model selection criterion [Debruyne et al., 2008]
  Case of order \(k\), least squares:

\[
C_{IF} = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - f_{\lambda,K,P_n}(x_i) - \sum_{j=1}^{k-1} \frac{1}{(1-n)^j j!} [IFM_j]_{i,i} - \frac{1}{(1-n)^k k!} \frac{[IFM_k]_{i,i}}{1-[H]_{i,i}} \right)^2
\]

where \(H\) denotes the smoother matrix.
Conclusions

• Towards integrative understanding and systematic constructive design for supervised, unsupervised learning and beyond

• Kernel methods: complementary views (LS-)SVM, RKHS, GP

• Least squares support vector machines as “core models”: provides methodology for “optimization modelling”

• Bridging gaps between fundamental theory, algorithms and applications

• Reliable methods: numerically, computationally, statistically

Supplementary material on display for MFO:
Acknowledgements

• Current and former colleagues:

• Many people for joint work, discussions, invitations, joint organization of meetings.

• Support from K.U. Leuven, GOA-Ambiorics, COE Optimization in Engineering, IAP V, FWO projects, IWT.
Books