

Kernel-based Modelling for Complex Networks

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/>

NOLTA 2014, Luzern

Overview

- Coupled oscillators, synchronization and clustering
- From spectral clustering to kernel spectral clustering
- Sparse kernel models
- Applications in complex networks
- Towards black-box weather forecasting

Coupled oscillators and clustering

Link between synchronization and spectral clustering

- (generalized) Kuramoto model: N coupled phase oscillators

$$\frac{d\theta_i}{dt} = \omega_i + \sum_j \kappa_{ij} \sin(\theta_j - \theta_i), \quad i = 1, \dots, N$$

Special case: $\omega_i = \omega$, $\kappa_{ij} = \sigma a_{ij}$ with adjacency matrix $[a_{ij}]$

- Linearized dynamics (Laplacian matrix L)

$$\frac{d\theta_i}{dt} = -\sigma \sum_j L_{ij} \theta_j, \quad i = 1, \dots, N$$

- Relationship between topological scales and dynamic time scales.
Modular structures emerge at different time scales.

[Arenas et al., PRL 2006, PR 2008]

Community detection from synchronization

- **Kuramoto model:** $\dot{\theta}_i = \omega + \sigma \sum_j a_{ij} \sin(\theta_j - \theta_i)$
- Follow the evolution of

$$\rho_{ij}(t) = \langle \cos[\theta_i(t) - \theta_j(t)] \rangle$$

averaged over different initial conditions.

- Community detection based on a **binary dynamic connectivity matrix**

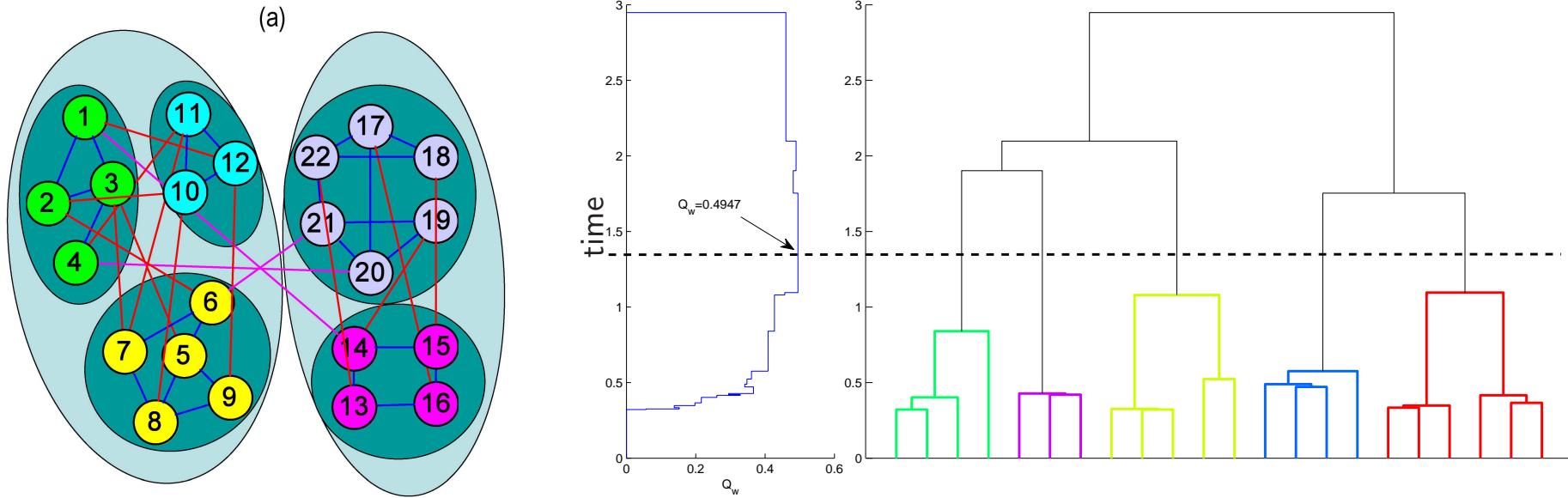
$$[\mathcal{D}_t(T)]_{ij} = 1 \text{ if } \rho_{ij}(t) > T, \text{ zero otherwise}$$

T large enough: one finds set of disconnected clusters

T smaller: inter-community connections become visible

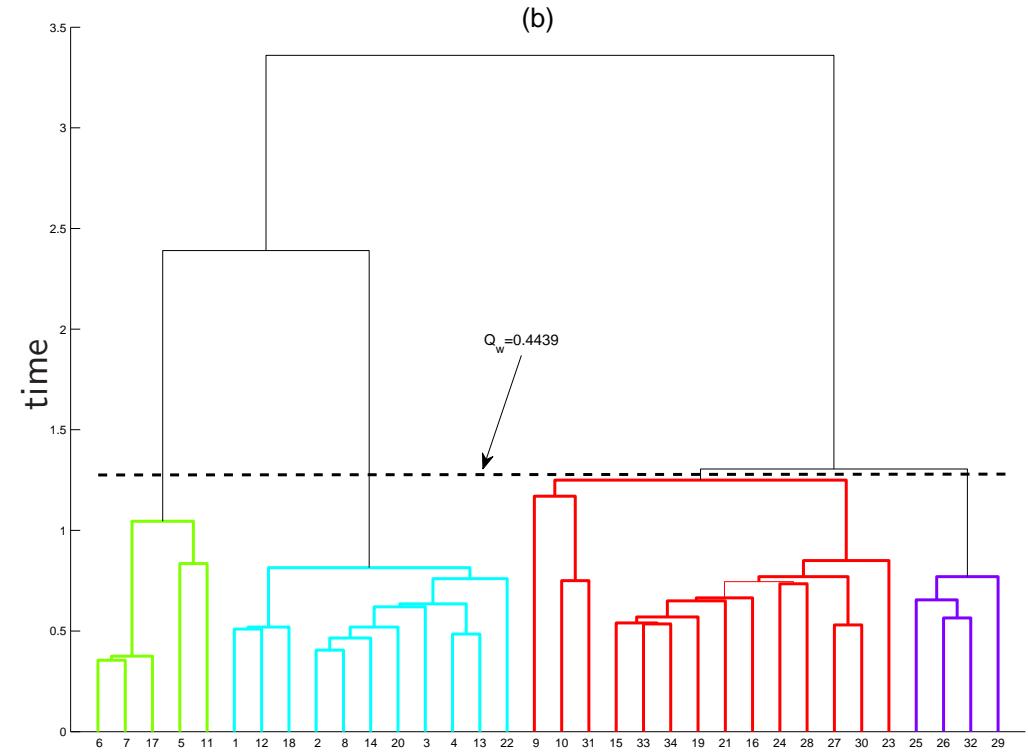
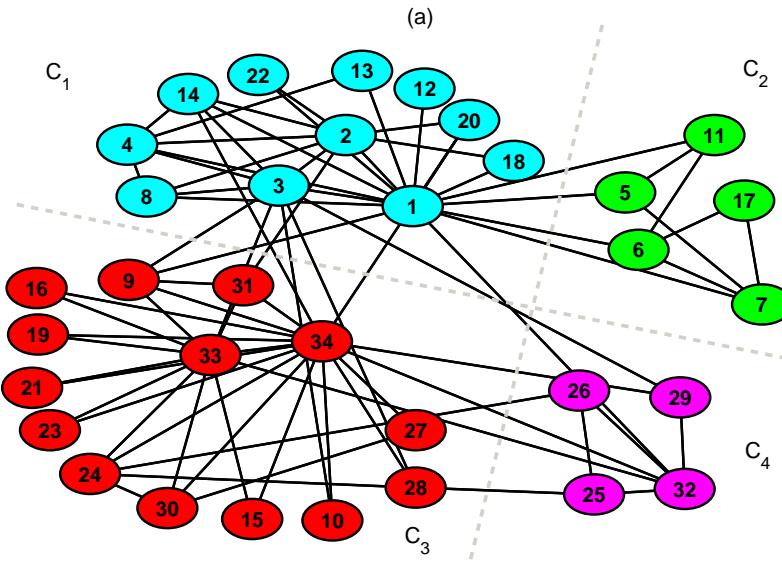
[Arenas et al., PRL 2006, PR 2008]

Finding communities in weighted networks (1)



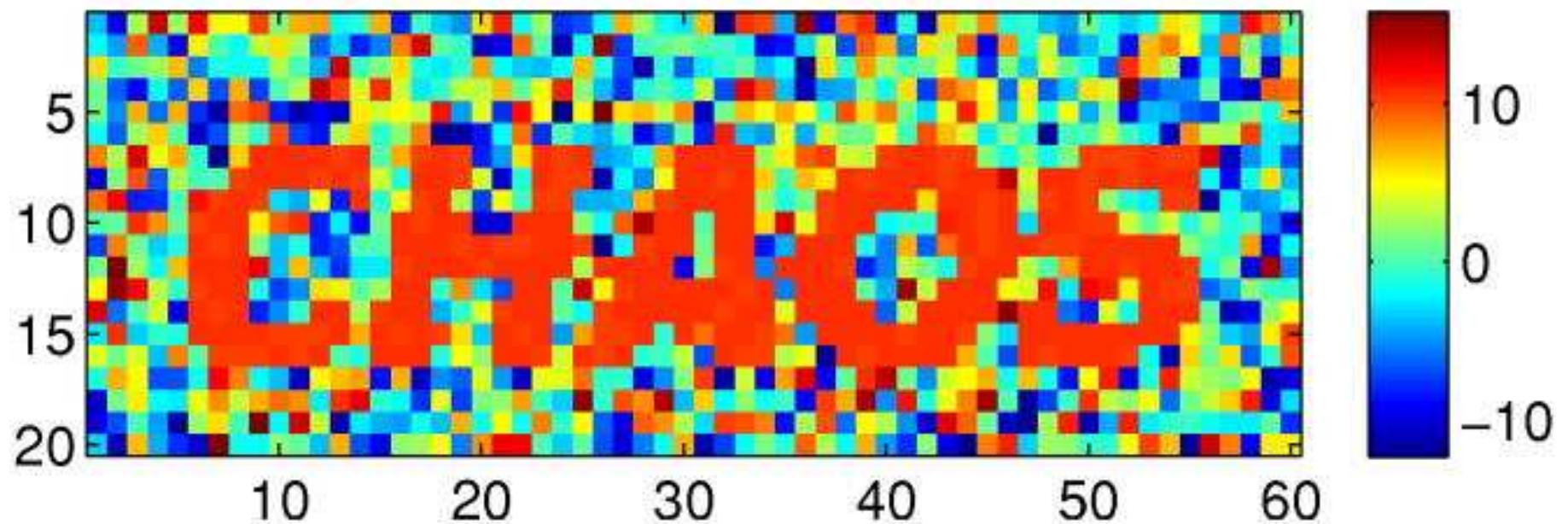
Synthetic example [Lou & Suykens, Chaos 2011]: community detection by considering $[D]_{ij} = t_{ij}$ if $\rho_{ij}(t) > T$ and zero otherwise, where t_{ij} is the time needed for nodes i and j to synchronization in the sense that $\rho_{ij}(t) > T$.

Finding communities in weighted networks (2)



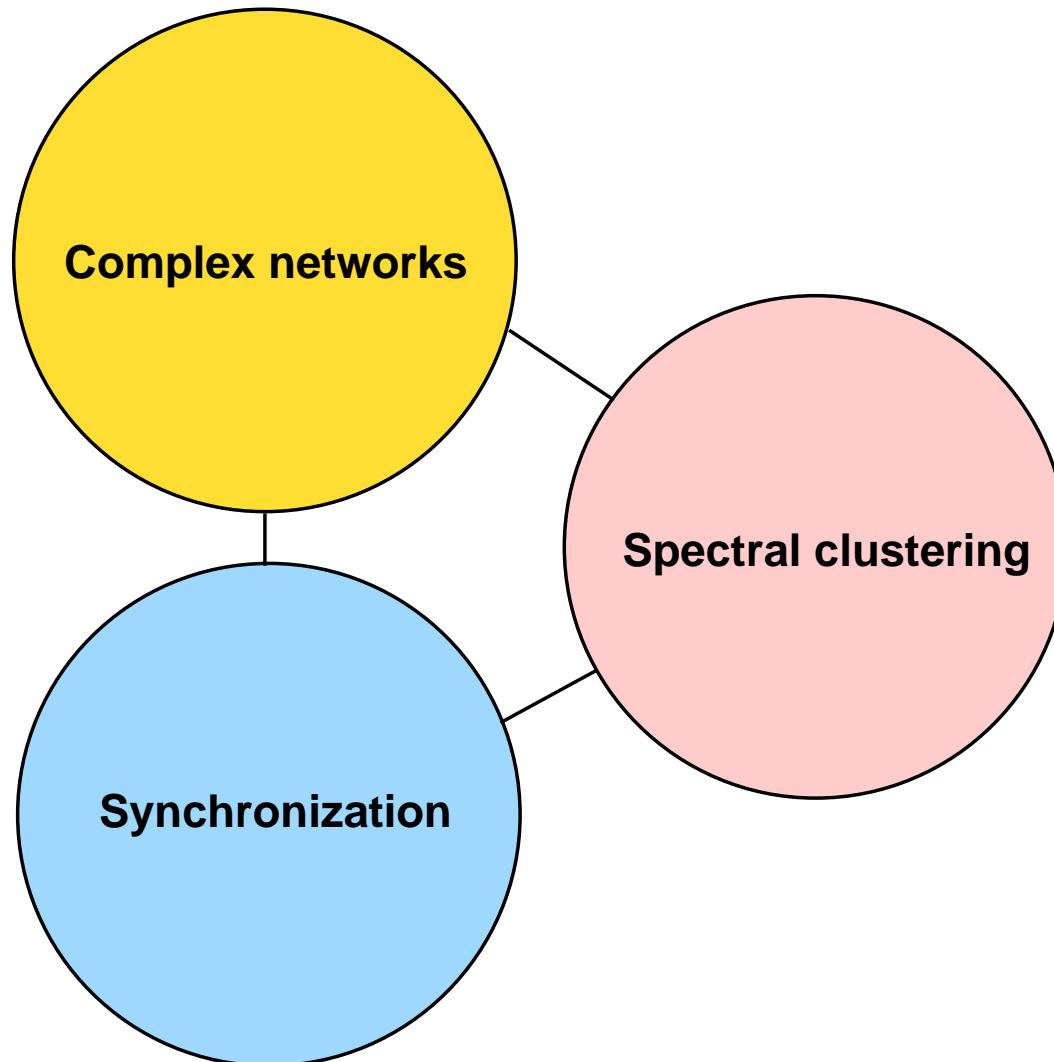
on the Zachary's karate club network [Lou & Suykens, Chaos 2011]

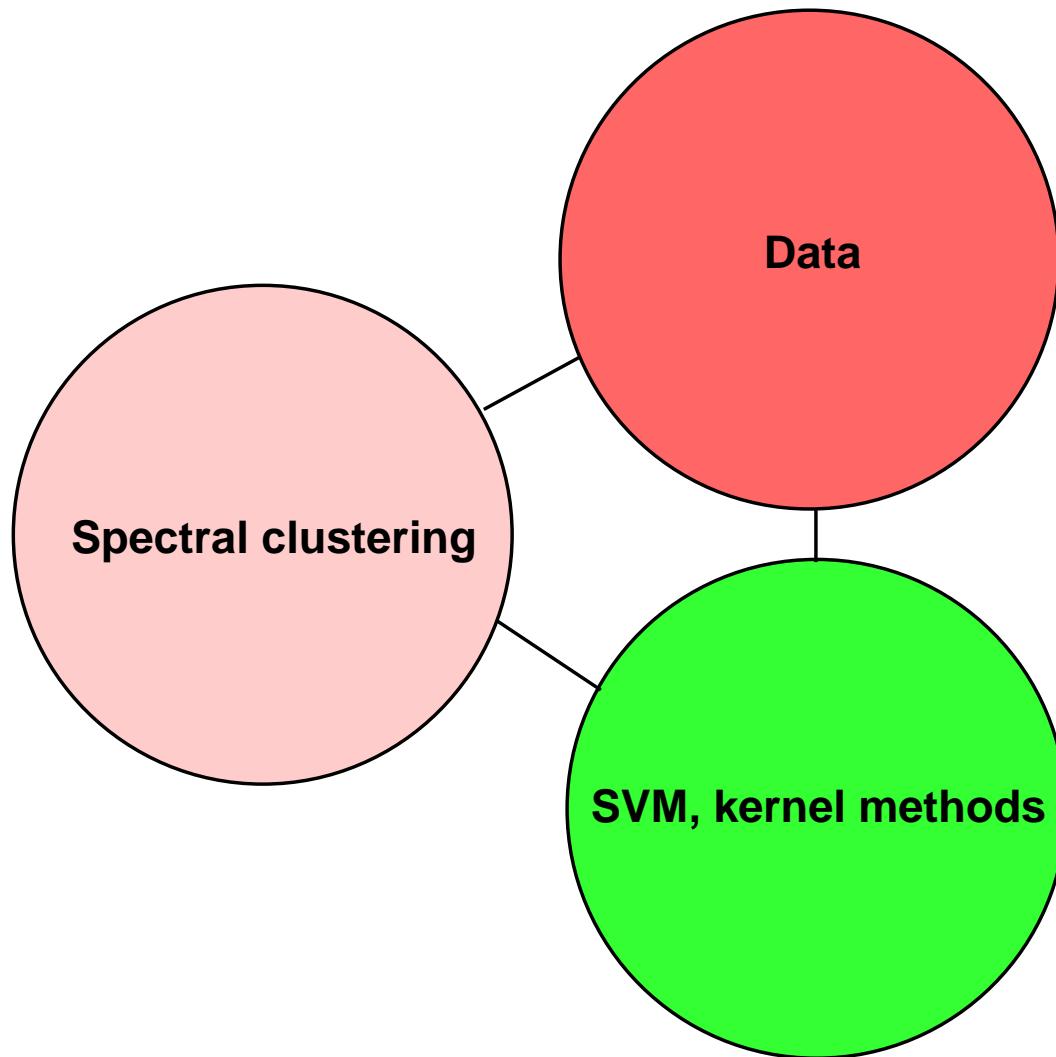
"Programming" clusters into complex networks



- cluster design on a 20×60 lattice of identical Rössler oscillators.
- cluster "CHAOS" obtained from randomly distributed initial conditions.

[Belykh, Osipov, Petrov, Suykens, Vandewalle, Chaos 2008]



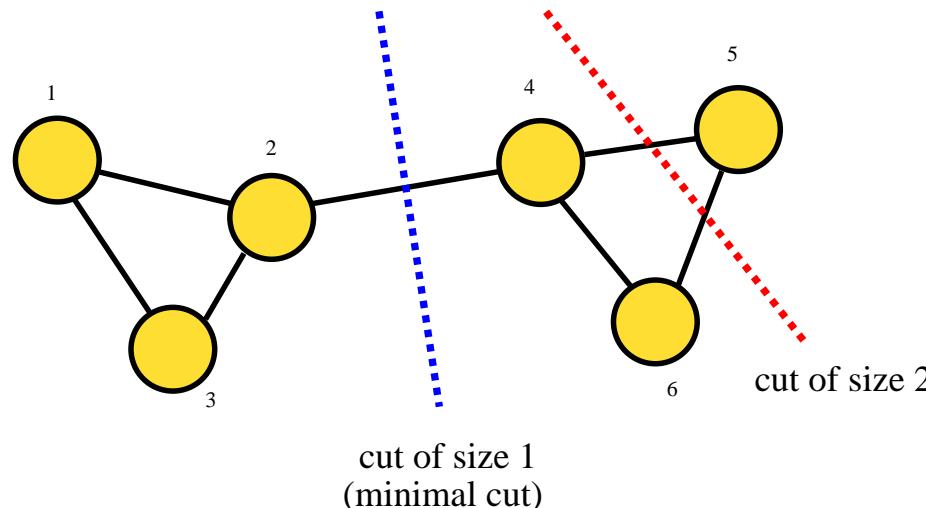


Spectral graph clustering (1)

Minimal cut: given the graph $\mathcal{G} = (V, E)$, find clusters $\mathcal{A}_1, \mathcal{A}_2$

$$\min_{q_i \in \{-1, +1\}} \frac{1}{2} \sum_{i,j} w_{ij} (q_i - q_j)^2$$

with cluster membership indicator q_i ($q_i = 1$ if $i \in \mathcal{A}_1$, $q_i = -1$ if $i \in \mathcal{A}_2$) and $W = [w_{ij}]$ the weighted adjacency matrix.



Spectral graph clustering (2)

- Relaxation to **Min-cut** spectral clustering problem

$$\min_{\tilde{q}^T \tilde{q} = 1} \tilde{q}^T L \tilde{q}$$

with $L = D - W$ the unnormalized graph Laplacian, degree matrix $D = \text{diag}(d_1, \dots, d_N)$, $d_i = \sum_j w_{ij}$, giving

$$L \tilde{q} = \lambda \tilde{q}.$$

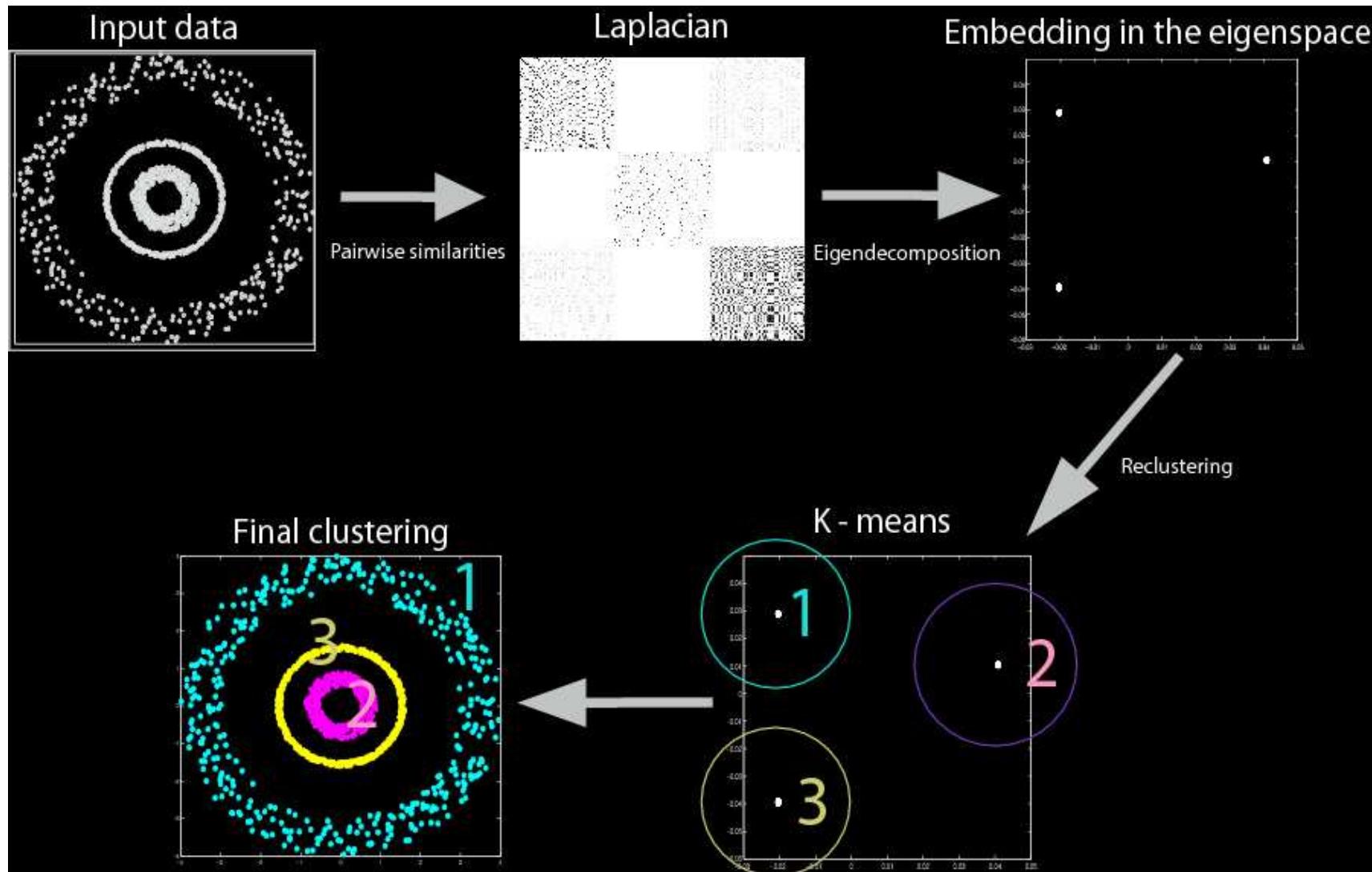
Cluster member indicators: $\hat{q}_i = \text{sign}(\tilde{q}_i - \theta)$ with threshold θ .

- **Normalized cut**

$$L \tilde{q} = \lambda D \tilde{q}$$

[Fiedler, 1973; Shi & Malik, 2000; Ng et al. 2002; Chung, 1997; von Luxburg, 2007]

Spectral clustering + K-means



Kernel-based modelling approach to spectral clustering

Kernel Spectral Clustering (KSC): case of two clusters

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

$$\begin{aligned} \min_{w,b,e} \quad & \frac{1}{2} w^T w - \gamma \frac{1}{2} \sum_{i=1}^N \textcolor{red}{v}_i e_i^2 \\ \text{subject to} \quad & e_i = w^T \varphi(x_i) + b, \quad i = 1, \dots, N \end{aligned}$$

with **positive weights** v_i (will be related to inverse degree matrix), $V = \text{diag}\{v_i\}$ and $\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^h$ the feature map.

- **Dual:**

$$VM_V\Omega\alpha = \lambda\alpha$$

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

[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:

$$\left\{ \begin{array}{l} \frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_i \alpha_i \varphi(x_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_i \alpha_i = 0 \\ \frac{\partial \mathcal{L}}{\partial e_i} = 0 \Rightarrow \alpha_i = \gamma v_i e_i, \quad i = 1, \dots, N \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \Rightarrow e_i = w^T \varphi(x_i) + b, \quad i = 1, \dots, N \end{array} \right.$$

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

Choice of weights v_i

- Take $V = D^{-1}$ where $D = \text{diag}\{d_1, \dots, d_N\}$ and $d_i = \sum_{j=1}^N \Omega_{ij}$
- This gives the **generalized eigenvalue problem**:

$$M_D \Omega \alpha = \lambda D \alpha$$

$$\text{with } M_D = I_N - \frac{1}{1_N^T D^{-1} 1_N} 1_N 1_N^T D^{-1}$$

This is a modified version of random walks spectral clustering.

- Corresponds to weighted kernel PCA.

Kernel spectral clustering: more clusters

- Case of k clusters: additional sets of constraints

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

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

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

[Alzate & Suykens, IEEE-PAMI, 2010]

Primal and dual model representations

k clusters

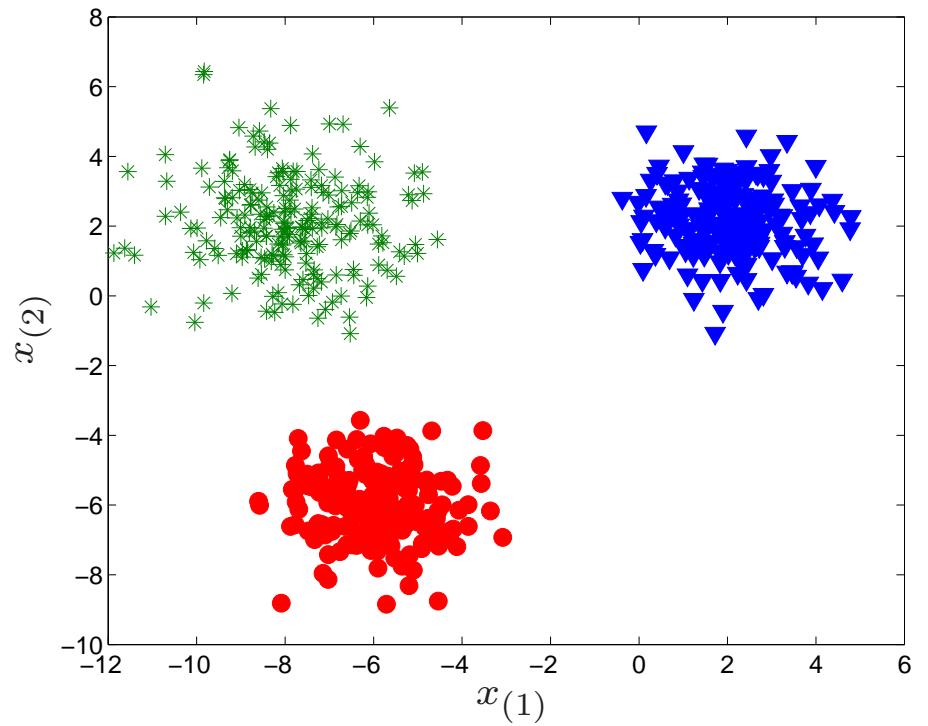
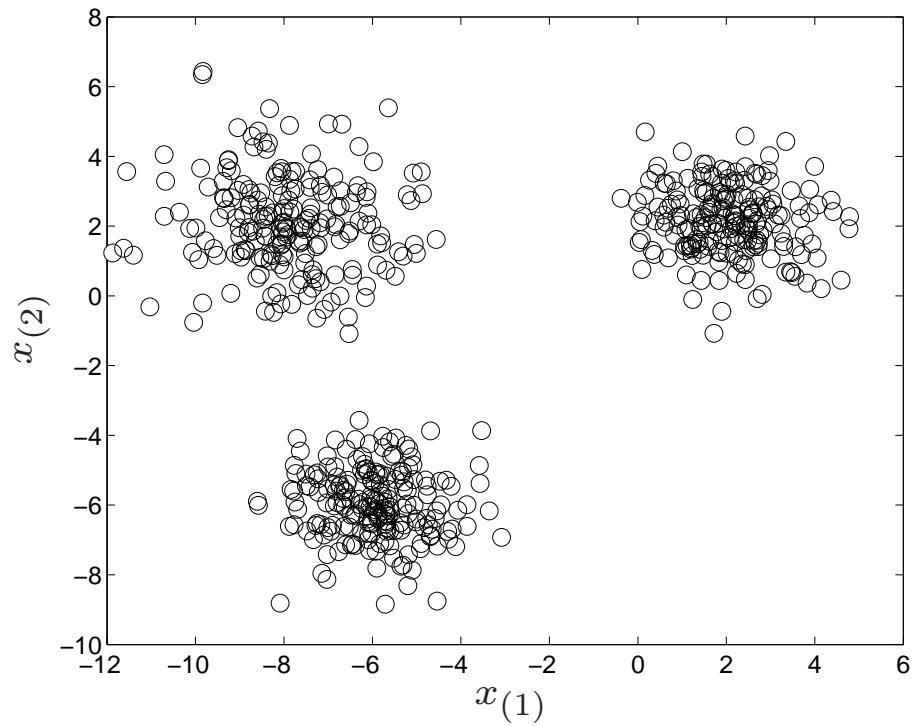
$k - 1$ sets of constraints (index $l = 1, \dots, k - 1$)

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

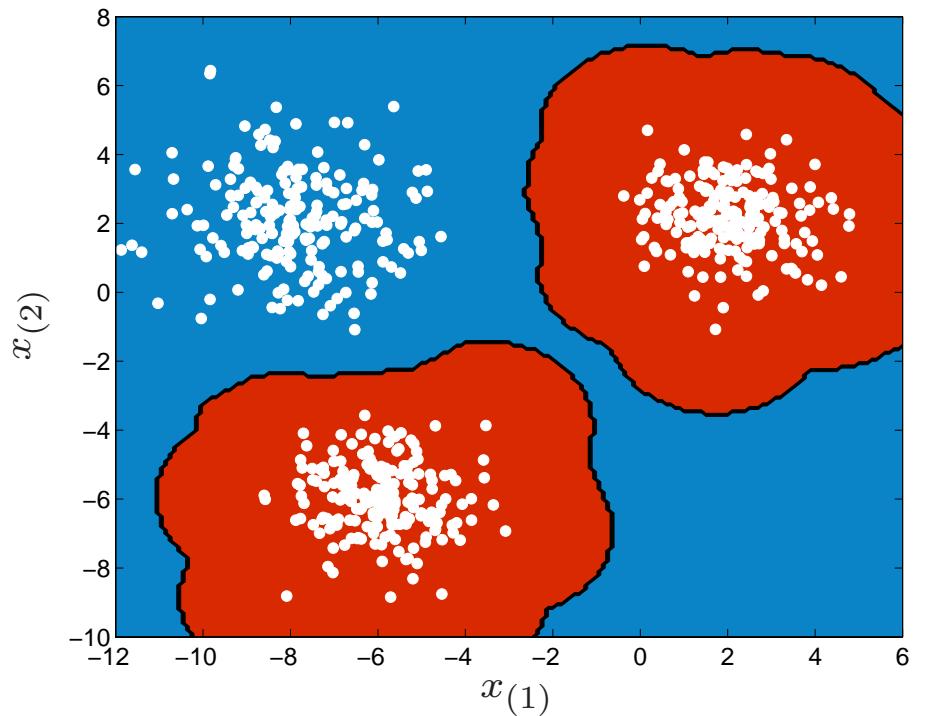
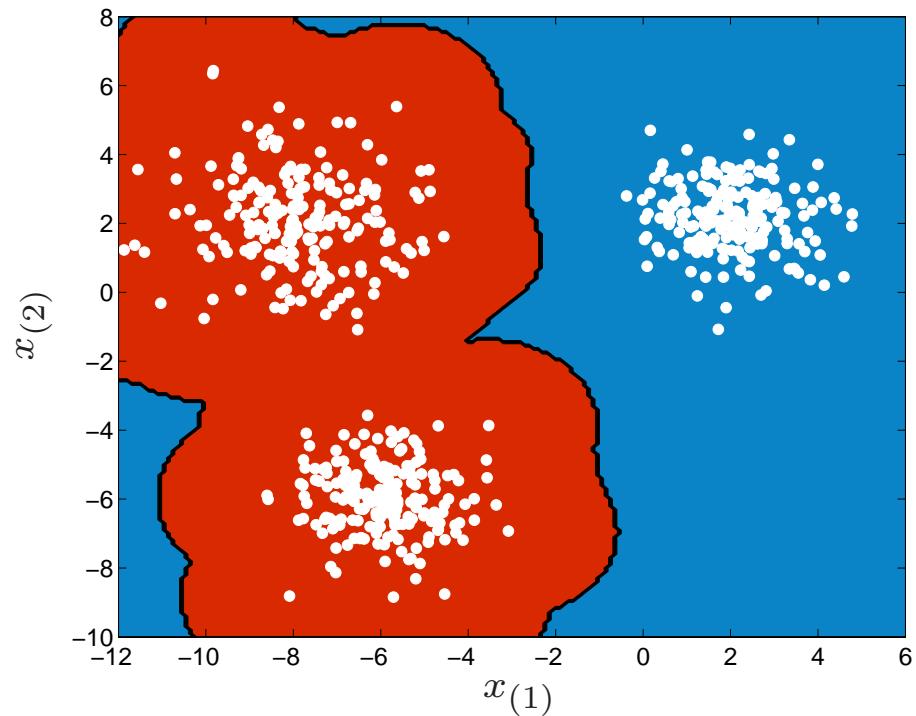
Kernel-based approach: advantages

- underlying model representations (primal and dual)
- **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**

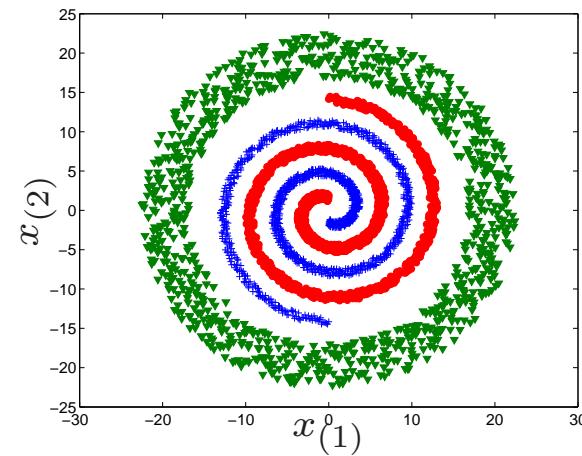
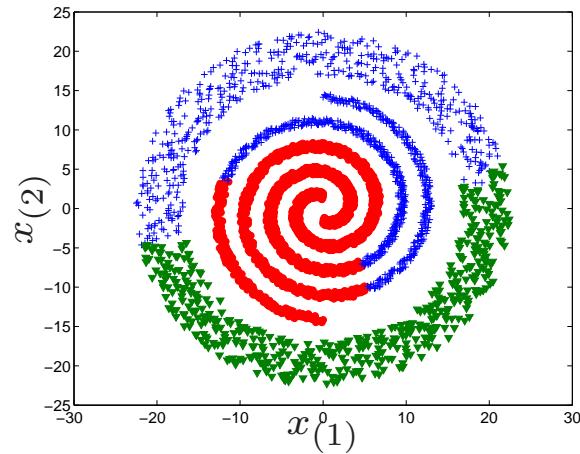
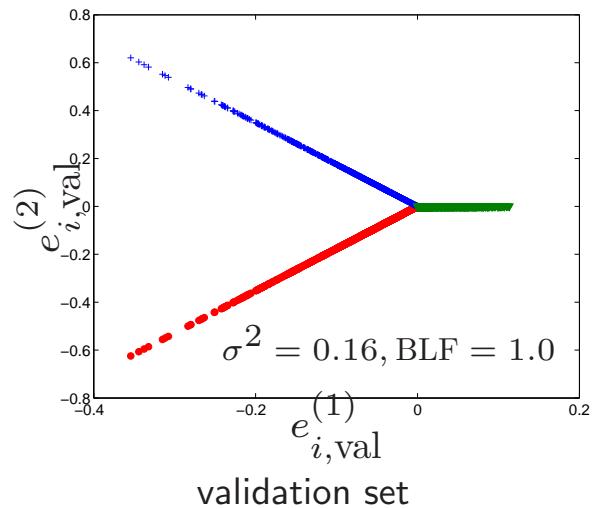
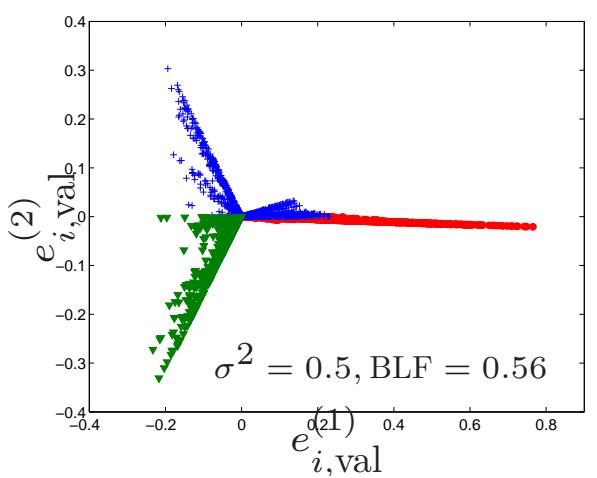
Out-of-sample extension and coding



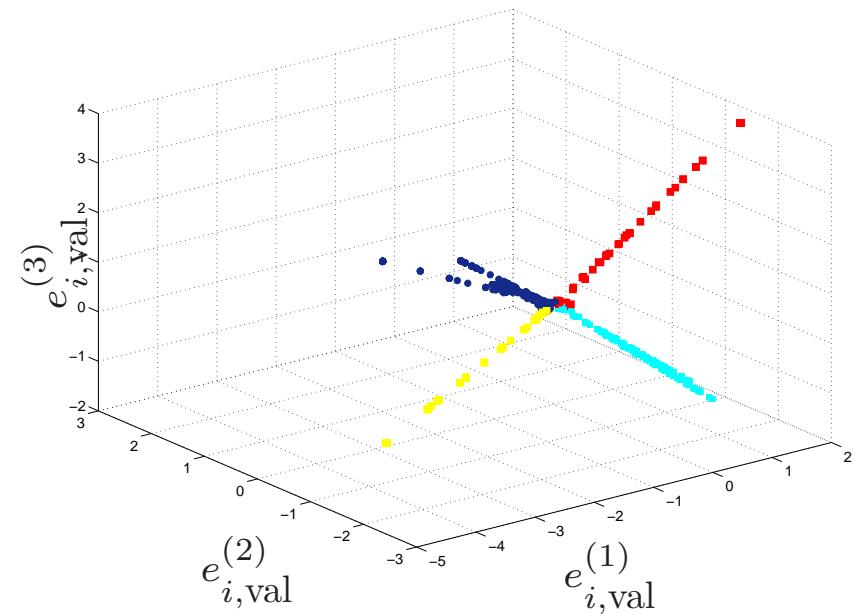
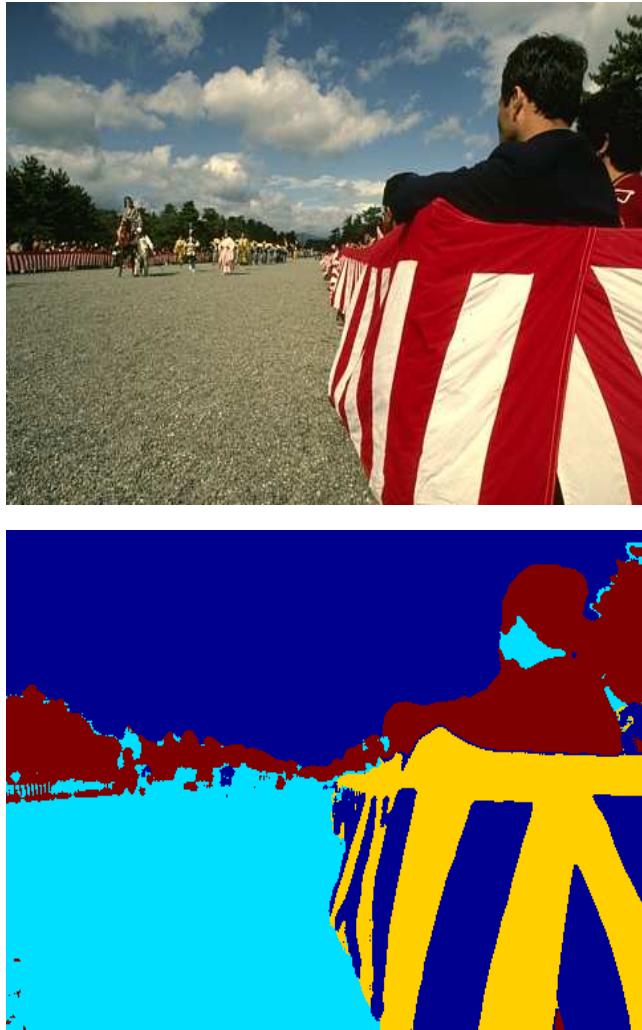
Out-of-sample extension and coding



Model selection: toy problem



Example: image segmentation

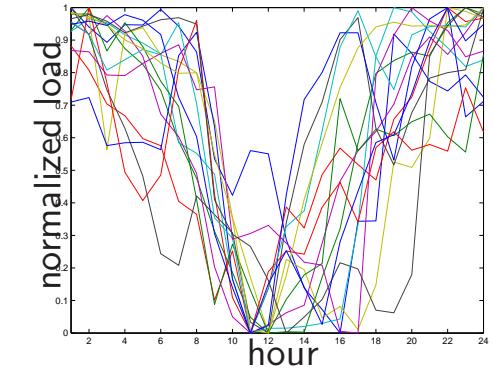
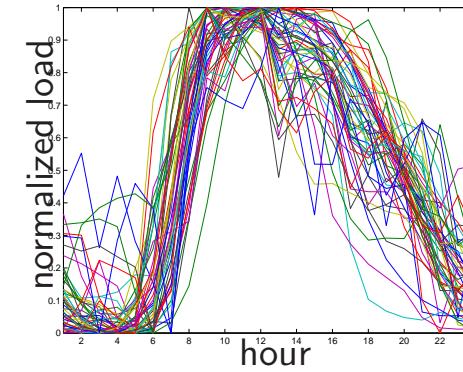
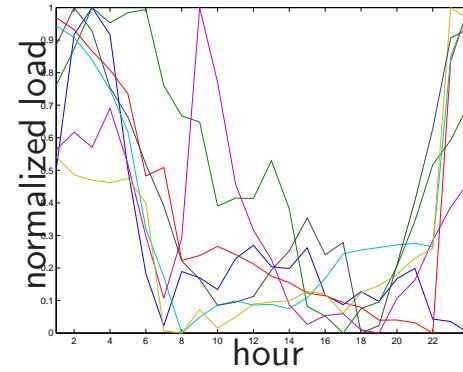
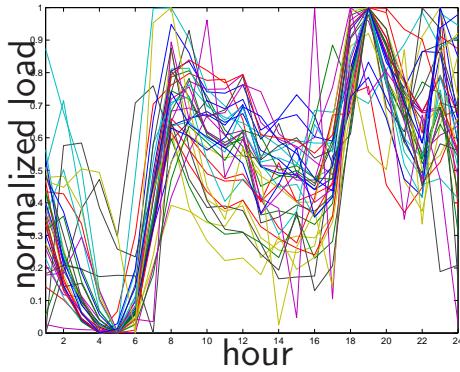
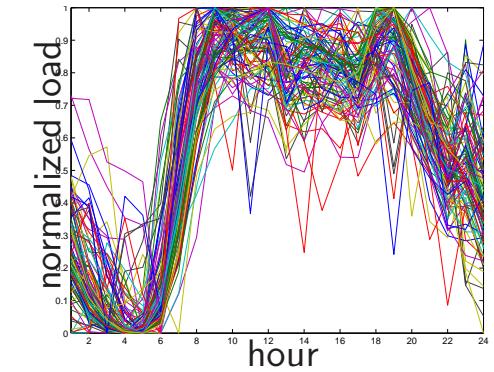
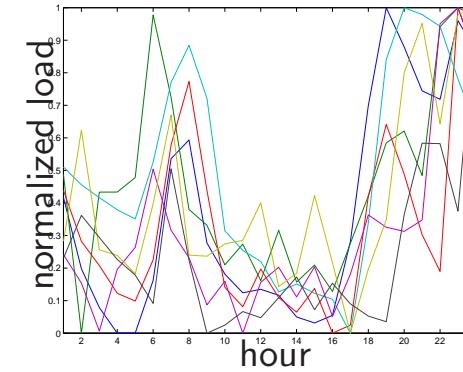
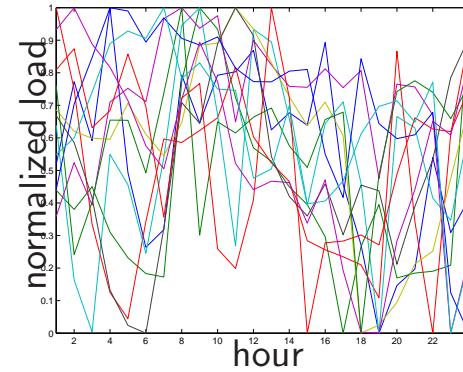
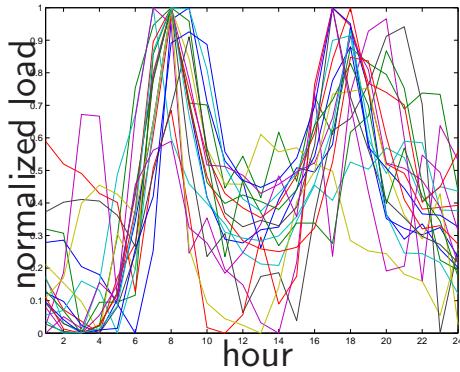


Example: power grid - identifying customer profiles (1)

Power load: 245 substations, hourly (5 years)

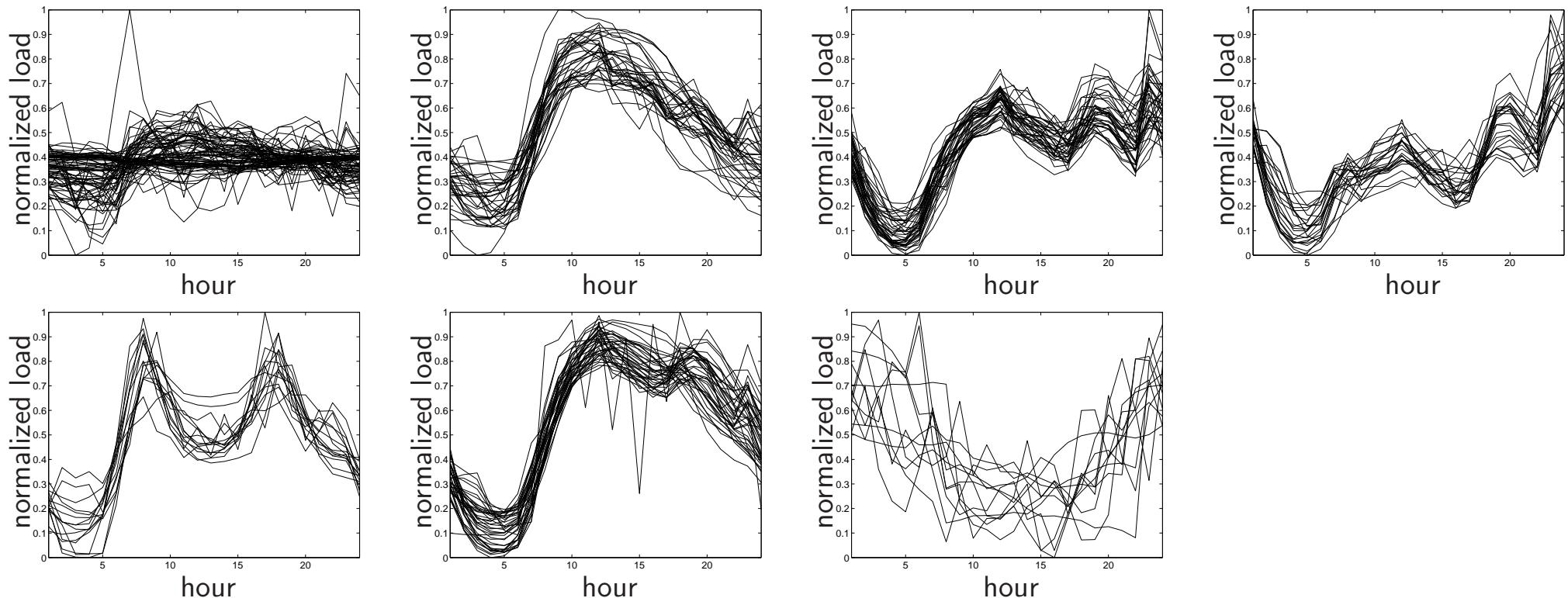
Periodic AR modelling: dim reduction $43.824 \rightarrow 24$

k-means applied after dimensionality reduction



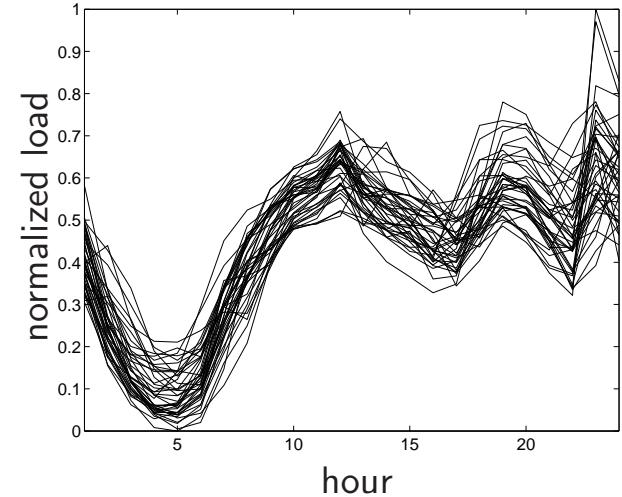
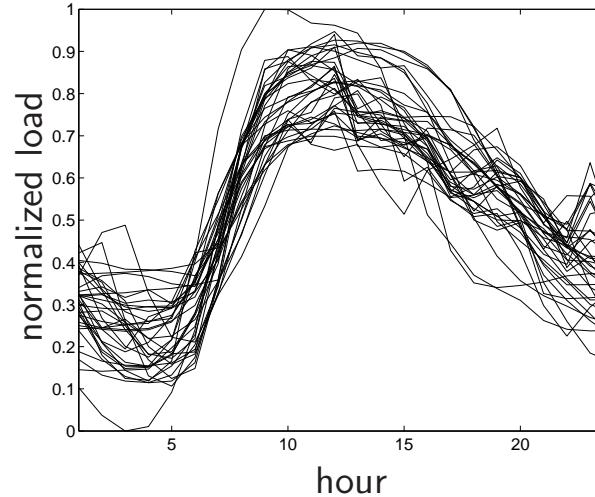
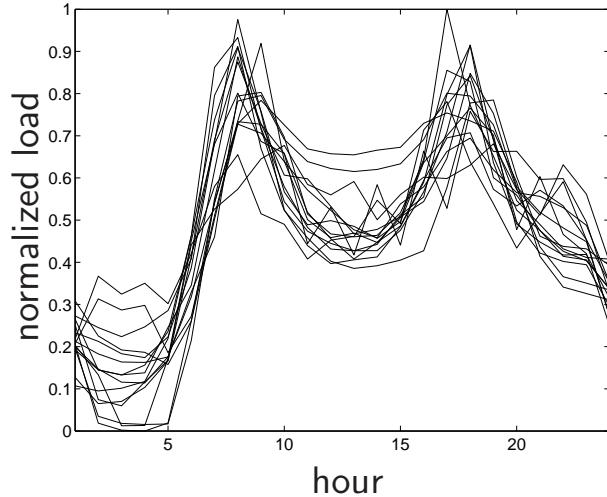
Example: power grid - identifying customer profiles (2)

Application of kernel spectral clustering, directly in high dim $d = 43.824$
Model selection on kernel parameter and number of clusters



[Alzate, Espinoza, De Moor, Suykens, 2009]

Example: power grid - identifying customer profiles (3)



Electricity load: 245 substations in Belgian grid (1/2 train, 1/2 validation)

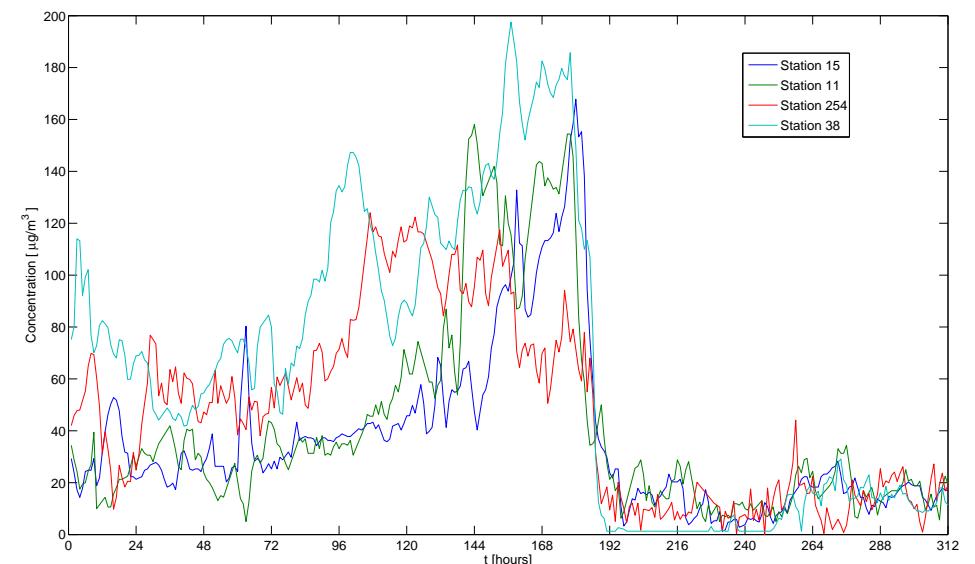
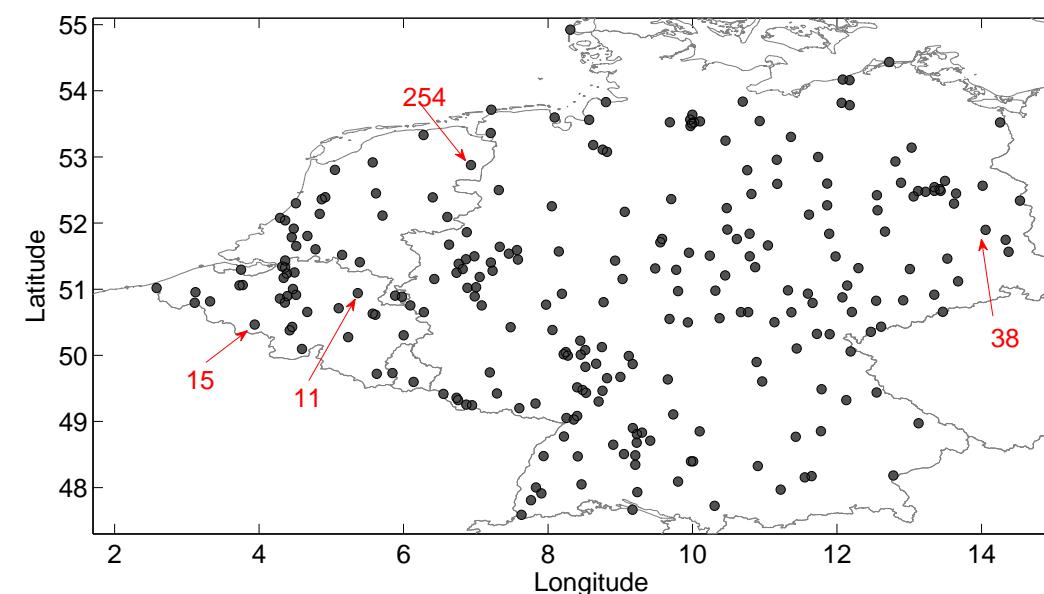
$x_i \in \mathbb{R}^{43,824}$: spectral clustering on high dimensional data (5 years)

3 of 7 detected clusters:

- 1: *Residential profile*: morning and evening peaks
- 2: *Business profile*: peaked around noon
- 3: *Industrial profile*: increasing morning, oscillating afternoon and evening

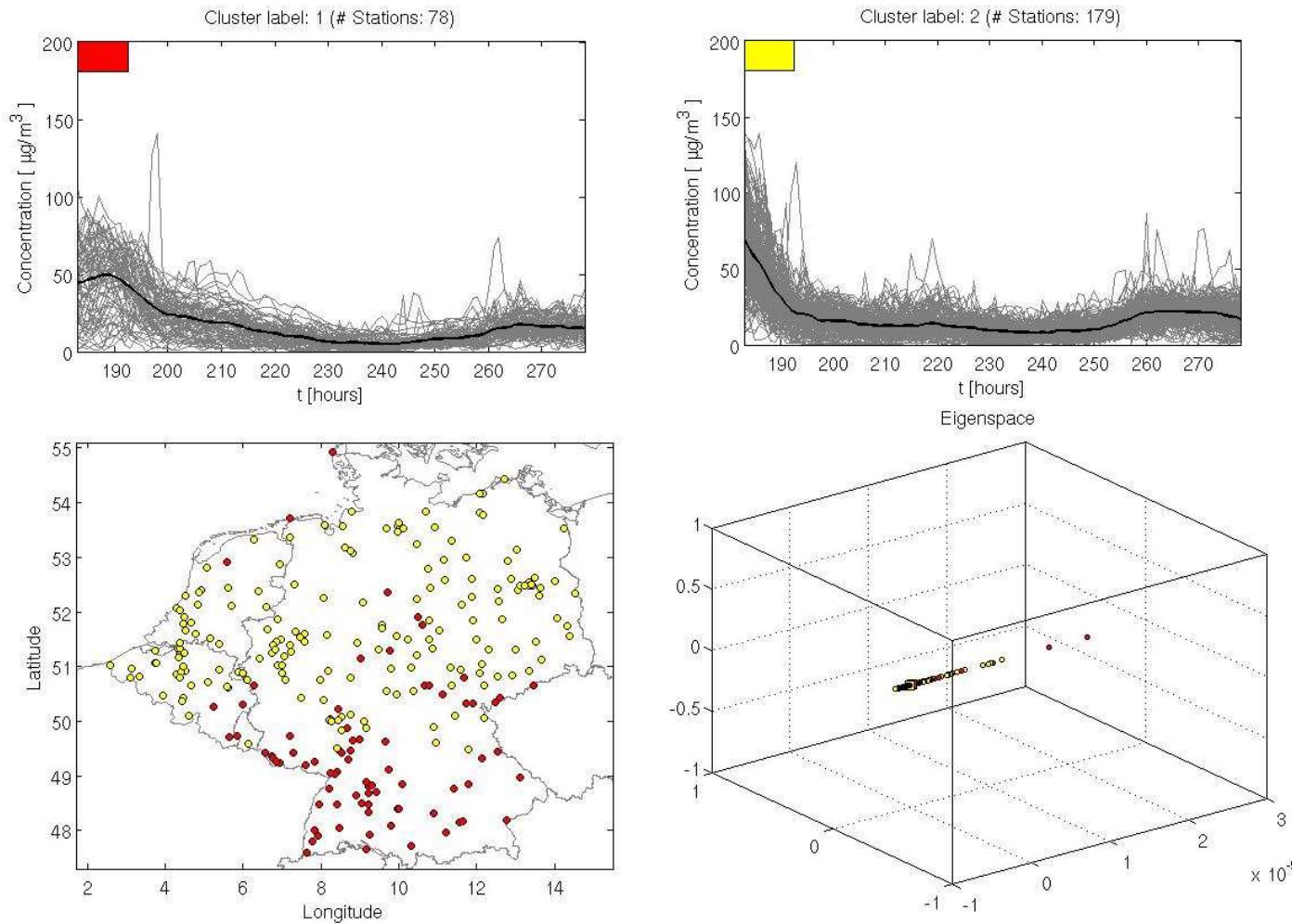
Example: dynamic clustering of PM10 concentrations (1)

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]

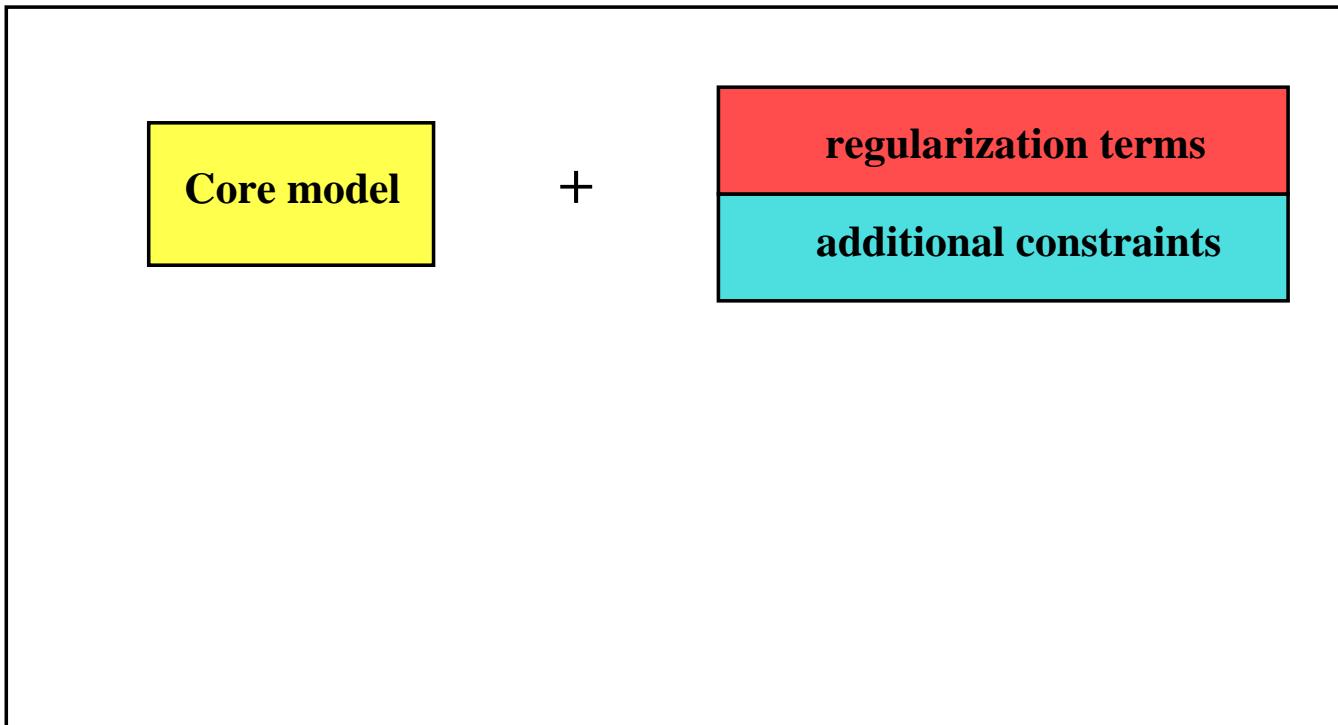
Example: dynamic clustering of PM10 concentrations (2)



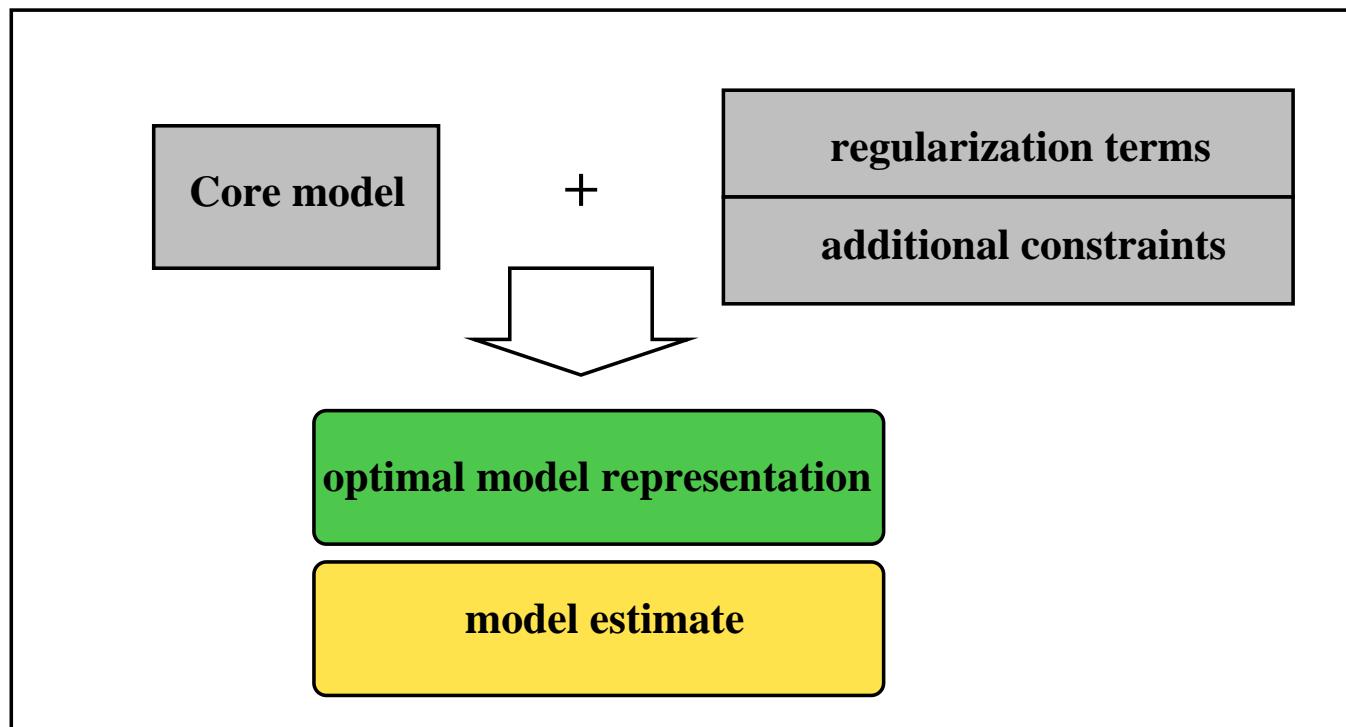
video - [Langone, Agudelo, De Moor, Suykens, Neurocomputing, 2014]

Knowledge incorporation by adding constraints and regularization

Core models + constraints



Core models + constraints



Kernel spectral clustering: adding prior knowledge

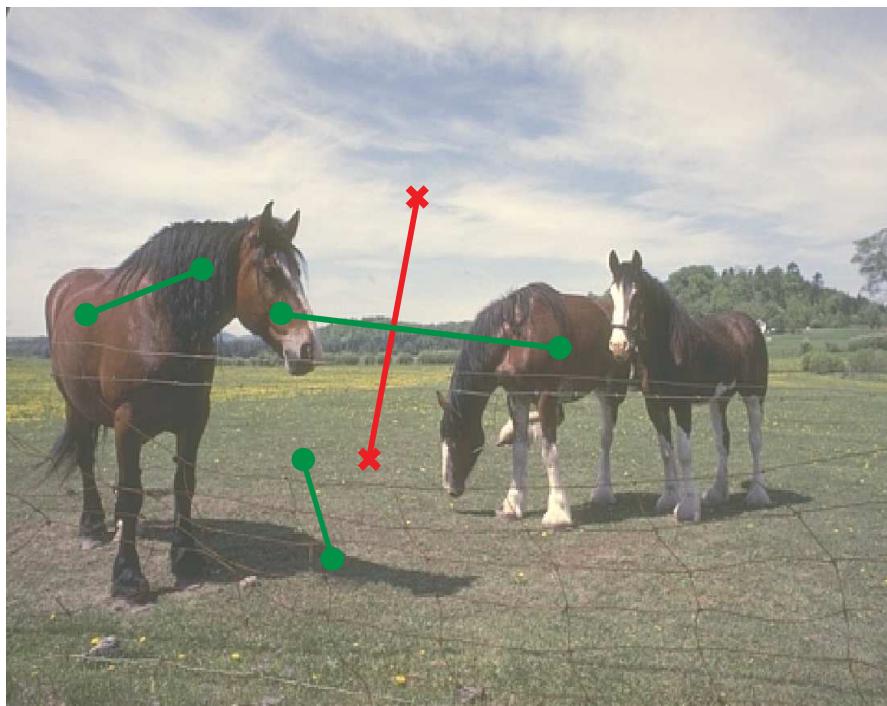
- Pair of points x_\dagger, x_\ddagger : $c = 1$ must-link, $c = -1$ cannot-link
- Primal problem [Alzate & Suykens, IJCNN 2009]

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

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

Adding prior knowledge

original image

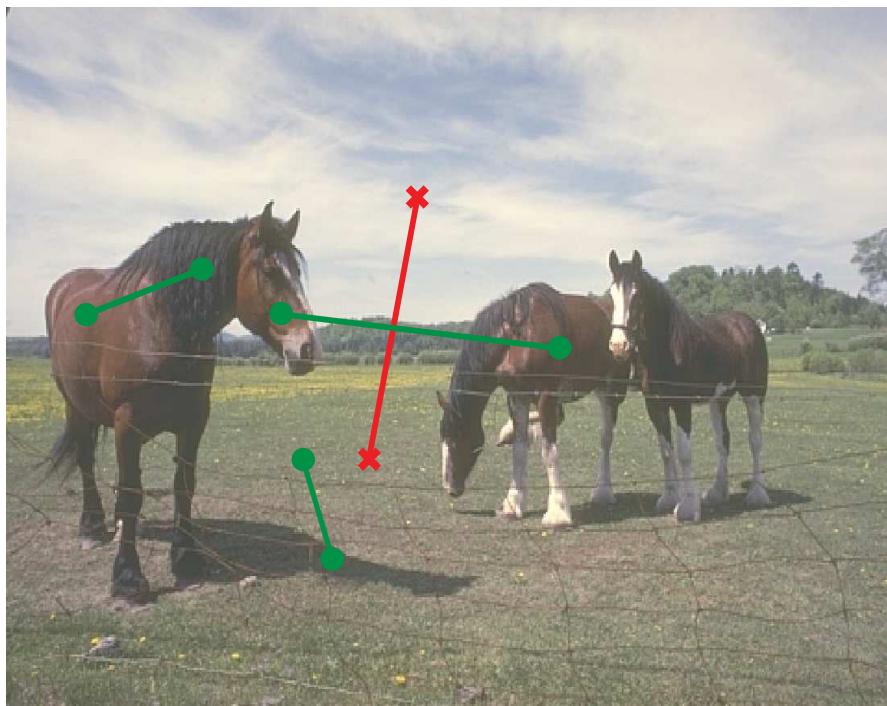


without constraints

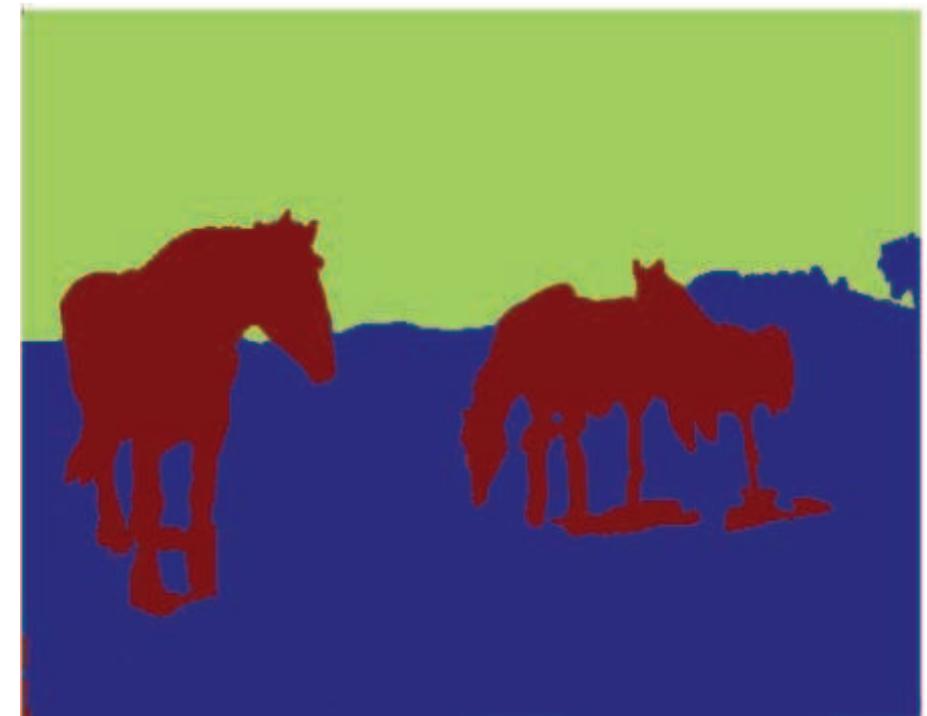


Adding prior knowledge

original image



with constraints



Semi-supervised learning using KSC (1)

- N unlabeled data, but additional labels on $M - N$ data
 $\mathcal{X} = \{x_1, \dots, x_N, \textcolor{red}{x_{N+1}}, \dots, x_M\}$
- Binary classification by using a binary spectral clustering core model [Alzate & Suykens, WCCI 2012; Mehrkanoon et al., 2014]:

$$\min_{w, e, b} \quad \frac{1}{2} w^T w - \gamma \frac{1}{2} e^T D^{-1} e + \rho \frac{1}{2} \sum_{m=N+1}^M (e_m - y_m)^2$$

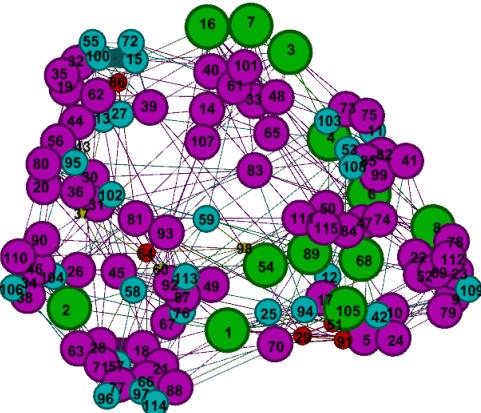
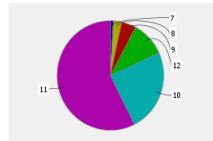
subject to $e_i = w^T \varphi(x_i) + b, \quad i = 1, \dots, M$

Dual solution is characterized by a linear system.

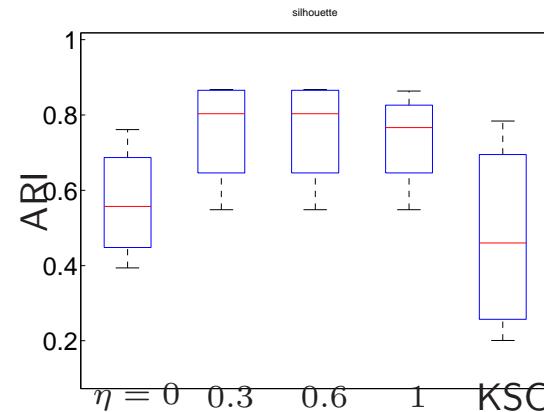
- Other approaches in semi-supervised learning and manifold learning, e.g. [Belkin et al., 2006]

Semi-supervised learning (2)

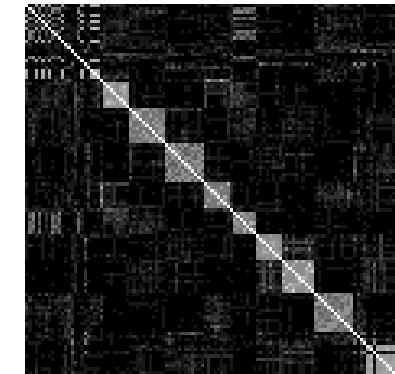
American college football network



ARI



community structure



network: 115 nodes (teams), 616 edges (games), 12 communities

semi-supervised classification: labeled (40%) and unlabeled nodes

prior knowledge incorporation improves test performance (on complete network)

[Mehrkanon, Alzate, Mall, Langone, Suykens, IEEE-TNNLS, in press]

Evolving networks - temporal smoothness

- Binary clustering case: **adding a memory effect**

$$\begin{aligned} \min_{w,e,b} \quad & \frac{1}{2} w^T w - \gamma \frac{1}{2} e^T D^{-1} e - \nu w^T w_{\text{old}} \\ \text{subject to} \quad & e_i = w^T \varphi(x_i) + b, \quad i = 1, \dots, N \end{aligned}$$

with w_{old} the previous result in time.

- Aims at including **temporal smoothness**
- Smoothed modularity criterion

[Langone, Alzate, Suykens, Physica A, 2013]

Sparse kernel models within the primal-dual setting

Least Squares Support Vector Machines: “core models”

- Regression

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

- Classification

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

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

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

- Kernel canonical correlation analysis/partial least squares

$$\min_{w,v,b,d,e,r} w^T w + v^T v + \nu \sum_i (e_i - r_i)^2 \quad \text{s.t.} \quad \begin{cases} e_i &= w^T \varphi_1(x_i) + b \\ r_i &= v^T \varphi_2(y_i) + d \end{cases}$$

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

Probability and quantum mechanics

- **Kernel pmf estimation**

- *Primal:*

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

- *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

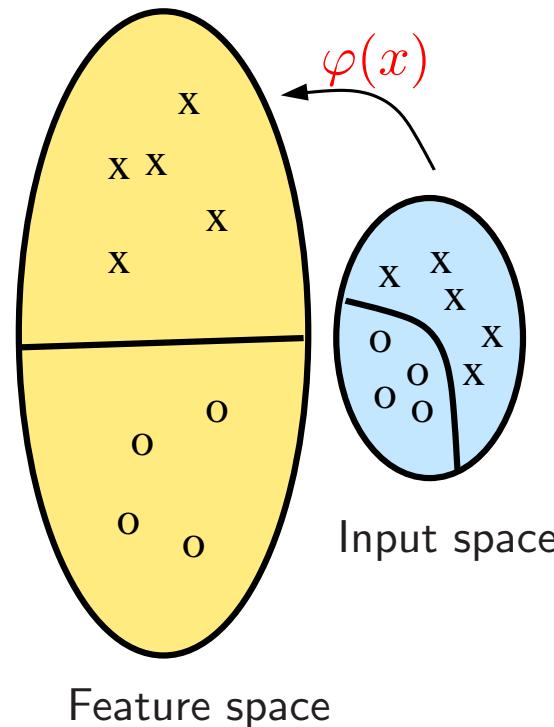
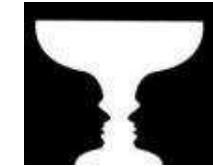
- *Primal:*

$$\min_{|w\rangle, p_i} \frac{1}{2} \langle w | w \rangle \text{ subject to } p_i = \text{Re}(\langle w | M_i \psi \rangle), i = 1, \dots, N \text{ and } \sum_{i=1}^N p_i = 1$$

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

[Suykens, Physical Review A, 2013]

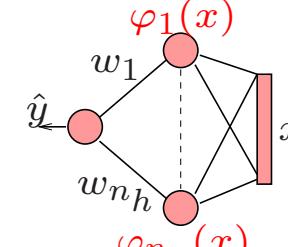
SVMs: living in two worlds ...



Primal space

Parametric

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

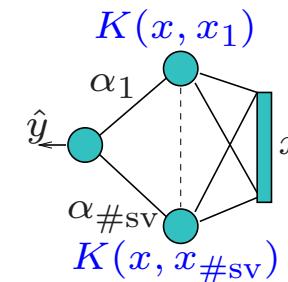


$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \text{ ("Kernel trick")}$$

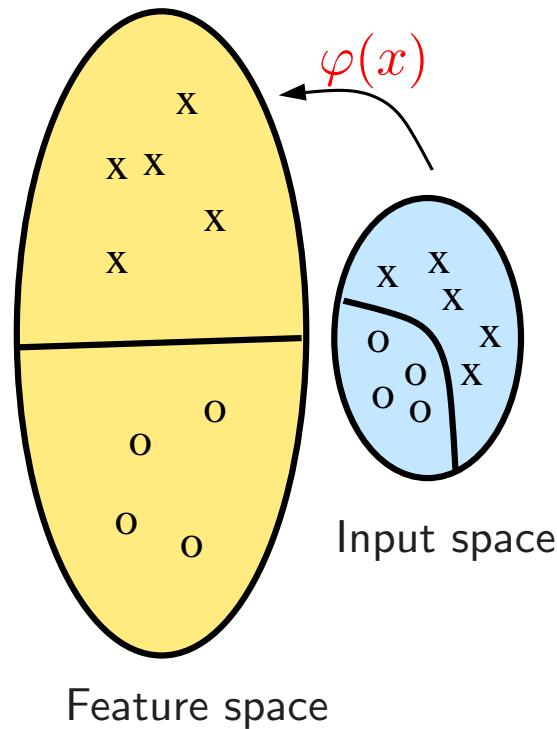
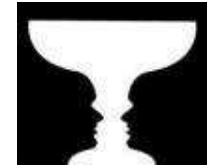
Dual space

Nonparametric

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



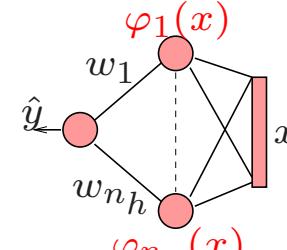
SVMs: living in two worlds ...



Primal space

Parametric

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



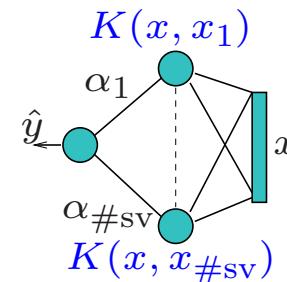
Parametric

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \text{ ("Kernel trick")}$$

Dual space

Nonparametric

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



Non-parametric

Linear model: solving in primal or dual?

inputs $x \in \mathbb{R}^d$, output $y \in \mathbb{R}$

training set $\{(x_i, y_i)\}_{i=1}^N$

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

↗
Model

Linear model: solving in primal or dual?

inputs $x \in \mathbb{R}^d$, output $y \in \mathbb{R}$

training set $\{(x_i, y_i)\}_{i=1}^N$

Model

$$(P) : \hat{y} = \mathbf{w}^T x + b, \quad w \in \mathbb{R}^d$$
$$(D) : \hat{y} = \sum_i \alpha_i x_i^T x + b, \quad \alpha \in \mathbb{R}^N$$

Linear model: solving in primal or dual?

few inputs, many data points: (e.g. $20 \times 1.000.000$)

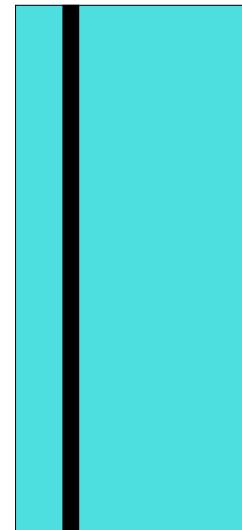


primal: $w \in \mathbb{R}^{20}$

dual: $\alpha \in \mathbb{R}^{1.000.000}$ (kernel matrix: $1.000.000 \times 1.000.000$)

Linear model: solving in primal or dual?

many inputs, few data points: (e.g. 10.000×50)



primal: $w \in \mathbb{R}^{10.000}$

dual: $\alpha \in \mathbb{R}^{50}$ (kernel matrix: 50×50)

Feature map and kernel

From linear to nonlinear model:

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

Kernel trick (Mercer theorem):

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

Feature map $\varphi(x) = [\varphi_1(x); \varphi_2(x); \dots; \varphi_{n_h}(x)]$

Kernel function $K(x, z)$ (e.g. linear, polynomial, RBF, ...)

(the use of a feature map in connection to a positive definite kernel became popular since SVMs [Cortes & Vapnik, 1995])

Fixed-size method

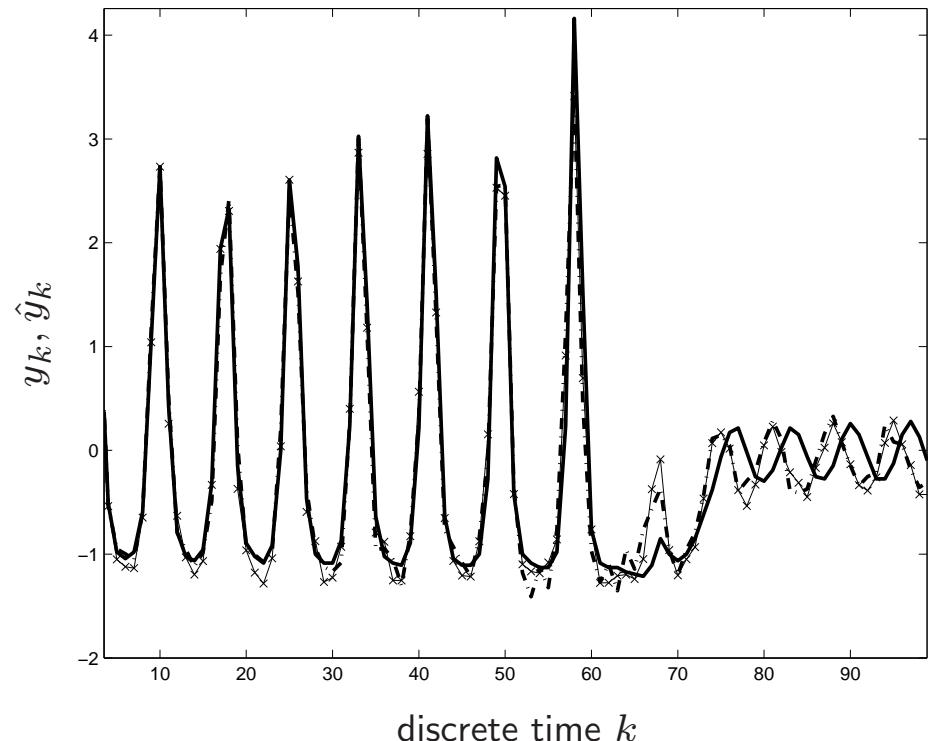
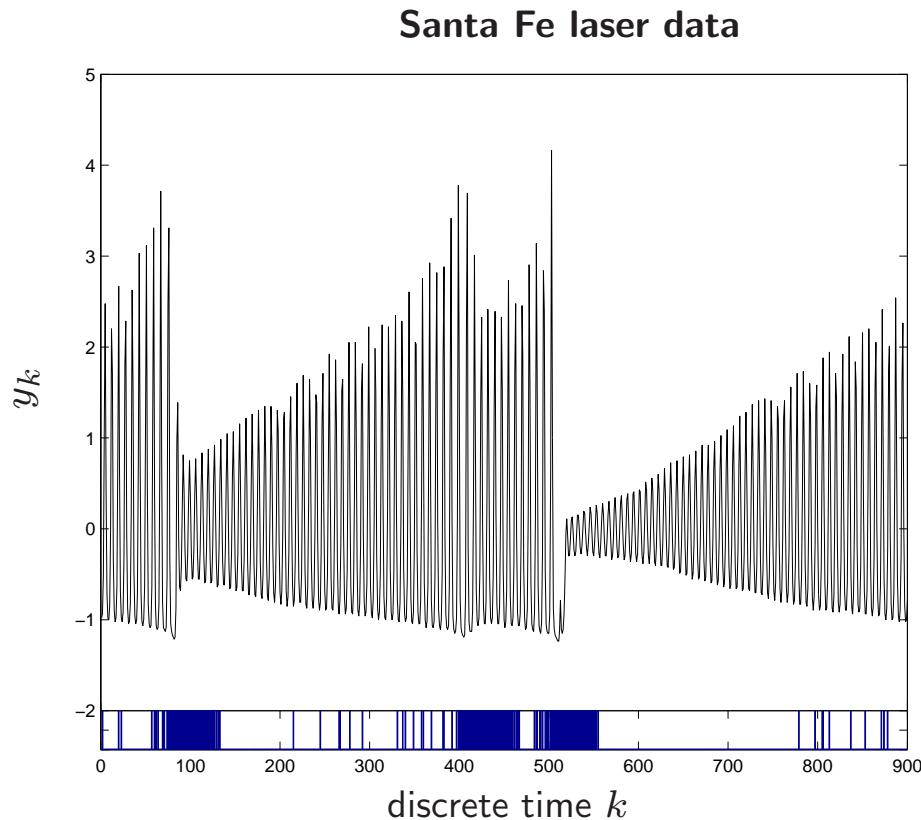
- Find finite dimensional approximation to feature map $\tilde{\varphi}(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}^M$ based on the eigenvalue decomposition of the kernel matrix (on a **subset** of size $M \ll N$).
- Based on [Williams & Seeger, 2001]: relates KPCA to a **Nyström approximation** of the integral equation

$$\int K(z, x)\phi_i(x)dP_X = \lambda_i\phi_i(z)$$

- **Fixed-size method** [Suykens et al., 2002; De Brabanter et al., 2009]:
 - selects subset such that it represents the data distribution P_X
 - optimizes quadratic Renyi entropy criterion (instead of random subset)
 - estimate in **primal** by ridge regression (**sparse** representation):

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

Fixed-size method: example time-series prediction

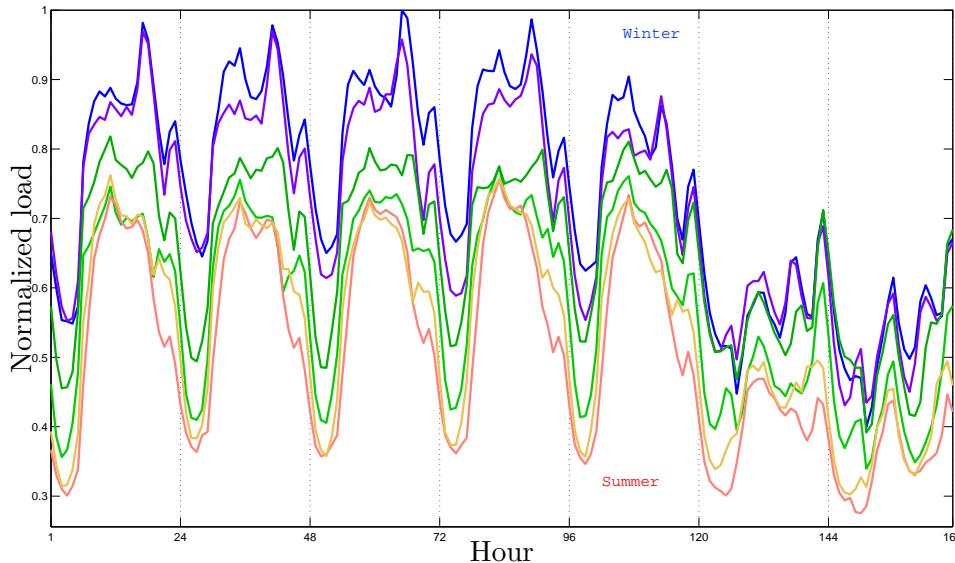


Training: $\hat{y}_{k+1} = f(y_k, y_{k-1}, \dots, y_{k-p})$

Iterative prediction: $\hat{y}_{k+1} = f(\hat{y}_k, \hat{y}_{k-1}, \dots, \hat{y}_{k-p})$

(works well for p large, e.g. $p = 50$) [Espinoza et al., 2003]

Electricity load forecasting



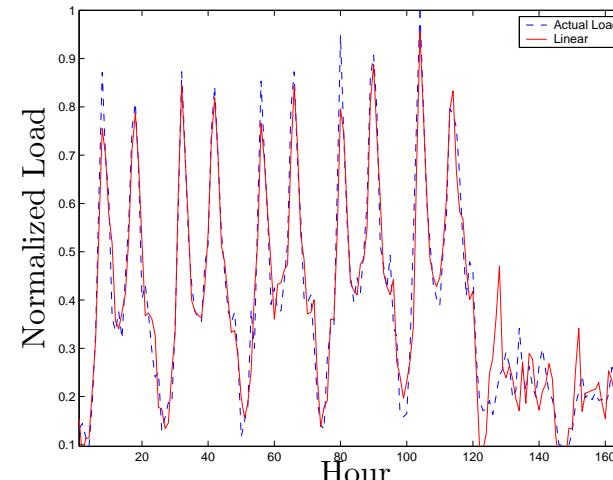
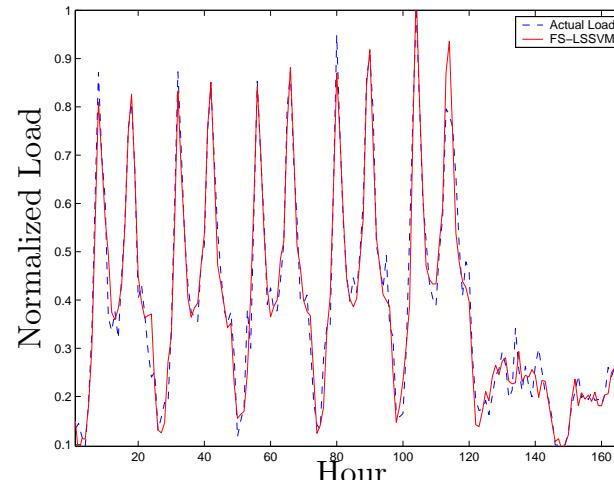
Short-term load forecasting, important for power generation decisions
Hourly load from substations in Belgian grid (ELIA transmission operator)
Seasonal/weekly/intra-daily patterns [Espinoza et al., IEEE CSM 2007]

NARX and AR-NARX model structures: 98 explanatory variables:

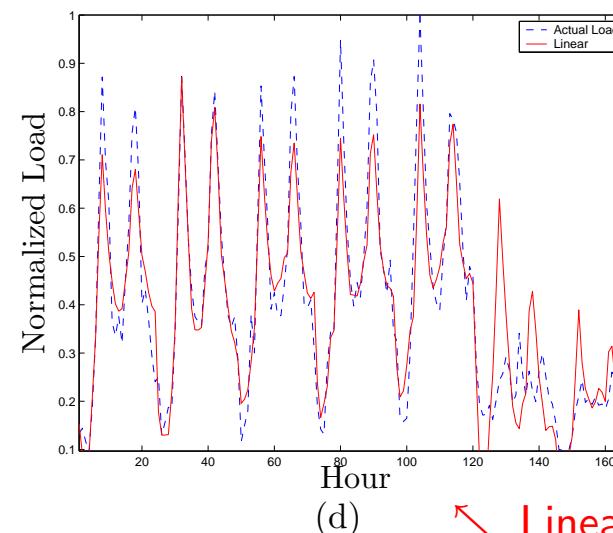
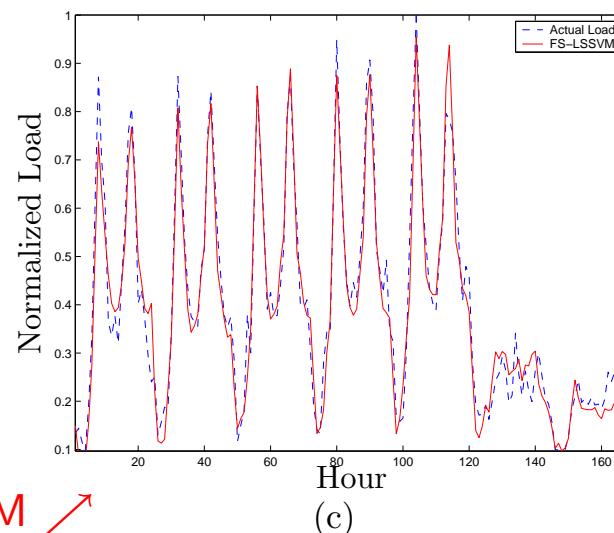
- lagged load values previous two days (48)
- effect of temperature on cooling and heating requirements (3)
- calendar information: month, day, hour indications (43)

Electricity load forecasting

1-hour ahead



24-hours ahead



Fixed-size LS-SVM ↗

↖ Linear ARX model

[Espinoza, Suykens, Belmans, De Moor, IEEE CSM 2007]

*sparse kernel models in community detection and
large graphs*

Modularity and community detection

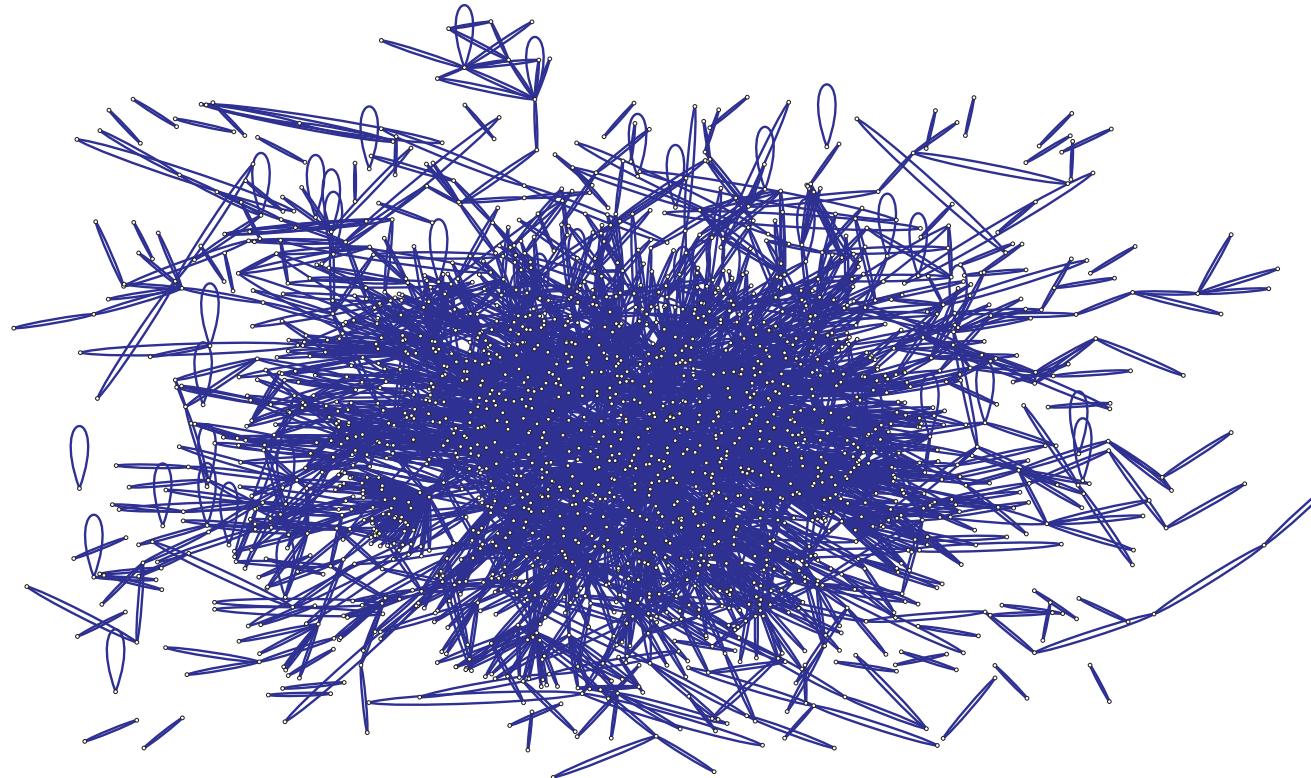
- **Modularity** for two-group case [Newman, 2006]:

$$Q = \frac{1}{4m} \sum_{i,j} (A_{ij} - \frac{d_i d_j}{2m}) q_i q_j$$

with A adjacency matrix, d_i degree of node i , $m = \frac{1}{2} \sum_i d_i$, $q_i = 1$ if node i belongs to group 1 and $q_i = -1$ for group 2.

- **Use of modularity within kernel spectral clustering** [Langone et al., 2012]:
 - use modularity at the level of model **validation**
 - finding **representative subgraph** using a **fixed-size** method by maximizing the expansion factor [Maiya, 2010] $\frac{|\mathcal{N}(\mathcal{G})|}{|\mathcal{G}|}$ with a subgraph \mathcal{G} and its neighborhood $\mathcal{N}(\mathcal{G})$.
 - definition data in unweighted networks: $x_i = A(:, i)$;
 - use of a community kernel function [Kang, 2009].

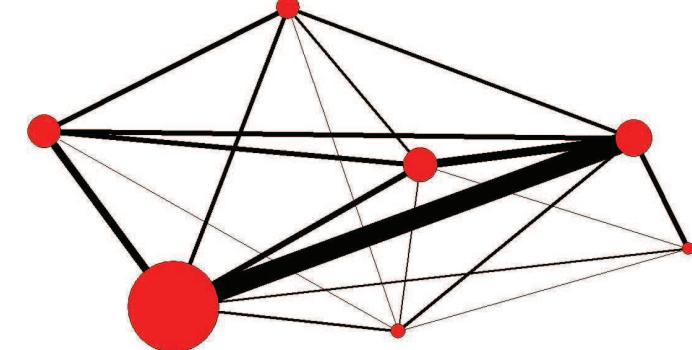
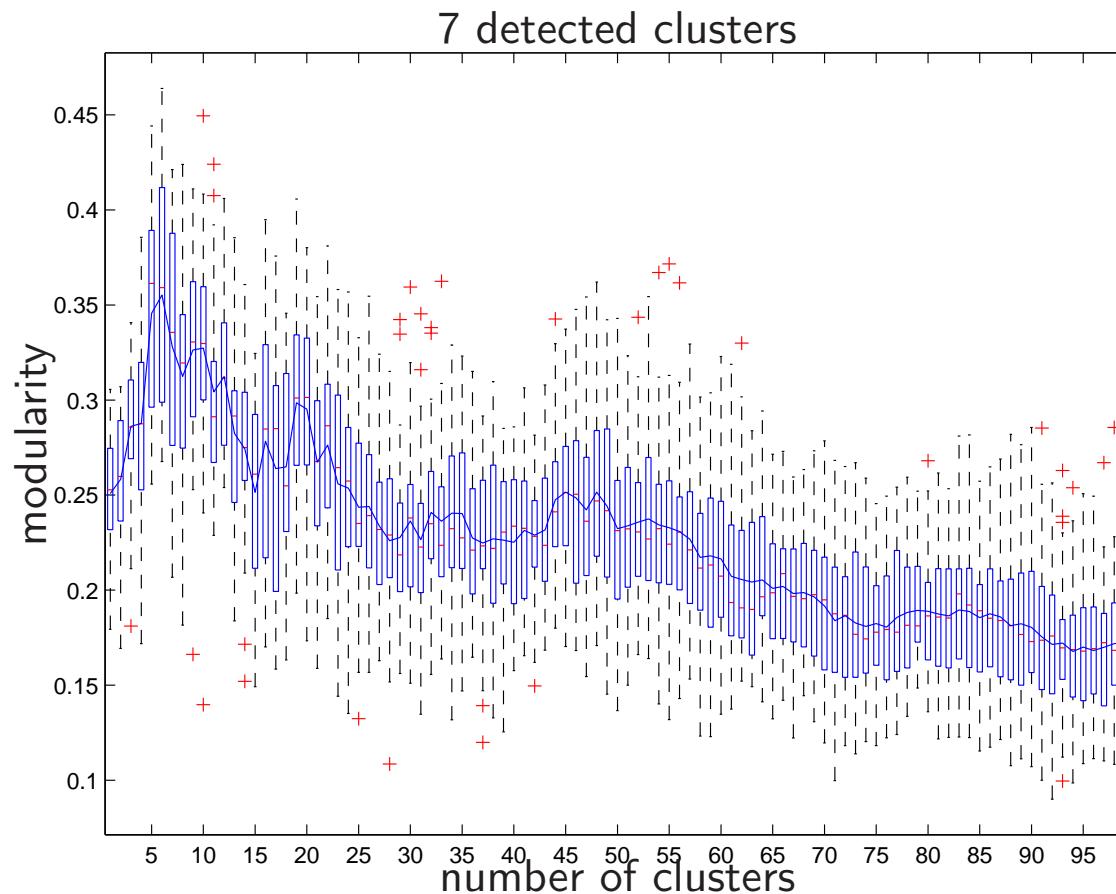
Protein interaction network (1)



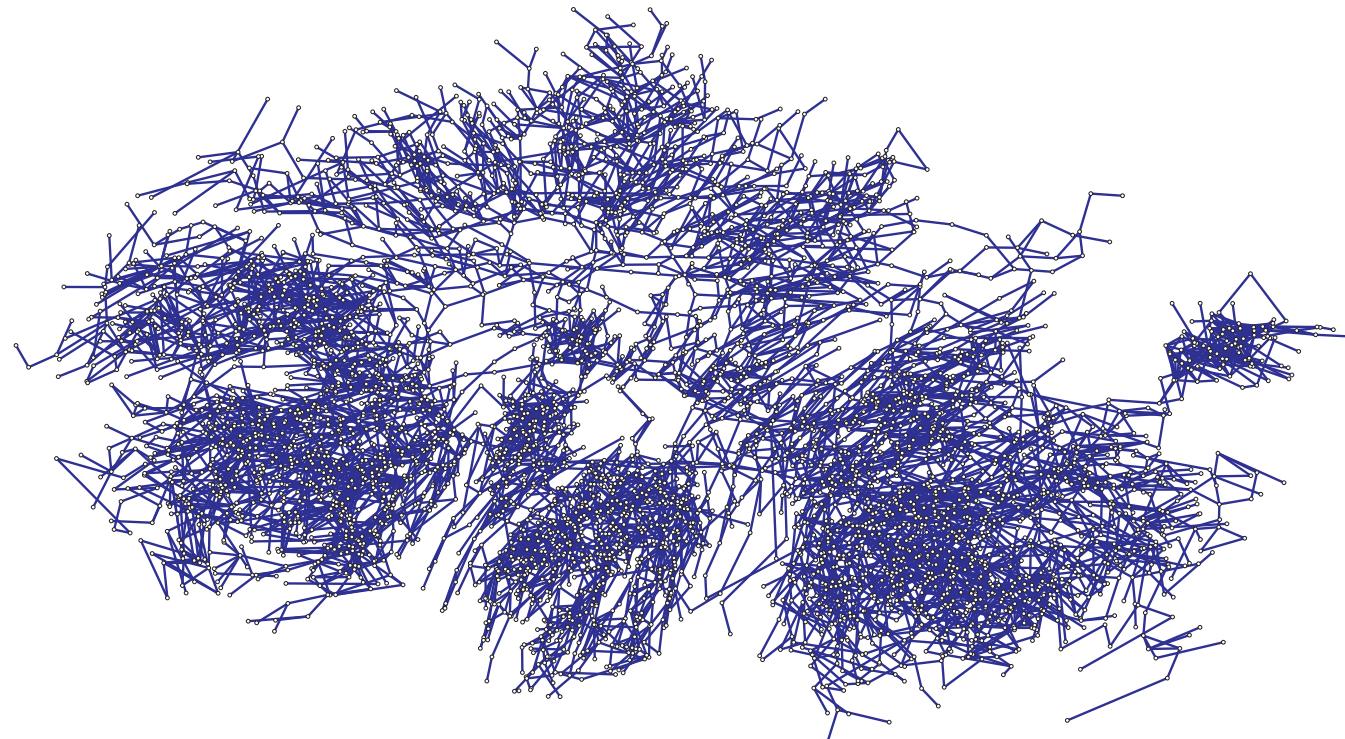
Yeast interaction network: 2114 nodes, 4480 edges [Barabasi et al., 2001]

Protein interaction network (2)

- Yeast interaction network: 2114 nodes, 4480 edges [Barabasi et al., 2001]
- KSC community detection, representative subgraph [Langone et al., 2012]



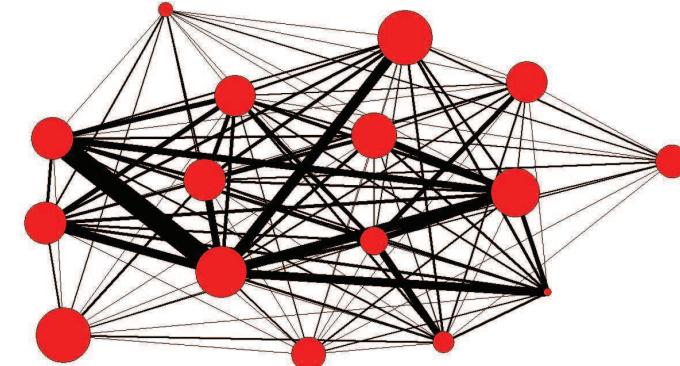
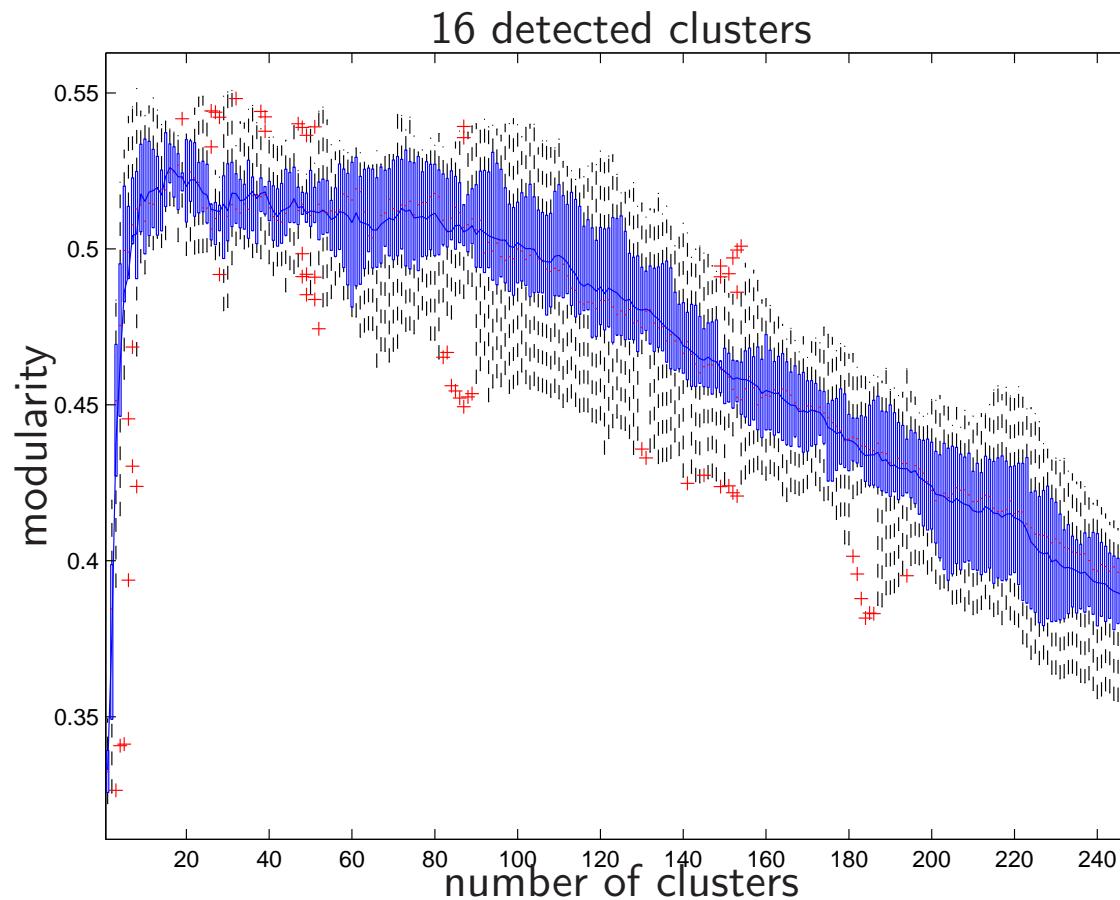
Power grid network (1)



Western USA power grid: 4941 nodes, 6594 edges [Watts & Strogatz, 1998]

Power grid network (2)

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



KSC for big data networks (1)

- **YouTube Network:** YouTube social network where users form friendship with each other and the users can create groups where others can join.
- **RoadCA Network:** a road network of California. Intersections and endpoints are represented by nodes and the roads connecting these intersections are represented as edges.
- **Livejournal Network:** free online social network where users are bloggers and they declare friendship among themselves.

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:

- Select representative training subgraph using FURS
(FURS = Fast and Unique Representative Subset selection [Mall et al., 2013]: selects nodes with high degree centrality belonging to different dense regions, using a deactivation and activation procedure)
- Perform model selection using BAF
(BAF = Balanced Angular Fit: makes use of a 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

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

KSC for big data networks (3)

BAF-KSC

[Mall, Langone, Suykens, 2013]

Infomap

[Lancichinetti, Fortunato, 2009]

Louvain

[Blondel et al., 2008]

CNM

[Clauset, Newman, Moore, 2004]

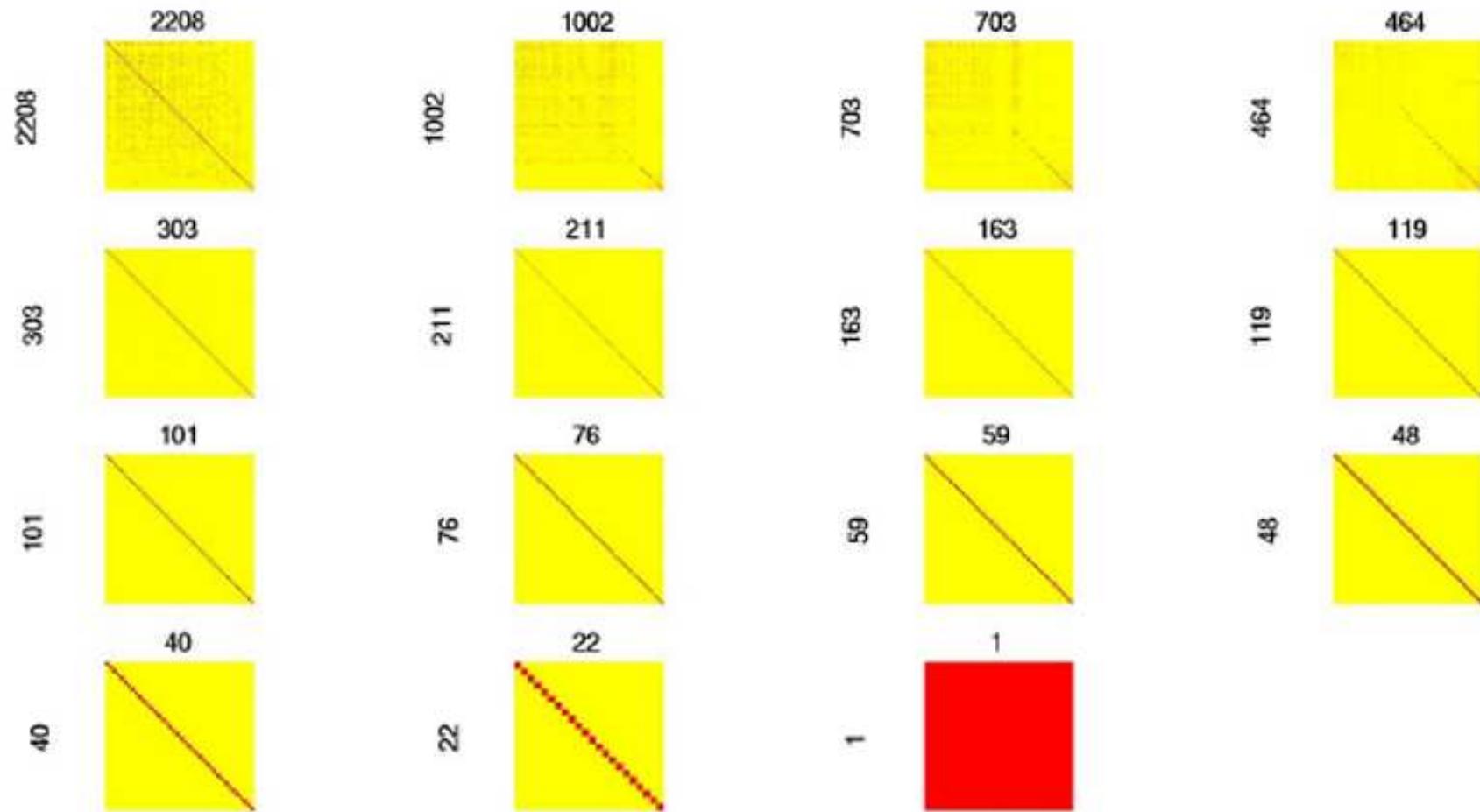
Dataset	BAF-KSC			Louvain			Infomap			CNM		
	CI	Q	Con	CI	Q	Con	CI	Q	Con	CI	Q	Con
Openflight	5	0.533	0.002	109	0.61	0.02	18	0.58	0.005	84	0.60	0.016
PGPnet	8	0.58	0.002	105	0.88	0.045	84	0.87	0.03	193	0.85	0.041
Metabolic	10	0.22	0.028	10	0.43	0.03	41	0.41	0.05	11	0.42	0.021
HepTh	6	0.45	0.0004	172	0.65	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]

CI = 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 (Enron email network)

Multilevel Hierarchical KSC for complex networks (2)

- Start at ground level 0 and compute cosine similarities on validation nodes $S_{val}^{(0)}(i, j) = \text{CosDist}(e_i, e_j)$
- Select projections e_j satisfying $S_{val}^{(0)}(i, j) < t^{(0)}$ with $t^{(0)}$ a threshold value. Keep these nodes as the first cluster at level 0 and remove the nodes from matrix $S_{val}^{(0)}$ to obtain a reduced matrix.
- **Iterative procedure over different levels h :**
Communities at level h become nodes for the next level $h+1$, by creating a set of matrices at different levels of hierarchy:

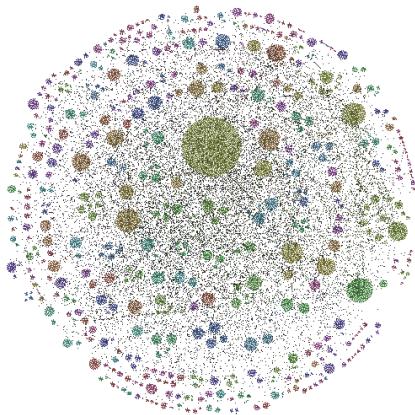
$$S_{val}^{(h)}(i, j) = \frac{1}{|C_i^{(h-1)}| |C_j^{(h-1)}|} \sum_{m \in C_i^{(h-1)}} \sum_{l \in C_j^{(h-1)}} S_{val}^{(h-1)}(m, l)$$

and working with suitable threshold values $t^{(h)}$ for each level.

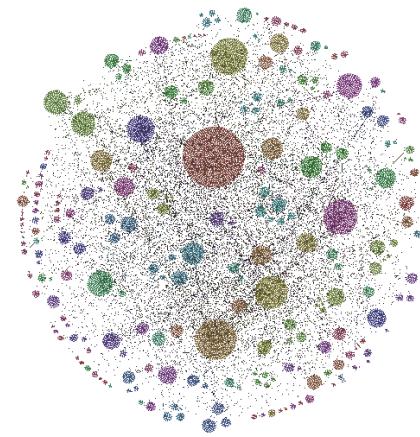
[Mall, Langone, Suykens, PLOS ONE, 2014]

Multilevel Hierarchical KSC for complex networks (3)

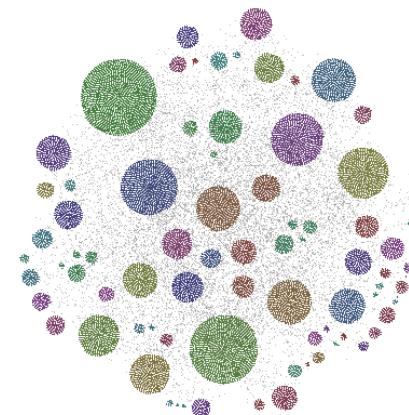
MH-KSC on PGP network:



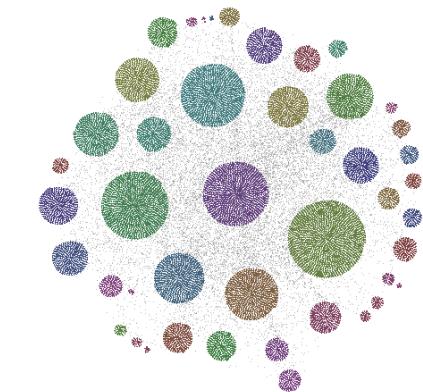
fine



intermediate



intermediate

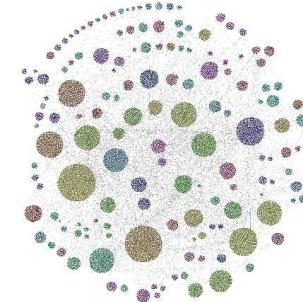
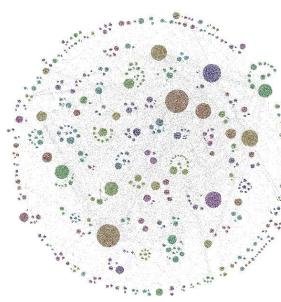
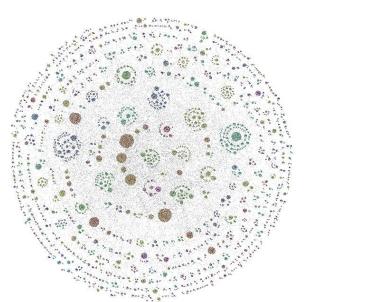


coarse

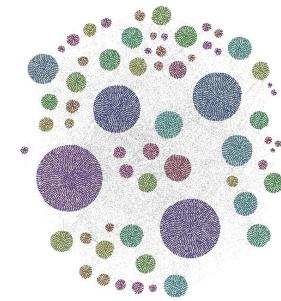
Multilevel Hierarchical KSC finds high quality clusters at coarse as well as fine and intermediate levels of hierarchy.

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

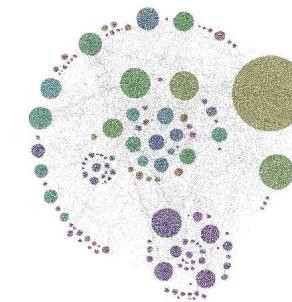
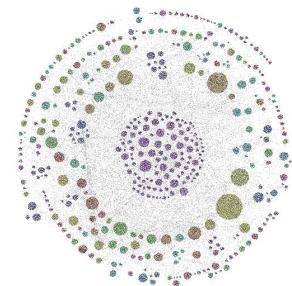
Multilevel Hierarchical KSC for complex networks (4)



Louvain



Infomap



OSLOM

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

Predicting complex networks using multitask learning

Black-box weather forecasting



Weather data NOAA
mid 1998 - late 2010
350 stations located in US

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

Preprocessing: kernel pca
(100 dominant components)

Black-box forecasting multiple weather stations simultaneously

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

Use of multitask learning (1)

- Time-varying model

$$y_t = f(x_t, t) + e_t$$

with $x_t = [y_{t-1}, \dots, y_{t-p}]^T$.

- Consider f in a RKHS with kernel $K((x, t), (x', t')) = x^T x' g(t, t')$ and periodic kernel $g(t, t') = \exp(-\sin(\pi|t - t'|/T)/\sigma^2)$ with $T = 365$.
- Multiple regression functions $f^{(i)}$ for $i = 1, \dots, I$
Task-specific datasets $\mathcal{D}^{(i)} = \{(z_n, y_n)\}_{n=1}^{N_i}$ with $z = (x, t)$
- Auxiliary operator $F : \mathbb{R}^I \rightarrow \mathcal{H}_K$ to jointly estimate functions $f^{(i)}$ such that

$$f^{(i)} = \langle K_{(x,t)}, F\epsilon^{(i)} \rangle$$

with $\epsilon^{(i)}$ the canonical basis, and $K_{(x,t)} : (x', t') \mapsto K((x, t), (x', t'))$.

Use of multitask learning (2)

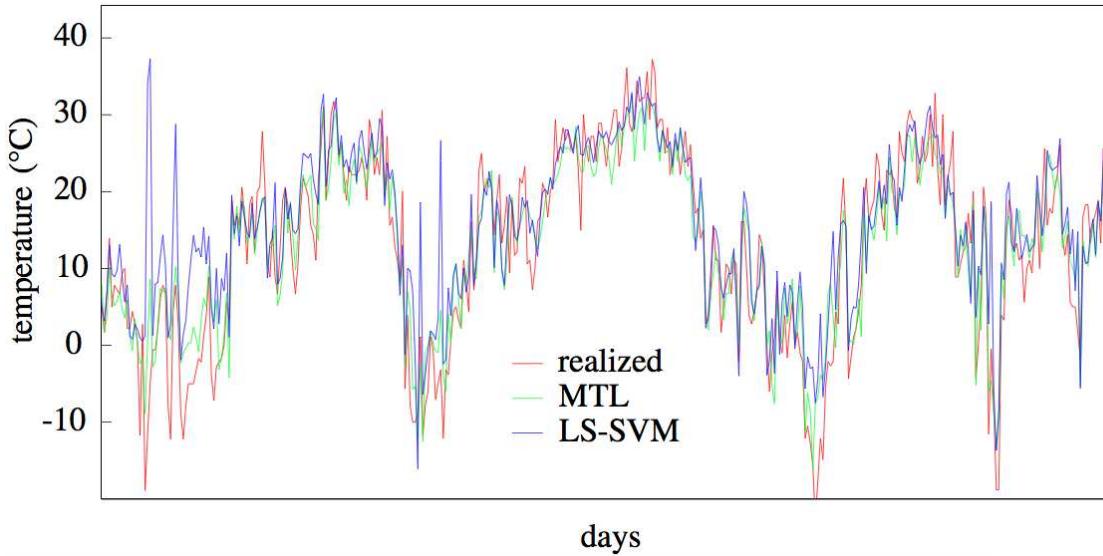
- Learning problem:

$$\min_{F \in \mathcal{F}} \sum_{i \in [I]} \sum_{(z,y) \in \mathcal{D}^{(i)}} (y - \langle K_z, F\epsilon^{(i)} \rangle)^2 \quad \text{s.t.} \quad \|F\|_1 \leq \tau$$

with \mathcal{F} the set of finite-rank operators from \mathbb{R}^I to \mathcal{H}_K and
nuclear norm $\|F\|_1 = \sum_{i \in [I]} \sigma_i(F)$.

- Representer theorem [Abernethy et al., 2009]:
 \hat{F} lies in the span of rank-1 operators constructed based upon input patterns and task indicators.
- Estimation of coefficients based on generalized Frank-Wolfe algorithm.

Use of multitask learning (3)



- Improved results by multitask learning (1-step ahead temperature prediction on test data) with respect to model estimation per station independently from the other stations.
- Use of just-in-time compiled versions for GPUs and CPUs.

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

Conclusions

- Kernel-based **modelling approach** to spectral clustering
- It provides a **powerful framework** for out-of-sample extensions, extracting representative subgraphs, incorporating prior knowledge and handling large scale problems.
- Sparse kernel models using **fixed-size method**
- **Successful applications** to multilevel hierarchical clustering and community detection in complex networks, power load forecasting, PM10 concentrations pollution modelling, black-box weather forecasting

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. Alzate, A. Argyriou, R. Castro, J. De Brabanter, K. De Brabanter, J. de Haas, L. De Lathauwer, B. De Moor, M. Espinoza, Y. Feng, E. Frandi, D. Geubelen, X. Huang, V. Jumutc, Z. Karevan, X. Lou, R. Langone, R. Mall, S. Mehrkanoon, J. Puertas, M. Seslija, L. Shi, M. Signoretto, J. Vandewalle, 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