Learning with primal and dual model representations: a unifying picture

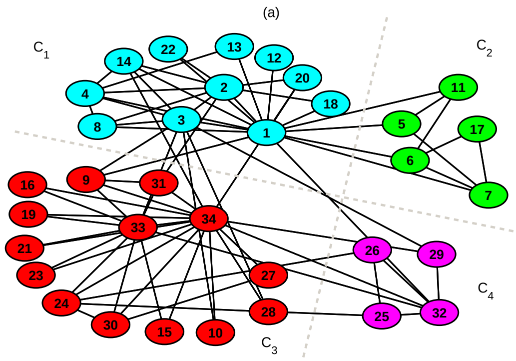
Johan Suykens

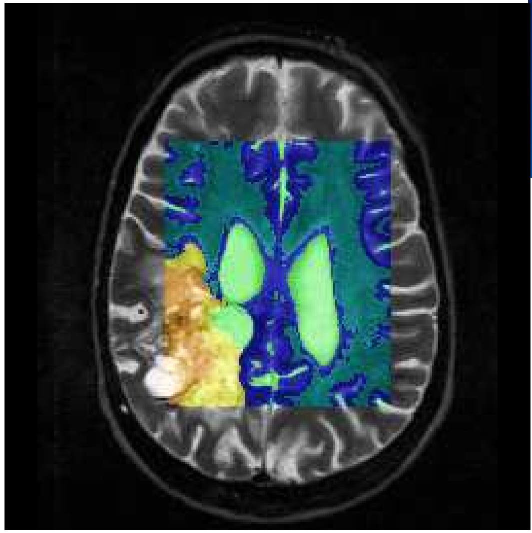
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.be
http://www.esat.kuleuven.be/stadius/

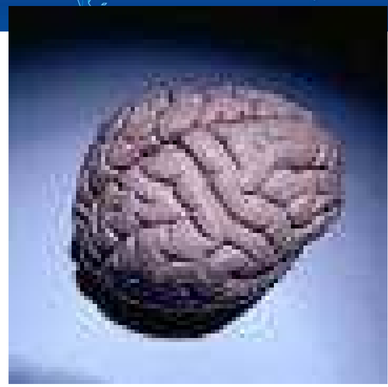
Plenary talk ICASSP 2016 Shanghai

This talk is dedicated to all victims of war and terrorism.
Our thoughts are with the victims and their families.

Data & Signals World







Black-box weather forecasting

Weather data
350 stations located in US

Features:
Tmax, Tmin, precipitation, wind speed, wind direction ,...

Black-box forecasting multiple weather stations simultaneously

[Signoretto, Frandi, Karevan, Suykens, IEEE-SCCI, 2014]

Challenges

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

Outline talk

- Sparsity in parametric and kernel based models
- Learning with primal and dual representations:
  - Supervised and unsupervised learning, and beyond
  - Sparsity, robustness, networks, big data
- New variational principle for SVD
- New unifying theory for deep learning and kernel machines

Different paradigms

SVM &

Kernel methods

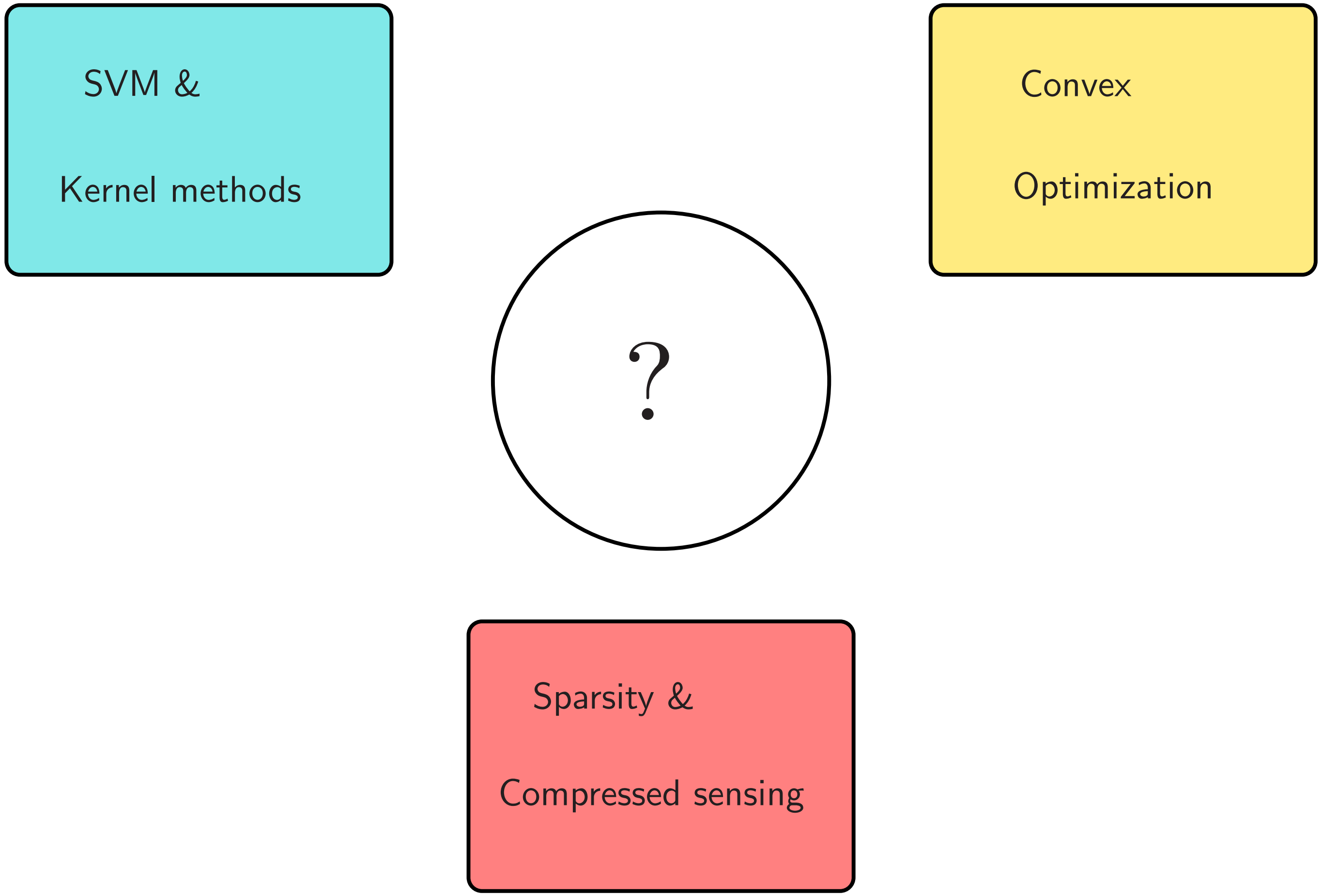
Convex

Optimization

Sparsity &

Compressed sensing

Different paradigms



Sparsity through regularization or loss function

Sparsity: through regularization or loss function

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

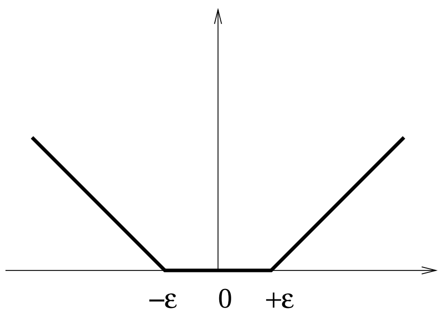
$$\min \sum_j |w_j| + \gamma \sum_i e_i^2$$

$\Rightarrow$ 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

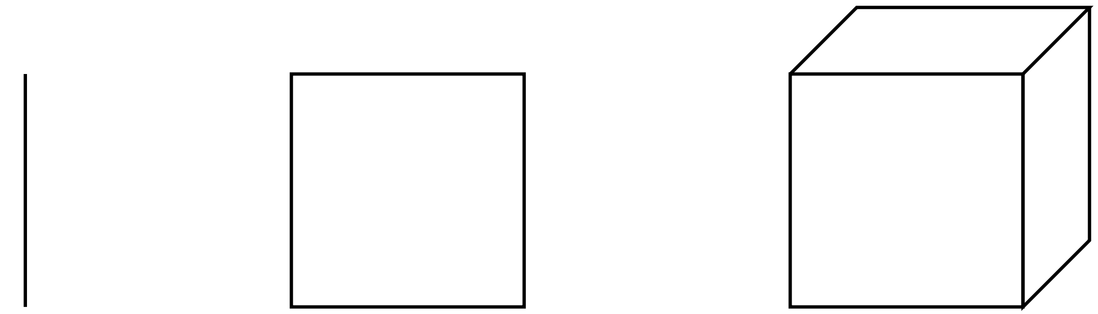


Sparsity: matrices and tensors

neuroscience: EEG data
(time samples $\times$ frequency $\times$ electrodes)

computer vision: image (/video) compression/completion/$\cdots$
(pixel $\times$ illumination $\times$ expression $\times \cdots$)

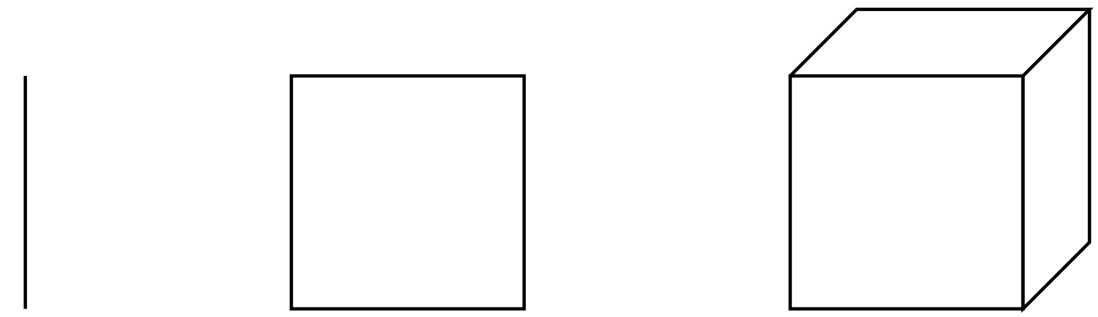
web mining: analyze users behaviors
(users $\times$ queries $\times$ webpages)



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:

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:
$\sum_j \|w_j\|$		$\\|W\\|_*$		$\\|\mathcal{W}\\|_*$

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]

Kernels

Wide range of positive definite kernel functions possible:

- linear $K(x, z) = x^T z$
- polynomial $K(x, z) = (\eta + x^T z)^d$
- radial basis function $K(x, z) = \exp(-\|x - z\|_2^2/\sigma^2)$
- splines
- wavelets
- string kernel
- kernels from graphical models
- Fisher kernels
- graph kernels
- data fusion kernels
- tensorial kernels
- other

[Schölkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004; Jebara et al., 2004; other]

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

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

- 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

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

- Classification

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

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

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

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

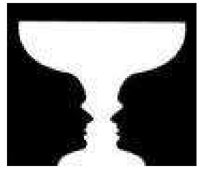
- **Quantum measurement**: state vector $|\psi\rangle$, measurement operators M_i
  – *Primal:*

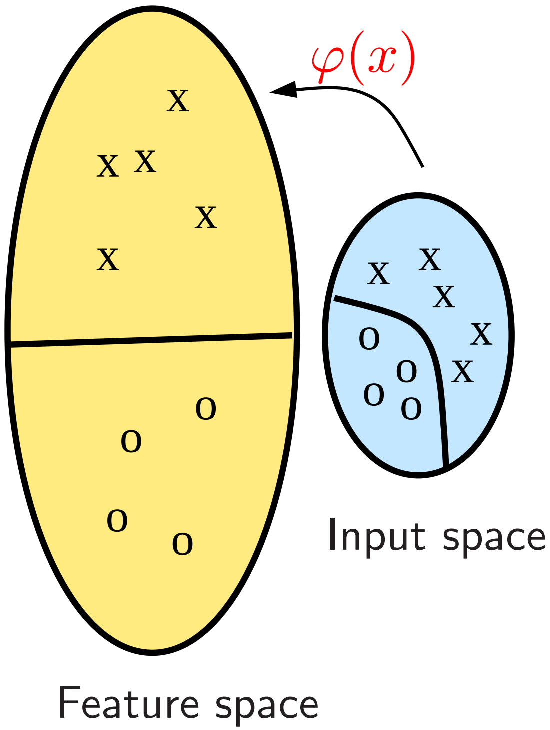
$$\min_{|w\rangle, p_i} \tfrac{1}{2}\langle w|w\rangle \text{ subject to } \quad p_i = \mathrm{Re}(\langle w|M_i\psi\rangle), \, i = 1, ..., N \text{ and } \textstyle\sum_{i=1}^N p_i = 1$$

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

[Suykens, Physical Review A, 2013]

SVMs: living in two worlds ...

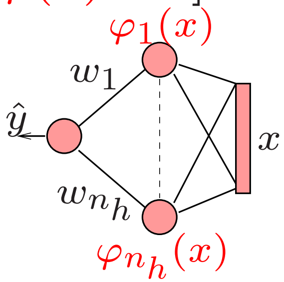




Primal space

Parametric

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

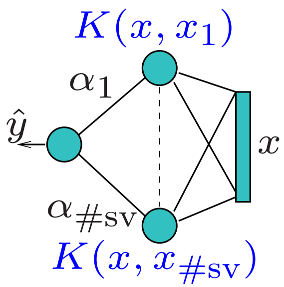


$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \quad \textbf{(Mercer)}$$

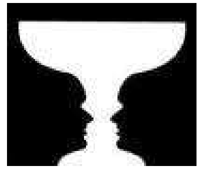
Dual space

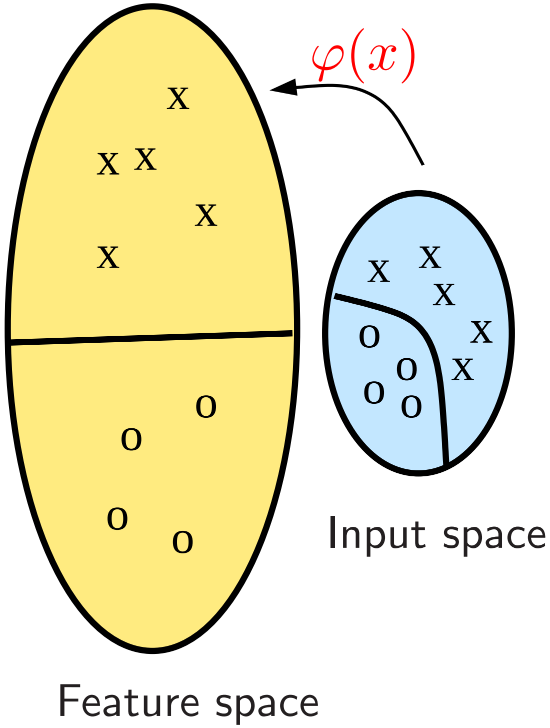
Nonparametric

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



SVMs: living in two worlds ...

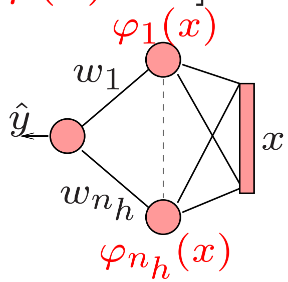




Primal space

Parametric

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

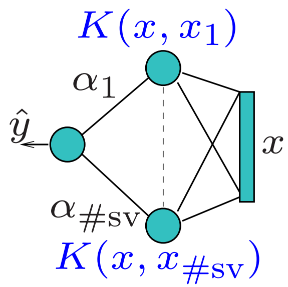


$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$$ **("Kernel trick")**

Dual space

Nonparametric

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



Parametric

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?

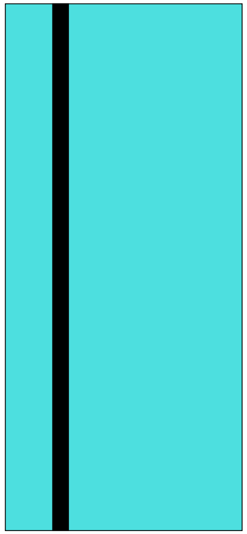
few inputs, many data points: $d \ll N$

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]

Hilbert space to RKHS Transform

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

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

-

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

Hilbert space to RKHS 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:** $W_\eta : \mathcal{H} \to \mathcal{H}_K : |w\rangle \mapsto \langle \eta_\cdot|w\rangle_{\mathcal{H}}$

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

[Fanuel & Suykens, TR15-101, 2015]: including wavelet transform, graph wavelets

Learning in Krein spaces: indefinite kernels

- LS-SVM classifier for indefinite kernel case:

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

with **indefinite kernel** K

$$K(x_i, x_j) = K_+(x_i, x_j) - K_-(x_i, x_j)$$

with positive definite kernels K_+, K_-

$$K_+(x_i, x_j) = \varphi_+(x_i)^T \varphi_+(x_j) \text{ and } K_-(x_i, x_j) = \varphi_-(x_i)^T \varphi_-(x_j)$$

- similarly also for kernel PCA with indefinite kernel

[X. Huang, Maier, Hornegger, Suykens, TR15-214, 2015]

Related work of RKKS: [Ong et al 2004; Haasdonk 2005; Luss 2008; Loosli et al. 2015]

Learning in Banach spaces: generalized SVR

- Continuous representer theorem (from Fenchel-Rockafellar duality) for

$$\min_{(w,b,e)\in\mathcal{F}\times\mathbb{R}\times L^p(P)} G(w)+\gamma\int_{\mathcal{X}\times\mathcal{Y}} L(e(x,y))dP(x,y) \quad \text{s.t. } y-\langle w,\varphi(x)\rangle-b=e(x,y)$$
$$\forall P-a.a.(x,y)\in\mathcal{X}\times\mathcal{Y}$$

- Special case:

$$\min_{(w,b,e)\in\ell^r(\mathbb{K})\times\mathbb{R}\times\mathbb{R}^N} \rho(\|w\|_r)+\frac{\gamma}{N}\sum_{i=1}^N L(e_i) \text{ s.t. } y_i-\langle w,\varphi(x_i)\rangle-b=e_i\,,\forall i$$

with $\mathcal{F}=\ell^r(\mathbb{K})$ and $r=\frac{m}{m-1}$ for even $m\geq 2$, ρ convex and even (approaches ℓ^1 regularization for m large)

- **Tensor-kernel representation** ($b=0$), matrix case $K(x_i,x)$ for $m=2$:

$$\hat{y}=\langle w,\varphi(x)\rangle_{r,r^*}=\frac{1}{N^{m-1}}\sum_{i_1,\ldots,i_{m-1}=1}^N u_{i_1}\ldots u_{i_{m-1}}K(x_{i_1},\ldots,x_{i_{m-1}},x)$$

[Salzo & Suykens, arXiv 1603.05876], related: RKBS [Zhang 2013; Fasshauer et al. 2015]

Sparsity by fixed-size kernel method

Fixed-size method: steps

1. **selection of a subset** from the data (random, quadratic Renyi entropy, incomplete Cholesky factorization, other)

2. kernel matrix on the subset

3. eigenvalue decomposition of kernel matrix

4. **approximation of the feature map** based on the eigenvectors (Nyström approximation) [Williams & Seeger, 2001]

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)*

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 [De Brabanter et al., CSDA 2010]
- Challenging to achieve high performance by very sparse models:
  - Mall & Suykens [TNNLS 2015]: Very Sparse LSSVM Reductions
  - Gauthier & Suykens [KU Leuven TR16-26, 2016]: Energy and Discrepancy SVMs

Kernel PCA and kernel 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]

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

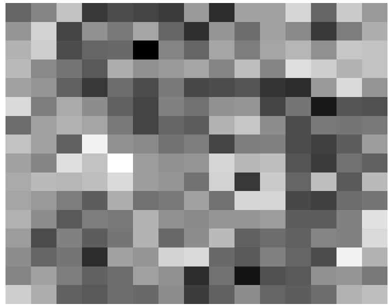
with $\Omega_{c,ij} = (\varphi(x_i) - \hat{\mu}_\varphi)^T (\varphi(x_j) - \hat{\mu}_\varphi)$ the *centered kernel matrix*.

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

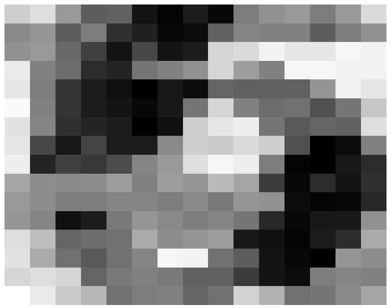
Robustness: Kernel Component Analysis

original image

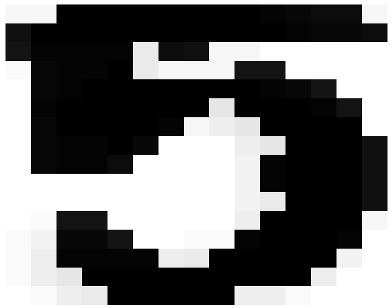
corrupted image



KPCA reconstruction



KCA reconstruction



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

Kernel Spectral Clustering (KSC)

- **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 $\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$

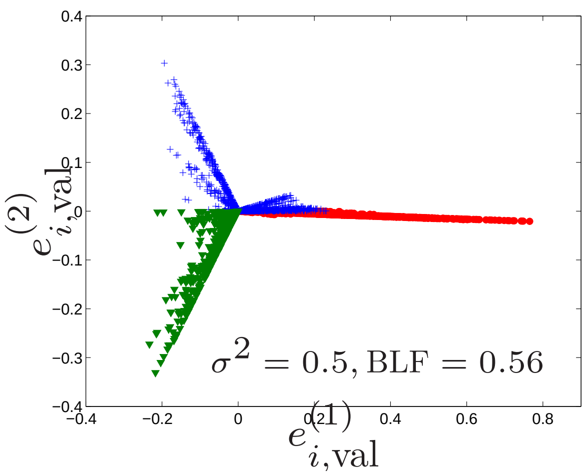
- Taking $V = D^{-1}$ with degree matrix $D = \text{diag}\{d_i\}$, $d_i = \sum_{j=1}^N \Omega_{ij}$ relates to random walks algorithm [Chung, 1997; Shi & Malik, 2000; Ng 2002]

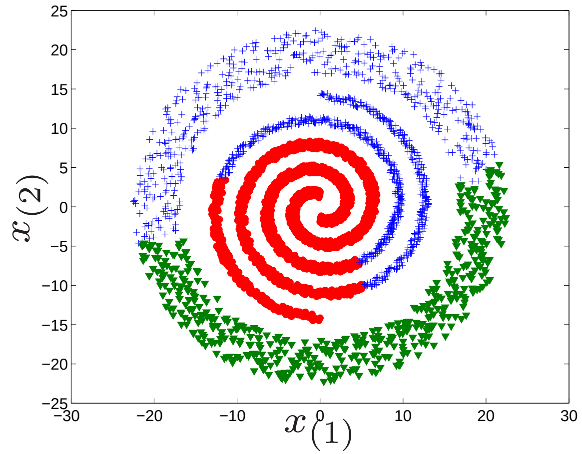
[Alzate & Suykens, IEEE-PAMI, 2010]

Advantages of kernel-based setting

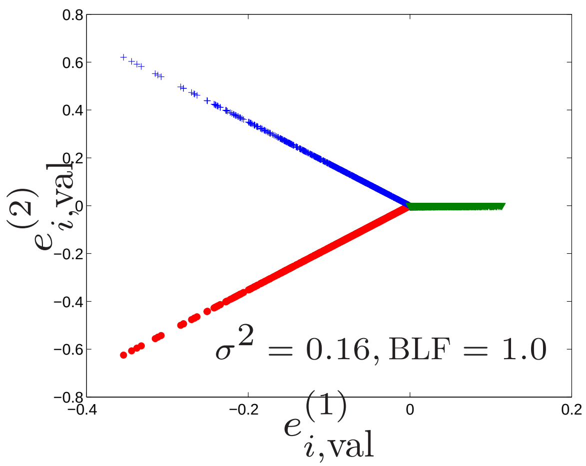
- **model-based** approach
- **out-of-sample extensions**, applying model to new data
- consider **training, validation and test data**
  (training problem corresponds to eigenvalue decomposition problem)
- model selection procedures
- **sparse representations and large scale methods**

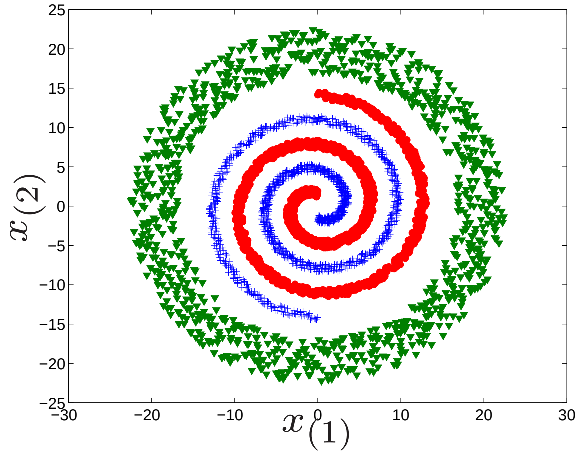
Model selection: toy example





BAD



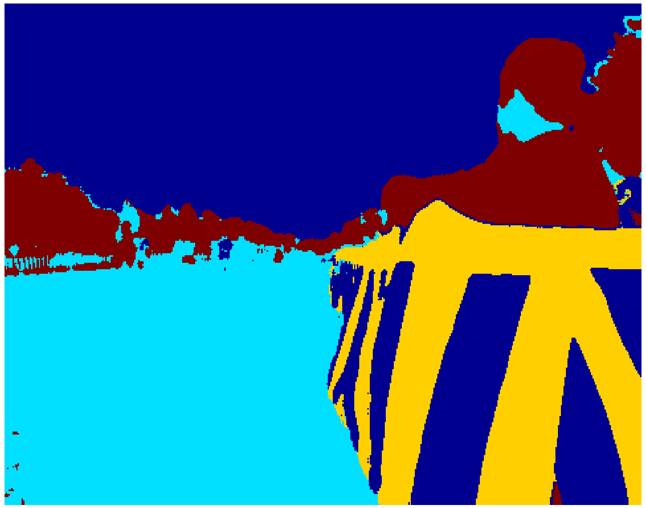


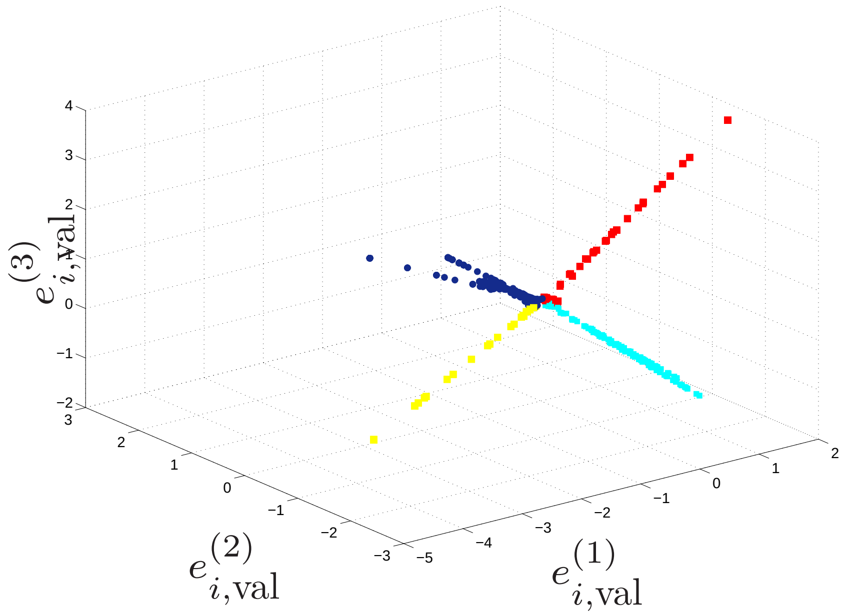
GOOD

validation set

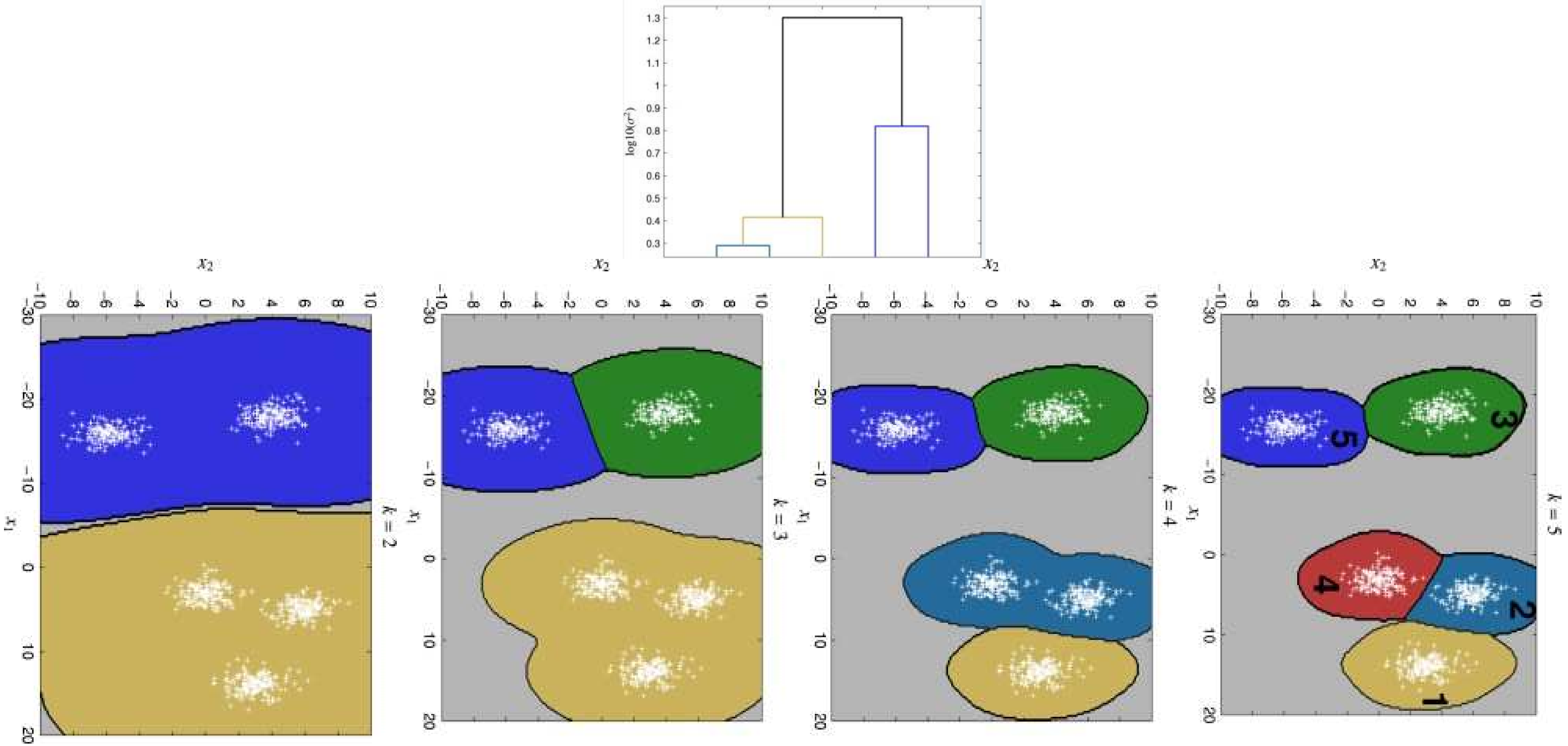
train + validation + test data

Example: image segmentation



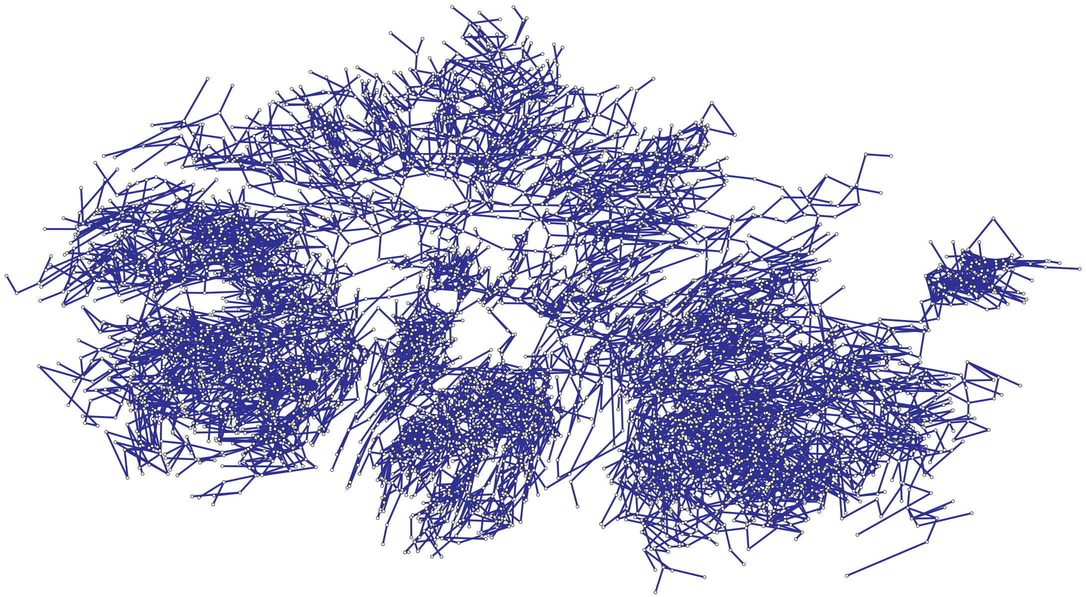


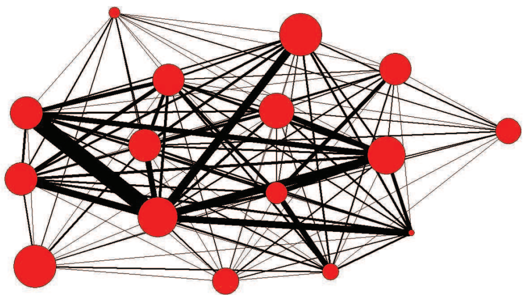
Hierarchical KSC



[Alzate & Suykens, 2012]

Representative subgraphs - Power grid network





KSC community detection, representative subgraphs [Langone et al., 2012]
Western USA power grid: 4941 nodes, 6594 edges [Watts & Strogatz, 1998]

KSC for big data networks (1)

- Select **representative training subgraph**
- Perform **model selection** using Balanced Angular Fit (BAF) (cosine similarity measure) related to the e-projection values on validation nodes)
- Train the KSC model by solving a small eigenvalue problem of size $\min(0.15N, 5000)^2$
- Apply **out-of-sample extension** to find cluster memberships of the remaining nodes

Dataset	Nodes	Edges
YouTube	1,134,890	2,987,624
roadCA	1,965,206	5,533,214
Livejournal	3,997,962	34,681,189

[Mall, Langone, Suykens, Entropy, special issue Big data, 2013]

KSC for big data networks (2)

BAF-KSC	[Mall, Langone, Suykens, 2013]	**Louvain**	[Blondel et al., 2008]
Infomap	[Lancichinetti, Fortunato, 2009]	**CNM**	[Clauset, Newman, Moore, 2004]

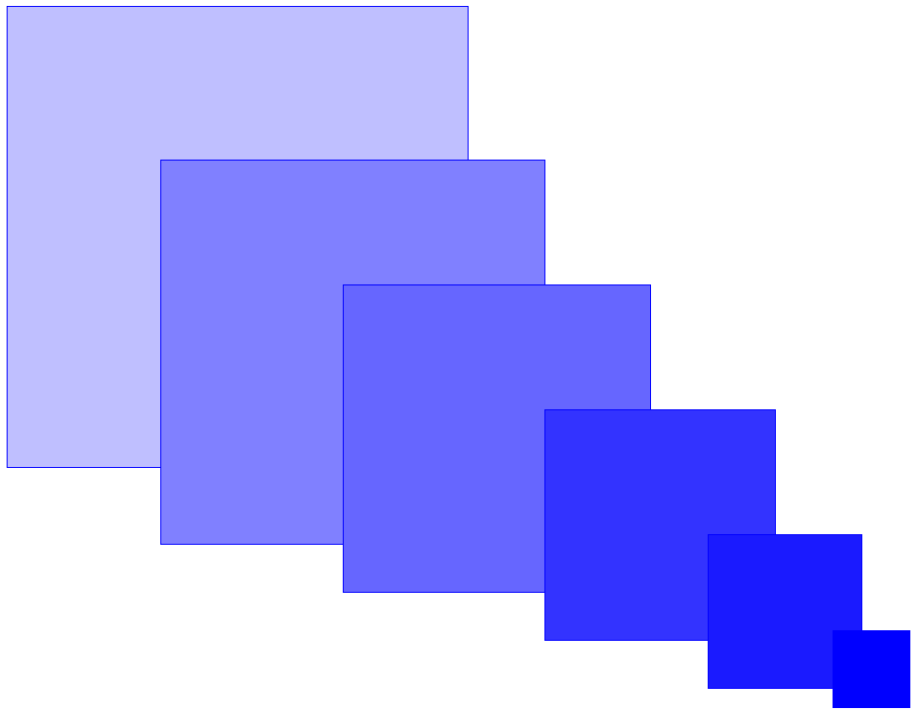
Dataset	BAF-KSC			Louvain			Infomap			CNM		
	Cl	Q	Con	Cl	Q	Con	Cl	Q	Con	Cl	Q	Con
Openflight	5	0.533	**0.002**	109	**0.61**	0.02	18	0.58	0.005	84	0.60	0.016
PGPnet	8	0.58	**0.002**	105	**0.88**	0.045	84	0.87	0.03	193	0.85	0.041
Metabolic	10	0.22	0.028	10	**0.43**	0.03	41	0.41	0.05	11	0.42	0.021
HepTh	6	0.45	**0.0004**	172	**0.65**	0.004	171	0.3	0.004	6	0.423	0.0004
HepPh	5	0.56	0.0004	82	**0.72**	0.007	69	0.62	0.06	6	0.48	0.0007
Enron	10	0.4	0.002	1272	**0.62**	0.05	1099	0.37	0.27	6	0.25	0.0045
Epinion	10	**0.22**	0.0003	33	0.006	0.0003	17	0.18	0.0002	10	0.14	**0.0**
Condmat	6	0.28	**0.0002**	1030	**0.79**	0.03	1086	0.79	0.025	8	0.38	0.0003

Flight network (Openflights), network based on trust (PGPnet), biological network (Metabolic), citation networks (HepTh, HepPh), communication network (Enron), review based network (Epinion), collaboration network (Condmat) [snap.stanford.edu]

Cl = Clusters, Q = modularity, Con = Conductance

BAF-KSC usually finds a smaller number of clusters and achieves lower conductance

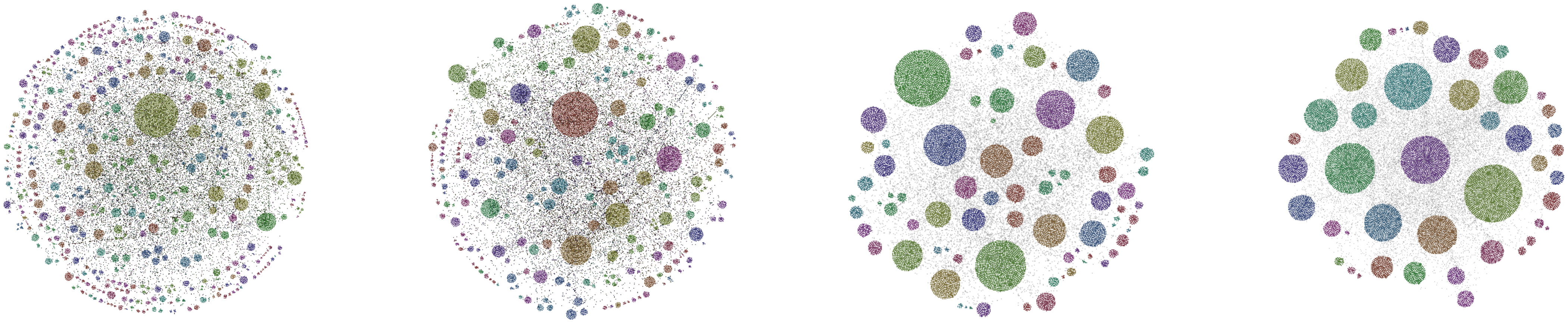
Multilevel Hierarchical KSC for complex networks (1)



Generating a series of affinity matrices over different levels:
communities at level h become nodes for next level $h+1$

Multilevel Hierarchical KSC for complex networks (2)

MH-KSC on PGP network:

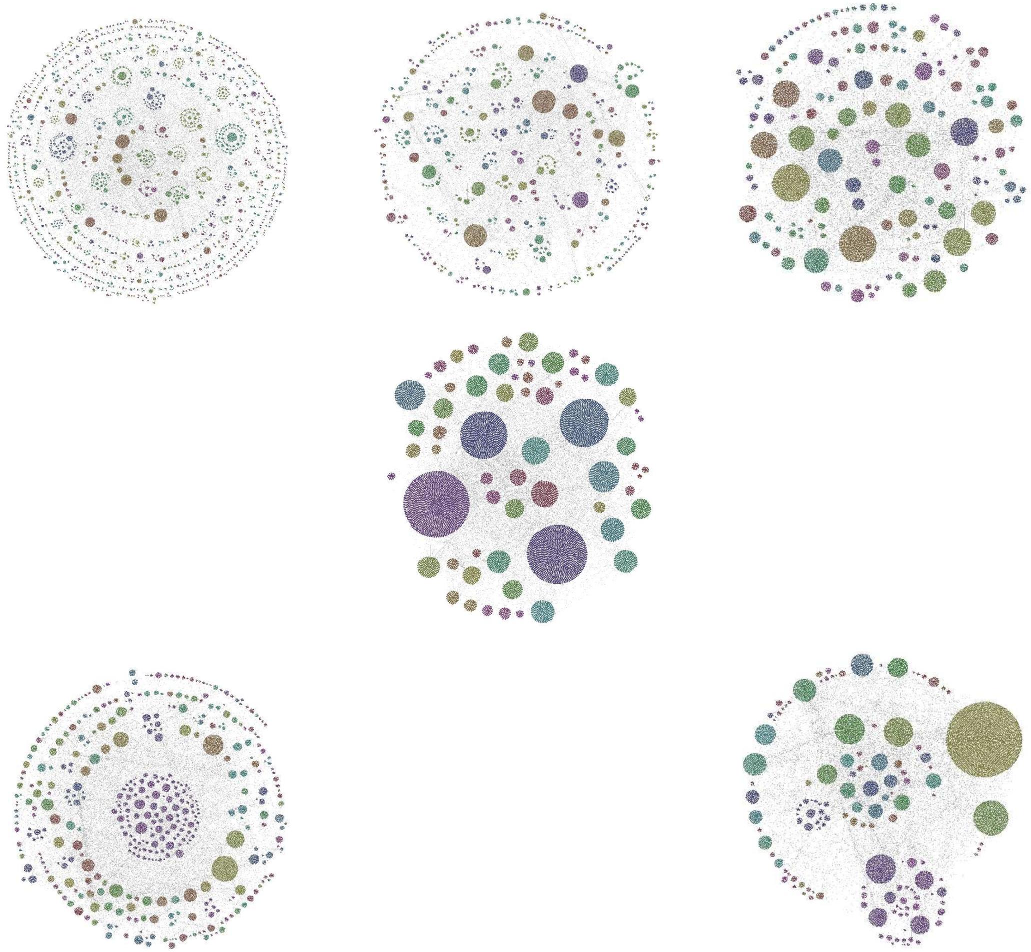


fine intermediate intermediate coarse

Multilevel Hierarchical KSC finds high quality clusters at coarse as well as fine and intermediate levels of hierarchy.

[Mall, Langone, Suykens, PLOS ONE, 2014]

Multilevel Hierarchical KSC for complex networks (3)



Louvain

Infomap

OSLOM

Louvain, Infomap, and OSLOM seem biased toward a particular scale in comparison with MH-KSC, based upon ARI, VI, Q metrics

Big data: representative subsets using k-NN graphs (1)

- Convert the large scale dataset into a sparse undirected k-NN graph using a distributed network generation framework

- Julia language (http://julialang.org/)

- Large $N \times N$ kernel matrix Ω for data set $\mathcal{D}$ with N data points

- Batch cluster-based approach: a batch subset $\mathcal{D}_p \subset \mathcal{D}$ is loaded per node with $\cup_{p=1}^{P} \mathcal{D}_p = \mathcal{D}$; related matrix slice X_p and Ω_p.

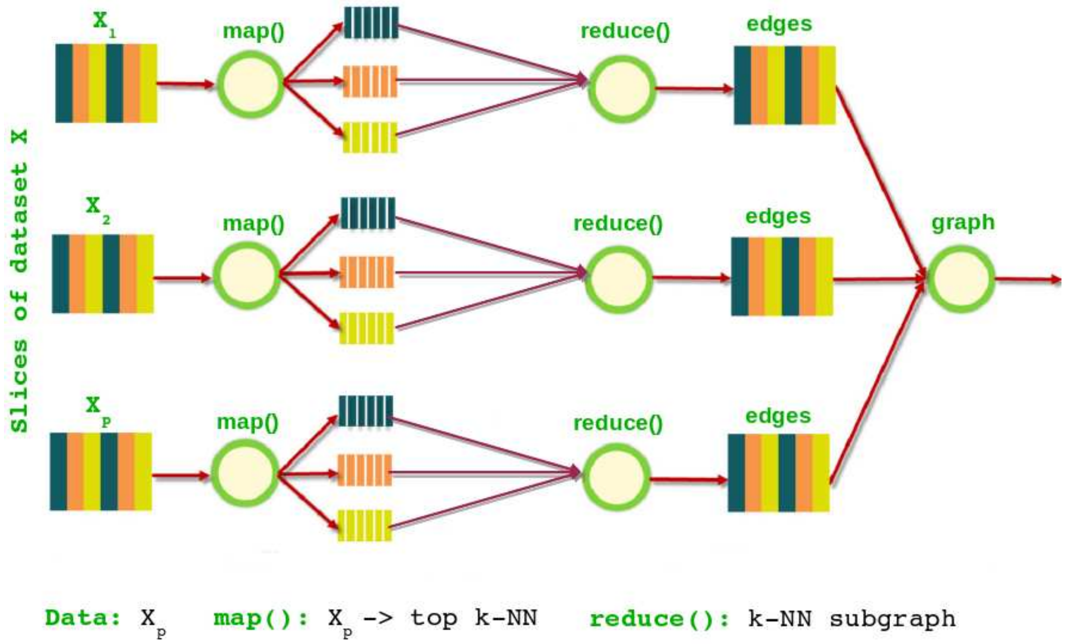
- MapReduce and AllReduce settings implementable using Hadoop or Spark (see also [Agarwal et al., JMLR 2014] Terascale linear learning)

- Computational complexity: complexity for construction of the kernel matrix reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N^2(1 + \log N)/P)$ for P nodes

[Mall, Jumutc, Langone, Suykens, IEEE Bigdata 2014]

Big data: representative subsets using k-NN graphs (2)

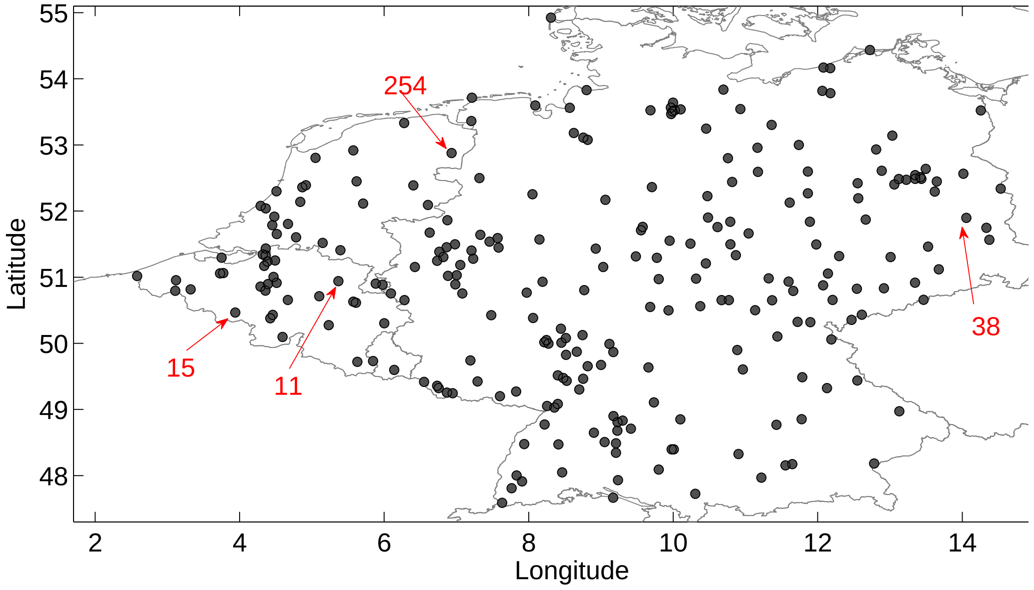
Map: for Silverman's Rule of Thumb, compute mean and standard deviation of the data per node; compute slice Ω_p; sort in ascending order the columns of Ω_p (sortperm in Julia); pick indices for top k values.
Reduce: merge k-NN subgraphs into aggregated k-NN graph

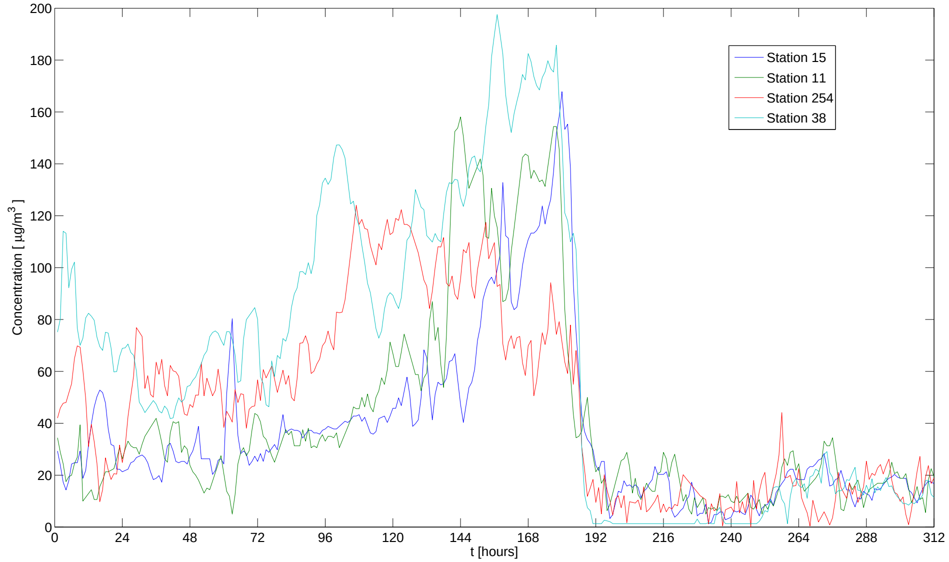


[Mall, Jumutc, Langone, Suykens, IEEE Bigdata 2014]

Incremental KSC clustering of PM10 concentrations (1)

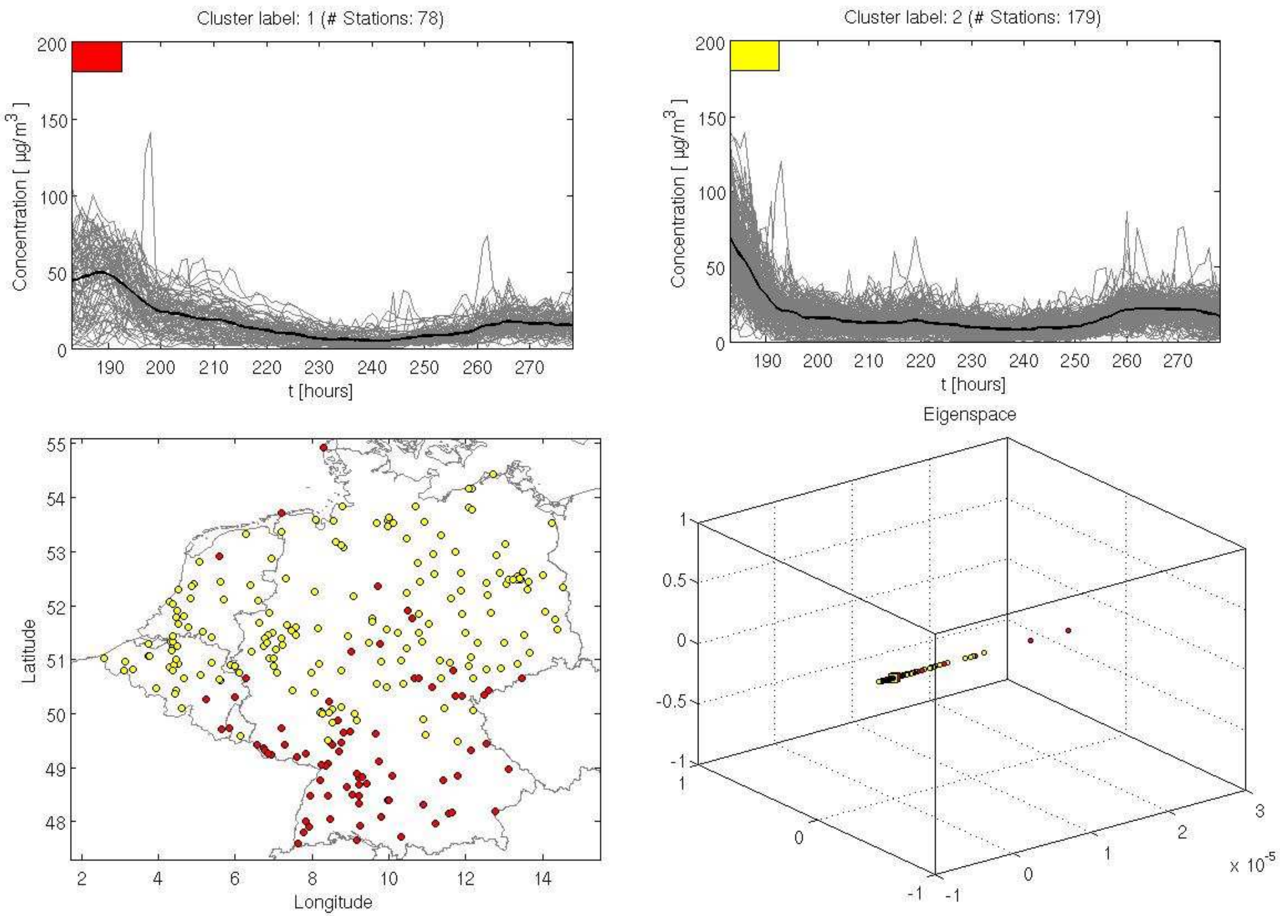
PM10 time-series: PM10 data (Particulate Matter) registered during a heavy pollution episode (Jan 20 2010 - Feb 1 2010) in Europe.





[Langone, Agudelo, De Moor, Suykens, Neurocomputing, 2014]

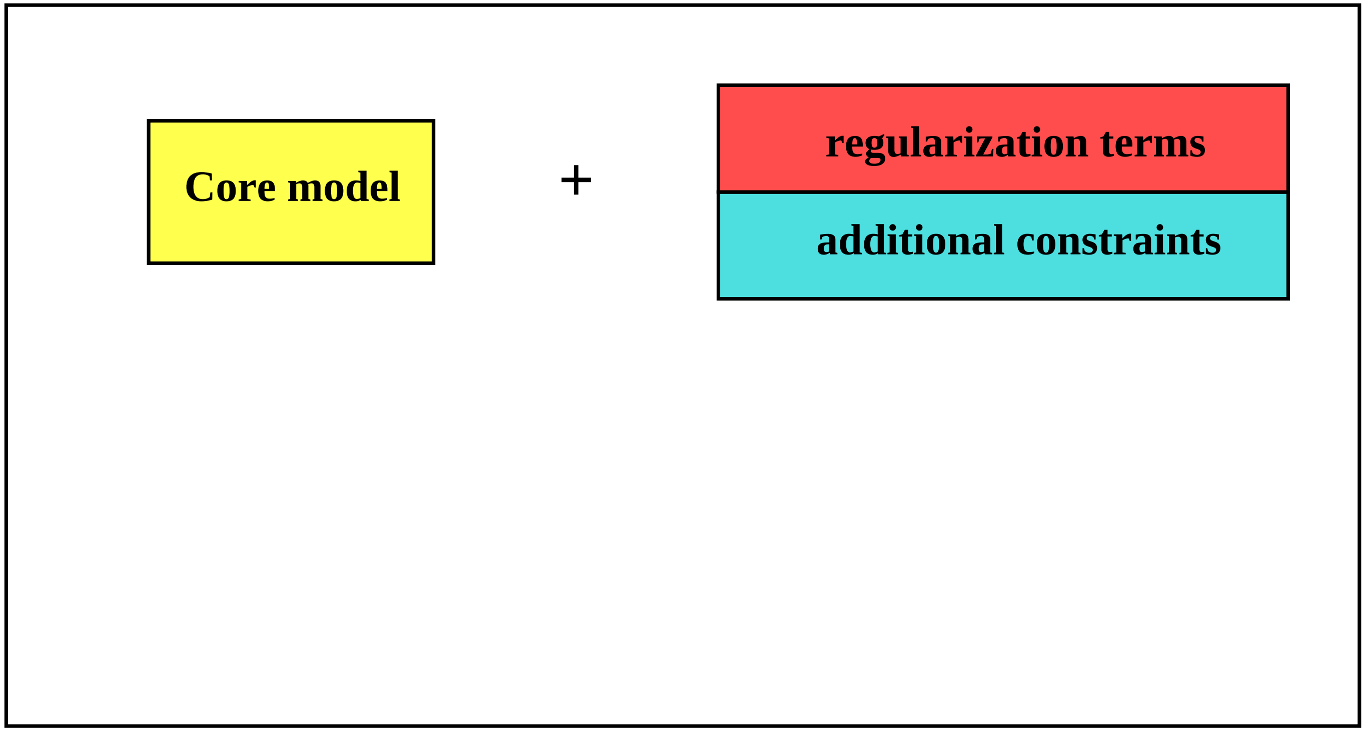
Incremental KSC clustering of PM10 concentrations (2)



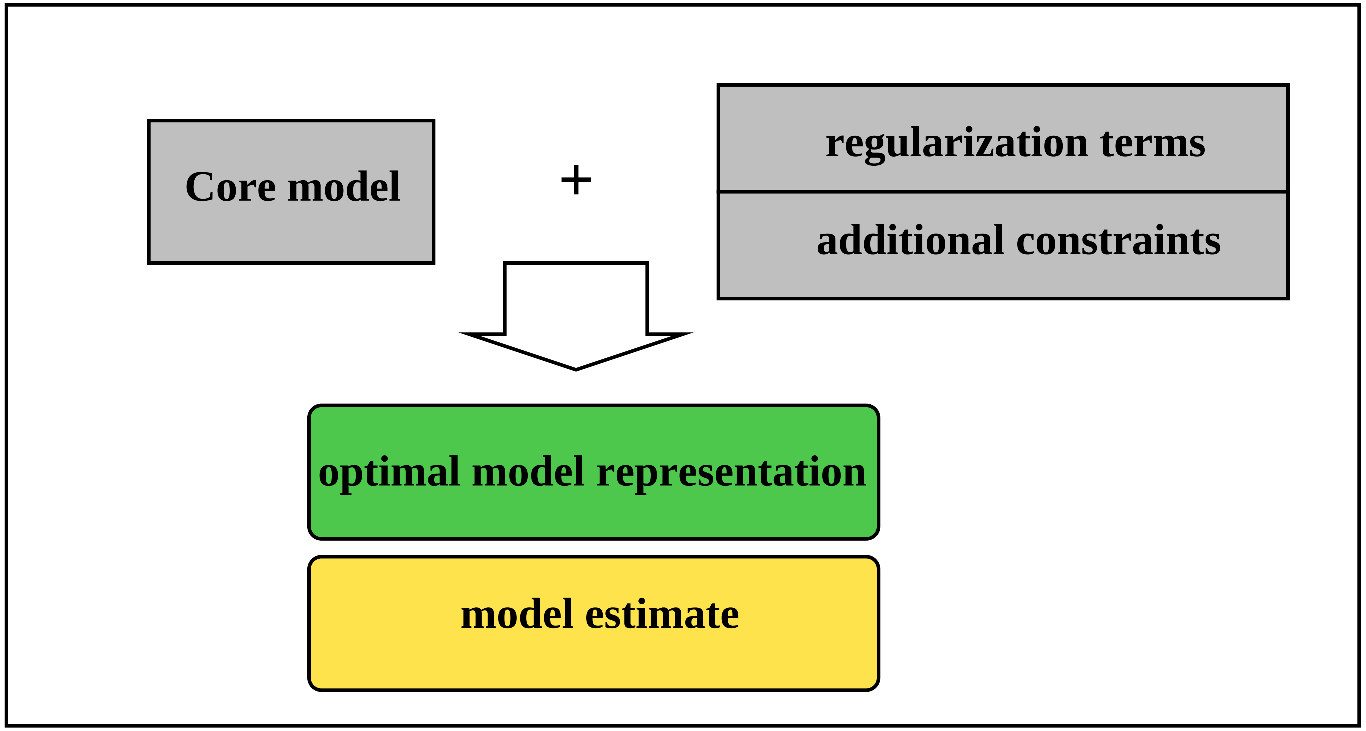
Applies *out-of-sample eigenvectors* for fast incremental KSC learning

video - [Langone, Agudelo, De Moor, Suykens, Neurocomputing, 2014]

Core models + constraints

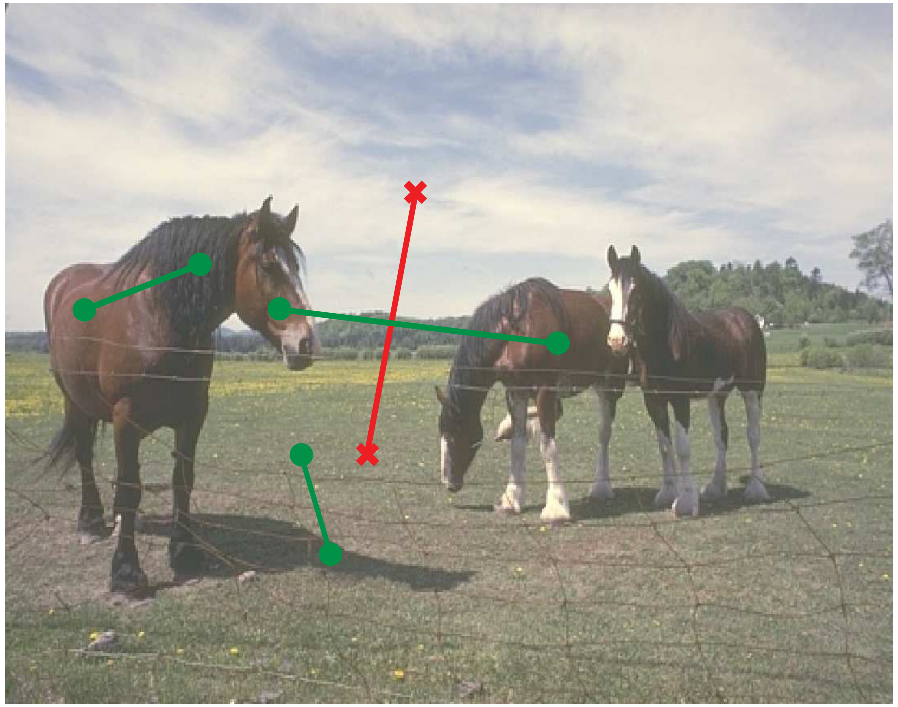


Core models + constraints



Adding prior knowledge: example

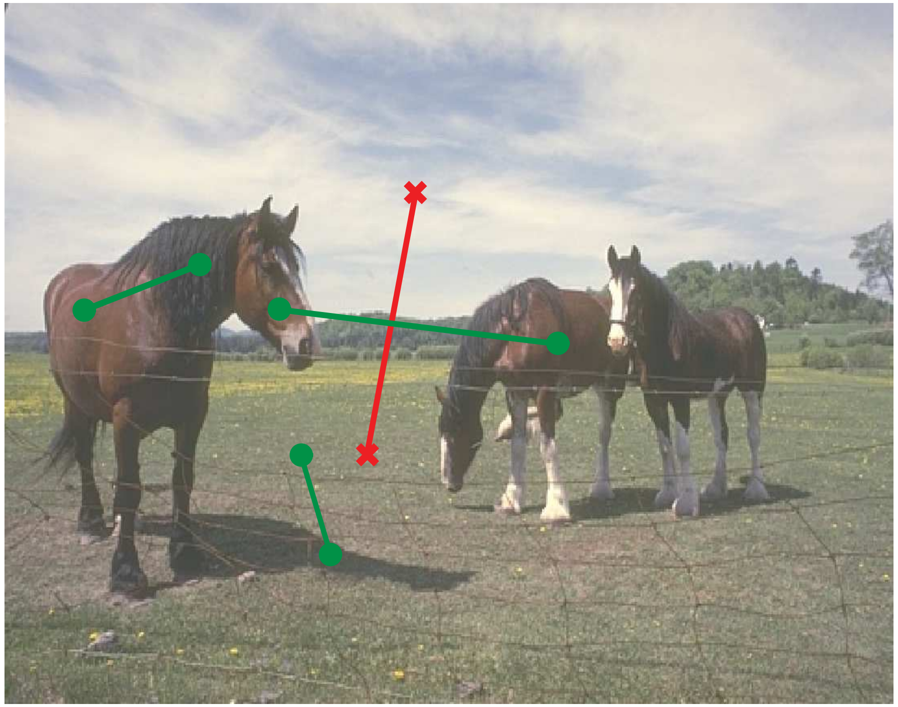
original image



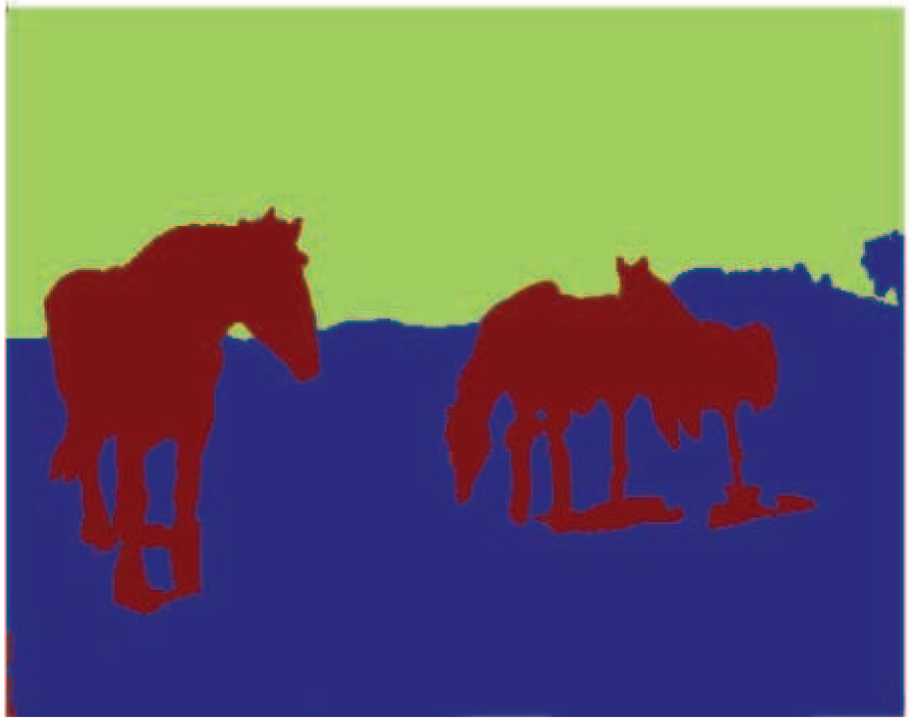
without constraints

Adding prior knowledge: example

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])

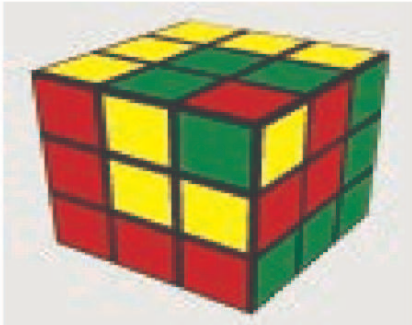
$$\min_{w,e,b} \quad \frac{1}{2}w^T w - \gamma \frac{1}{2}e^T D^{-1} e + \rho \frac{1}{2} \sum_{m=N+1}^{M} (e_m - y_m)^2$$
$$\text{subject to} \quad e_i = w^T \varphi(x_i) + b, \; i = 1, ..., M$$

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

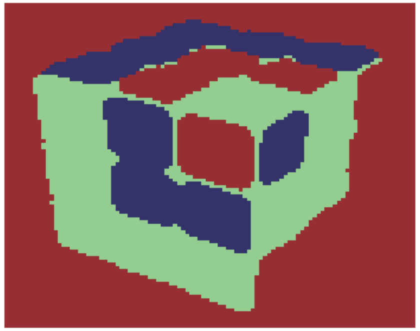
- Other approaches in semi-supervised learning and manifold learning, e.g. [Belkin et al., 2006]

Semi-supervised learning using KSC (2)

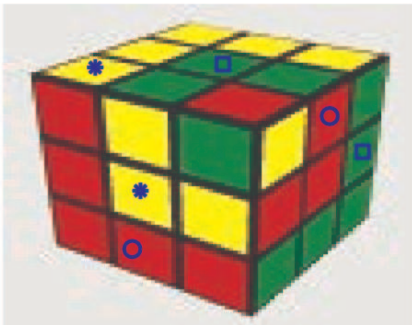
original image



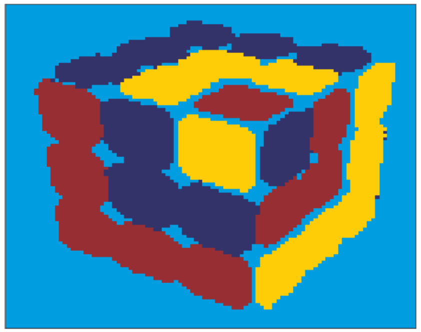
KSC



given a few labels



semi-supervised KSC



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

SVD from LS-SVM setting

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

- 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:** $\varphi : \mathbb{R}^M \rightarrow \mathbb{R}^N$, $\psi : \mathbb{R}^N \rightarrow \mathbb{R}^N$ 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 2016]

SVD within the LS-SVM setting (2)

- **Primal problem** (new variational principle):

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

- From the Lagrangian and conditions for optimality one obtains:

$$\left[\begin{array}{cc} 0 & [\varphi(x_i)^T \psi(z_j)] \\ [\psi(z_j)^T \varphi(x_i)] & 0 \end{array} \right] \left[\begin{array}{c} \alpha \\ \beta \end{array} \right] = (1/\gamma) \left[\begin{array}{c} \alpha \\ \beta \end{array} \right]$$

- **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 2016]

Linear versus nonlinear SVD: example

SVD

lin+pol

exp, $\eta = 1$

exp, $\eta = -1$

original | 20 comp. | 100 comp.

New theory Deep Learning with Kernel Machines

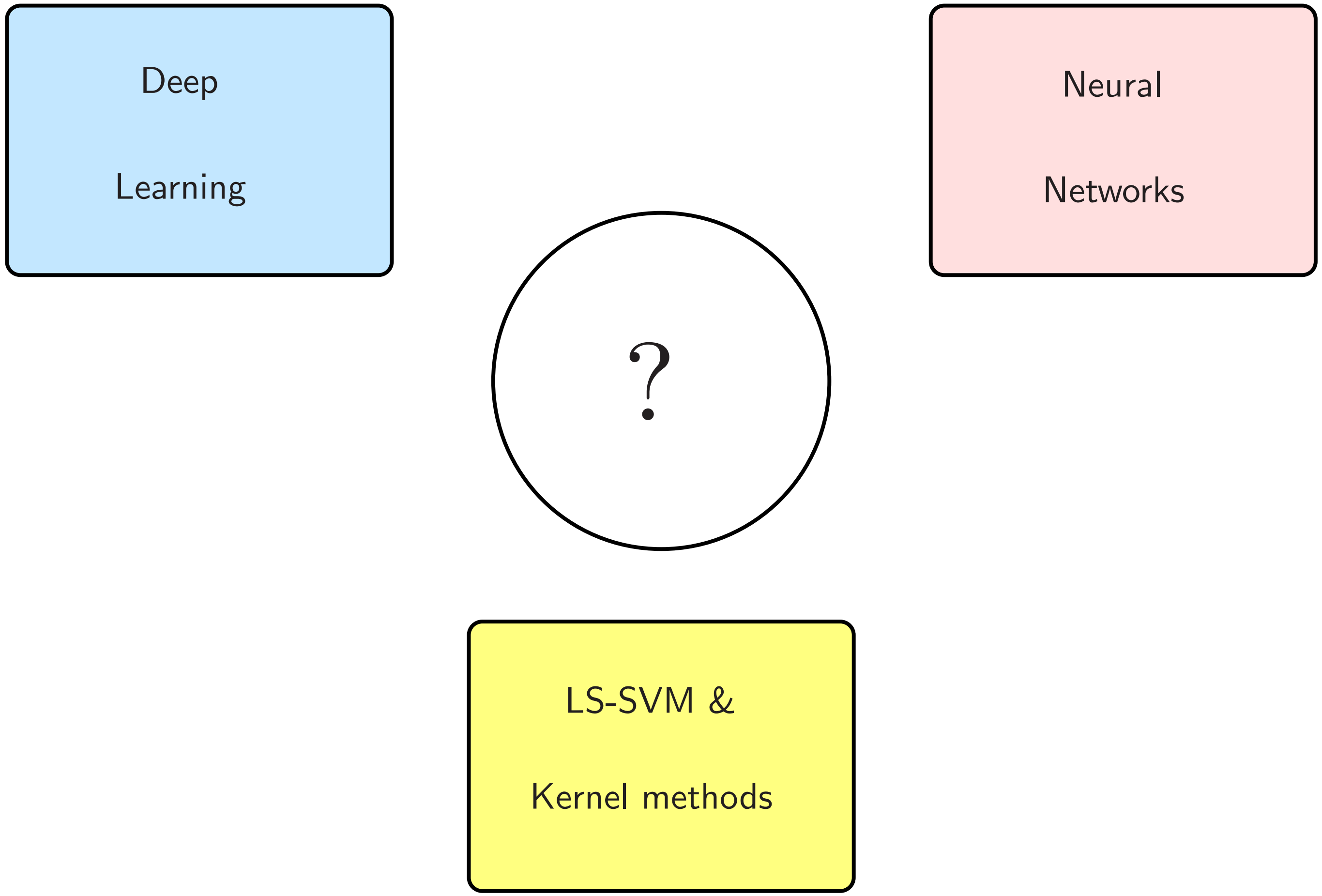
Different paradigms

Deep Learning

Neural Networks

LS-SVM & Kernel methods

Different paradigms



Deep learning

- Learning feature hierarchies
- Deep networks versus shallow networks
- Excellent performance e.g. computer vision, speech recognition, language processing
- Deep belief networks
  Deep Boltzmann machines
  Convolutional neural networks
  Stacked autoencoders with pretraining and finetuning

[LeCun, Bengio, Hinton, Nature 2015; Hinton 2005; Bengio 2009; Salakhutdinov 2015]

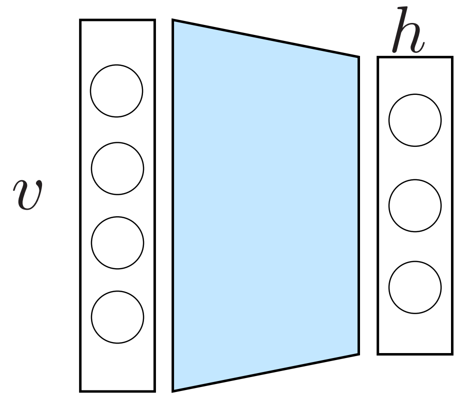
New theory Deep Learning with Kernel Machines

Main characteristics:

- Based on **conjugate feature duality**
- Interpretation of **visible and hidden units** for several kernel machines (LS-SVM regression/classification, Kernel PCA, SVD, Parzen-type)
- **Restricted Kernel Machine (RKM) representation**, related to RBM
- Neural networks interpretations (hidden layer corresponds to feature map)
- **Deep RKM** by coupling RKMs over different levels

[Suykens J.A.K., "Deep Restricted Kernel Machines using Conjugate Feature Duality", Internal Report 16-50, ESAT-SISTA, KU Leuven 2016]

Restricted Boltzmann Machines (RBM)



- Markov random field, characterized by a bipartite graph with layer of visible units v and layer of hidden units h; stochastic binary units
- **No hidden-to-hidden connections**
  Energy:
  $$E(v, h; \theta) = -v^T W h - c^T v - a^T h$$
- Joint distribution: $P(v, h; \theta) = \frac{1}{Z(\theta)} \exp(-E(v, h; \theta))$ with partition function $Z(\theta) = \sum_v \sum_h \exp(-E(v, h; \theta))$ for normalization.
- RBMs used for deep belief networks.

[Hinton, Osindero, Teh, NC 2006]

Restricted Kernel Machines (RKM) - Example LS-SVM (1)

Multi-output model $\hat{y} = W^T x + b$, $e = y - \hat{y}$

Objective in LS-SVM regression (linear case)

$$J = \frac{\eta}{2}\mathrm{Tr}(W^T W) + \frac{1}{2\lambda}\sum_{i=1}^{N} e_i^T e_i \text{ s.t. } e_i = y_i - W^T x_i - b, \forall i$$

Restricted Kernel Machines (RKM) - Example LS-SVM (1)

Multi-output model $\hat{y} = W^T x + b$, $e = y - \hat{y}$

Objective in LS-SVM regression (linear case)

$$
\begin{aligned}
J &= \frac{\eta}{2}\mathrm{Tr}(W^T W) + \frac{1}{2\lambda}\sum_{i=1}^{N} e_i^T e_i \ \text{ s.t. } e_i = y_i - W^T x_i - b, \forall i \\
&\geq \sum_{i=1}^{N} e_i^T h_i - \frac{\lambda}{2}\sum_{i=1}^{N} h_i^T h_i + \frac{\eta}{2}\mathrm{Tr}(W^T W) \ \text{ s.t. } e_i = y_i - W^T x_i - b, \forall i
\end{aligned}
$$

Restricted Kernel Machines (RKM) - Example LS-SVM (1)

Multi-output model $\hat{y} = W^T x + b$, $e = y - \hat{y}$

Objective in LS-SVM regression (linear case)

$$
\begin{aligned}
J &= \frac{\eta}{2}\mathrm{Tr}(W^T W) + \frac{1}{2\lambda}\sum_{i=1}^{N} e_i^T e_i \;\; \text{s.t. } e_i = y_i - W^T x_i - b, \forall i \\
&\geq \sum_{i=1}^{N} e_i^T h_i - \frac{\lambda}{2}\sum_{i=1}^{N} h_i^T h_i + \frac{\eta}{2}\mathrm{Tr}(W^T W) \;\; \text{s.t. } e_i = y_i - W^T x_i - b, \forall i \\
&= \sum_{i=1}^{N} (y_i^T - x_i^T W - b^T) h_i - \frac{\lambda}{2}\sum_{i=1}^{N} h_i^T h_i + \frac{\eta}{2}\mathrm{Tr}(W^T W) \triangleq \underline{J}(h_i, W, b) \\
&= R_{\mathrm{RKM}}^{\mathrm{train}} - \frac{\lambda}{2}\sum_{i=1}^{N} h_i^T h_i + \frac{\eta}{2}\mathrm{Tr}(W^T W)
\end{aligned}
$$

Restricted Kernel Machines (RKM) - Example LS-SVM (2)

- Based on **property**: $\frac{1}{2\lambda} e^T e \geq e^T h - \frac{\lambda}{2} h^T h, \ \forall e, h$ and $\frac{1}{2\lambda} e^T e = \max_h (e^T h - \frac{\lambda}{2} h^T h)$.

- **Conjugate feature duality:** hidden features h_i are conjugated to the e_i

- Interpretation in terms of **visible and hidden units**

$$
\begin{aligned}
R_{\mathrm{RKM}}^{\mathrm{train}} &= \sum_{i=1}^{N} R_{\mathrm{RKM}}(v_i, h_i) \\
&= -\sum_{i=1}^{N} (x_i^T W h_i + b^T h_i - y_i^T h_i) = \sum_{i=1}^{N} e_i^T h_i
\end{aligned}
$$

with $R_{\mathrm{RKM}}(v, h) = -v^T \tilde{W} h = -(x^T W h + b^T h - y^T h) = e^T h$.

Restricted Kernel Machines (RKM) - Example LS-SVM (3)

- Stationary points of $\underline{J}(h_i, W, b)$ (nonlinear case, feature map $\varphi(\cdot)$)

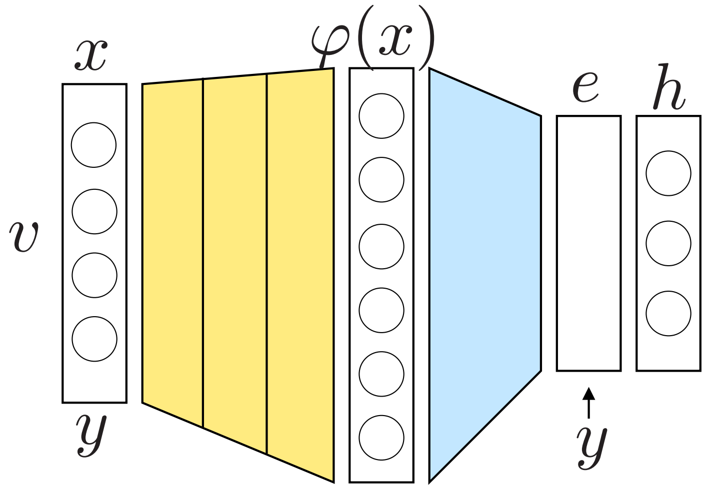
$$\begin{cases} \frac{\partial \underline{J}}{\partial h_i} = 0 & \Rightarrow \quad y_i = W^T \varphi(x_i) + b + \lambda h_i, \ \forall i \\ \frac{\partial \underline{J}}{\partial W} = 0 & \Rightarrow \quad W = \frac{1}{\eta} \sum_i \varphi(x_i) h_i^T \\ \frac{\partial \underline{J}}{\partial b} = 0 & \Rightarrow \quad \sum_i h_i = 0. \end{cases}$$

- Solution in h_i and b with positive definite kernel $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$

$$\left[\begin{array}{c|c} \frac{1}{\eta} K + \lambda I_N & 1_N \\ \hline 1_N^T & 0 \end{array} \right] \left[\begin{array}{c} H^T \\ \hline b^T \end{array} \right] = \left[\begin{array}{c} Y^T \\ \hline 0 \end{array} \right]$$

with $K = [K(x_i, x_j)]$, $H = [h_1 ... h_N]$, $Y = [y_1 ... y_N]$.

Restricted Kernel Machines (RKM) - Example LS-SVM (4)

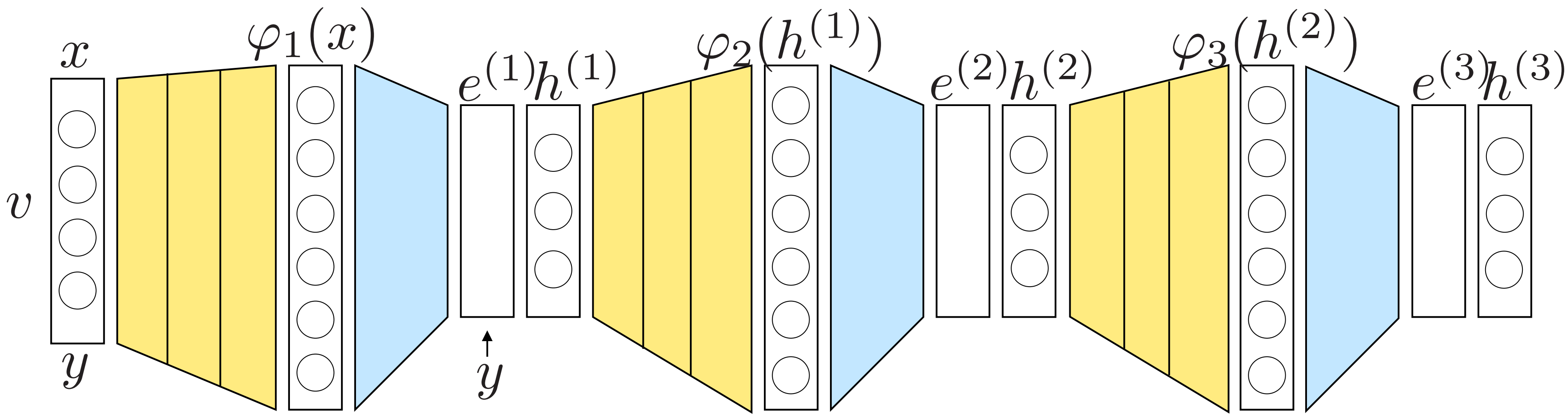


Primal and dual model representations:

$$\mathcal{M} \begin{array}{l} \nearrow \quad (P)_{\mathrm{RKM}}: \quad \hat{y} = W^T \varphi(x) + b \\ \\ \searrow \quad (D)_{\mathrm{RKM}}: \quad \hat{y} = \frac{1}{\eta} \sum_j h_j K(x_j, x) + b. \end{array}$$

[Suykens J.A.K., Internal Report 16-50, ESAT-SISTA, KU Leuven 2016]

Deep RKM



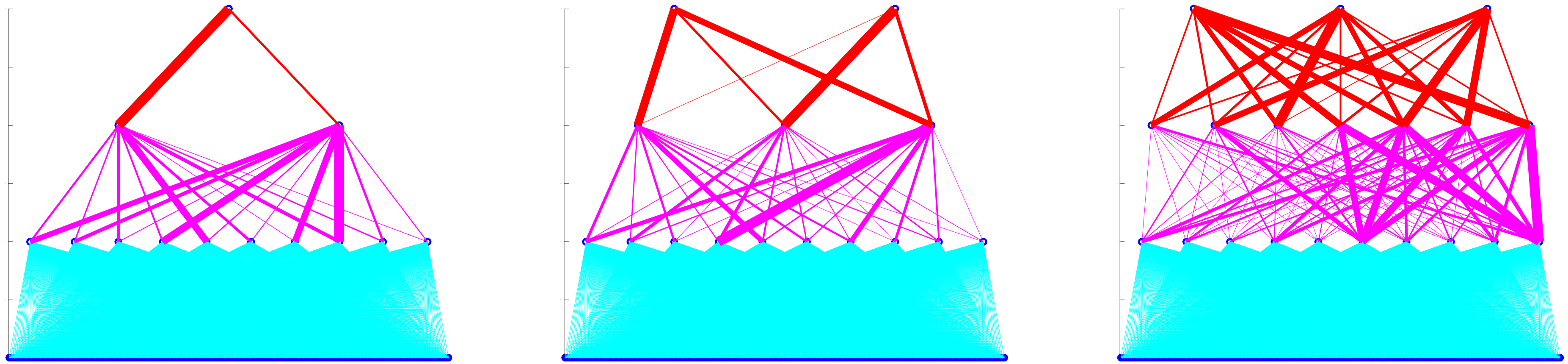
Deep RKM: LSSVM + KPCA + KPCA

Coupling of RKMs by taking sum of the objectives

$$J_{\text{deep}} = \underline{J}_1 + \overline{J}_2 + \overline{J}_3$$

with inner pairings $\sum_{i=1}^{N} e_i^{(1)^T} h_i^{(1)}, \sum_{i=1}^{N} e_i^{(2)^T} h_i^{(2)}, \sum_{i=1}^{N} e_i^{(3)^T} h_i^{(3)}$

Deep RKM - Example USPS data



USPS (10 classes): Deep RKM: LSSVM ($K_{\rm rbf}$) + KPCA ($K_{\rm lin}$) + KPCA ($K_{\rm lin}$)

Training algorithm: forward & backward phases, kernel fusion between levels

$N = 2000$: test error 3.26% (basic) - 3.18% (deep) $\quad$ ($N_{\rm test} = 5000$)

$N = 4000$: test error 2.14% (basic) - 2.12% (deep) $\quad$ ($N_{\rm test} = 5000$)

[Suykens J.A.K., Internal Report 16-50, ESAT-SISTA, KU Leuven 2016]

Conclusions

- **Synergies** parametric and kernel based-modelling
- **Primal and dual** representations
- Sparsity, robustness, networks, big data
- **SVD from LS-SVM**, nonlinear extensions to SVD
- **Beyond Mercer kernels**
- **Deep learning and kernel machines:** Deep RKM

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, S. Salzo, 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