Gaussian Mixture Model - method and application

Presentation · November 2017

DOI: 10.13140/RG.2.2.32667.77602

1 author:

Jesús Zambrano
The MathWorks

Gaussian Mixture Models – method and applications

Jesús Zambrano

PostDoctoral Researcher

School of Business, Society and Engineering

www.mdh.se

Outline

- Method
  - Introduction to Gaussian Mixture Process (GMM)
  - Standard construction of GMM
  - Clustering (Silhouette and Akaike criterion)
- Case studies
  - Monitoring a secondary settler tank
  - Residual and fault detection criteria
- Conclusions

Gaussian Mixture Model (GMM) - standard construction

μ_k: mean
σ_k: covariance

A linear superposition of K-Gaussians

$$p(\mathrm{x}_i) = \sum_{k=1}^{K} \underbrace{\pi_k}_{p(k)} \underbrace{\mathcal{N}(\mathrm{x}_i|\mu_k, \sigma_k)}_{p(\mathrm{x}_i|k)}, \quad i = 1, \dots, N$$

is called a **Gaussian mixture (GM)**. The mixture coefficient π_k satisfies

$$\sum_{k=1}^{K} \pi_k = 1, \qquad 0 \leq \pi_k \leq 1$$

Interpretation: The density $p(\mathrm{x}|k) = \mathcal{N}(\mathrm{x}|\mu_k, \sigma_k)$ is the probability of x, given that component k was chosen. The probability of choosing component k is given by the prior probability $p(k)$.

GMM - standard construction (cont.)

For example, consider the following GMM:

$$p(x) = \underbrace{0.3}_{\pi_1} \mathcal{N}\left(x \mid \underbrace{\begin{pmatrix} 4 \\ 4.5 \end{pmatrix}}_{\mu_1}, \underbrace{\begin{pmatrix} 1.2 & 0.6 \\ 0.6 & 0.5 \end{pmatrix}}_{\Sigma_1}\right) + \underbrace{0.5}_{\pi_2} \mathcal{N}\left(x \mid \underbrace{\begin{pmatrix} 8 \\ 1 \end{pmatrix}}_{\mu_2}, \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\Sigma_2}\right) + \underbrace{0.2}_{\pi_3} \mathcal{N}\left(x \mid \underbrace{\begin{pmatrix} 9 \\ 8 \end{pmatrix}}_{\mu_3}, \underbrace{\begin{pmatrix} 0.6 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}}_{\Sigma_3}\right)$$

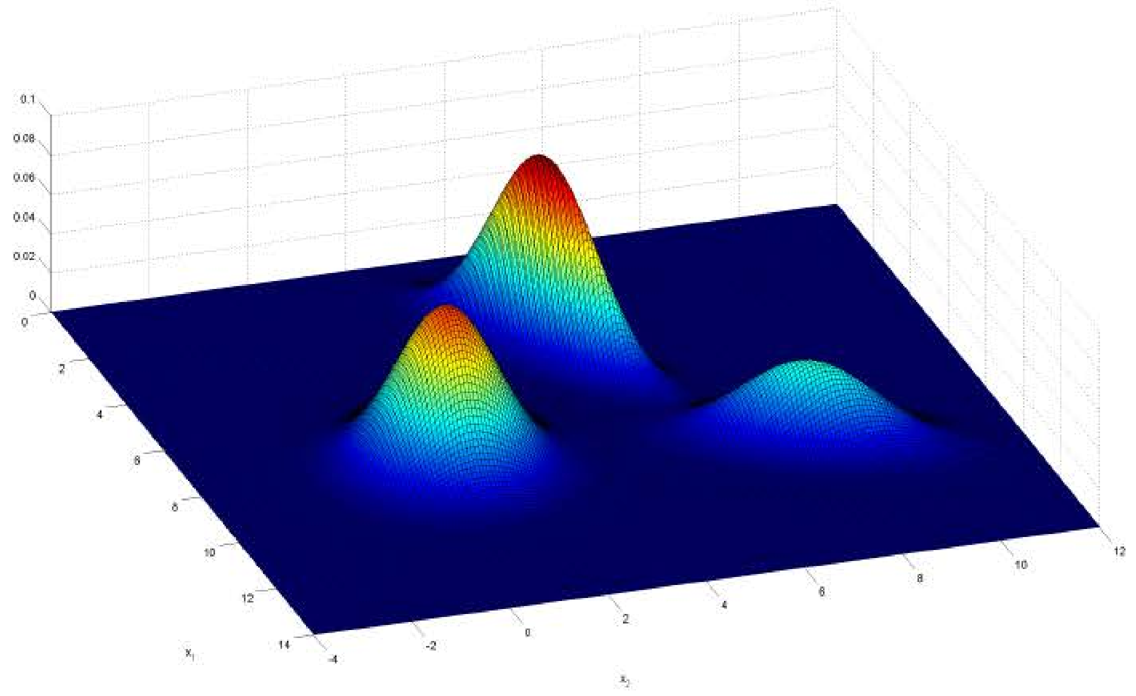


Figure: Probability density function.

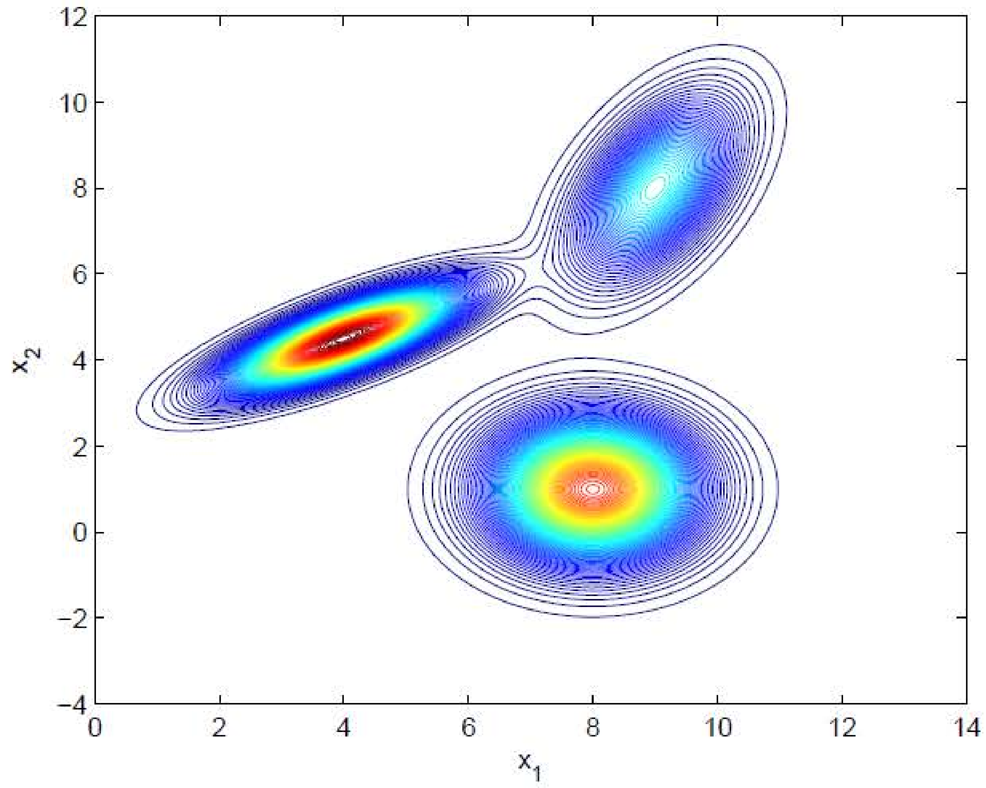


Figure: Contour plot.

GMM - standard construction (cont.)

The form of the GM distribution is governed by the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. One way to get them is by **maximum likelihood**.

Given N observations $\{\mathrm{x}_n\}_{n=1}^{N}$, the log-likelihood function is

$$\ln p\left(X; \pi_{1:K}, \mu_{1:K}, \sigma_{1:K}\right) = \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathrm{x}_n | \mu_k, \sigma_k\right)\right)$$

There is **no closed-form solution** available (due to the sum inside the logarithm).

This problem can be separated into two simple problems using the *expectation-maximization (EM)* algorithm.

GMM - standard construction (cont.)

Conditions to be satisfied at a maximum of the likelihood function

$$\frac{\mathrm{d}}{\mathrm{d}\mu_k}\left[\ln p(\mathrm{x}|\pi,\mu,\sigma)\right]=0 \quad \rightarrow \quad 0=-\sum_{n=1}^{N}\underbrace{\frac{\pi_k\mathcal{N}(\mathrm{x}_n|\mu_k,\sigma_k)}{\sum_j \pi_j\mathcal{N}(\mathrm{x}_n|\mu_j,\sigma_j)}}_{\gamma(z_{nk})}\sigma_k(\mathrm{x}_n-\mu_k)$$

which gives $\rightarrow \quad \mu_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathrm{x}_n$

$$\frac{\mathrm{d}}{\mathrm{d}\sigma_k}\left[\ln p(\mathrm{x}|\pi,\mu,\sigma)\right]=0 \quad \rightarrow \quad \sigma_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathrm{x}_n-\mu_k\right)\left(\mathrm{x}_n-\mu_k\right)^T$$

Maximize $\ln p(\mathrm{x}|\pi,\mu,\sigma)$ with respect to π_k (using Lagrange multipliers) gives

$$\pi_k=\frac{N_k}{N}, \quad \text{where} \quad N_k=\sum_{n=1}^{N}\gamma(z_{nk})$$

For more details of EM and GMM see: C. Bishop, Pattern Recognition and Machine Learning, Springer, 2007.

GMM - standard construction (cont.)

Algorithm 1 EM for Gaussian mixtures

Initialize $\mu_k^1, \sigma_k^1, \pi_k^1$ and set $i = 1$.
while not converged **do**
    Compute $\gamma(z_{nk})$. ▷ Expectation step
    Compute $\mu_k^{i+1}; \pi_k^{i+1}; N_k; \sigma_k^{i+1}$. ▷ Maximization step
    $i \leftarrow i + 1$.
end while

$$\gamma(z_{nk}) = \frac{\pi_k^i \mathcal{N}(\mathrm{x}_n | \mu_k^i, \sigma_k^i)}{\sum_{j=1}^{K} \pi_j^i \mathcal{N}(\mathrm{x}_n | \mu_j^i, \sigma_j^i)}, n = 1, ..., N; k = 1, ..., K$$

$$\mu_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n,$$

$$\pi_k^{i+1} = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}),$$

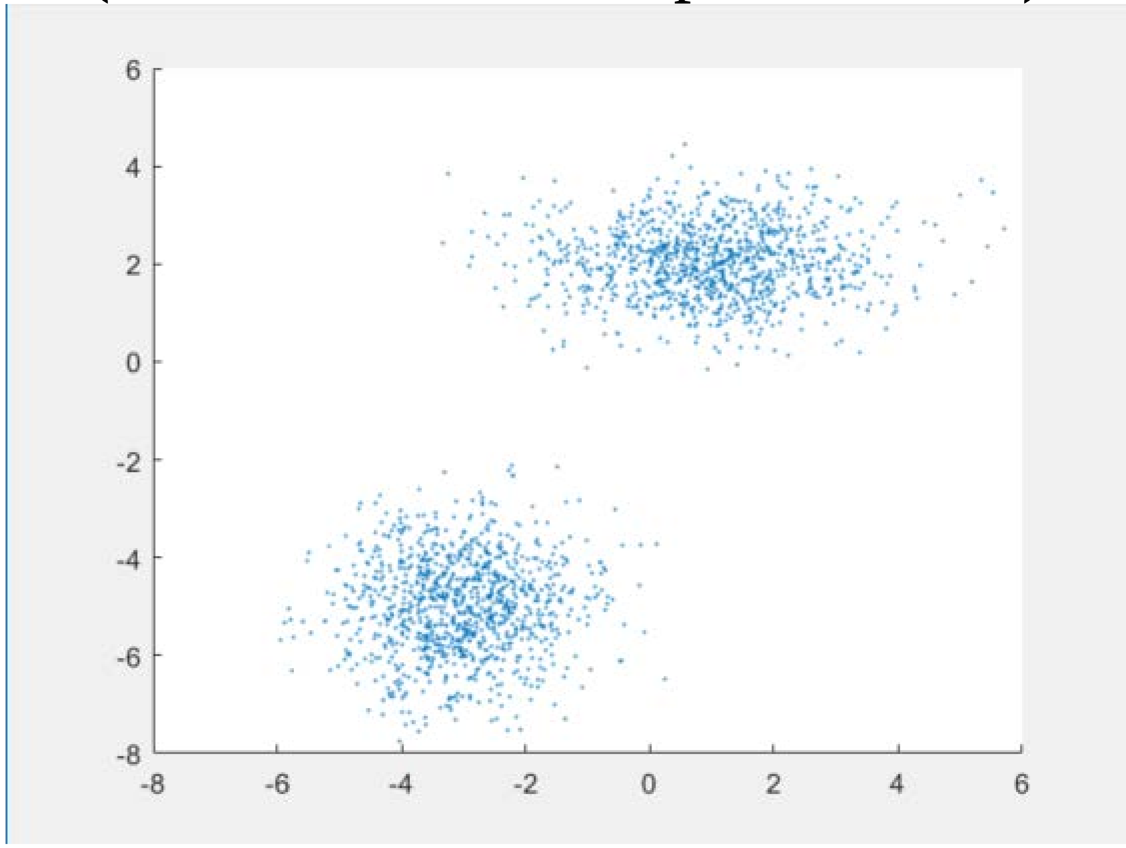
$$\sigma_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathrm{x}_n - \mu_k^{i+1}\right) \left(\mathrm{x}_n - \mu_k^{i+1}\right)^T.$$

A simple Matlab example

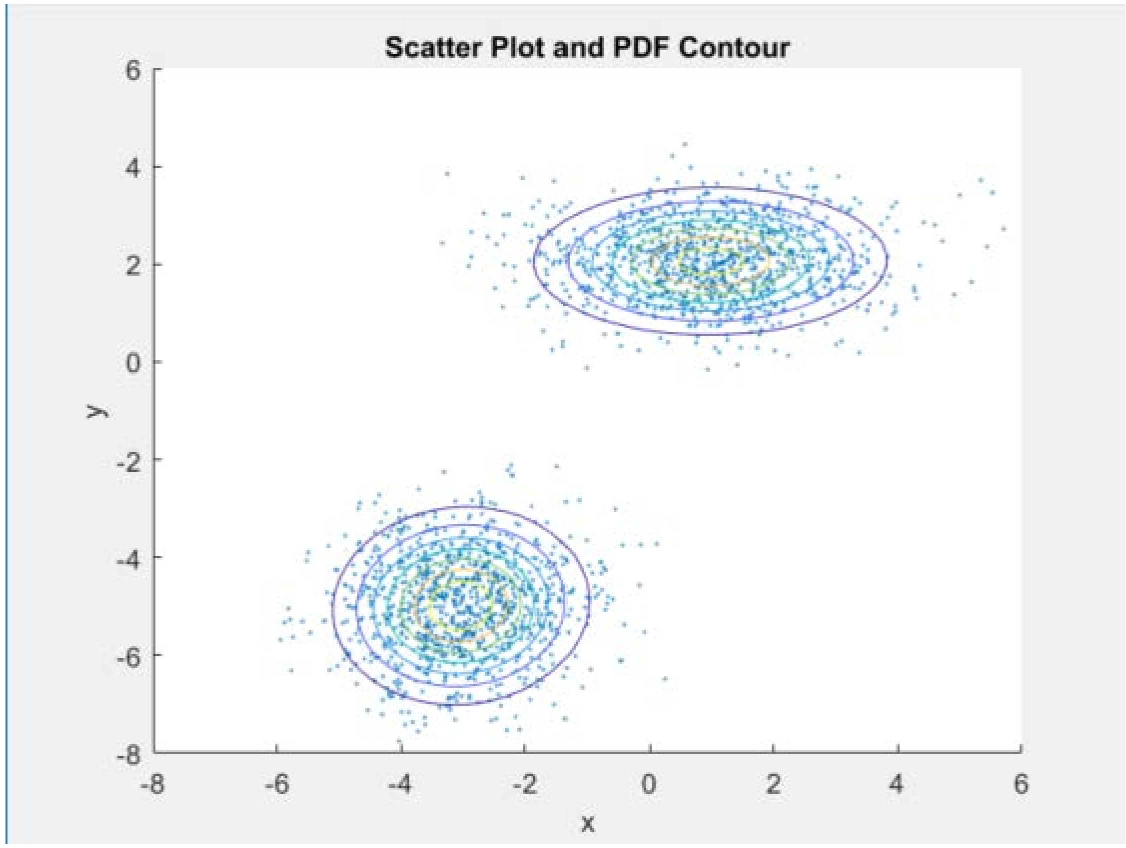
- Matlab functions:
  - `fitgmdist` (Fit a Gaussian mixture distribution to data)
  - `pdf` (Density function of a specific ditribution)

Raw data

(2 clusters of 1000 points each)



Data model with 2 Gaussian Mixture distributions



Run: gmm_example.m

A simple Matlab example (cont.)

- Silhouette value (S)

It is a measure of how similar a point is to a point in its own cluster.

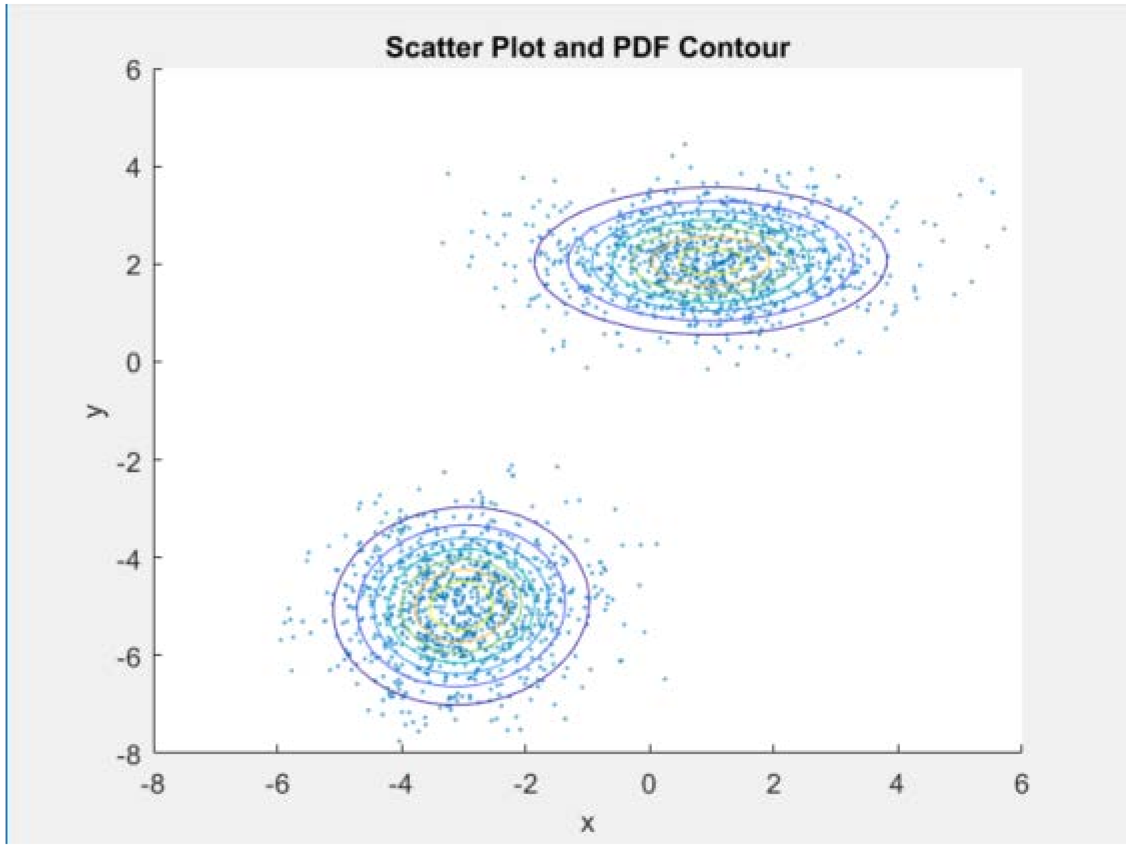
Minimum average distance from the i^{th} point to points in a different cluster

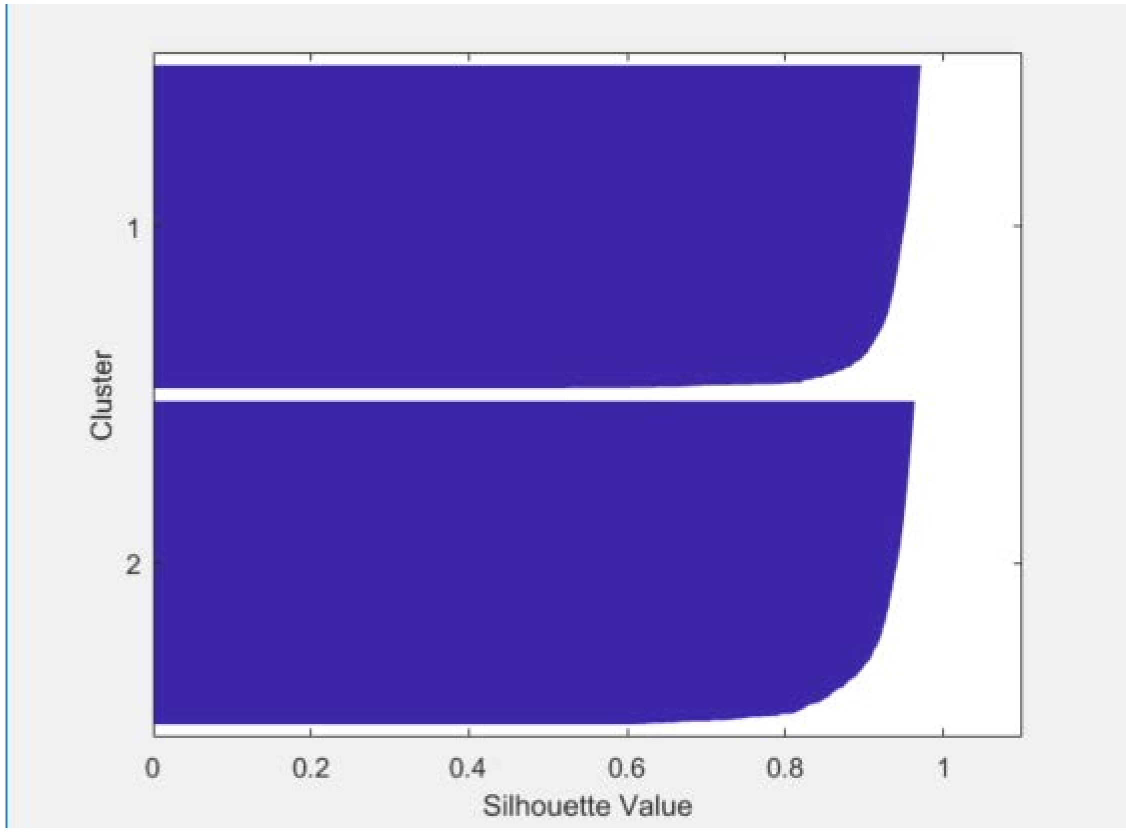
Average distance from i^{th} point to other points in the same cluster

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

For well match of i in its own cluster, b_i should be large and a_i small.

S_i ranges between -1 to +1. High S_i indicates that i is well-matched to its own cluster, and poorly-matched to neighboring clusters.

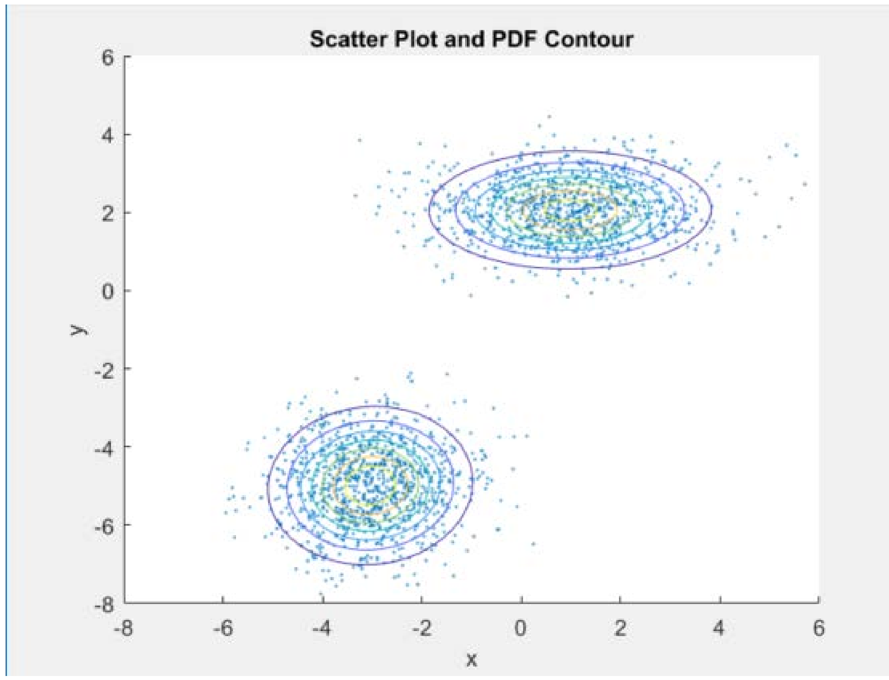


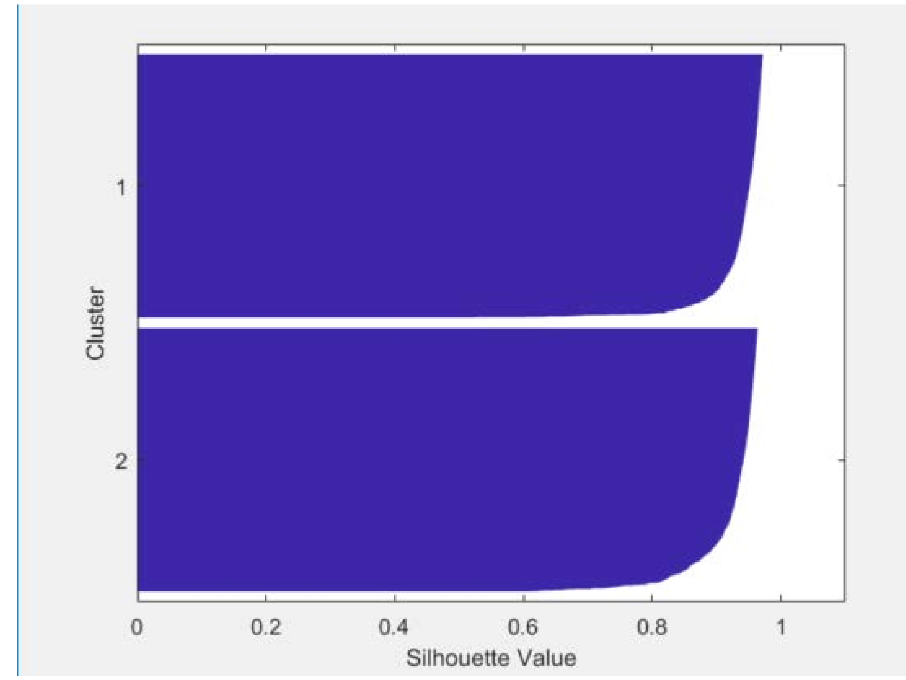


A simple Matlab example (cont.)

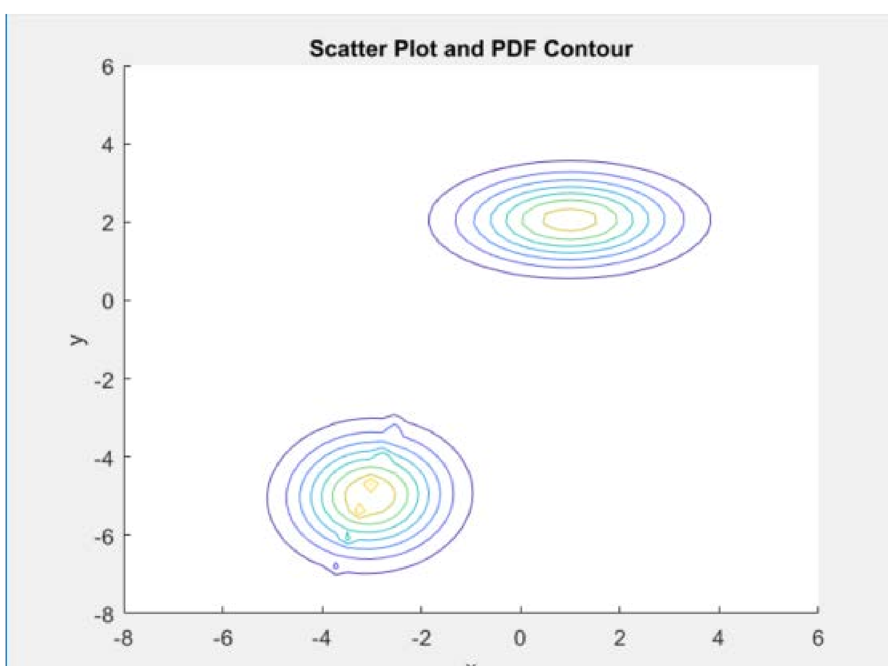
- Silhouette value (S)

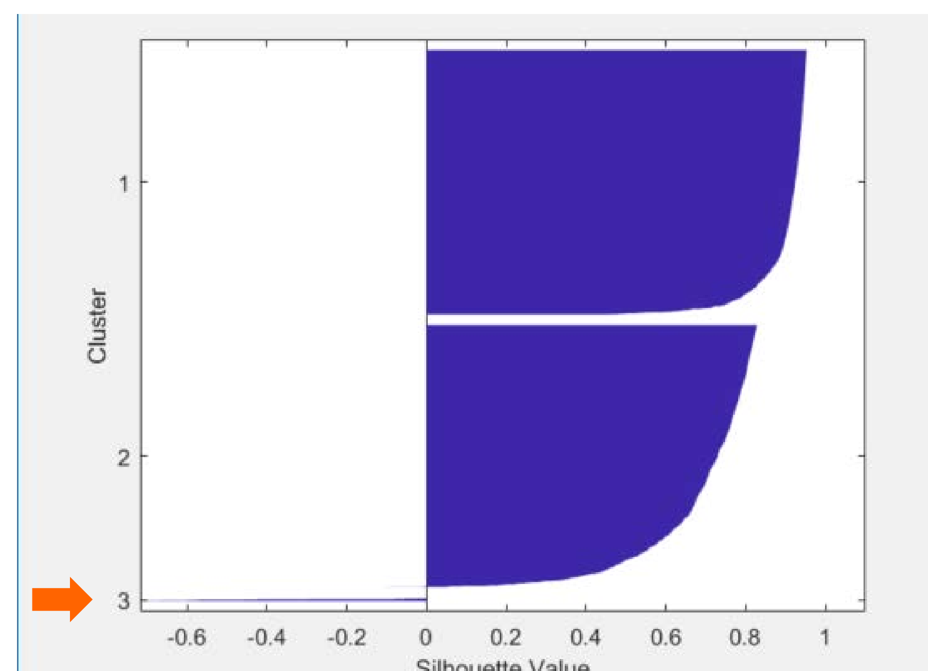
K=2 GM





K=3 GM





A simple Matlab example (cont.)

- Akaike's Information Criterion (AIC)

Provides a measure of the relative quality of a model for a given set of data.

Number of estimated parameters

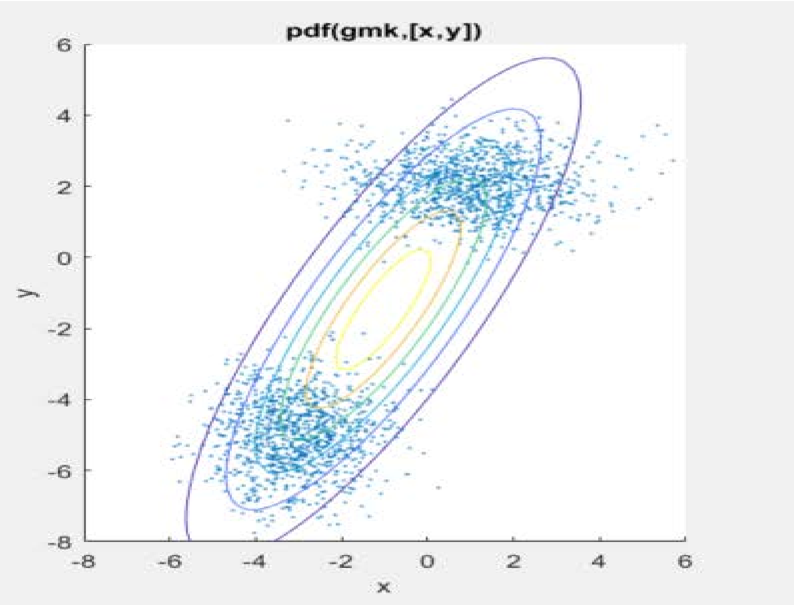
Model parameters

Then, the aim is to get: $\min_{n_p,\theta}\left(1+\frac{2n_p}{N}\right)\sum_{t=1}^{N}\varepsilon^2(t,\theta)$

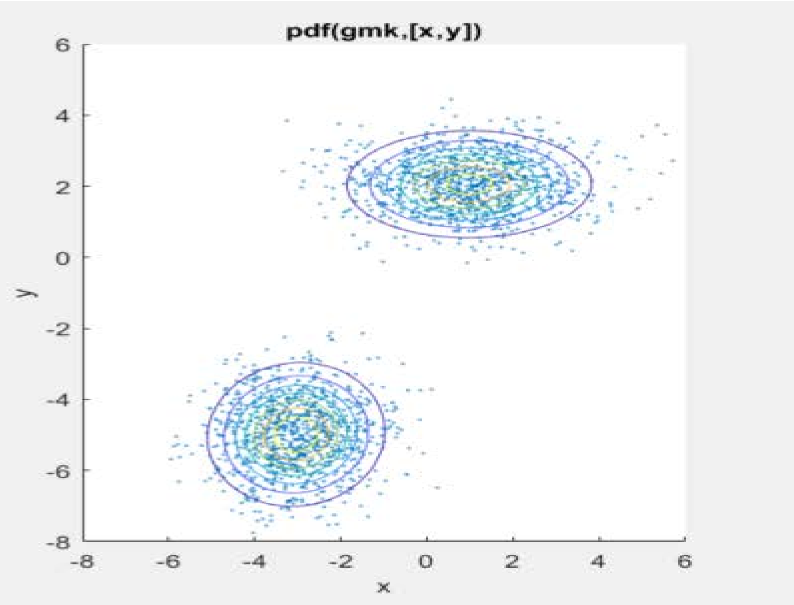
Number of values in the estimation data set

Prediction error

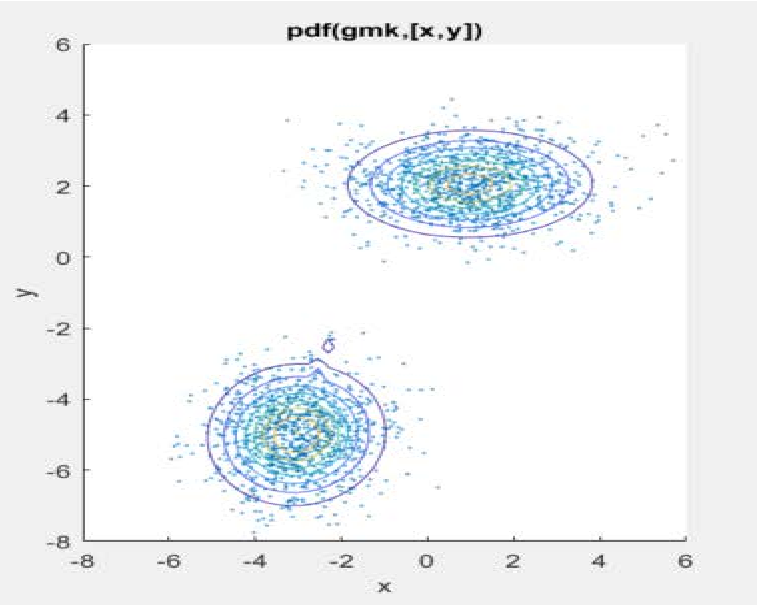
The most accurate model has the smallest AIC.



AIC=17584



AIC=14233

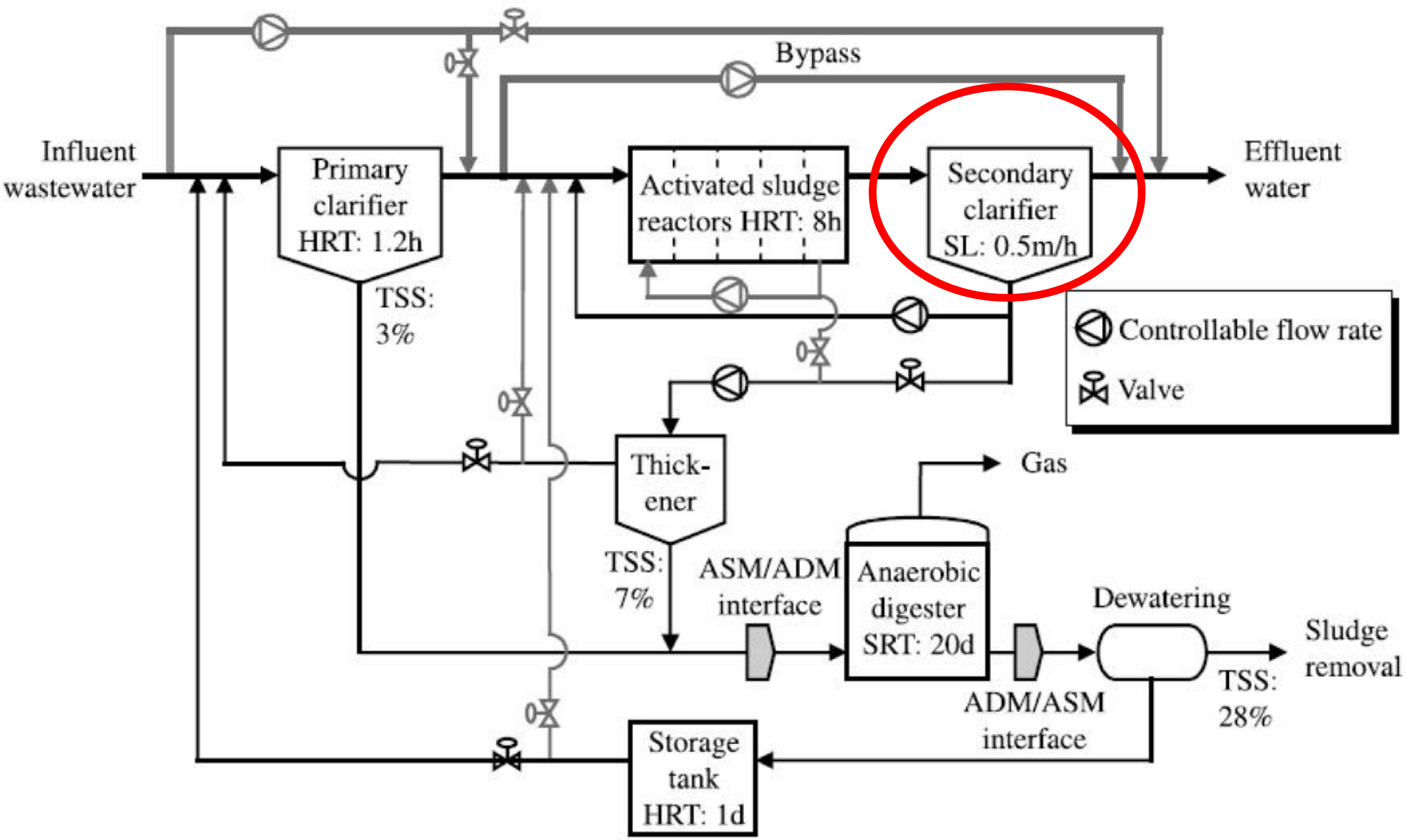


AIC=14238

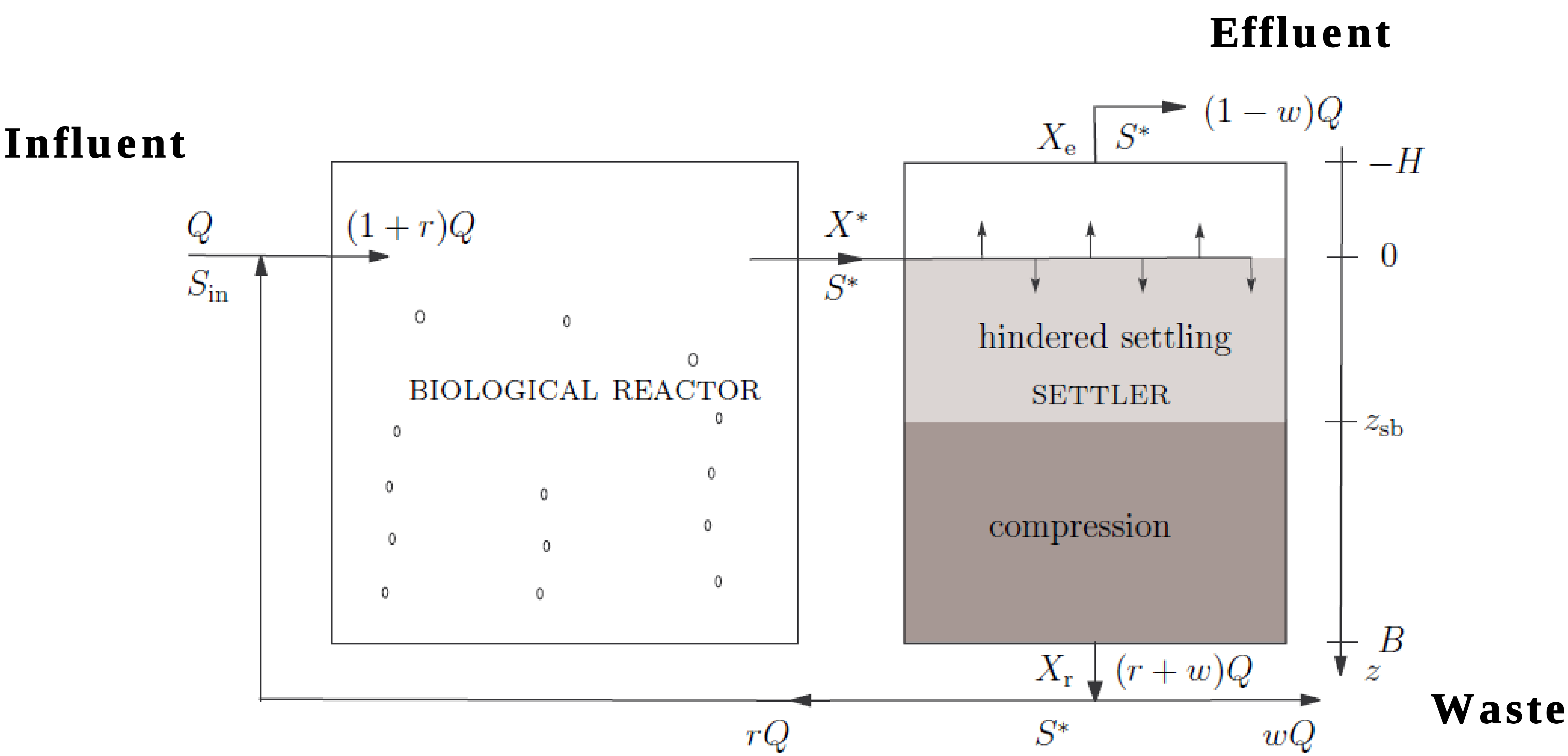
Case study

A wastewater treatment plant

A wastewater treatment plant (cont.)

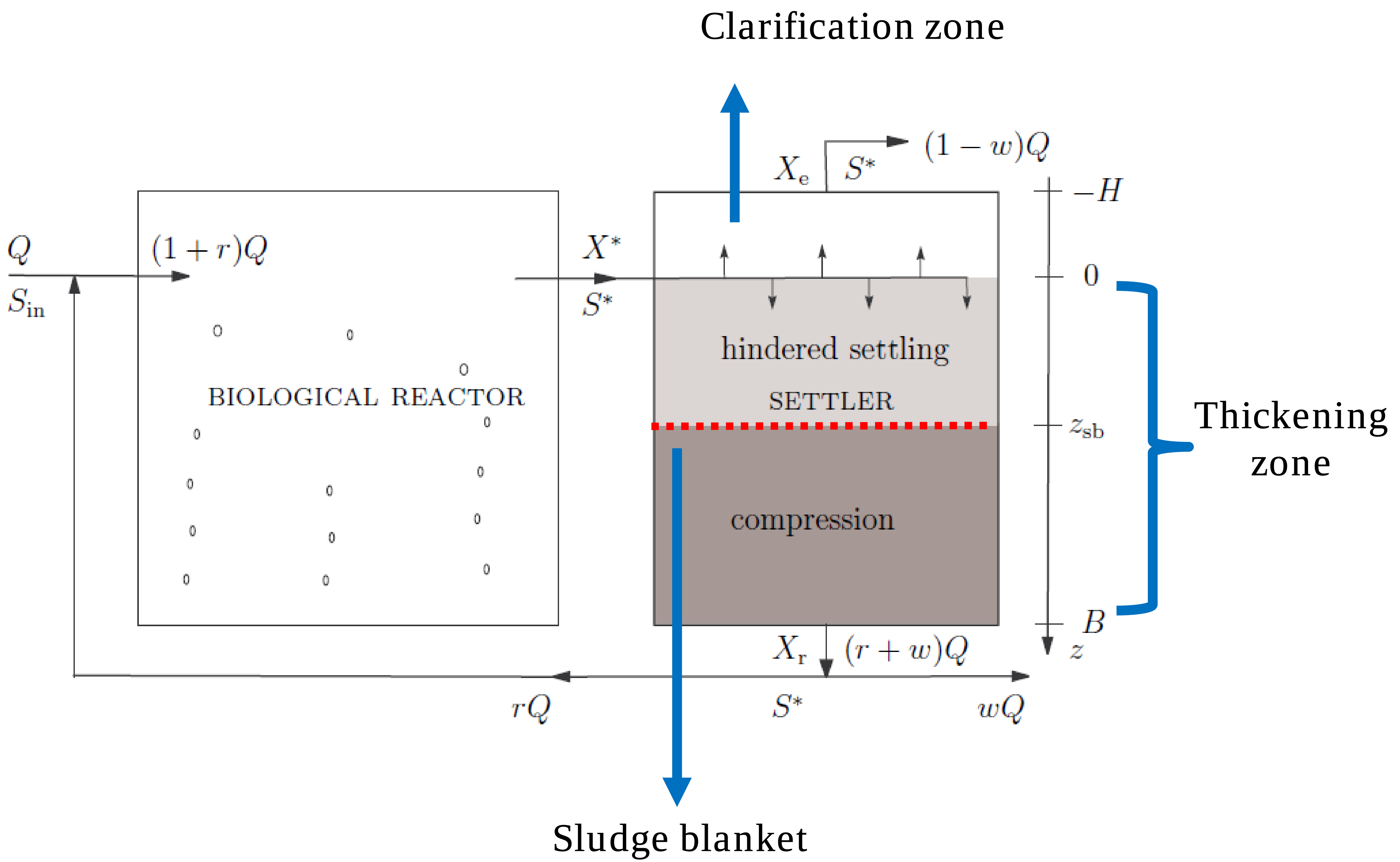


The Process

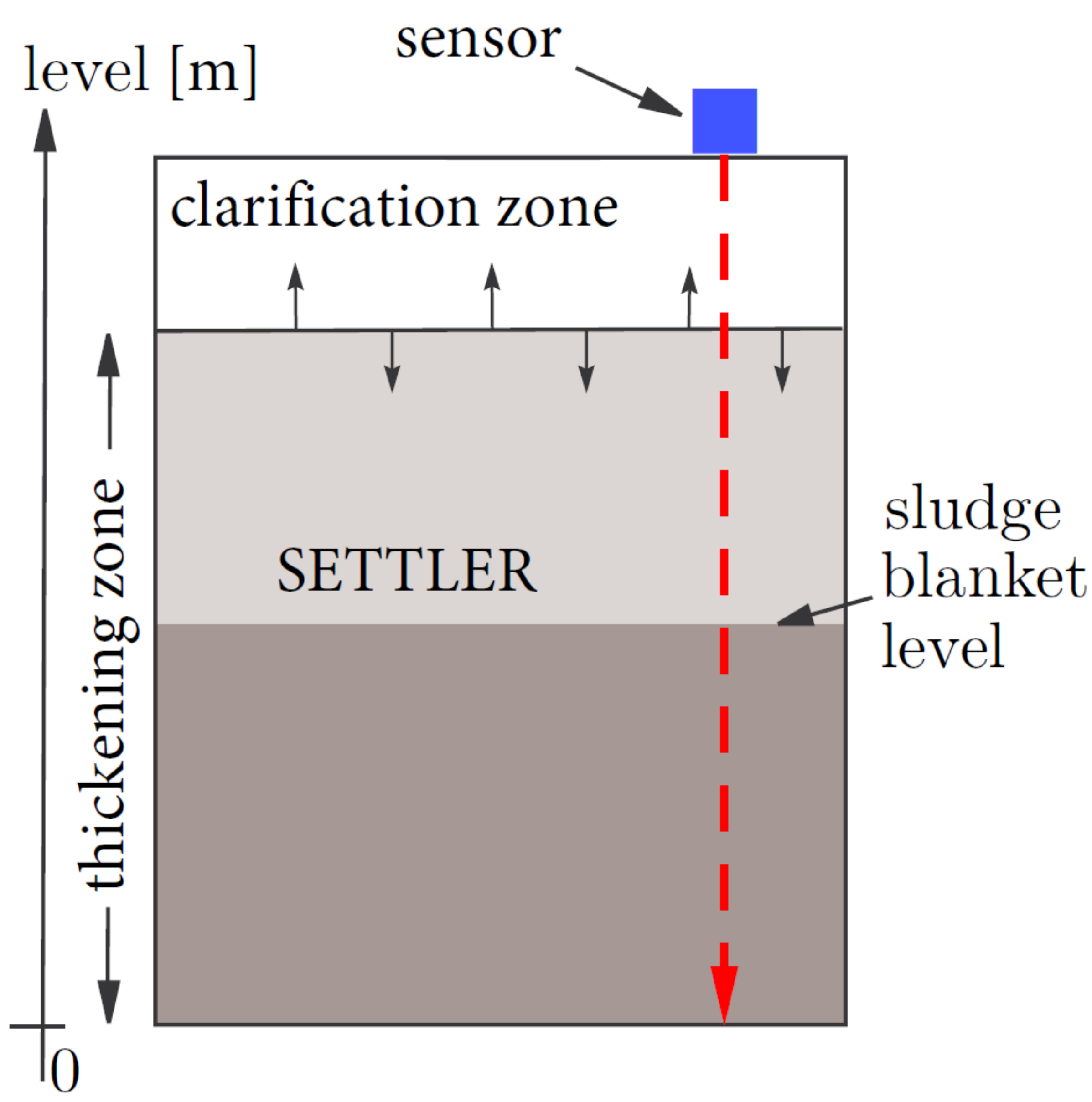


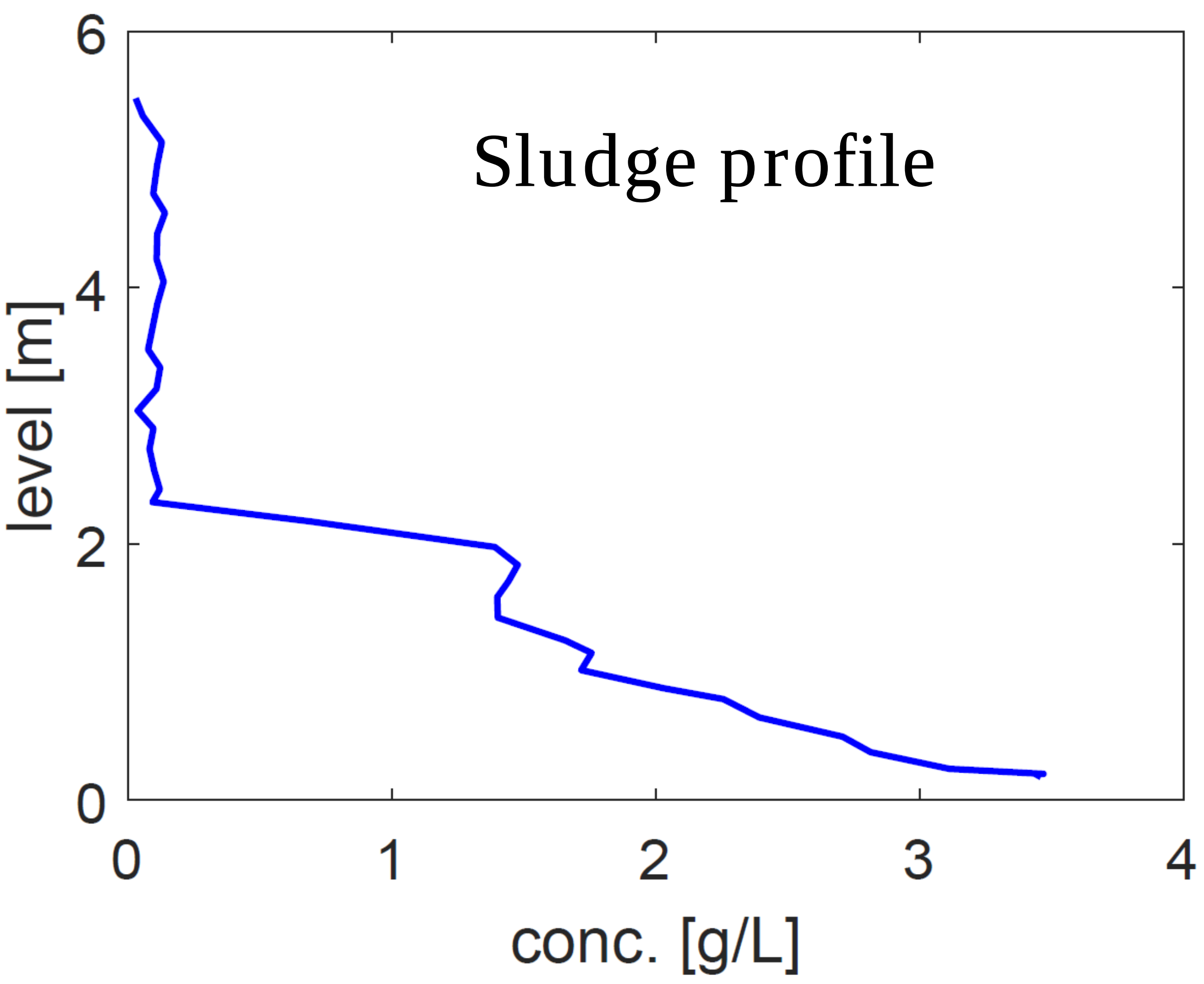
Q: flowrate
S: conc. soluble substrate
X: conc. biomass
r: recycle ratio
w: wastage ratio

The Process (cont.)



Scanning a secondary settler

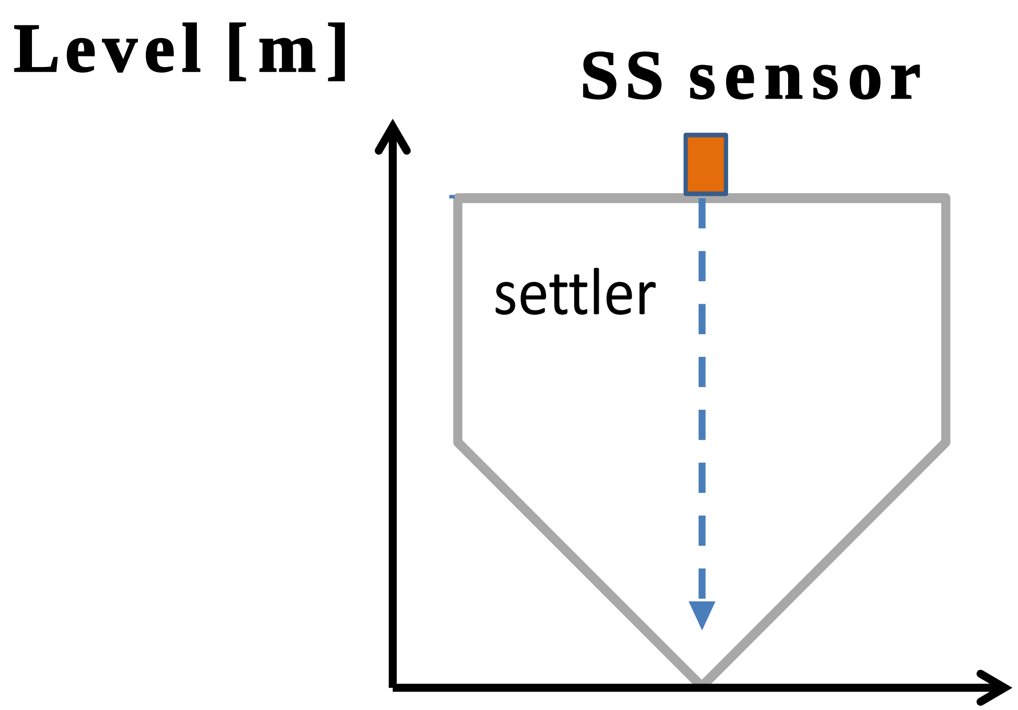


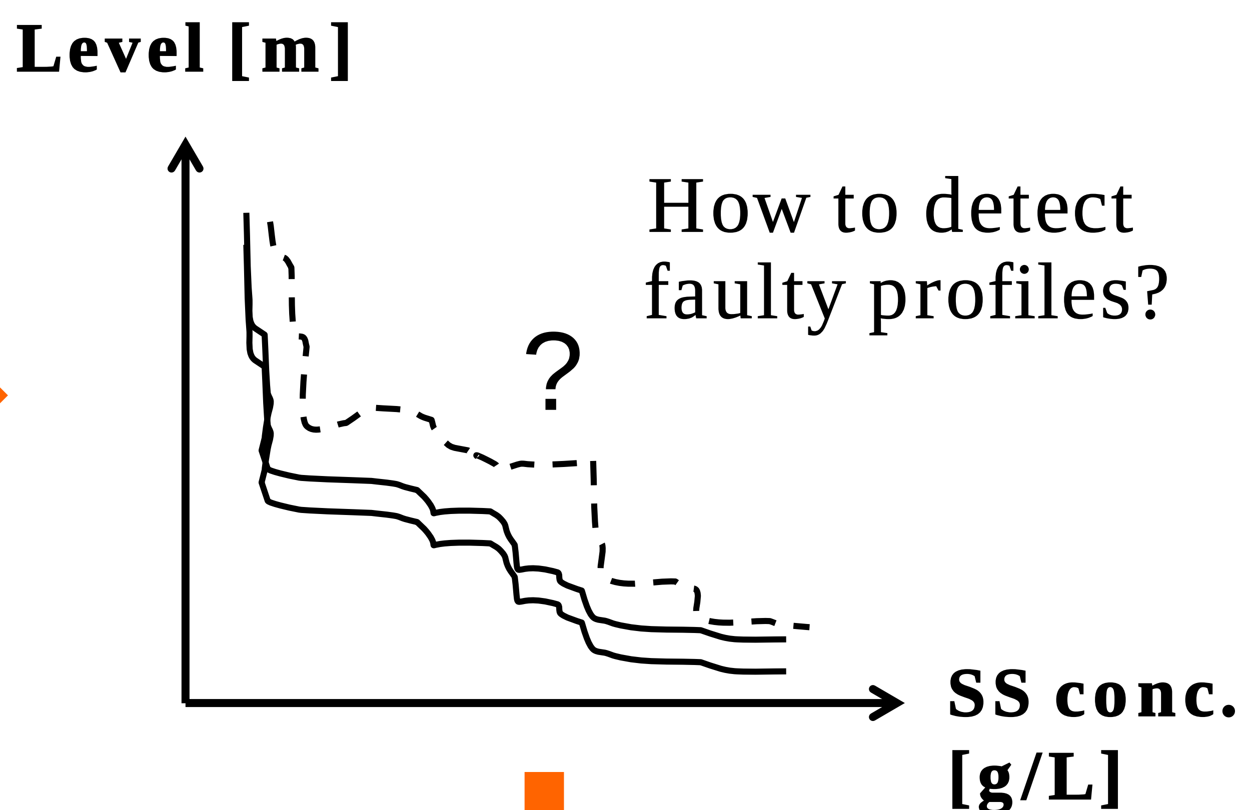


The Problem

Scanning

Sludge profiles

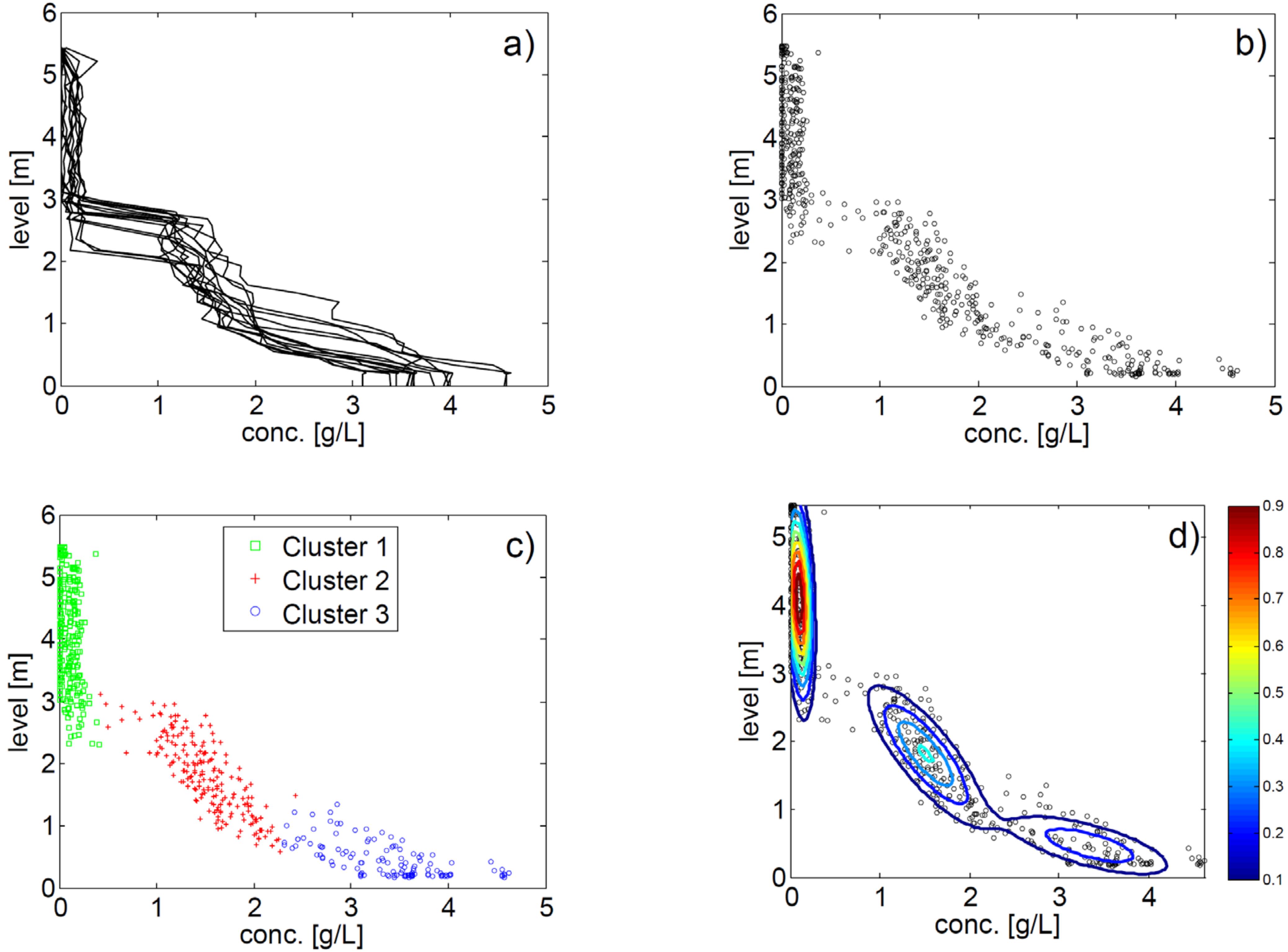




Let’s apply Gaussian Mixture Models!

GMM for the settler

15 sludge profiles in non-faulty conditions



GMM for the settler (cont.)

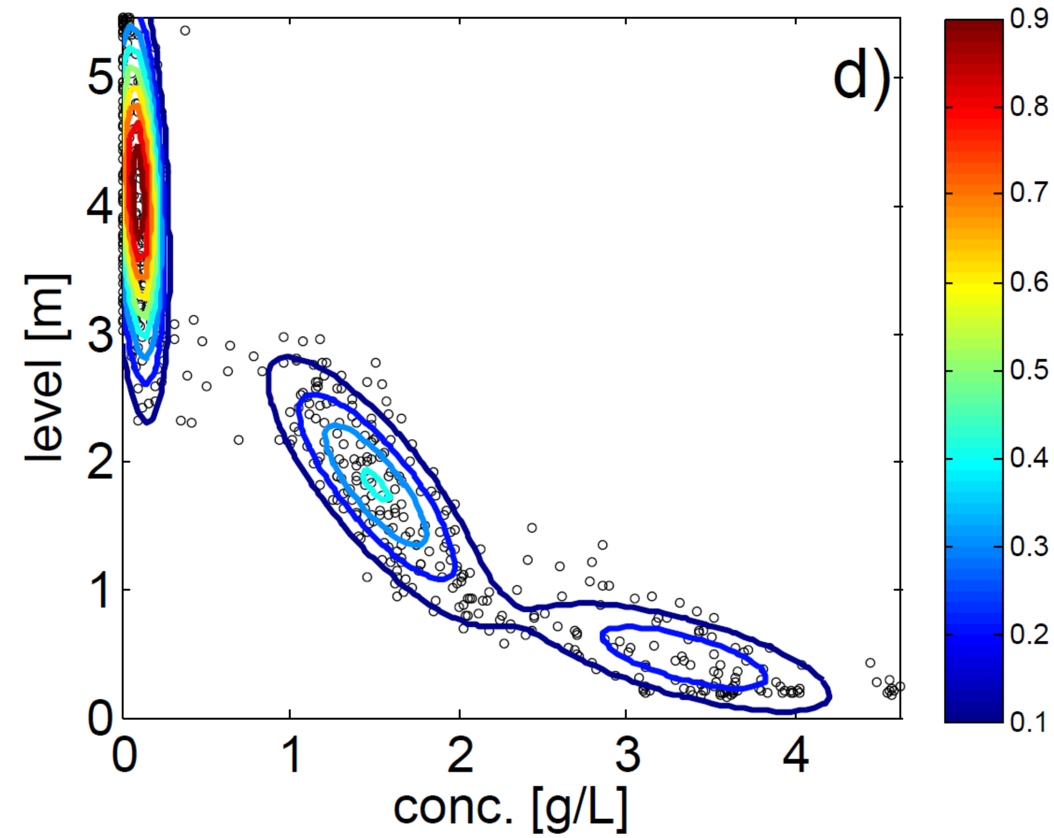
GMM parameters π_k, μ_k, σ_k:

We denote

$$x_1 = \{\text{SS conc.}\} \text{ and } x_2 = \{\text{level}\}$$

$$\mu_k = \begin{bmatrix} \text{mean}(x_1) \\ \text{mean}(x_2) \end{bmatrix},$$

$$\sigma_k = \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) \end{bmatrix},$$

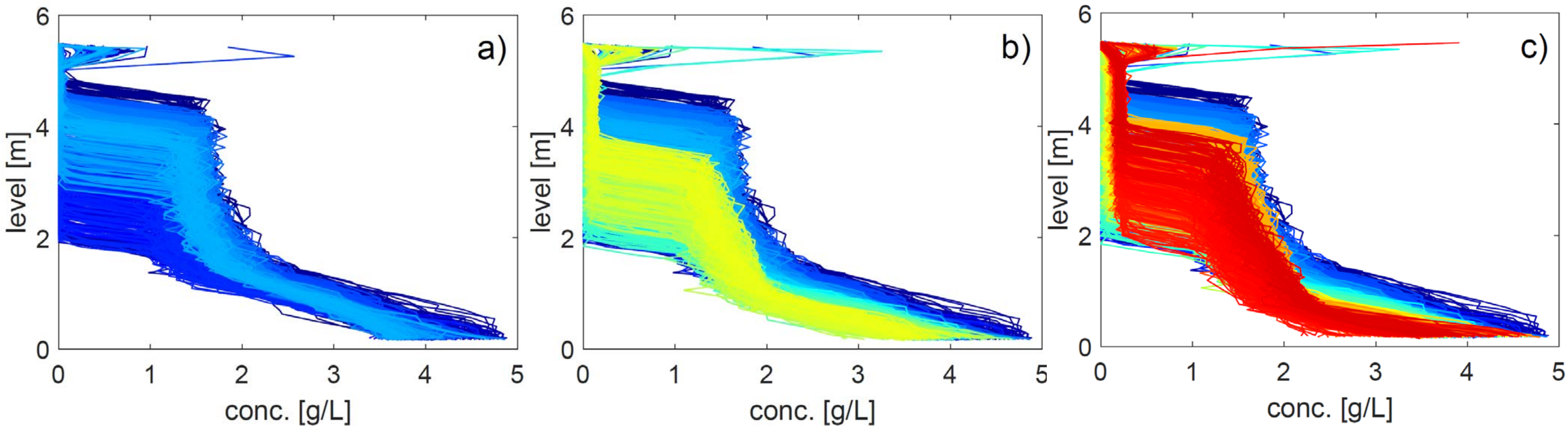


Weight	Mean	Covariance
$\pi_1 = 0.4329$	$\mu_1 = \begin{bmatrix} 0.0958 \\ 4.1102 \end{bmatrix}$	$\sigma_1 = \begin{bmatrix} 0.0074 & -0.0223 \\ -0.0223 & 0.7084 \end{bmatrix}$
$\pi_2 = 0.3405$	$\mu_2 = \begin{bmatrix} 1.5065 \\ 1.8203 \end{bmatrix}$	$\sigma_2 = \begin{bmatrix} 0.1446 & -0.1840 \\ -0.1840 & 0.3550 \end{bmatrix}$
$\pi_3 = 0.2265$	$\mu_3 = \begin{bmatrix} 3.3421 \\ 0.4691 \end{bmatrix}$	$\sigma_3 = \begin{bmatrix} 0.3612 & -0.1208 \\ -0.1208 & 0.0866 \end{bmatrix}$

Settler monitoring

- Sludge profiles from day 1 (blue) to day 33 (red).
- New profile every 15 minutes = 3168 profiles.

Day 1 -10 Day 11 - 20 Day 21 -33



(Red does not mean alarm!)

Residual and Fault detection criteria

Algorithm 2 GMM-based residual calculation

1: Collect a group of M-profiles in non-faulty conditions.
2: Set K and compute the iterative EM algorithm (see Algorithm 1) to get π_k, μ_k, σ_k.
3: **while** monitoring a new profile **do**
4: **for** every profile **do**
5:

threshold

$$r = \frac{1}{p(\mathrm{x}; \pi_{1:K}, \mu_{1:K}, \sigma_{1:K})}, \tag{7}$$

where

$$p(\mathrm{x}; \pi_{1:K}, \mu_{1:K}, \sigma_{1:K}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathrm{x}_n | \mu_k, \sigma_k). \tag{8}$$

6: **end for**
7: **end while**

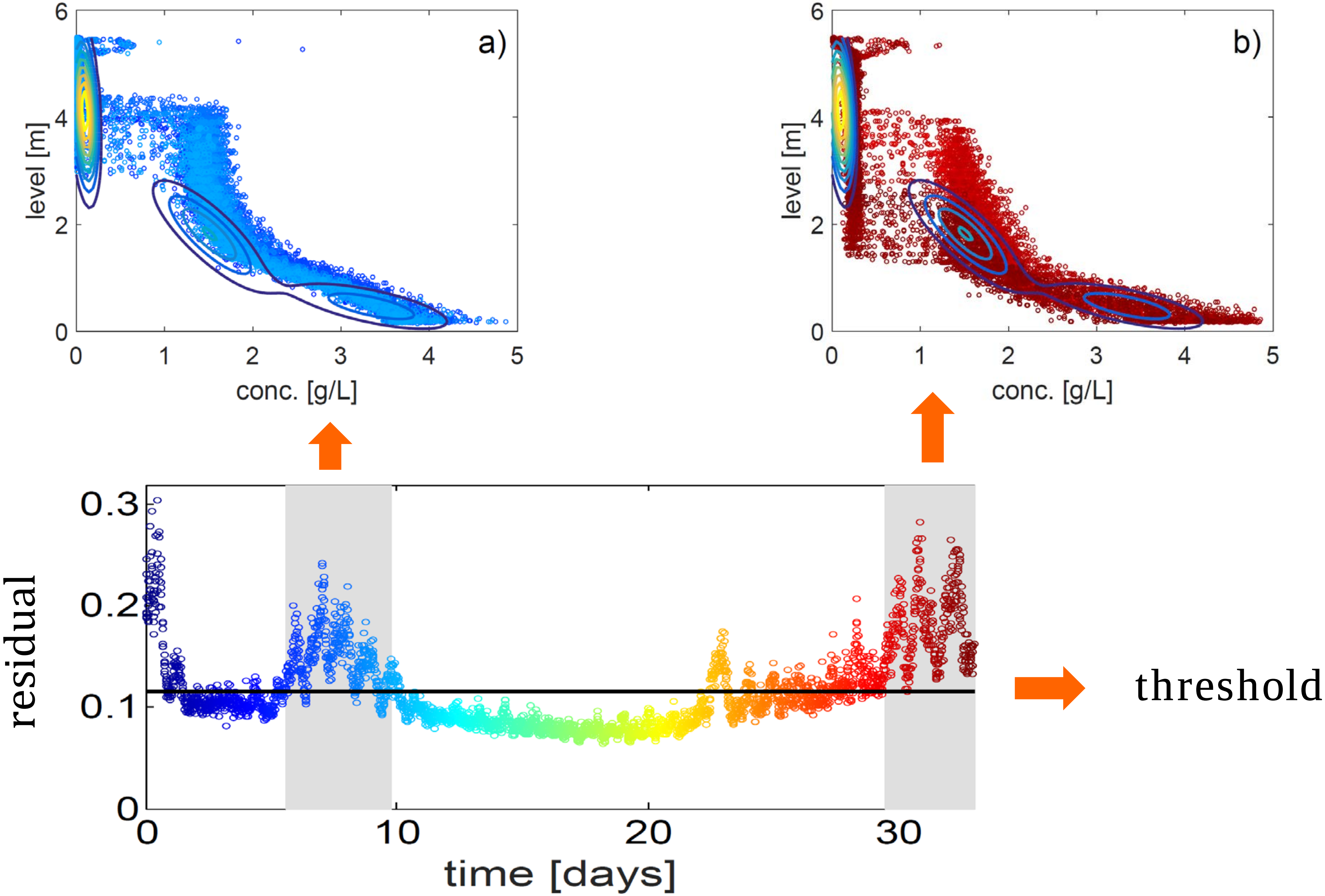
$H_0 : r \le h$ normal
$H_1 : r > h$ faulty!

where

$$h = \max\{r\}\Big|_{t \in H_0}$$

Classical binary hypothesis testing problem

Settler monitoring (cont.)



Conclusions

- Valuable information can be obtained by monitoring a Secondary Settler in a wastewater treatment plant.

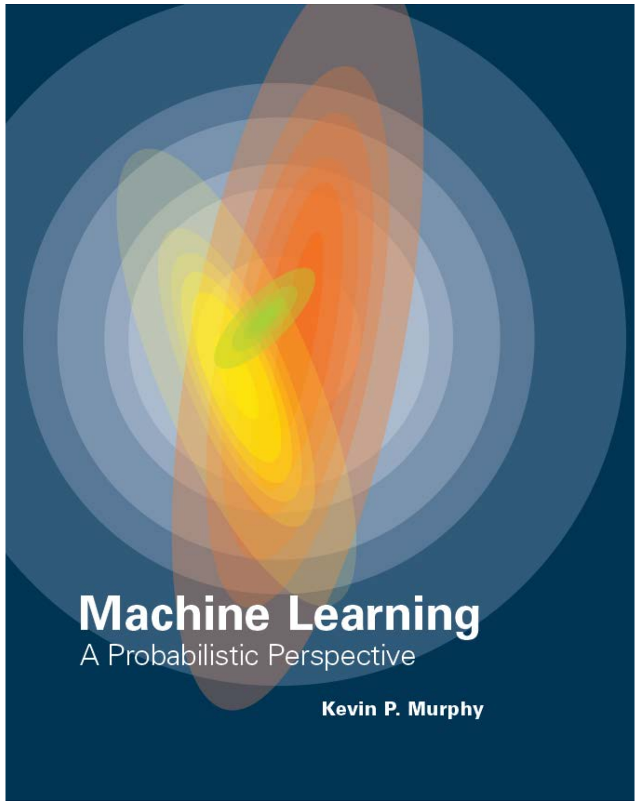
- Gaussian Mixture Models provide a novel tool for fault detection in this process.

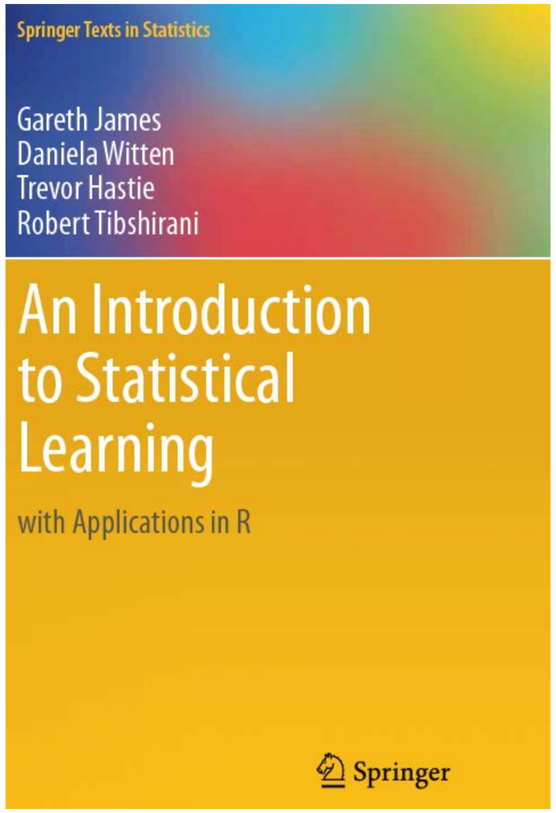
- The proposed method is general and could be implemented in settlers with different geometries and sludge profiles.

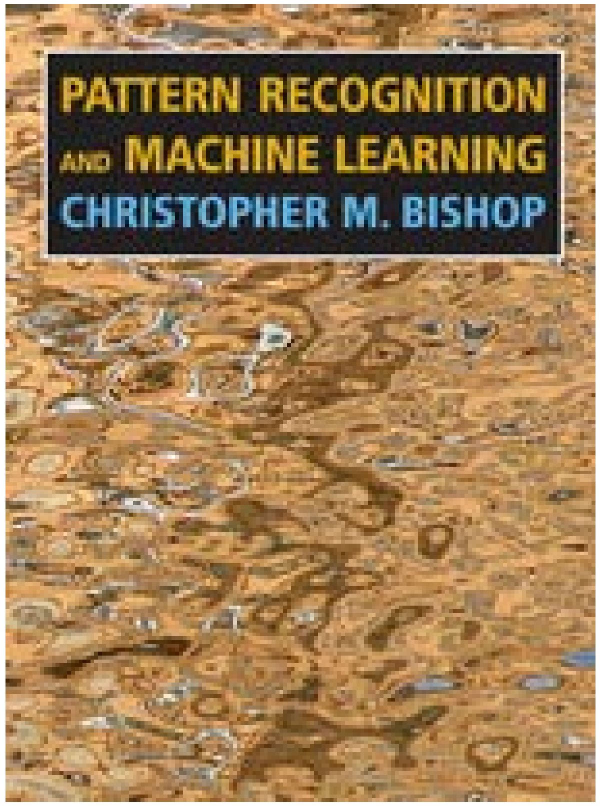
- The method is also suitable for monitoring deviations in a process with repetitive data profiles.

Sources of information

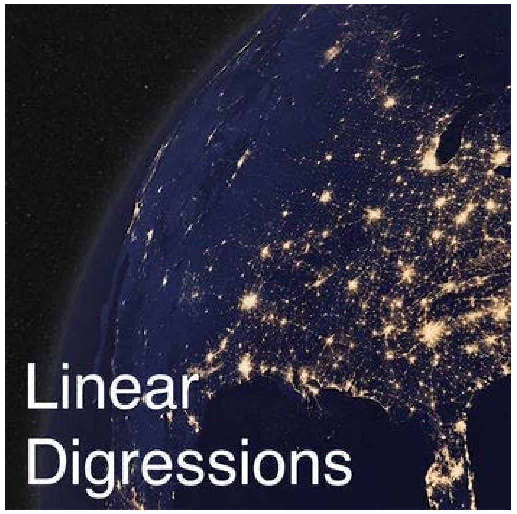
- Books:



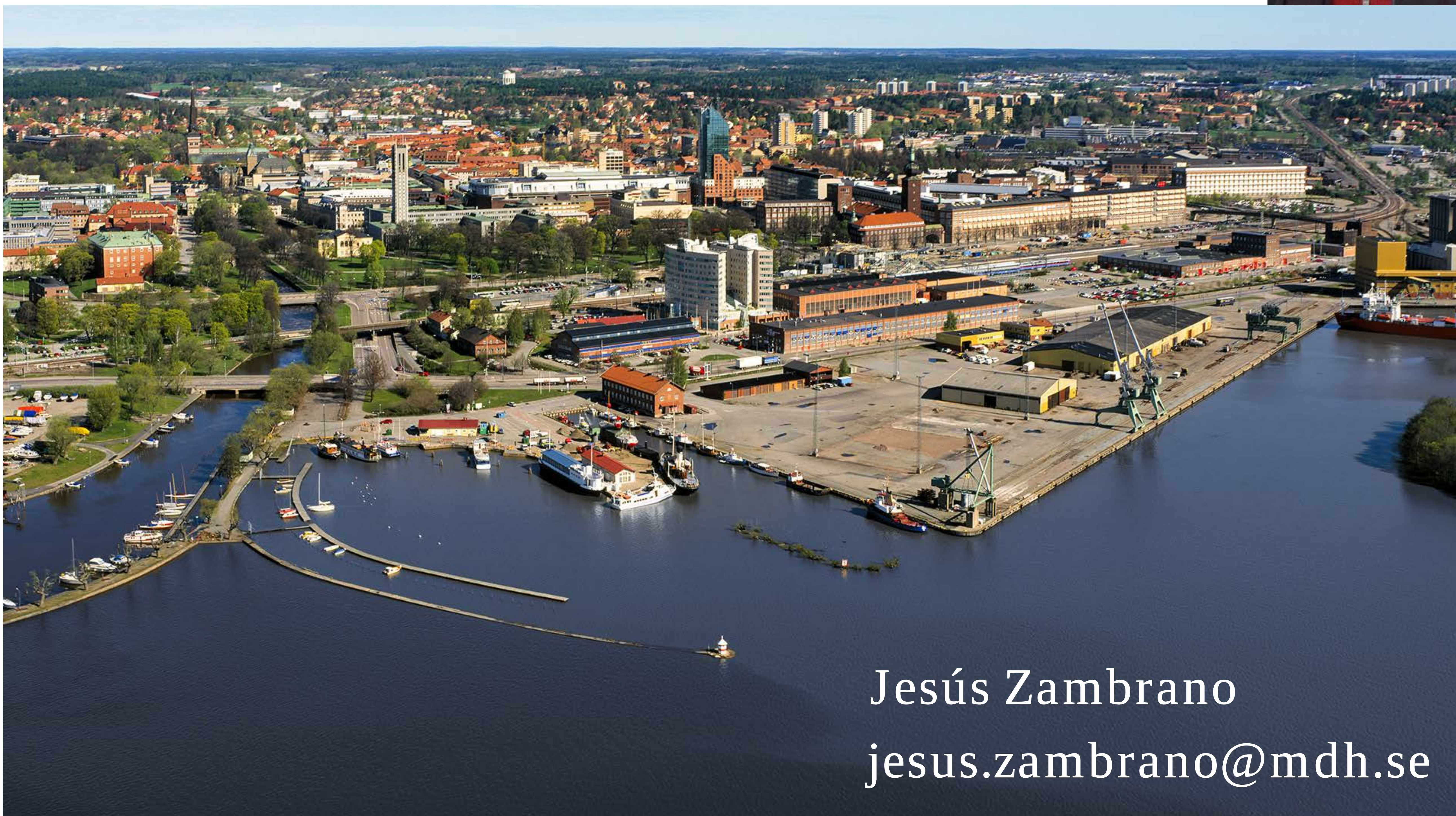




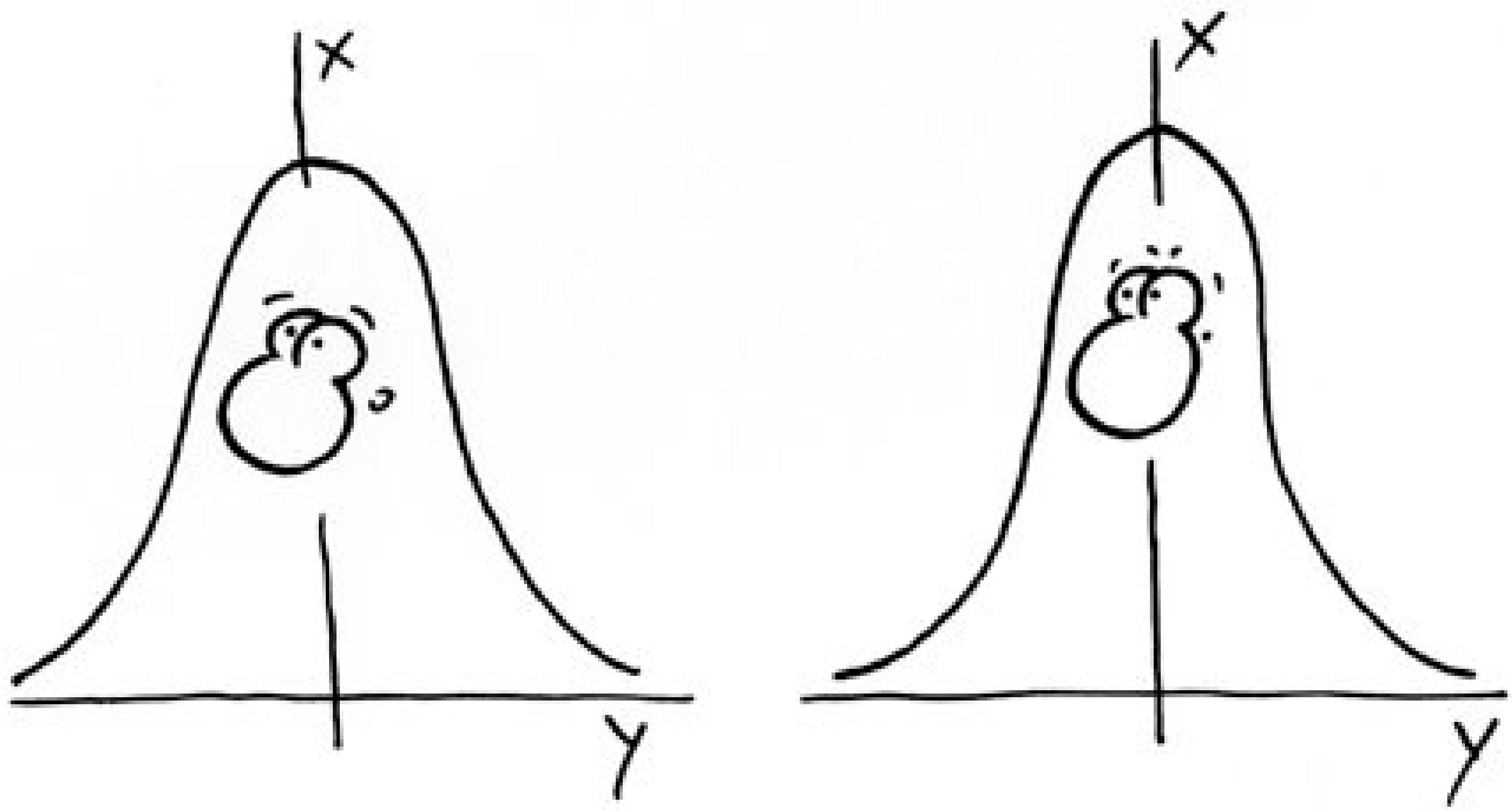
- Podcasts:



Thanks for your attention!



Jesús Zambrano
jesus.zambrano@mdh.se



"I always feel so normal, so bored, you know. Sometimes I would like to do something... you know... something... mmm... Poissonian."