veritas
iustitia
libertas

# Numerics of Stochastic Processes VI (19.12.2008)

***Illia Horenko***

Research Group “***Computational Time Series Analysis*** “
Institute of Mathematics
Freie Universität  Berlin (FU)

DFG Research Center ***MATHEON***
„Mathematics in key technologies“

# Model Distance Functional

Let $x(t) : \mathbf{R}^1 \longrightarrow \Psi \subset \mathbf{R}^n$ be the *observed process* $t \in [0, T]$

Define $\mathbf{K}$ local models by a *model distance functional*:

$$g(x, \theta_i) \quad : \quad \Psi \times \Omega \longrightarrow [0, \bar{g}], \quad 0 < \bar{g} < +\infty,$$

$$\theta_1, \ldots, \theta_{\mathbf{K}} \in \Omega \subset \mathbf{R}^d$$

***Examples***

- *Geometrical clustering*: $\theta_i \in \Psi$ - cluster centers

$$g(x, \theta_i) \quad = \quad \| x - \theta_i \|^2,$$

# Model Distance Functional

Let $x(t) : \mathbf{R}^1 \longrightarrow \Psi \subset \mathbf{R}^n$ be the *observed process* $t \in [0, T]$

Define $\mathbf{K}$ local models by a *model distance functional*:

$$g\left(x, \theta_i\right) \quad : \quad \Psi \times \Omega \longrightarrow [0, \bar{g}]\,, \quad 0 < \bar{g} < +\infty,$$

$$\theta_1, \ldots, \theta_{\mathbf{K}} \in \Omega \subset \mathbf{R}^d$$

***Examples***

- *Geometrical clustering*: $\theta_i \in \Psi$ - cluster centers

$$g\left(x, \theta_i\right) \quad = \quad \|\, x - \theta_i \,\|^2,$$

- *Gaussian clustering*: $\theta_i = (\mu_i, \Sigma_i)$ - Gaussian parameters

$$g\left(x, \theta_i\right) \quad = \quad \|\, x - \mu_i \,\|^2_{\Sigma_i^{-1}}$$

# Clustering Problem: definition

## *What is "clustering"?*

Find $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_{\mathbf{K}}(t))$ such that for each $t$ :

$$\sum_{i=1}^{\mathbf{K}} \gamma_i(t) g(x, \theta_i) \rightarrow \min_{\Gamma(t), \Theta},$$

subjected to constraints:

$$\sum_{i=1}^{\mathbf{K}} \gamma_i(t) = 1, \quad \forall t \in [0, T]$$

$$\gamma_i(t) \geq 0, \quad \forall t \in [0, T], i = 1, \ldots, \mathbf{K}$$

# Averaged Clustering Functional

Find $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_{\mathbf{K}}(t))$ such that for each $t$ :

$$\mathbf{L}(\Theta, \Gamma(t)) = \int_0^T \sum_{i=1}^{\mathbf{K}} \gamma_i(t) g\left(x_t, \theta_i\right) \rightarrow \min_{\Gamma(t),\Theta},$$

subjected to constraints:

$$\begin{aligned} \sum_{i=1}^{\mathbf{K}} \gamma_i(t) &= 1, \quad \forall t \in [0, T] \\ \gamma_i(t) &\geq 0, \quad \forall t \in [0, T], i = 1, \ldots, \mathbf{K} \end{aligned}$$

Numerical Method: Subspace Iteration (splitting scheme)

No global convergence (non-convex optimization, simulated annealing)

# K-Means clustering: problems

*Geometrical distance*: $\theta_i \in \Psi$ - *time-independent* cluster centers

$$g(x, \theta_i) = \| x - \theta_i \|^2,$$

$$t_j, j = 1, \ldots, n \in [0, T]$$

$$\sum_{i=1}^{\mathbf{K}} \sum_{j=1}^{n} \gamma_i(t_j) \| x_{t_j} - \theta_i \|^2 \rightarrow \min_{\Gamma(t), \Theta}$$

*(Bezdek1981, Höppner et.al. 1999)*

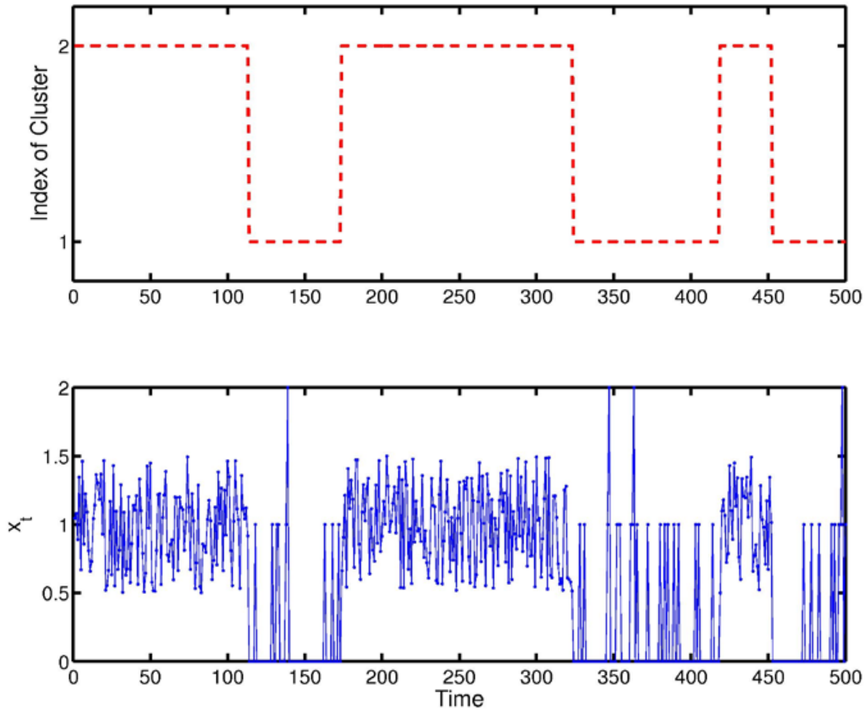
Iteration number $(l)$ :

$$\gamma_i^{(l)}(t_j) = \begin{cases} 1 & i = \arg\min \| x_{t_j} - \theta_i^{(l-1)} \|^2, \\ 0 & \text{otherwise}, \end{cases}$$

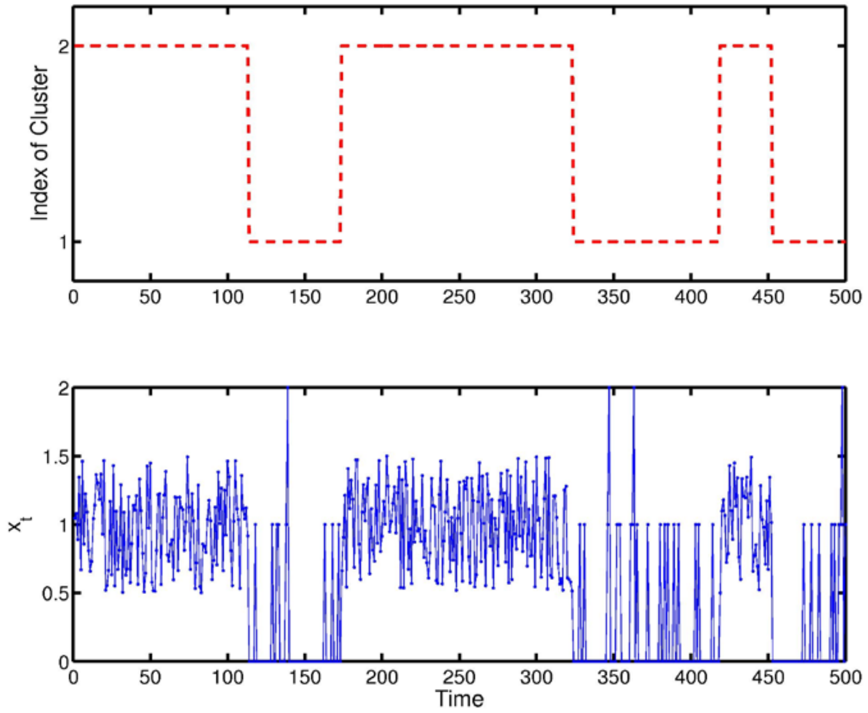
$$\theta_i^{(l)} = \frac{\sum_{j=1}^{n} \gamma_i^{(l)}(t_j) x_{t_j}}{\sum_{j=1}^{n} \gamma_i^{(l)}(t_j)}.$$

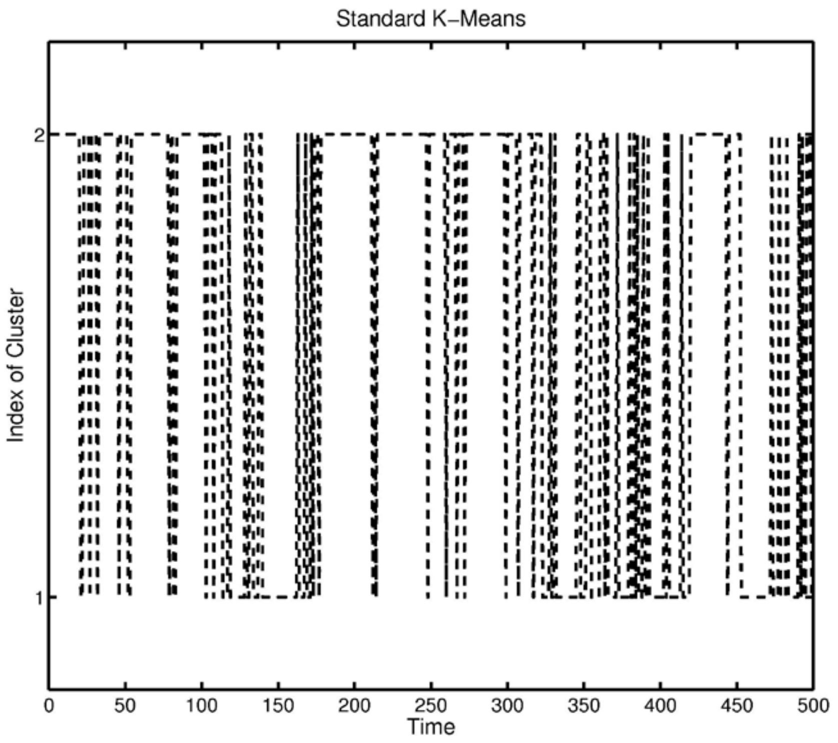
*"Sharp" assigment :* ***problem for overlapping data***

# K-Means clustering: Toy Example I



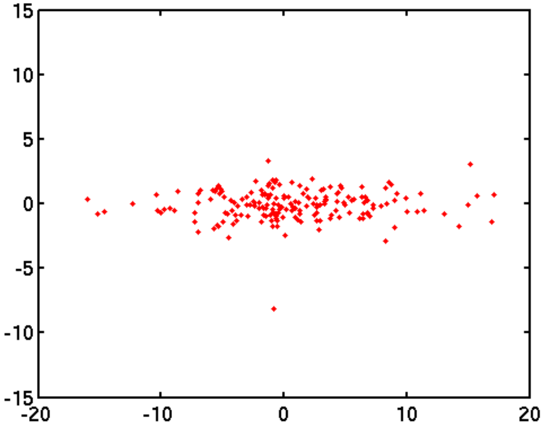
# K-Means clustering: Toy Example I



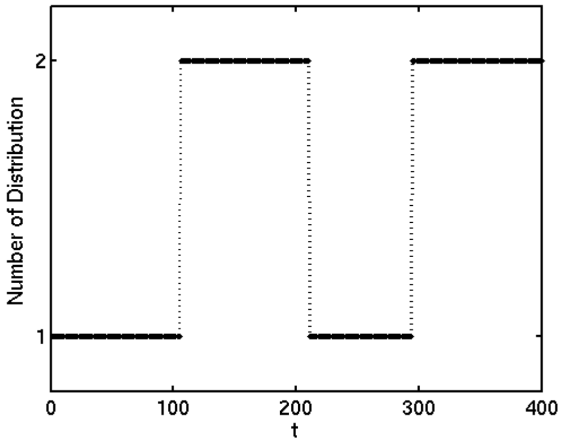



# K-Means clustering: Toy Example II

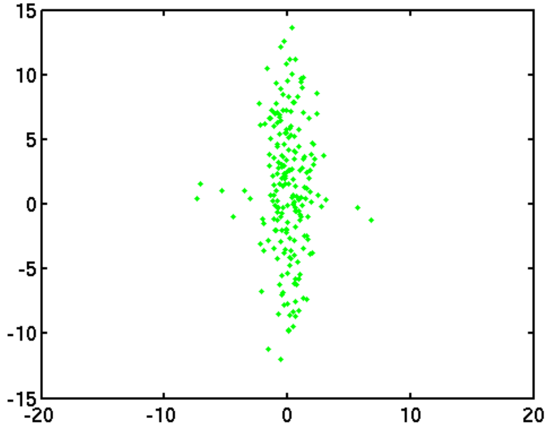
Distribution 1




Switching between the distributions




Distribution 2



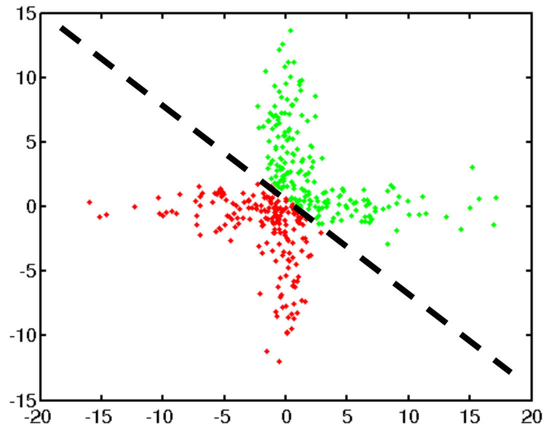

$$\mathbf{L}(\Theta,\Gamma(t)) = \int_0^T \sum_{i=1}^{\mathbf{K}} \gamma_i(t) g\left(x_t, \theta_i\right) \to \min_{\Gamma(t),\Theta},$$

$$g\left(x, \theta_i\right) = \| x - \theta_i \|^2,$$

*K-Means-Algorithm*

# K-Means clustering: Toy Example II

colouring from
geometrical clustering




$$\mathbf{L}(\Theta, \Gamma(t)) = \int_0^T \sum_{i=1}^{\mathbf{K}} \gamma_i(t) g(x_t, \theta_i) \rightarrow \min_{\Gamma(t), \Theta},$$

$$g(x, \theta_i) = \| x - \theta_i \|^2,$$

*K-Means-Algorithm*

*Problems:*

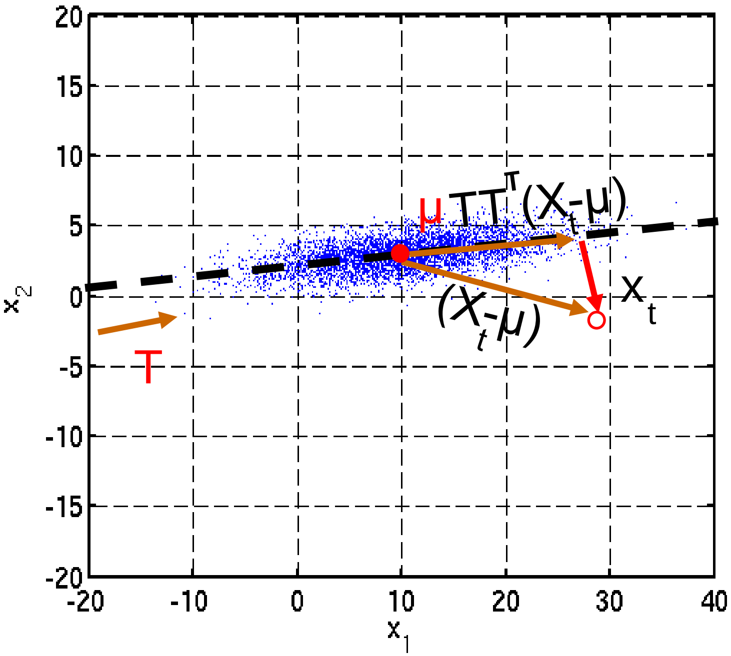
1. *Euclidean distance may be not appropriate*
2. *geometrical clustering gets no use of temporal information*

# Manifold Clustering (H.06-07)

*Problem 1:* $$\mathbf{L}\left(\gamma_i(t), \mathbf{T}_i, \mu_i\right) \longrightarrow \min$$

$$g\left(x, \theta_i\right) = \| x - \mathcal{T}_i \mathcal{T}_i^{\mathbf{T}} x \|^2$$

$\theta_i = \mathcal{T}_i \in \mathbf{R}^{nxm}$ ***m << n***

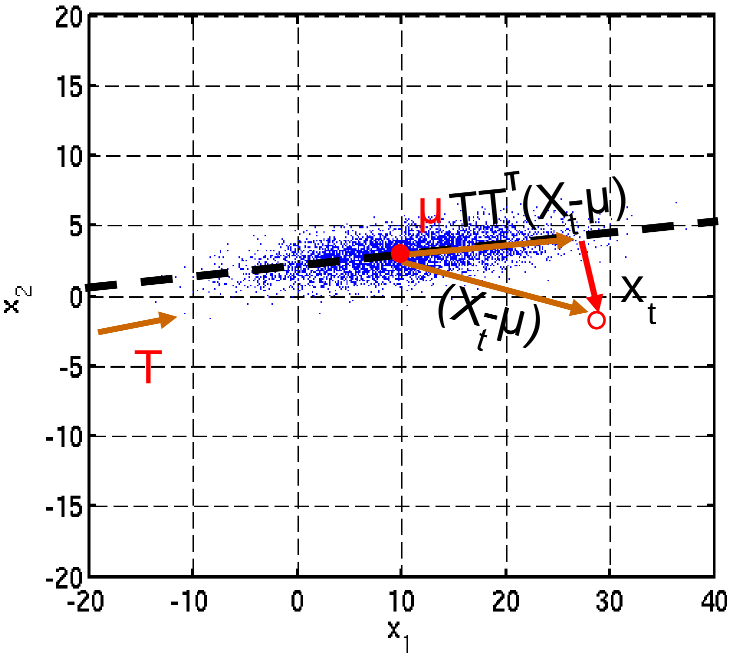


# Manifold Clustering (H.06-07)

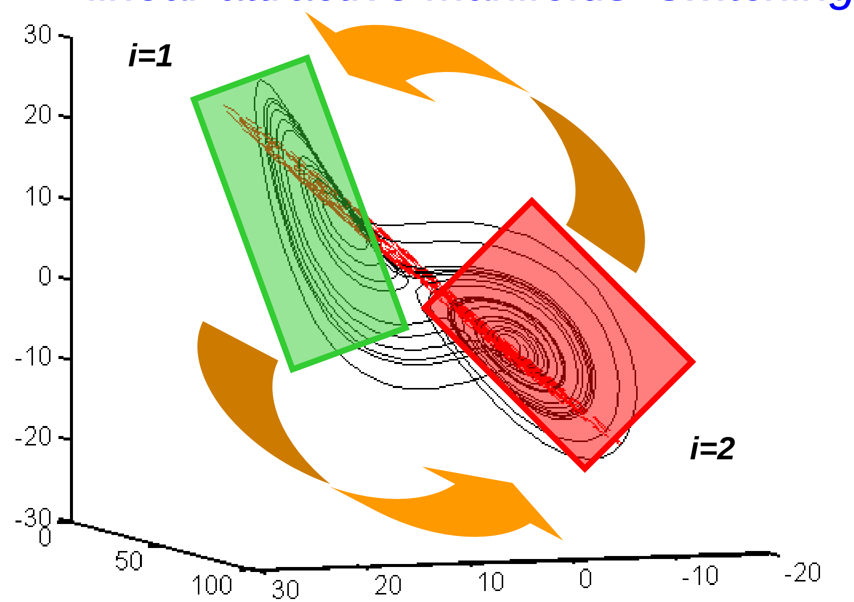
*Problem 1:* $\mathbf{L}\left(\gamma_i(t), \mathbf{T}_i, \mu_i\right) \rightarrow \min$

$$g\left(x, \theta_i\right) \quad = \quad \| \, x - \mathcal{T}_i \mathcal{T}_i^{\mathbf{T}} x \, \|^2$$

$\theta_i = \mathcal{T}_i \in \mathbf{R}^{nxm}$ ***m << n***




***Idea**: Essential manifold*
can be appproximated by
*linear attractive manifolds+switching*




*H./Schmidt-Ehrenberg/Schütte 06*
*H. 07*

# Fuzzy C-Means Algorithm

*Geometrical distance*: $\theta_i \in \Psi$ - *time-independent* cluster centers,

introduce the *"fuzzifier"* $m > 1$

$$\sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_i^m(t_j) \parallel x_{t_j} - \theta_i \parallel^2 \quad \longrightarrow \quad \min_{\Gamma(t),\Theta} \qquad \textit{(Bezdek1987)}$$

$$\mathbf{I}_{x_{t_j}} = \{p \in \{1, \ldots, K\} | \parallel x_{t_j} - \theta_p^{(l-1)} \parallel^2 = 0\}$$

$$\gamma_i^{(l)}(t_j) = \begin{cases} \dfrac{1}{\sum_{p=1}^{K} \left( \frac{\parallel x_{t_j} - \theta_i^{(l-1)} \parallel^2}{\parallel x_{t_j} - \theta_p^{(l-1)} \parallel^2} \right)^{\frac{1}{m-1}}} & \text{if } \mathbf{I}_{x_{t_j}} \text{ is empty}, \\ \sum_{r \in \mathbf{I}_{x_{t_j}}} \gamma_r^{(l)}(t_j) = 1 & \text{if } \mathbf{I}_{x_{t_j}} \text{ is not empty}, i \in \mathbf{I}_{x_{t_j}}, \\ 0 & \text{if } \mathbf{I}_{x_{t_j}} \text{ is not empty}, i \notin \mathbf{I}_{x_{t_j}}, \end{cases}$$

$$\theta_i^{(l)} = \frac{\sum_{j=1}^{n} \gamma_i^{(l)}(t_j) x_{t_j}}{\sum_{j=1}^{n} \gamma_i^{(l)}(t_j)}.$$

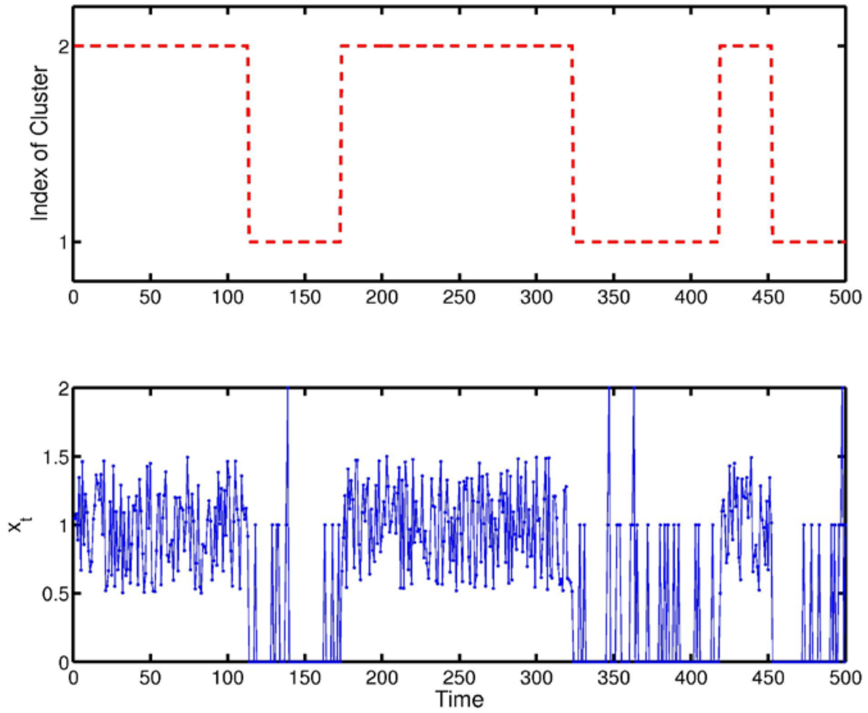
# Fuzzy C-Means Algorithm

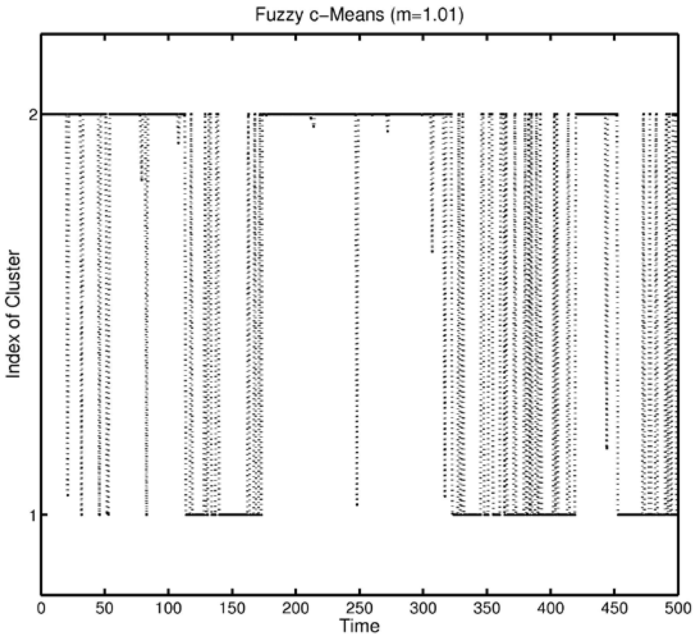
*Geometrical distance*: $\theta_i \in \Psi$ - *time-independent* cluster centers,

introduce the *"fuzzifier"* $m > 1$

$$\sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_i^m(t_j) \parallel x_{t_j} - \theta_i \parallel^2 \longrightarrow \min_{\Gamma(t),\Theta}$$

*(Bezdek1987)*



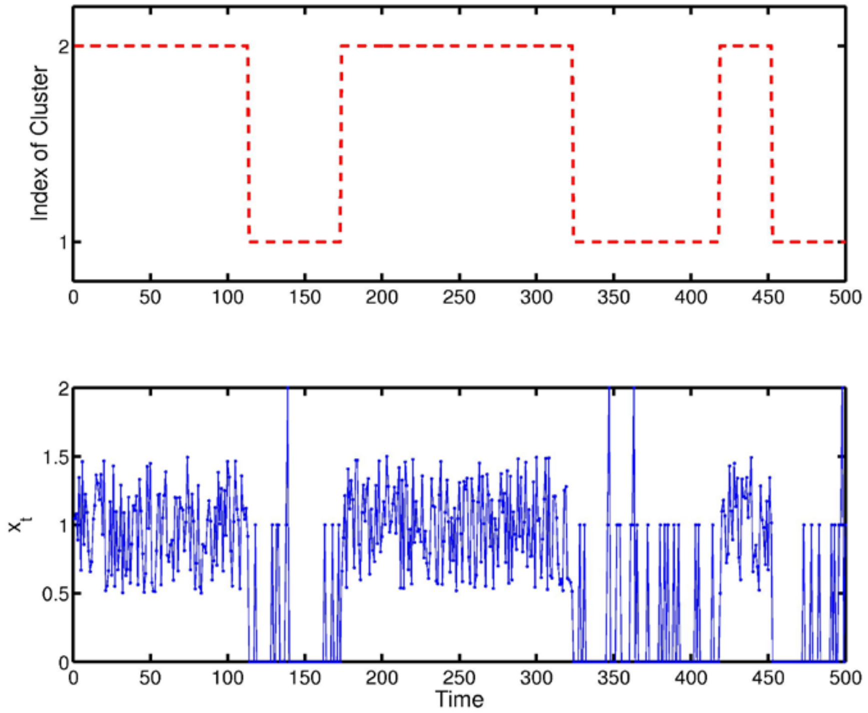
# Fuzzy C-Means Algorithm

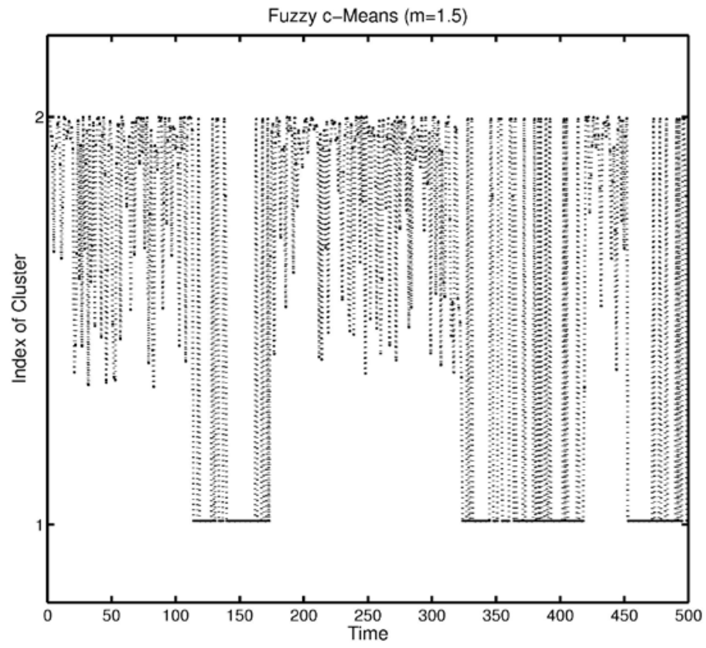
*Geometrical distance*: $\theta_i \in \Psi$ - *time-independent* cluster centers,

introduce the *"fuzzifier"* $m > 1$

$$\sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_i^m(t_j) \parallel x_{t_j} - \theta_i \parallel^2 \quad \longrightarrow \quad \min_{\Gamma(t),\Theta}$$

*(Bezdek1987)*



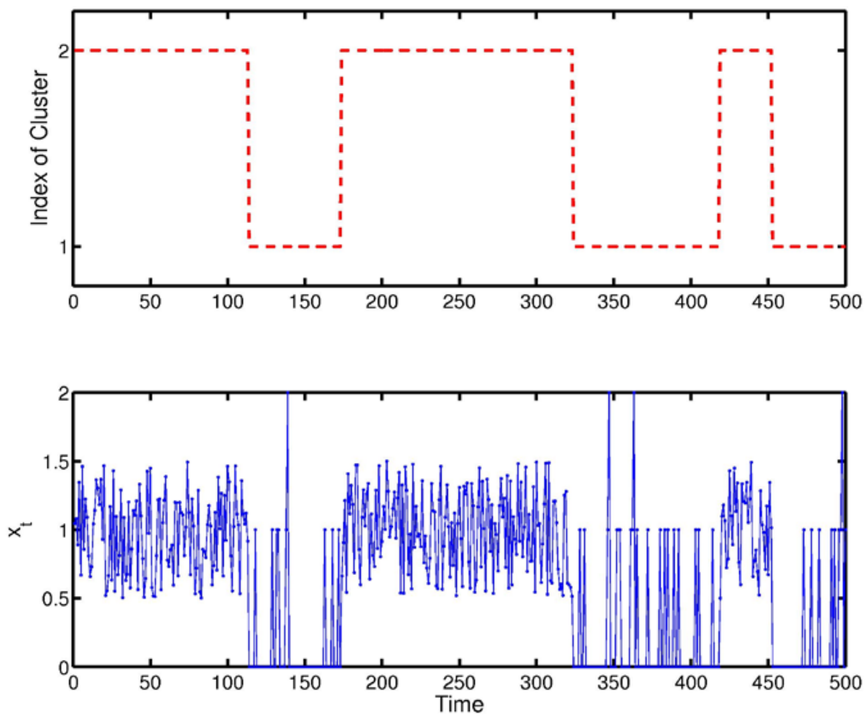
# Fuzzy C-Means Algorithm

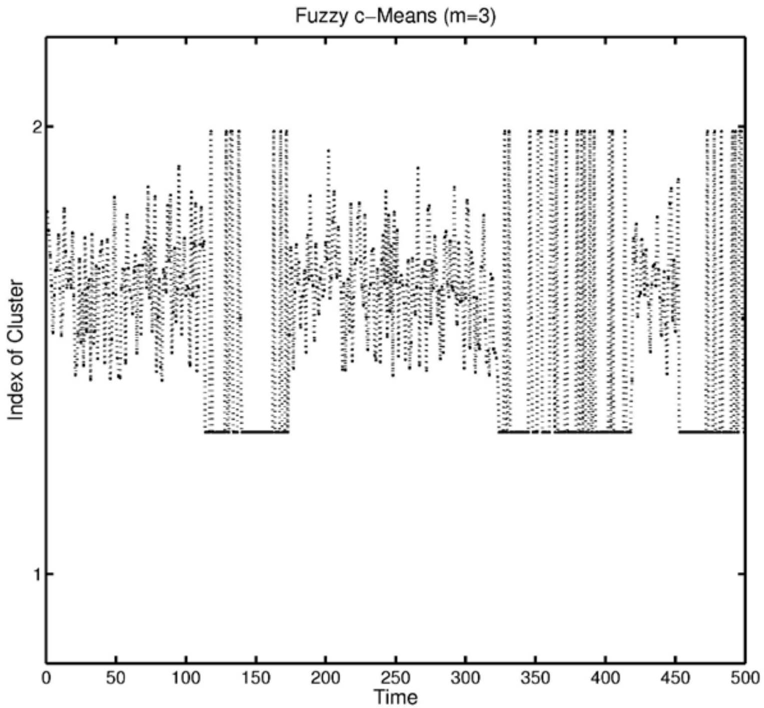
*Geometrical distance*: $\theta_i \in \Psi$ - *time-independent* cluster centers,

introduce the *"fuzzifier"* $m > 1$

$$\sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_i^m(t_j) \parallel x_{t_j} - \theta_i \parallel^2 \longrightarrow \min_{\Gamma(t),\Theta}$$

*(Bezdek1987)*



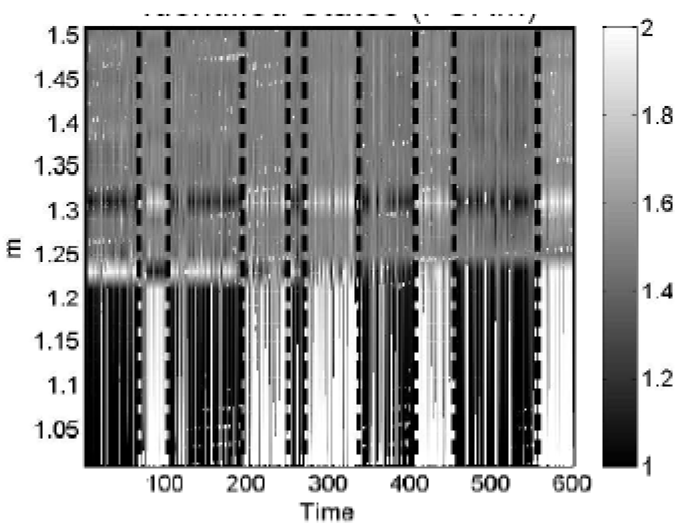
# Fuzzy C-Means Algorithm

*Geometrical distance*: $\theta_i \in \Psi$ - *time-independent* cluster centers,

introduce the *"fuzzifier"* $m > 1$

$$\sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_i^m(t_j) \parallel x_{t_j} - \theta_i \parallel^2 \quad \longrightarrow \quad \min_{\Gamma(t),\Theta}$$

*(Bezdek1987)*




*Just to "fuzzify" is not enough!*

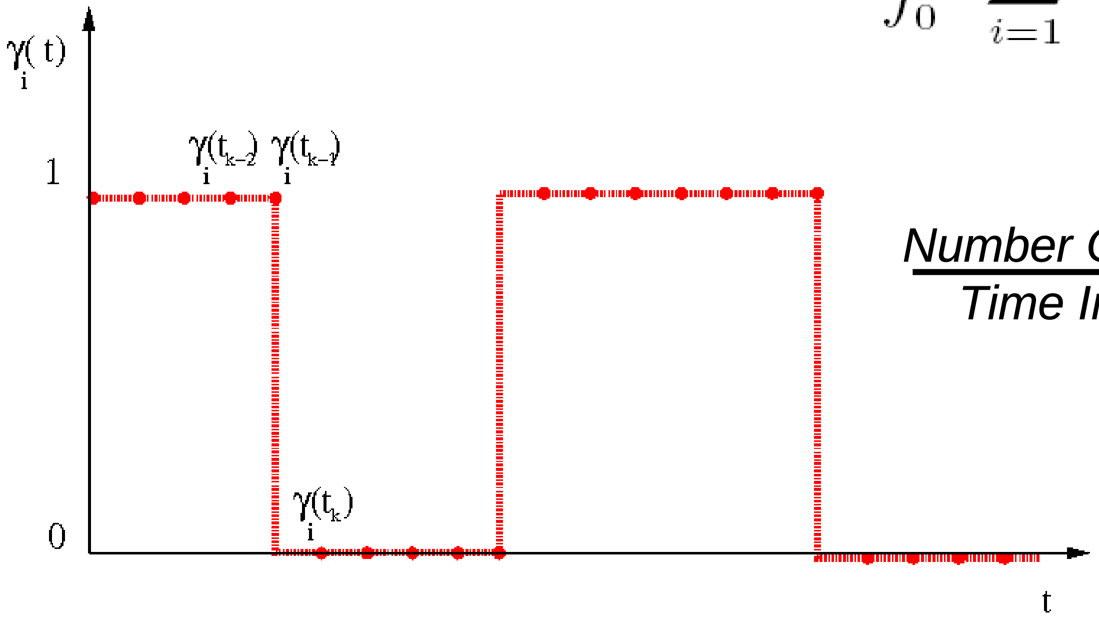
# Incorporation of Temporal Information

*Problem 2:* *identification of "persistent states"*

Let $\gamma_i(\cdot)$ on $t \in [0,T], i = 1,\ldots,\mathbf{K}$ be differentiable and

$\partial_t \gamma_i \in \mathcal{L}_2(0,T)$, i. e. :

$$\mathbf{L}(\Theta,\Gamma(t)) = \int_0^T \sum_{i=1}^{\mathbf{K}} \gamma_i(t) g(x_t,\theta_i) \rightarrow \min_{\Gamma(t),\Theta},$$




$$\frac{\textit{Number Of Jumps}}{\textit{Time Interval}} =$$

$$= \sum_{k=1} \frac{\left(\gamma_i(t_{k+1}) - \gamma_i(t_k)\right)^2}{\Delta t}$$

# Tikhonov-type Regularization

## *Problem 2:* *incorporation of temporal information*

Let $\gamma_i(\cdot)$ on $t \in [0,T]\,, i = 1,\ldots,\mathbf{K}$ be differentiable and

$\partial_t \gamma_i \in \mathcal{L}_2(0,T)$ , i. e. :

$$\mathbf{L}(\Theta,\Gamma(t)) = \int_0^T \sum_{i=1}^{\mathbf{K}} \gamma_i(t) g(x_t,\theta_i) \rightarrow \min_{\Gamma(t),\Theta},$$

subjected to

$$|\gamma_i|_{\mathcal{H}^1(0,T)} = \| \partial_t \gamma_i(\cdot) \|_{\mathcal{L}_2(0,T)} = \int_0^T (\partial_t \gamma_i(t))^2 \, dt \le C_\epsilon^i < +\infty,$$

*Regularized clustering functional*:

$$\mathbf{L}^\epsilon(\Theta,\Gamma(t),\epsilon^2) = \mathbf{L}(\Theta,\Gamma(t)) + \epsilon^2 \sum_{i=1}^{\mathbf{K}} \int_0^T (\partial_t \gamma_i(t))^2 \, dt \rightarrow \min_{\Gamma(t),\Theta}$$

*(H. 08, to appear in SISC)*

# FEM: Regularized Clustering Functional

*Regularized clustering functional:* *(H. 08, to appear in SISC)*

$$\mathbf{L}^{\epsilon}(\Theta, \Gamma(t), \epsilon^2) = \mathbf{L}(\Theta, \Gamma(t)) + \epsilon^2 \sum_{i=1}^{\mathbf{K}} \int_0^T (\partial_t \gamma_i(t))^2 \, dt \rightarrow \min_{\Gamma(t), \Theta},$$

# FEM: Regularized Clustering Functional

*Regularized clustering functional:* (H. 08, to appear in SISC)

$$\mathbf{L}^{\epsilon}(\Theta,\Gamma(t),\epsilon^2) \quad = \quad \mathbf{L}(\Theta,\Gamma(t)) + \epsilon^2 \sum_{i=1}^{\mathbf{K}} \int_0^T \left(\partial_t \gamma_i\left(t\right)\right)^2 dt \longrightarrow \min_{\Gamma(t),\Theta},$$

*Galerkin-Ansatz* :

$$\begin{aligned} \gamma_i\left(t\right) &= \tilde{\gamma}_i\left(t\right) + \delta_N \\ &= \sum_{k=1}^{N} \tilde{\gamma}_i^{(k)} v_k\left(t\right) + \delta_N \end{aligned}$$

where $\tilde{\gamma}_i^{(k)} = \int_0^T \gamma_i\left(t\right) v_k\left(t\right) dt$, and

$$\mathbf{L}^{\epsilon} \quad = \quad \tilde{\mathbf{L}}^{\epsilon} + \mathcal{O}\left(\delta_N\right) \longrightarrow \min_{\tilde{\gamma}_i(t),\Theta},$$

$$\tilde{\mathbf{L}}^{\epsilon} = \sum_{i=1}^{\mathbf{K}} \int_0^T \left[\tilde{\gamma}_i\left(t\right) g(x_t,\theta_i) + \epsilon^2 \left(\partial_t \tilde{\gamma}_i\left(t\right)\right)^2\right] dt.$$

# FEM-Discretized Clustering Functional

*(H. 08, to appear in SISC)*

$$\tilde{\mathbf{L}}^{\epsilon} = \sum_{i=1}^{\mathbf{K}} \left[ a^{\mathbf{T}}(\theta_i)\bar{\gamma}_i + \epsilon^2 \bar{\gamma}_i^{\mathbf{T}} \mathbf{H} \bar{\gamma}_i \right] \rightarrow \min_{\bar{\gamma}_i, \Theta}$$

subjected to

$$\sum_{i=1}^{\mathbf{K}} \tilde{\gamma}_i^{(k+1)} = 1, \quad \forall k = 1, \ldots, N,$$

$$\tilde{\gamma}_i^{(k+1)} \geq 0, \quad \forall k = 1, \ldots, N; i = 1, \ldots, \mathbf{K}.$$

*Iterative Subspace Minimization:* ***sparse QP*** can be used

where $a(\theta_i) = \left( \int_{t_1}^{t_2} v_1(t) g(x_t, \theta_i) dt, \ldots, \int_{t_{N-1}}^{t_N} v_N(t) g(x_t, \theta_i) dt \right)$

is a vector of *FEM-discretized model distances* and $\mathbf{H}$ is

a *mass-matrix* of the *FEM-basis*

# Algorithm: monotony conditions

**Theorem** *Let for a given observed time series* $x(t) : \mathbf{R}^1 \to \Psi \subset \mathbf{R}^n$, *the model distance functional is chosen such that it satisfies (2),* $\Psi$ *and* $\Omega$ *are compact,* $g(x_t, \cdot)$ *is continuously differentiable function of* $\theta$ *and*

$$\frac{\partial}{\partial \Theta} \tilde{\mathbf{L}}^{\epsilon} \left(\Theta^*, \bar{\gamma}\right) \quad = \quad 0,$$

*has a unique solution* $\Theta^* = \left(\theta_1^*, \ldots, \theta_{\mathbf{K}}^*\right), \theta_i * \in \Omega$ *for any fixed* $\bar{\gamma}$ *satisfying (18-19) and* $\frac{\partial^2}{\partial \Theta^2} \tilde{\mathbf{L}}^{\epsilon} \left(\Theta^*, \bar{\gamma}\right)$ *is positive definite. Then for any* $\epsilon^2 \geq 0$ *and any finite continuous non-negative finite elements set* $\{v_1(t), v_2(t), \ldots, v_N(t)\} \in \mathcal{L}_2(0, T)$ *such that the respective* mass-matrix $\mathcal{H}$ *is positive definite, the above algorithm is* monotonous, *i. e., for any* $j \geq 1$

$$\tilde{\mathbf{L}}^{\epsilon} \left(\Theta^{[j+1]}, \bar{\gamma}_i^{[j+1]}\right) \quad \leq \quad \tilde{\mathbf{L}}^{\epsilon} \left(\Theta^{[j]}, \bar{\gamma}_i^{[j]}\right).$$

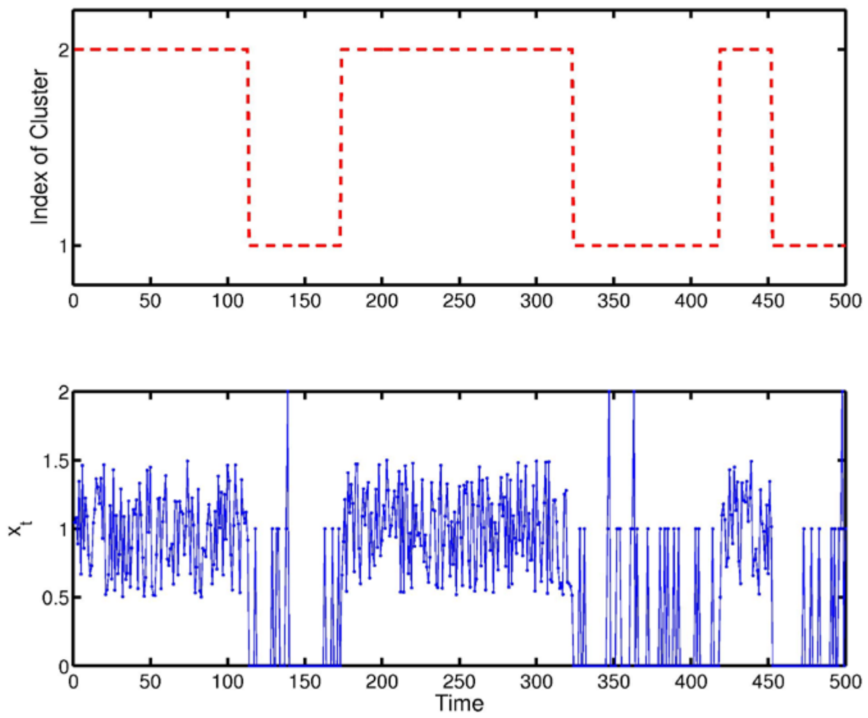
Convergence to a local optimum only!

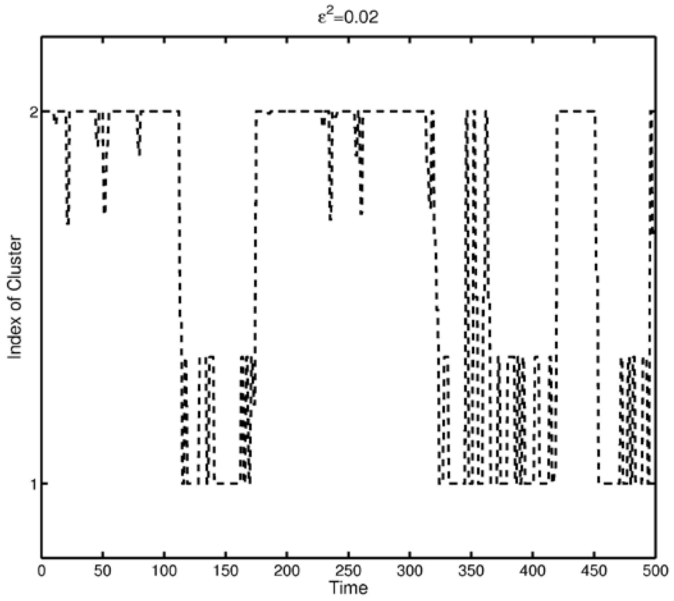
(coupling to some *global opimizer* necessary)

# Toy Example I

$$\mathbf{L}\left(\gamma_i(t), \mathbf{T}_i, \mu_i\right) \longrightarrow \min$$

$$g\left(x, \theta_i\right) \quad = \quad \parallel x - \theta_i \parallel^2,$$

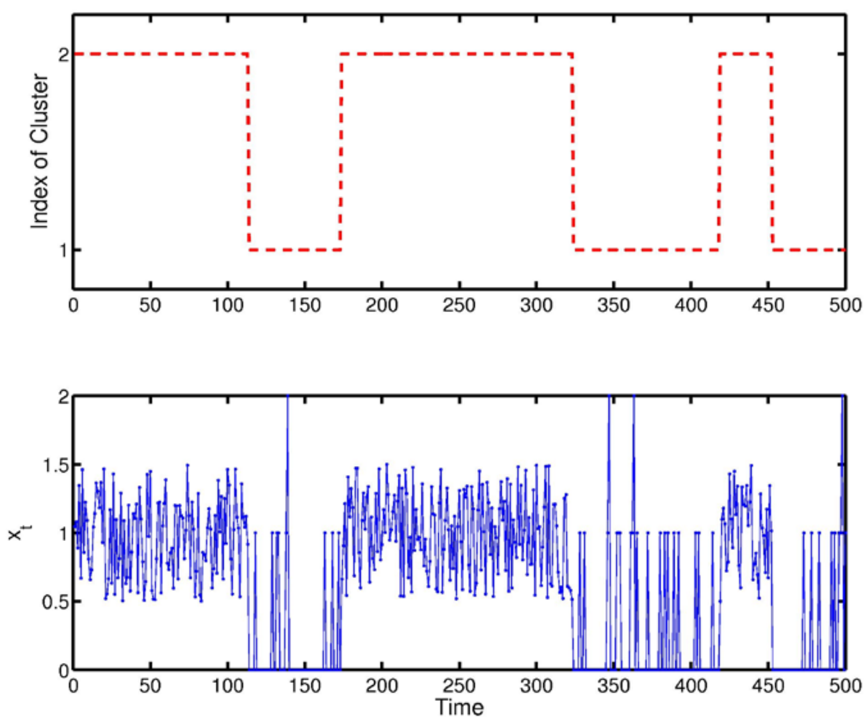


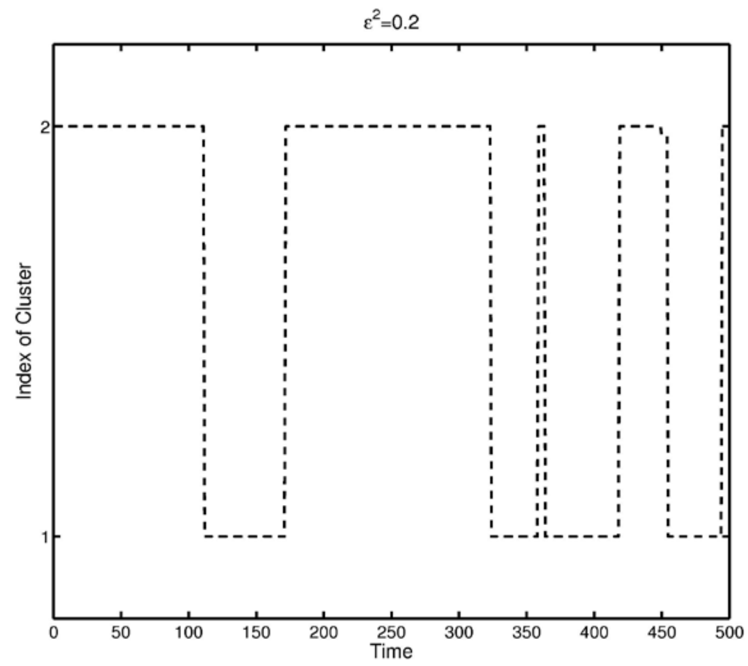



# Toy Example I

$$\mathbf{L}\left(\gamma_i(t), \mathbf{T}_i, \mu_i\right) \longrightarrow \min$$

$$g\left(x, \theta_i\right) \quad = \quad \| \, x - \theta_i \, \|^2,$$

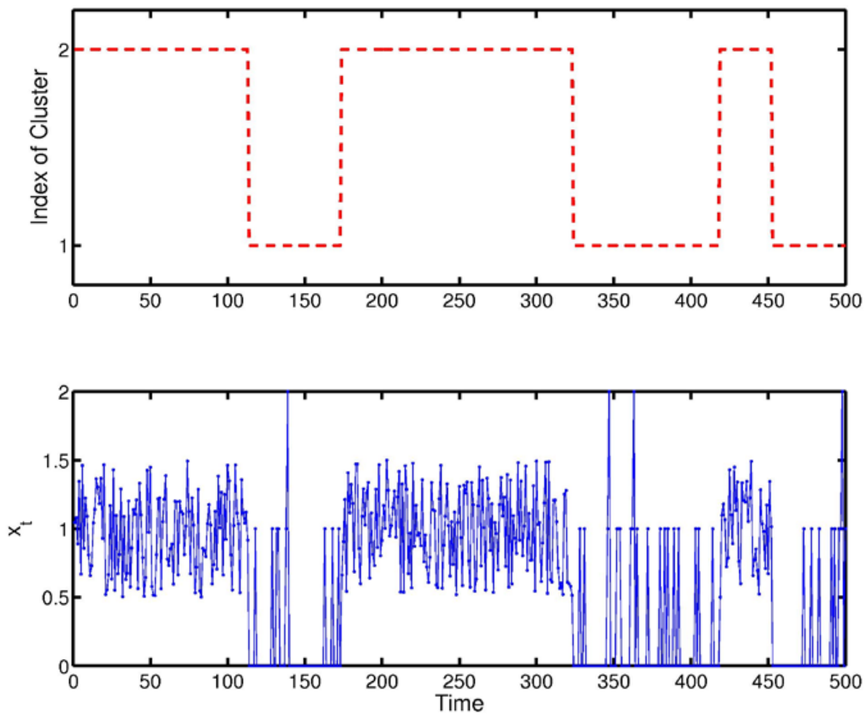


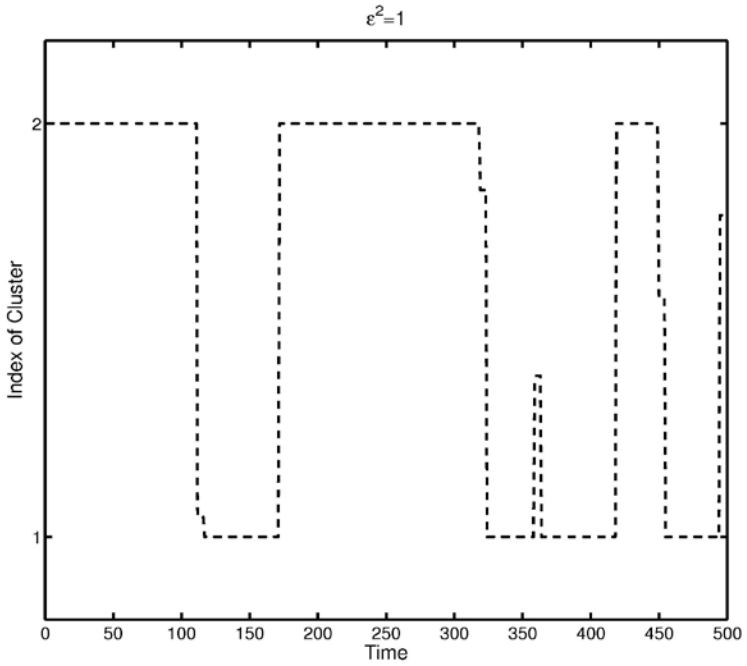



# Toy Example I

$$\mathbf{L}\left(\gamma_i(t), \mathbf{T}_i, \mu_i\right) \rightarrow \min$$

$$g\left(x, \theta_i\right) = \| x - \theta_i \|^2,$$

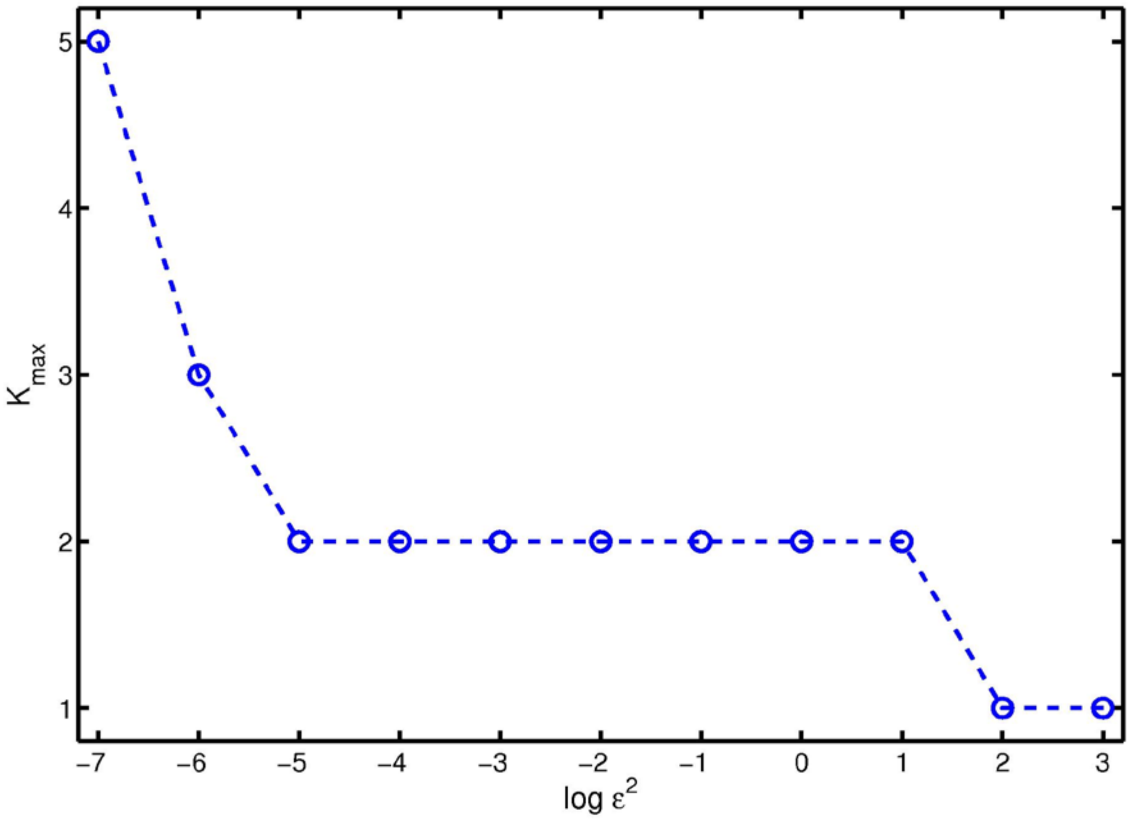


# Toy Example I

How to determine the optimal K: probabilistic model assumptions a posteriory

$$g(x, \theta_i) \quad = \quad \| x - \theta_i \|^2,$$



# Toy Example I

How to determine the optimal ε: standard L-Curve approach from Tikhonov-regularized linear least-squares problems (Cullum(79), Hansen(99))