Anomaly Detection in large scale networks
State of the art and trend

Kavé Salamatian, Lancaster University, ECODE Project

Outline

- Introduction & Methodologies
- Basics Anomaly
- Parametric methods
- Non Parametric Methods
- Distributed anomaly detection

Introduction and Methodologies

Network state

- Network state results from
  - Traffic demand
    - Traffic matrix
      - Not observable directly
  - Capacity offer
    - Routing matrix, link cap., Traffic Engineering, etc.
      - Monitored by SNMP, etc.
- Network manager goal
  - To drive this equilibrium to the best beneficial point
    - By managing capacity offer
      - Traffic engineering is the art of managing offered capacity

Network monitoring

- Monitoring ?
  - Being able to separate
    - What is predictable
      - Expected, normal, under control, ...
    - What is not predictable
      - Unexpected, abnormal, ...
- Interpretation framework
  - Only what is unpredictable have a meaning
    - What can be predicted does not have a information

Classes of anomaly detections

- Deterministic approaches
  - Signature based
    - Complexity vs. exhaustivity trade-off
- Statistical approaches
  - Probabilistic
    - False positive vs. detection rate trade - off

Deterministic approaches

- Each observation is assumed to result from a known causality chain
  - All anomalies are characterized by causality chain that describes the signature of an anomaly
  - Anomaly detection consist of back tracking the causality sequence
- Need exhaustive anomaly signature database
  - Have to check all possible signature
    - Not scalable

Statistical approach

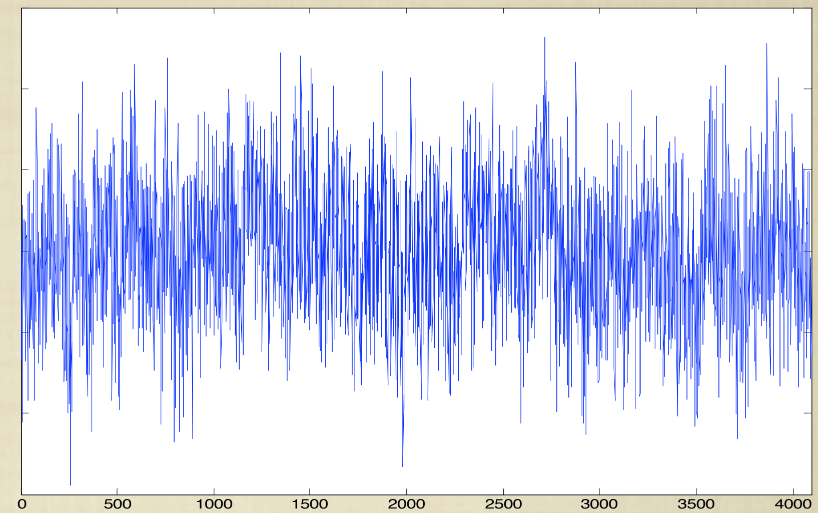
- a dataset is a unique sequence of deterministic observation that is not anymore perfectly reproducible
- Statistical approach assumes that
  - Observation results from a random function
    - they come from a random choice in a (in)finite set of «possible» observations.
  - Is this assumption sound ?
- Gives access to an arsenal of probabilistic methods
  - Stationarity : Intrinsic hypothesis
    - Statistical properties hold over time
      - Bet on future
    - Possibility of radical error
  - Essential False alarm vs. misdetection trade-off

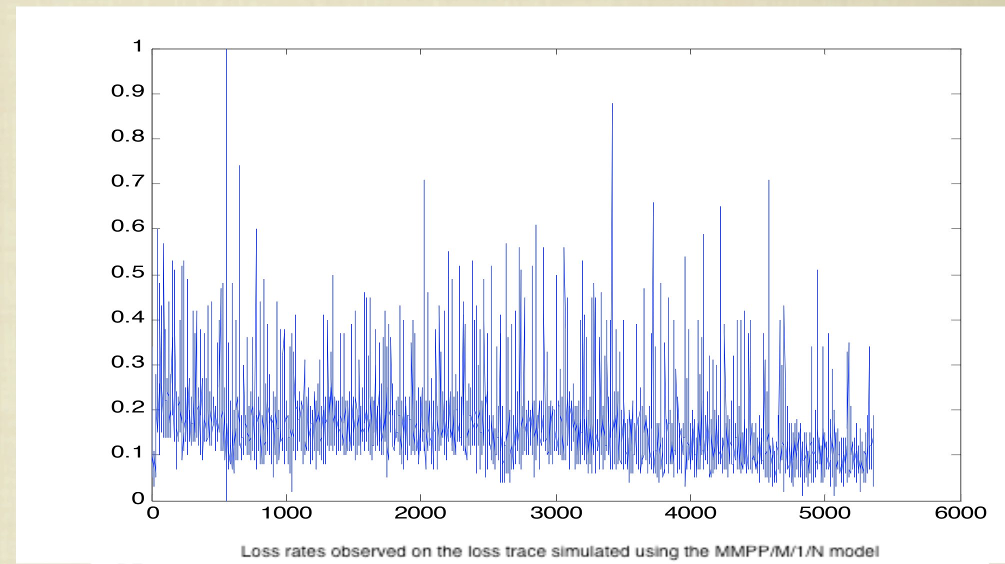
Empirical modeling Challenge

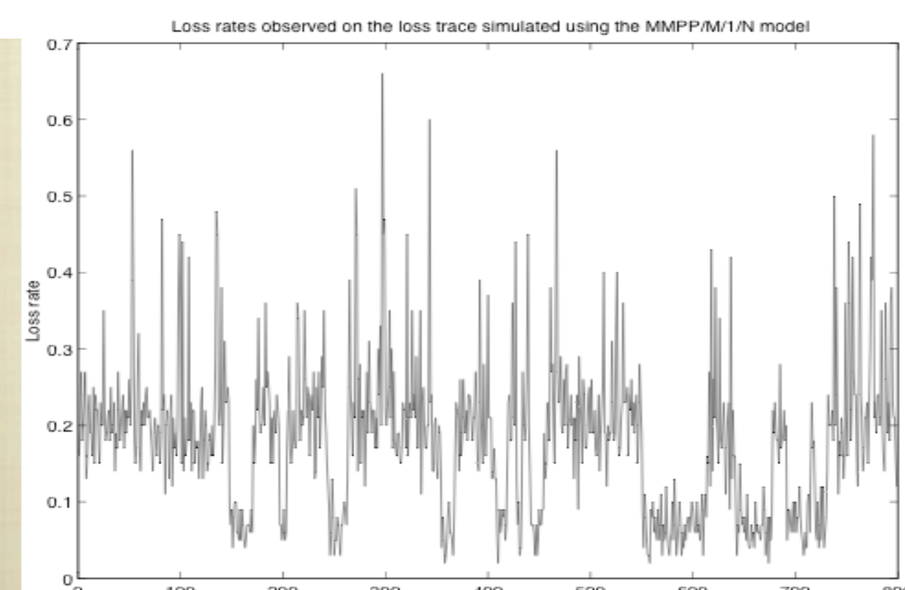
- Measurement contains two components
  - A structured component that reflects essential properties of the phenomenon under study
  - A random component that represent fluctuations that are put aside from the model
- Modeling challenge
  - Separate structure from randomness and characterize it

Interpretation

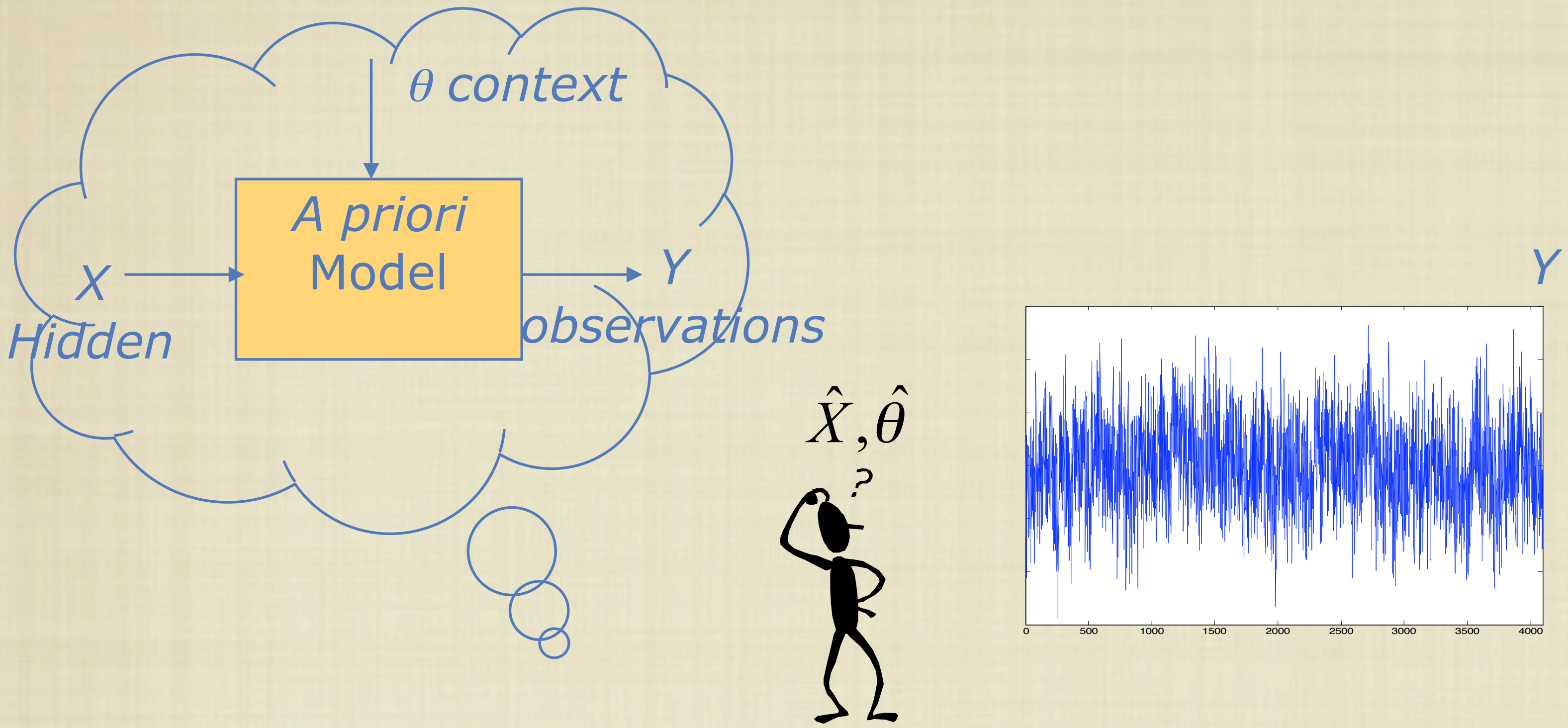
- To relate effect to causes
  - Enable prediction
    - At different time scale
  - Reaction
    - When causes are known one can affect them







Interpretation framework



- **What are the hidden causes and the hidden context that leads to the observation**
  - The understanding of the phenomenon is condensed in the a priori model $Y=M(x,\theta)$

Interpretation framework

- Two statistical inverse problems
  - Modeling problem
    - What is the context parameter θ
  - Interpretation problem
    - What is the hidden input X

Statistical Anomaly detection Basics

Anomalies ???

- What is an anomaly?
  - What is normal?
    - What is likely?
- Assumptions
  - Anomaly are anomalous
    - different from the norm
  - Anomalies are rare
  - Anomalous activity is malicious
    - Reverse assumption : anomaly represents attacks or malicious behavior
  - Are these assumptions correct ???

What is Likely ?

- Typical set
  - The set of observed sequence Y_k^{n+k} such that

$$2^{-n(\overline{H}(Y_k^{k-n})-\varepsilon)} \leq \Pr\left\{Y_k^{k+n}\right\} \leq 2^{-n(\overline{H}(Y_k^{k-n})+\varepsilon)}$$

$$\Pr\left\{A_\varepsilon^{(n)}\right\} \geq 1-\varepsilon$$

  - From AEP

$$(1-\varepsilon)2^{-n(\overline{H}(Y_k^{k-n})+\varepsilon)} \leq \left|A_\varepsilon^{(n)}\right| \leq 2^{n(\overline{H}(Y_k^{k-n})+\varepsilon)}$$

Universal anomaly detector !

- A simple anomaly detector
  - Calculate the likelihood $\Pr\left\{Y_k^{k+n}\right\}$ of an observed sequence
  - Check if $2^{-n(\overline{H}(Y_k^{k-n})-\varepsilon)} \leq \Pr\left\{Y_k^{k+n}\right\} \leq 2^{-n(\overline{H}(Y_k^{k-n})+\varepsilon)}$
  - If not, it is an anomaly
- Non parametric and universal anomaly detector
  - How to calculate the likelihood ?
    - Need simplifying hypothesis

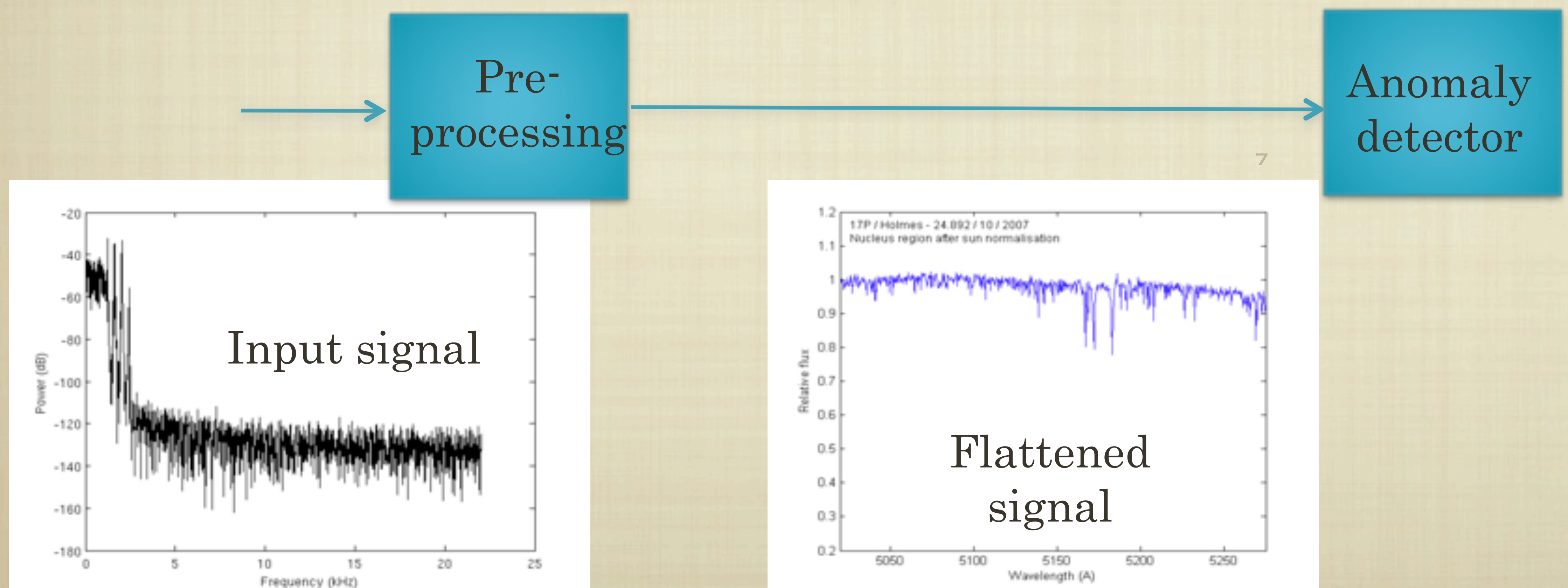
Filtering

- What if the observations are independent !

$$\Pr\left\{Y_k^{k+n}\right\} = \left(\Pr\left\{Y_k\right\}\right)^n$$

- We have to transform the observation so that it become independent !

- Basic signal processing question



Anomaly detector structure

- two generic steps
  - Entropy reduction/Filtering
    - Remove dependencies in data
      - Compress the data:Information bottleneck
        - Remove predictable component
  - Decision/Detection
    - Apply the typicality test on the resulting hopefully independent signal
      - Classical test in the Neyman-Pearson Framework

Taxonomy of approaches

- Parametric
  - Assume a parametric dependencies model
- Non Parametric
  - Broke the dependencies in a universal way

Parametric Approaches

- Parametric Approaches
  - Assumes a parametric structure for dependencies
    - Normal behavior model
  - Calibrate the Model
    - MLE estimation
  - Filter the model prediction from observation
    - Results in an IID gaussian innovation process
  - Apply a Neyman-Pearson statistical test with H0: the observation is compatible with a zero mean with given variance

Non parametric approach

- Apply a de-correlating transform
  - Random projection, hashing, Sketch, etc...
- Learn the resulting distribution of normal behavior
  - Clustering, SVM, etc...
- detect anomalies by checking cluster membership

Parametric techniques

ANOMALY DETECTION STEPS

- Modeling
- Filtering
- Decision

Model structure

- First step is to define a correlation/model structure
  - Temporal, spatial correlationS
  - The traffic is a dynamic entity
    - Generic dynamical model
    - Linear approximation
      - LTI Model

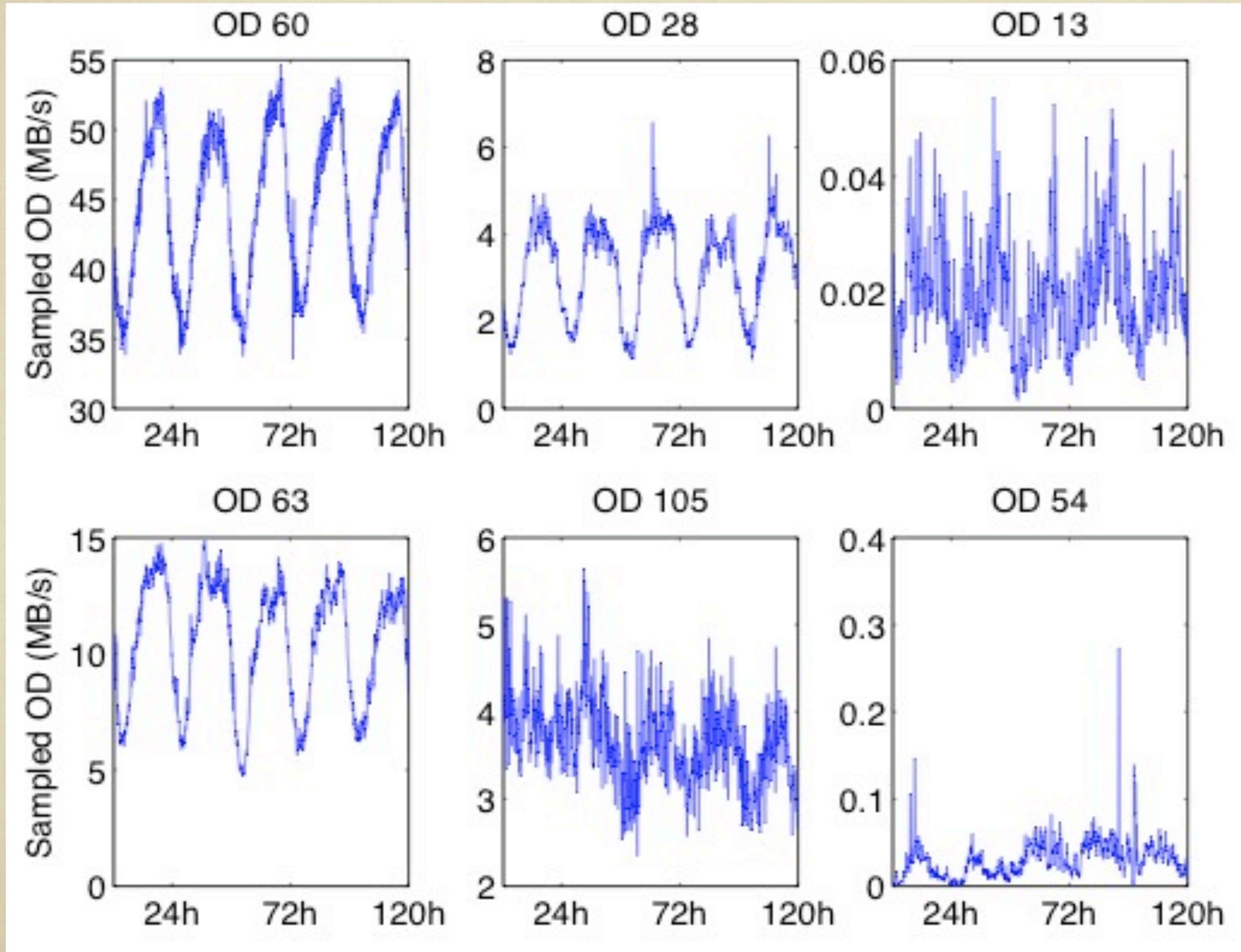
$$\begin{cases} X_{k+1} = f(X_k) + g(U_k) \\ Y_k = h(X_k) + i(U_k) \end{cases}$$

$$\begin{cases} X_{k+1} = AX_k + BU_k \\ Y_k = CX_k + DU_k \end{cases}$$

$$X_k = \left(x_k^1, x_{k-1}^1, \ldots., x_{k-T}^1, x_k^2, x_{k-1}^2, \ldots., x_{k-T}^2, x_k^1, \ldots., x_{k-1}^L, \ldots., x_{k-T}^L\right)$$

- A well behaved non-linear process with any memory can be approximated by a large enough linear model

Traffic Dynamic



Stochastic models

- State model:

$$\begin{cases} X(t+1) & = & C*X(t)+W(t) \\ Y(t) & = & A*X(t)+V(t) \end{cases} \begin{cases} W & \sim & N(0,Q) \\ V & \sim & N(0,R) \end{cases}$$

- How to calibrate C, Q and R?
  - Maximum Likelihood method (EM)
    - Find values of C, Q, R that maximize the likelihood of the observations
  - PCA Method
    - Find values of C, Q, R that give the best k-dimensional approximation

PCA Method

- PCA Theorem

$$\mathbf{X} = \sum_{i=1}^{K} Y_i \phi_i$$

- Karhunen-Loeve Theorem

$$X_l(t) = \sum_{i=1}^{K} \sum_{j=1}^{\infty} Y_{i,j}^{l} \Phi_{i,j}(t)$$

- KLT/PCA based model

$$\hat{X}_l(kT) = \sum_{i=1}^{L} \sum_{j=1}^{M} Y_{i,j}^{l} \Phi_{i,j}^{k}.$$

PCA METHOD

- How to derive the basis
  - PCA

$$\Sigma\phi_i = \lambda_i \phi_i$$

  - KLT

$$\sum_{i=1}^{K} \int_a^b \sigma_{i,l}(s)\Phi_{i,j}(s-t)ds = \lambda_{l,j}\Phi_{l,j}(t),\ j > 0.$$

    - Galerkin method : Apply pca to

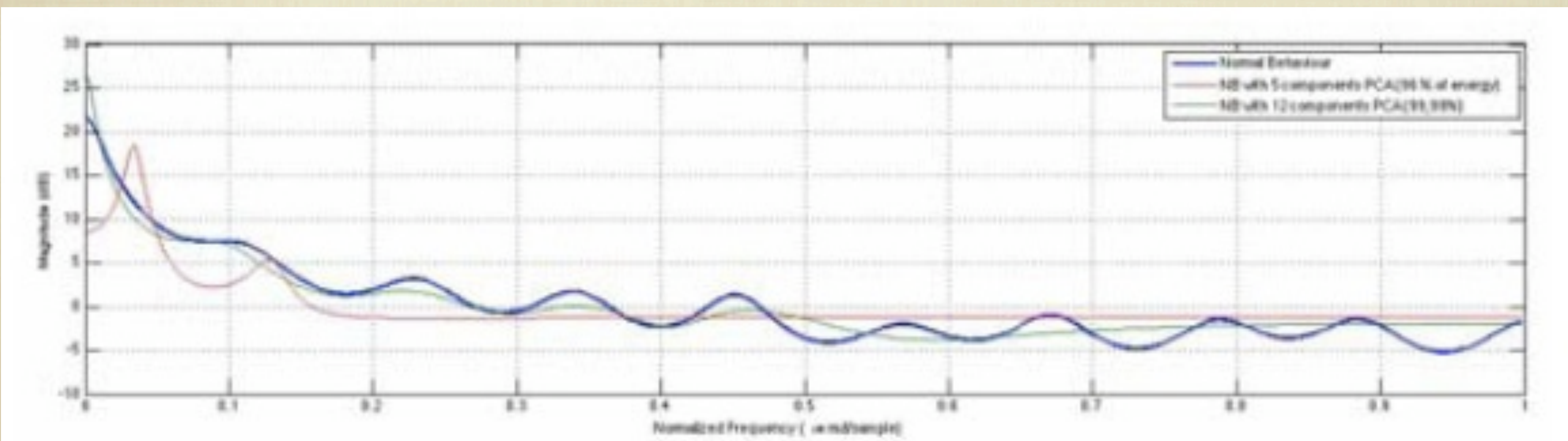
$$\mathbf{x} = \begin{pmatrix} x_1(1) & \dots & x_1(n-N) \\ x_1(2) & \dots & x_1(n-N+1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \dots & x_1(n) \\ x_2(1) & \dots & x_2(n-N) \\ \vdots & \ddots & \vdots \\ x_2(N) & \dots & x_2(n) \\ \vdots & \ddots & \vdots \\ x_K(1) & \dots & x_K(n-N) \\ \vdots & \ddots & \vdots \\ x_K(N) & \dots & x_k(n) \end{pmatrix}$$

Demonstration

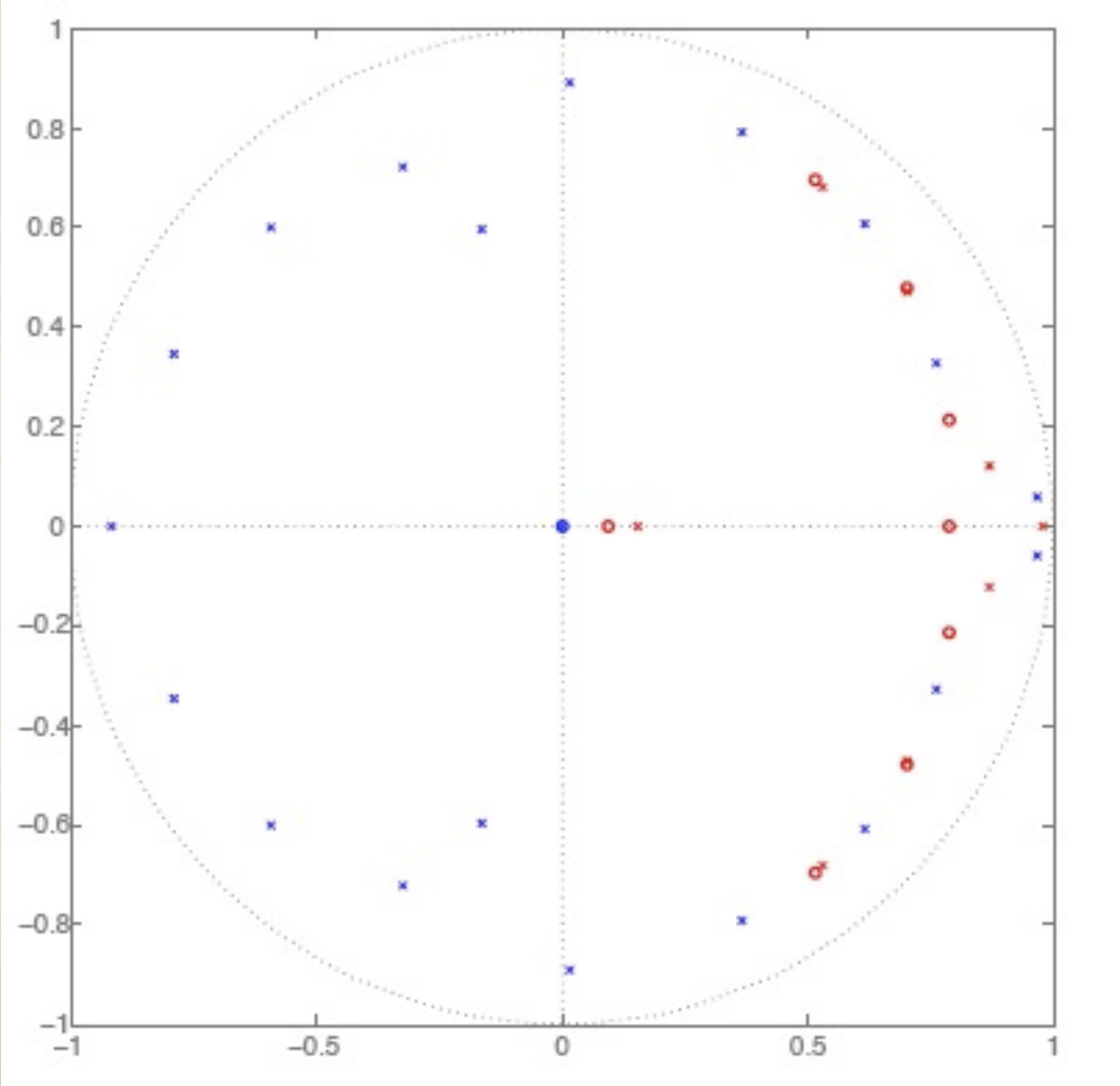
- MATLAB

Normal behavior ?

- What is the normal behavior characterized by the model
  - Time and frequency representation are equivalent



PCA Model

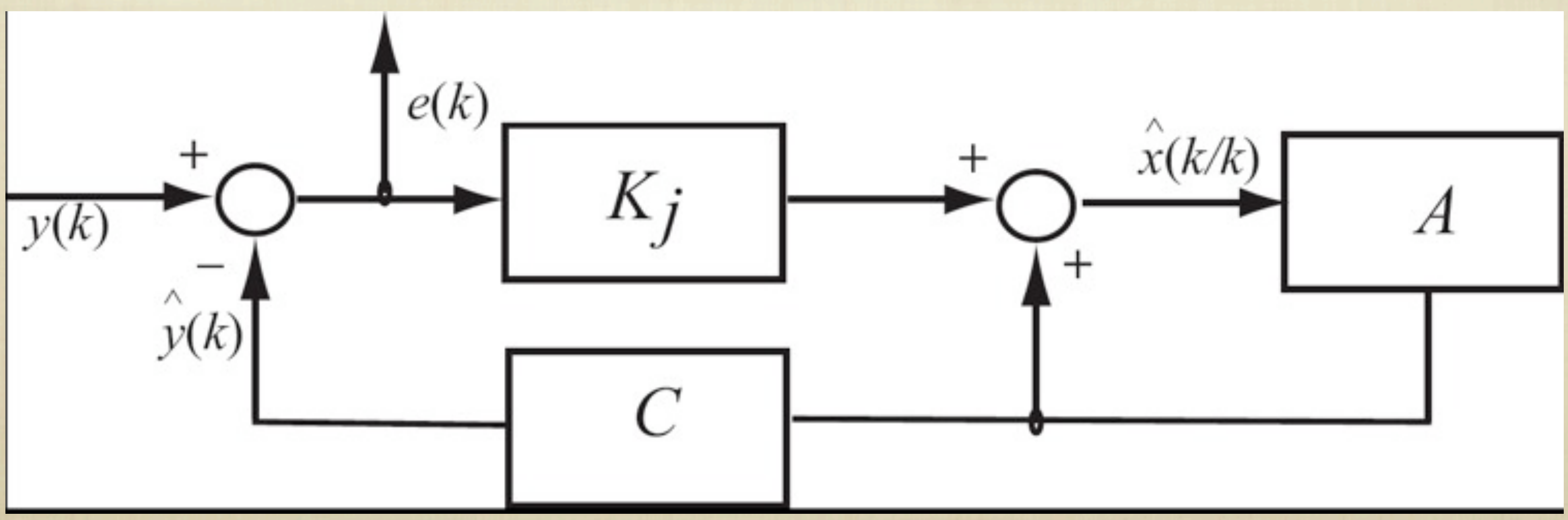


Filtering

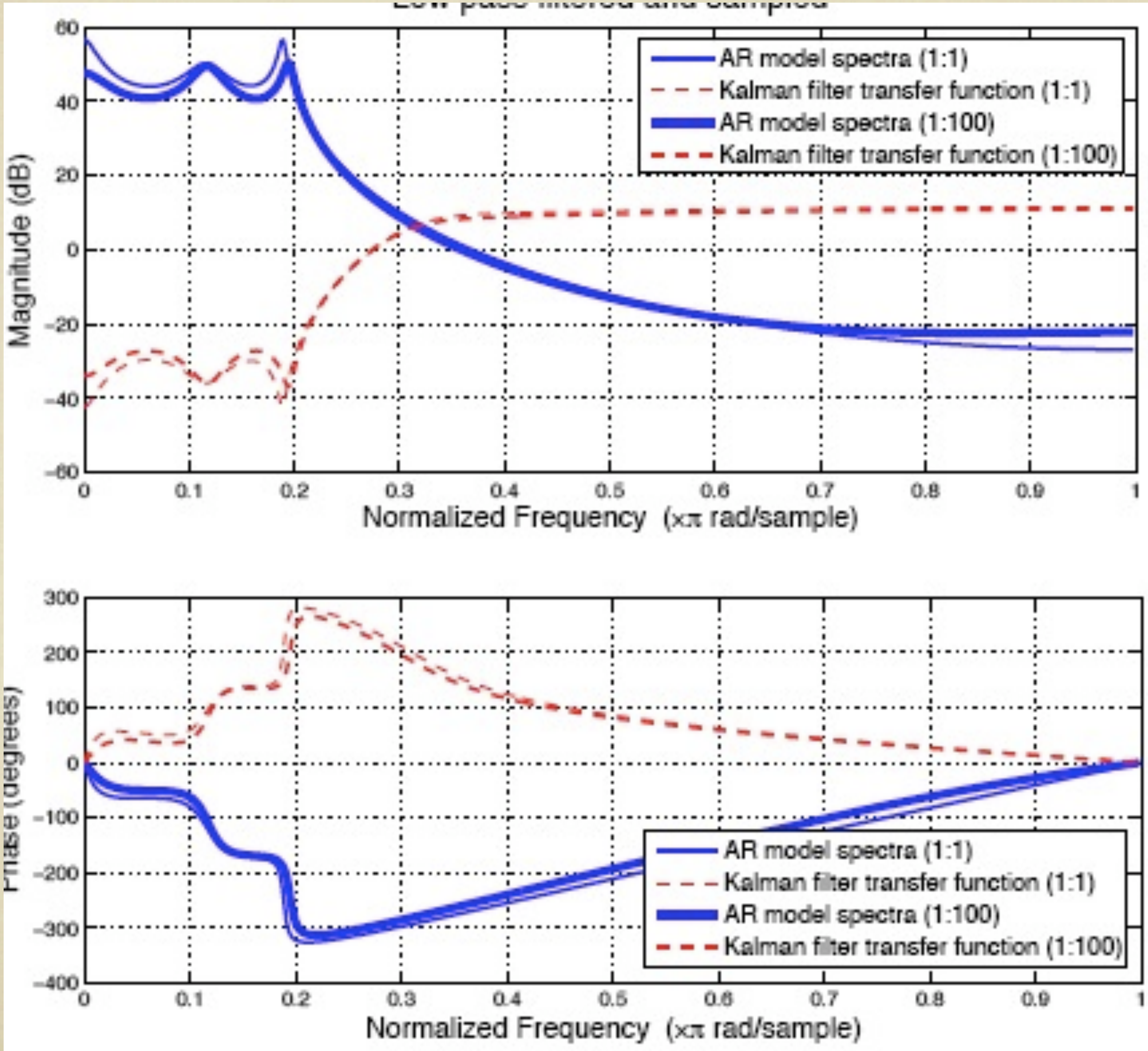
- Filter separate the signal space into two components
  - signals that pass through the filter
  - Signals that are rejected by the filter
- Anomaly detection Filter
  - pass everything is compatible with the normal behavior
  - block everything is not compatible with the normal behavior

Kalman Filter

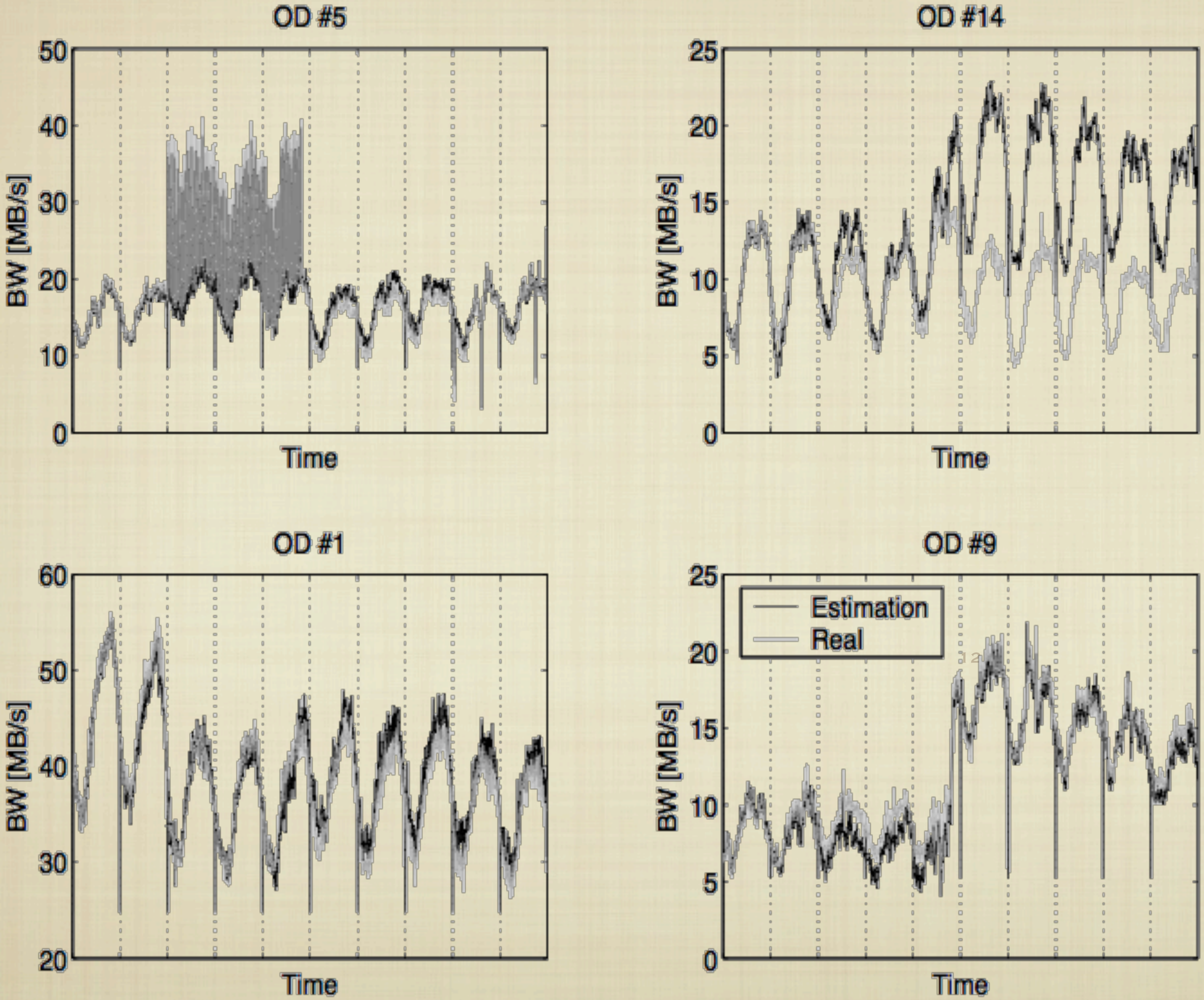
- Filtering what is compatible with the model
  - Two steps
    - Prediction $\hat{X}^{-}(t+1) = C\hat{X}(t)$
    - Correction $\hat{X}(t+1) = \hat{X}^{-}(t+1) + K(t+1)\left(Y(t+1) - A\hat{X}^{-}(t+1)\right)$
  - Innovation $\eta(t+1) = Y(t+1) - A\hat{X}^{-}(t+1)$



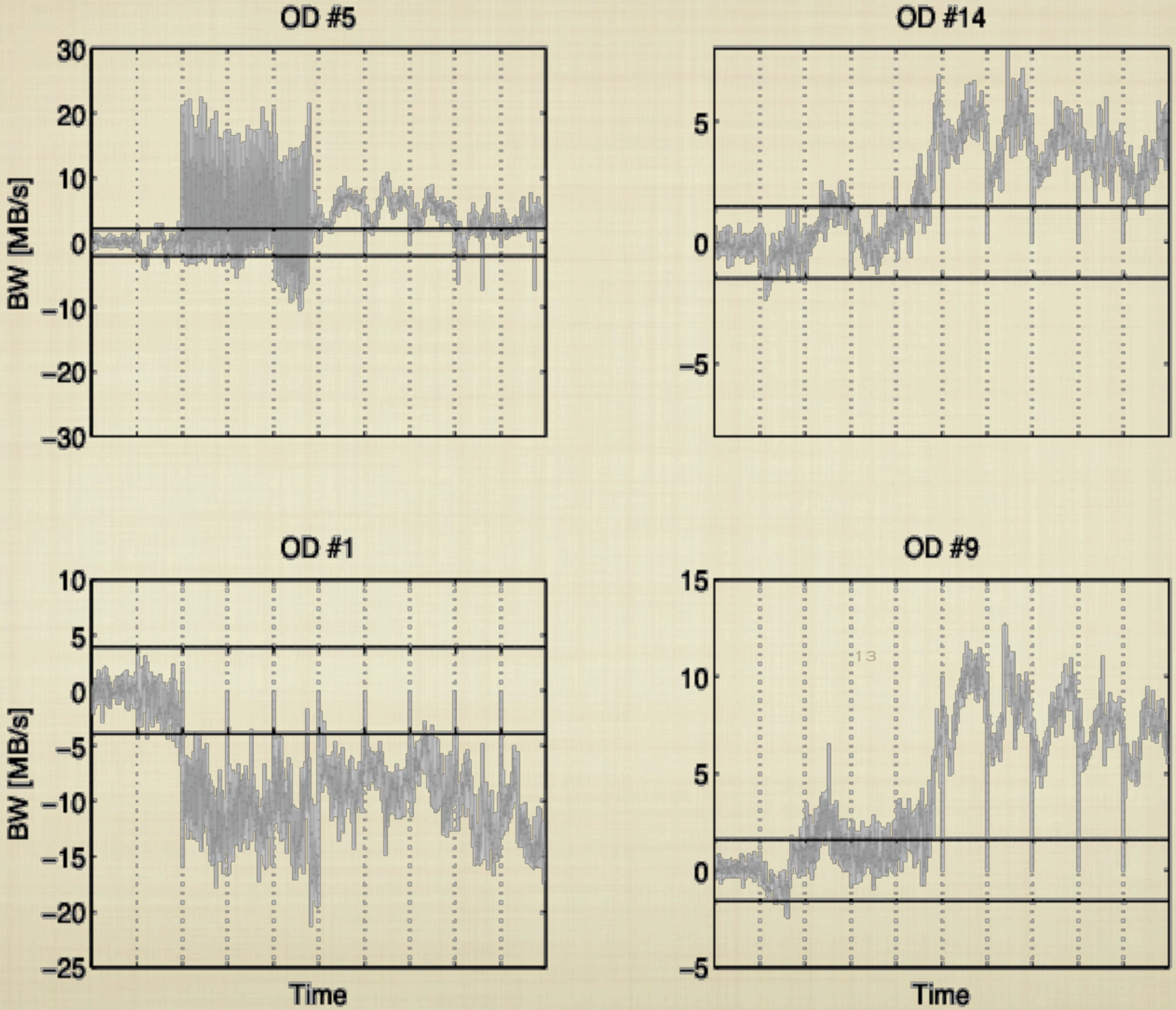
Kalman filter



Kalman filter results

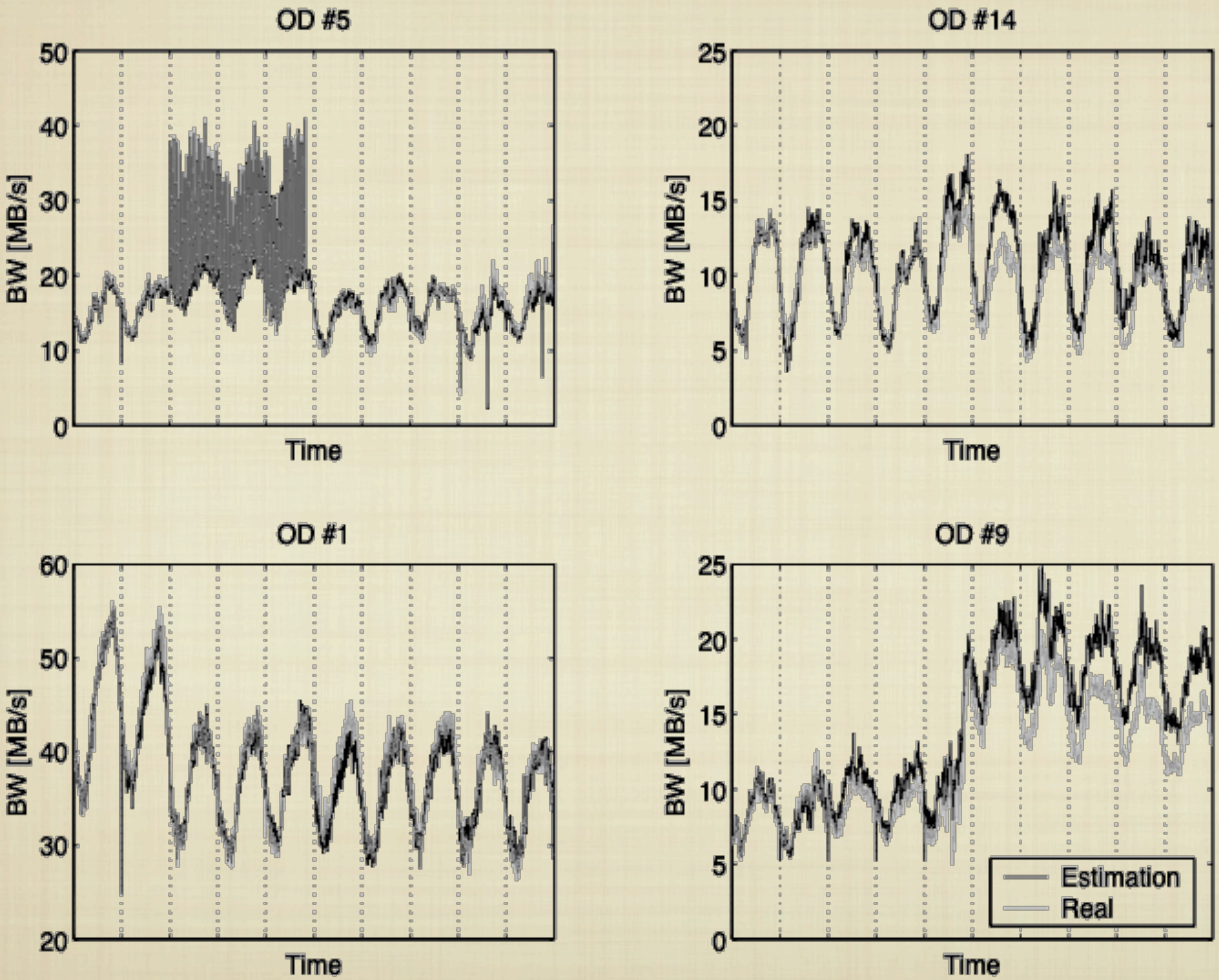


Innovation

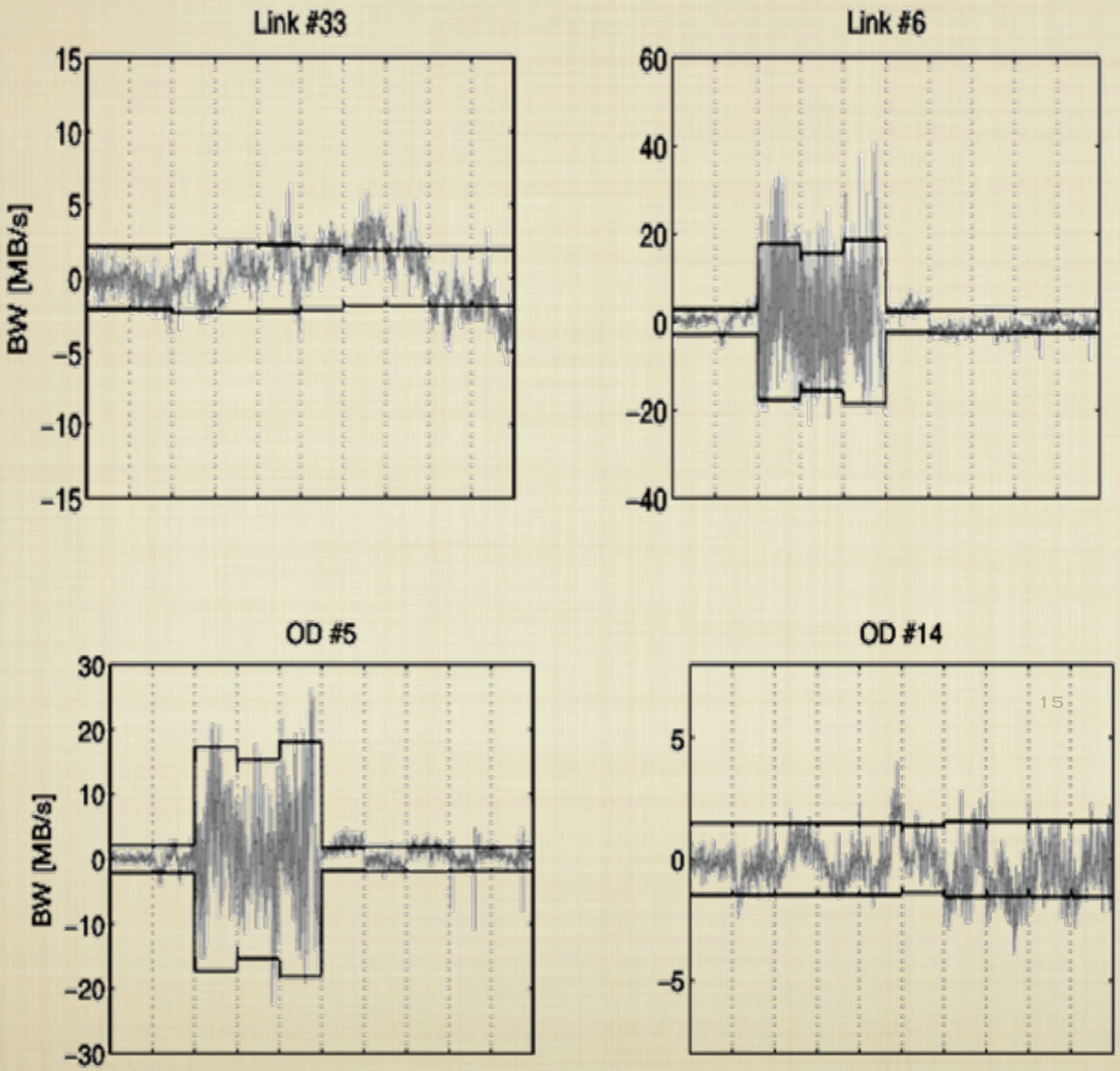


OD #5
OD #14
OD #1
OD #9
BW [MB/s]
BW [MB/s]
Time
Time
30
20
10
0
−10
−20
−30
5
0
−5
10
5
0
−5
−10
−15
−20
−25
15
10
5
0
−5

Recalibration !

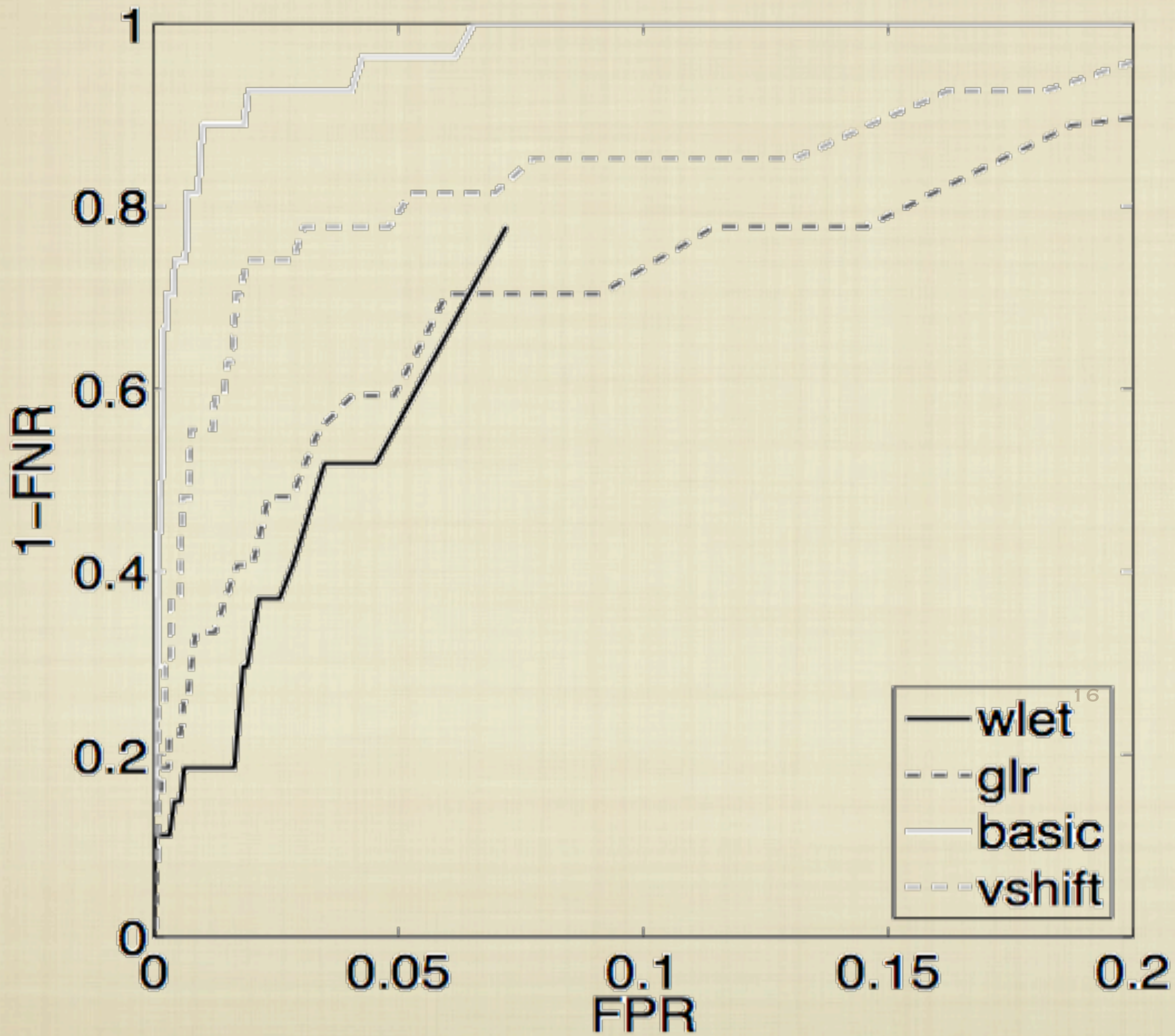


Innovation after recalibration

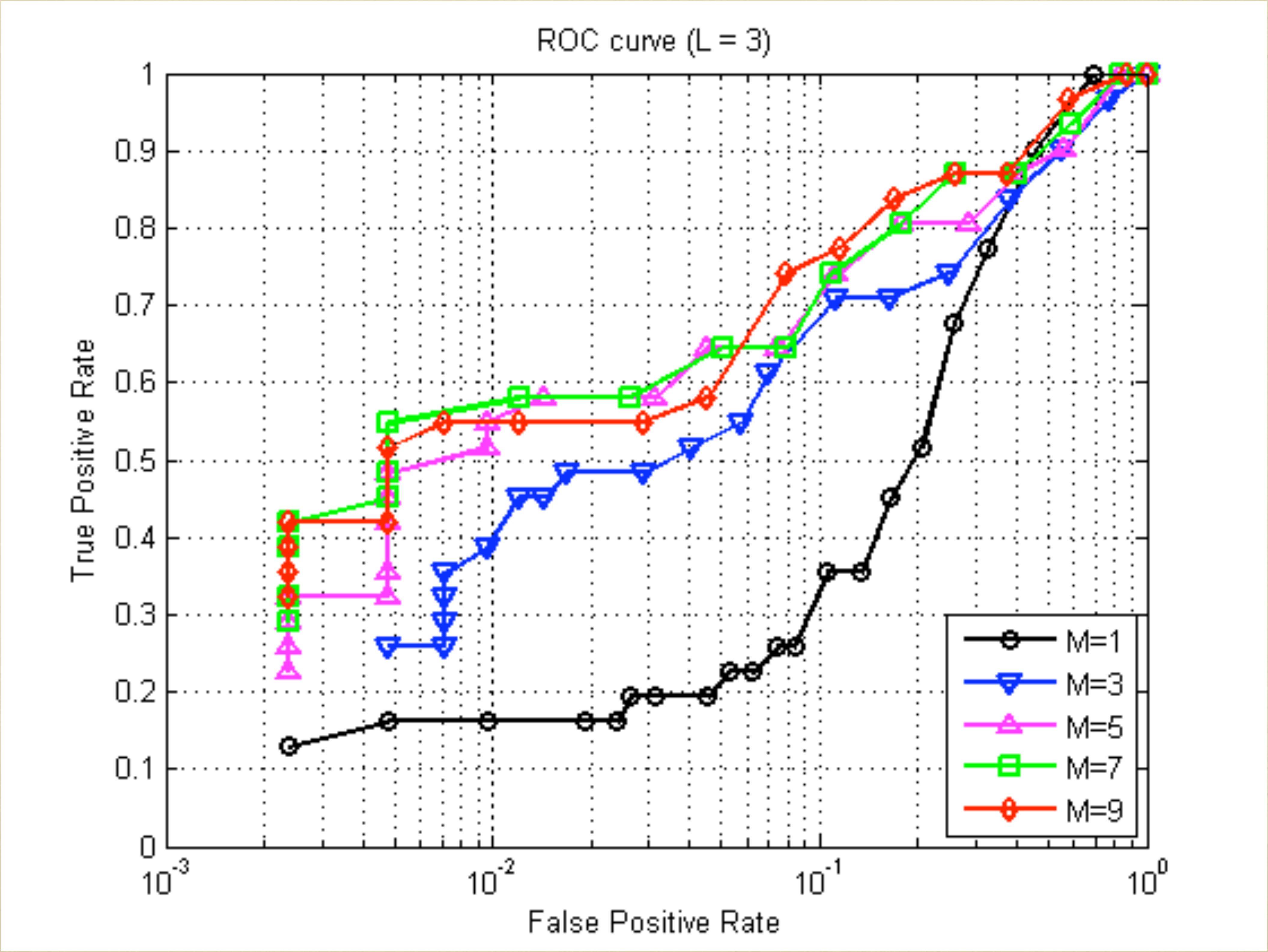


ROC curve

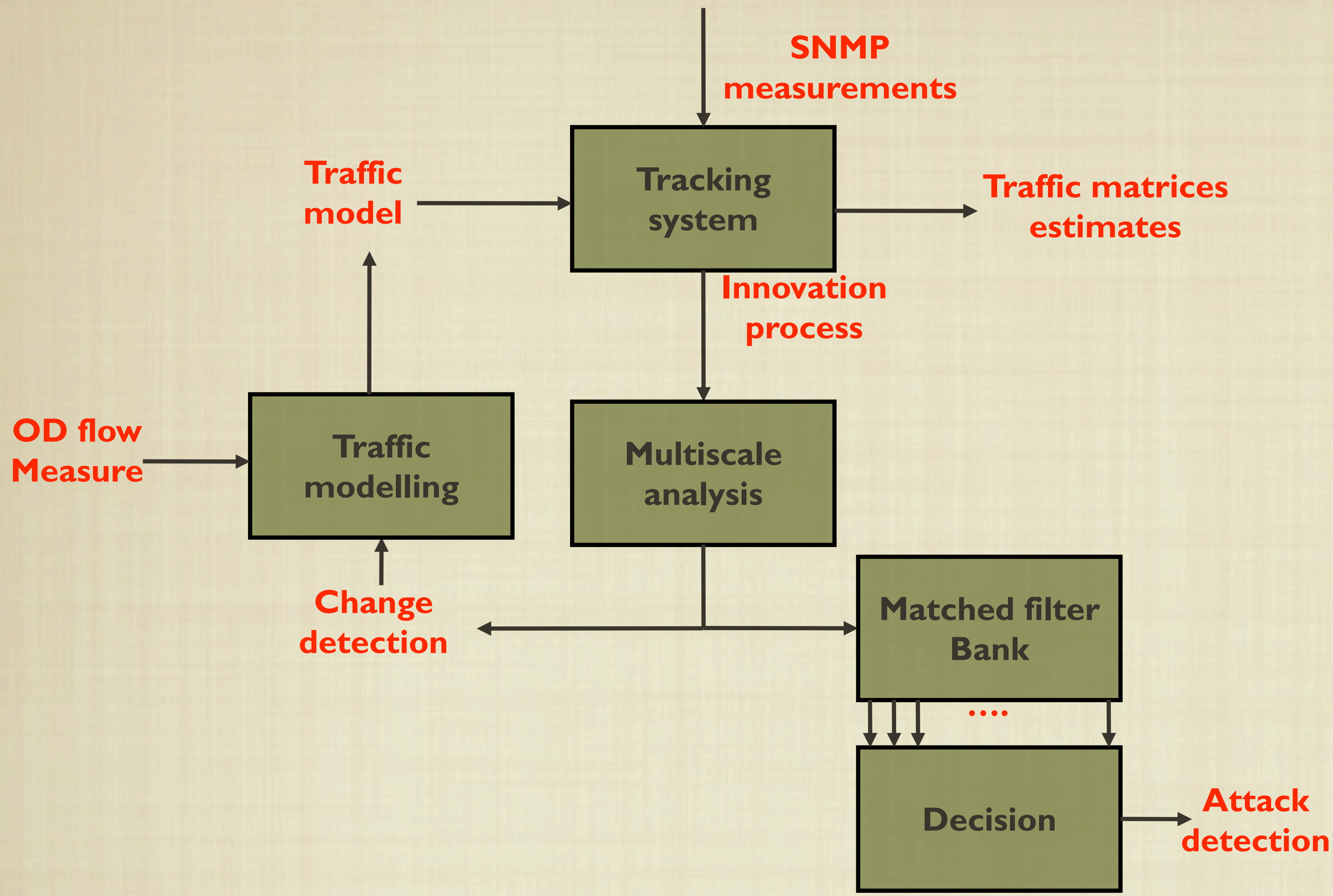
Receive Operating Characteristics (ROC) Curve



ROC Curve PCA



Anomaly



SNMP measurements
Traffic model
Tracking system
Traffic matrices estimates
Innovation process
OD flow Measure
Traffic modelling
Multiscale analysis
Change detection
Matched filter Bank
....
Decision
Attack detection

Non Parametric techniques

Anomaly detection

- The normal behaviour of a network should be assessed as a global joint distribution
  - An anomaly will have effects at different place of the network
- Basic underlying question how to infer the joint distribution in a distributed way ?
  - In an efficient way

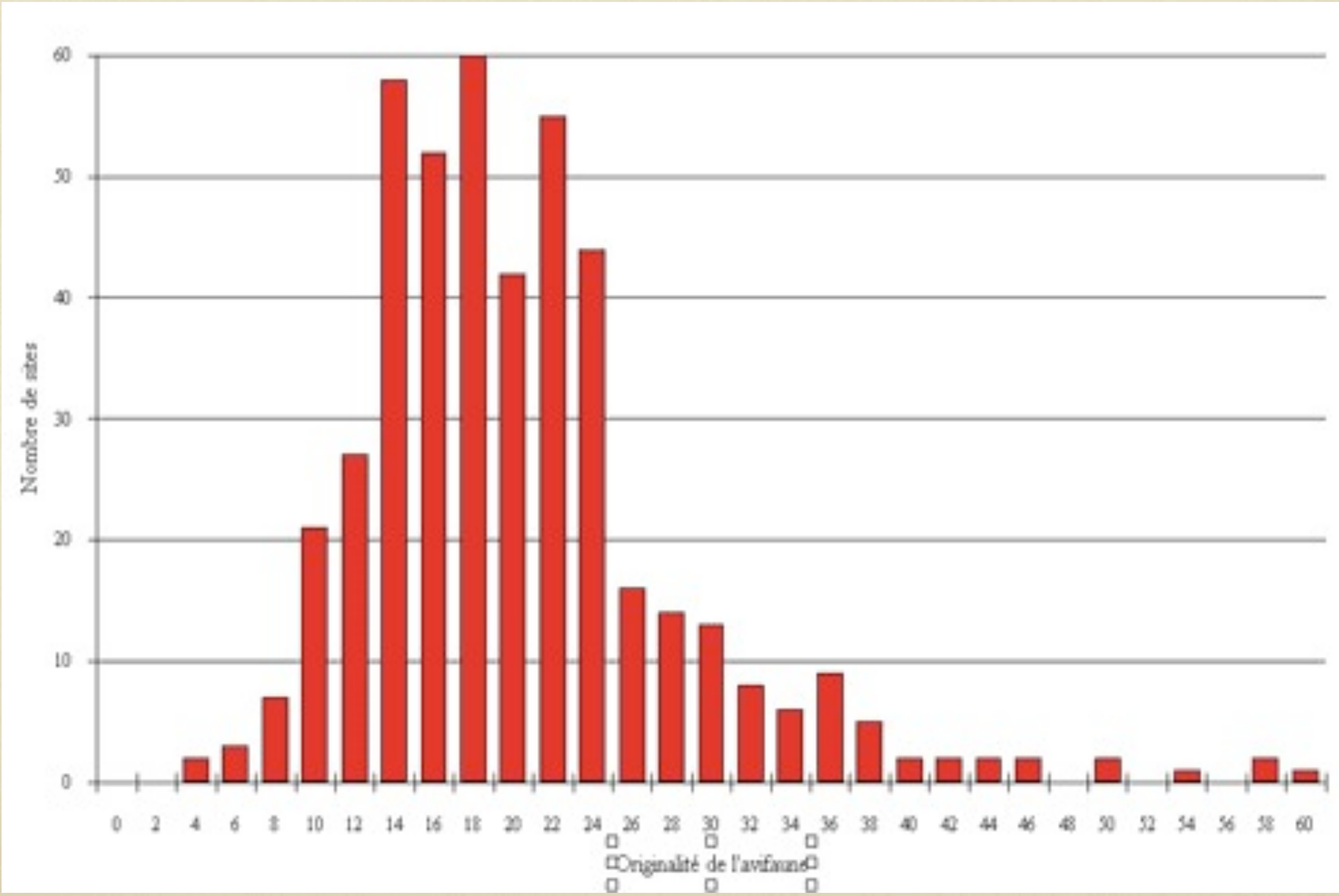
Distributed density

- Naive, costly solution
  - Collect all data sources to a central analysis point. Then perform density estimation
  - Infeasible especially if the procedure has to be repeated periodically

- Distributed Density estimation
  - Each data source makes a local analysis
  - Send the results to a central point
    - Aggregation of local analyses to obtain a global analysis of data

Classical distribution estimation

- Simplest density estimator : multi-dimensional histogram

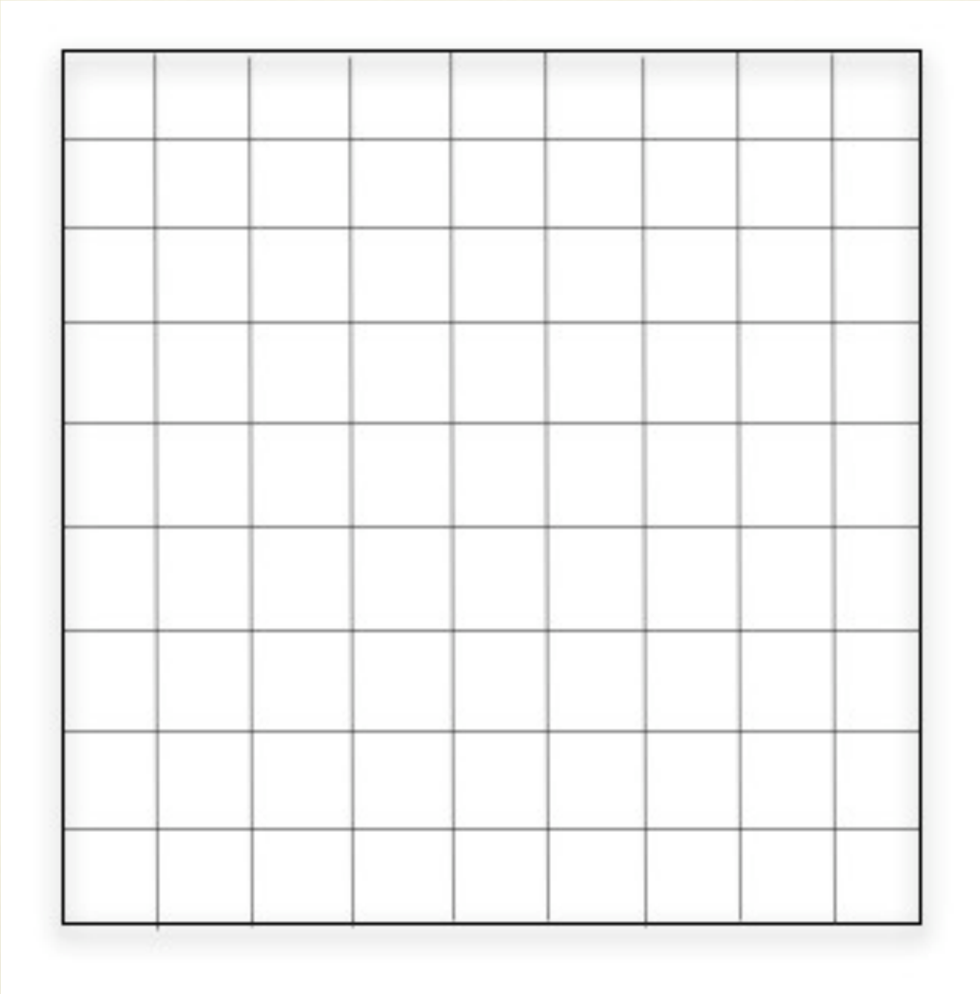


Rectangular grid issues

- Quality of the binning depends of
  - bin size
  - origin
  - grid orientation
    - following principal directions
    - A single grid could either be quite good or quite bad, depending how much it is oriented like the data.

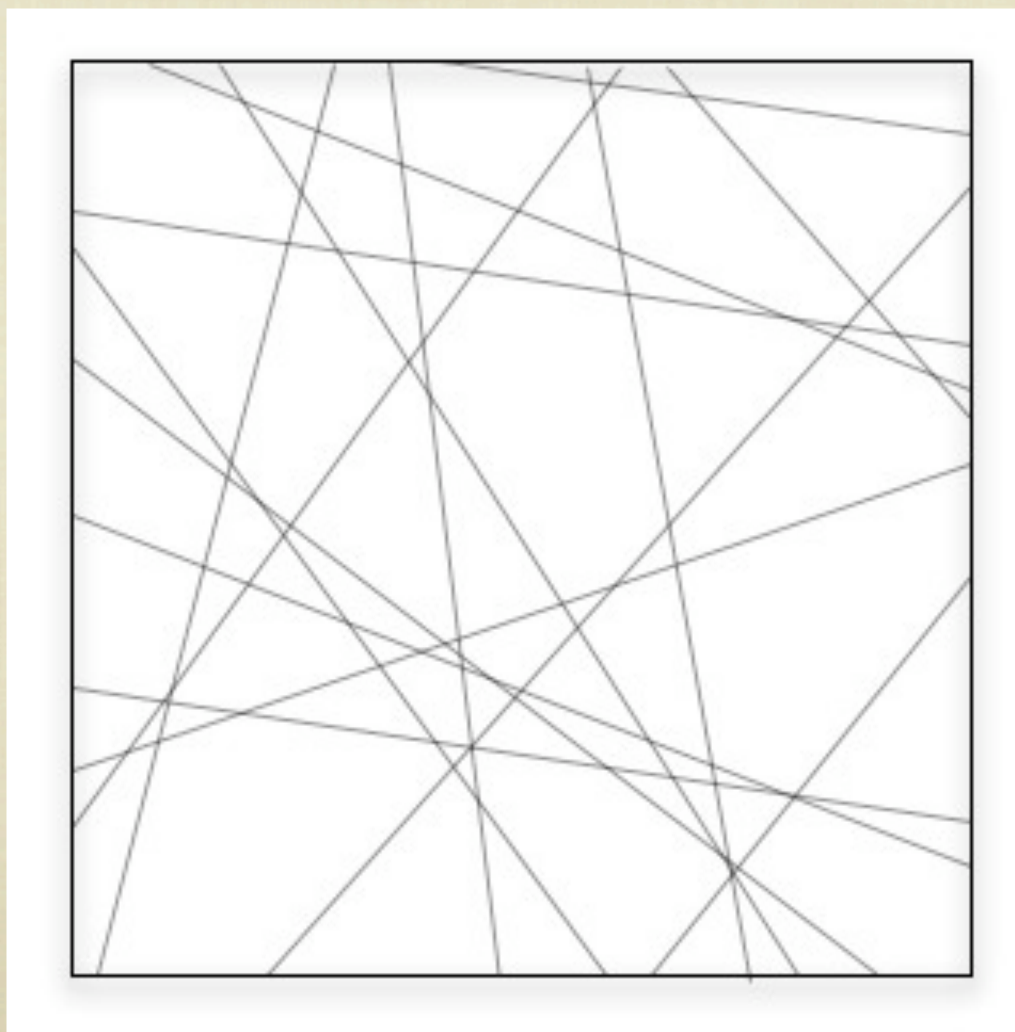
Mosaic-based histogram

- A rectangular binning can be viewed as a set of regularly-spaced hyperplanes on each dimension, each bin being an hyper-rectangle delimited by hyperplanes



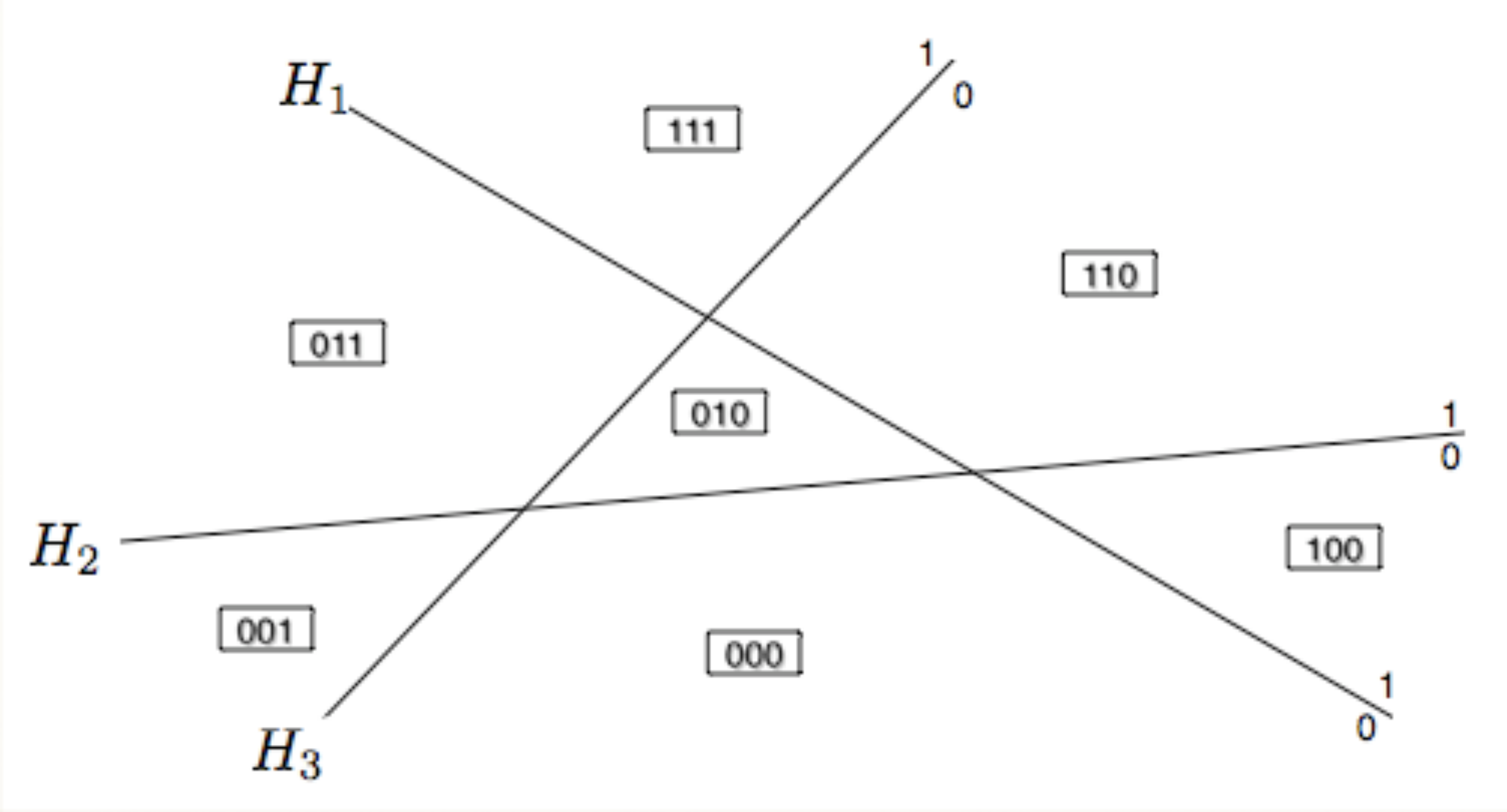
Random Mosaic

- Throw randomly the same number M of hyperplanes to obtain a mosaic covering the data space



Mosaic-based binning

- We can associate a binary number to each of the regions (convex polytopes) in a natural way



Mosaic-based binning

- The binning Y_i of the data sample X_i on each local site is made very simply

$$Y_i = sgn(A \cdot X_i - b)$$

where A and b are random matrices.

- Only need to transmit number of hyperplanes and random seed to each local site !

Mosaic-based binning

- The binary codewords are transmitted to the central node
- Central node must compute the volume of each non-empty polytope of the tessellation
- Finally we can build the histogram

Density estimation

- Let the system of cells be $\{C_j\}$ Then the estimate at a point $t \in \mathbb{R}^d$ is

$$\hat{f}(t) = \frac{\mathbf{1}(X \in C_j)}{N_X \cdot Vol(C_j)},\ t \in C_j$$

Back to anomaly detection

- This performs better than a fixed grid when the local distributions differ

- Anomaly detection
  - Fix a mosaic
  - Periodically update the density estimate $f_t(\cdot)$
  - If the current density estimate $f_0(\cdot)$ is "too far" from the density of the "normal" state

=> Anomaly

Desired qualities of the estimator

- Quality of the global aggregate straightly depend on the quality of local analyses
- Local analyses : quality / size tradeoff
- A good distributed estimator must also allow tunability of this tradeoff

Anomaly detection

- "Too far" may be the quadratic error between the distributions

$$\int (f_t(u) - f_0(u))^2 du$$

  - but this could be rather costly to evaluate

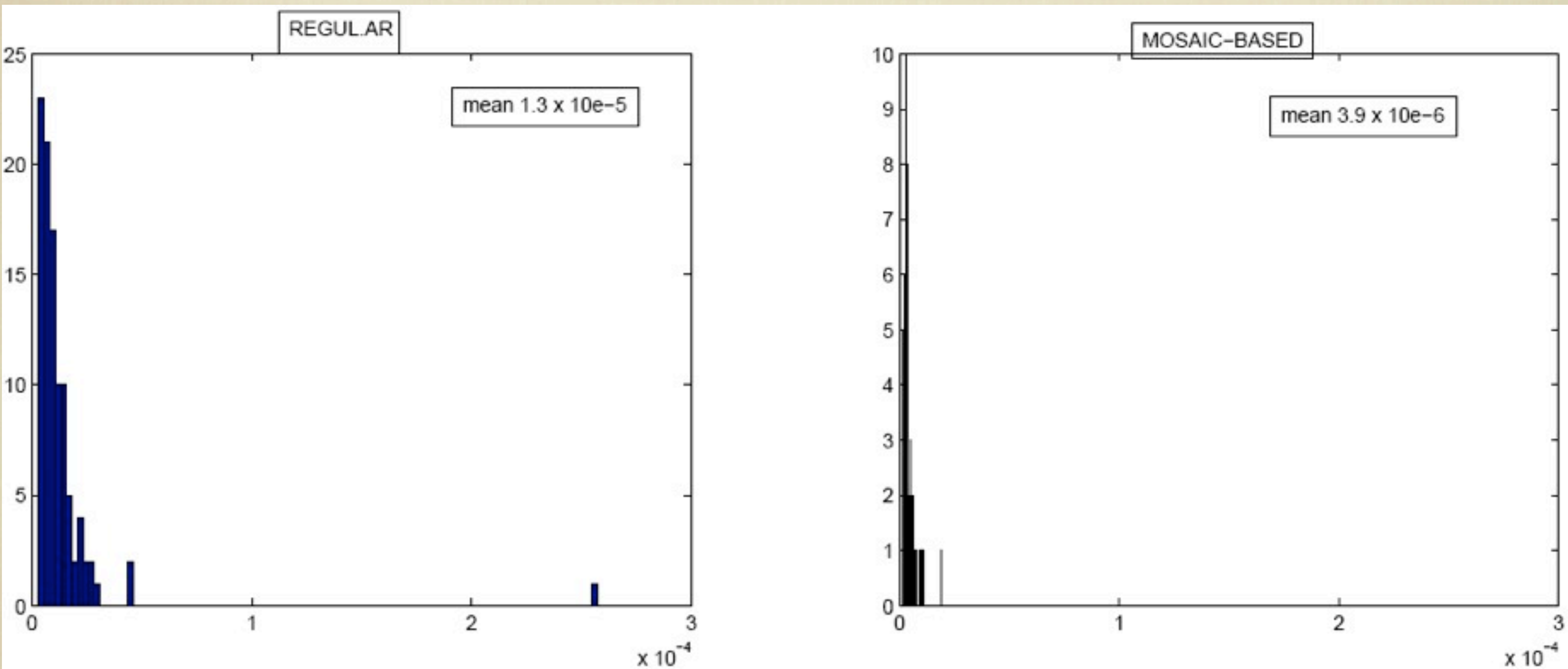
Anomaly detection

- Kullback-Leibler distance is easier to evaluate

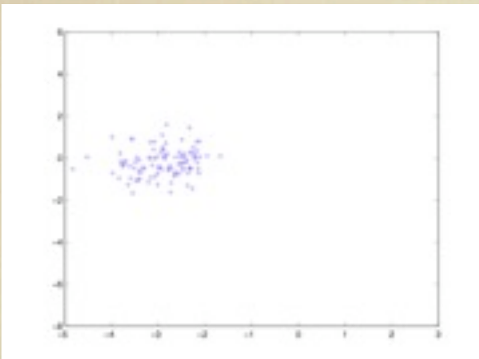
$$D(f_0||f_t) = \sum_{cells\ C_j} f_0(C_j) \cdot log\left(\frac{f_0(C_j)}{f_t(C_j)}\right)$$

since we can take it as a discrete distribution

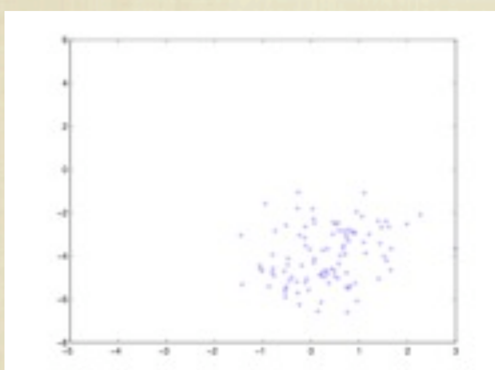
over the bins.


REGUL.AR
mean 1.3 x 10e-5
0
5
10
15
20
25
1
2
3
x 10^{-4}
MOSAIC-BASED
mean 3.9 x 10e-6
0
1
2
3
4
5
6
7
8
9
10
1
2
3
x 10^{-4}

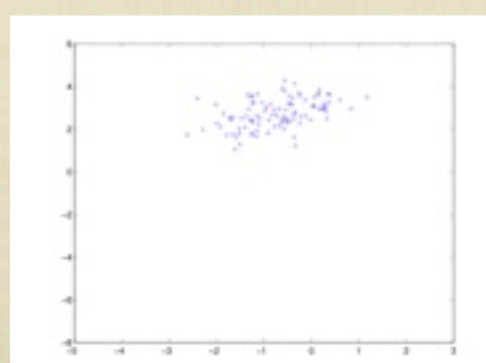
Global distribution



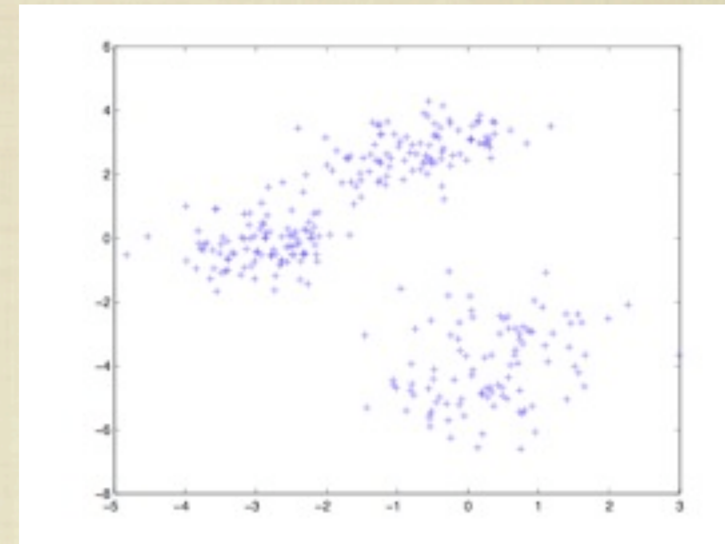
$f_1(x), \frac{N_1}{N}\ points$



$f_2(x), \frac{N_2}{N}\ points$

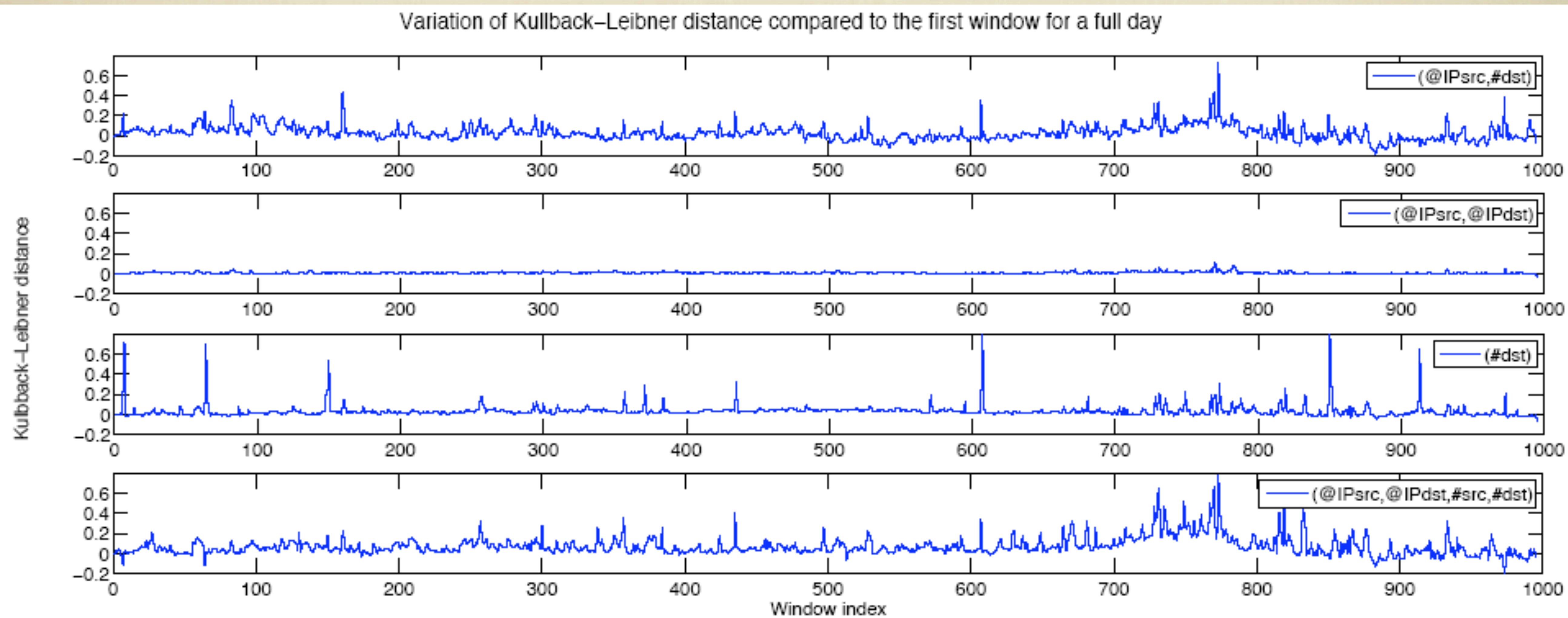


$f_3(x), \frac{N_3}{N}\ points$



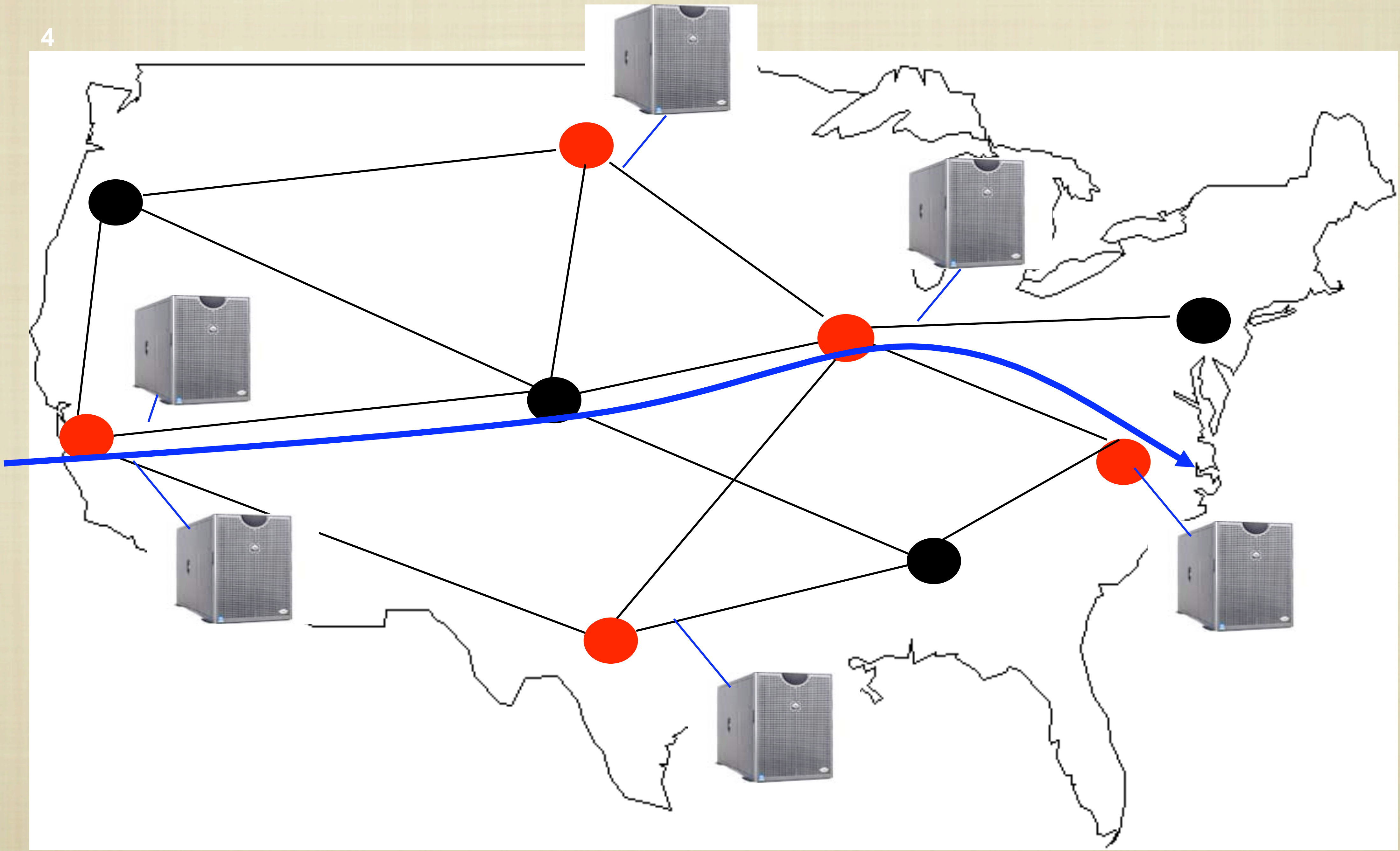
$$f(x) = \sum_i \frac{N_i}{N} f_i(x)$$

Non Parametric AD



Distributed Anomaly detection

Distributed anomaly detection



Distributed anomaly detection !

- Same as centralized
  - All the steps should be run in a distributed way
    - Distributed model identification
    - Distributed filtering
    - Distributed decision
  - For parametric and non parametric approaches !
    - Parametric : Steps 1 and 2 distributed, step 3 local
    - Non parametric : Steps 1 and 2 local, step 3 distributed

Parametric vs. Non Parametric

- Parametric case
  - Knowing the autocovariance what are the best projections
    - Local KLT gives best local projections
      - Not optimal
    - Distributed KLT
      - Solve global optimization by an iterative algorithm
  - Non parametric case
    - Use random projection
      - If enough used converges to KLT
        - Asymptotically optimal
        - Compressed sensing ideA

Distributed KLT

- 4 nodes with 10 observations each
- Exchanging 2 projections each
- Joint KLT MSE= 8.5
- Distributed KLT MSE= 15.2
- 90% of AD coincidence