Towards a low-cost scalable data-driven modeling of coupled systems and non-stationary time series analysis

IPAM Workshop, Los Angeles
October 28 - November 1, 2019

Illia Horenko
USI Lugano, Switzerland

Susanne Gerber
JGU Mainz, Germany

Lukáš Pospíšil
VSB-TU Ostrava, Czech Republic

Patrick Gagliardini
Will Sawyer
Ganna Marchenko
Karin Everschor-Sitte
Terry O'Kane
et al.

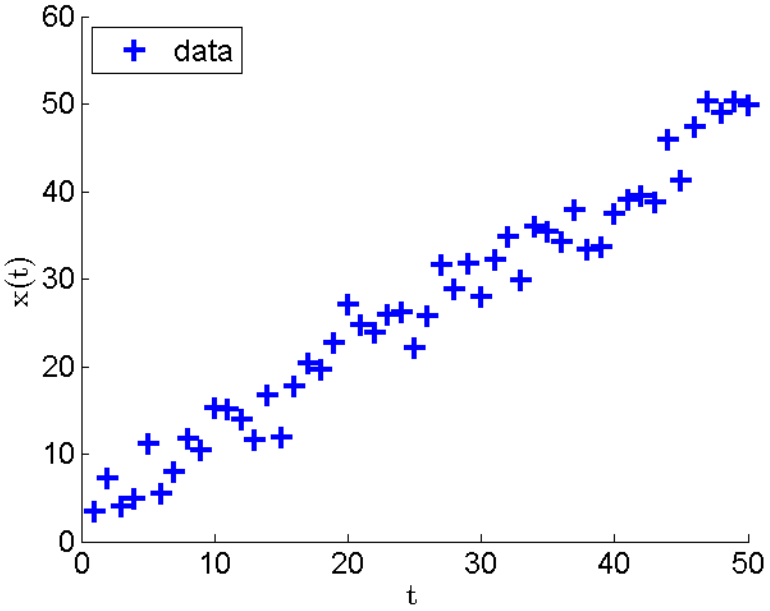
Outline

1. Non-stationary time series analysis
   - General talk about time series (polynomial regression, AIC,…)
   - Non-stationary modeling
   - K-means: unsupervised clustering algorithm
   - FEM-H1, FEM-BV

2. Scalable Probabilistic Approximation (SPA)
   - Exact Law of the Total Probability
   - Clustering = the cure for the curse of dimensionality?
   - Soft (probabilistic) K-means without Jensen inequality
   - Compressing the pipeline:
     the combination of optimal discretization with Markov model

Regime-based non-stationary time series analysis with regularization
(FEM-H1, FEM-BV)

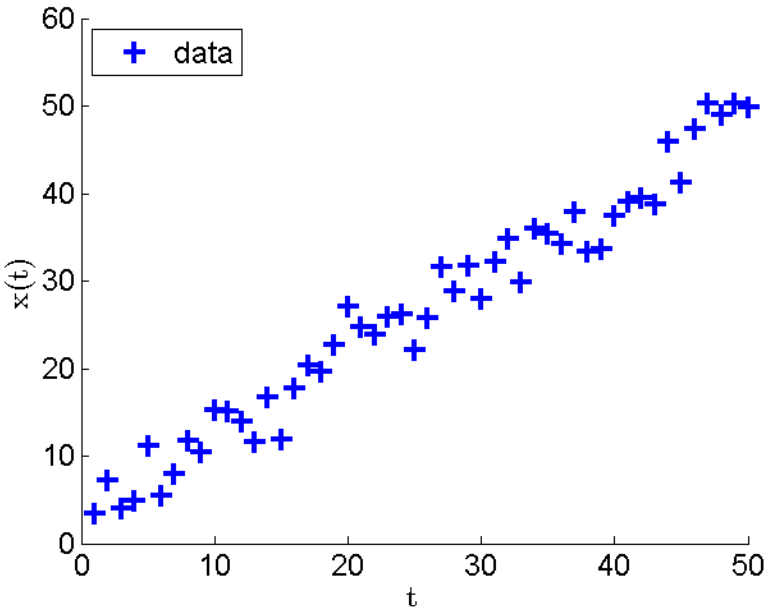
Time series analysis



- time series = data measured in intervals over a period of time

$$x_t \in \mathbb{R}^n, t = 1, \ldots, T$$

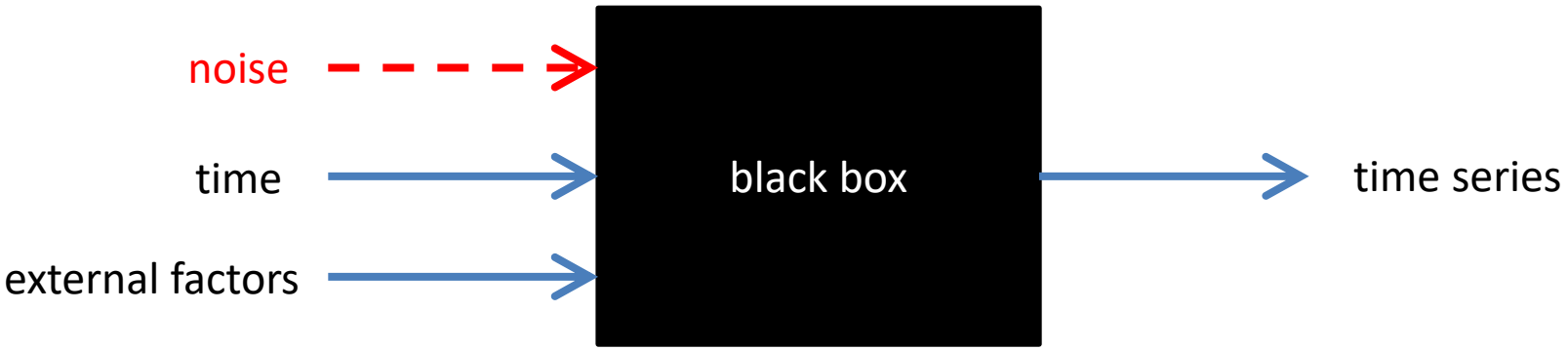
Time series analysis



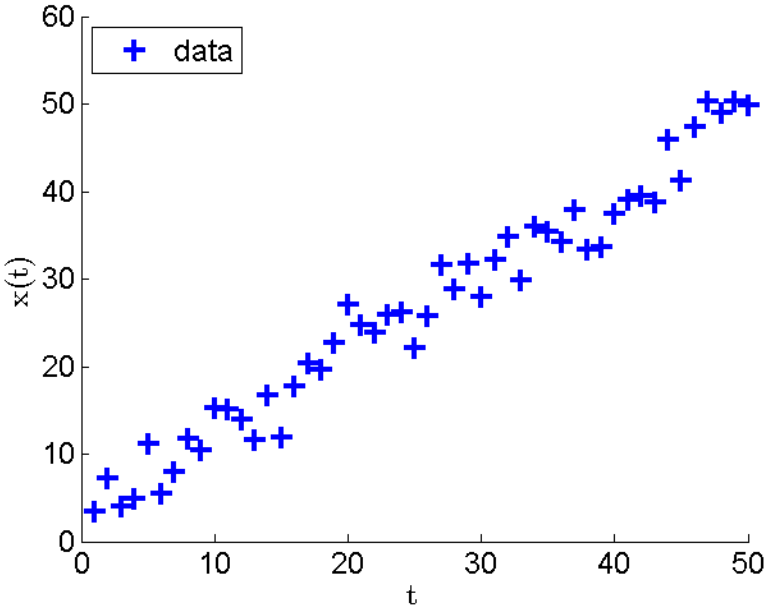
$$x_t = \underbrace{\mu(t, \Theta)}_{\text{model function}} + \underbrace{\varepsilon_t}_{\text{noise}}$$

- time series = data measured in intervals over a period of time

$$x_t \in \mathbb{R}^n, t = 1, \ldots, T$$



Time series analysis



$$x_t = \mu(t, \Theta) + \varepsilon_t$$

- Example: linear regression

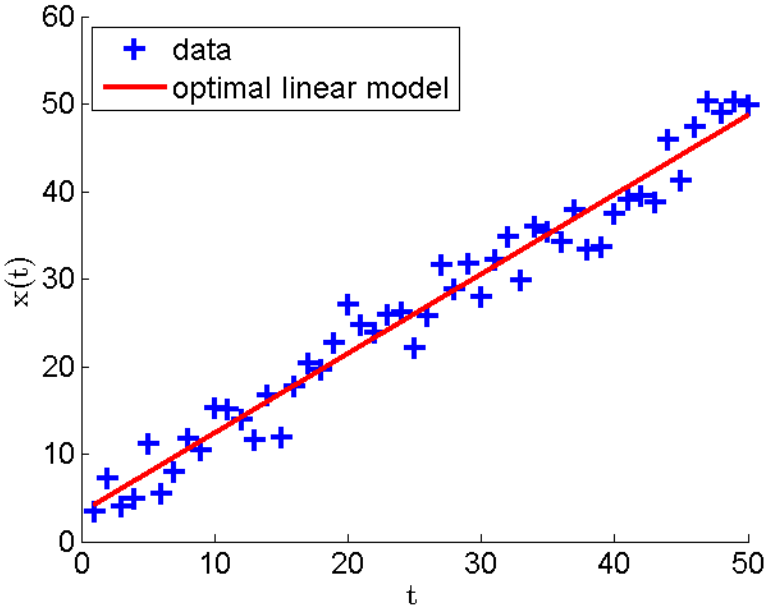
$$\mu(t, \theta_0, \theta_1) = \theta_1 t + \theta_0$$

- time series = data measured in intervals over a period of time

$$x_t \in \mathbb{R}^n, t = 1, \ldots, T$$

- choose model

Time series analysis



$$x_t = \mu(t, \Theta) + \varepsilon_t$$

- Example: linear regression

$$\mu(t, \theta_0, \theta_1) = \theta_1 t + \theta_0$$

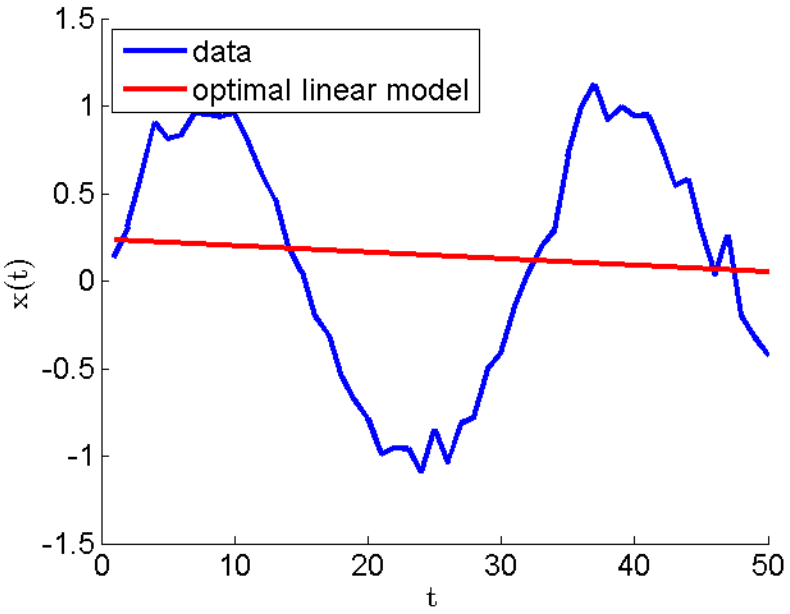
$$[\bar{\theta}_0, \bar{\theta}_1] = \arg\min_{\theta_0, \theta_1} \sum_{t=1}^{T} \|x_t - \mu(t, \theta_0, \theta_1)\|^2$$

- time series = data measured in intervals over a period of time

$$x_t \in \mathbb{R}^n, t = 1, \ldots, T$$

- choose model
- find parameters of (optimal) model

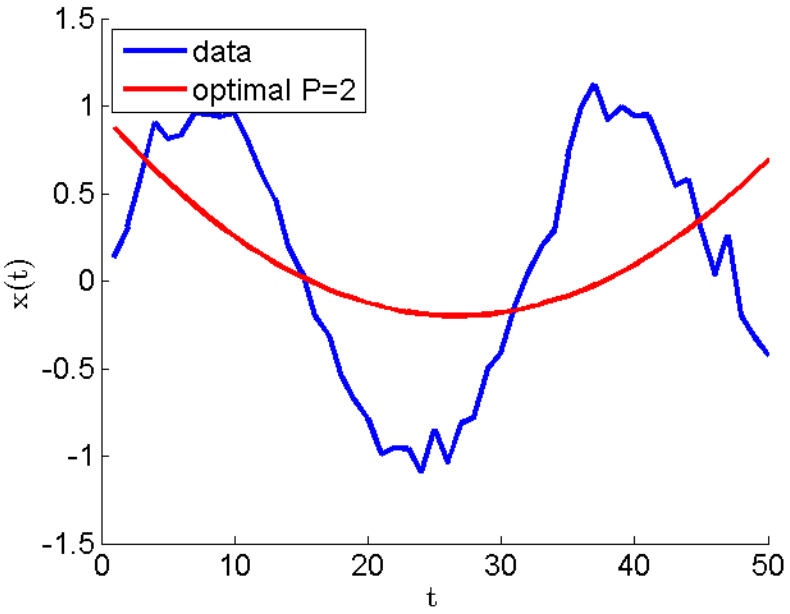
Time series analysis



$$\bar{\Theta} = \arg\min_{\Theta \in \mathcal{P}} \sum_{t=1}^{T} \rho(x_t, \mu(t, \Theta))$$

- more "sophisticated" parametric models:
  - general polynomial regression
  - autoregressive models
  - Hidden Markov models
  - Neural Networks
  - ...

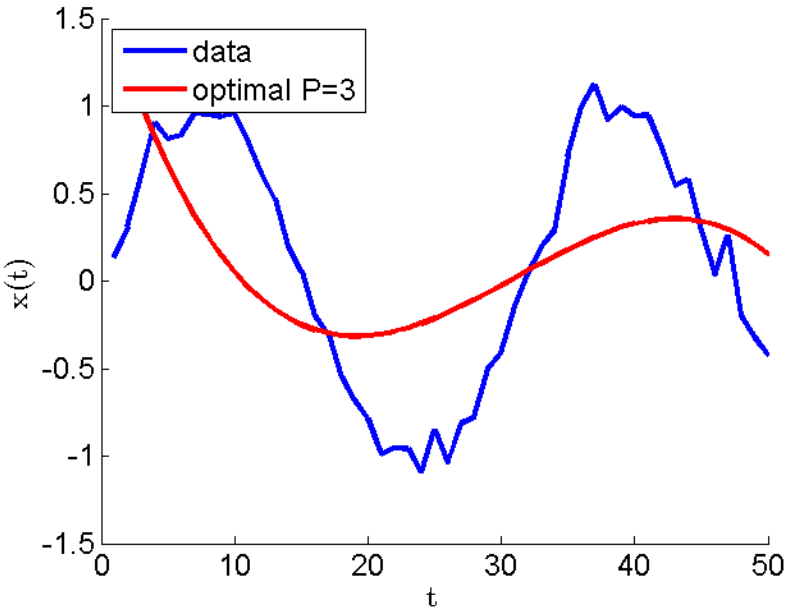
Time series analysis



$$\bar{\Theta} = \arg\min_{\Theta \in \mathcal{P}} \sum_{t=1}^{T} \rho(x_t, \mu(t, \Theta))$$

- more "sophisticated" parametric models:
  - general polynomial regression
  - autoregressive models
  - Hidden Markov models
  - Neural Networks
  - ...

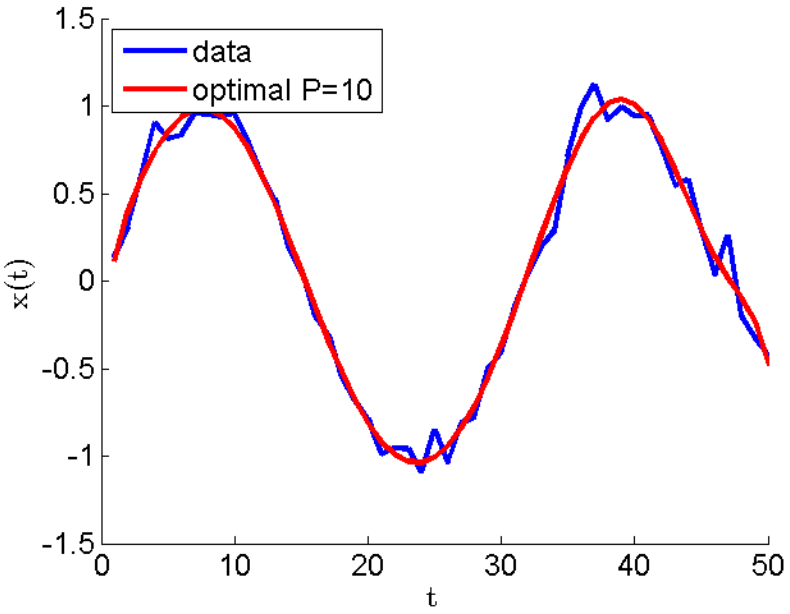
Time series analysis



$$\bar{\Theta} = \arg\min_{\Theta \in \mathcal{P}} \sum_{t=1}^{T} \rho(x_t, \mu(t, \Theta))$$

- more "sophisticated" parametric models:
  - general polynomial regression
  - autoregressive models
  - Hidden Markov models
  - Neural Networks
  - ...

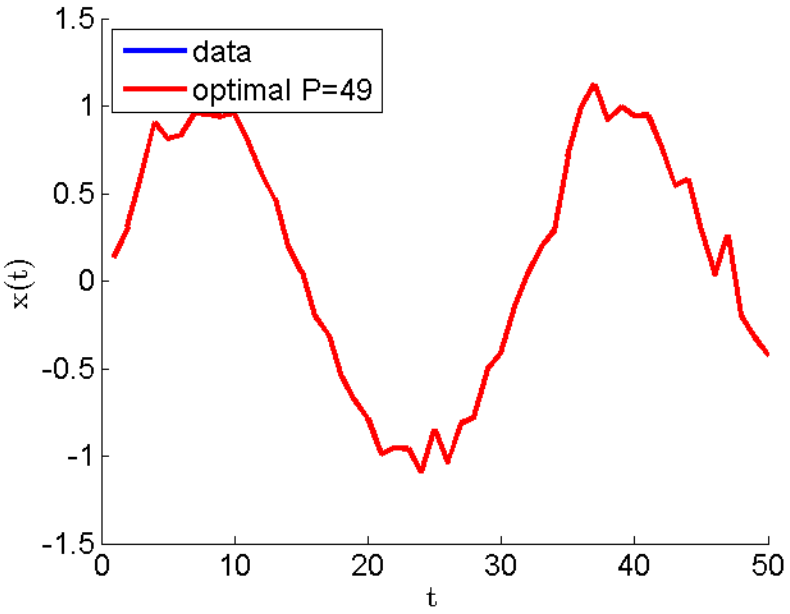
Time series analysis



$$\bar{\Theta} = \arg\min_{\Theta \in \mathcal{P}} \sum_{t=1}^{T} \rho(x_t, \mu(t, \Theta))$$

- more "sophisticated" parametric models:
  - general polynomial regression
  - autoregressive models
  - Hidden Markov models
  - Neural Networks
  - ...

Time series analysis



$$\bar{\Theta} = \arg\min_{\Theta \in \mathcal{P}} \sum_{t=1}^{T} \rho(x_t, \mu(t, \Theta))$$

- more "sophisticated" parametric models:
  - general polynomial regression
  - autoregressive models
  - Hidden Markov models
  - Neural Networks
  - ...

Time series analysis

- in general:

more parameters = lower modeling error

Time series analysis

- in general:

more parameters = lower modeling error

- noise reduction ?
- overfitting ?
- interpretation of results ?

Time series analysis

- in general:

more parameters = lower modeling error 

more parameters = more unknowns 

  - larger optimization problem
  - harder to solve

Time series analysis

- in general:

more parameters = lower modeling error

more parameters = more unknowns

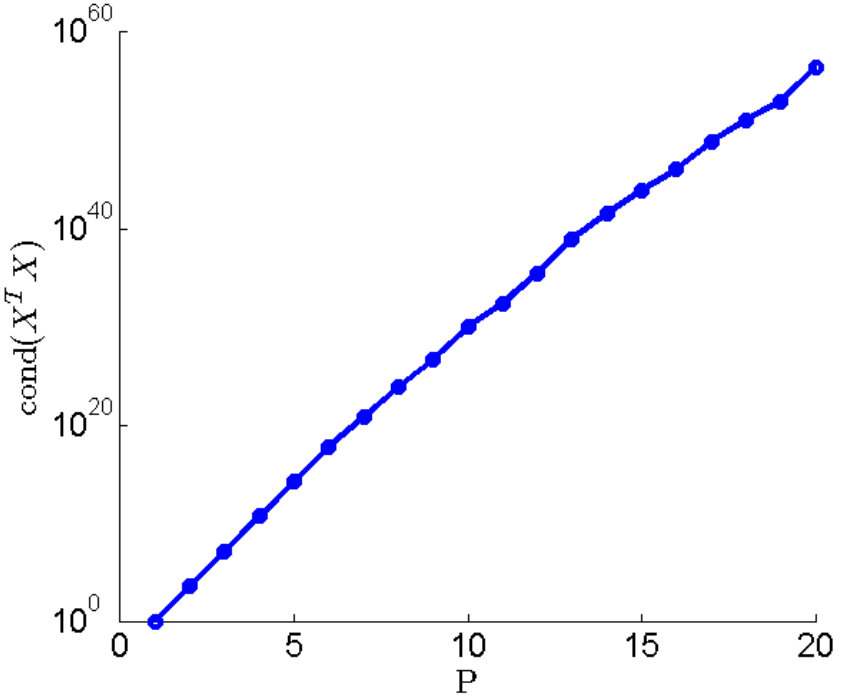
- larger optimization problem
- harder to solve

Example - polynomial regression

$$\mu(t,\Theta) = \sum_{p=0}^{P} \theta_p t^p$$

$$\bar{\Theta} = (X^T X)^{-1} X^T y$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^P \\ 1 & x_2 & x_2^2 & \dots & x_2^P \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_T & x_T^2 & \dots & x_T^P \end{bmatrix}$$



Time series analysis

- in general:

more parameters = lower modeling error

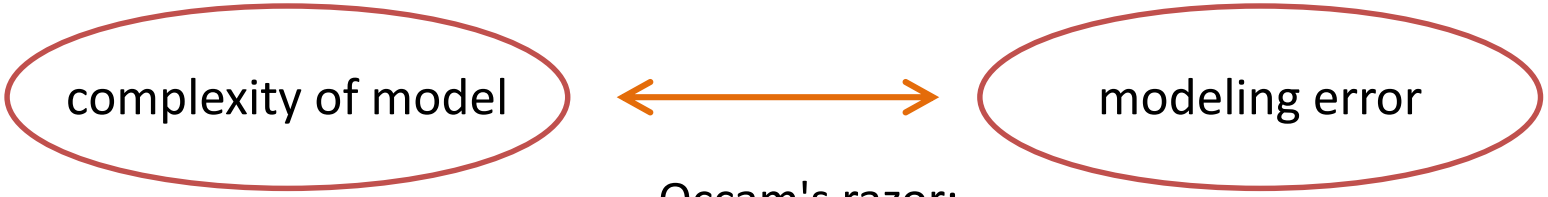
more parameters = more unknowns

Time series analysis

- in general:

more parameters = lower modeling error

more parameters = more unknowns



Occam's razor:

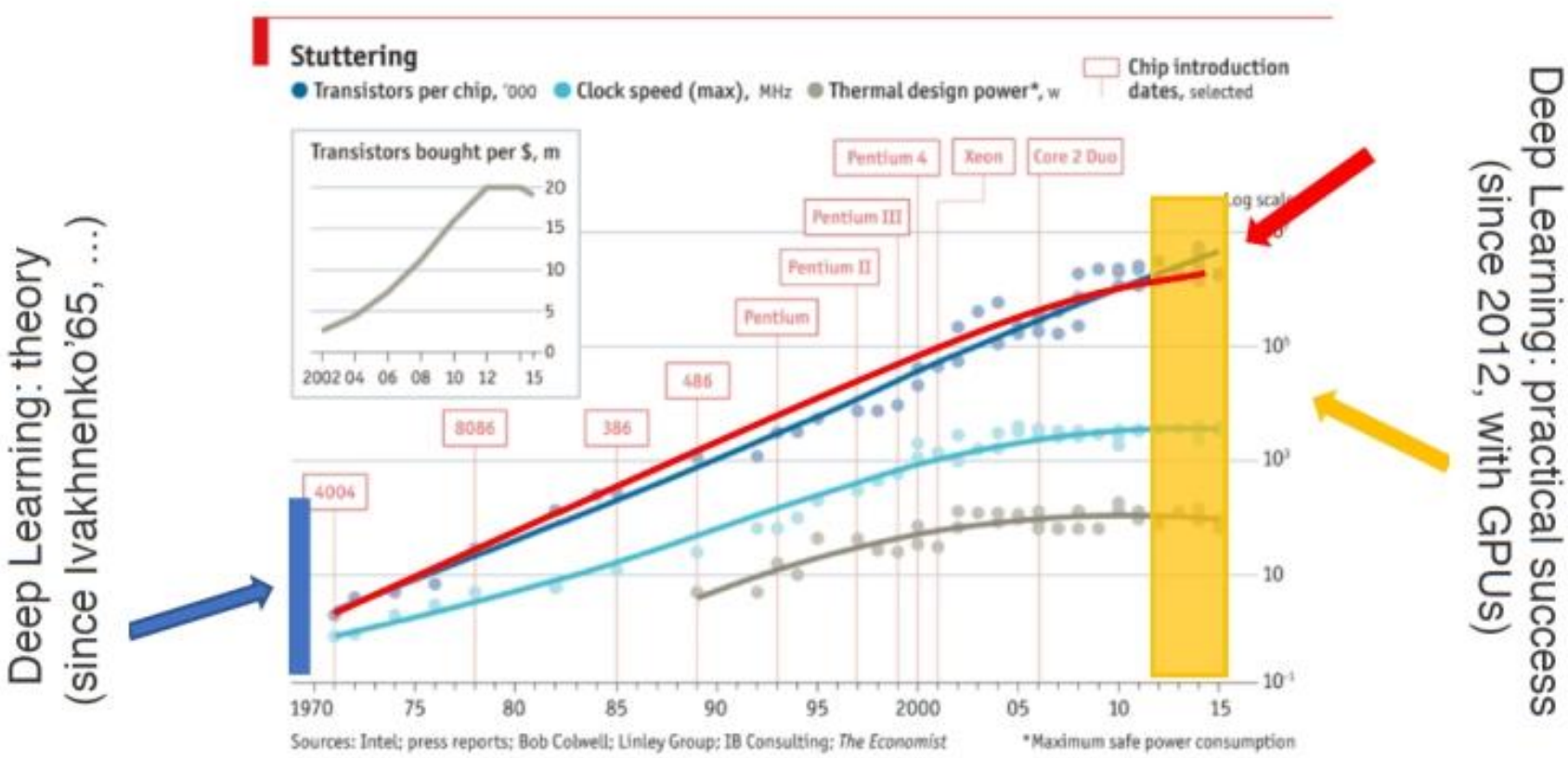
Among competing hypotheses, the one with the fewest assumptions should be selected.

- *Akaike H.:* Information theory and an extension of the maximum likelihood principle*, (1971)*

Example: Moore’s law

Moore's law:

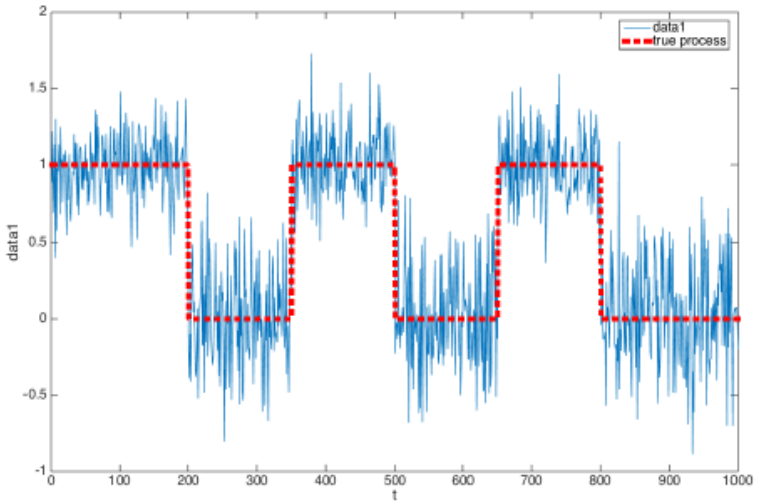
„the number of transistors in a dense integrated circuit doubles about every two years.“



- *H. Khan, D. Hounshell, E. Fuchs:* Science and research policy at the end of Moore’s law, *Nature Electronics, 1: 14–21, (2018)*

Example

Filtering Example: signal/noise ≈ 2/1

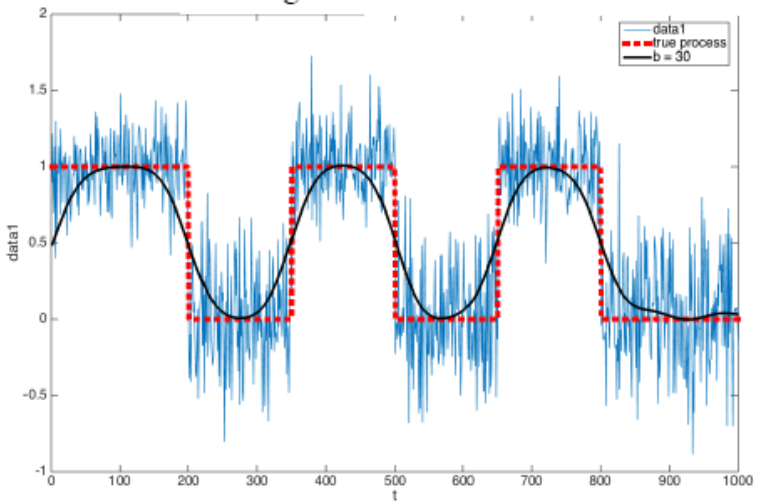


true signal +noise

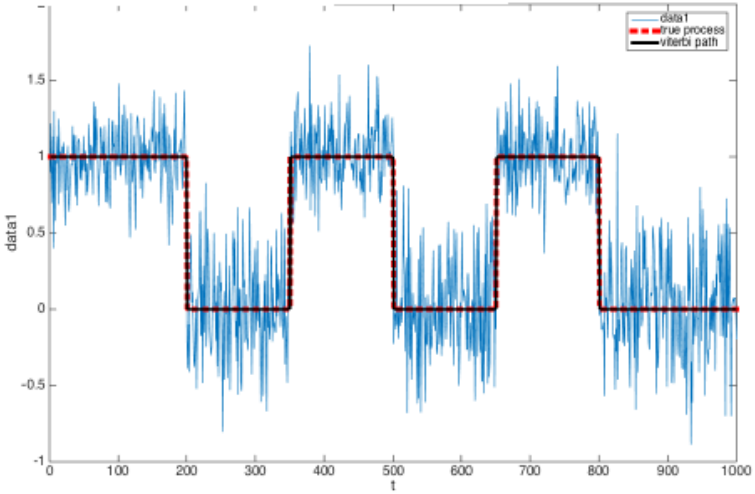
true signal

extracted signal

Moving Gaussian Window



HMM-Gauss

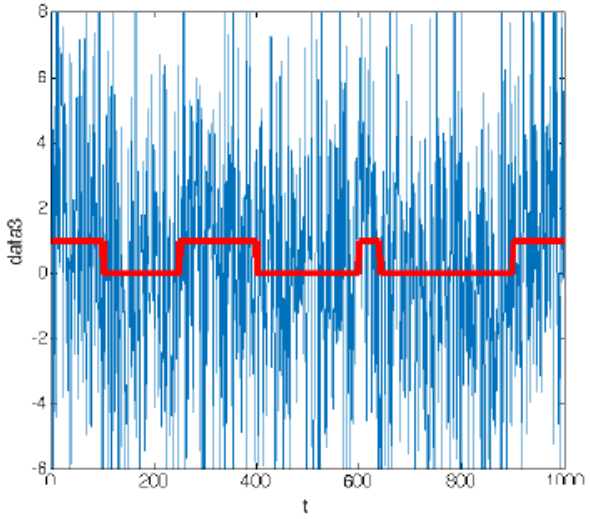


MIT Machine Learning Toolbox: **Gaussian smoothing** and **HMM-Gauss** in optimal setting

Example

Filtering Example: signal/noise ≈ 1/20

Neurosciences: signal/noise ≈ 1/10-1/100

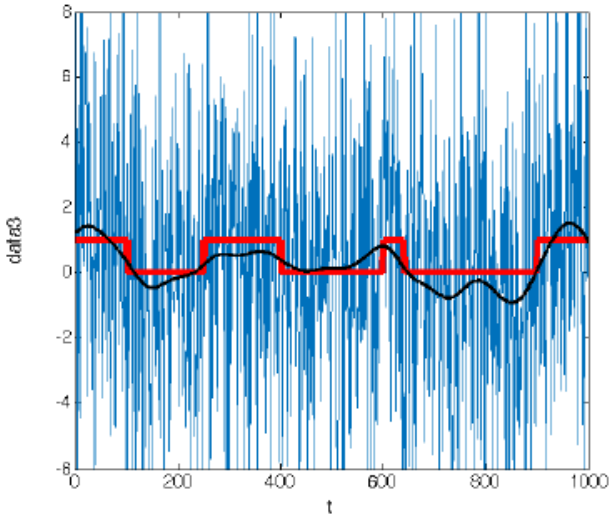


true signal +noise

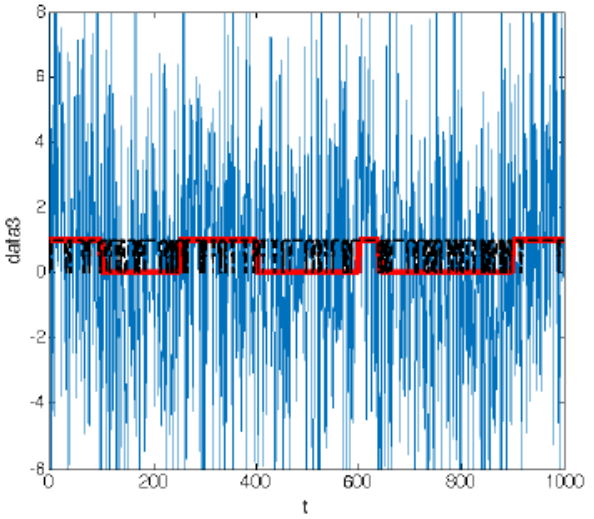
true signal

extracted signal

Moving Gaussian Window

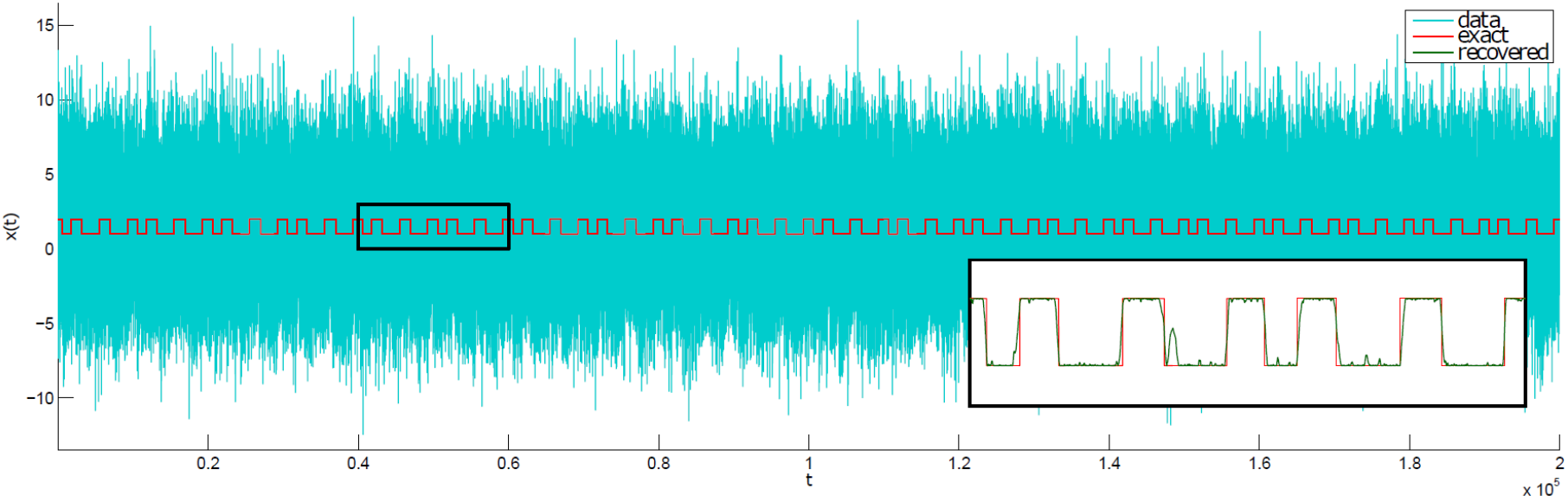


HMM-Gauss



MIT Machine Learning Toolbox: **Gaussian smoothing** and **HMM-Gauss** in optimal setting

FEM-H1: demonstration by results



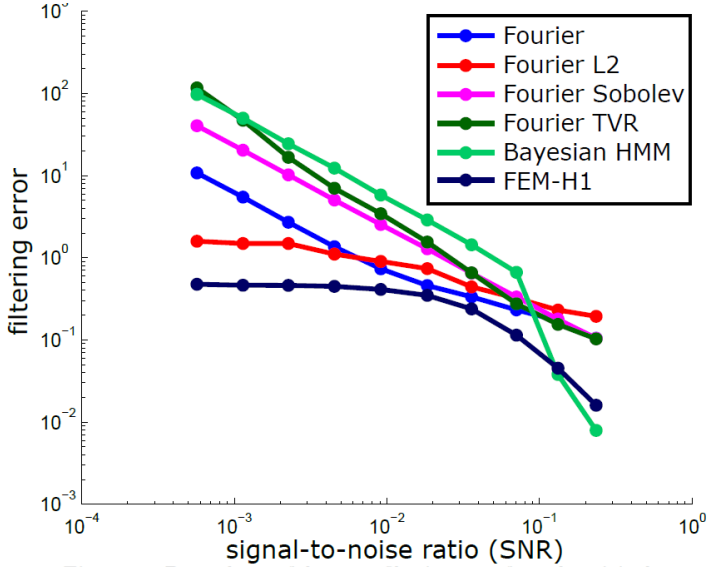


Figure Dependence of the mean filtering error from the original signal-to-noise ratio (SNR). Results were obtained for the signal with Gaussian noise by averaging 100 independent realizations of the noisy signal data for each of the methods.

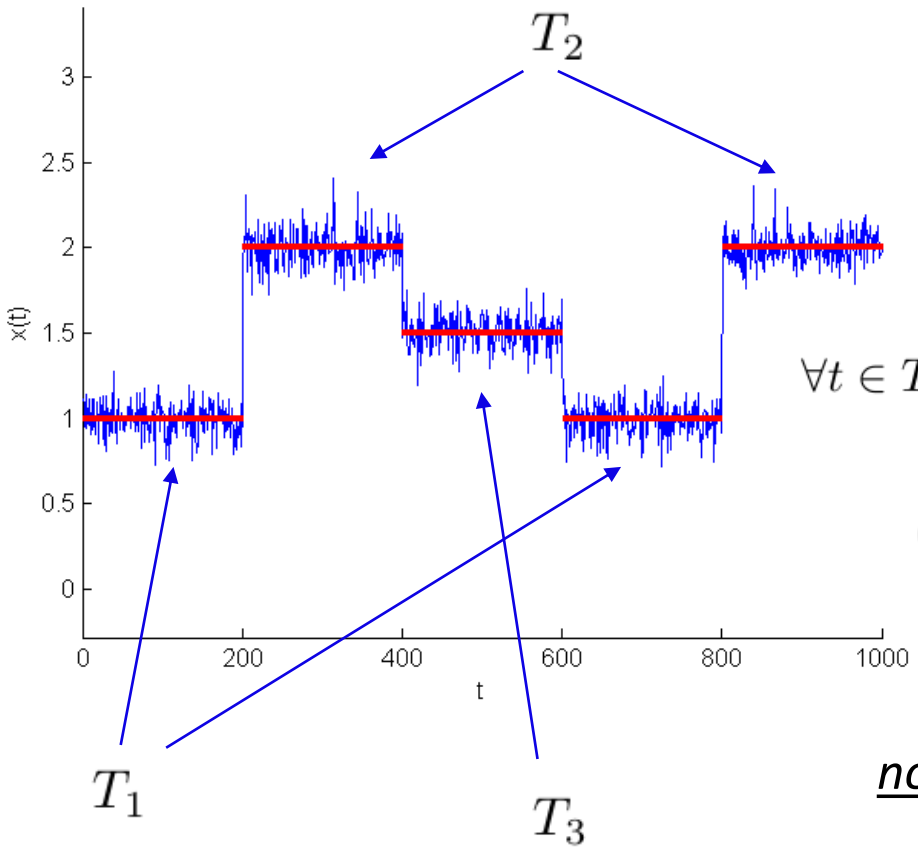
- Pospíšil L., Gagliardini P., Sawyer W., Horenko I.: *On a scalable nonparametric denoising of time series signals.* Communications in Applied Mathematics and Computational Science, 2018.

From non-stationary models to clustering

$$\bar{\Theta}(t) = \arg\min_{\Theta} \sum_{t=1}^{T} \|x_t - \mu(t, \Theta(t))\|^2$$

(non-stationary model)

"constant" regression in regimes



$$x(t) = \theta(t) + \varepsilon_t$$

$$\forall t \in T_k : \ \theta(t) = \theta_k \quad \text{(constant on } k\text{-th cluster)}$$

$$\bar{\Theta} = \arg\min \sum_{k=1}^{K} \sum_{t \in T_k} \|x(t) - \theta_k\|^2$$

non-stationary ⟷ locally stationary

From non-stationary models to clustering

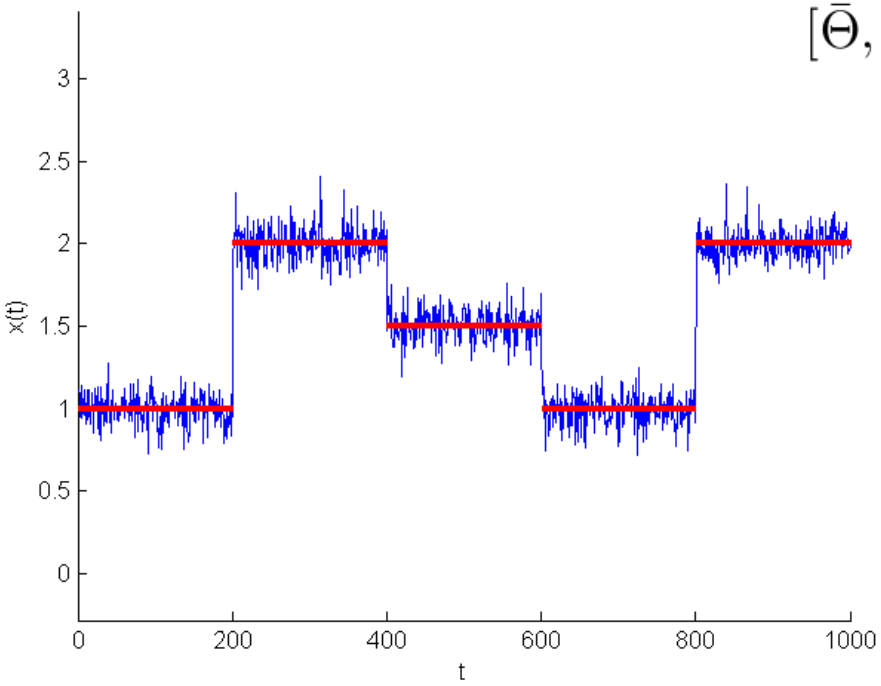
$$\bar{\Theta}(t) = \arg\min_{\Theta} \sum_{t=1}^{T} \|x_t - \mu(t, \Theta(t))\|^2$$

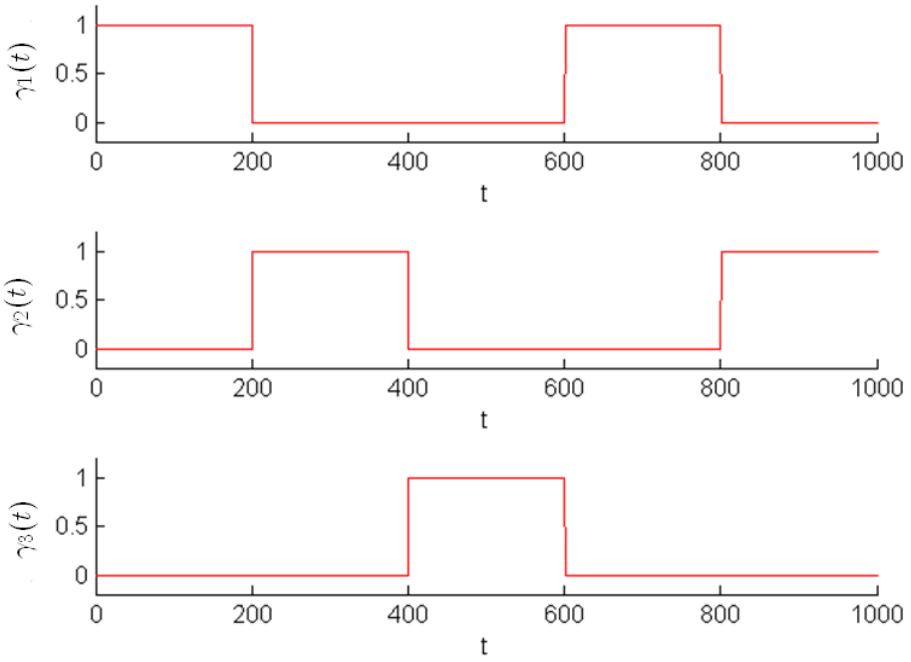
regime indicator function:

$$\gamma_k(t) = \begin{cases} 1 & \text{if } x_t \text{ belongs to } k\text{-th regime} \\ 0 & \text{elsewhere} \end{cases}$$

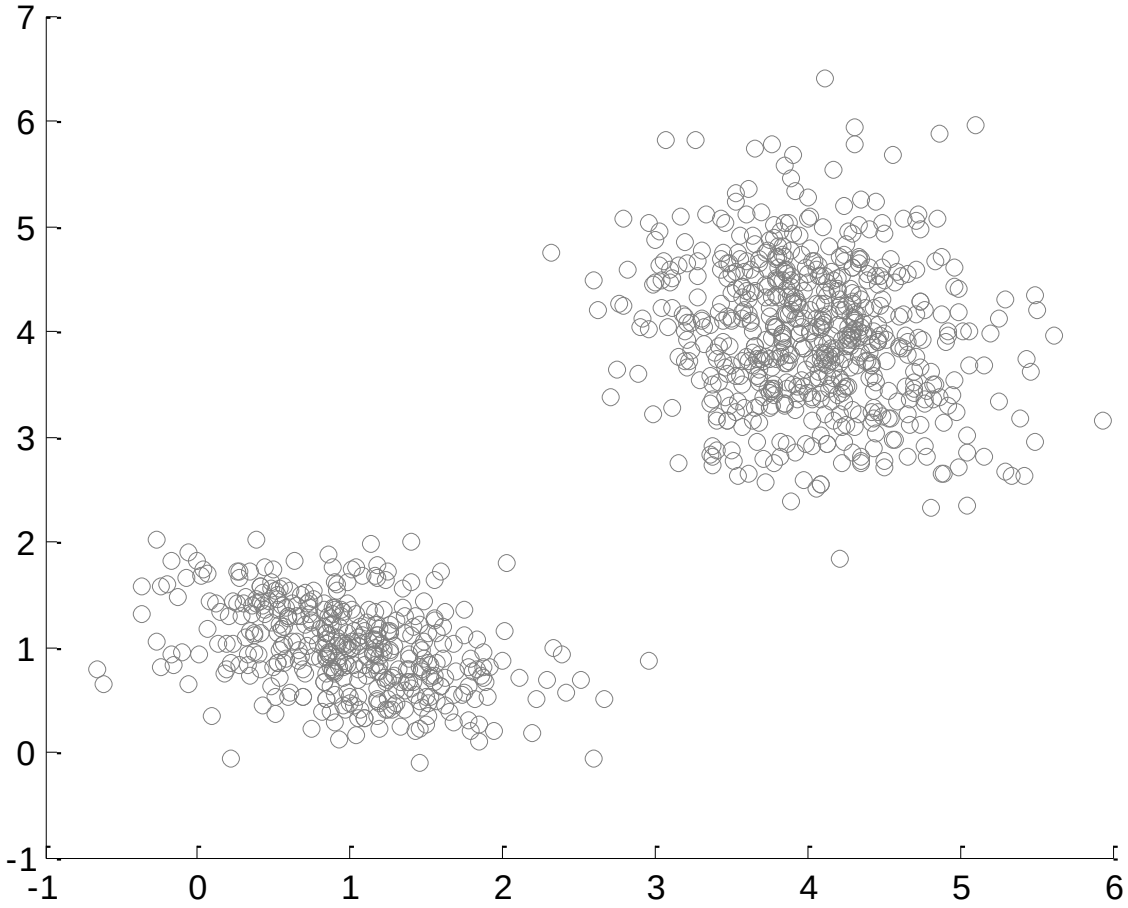
"constant" regression in regimes

$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \theta_k\|^2$$





K-means algorithm: unsupervised clustering



- *Lloyd, Stuart P. :* Least squares quantization in PCM. *Information Theory, IEEE Transactions on 28.2 (1982): 129-137.*

K-means algorithm: unsupervised clustering

- we don't have any apriori classification for supervised learning
- each cluster consists of ***similar*** points (points in cluster are close to each other)
- each cluster can be characterised by ***mean value*** (“centroid”) of points inside it

K-means algorithm

$$[\underbrace{\hat{\mu}_1, \ldots, \hat{\mu}_K}_{\hat{\theta}}, \underbrace{\hat{\gamma}_1, \ldots, \hat{\gamma}_K}_{\hat{\Gamma}}] = \arg \min_{\mu_1, \mu_2, \ldots, \mu_K} \underbrace{\sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_k(x_t) \|x_t - \mu_k\|^2}_{L(\theta, \Gamma)}$$

$$\Omega_\Gamma \left[\begin{array}{c} \forall t, k : \gamma_k(x_t) \in \{0, 1\} \\ \forall t : \sum_{k=1}^{K} \gamma_k(x_t) = 1 \end{array} \right]$$

set feasible initial approximation Γ_0

while $\|L(\Gamma_{it}, \theta_{it}) - L(\Gamma_{it-1}, \theta_{it-1})\| > \varepsilon$

$\quad$ *solve* $\theta_{it} = \arg \min_{\theta} L(\theta, \Gamma_{it})$ $\quad$ *(with fixed* Γ_{it}*)*

$\quad$ *solve* $\Gamma_{it} = \arg \min_{\Gamma \in \Omega_\Gamma} L(\theta_{it}, \Gamma)$ $\quad$ *(with fixed* θ_{it}*)*

$\quad$ $it = it + 1$

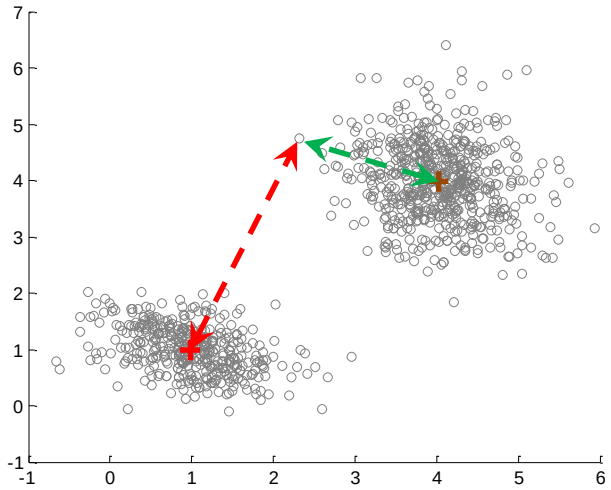
endwhile

K-means algorithm

$$\Gamma_{it} = \arg\min_{\Gamma \in \Omega_\Gamma} \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_k(x_t) \|x_t - \mu_k\|^2$$

subject to $\gamma_k(x_t) \in \{0, 1\}, \forall t$: exactly one of $\gamma_k(x_t)$ is equal to 1

$$\hat{\gamma}_{\hat{k}}(x_t) := \begin{cases} 1 & \text{if } \hat{k} = \arg_k \min \|x_t - \mu_k\|^2 \\ 0 & \text{otherwise} \end{cases}$$



set feasible initial approximation Γ_0

while $\|L(\Gamma_{it}, \theta_{it}) - L(\Gamma_{it-1}, \theta_{it-1})\| > \varepsilon$

solve $\theta_{it} = \arg\min_{\theta} L(\theta, \Gamma_{it})$ *(with fixed* Γ_{it}*)*

solve $\Gamma_{it} = \arg\min_{\Gamma \in \Omega_\Gamma} L(\theta_{it}, \Gamma)$ *(with fixed* θ_{it}*)*

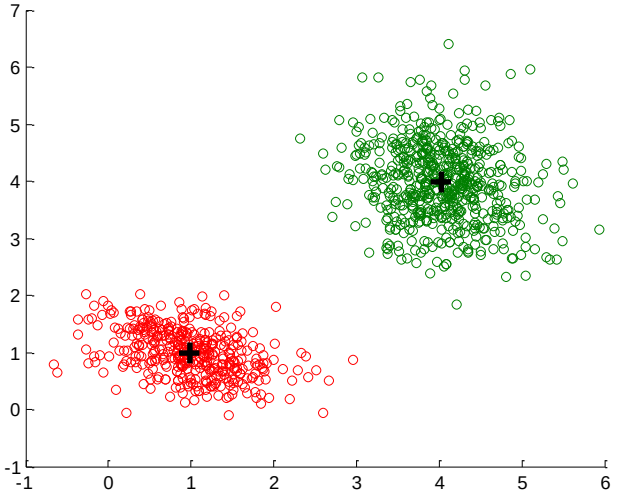
$it = it + 1$

endwhile

K-means algorithm

$$\theta_{it} = \arg\min_{\theta} \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_k(x_t) \|x_t - \mu_k\|^2$$

$$\hat{\mu}_k = \frac{\sum_{t=1}^{T} \gamma_k(x_t) x_t}{\sum_{t=1}^{T} \gamma_k(x_t)}$$



$$\mu = \frac{1}{N} \sum_t x_t \qquad \mu = \frac{1}{N} \sum_t x_t$$

set feasible initial approximation Γ_0

while $\|L(\Gamma_{it}, \theta_{it}) - L(\Gamma_{it-1}, \theta_{it-1})\| > \varepsilon$

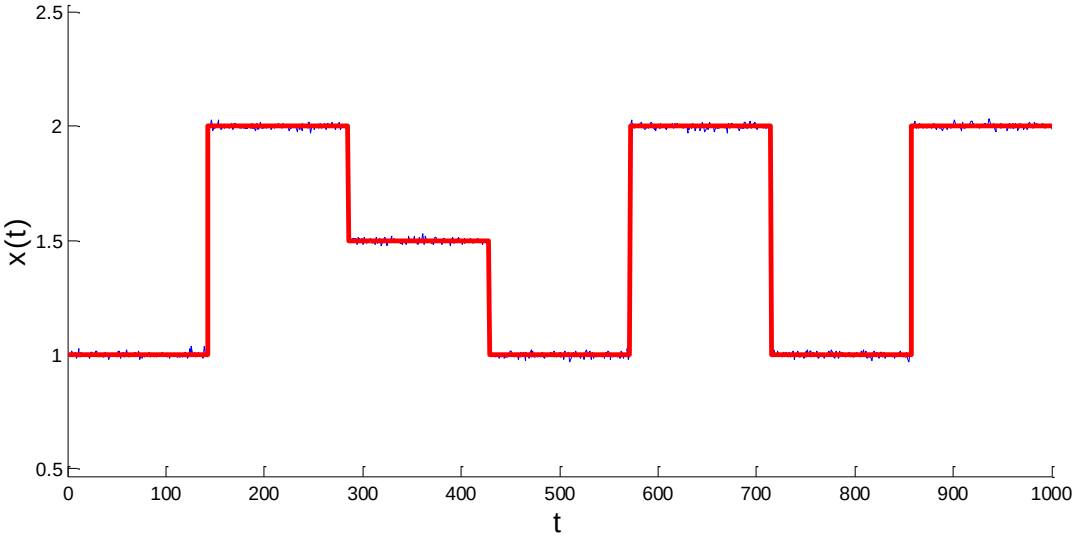
solve $\theta_{it} = \arg\min_{\theta} L(\theta, \Gamma_{it})$ *(with fixed* Γ_{it}*)*

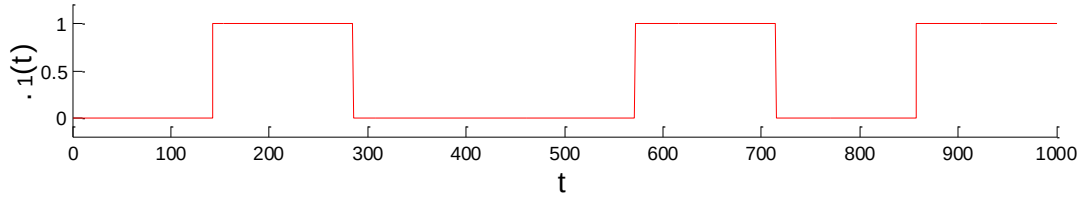
solve $\Gamma_{it} = \arg\min_{\Gamma \in \Omega_\Gamma} L(\theta_{it}, \Gamma)$ *(with fixed* θ_{it}*)*

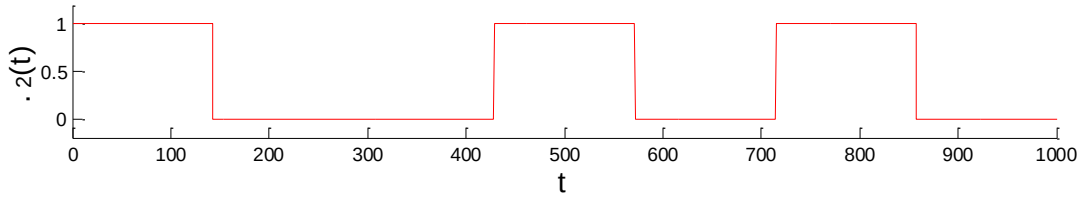
$it = it + 1$

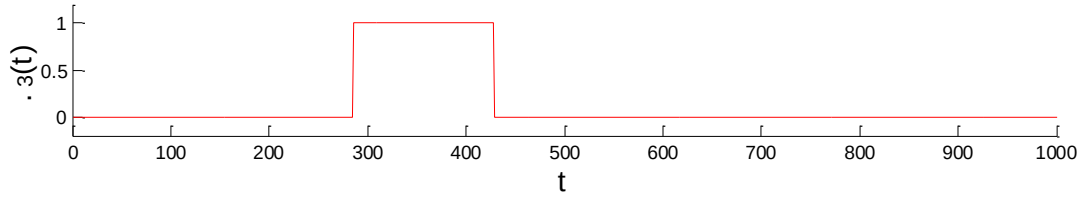
endwhile

K-means algorithm for time-series problem

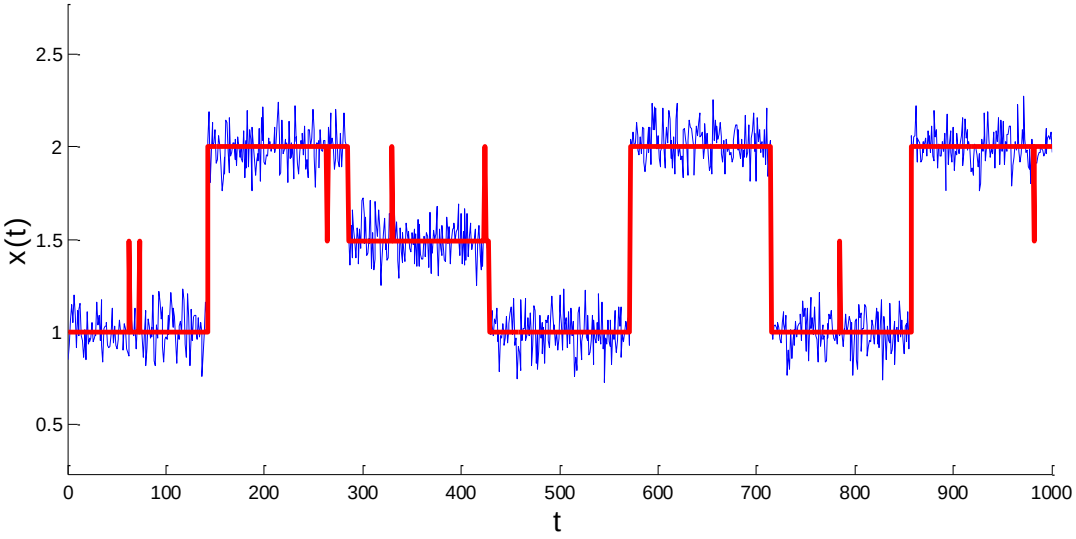


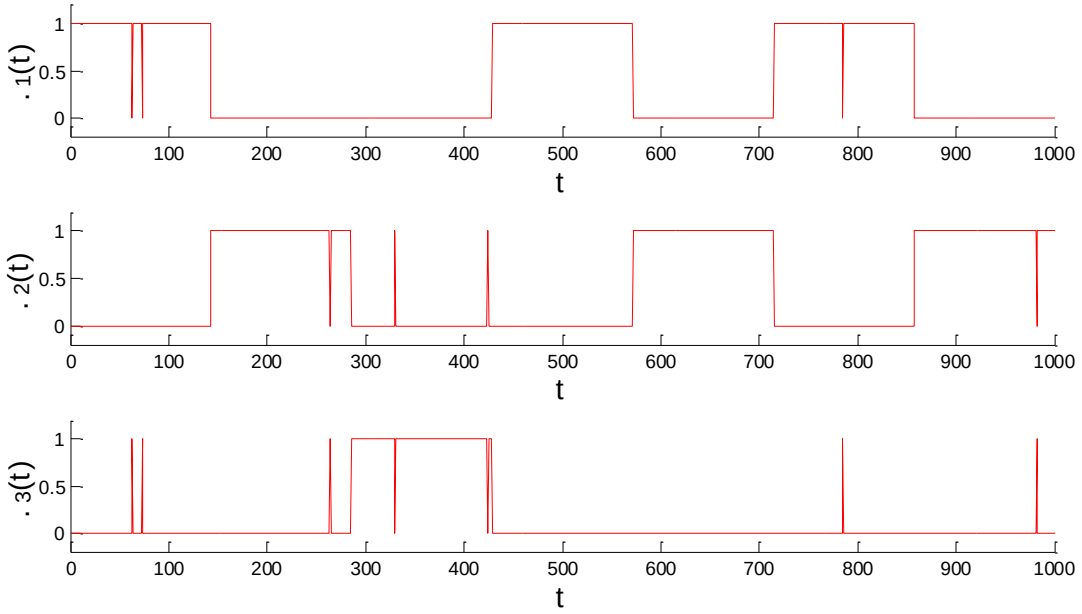




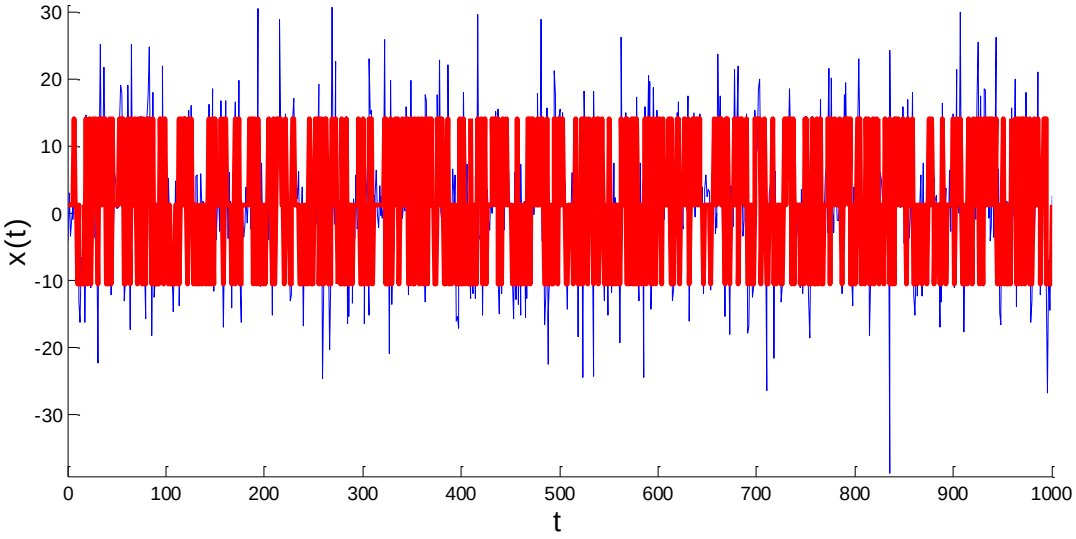


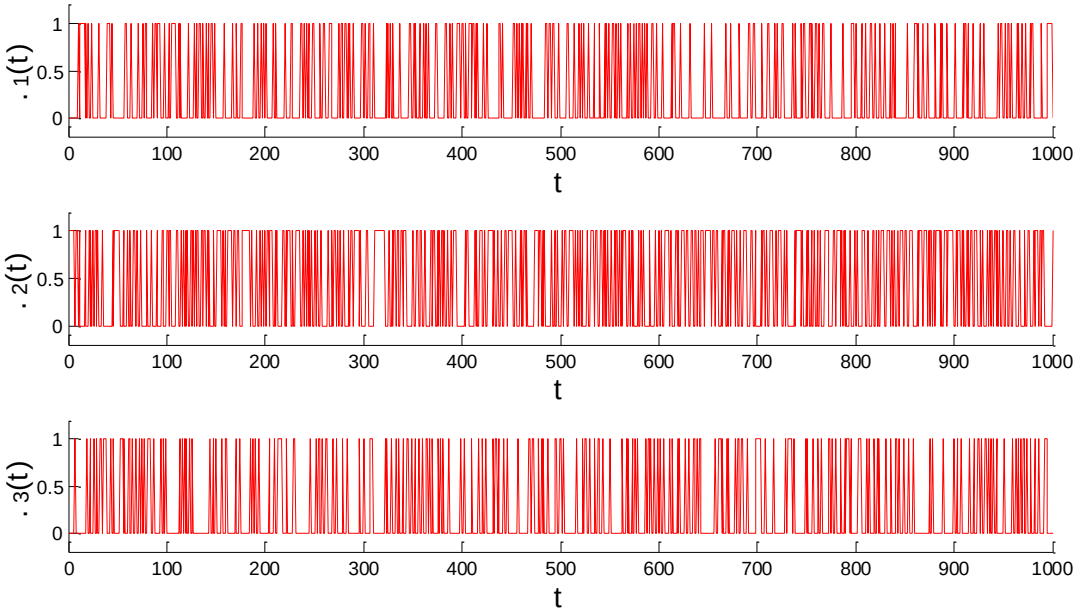
K-means algorithm for time-series problem





K-means algorithm for time-series problem





K-means algorithm for time-series problem

- K-means ignores (doesn't take into account) the time

- we want to implement "rationality" of the cluster changes during time
  (i.e. the time series is not jumping between clusters in crazy way)

$$\|\text{change of } \gamma_k(x_t) \text{ during } t\| \quad \leq \quad \|\text{change of } x_t \text{ during } t\|$$

caused by the change of cluster (model)

typically caused by the noise

K-means algorithm for time-series problem

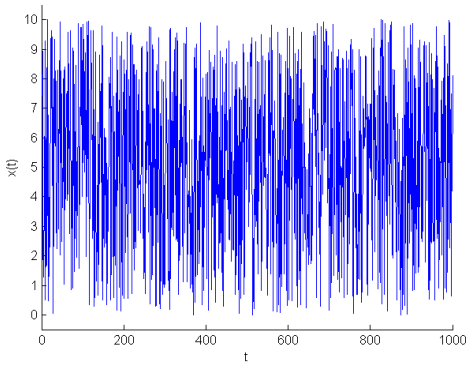
Assumption: $\text{smooth } \gamma_k(t) \quad \Rightarrow \quad \text{regularisation}$

- enforce the persistency of underlying hidden switching process

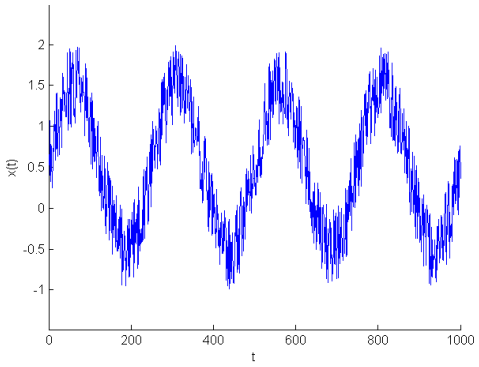
$$\|\gamma_k\|_{\mathcal{H}_1} := \sqrt{\sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2} \leq C_k \qquad \text{FEM-H1}$$

$$\|\gamma_k\|_{BV} := \sum_{t=1}^{T-1} |\gamma_k(t+1) - \gamma_k(t)| \leq C_k \qquad \text{FEM-BV}$$

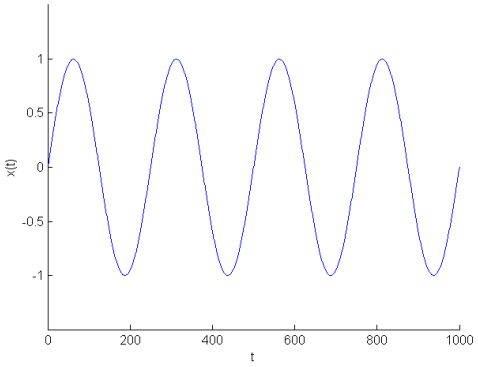
Examples:



$\|x\|_{BV} = 3.4290 \cdot 10^3$
$\|x\|_{H1} = 1.7453 \cdot 10^4$



$\|x\|_{BV} = 319.2523$
$\|x\|_{H1} = 153.7537$



$\|x\|_{BV} = 15.9736$
$\|x\|_{H1} = 0.3152$

$$[\Theta, \Gamma(t)] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \theta_k\|^2 + \varepsilon^2 \sum_{k=1}^{K} \sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$

From non-stationaty models to clustering

$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t,\theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K}\sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$

$$\text{s.t.} \;\; \gamma_k(t) \in \{0,1\}, \;\; \forall t : \sum_{k=1}^{K} \gamma_k(t) = 1$$

Assumption: continuous real-valued functions $\gamma_k(t) \in [0,1]$ (probabilities)

$$0 \le \gamma_k(t) \le 1, \quad \forall t : \sum_{k=1}^{K} \gamma_k(t) = 1$$

- *Horenko I.:* Finite Element Approach to Clustering of Multidimensional Time Series*, SIAM J. Sci. Comp. 32(1), 62-83 (2010)*

Optimization: solving the problem

$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \underbrace{\sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t,\theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K}\sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2}_{=: L(\Theta,\Gamma)} \quad \text{s.t. } 0 \le \gamma_k(t) \le 1,\ \forall t : \sum_{k=1}^{K} \gamma_k(t) = 1$$

while $\|L(\Theta_{it}, \Gamma_{it}) - L(\Theta_{it-1}, \Gamma_{it-1})\| > \varepsilon_L$

$\quad$ *solve* $\Theta_{it} = \arg\min_{\Theta} L(\Theta, \Gamma_{it})$ *(with fixed* Γ_{it}*)*

$\quad$ *solve* $\Gamma_{it} = \arg\min_{\Gamma} L(\Theta_{it}, \Gamma)$ *(with fixed* Θ_{it}*)*

$\quad$ $it = it + 1$

endwhile

depends on model

solve QP problem

$$\Gamma^* = \arg\min_{\gamma} \gamma^T H \gamma + g^T \gamma$$

$$\text{subject to } B\gamma = c \text{ and } \gamma \ge 0$$

Spectral Projected Gradient method for QP

Given cost function $f : \mathbb{R}^n \to \mathbb{R}$, *initial approximation* $x^0 \in \Omega$, *projection onto feasible set* $P_\Omega(x)$, *parameters* $m \in \mathbb{N}, \tau \in (0,1)$, *safeguarding parameters* $\sigma_1, \sigma_2 \in \mathbb{R} : 0 < \sigma_1 < \sigma_2 < 1$, *precision* $\varepsilon > 0$, *and initial step-size* $\alpha_0 > 0$.

$k := 0$
$g^0 := Ax^0 - b$
$f^0 := 1/2\langle g^0 - b, x^0\rangle$

for $k = 0, 1, \ldots$
$d^k := P_\Omega(x^k - \alpha_k g^k) - x^k$

compute matrix-vector multiplication Ad^k
compute multiple dot-product $\langle d^k, \{d^k, Ad^k, g^k\}\rangle$

if $\sqrt{\langle d^k, d^k\rangle} \leq \varepsilon$ *then stop.*

$f_{\max} := \max\{f(x^{k-j}) : 0 \leq j \leq \min\{k, m-1\}\}$
$\xi := (f_{\max} - f^k)/\langle d^k, Ad^k\rangle$
$\bar{\beta} := -\langle g^k, d^k\rangle/\langle d^k, Ad^k\rangle$
$\hat{\beta} := \tau\bar{\beta} + \sqrt{\tau^2\bar{\beta}^2 + 2\xi}$
choose $\beta_k \in \langle \sigma_1, \min\{\sigma_2, \hat{\beta}\}\rangle$

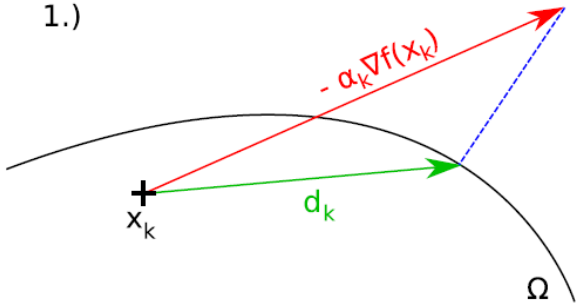
$x^{k+1} := x^k + \beta_k d^k$
$g^{k+1} := g^k + \beta_k Ad^k$
$f^{k+1} := f^k + \beta_k\langle d^k, g^k\rangle + \frac{1}{2}\beta_k^2\langle d^k, Ad^k\rangle$

$\alpha_{k+1} := \langle d^k, d^k\rangle/\langle d^k, Ad^k\rangle$

$k := k + 1$
endwhile

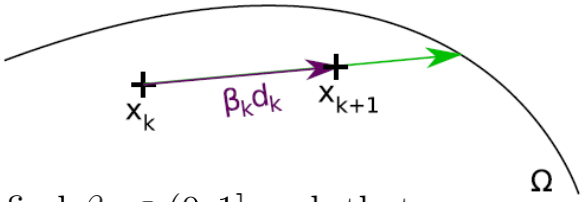
Return approximation of solution x^k.

1.)



α_k - Barzilai-Borwein step-length

2.)



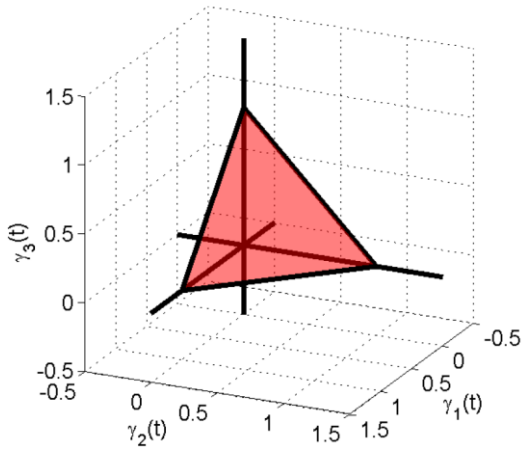
find $\beta_k \in (0, 1]$ such that

$$f(x^k + \beta_k d^k) \leq f_{\max} + \gamma\beta_k\langle \nabla f(x^k), d^k\rangle$$

- *Birgin E.G., Martínez J.M., Raydan M.:* Nonmonotone spectral projected gradient methods on convex sets, *(2000)*
- *Birgin E.G., Raydan M., Martínez J.M.:* Spectral Projected Gradient Methods: Review and Perspectives, *(2014)*
- *Pospíšil L.:* Development of Algorithms for Solving Minimizing Problems with Convex Quadratic Function on Special Convex Sets and Applications, *PhD thesis, supervised by Z. Dostál (2015)*

Feasible set – separable simplexes

$$0 \leq \gamma_k(t) \leq 1, \quad \forall t : \sum_{k=1}^{K} \gamma_k(t) = 1$$



T separable simplexes (hypertriangles) = fully parallel projection

$$B\gamma = c \text{ and } \gamma \geq 0$$

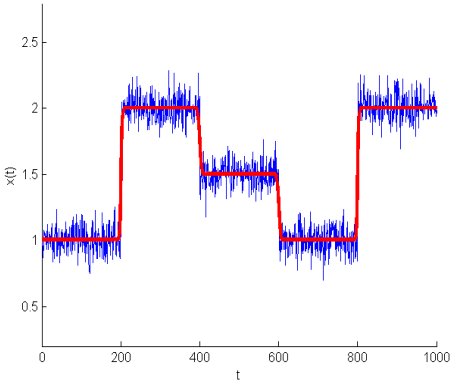
$$B = \begin{bmatrix} 1 & & & 1 & & & 1 & & \\ & 1 & & & 1 & & & 1 & \\ & & \ddots & & & \ddots & & & \ddots \\ & & 1 & & & 1 & & & 1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \} T$$

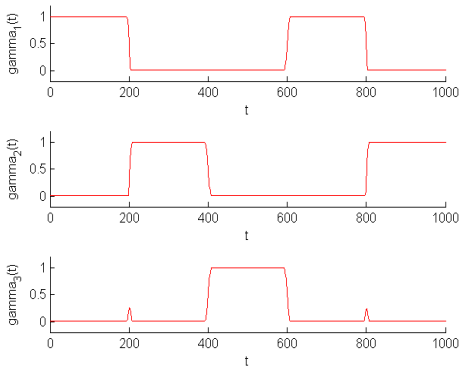
$$\underbrace{\underbrace{\quad}_{T}\ \underbrace{\quad}_{T}\ \underbrace{\quad}_{T}}_{KT}$$

Regularization parameter

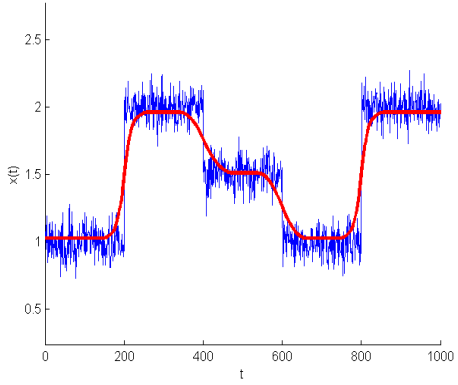
$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t, \theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K} \sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$

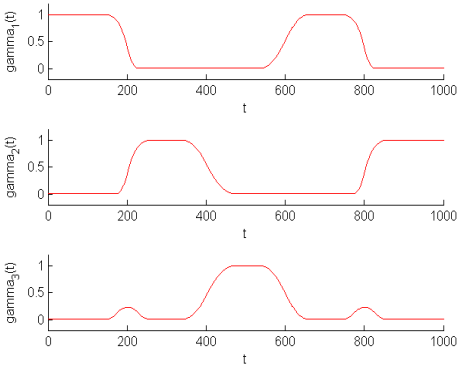
$\varepsilon^2 = 10$





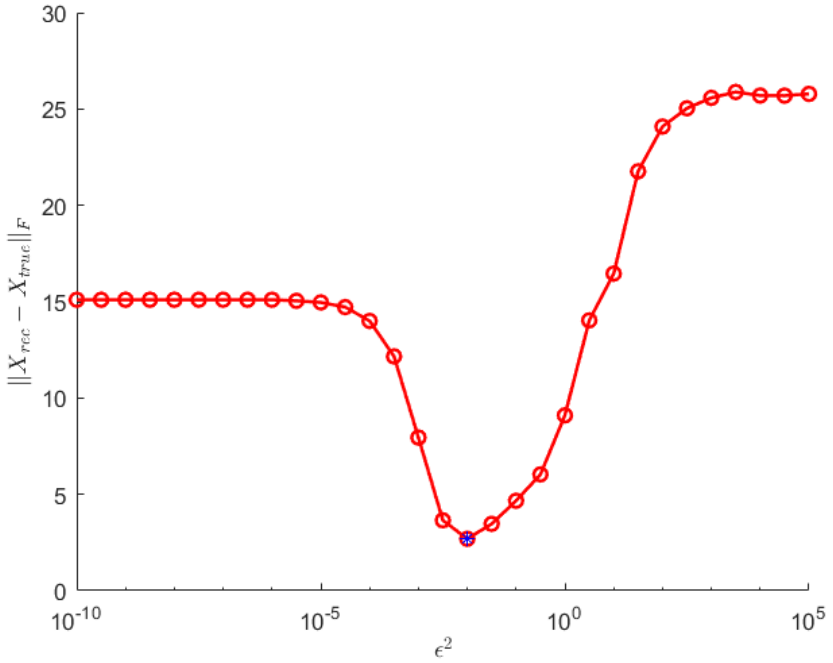
$\varepsilon^2 = 500$





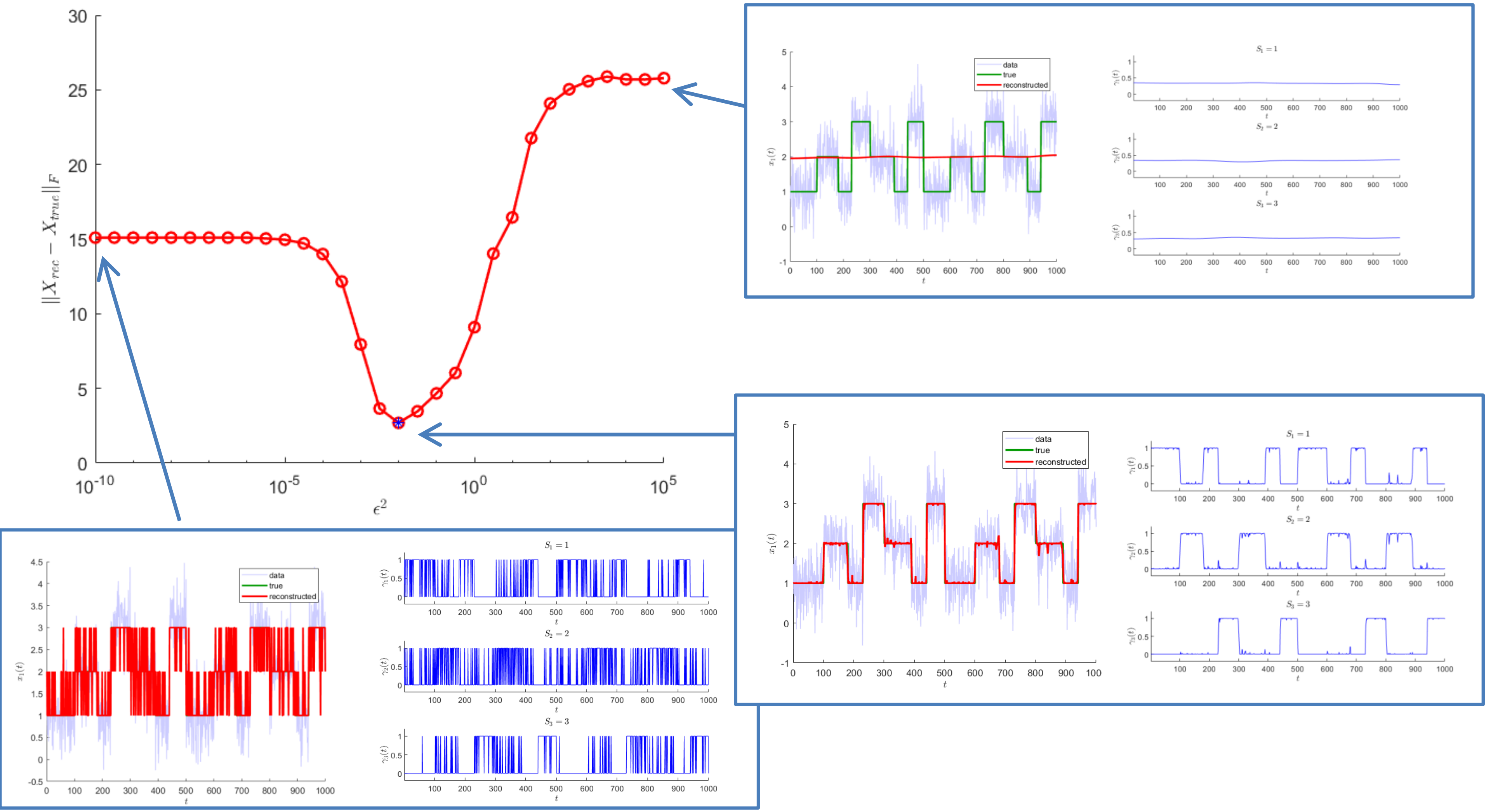
Regularization parameter: error curve

$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t,\theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K}\sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$



Regularization parameter: error curve

$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t, \theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K} \sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$

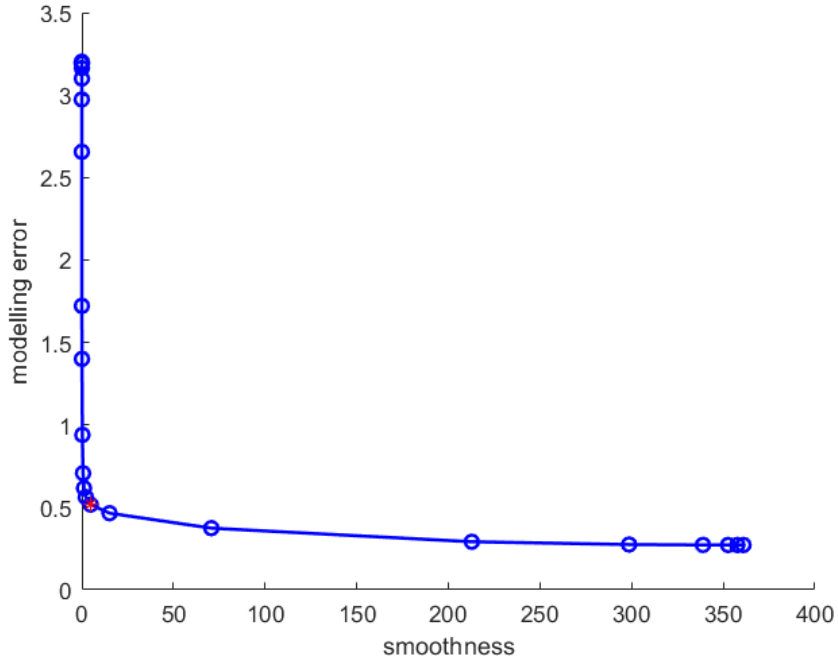


Regularization parameter: L-curve

$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t,\theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K}\sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$

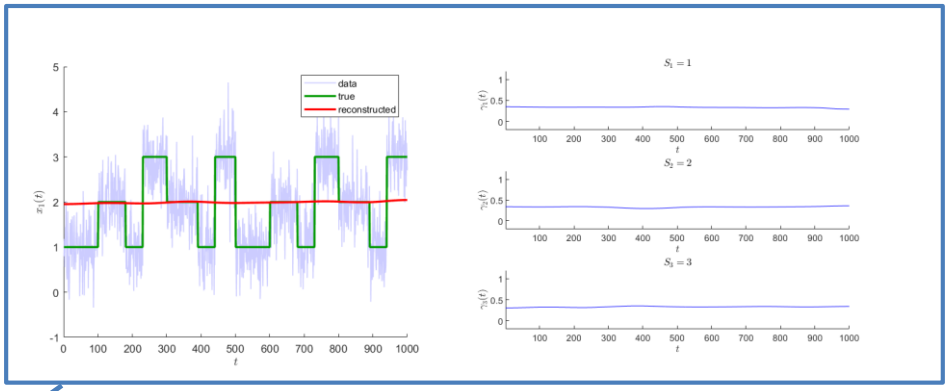
linear term = modeling error

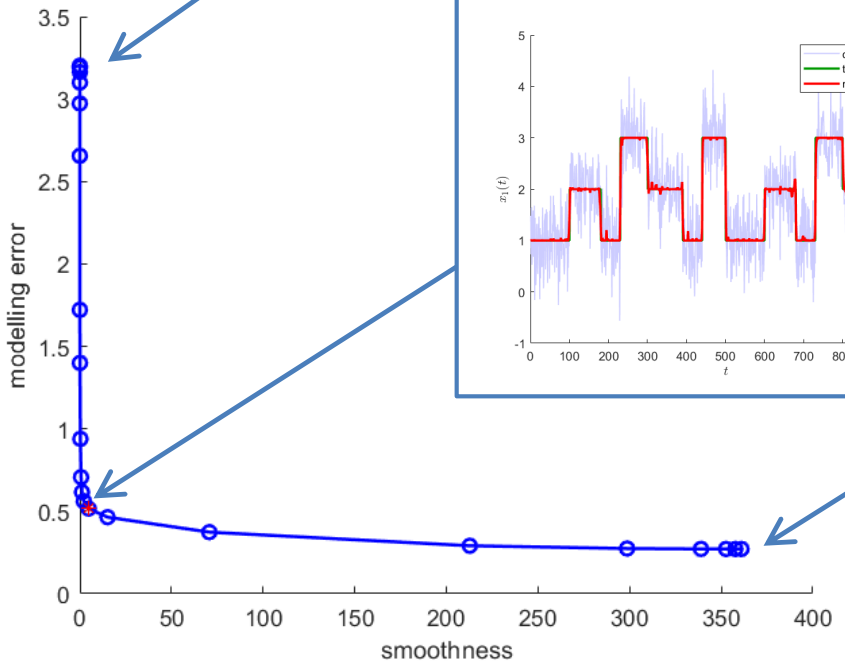
quadratic term = (non)smoothness

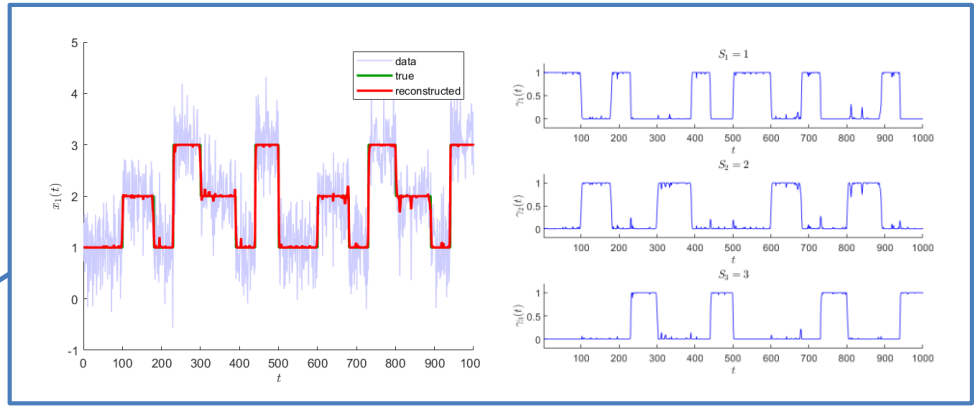


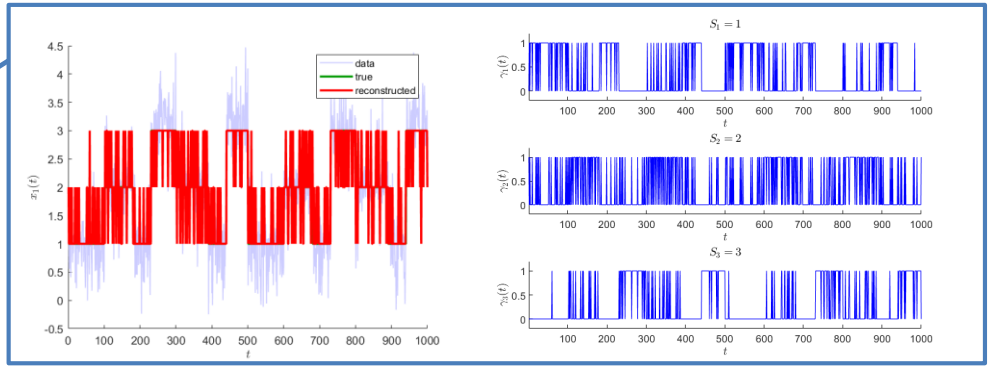
- *Hansen P.C. and D. P. O'Leary D.P.*: The use of the L-curve in the regularization of discrete ill-posed problems, *SIAM J. Sci. Comp. 14, (1993)*

Regularization parameter: L-curve



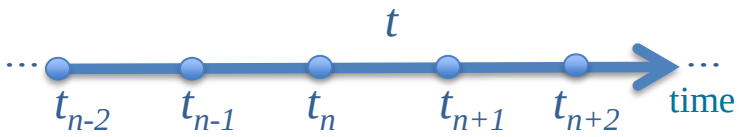






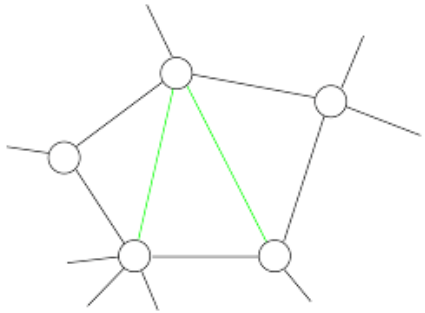
Possible extension: spatial regularization

time-persistent/metastable models

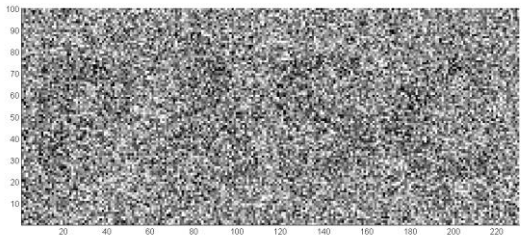


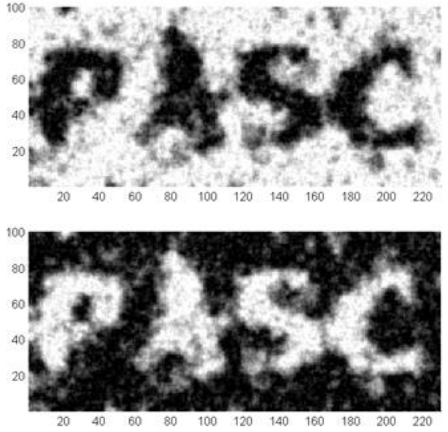
Regularization: parameter values of θ*(t)* should be similar if they are close in time

graph-persistent/metastable models



Regularization: parameter values of θ*(s)* should be similar if they are close on the graph





- *Gerber S., Horenko I.: Improving Clustering by Imposing Network Information, Sciences Advances (AAAS), 1(7):e1500163, (2015)*

Possible extension: general non-stationary models

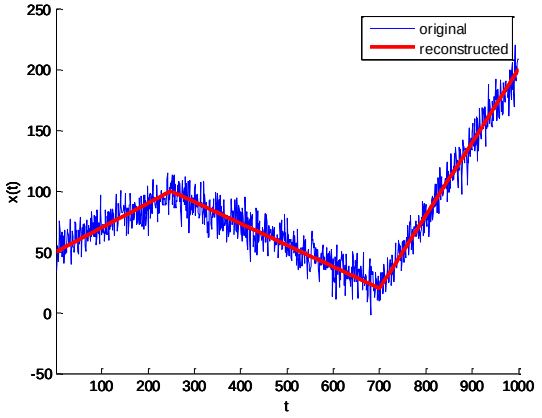
$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t, \theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K} \sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$

Piecewise linear regression

$$x(t) = \begin{cases} a_{0,1} + a_{1,1}t + \varepsilon(t) & \text{for } x \in \mathcal{T}_1 \\ a_{0,2} + a_{1,2}t + \varepsilon(t) & \text{for } x \in \mathcal{T}_2 \\ a_{0,3} + a_{1,3}t + \varepsilon(t) & \text{for } x \in \mathcal{T}_3 \end{cases}$$

$$error = \|x(t) - \hat{x}(t)\| = 307.1922$$

$$AIC_K = 2 \cdot 6 - 2 \cdot \log(307.1922) = 0.5451$$

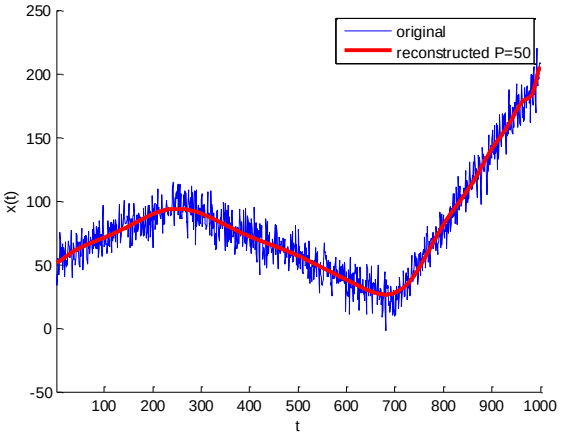


To compare: Polynomial regression

$$x(t) = a_0 + a_1 t + a_2 t^2 + \ldots a_P t^P + \varepsilon(t)$$

$$error = \|x(t) - \hat{x}(t)\| = 309.4182$$

$$AIC_P = 2 \cdot 51 - 2 \cdot \log(309.4182) = 90.5306$$



Possible extension: general non-stationary models

$$[\bar{\Theta}, \bar{\Gamma}] = \arg\min_{\Theta,\Gamma} \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \cdot \|x_t - \mu(t,\theta_k)\|^2 + \varepsilon^2 \sum_{k=1}^{K}\sum_{t=1}^{T-1} (\gamma_k(t+1) - \gamma_k(t))^2$$

TV-Entropy (regime-based non-stationary Entropy)

$$\max_{f_t(x),\forall t} \left\{ \mathop{\mathbb{E}}_{t}\Big[H\Big[f_t(x)\Big]\Big] = \mathop{\mathbb{E}}_{t}\Big[-\int_{\mathcal{X}} f_t(x)\ln f_t(x)\,dx\Big] \right\}$$

s.t. $$\int_{\mathcal{X}} x^j f_t(x)\,dx = \mu_j(t) \qquad \forall j \in \{0,\dots,k\},$$

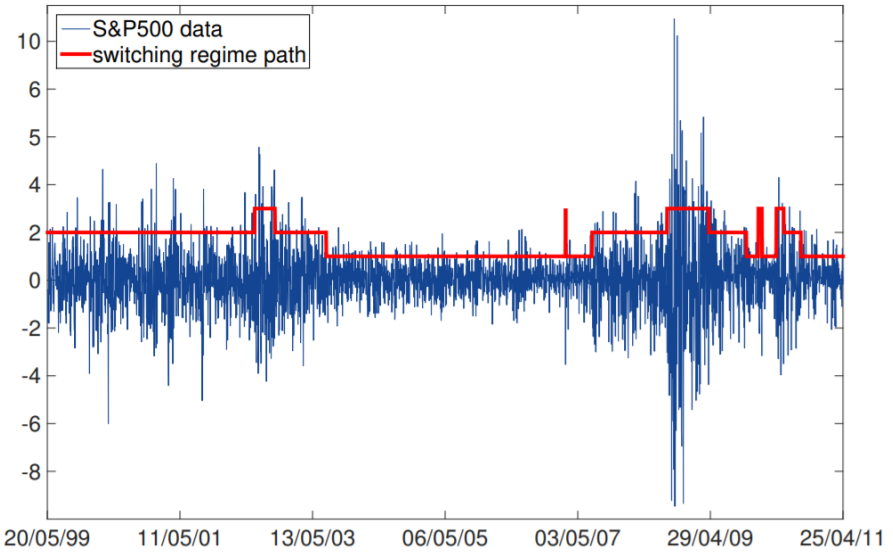
$$f_t^*(x) = \exp\Big[-\sum_{i=0}^{k} \lambda_j(t) x^j\Big]$$

$$\lambda(t) = \sum_{i=1}^{K} \gamma_i(t)\lambda^{(i)},$$

$$\gamma_i(t) \in \{0,1\}, \sum_{i=1}^{K} \gamma_i(t) = 1.$$

TV-Entropy identifies memoryless models that are simpler and better then the state-of-the-art for all of the financial benchmark data considered.

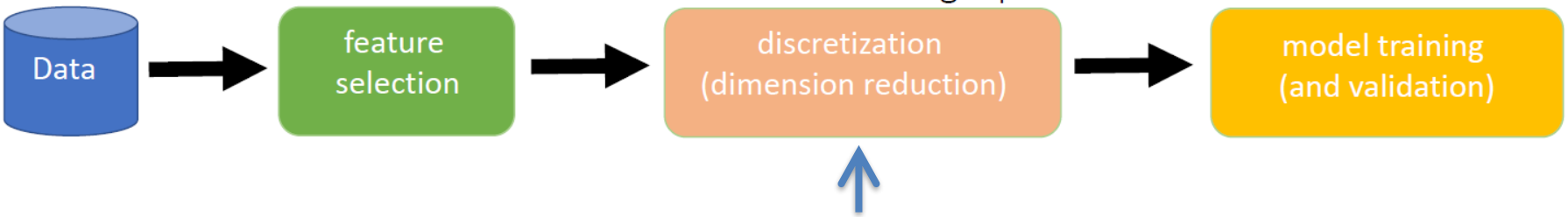
- *Marchenko G., Gagliardini P., Horenko I.:* Towards a Computationally Tractable Maximum Entropy Principle for Nonstationary Financial Time Series*, SIAM J. Finan. Math., 9(4), 1249–1285, (2018)*



Scalable Probabilistic Approximation (SPA)

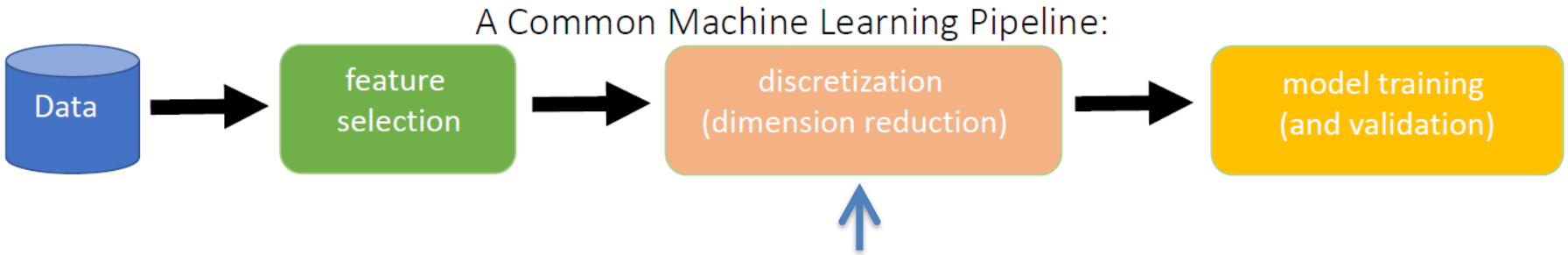
Reduction of the data – K-means revisited

A Common Machine Learning Pipeline:



$$x_t \in \mathbb{R}^n, t = 1, \ldots, T \quad \approx \quad S_k \in \mathbb{R}^n, k = 1, \ldots, K \qquad K \ll T$$

Reduction of the data – K-means revisited



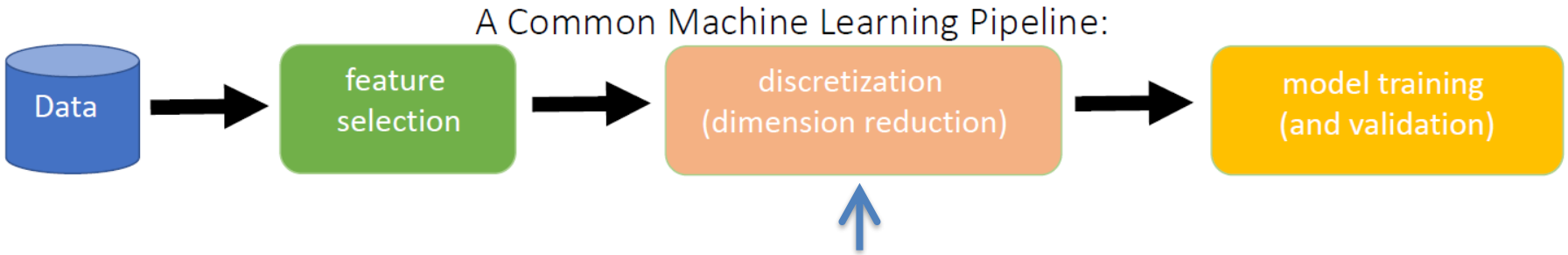
$$x_t \in \mathbb{R}^n, t = 1, \ldots, T \quad \approx \quad S_k \in \mathbb{R}^n, k = 1, \ldots, K \qquad K \ll T$$

K-means?

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \Gamma_{k,t} \|x_t - S_k\|_2^2$$

$\Gamma_{k,t}$ - probability for the system to be in particular state S_k at the instance t
(K-means: = probability of x_t to belong to k-th cluster)

Reduction of the data – K-means revisited



$$x_t \in \mathbb{R}^n, t = 1, \ldots, T \quad \approx \quad S_k \in \mathbb{R}^n, k = 1, \ldots, K \qquad K \ll T$$

K-means?

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \Gamma_{k,t} \|x_t - S_k\|_2^2$$

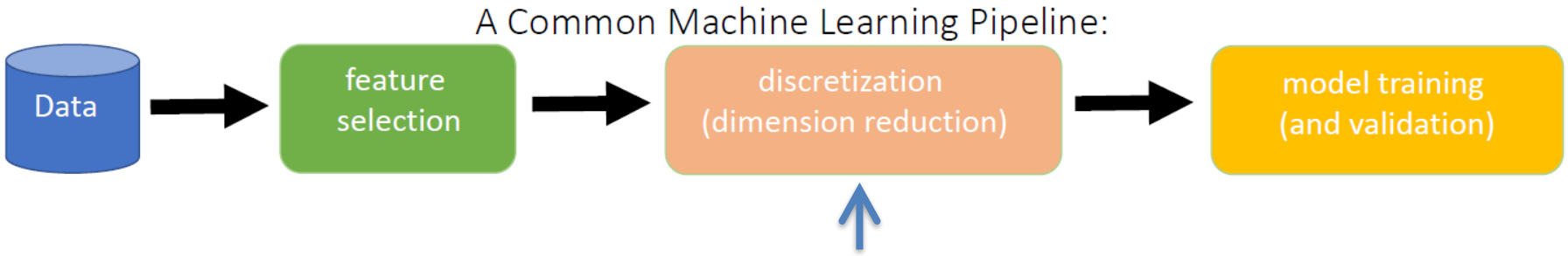
$\Gamma_{k,t}$ - probability for the system to be in particular state S_k at the instance t

~~(K-means: = probability of x_t to belong to k-th cluster)~~

(K-means: = probability of observing error $\|x_t - S_k\|_2^2$)

$$x_t = \sum_{k=1}^{K} \Gamma_{k,t} S_k$$

Reduction of the data – K-means revisited



$$x_t \in \mathbb{R}^n, t = 1, \ldots, T \quad \approx \quad S_k \in \mathbb{R}^n, k = 1, \ldots, K \qquad K \ll T$$

K-means?

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \sum_{k=1}^{K} \Gamma_{k,t} \|x_t - S_k\|_2^2$$

$\Gamma_{k,t}$ - probability for the system to be in particular state S_k at the instance t

~~(K-means: = probability of x_t to belong to k-th cluster)~~

~~(K-means: = probability of observing error $\|x_t \quad S_k\|_2^2$)~~

$$x_t = \sum_{k=1}^{K} \Gamma_{k,t} S_k$$

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \|x_t - \sum_{k=1}^{K} \Gamma_{k,t} S_k\|_2^2$$

- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, accepted in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

Scalable Probabilistic Approximation (SPA)

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \|x_t - \sum_{k=1}^{K} \Gamma_{k,t} S_k\|_2^2 \qquad \text{s.t.} \quad \forall t : \sum_{k=1}^{K} \Gamma_{k,t} = 1, \ \forall k : \Gamma_{k,t} \in [0,1]$$

Jensen inequality: $$\leq \sum_{t=1}^{T} \sum_{k=1}^{K} \Gamma_{k,t} \|x_t - S_k\|_2^2$$

(K-means is suboptimal)

$$\Gamma_{k,t} \in \{0,1\} : \qquad = \sum_{t=1}^{T} \sum_{k=1}^{K} \Gamma_{k,t} \|x_t - S_k\|_2^2$$

- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, accepted in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

Scalable Probabilistic Approximation (SPA)

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \|x_t - \sum_{k=1}^{K} \Gamma_{k,t} S_k\|_2^2 \qquad \text{s.t.} \quad \forall t : \sum_{k=1}^{K} \Gamma_{k,t} = 1, \ \forall k : \Gamma_{k,t} \in [0,1]$$

General SPA:

$$[S^*, \Gamma^*] := \arg\min_{\Gamma \in \Omega_\Gamma} L(S, \Gamma)$$

$$L(S, \Gamma) := \sum_{t=1}^{T} \mathrm{dist}_S(X(t), \Gamma(t)) + \varepsilon_S^2 \Phi_S(S) + \varepsilon_\Gamma^2 \Phi_\Gamma(\Gamma)$$

Set a feasible initial approximation $\Gamma^0 \in \Omega_\Gamma$

while $\|L(S^k, \Gamma^k) - L(S^{k-1}, \Gamma^{k-1})\| \geq \varepsilon$

solve $S^k = \arg\min_S L(S, \Gamma^{k-1})$ *(with fixed* Γ^{k-1}*)*

solve $\Gamma^k = \arg\min_{\Gamma \in \Omega_\Gamma} L(S^k, \Gamma)$ *(with fixed* S^k*)*

$k = k + 1$

endwhile

Return an approximation of the data representation vectors S^k *and an approximation of cluster affiliation probability vectors* Γ^k.

- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, accepted in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

Scalable Probabilistic Approximation (SPA)

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \|x_t - \sum_{k=1}^{K} \Gamma_{k,t} S_k\|_2^2 \qquad \text{s.t.} \quad \forall t : \sum_{k=1}^{K} \Gamma_{k,t} = 1, \ \forall k : \Gamma_{k,t} \in [0,1]$$

$$= \|X - S\Gamma\|_F^2$$

$$S^* = X\Gamma^T \left(\Gamma\Gamma^T\right)^+ + \alpha^T R^T, \quad \textit{with parameter } \alpha \in \mathbb{R}^{r,n},$$

$$\gamma_t^* := \arg\min_{\gamma \in \Omega_\gamma} \frac{1}{2}\gamma^T A\gamma - b_t^T\gamma,$$

$$\Omega_\gamma := \{\gamma \in \mathbb{R}^K \mid B\gamma = c, \gamma \geq 0\}$$

$$A := 2S^TS, \ b_t := S^Tx_t, \ B := \mathbb{1}_K^T, \ c := 1$$

Set a feasible initial approximation $\Gamma^0 \in \Omega_\Gamma$

while $\|L(S^k, \Gamma^k) - L(S^{k-1}, \Gamma^{k-1})\| \geq \varepsilon$

solve $S^k = \arg\min_S L(S, \Gamma^{k-1})$ *(with fixed* Γ^{k-1}*)*

solve $\Gamma^k = \arg\min_{\Gamma \in \Omega_\Gamma} L(S^k, \Gamma)$ *(with fixed* S^k*)*

$k = k + 1$

endwhile

Return an approximation of the data representation vectors S^k *and an approximation of cluster affiliation probability vectors* Γ^k.

- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, accepted in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

Scalable Probabilistic Approximation (SPA)

$$[S^*, \Gamma^*] = \arg\min_{S,\Gamma} \sum_{t=1}^{T} \|x_t - \sum_{k=1}^{K} \Gamma_{k,t} S_k\|_2^2 \qquad \text{s.t.} \quad \forall t : \sum_{k=1}^{K} \Gamma_{k,t} = 1, \ \forall k : \Gamma_{k,t} \in [0,1]$$

$$= \|X - S\Gamma\|_F^2$$

$$S^* = X\Gamma^T \left(\Gamma\Gamma^T\right)^+ + \alpha^T R^T, \ \textit{with parameter } \alpha \in \mathbb{R}^{r,n},$$

Introducing regularization:

$$\Phi_S(S) := \frac{1}{nK(K-1)} \sum_{i=1}^{n} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} (S_{i,k_1} - S_{i,k_2})^2$$

$$S^* = X\Gamma^T H_\epsilon^{-1}, H_\varepsilon := \Gamma\Gamma^T + \frac{2\varepsilon^2}{nK(K-1)}(KI_{K,K} - \mathbb{1}_{K,K})$$

- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, accepted in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

SPA in action

Exact Law of the Total Probability in a matrix notation:

$\pi_Y(t)$ = Λ $\pi_X(t)$

$\pi_Y(t)$	Λ				$\pi_X(t)$
$P[Y^t=y_1]$	$P[Y^t=y_1 \mid X^t=x_1]$	$P[Y^t=y_1 \mid X^t=x_2]$	…	$P[Y^t=y_1 \mid X^t=x_n]$	$P[X^t=x_1]$
$P[Y^t=y_2]$	$P[Y^t=y_2 \mid X^t=x_1]$	$P[Y^t=y_2 \mid X^t=x_2]$	…	$P[Y^t=y_2 \mid X^t=x_n]$	$P[X^t=x_2]$
…	…	…	…	…	…
$P[Y^t=y_m]$	$P[Y^t=y_m \mid X^t=x_1]$	$P[Y^t=y_m \mid X^t=x_2]$	…	$P[Y^t=y_m \mid X^t=x_n]$	$P[X^t=x_n]$

SPA in action

Exact Law of the Total Probability in a matrix notation:

$\pi_Y(t)$ | Λ | $\pi_X(t)$

$P[Y^t=y_1]$		$P[Y^t=y_1 \mid X^t=x_1]$	$P[Y^t=y_1 \mid X^t=x_2]$	…	$P[Y^t=y_1 \mid X^t=x_n]$	$P[X^t=x_1]$
$P[Y^t=y_2]$		$P[Y^t=y_2 \mid X^t=x_1]$	$P[Y^t=y_2 \mid X^t=x_2]$	…	$P[Y^t=y_2 \mid X^t=x_n]$	$P[X^t=x_2]$
…	=	…	…	…	…	…
$P[Y^t=y_m]$		$P[Y^t=y_m \mid X^t=x_1]$	$P[Y^t=y_m \mid X^t=x_2]$	…	$P[Y^t=y_m \mid X^t=x_n]$	$P[X^t=x_n]$

$\Gamma_{k,t}$ - probability for the system to be in particular state S_k at the instance t $\quad \Gamma_{k,t} := P(x_t = S_k)$

$$\forall t : \Gamma^Y_{:,t} = \Lambda \Gamma^X_{:,t}$$

Markov model: $\quad Y(t) := X(t+1)$

SPA in action

Exact Law of the Total Probability in a matrix notation:

$\pi_Y(t)$ | Λ | $\pi_X(t)$

$\pi_Y(t)$		Λ			$\pi_X(t)$
$P[Y^t=y_1]$	=	$P[Y^t=y_1 \mid X^t=x_1]$	$P[Y^t=y_1 \mid X^t=x_2]$	…	$P[Y^t=y_1 \mid X^t=x_n]$	$P[X^t=x_1]$
$P[Y^t=y_2]$		$P[Y^t=y_2 \mid X^t=x_1]$	$P[Y^t=y_2 \mid X^t=x_2]$	…	$P[Y^t=y_2 \mid X^t=x_n]$	$P[X^t=x_2]$
…		…	…	…	…	…
$P[Y^t=y_m]$		$P[Y^t=y_m \mid X^t=x_1]$	$P[Y^t=y_m \mid X^t=x_2]$	…	$P[Y^t=y_m \mid X^t=x_n]$	$P[X^t=x_n]$

$\Gamma_{k,t}$ - probability for the system to be in particular state S_k at the instance t $\quad \Gamma_{k,t} := P(x_t = S_k)$

$$Y(t) \in \mathbb{R}^m, t = 1, \ldots, T$$
$$S_k^Y \in \mathbb{R}^m, k = 1, \ldots, K_Y$$
$$\|Y - S^Y \Gamma^Y\|_F \rightarrow \min$$

$$\forall t : \Gamma^Y_{:,t} = \Lambda \Gamma^X_{:,t}$$

$$X(t) \in \mathbb{R}^n, t = 1, \ldots, T$$
$$S_k^X \in \mathbb{R}^n, k = 1, \ldots, K_X$$
$$\|X - S^X \Gamma^X\|_F \rightarrow \min$$

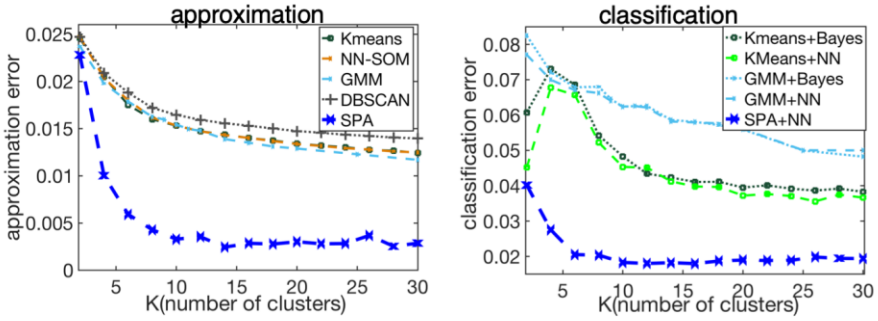
Markov model: $\quad Y(t) := X(t+1)$

Results

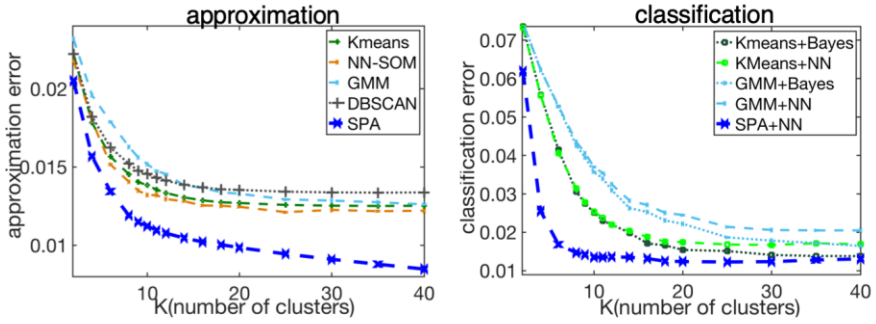
- X(t) is continuous (and real valued) set of collected 32 image features
- index t denotes patients and goes from 1 to 569
- Y(t) is binary: ‘benign’ or ‘malignant’

- X(t) contains genetic expression levels for 25’000 genes
- index t goes from 1 to 300 (there are 300 single cell probes),
- Y(t) is a label denoting one of the 11 cell types (e.g., ‘blood cell’, ‘glia 25 cell’, etc.)

a) breast cancer diagnostics (569 patients, 32 features, WDBC data)

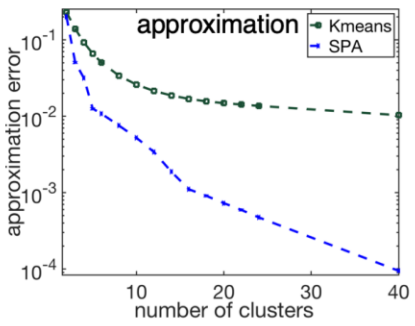


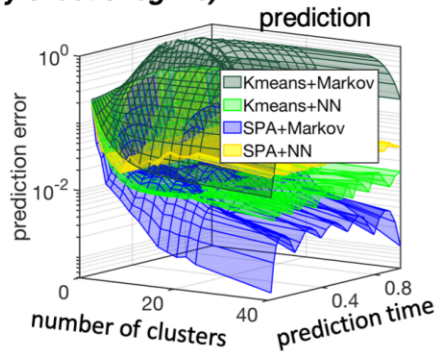
b) single cell human mRNA classification (25`000 genes, 11 cell types, 300 cell samples)



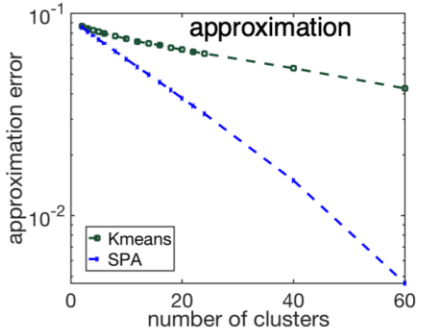
- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, accepted in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

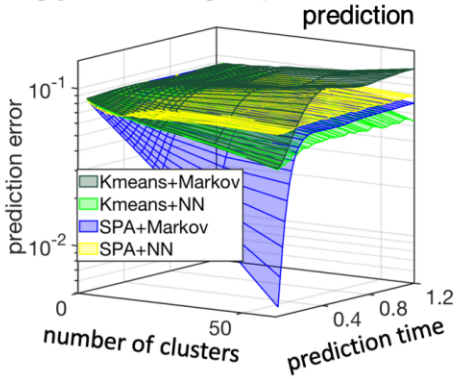
a) Lorenz-96 1D turbulence model (weakly-chaotic regime)



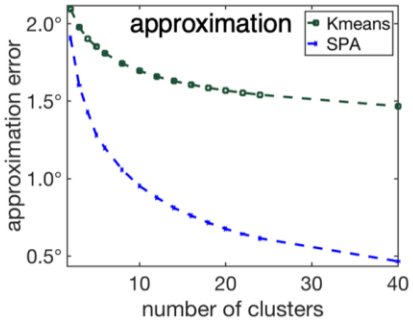


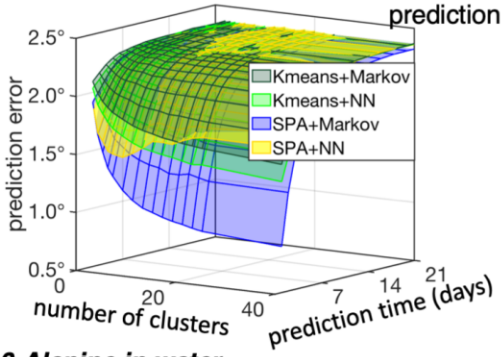
b) Lorenz-96 1D turbulence model (strongly-chaotic regime)



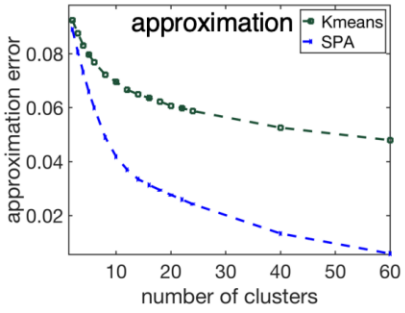


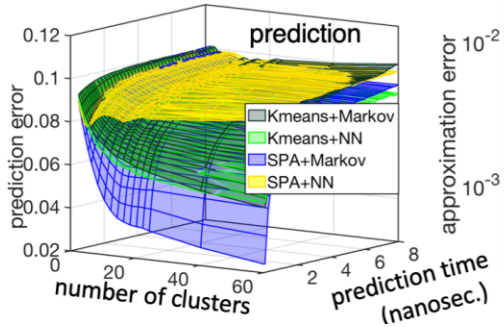
c) surface temperature dynamics over Europe (1979-2010, 20x30 grid ECMWF resimulation data)



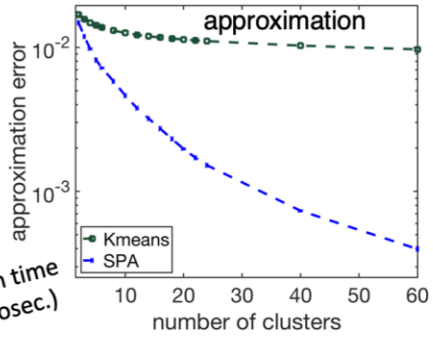


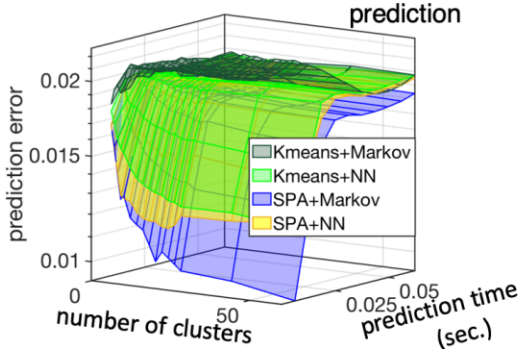
d) molecular dynamics simulation of 10-Alanine in water





e) EEG dynamics in a brain-computer interface (BCI2000 data)





- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, accepted in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

SPA in action: compressing the pipeline

Exact Law of the Total Probability in a matrix notation:

$\pi_Y(t)$ = Λ $\pi_X(t)$

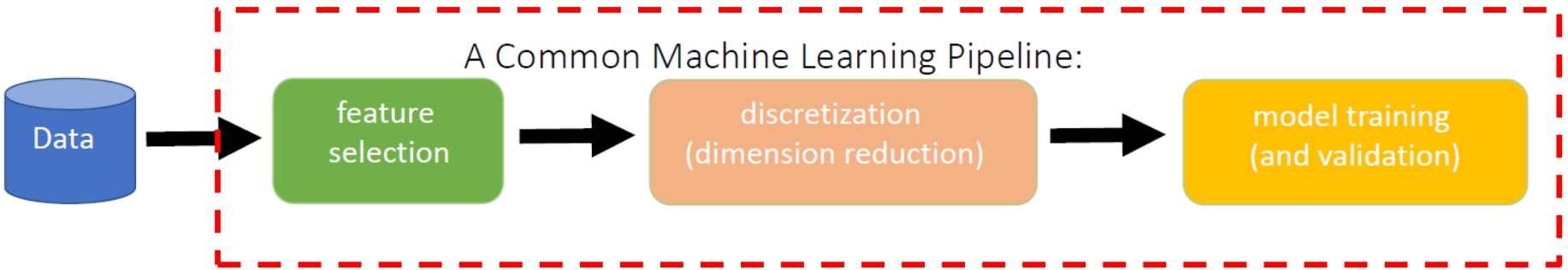
$\pi_Y(t)$	Λ				$\pi_X(t)$
$P[Y^t=y_1]$	$P[Y^t=y_1 \mid X^t=x_1]$	$P[Y^t=y_1 \mid X^t=x_2]$	…	$P[Y^t=y_1 \mid X^t=x_n]$	$P[X^t=x_1]$
$P[Y^t=y_2]$	$P[Y^t=y_2 \mid X^t=x_1]$	$P[Y^t=y_2 \mid X^t=x_2]$	…	$P[Y^t=y_2 \mid X^t=x_n]$	$P[X^t=x_2]$
…	…	…	…	…	…
$P[Y^t=y_m]$	$P[Y^t=y_m \mid X^t=x_1]$	$P[Y^t=y_m \mid X^t=x_2]$	…	$P[Y^t=y_m \mid X^t=x_n]$	$P[X^t=x_n]$

$\Gamma_{k,t}$ - probability for the system to be in particular state S_k at the instance t $\quad \Gamma_{k,t} := P(x_t = S_k)$

$$Y(t) \in \mathbb{R}^m, t = 1, \ldots, T$$
$$S_k^Y \in \mathbb{R}^m, k = 1, \ldots, K_Y$$
$$\|Y - S^Y \Gamma^Y\|_F \rightarrow \min$$

$$\forall t : \Gamma^Y_{:,t} = \Lambda \Gamma^X_{:,t}$$

$$X(t) \in \mathbb{R}^n, t = 1, \ldots, T$$
$$S_k^X \in \mathbb{R}^n, k = 1, \ldots, K_X$$
$$\|X - S^X \Gamma^X\|_F \rightarrow \min$$



SPA in action: compressing the pipeline

$$[S_x^*, \Gamma_x^*, S_y^*, \Gamma_y^*] = \arg\min_{\Gamma_x, \Gamma_y \in \Omega_\Gamma} \underbrace{\|Y - S_y\Gamma_y\|_F^2 + \varepsilon\|X - S_x\Gamma_x\|_F^2}$$

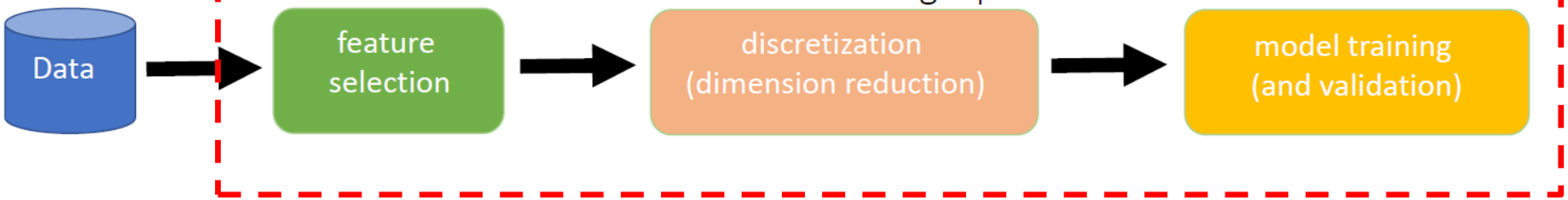
$$= \left\| \begin{bmatrix} Y \\ \varepsilon X \end{bmatrix} - \begin{bmatrix} S_y\Gamma_y \\ \varepsilon S_x\Gamma_x \end{bmatrix} \right\|_F^2 = \underbrace{\left\| \begin{bmatrix} Y \\ \varepsilon X \end{bmatrix} - \begin{bmatrix} S_y\Lambda \\ \varepsilon S_x \end{bmatrix} \Gamma_x \right\|_F^2}$$

$$= \|\hat{X} - \hat{S}\hat{\Gamma}\|_F^2$$

$Y(t) \in \mathbb{R}^m, t = 1, \ldots, T$
$S_k^Y \in \mathbb{R}^m, k = 1, \ldots, K_Y$
$\|Y - S^Y\Gamma^Y\|_F \rightarrow \min$

$\forall t : \Gamma_{:,t}^Y = \Lambda\Gamma_{:,t}^X$

$X(t) \in \mathbb{R}^n, t = 1, \ldots, T$
$S_k^X \in \mathbb{R}^n, k = 1, \ldots, K_X$
$\|X - S^X\Gamma^X\|_F \rightarrow \min$



- *Horenko I.:* Finite Element Approach to Clustering of Multidimensional Time Series*, SIAM J. Sci. Comp. 32(1), 62-83 (2010)*
- *Metzner P., Putzig L., Horenko I. :* Analysis of persistent non-stationary time series and applications*. Communications in Applied Mathematics and Computational Science, 7(2):175-229, (2012)*
- *Gerber S., Horenko I.:* Improving Clustering by Imposing Network Information*, Sciences Advances (AAAS), 1(7):e1500163, (2015)*
- *Pospisil L., Gagliardini P., Sawyer W., Horenko I.:* On a scalable nonparametric denoising of time series signals. *Communications in Applied Mathematics and Computational Science, 13(1), (2018)*
- *Marchenko G., Gagliardini P., Horenko I.:* Towards a Computationally Tractable Maximum Entropy Principle for Nonstationary Financial Time Series*, SIAM J. Finan. Math., 9(4), 1249–1285, (2018)*

- Gerber S., Pospíšil L., Navandar M., Horenko I.: *Low-cost scalable discretization, prediction and feature selection for complex systems*, almost published in Science Advances, (2019), http://www.biorxiv.org/content/10.1101/720441v1

… thank you for your attention!