Learning with primal and dual model representations

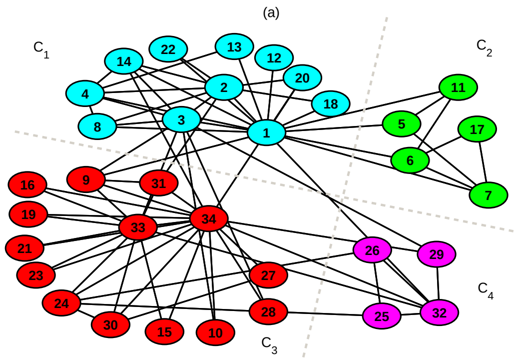
Johan Suykens

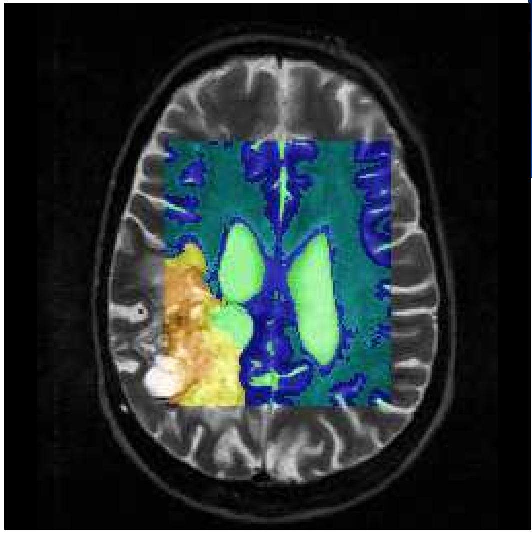
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.be
http://www.esat.kuleuven.be/stadius/

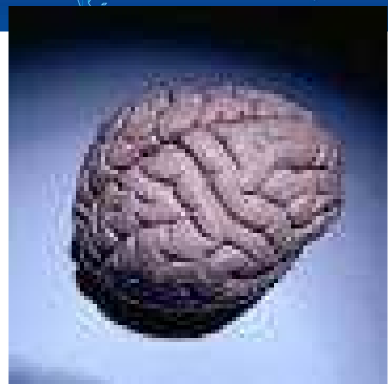
CIMI 2015, Toulouse

Introduction and motivation

Data world


(a)
C1
C2
C3
C4






CampusInsight
MIND
SPELLER
AWARD-WINNING DEVICE CONVERTS BRAIN SIGNALS INTO WORDS
KATHOLIEKE UNIVERSITEIT
LEUVEN

Challenges

- data-driven
- general methodology
- scalability
- need for new mathematical frameworks

Different paradigms

SVM &

Kernel methods

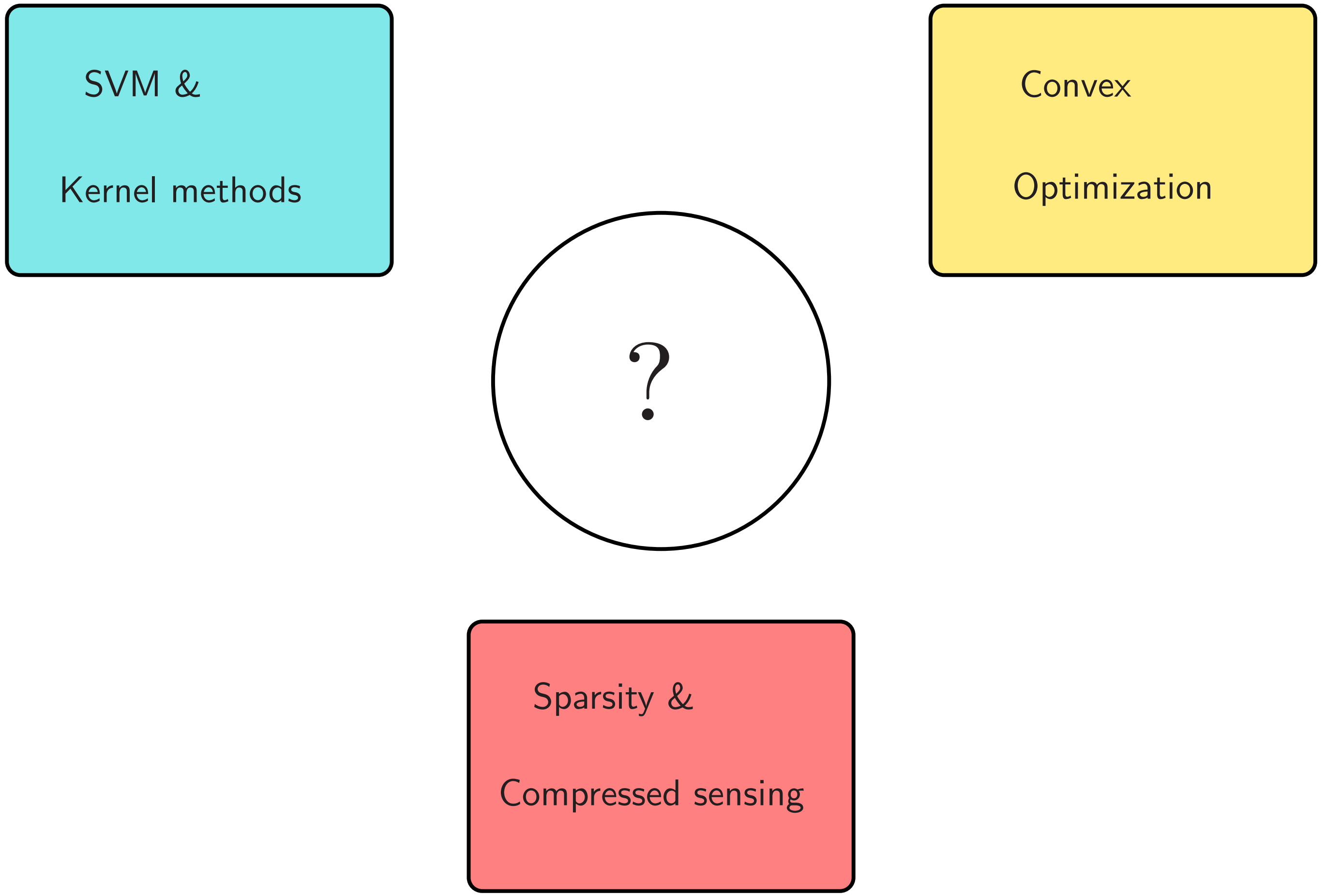
Convex

Optimization

Sparsity &

Compressed sensing

Different paradigms



Sparsity through regularization or loss function

Sparsity: through regularization or loss function

- through regularization: model $\hat{y} = w^T x + b$

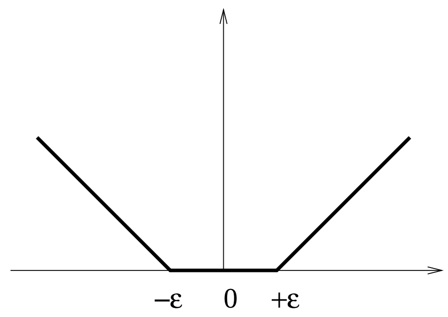
$$\min \sum_j |w_j| + \gamma \sum_i e_i^2$$

$\Rightarrow$ sparse w

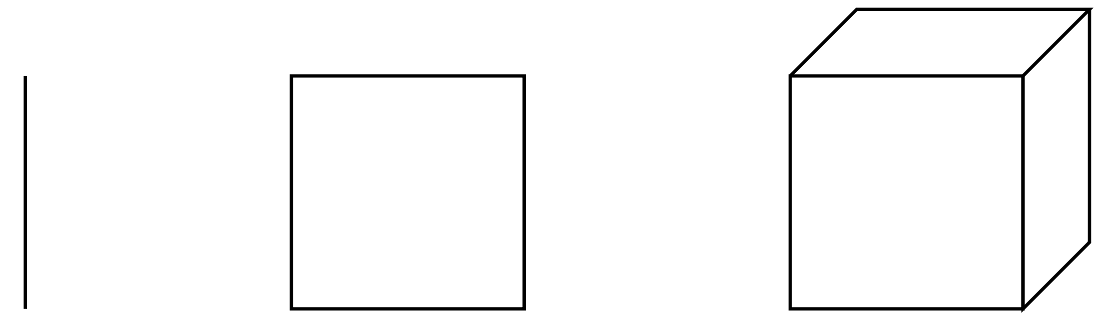
- through loss function: model $\hat{y} = \sum_i \alpha_i K(x, x_i) + b$

$$\min \ w^T w + \gamma \sum_i L(e_i)$$

$\Rightarrow$ sparse α



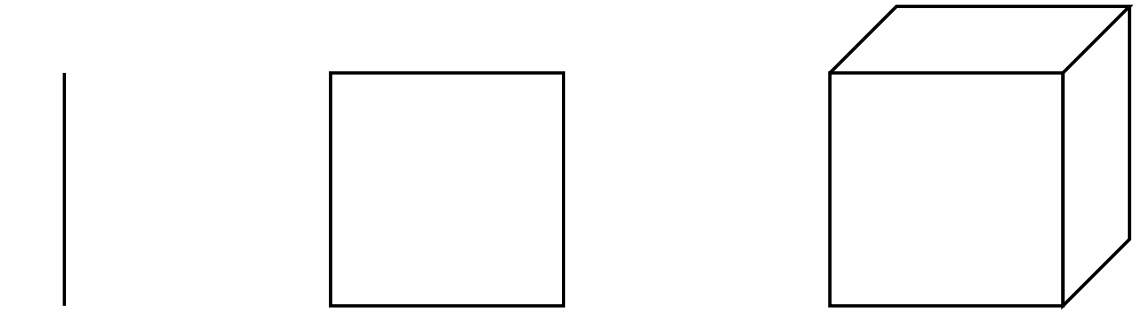
Sparsity: matrices and tensors



vector x matrix X tensor $\mathcal{X}$

data vector x		data matrix X		data tensor $\mathcal{X}$
vector model:	$\longrightarrow$	matrix model:	$\longrightarrow$	tensor model:
$\hat{y} = w^T x$		$\hat{y} = \langle W, X \rangle$		$\hat{y} = \langle \mathcal{W}, \mathcal{X} \rangle$

Sparsity: matrices and tensors



vector x matrix X tensor $\mathcal{X}$

data vector x		data matrix X		data tensor $\mathcal{X}$
vector model:	$\longrightarrow$	matrix model:	$\longrightarrow$	tensor model:
$\hat{y} = w^T x$		$\hat{y} = \langle W, X \rangle$		$\hat{y} = \langle \mathcal{W}, \mathcal{X} \rangle$
sparsity:		sparsity:		sparsity:
$\sum_j \vert w_j \vert$		$\Vert W \Vert_*$		$\Vert \mathcal{W} \Vert_*$

Learning with tensors [Signoretto, Tran Dinh, De Lathauwer, Suykens, ML 2014]
Robust tensor completion [Yang, Feng, Suykens, 2014]

Function estimation in RKHS

- Find function f such that [Wahba, 1990; Evgeniou et al., 2000]

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

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

- Sparse representation by ϵ-insensitive loss [Vapnik, 1998]

Learning with primal and dual model representations

Learning models from data: alternative views

- Consider model $\hat{y} = f(x; w)$, given input/output data $\{(x_i, y_i)\}_{i=1}^N$:

$$\min_{w} w^T w + \gamma \sum_{i=1}^{N} (y_i - f(x_i; w))^2$$

Learning models from data: alternative views

- Consider model $\hat{y} = f(x; w)$, given input/output data $\{(x_i, y_i)\}_{i=1}^N$:

$$\min_w \; w^T w + \gamma \sum_{i=1}^{N} (y_i - f(x_i; w))^2$$

- Rewrite the problem as

$$\begin{array}{ll} \displaystyle\min_{w,e} & w^T w + \gamma \sum_{i=1}^N e_i^2 \\ \text{subject to} & e_i = y_i - f(x_i; w), i = 1, ..., N \end{array}$$

- Express the solution and the model in terms of Lagrange multipliers α_i

- For a model $f(x; w) = \sum_{j=1}^h w_j \varphi_j(x) = w^T \varphi(x)$ one obtains then $\hat{f}(x) = \sum_{i=1}^N \alpha_i K(x, x_i)$ with $K(x, x_i) = \varphi(x)^T \varphi(x_i)$.

Least Squares Support Vector Machines: “core models”

- 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

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

- Kernel pca $(V = I)$, Kernel spectral clustering $(V = D^{-1})$

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

- Kernel canonical correlation analysis/partial least squares

$$\min_{w,v,b,d,e,r} w^T w + v^T v + \nu \sum_i (e_i - r_i)^2 \ \text{s.t.} \left\{ \begin{array}{lcl} e_i & = & w^T \varphi^{(1)}(x_i) + b \\ r_i & = & v^T \varphi^{(2)}(y_i) + d \end{array} \right.$$

[Suykens & Vandewalle, 1999; Suykens et al., 2002; Alzate & Suykens, 2010]

Probability and quantum mechanics

- **Kernel pmf estimation**
  – *Primal:*

$$\min_{w,p_i} \tfrac{1}{2}\langle w, w\rangle \text{ subject to } \quad p_i = \langle w, \varphi(x_i)\rangle,\ i = 1, ..., N \text{ and } \textstyle\sum_{i=1}^N p_i = 1$$

  – *Dual:* $p_i = \frac{\sum_{j=1}^N K(x_j,x_i)}{\sum_{i=1}^N \sum_{j=1}^N K(x_j,x_i)}$

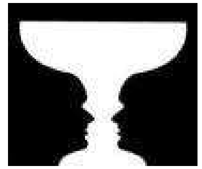
- **Quantum measurement**: state vector $|\psi\rangle$, measurement operators M_i
  – *Primal:*

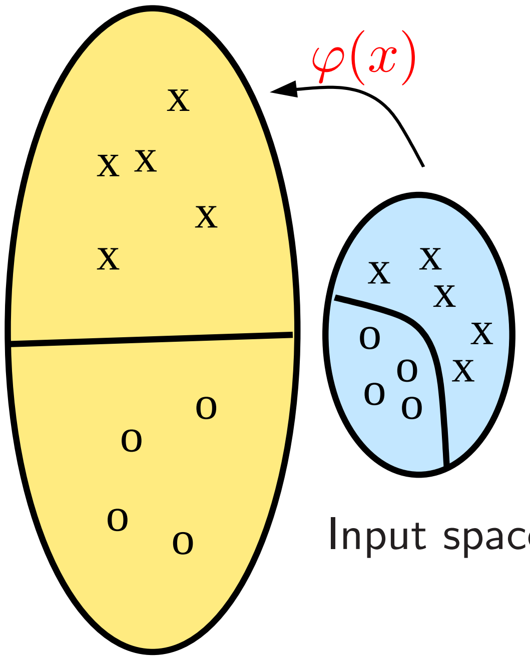
$$\min_{|w\rangle,p_i} \tfrac{1}{2}\langle w|w\rangle \text{ subject to } \quad p_i = \mathrm{Re}(\langle w|M_i\psi\rangle),\ i = 1, ..., N \text{ and } \textstyle\sum_{i=1}^N p_i = 1$$

  – *Dual:* $p_i = \langle \psi|M_i|\psi\rangle$ (Born rule, orthogonal projective measurement)

[Suykens, Physical Review A, 2013]

SVMs: living in two worlds ...

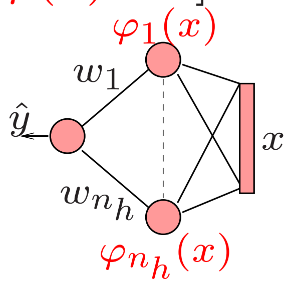




Primal space

Parametric

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

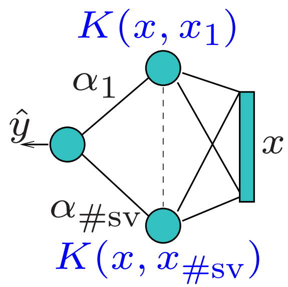


$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \quad \textbf{(Mercer)}$$

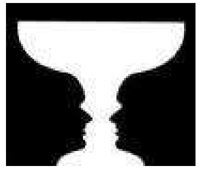
Dual space

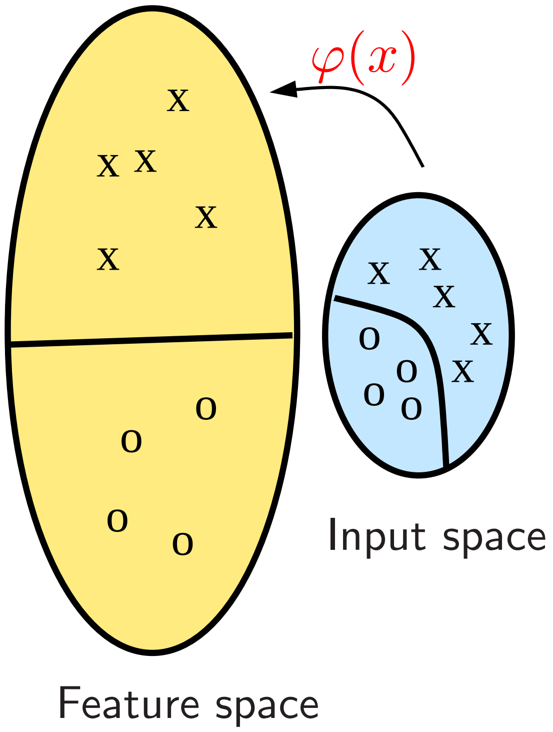
Nonparametric

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



SVMs: living in two worlds ...

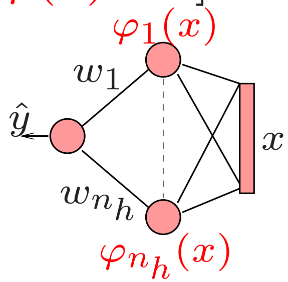




Primal space

Parametric

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



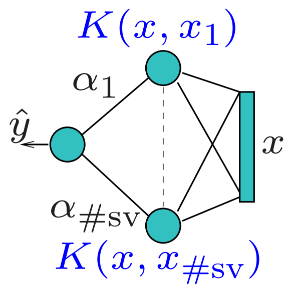
Parametric

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$$ **("Kernel trick")**

Dual space

Nonparametric

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



Non-parametric

Linear model: solving in primal or dual?

inputs $x \in \mathbb{R}^d$, output $y \in \mathbb{R}$
training set $\{(x_i, y_i)\}_{i=1}^N$

$$(P): \quad \hat{y} = w^T x + b, \qquad w \in \mathbb{R}^d$$

Model $\nearrow$

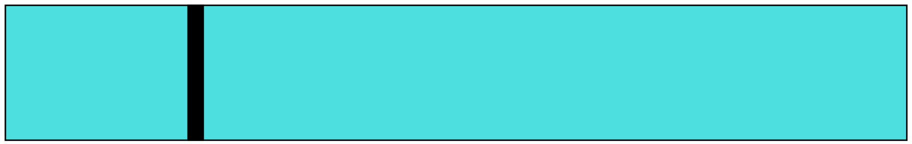
Linear model: solving in primal or dual?

inputs $x \in \mathbb{R}^d$, output $y \in \mathbb{R}$
training set $\{(x_i, y_i)\}_{i=1}^N$

$$\text{Model} \begin{array}{l} \nearrow \quad (P): \quad \hat{y} = w^T x + b, \qquad w \in \mathbb{R}^d \\ \\ \searrow \quad (D): \quad \hat{y} = \sum_i \alpha_i \, x_i^T x + b, \qquad \alpha \in \mathbb{R}^N \end{array}$$

Linear model: solving in primal or dual?

few inputs, many data points: $d \ll N$

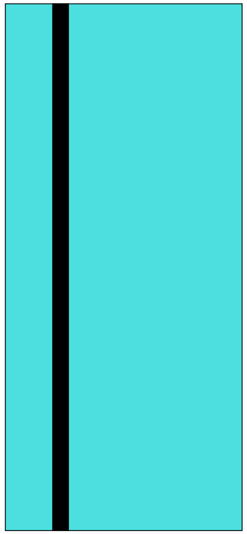


$\boxed{\textbf{primal}}$: $w \in \mathbb{R}^d$

dual: $\alpha \in \mathbb{R}^N$ (large kernel matrix: $N \times N$)

Linear model: solving in primal or dual?

many inputs, few data points: $d \gg N$



primal: $w \in \mathbb{R}^d$

$\boxed{\textbf{dual}}$: $\alpha \in \mathbb{R}^N$ (small kernel matrix: $N \times N$)

Feature map and kernel

From linear to nonlinear model:

$$\text{Model} \begin{array}{l} \nearrow \quad (P): \quad \hat{y} = w^T \varphi(x) + b \\ \\ \searrow \quad (D): \quad \hat{y} = \sum_i \alpha_i K(x_i, x) + b \end{array}$$

Mercer theorem:

$$K(x, z) = \varphi(x)^T \varphi(z)$$

Feature map $\varphi(x) = [\varphi_1(x); \varphi_2(x); ...; \varphi_h(x)]$
Kernel function $K(x, z)$ (e.g. linear, polynomial, RBF, ...)

- Use of feature map and positive definite kernel [Cortes & Vapnik, 1995]
- Extension to infinite dimensional case:
  - LS-SVM formulation [Signoretto, De Lathauwer, Suykens, 2011]
  - HHK Transform, coherent states, wavelets [Fanuel & Suykens, 2015]

HHK transform

- Coherent states $\{|\eta_x\rangle \in \mathcal{H}\}_{x \in X}$ in

$$\min_{|w\rangle \in \mathcal{H}, e_i, b} \frac{1}{2} \langle w|w\rangle_{\mathcal{H}} + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad y_i = \langle \eta_{x_i}|w\rangle_{\mathcal{H}} + b + e_i, \quad i = 1, ..., N$$

-

$$\begin{array}{lll} & & (P): \quad \hat{y} = \langle \eta_x|w\rangle_{\mathcal{H}} + b \quad \boxed{\rightarrow \text{transform}} \\ & \nearrow & \\ \mathcal{M} & & \qquad \downarrow K(x,z) = \langle \eta_x|\eta_z\rangle_{\mathcal{H}} \\ & \searrow & \\ & & (D): \quad \hat{y} = \sum_i \alpha_i K(x_i, x) + b \end{array}$$

[Fanuel & Suykens, TR15-101, 2015]

HHK transform

- Coherent states $\{|\eta_x\rangle \in \mathcal{H}\}_{x \in X}$ in

$$\min_{|w\rangle \in \mathcal{H}, e_i, b} \frac{1}{2} \langle w|w\rangle_{\mathcal{H}} + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad y_i = \langle \eta_{x_i}|w\rangle_{\mathcal{H}} + b + e_i, \quad i = 1, ..., N$$

- **HHK Transform:** $W_\eta : \mathcal{H} \to \mathcal{H}_K : |w\rangle \mapsto \langle \eta.|w\rangle_{\mathcal{H}}$

$$\begin{array}{llll} & (P): \quad \hat{y} = \langle \eta_x|w\rangle_{\mathcal{H}} + b & \boxed{\mathcal{H} \to \mathcal{H}_K} & \hat{y} = \langle W_\eta \eta_x|W_\eta w\rangle_K + b \\ \nearrow & & & \\ \mathcal{M} & \quad \downarrow K(x,z) = \langle \eta_x|\eta_z\rangle_{\mathcal{H}} & & \downarrow K(x,z) = \langle \xi_x|\xi_z\rangle_K \,,\; \xi_x = W_\eta \eta_x \\ \searrow & & & \\ & (D): \quad \hat{y} = \sum_i \alpha_i K(x_i, x) + b & & \hat{y} = \sum_i \alpha_i K(x_i, x) + b \end{array}$$

[Fanuel & Suykens, TR15-101, 2015]

Sparsity by fixed-size kernel method

Fixed-size method: steps

1. **selection of a subset** from the data

2. kernel matrix on the subset

3. eigenvalue decomposition of kernel matrix

4. **approximation of the feature map** based on the eigenvectors (Nyström approximation)

5. estimation of the model in the primal using the approximate feature map (applicable to large data sets)

[Suykens et al., 2002] *(ls-svm book)*

Selection of subset

- random
- quadratic Renyi entropy
- incomplete Cholesky factorization

Nyström method

- *"big"* kernel matrix: $\Omega_{(N,N)} \in \mathbb{R}^{N \times N}$
  "small" kernel matrix: $\Omega_{(M,M)} \in \mathbb{R}^{M \times M}$ (on subset)

- Eigenvalue decompositions: $\Omega_{(N,N)} \tilde{U} = \tilde{U} \, \tilde{\Lambda}$ and $\Omega_{(M,M)} \overline{U} = \overline{U} \, \overline{\Lambda}$

- Relation to eigenvalues and eigenfunctions of the integral equation

$$\int K(x, x') \phi_i(x) p(x) dx = \lambda_i \phi_i(x')$$

with

$$\hat{\lambda}_i = \frac{1}{M} \overline{\lambda}_i, \quad \hat{\phi}_i(x_k) = \sqrt{M} \, \overline{u}_{ki}, \quad \hat{\phi}_i(x') = \frac{\sqrt{M}}{\overline{\lambda}_i} \sum_{k=1}^{M} \overline{u}_{ki} K(x_k, x')$$

[Williams & Seeger, 2001] *(Nyström method in GP)*

Fixed-size method: estimation in primal

- For the feature map $\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^h$ obtain an approximation

$$\tilde{\varphi}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^M$$

  based on the eigenvalue decomposition of the kernel matrix with $\tilde{\varphi}_i(x') = \sqrt{\hat{\lambda}_i}\, \hat{\phi}_i(x')$ (on a **subset** of size $M \ll N$).

- Estimate in **primal**:

$$\min_{\tilde{w},\tilde{b}} \frac{1}{2}\tilde{w}^T\tilde{w} + \gamma \frac{1}{2}\sum_{i=1}^{N}(y_i - \tilde{w}^T\tilde{\varphi}(x_i) - \tilde{b})^2$$

  Sparse representation is obtained: $\tilde{w} \in \mathbb{R}^M$ with $M \ll N$ and $M \ll h$.

  [Suykens et al., 2002; De Brabanter et al., CSDA 2010]

Fixed-size method: performance in classification

	pid	spa	mgt	adu	ftc
N	768	4601	19020	45222	581012
N_{cv}	512	3068	13000	33000	531012
N_{test}	256	1533	6020	12222	50000
d	8	57	11	14	54
FS-LSSVM (# SV)	150	200	1000	500	500
C-SVM (# SV)	290	800	7000	11085	185000
ν-SVM (# SV)	331	1525	7252	12205	165205
RBF FS-LSSVM	76.7(3.43)	92.5(0.67)	86.6(0.51)	85.21(0.21)	81.8(0.52)
Lin FS-LSSVM	77.6(0.78)	90.9(0.75)	77.8(0.23)	83.9(0.17)	75.61(0.35)
RBF C-SVM	75.1(3.31)	92.6(0.76)	85.6(1.46)	84.81(0.20)	81.5(no cv)
Lin C-SVM	76.1(1.76)	91.9(0.82)	77.3(0.53)	83.5(0.28)	75.24(no cv)
RBF ν-SVM	75.8(3.34)	88.7(0.73)	84.2(1.42)	83.9(0.23)	81.6(no cv)
Maj. Rule	64.8(1.46)	60.6(0.58)	65.8(0.28)	83.4(0.1)	51.23(0.20)

- Fixed-size (FS) LSSVM: good performance and sparsity wrt C-SVM and ν-SVM
- Challenging to achieve high performance by very sparse models

[De Brabanter et al., CSDA 2010]

Two stages of sparsity

primal	
dual	subset selection Nyström approximation

Two stages of sparsity

	stage 1
primal	FS model estimation
	$\uparrow$
dual	subset selection Nyström approximation

Two stages of sparsity

	stage 1		stage 2
primal	FS model estimation	$\longrightarrow$	reweighted ℓ_1
	$\uparrow$		
dual	subset selection Nyström approximation		

Synergy between parametric & kernel-based models

[Mall & Suykens, IEEE-TNNLS 2015], reweighted ℓ_1 [Candes et al., 2008]

Two stages of sparsity

	stage 1		stage 2
primal	FS model estimation	$\longrightarrow$	reweighted ℓ_1
	$\uparrow$		
dual	subset selection Nyström approximation		

Synergy between parametric & kernel-based models

[Mall & Suykens, IEEE-TNNLS 2015], reweighted ℓ_1 [Candes et al., 2008]

Other possible approaches with improved sparsity: SCAD [Fan & Li, 2001]; coefficient-based ℓ_q $(0 < q \leq 1)$ [Shi et al., 2013]; two-level ℓ_1 [Huang et al., 2014]

Kernel-based models for spectral clustering

Kernel PCA

- Primal problem: [Suykens et al., 2002]

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

- Dual problem corresponds to kernel PCA [Scholkopf et al., 1998]

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

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

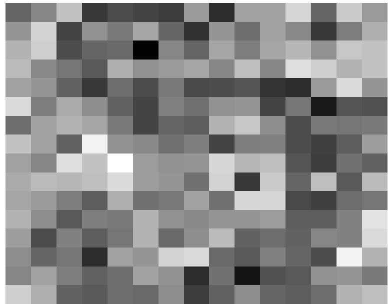
- Interpretation:
  1. pool of candidate components (objective function equals zero)
  2. select relevant components

- Robust and sparse versions [Alzate & Suykens, 2008]: by taking other loss functions

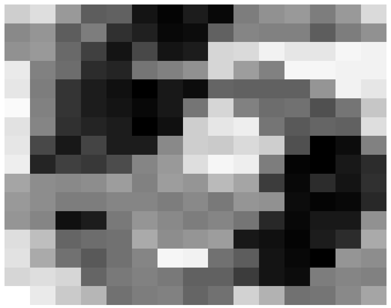
Robustness: Kernel Component Analysis

original image

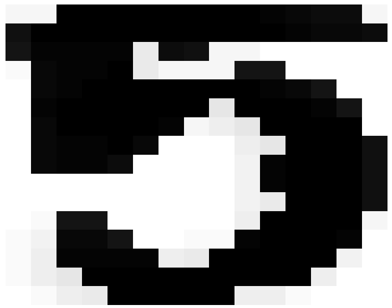
corrupted image



KPCA reconstruction



KCA reconstruction



Weighted LS-SVM [Alzate & Suykens, IEEE-TNN 2008]: robustness and sparsity

Kernel Spectral Clustering (KSC): case of two clusters

- **Primal problem:** training on given data $\{x_i\}_{i=1}^N$

$$\begin{array}{ll} \min\limits_{w,b,e} & \frac{1}{2}w^T w - \gamma \frac{1}{2} e^T V e \\ \text{subject to} & e_i = w^T \varphi(x_i) + b, \quad i = 1, ..., N \end{array}$$

with weighting matrix V and $\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^h$ the feature map.

- **Dual:**

$$V M_V \Omega \alpha = \lambda \alpha$$

with $\lambda = 1/\gamma$, $M_V = I_N - \frac{1}{1_N^T V 1_N} 1_N 1_N^T V$ weighted centering matrix, $\Omega = [\Omega_{ij}]$ kernel matrix with $\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$

- Taking $V = D^{-1}$ with degree matrix $D = \text{diag}\{d_1, ..., d_N\}$ and $d_i = \sum_{j=1}^N \Omega_{ij}$ relates to random walks algorithm.

[Alzate & Suykens, IEEE-PAMI, 2010]

Lagrangian and conditions for optimality

- Lagrangian:

$$\mathcal{L}(w,b,e;\alpha) = \frac{1}{2}w^T w - \gamma \frac{1}{2}\sum_{i=1}^{N} v_i e_i^2 + \sum_{i=1}^{N} \alpha_i (e_i - w^T \varphi(x_i) - b)$$

- Conditions for optimality:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 & \Rightarrow \quad w = \sum_i \alpha_i \varphi(x_i) \\ \dfrac{\partial \mathcal{L}}{\partial b} = 0 & \Rightarrow \quad \sum_i \alpha_i = 0 \\ \dfrac{\partial \mathcal{L}}{\partial e_i} = 0 & \Rightarrow \quad \alpha_i = \gamma v_i e_i, \;\; i = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_i} = 0 & \Rightarrow \quad e_i = w^T \varphi(x_i) + b, \;\; i = 1, ..., N \end{cases}$$

- Eliminate w, b, e, write solution in Lagrange multipliers α_i.

Kernel spectral clustering: more clusters

- Case of k clusters: additional sets of constraints

$$
\begin{array}{ll}
\min\limits_{w^{(l)},e^{(l)},b_l} & \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} - \frac{1}{2}\sum_{l=1}^{k-1} \gamma_l e^{(l)^T} D^{-1} e^{(l)} \\
\text{subject to} & e^{(1)} = \Phi_{N\times n_h} w^{(1)} + b_1 1_N \\
& e^{(2)} = \Phi_{N\times n_h} w^{(2)} + b_2 1_N \\
& \vdots \\
& e^{(k-1)} = \Phi_{N\times n_h} w^{(k-1)} + b_{k-1} 1_N
\end{array}
$$

where $e^{(l)} = [e_1^{(l)}; ...; e_N^{(l)}]$ and $\Phi_{N\times n_h} = [\varphi(x_1)^T; ...; \varphi(x_N)^T] \in \mathbb{R}^{N\times n_h}$.

- **Dual problem**: $M_D \Omega \alpha^{(l)} = \lambda D \alpha^{(l)}$, $l = 1, ..., k-1$.

[Alzate & Suykens, IEEE-PAMI, 2010]

Primal and dual model representations

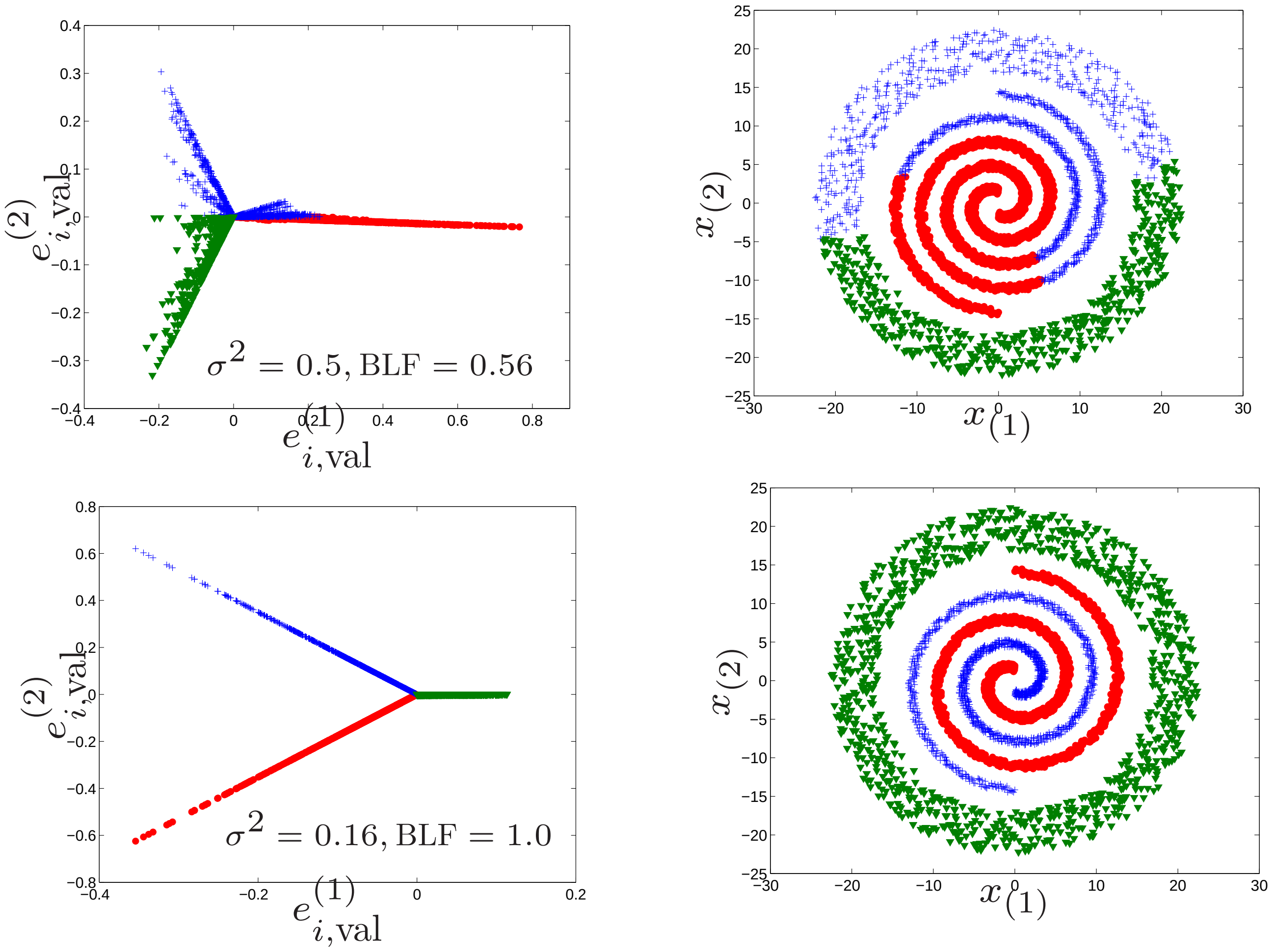
k clusters
$k-1$ sets of constraints (index $l = 1, ..., k-1$)

$$\mathcal{M} \begin{array}{l} \nearrow \quad (P): \quad \mathrm{sign}[\hat{e}_*^{(l)}] = \mathrm{sign}[{w^{(l)}}^T \varphi(x_*) + b_l] \\ \\ \searrow \quad (D): \quad \mathrm{sign}[\hat{e}_*^{(l)}] = \mathrm{sign}[\sum_j \alpha_j^{(l)} K(x_*, x_j) + b_l] \end{array}$$

Advantages of kernel-based setting

- **model-based** approach
- **out-of-sample extensions**, applying model to new data
- consider **training, validation and test data**
  (training problem corresponds to eigenvalue decomposition problem)
- model selection procedures
- **sparse representations and large scale methods**

Model selection: toy example



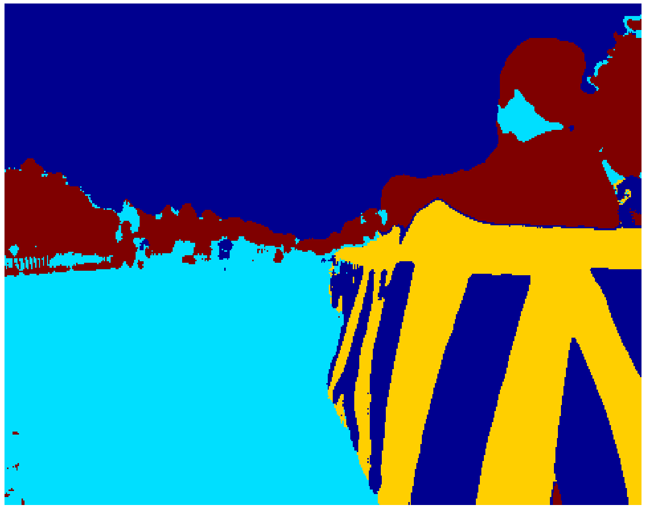
BAD

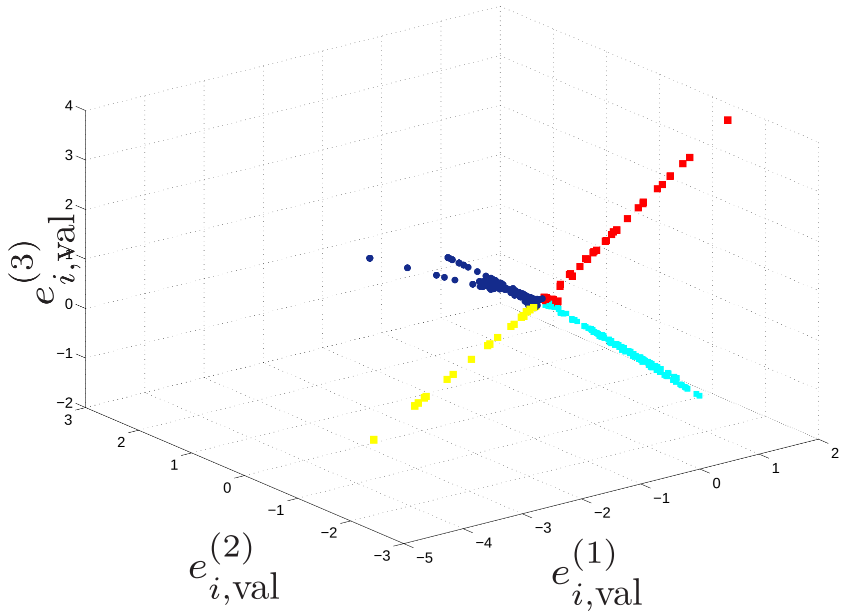
GOOD

validation set

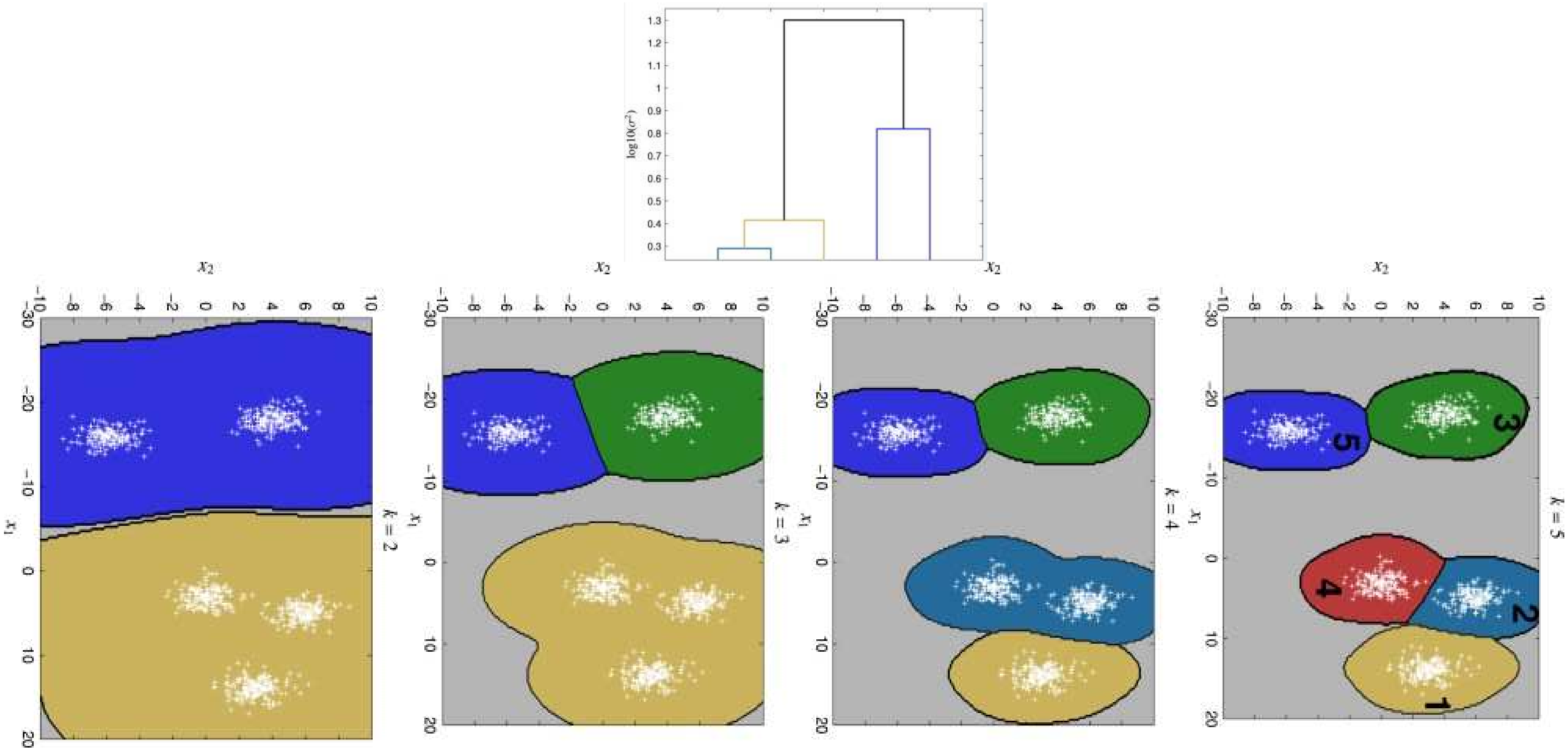
train + validation + test data

Example: image segmentation



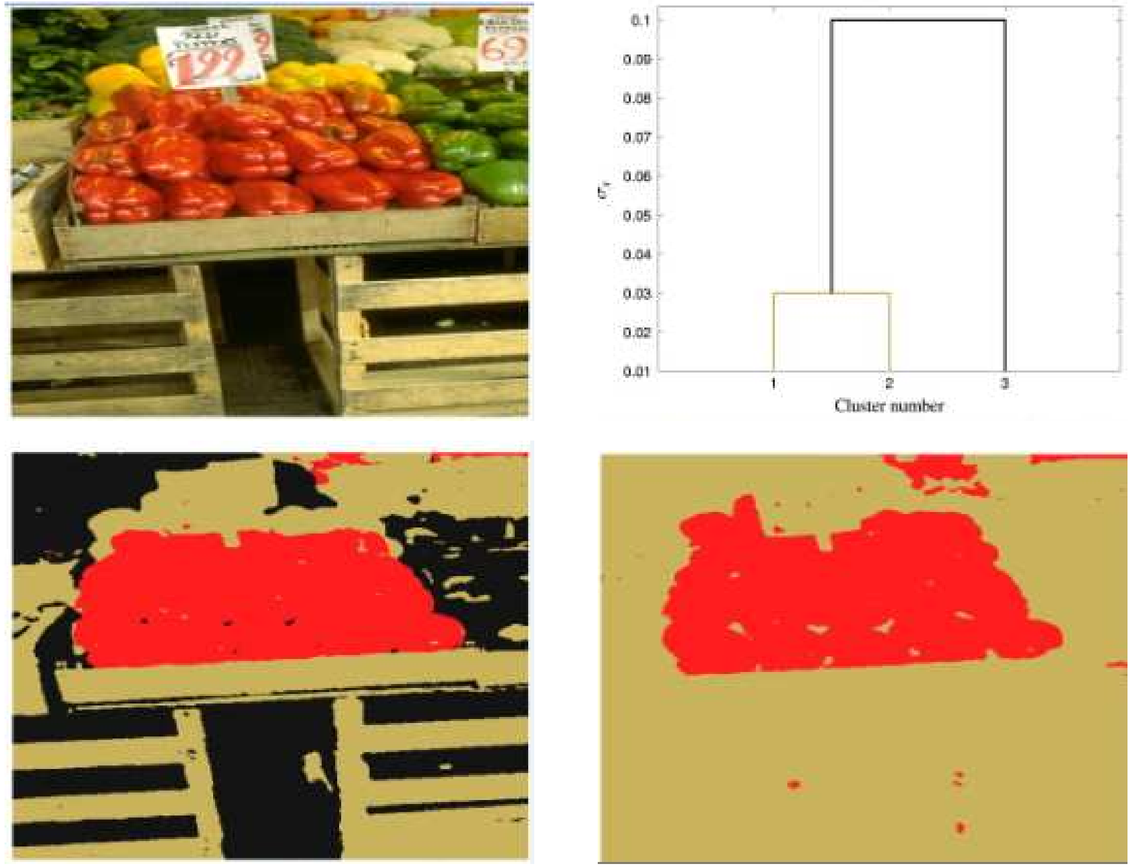


Hierarchical KSC



[Alzate & Suykens, 2012]

Hierarchical KSC



[Alzate & Suykens, 2012]

Kernel spectral clustering: sparse kernel models

original image

binary clustering

Incomplete Cholesky decomposition: $\Omega \simeq GG^T$ with $G \in \mathbb{R}^{N \times R}$ and $R \ll N$

Image (Berkeley image dataset): 321×481 $(154,401$ pixels), 175 SV

Kernel spectral clustering: sparse kernel models

original image

sparse kernel model

Incomplete Cholesky decomposition: $\Omega \simeq GG^T$ with $G \in \mathbb{R}^{N \times R}$ and $R \ll N$

Image (Berkeley image dataset): 321×481 ($154,401$ pixels), 175 SV

Time-complexity $O(R^2N^2)$ in [Alzate & Suykens, 2008]

Time-complexity $O(R^2N)$ in [Novak, Alzate, Langone, Suykens, 2014]

Incomplete Cholesky decomposition and reduced set

- For KSC problem $M_D \Omega \alpha = \lambda D \alpha$, solve the approximation

$$U^T M_D U \Lambda^2 \zeta = \lambda \zeta$$

from $\Omega \simeq G G^T$, singular value decomposition $G = U \Lambda V^T$ and $\zeta = U^T \alpha$. A smaller matrix of size $R \times R$ is obtained instead of $N \times N$.

- Pivots are used as subset $\{\tilde{x}_i\}$ for the data

- Reduced set method [Scholkopf et al., 1999]: approximation of $w = \sum_{i=1}^{N} \alpha_i \varphi(x_i)$ by $\tilde{w} = \sum_{j=1}^{M} \beta_j \varphi(\tilde{x}_j)$ in the sense

$$\min_{\beta} \|w - \tilde{w}\|_2^2$$

- Sparser solutions by adding ℓ_1 penalty, reweighted ℓ_1 or group Lasso.

[Alzate & Suykens, 2008, 2011; Mall & Suykens, 2014]

Incomplete Cholesky decomposition and reduced set

- For KSC problem $M_D \Omega \alpha = \lambda D \alpha$, solve the approximation

$$U^T M_D U \Lambda^2 \zeta = \lambda \zeta$$

  from $\Omega \simeq G G^T$, singular value decomposition $G = U \Lambda V^T$ and $\zeta = U^T \alpha$. A smaller matrix of size $R \times R$ is obtained instead of $N \times N$.

- Pivots are used as subset $\{\tilde{x}_i\}$ for the data

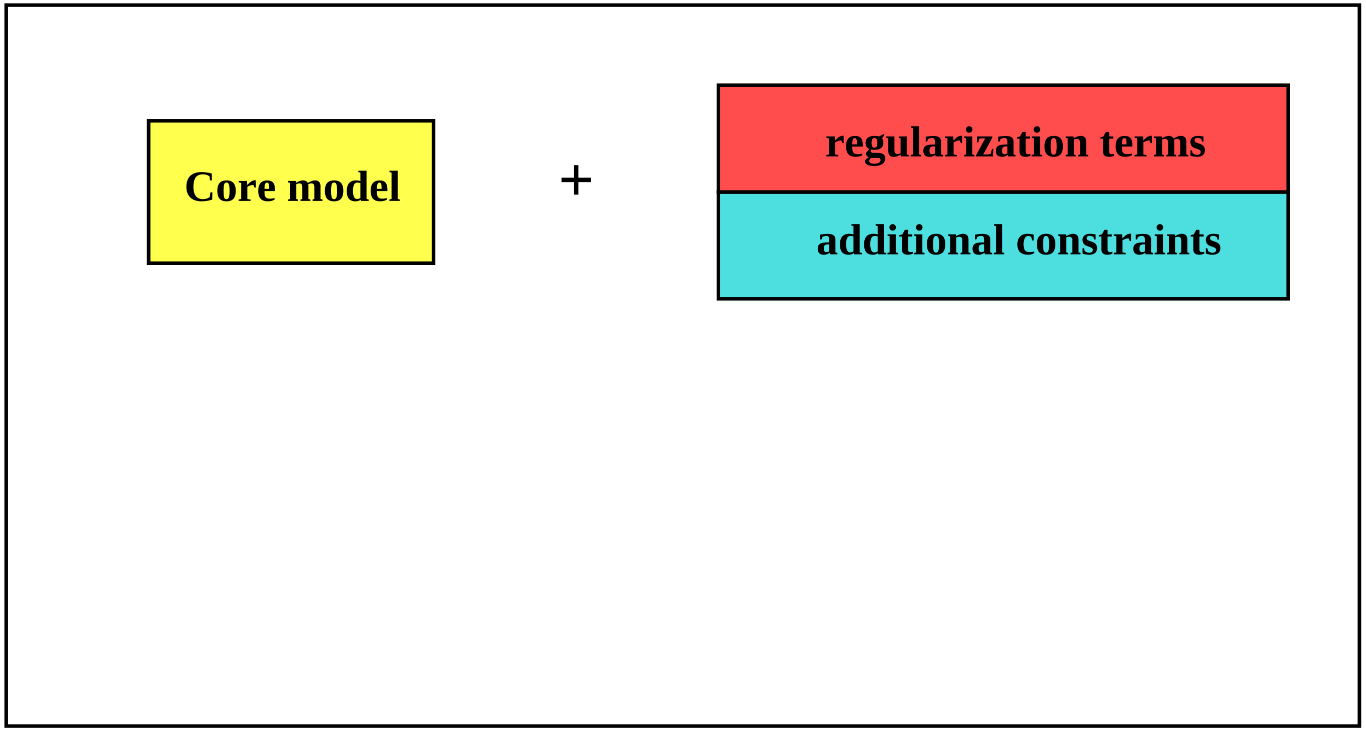
- Reduced set method [Scholkopf et al., 1999]: approximation of $w = \sum_{i=1}^{N} \alpha_i \varphi(x_i)$ by $\tilde{w} = \sum_{j=1}^{M} \beta_j \varphi(\tilde{x}_j)$ in the sense

$$\min_{\beta} \|w - \tilde{w}\|_2^2 + \nu \sum_j |\beta_j|$$

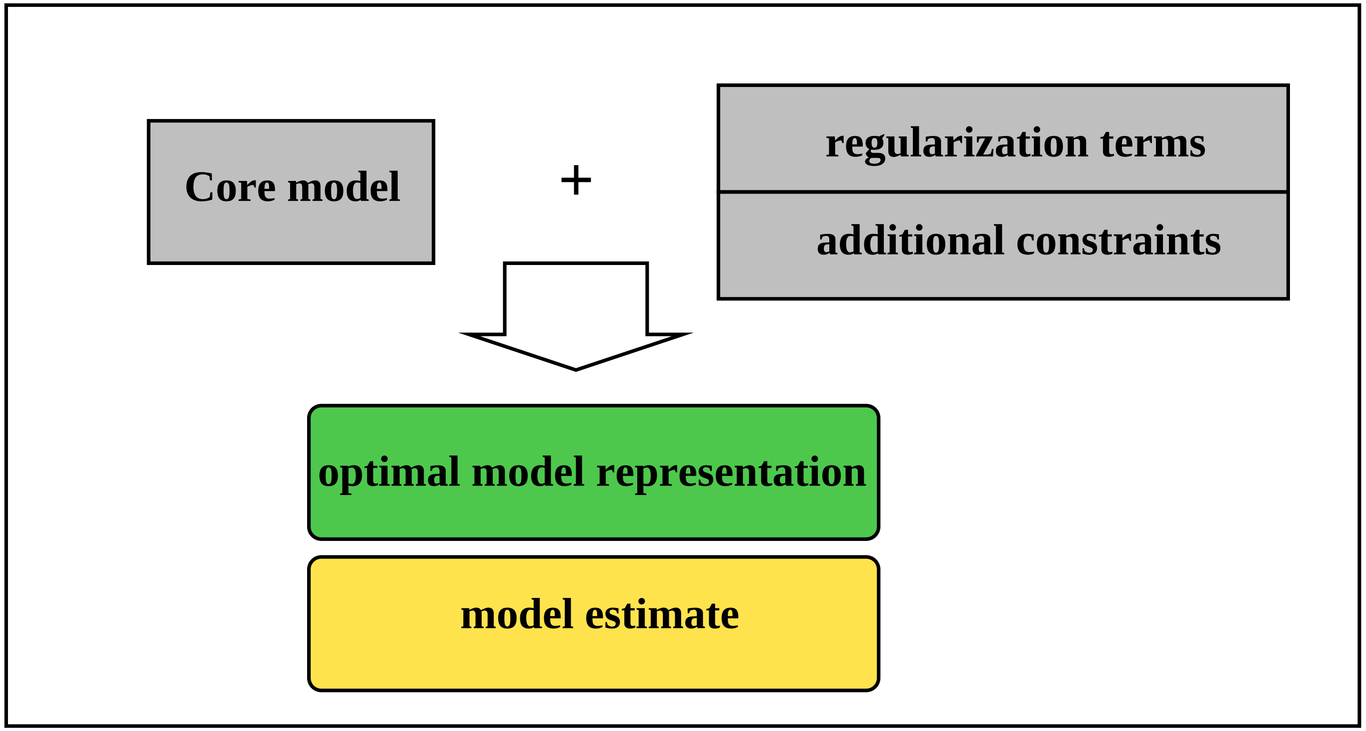
- Sparser solutions by adding ℓ_1 penalty, reweighted ℓ_1 or group Lasso.

[Alzate & Suykens, 2008, 2011; Mall & Suykens, 2014]

Core models + constraints



Core models + constraints



Kernel spectral clustering: adding prior knowledge

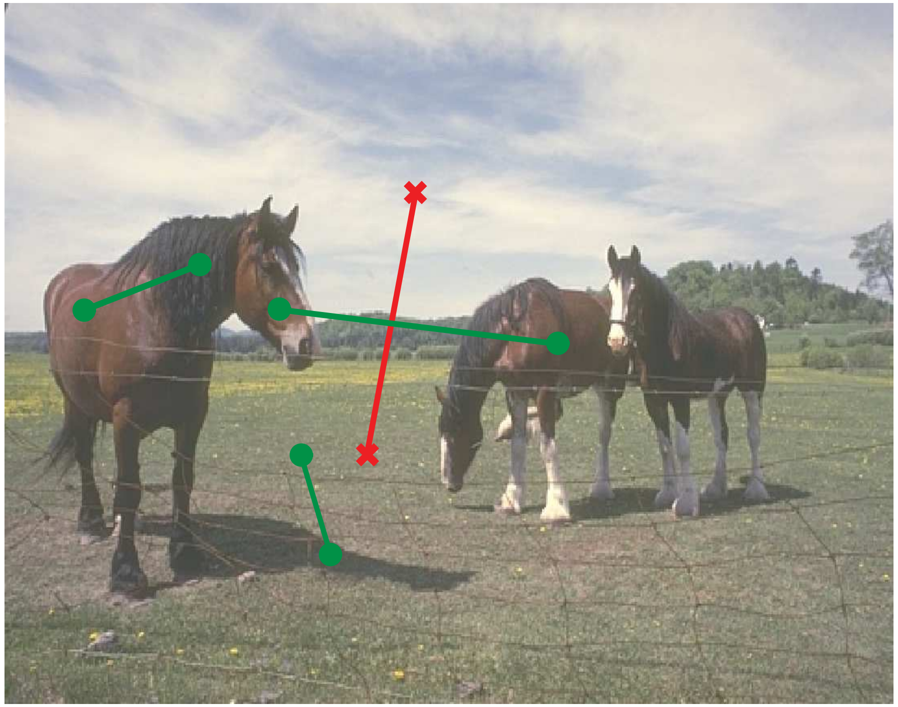
- Pair of points $x_\dagger, x_\ddagger$: $c = 1$ must-link, $c = -1$ cannot-link
- Primal problem [Alzate & Suykens, IJCNN 2009]

$$
\begin{array}{ll}
\displaystyle\min_{w^{(l)},e^{(l)},b_l} & \displaystyle -\frac{1}{2}\sum_{l=1}^{k-1} {w^{(l)}}^T w^{(l)} + \frac{1}{2}\sum_{l=1}^{k-1} \gamma_l {e^{(l)}}^T D^{-1} e^{(l)} \\
\text{subject to} & e^{(1)} = \Phi_{N\times n_h} w^{(1)} + b_1 1_N \\
 & \vdots \\
 & e^{(k-1)} = \Phi_{N\times n_h} w^{(k-1)} + b_{k-1} 1_N \\
 & {w^{(1)}}^T \varphi(x_\dagger) = c {w^{(1)}}^T \varphi(x_\ddagger) \\
 & \vdots \\
 & {w^{(k-1)}}^T \varphi(x_\dagger) = c {w^{(k-1)}}^T \varphi(x_\ddagger)
\end{array}
$$

- Dual problem: yields rank-one downdate of the kernel matrix

Adding prior knowledge

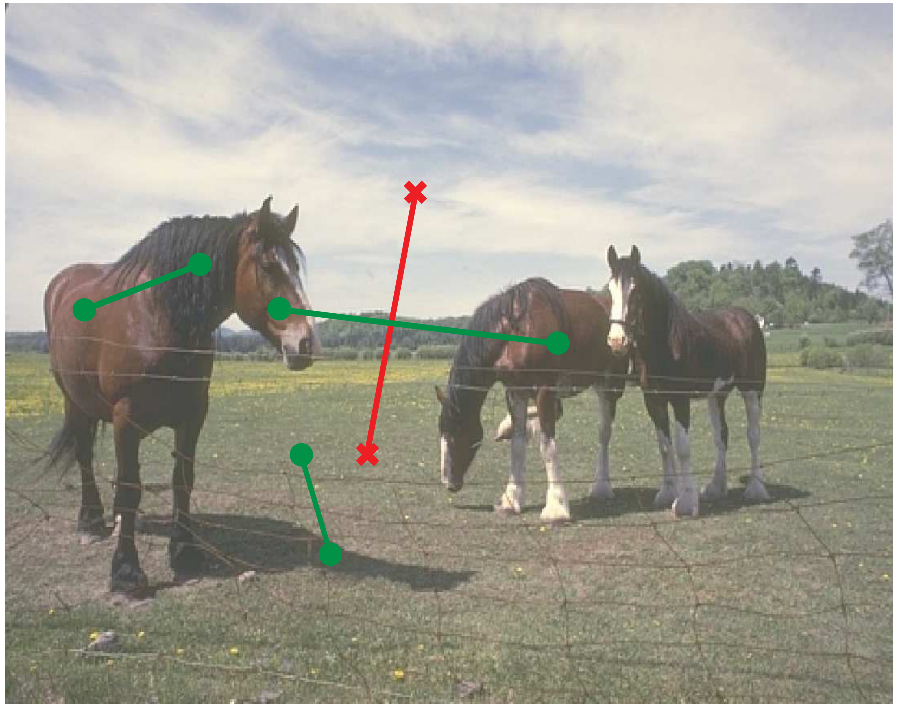
original image



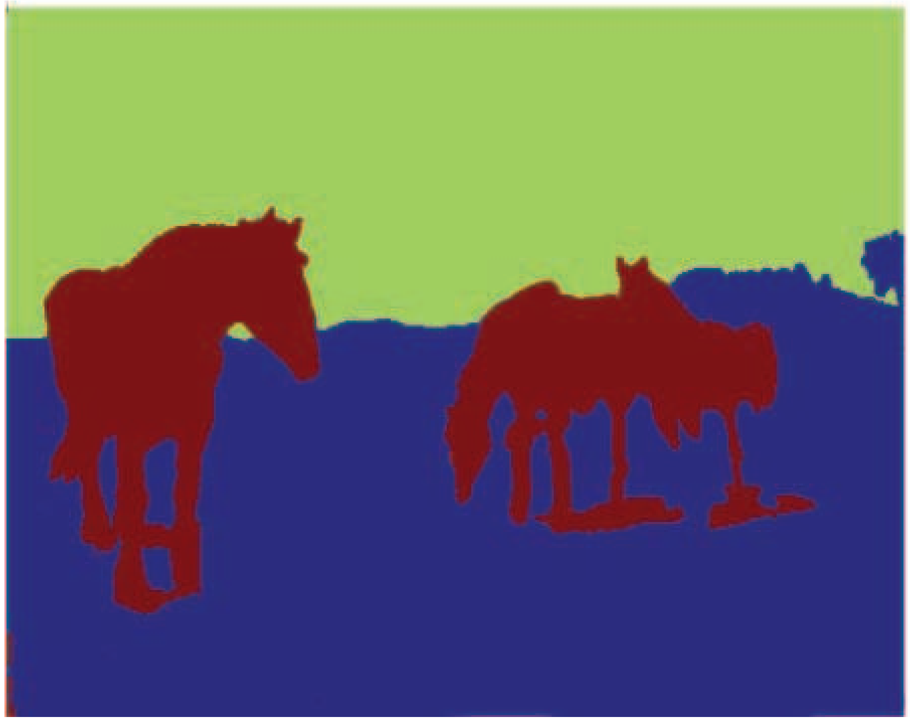
without constraints

Adding prior knowledge

original image



with constraints



Semi-supervised learning using KSC (1)

- N unlabeled data, but additional labels on $M - N$ data
  $\mathcal{X} = \{x_1, ..., x_N, x_{N+1}, ..., x_M\}$

- Kernel spectral clustering as core model (binary case [Alzate & Suykens, WCCI 2012], multi-way/multi-class [Mehrkanoon et al., TNNLS 2015])

$$\begin{array}{ll} \min\limits_{w,e,b} & \frac{1}{2}w^T w - \gamma \frac{1}{2} e^T D^{-1} e + \rho \frac{1}{2} \sum\limits_{m=N+1}^{M} (e_m - y_m)^2 \\ \text{subject to} & e_i = w^T \varphi(x_i) + b, \ i = 1, ..., M \end{array}$$

  Dual solution is characterized by a linear system. Suitable for clustering as well as classification.

- Other approaches in semi-supervised learning and manifold learning, e.g. [Belkin et al., 2006]

Semi-supervised learning using KSC (2)

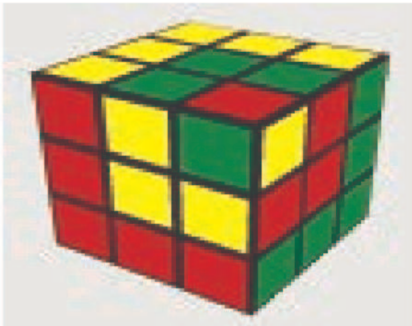
Dataset	size	n_L/n_U	test (%)	FS semi-KSC	RD semi-KSC	Lap-SVMp
Spambase	4597	368/736	919 (20%)	0.885 ± 0.01	0.883 ± 0.01	0.880 ± 0.03
Satimage	6435	1030/1030	1287 (20%)	0.864 ± 0.006	0.831 ± 0.009	0.834 ± 0.007
Ring	7400	592/592	1480 (20%)	0.975 ± 0.005	0.974 ± 0.005	0.972 ± 0.006
Magic	19020	761/1522	3804 (20%)	0.836 ± 0.006	0.829 ± 0.006	0.827 ± 0.005
Cod-rna	331152	1325/1325	66230 (20%)	0.957 ± 0.006	0.947 ± 0.008	0.951 ± 0.001
Covertype	581012	2760/2760	29050 (5%)	0.715 ± 0.005	0.684 ± 0.008	0.697 ± 0.001
		2760/27600		0.729 ± 0.04	0.709 ± 0.05	–
		2760/82800		0.739 ± 0.04	0.716 ± 0.03	–
		2760/138000		0.742 ± 0.05	0.723 ± 0.06	–

FS semi-KSC: Fixed-size semi-supervised KSC
RD semi-KSC: other subset selection related to [Lee & Mangasarian, 2001]
Lap-SVM: Laplacian support vector machine [Belkin et al., 2006]

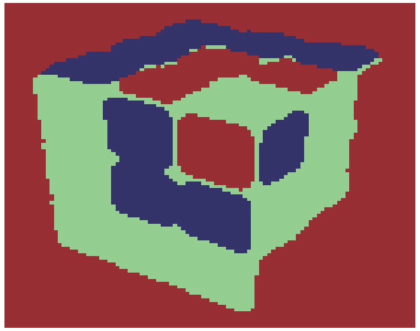
[Mehrkanoon & Suykens, 2014]

Semi-supervised learning using KSC (3)

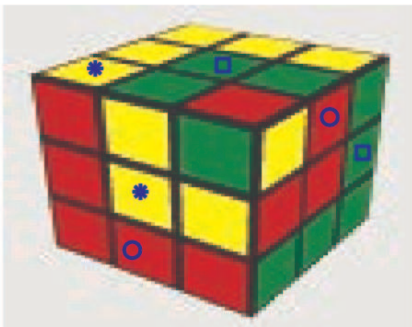
original image



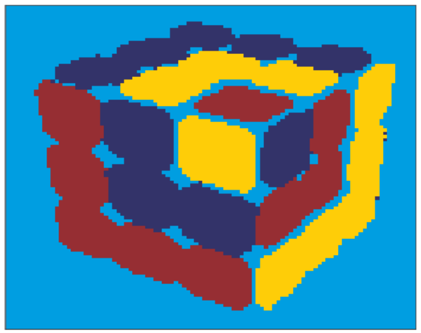
KSC



given a few labels



semi-supervised KSC



[Mehrkanoon, Alzate, Mall, Langone, Suykens, IEEE-TNNLS 2015], videos

SVD from LS-SVM

SVD within the LS-SVM setting (1)

- **Singular Value Decomposition (SVD)** of $A \in \mathbb{R}^{N \times M}$

$$A = U \Sigma V^T$$

  with $U^T U = I_N$, $V^T V = I_M$, $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_p) \in \mathbb{R}^{N \times M}$.

- Obtain two sets of data points (rows and columns): $x_i = A^T \epsilon_i$, $z_j = A \varepsilon_j$ for $i = 1, ..., N$, $j = 1, ..., M$ where $\epsilon_i, \varepsilon_j$ are standard basis vectors of dimension N and M.

- **Compatible feature maps:** $\varphi : \mathbb{R}^M \to \mathbb{R}^N$, $\psi : \mathbb{R}^N \to \mathbb{R}^N$ where

$$\begin{array}{rcl} \varphi(x_i) & = & C^T x_i = C^T A^T \epsilon_i \\ \psi(z_j) & = & z_j = A \varepsilon_j \end{array}$$

  with $C \in \mathbb{R}^{M \times N}$ a **compatibility matrix**.

[Suykens, ACHA, 2015, in press]

SVD within the LS-SVM setting (2)

- Primal problem:

$$\min_{w,v,e,r} -w^T v + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2 + \frac{1}{2}\gamma \sum_{j=1}^{M} r_j^2 \text{ subject to } \begin{array}{l} e_i = w^T \varphi(x_i), \ i = 1, ..., N \\ r_j = v^T \psi(z_j), \ j = 1, ..., M \end{array}$$

- From the Lagrangian and conditions for optimality one obtains:

$$\begin{array}{rcl} \left[\varphi(x_i)^T \psi(z_j)\right] [\beta] & = & [\alpha] \tilde{\Lambda} \\ \left[\psi(z_j)^T \varphi(x_i)\right] [\alpha] & = & [\beta] \tilde{\Lambda} \end{array}$$

- **Theorem**: If $ACA = A$ holds, this corresponds to the shifted eigenvalue problem in Lanczos' decomposition theorem.

- Goes beyond the use of Mercer theorem; extensions to nonlinear SVDs

[Suykens, ACHA, 2015, in press]

Conclusions

- **Synergies** parametric and kernel based-modelling
- **Primal and dual** representations
- Sparse kernel models using **fixed-size method**
- Applications in **supervised and unsupervised learning** and beyond
- **Finite and infinite** dimensional case
- **Beyond Mercer kernels**

Software: see ERC AdG A-DATADRIVE-B website
www.esat.kuleuven.be/stadius/ADB/software.php

Acknowledgements (1)

- Co-workers at ESAT-STADIUS:

  M. Agudelo, C. Alaiz, C. Alzate, A. Argyriou, R. Castro, J. De Brabanter, K. De Brabanter, L. De Lathauwer, B. De Moor, M. Espinoza, M. Fanuel, Y. Feng, E. Frandi, B. Gauthier, D. Geebelen, H. Hang, X. Huang, L. Houthuys, V. Jumutc, Z. Karevan, R. Langone, Y. Liu, R. Mall, S. Mehrkanoon, M. Novak, J. Puertas, L. Shi, M. Signoretto, V. Van Belle, J. Vandewalle, S. Van Huffel, C. Varon, X. Xi, Y. Yang, and others

- Many people for joint work, discussions, invitations, organizations

- Support from ERC AdG A-DATADRIVE-B, KU Leuven, GOA-MaNet, OPTEC, IUAP DYSCO, FWO projects, IWT, iMinds, BIL, COST

Acknowledgements (2)

Thank you