Change Detection in Multivariate Data

Likelihood and Detectability Loss

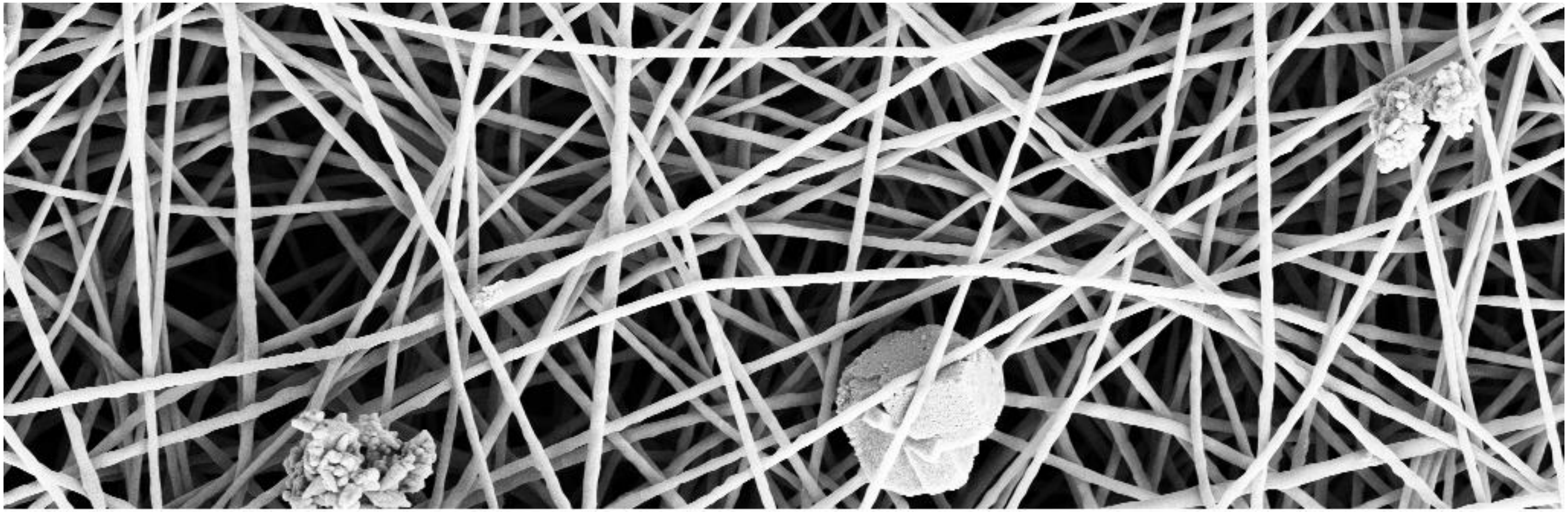
Giacomo Boracchi

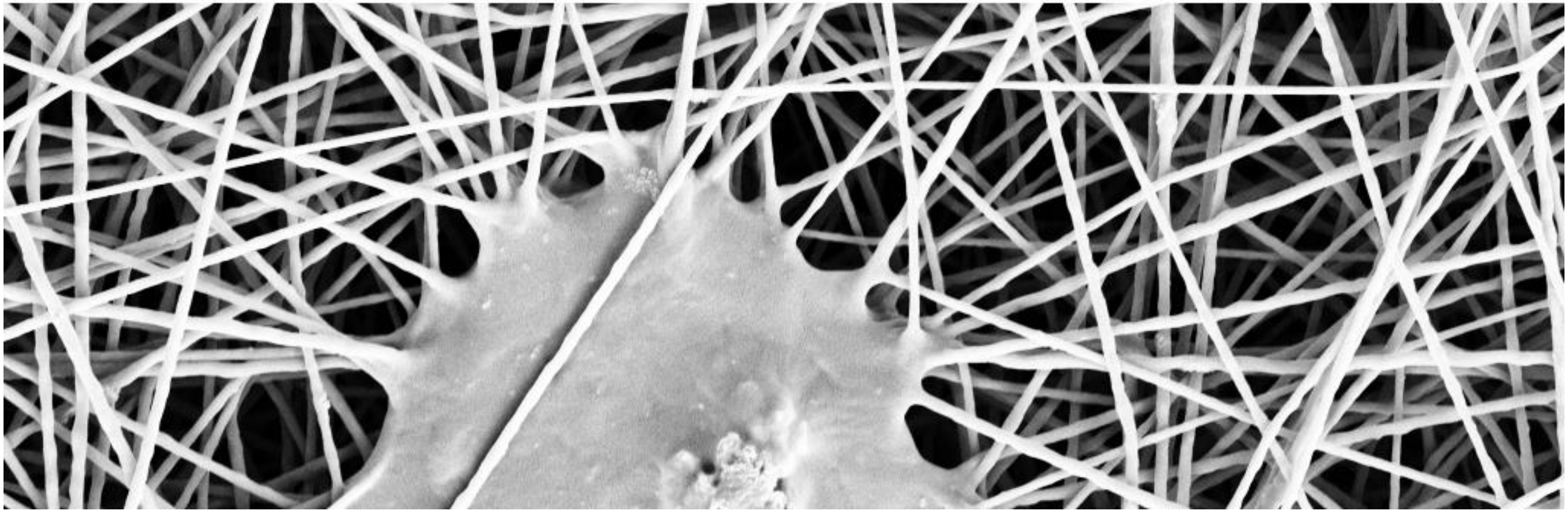
July, 8th , 2016

giacomo.boracchi@polimi.it

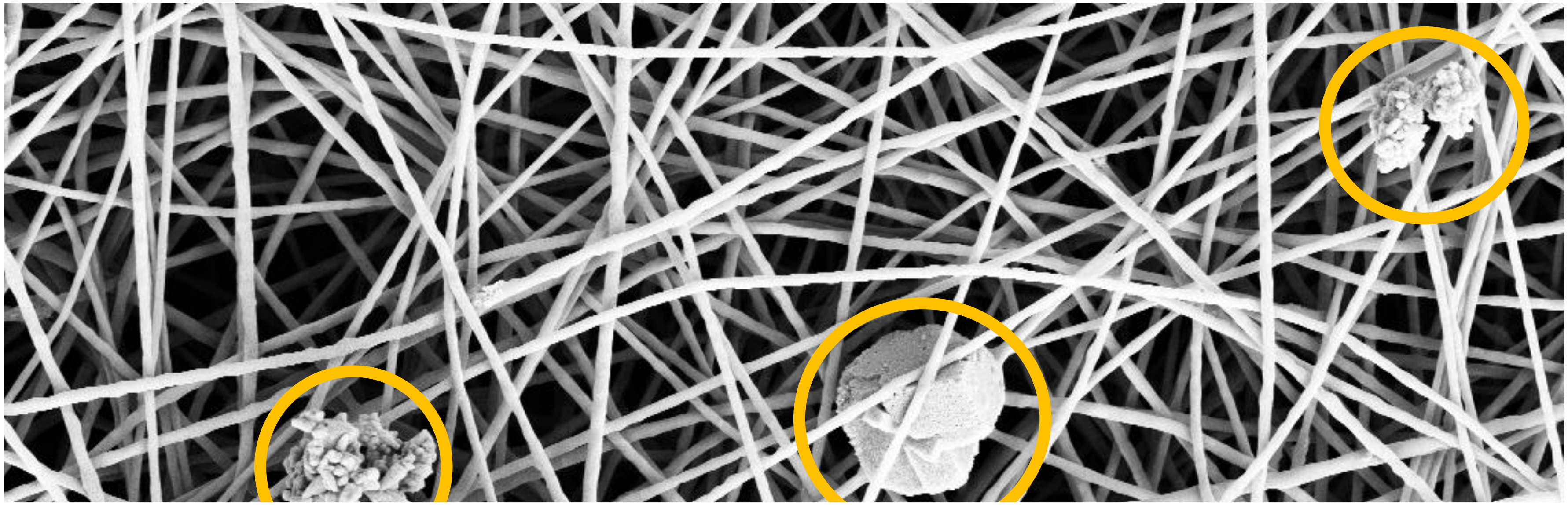
TJ Watson, IBM NY

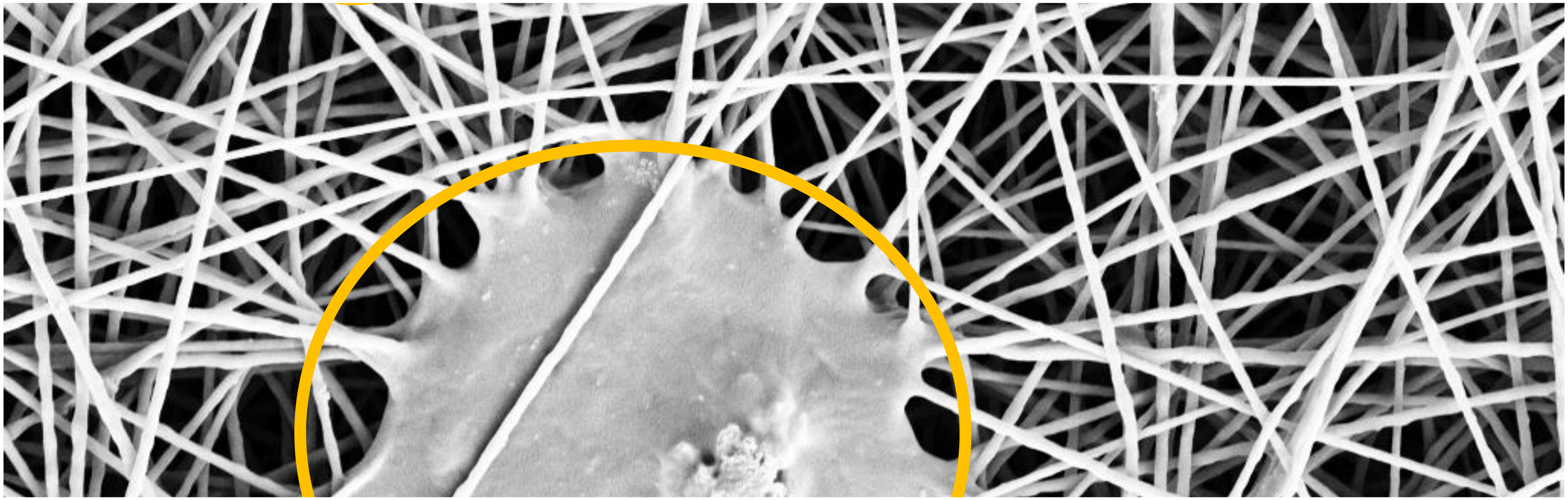
Examples of CD Problems: Anomaly Detection





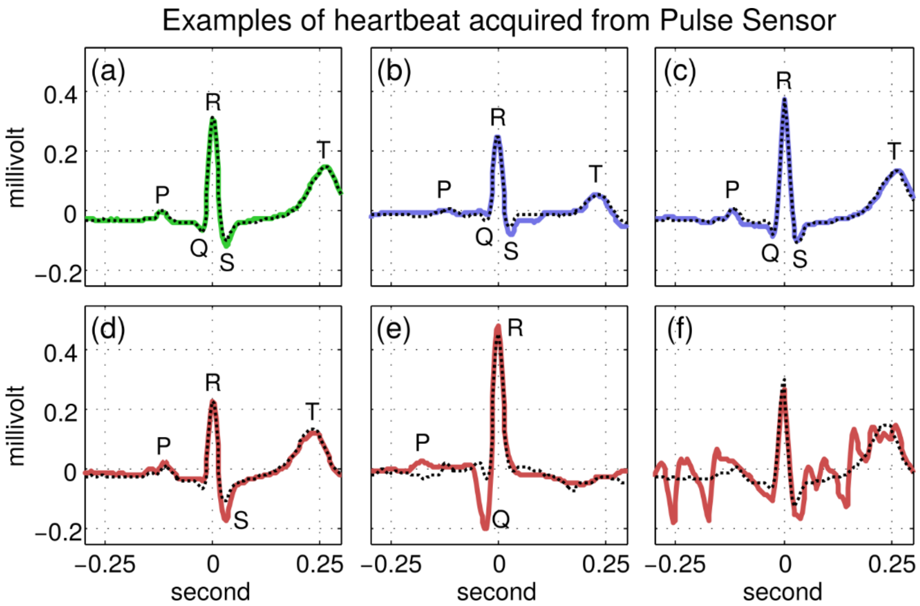
Examples of CD Problems: Anomaly Detection





Other Examples of CD Problems

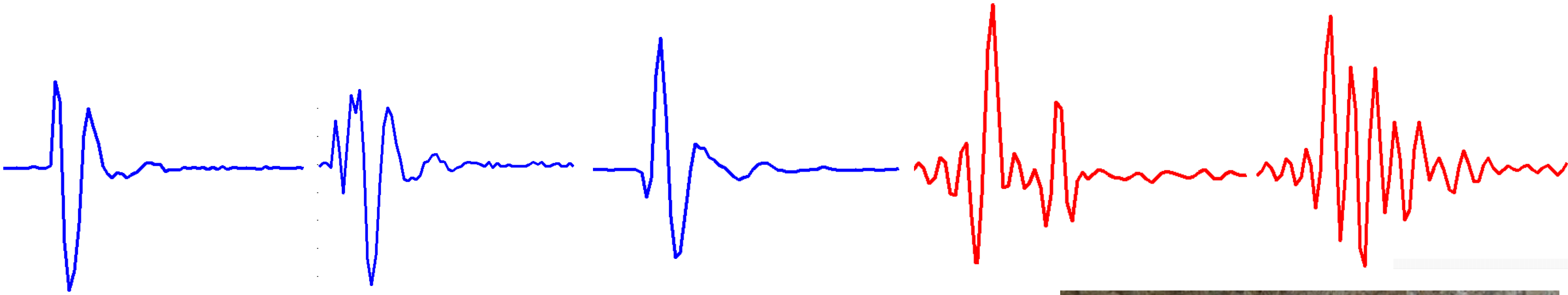
ECG monitoring: Detect arrhythmias / device mispositioning



Other Examples of CD Problems

ECG monitoring: Detect arrhythmias / device mispositioning

Environmental monitoring: detect changes in signals monitoring a rockface



Other Examples of CD Problems

ECG monitoring: Detect arrhythmias / device mispositioning

Environmental monitoring: detect changes in signals monitoring a rockface

Stream mining: Fraud Detection

Stream mining: Online Classification Systems

Spam Classification

Fraud Detection

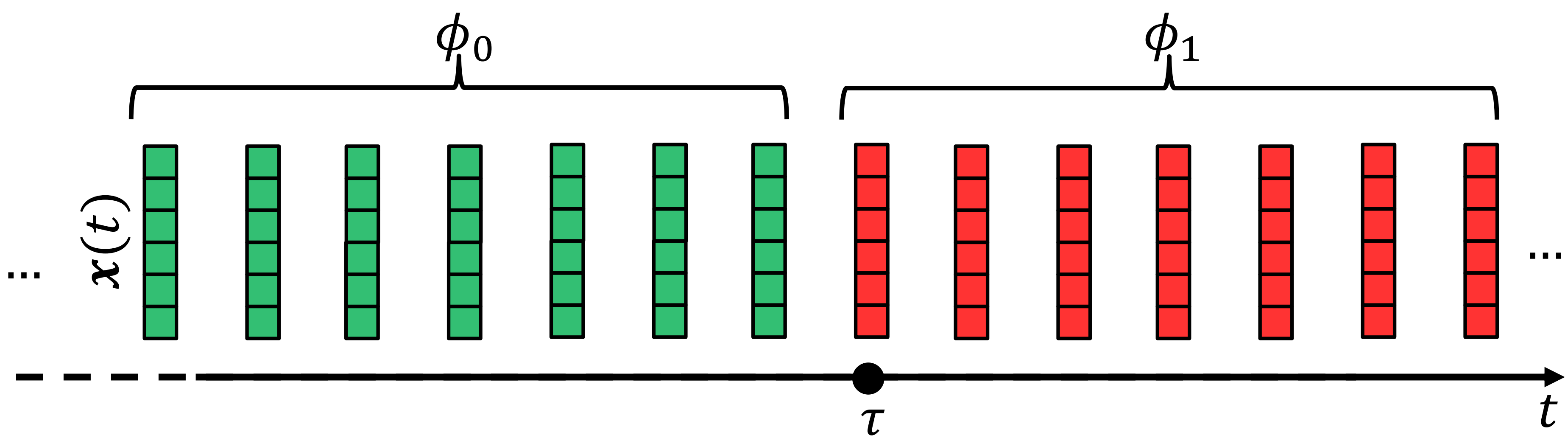
The Change-Detection Problem

Often, these problems boil down to:

a) Monitor a **stream** $\{\boldsymbol{x}(t), t = 1, \dots\}$, $\boldsymbol{x}(t) \in \mathbb{R}^d$ of realizations of a **random variable**, and **detect the change-point** τ,

$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & t < \tau \\ \phi_1 & t \geq \tau \end{cases},$$

where $\{\boldsymbol{x}(t),\ t < \tau\}$ are i.i.d. and $\phi_0 \neq \phi_1$, ϕ_1 is unknown and ϕ_0 can be possibly estimated from training data



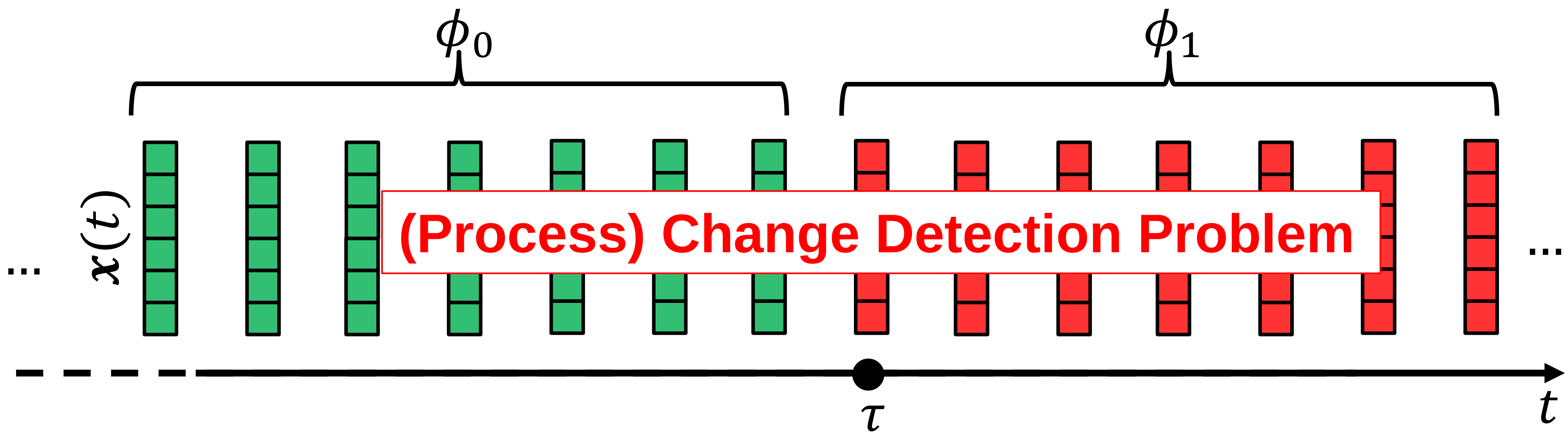
The Change-Detection Problem

Often, these problems boil down to:

a) Monitor a **stream** $\{\boldsymbol{x}(t), t = 1, \dots\}$, $\boldsymbol{x}(t) \in \mathbb{R}^d$ of realizations of a **random variable**, and **detect the change-point** τ,

$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & t < \tau \\ \phi_1 & t \geq \tau \end{cases},$$

where $\{\boldsymbol{x}(t),\ t < \tau\}$ are i.i.d. and $\phi_0 \neq \phi_1$, ϕ_1 is unknown and ϕ_0 can be possibly estimated from training data



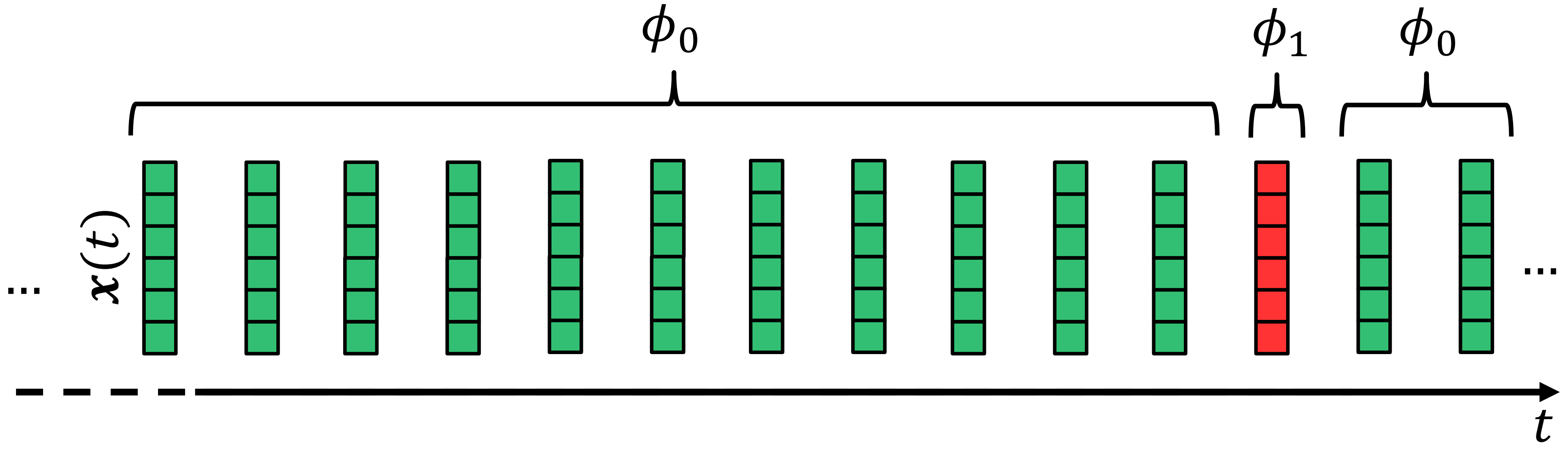
The Change-Detection Problem

Often, these problems boil down to:

b) Determining whether a set of data $\{\boldsymbol{x}(t),\ t = t_0, \dots, t_1\}$ **is generated from $\boldsymbol{\phi_o}$** and detect possible **outliers**

We refer to

- ϕ_0 pre-change distribution / normal (can be estimated)
- ϕ_1 post-change distribution / anomalous (unknown)



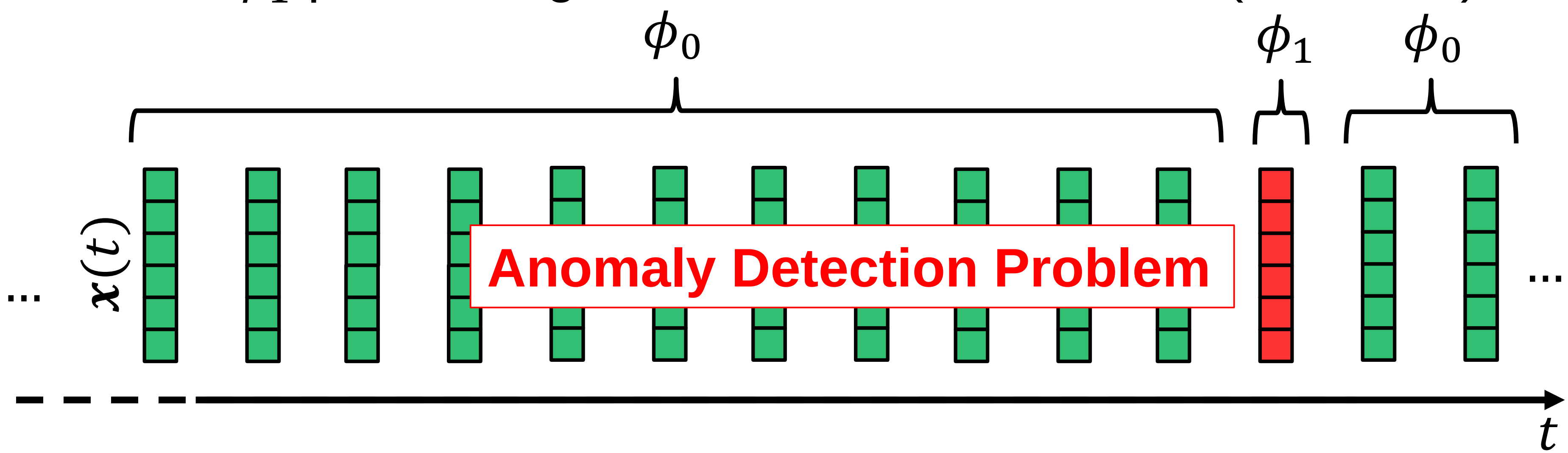
The Change-Detection Problem

Often, these problems boil down to:

b) Determining whether a set of data $\{\boldsymbol{x}(t),\ t = t_0, \dots, t_1\}$ **is generated from $\boldsymbol{\phi_o}$** and detect possible **outliers**

We refer to

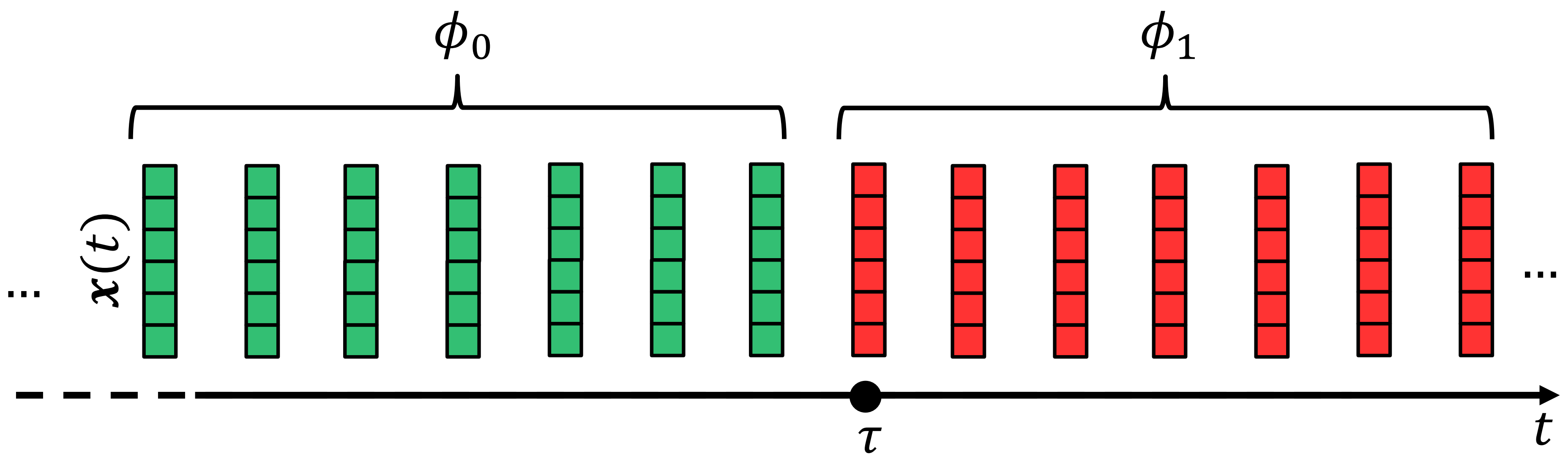
- ϕ_0 pre-change distribution / normal (can be estimated)
- ϕ_1 post-change distribution / anomalous (unknown)



THE ADDRESSED PROBLEM

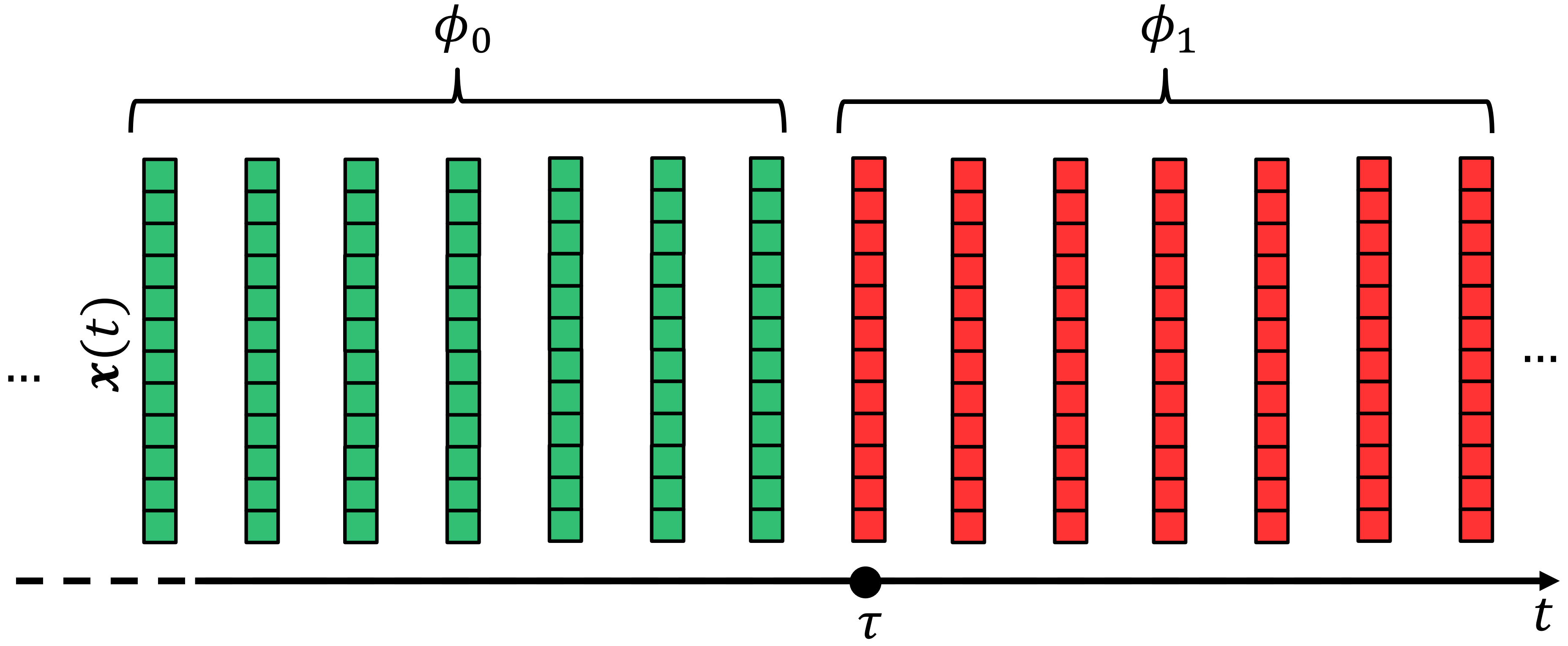
Our Goal

*Study how the **data dimension** d **influences** the **change detectability**, i.e., how difficult is to solve these two problems*



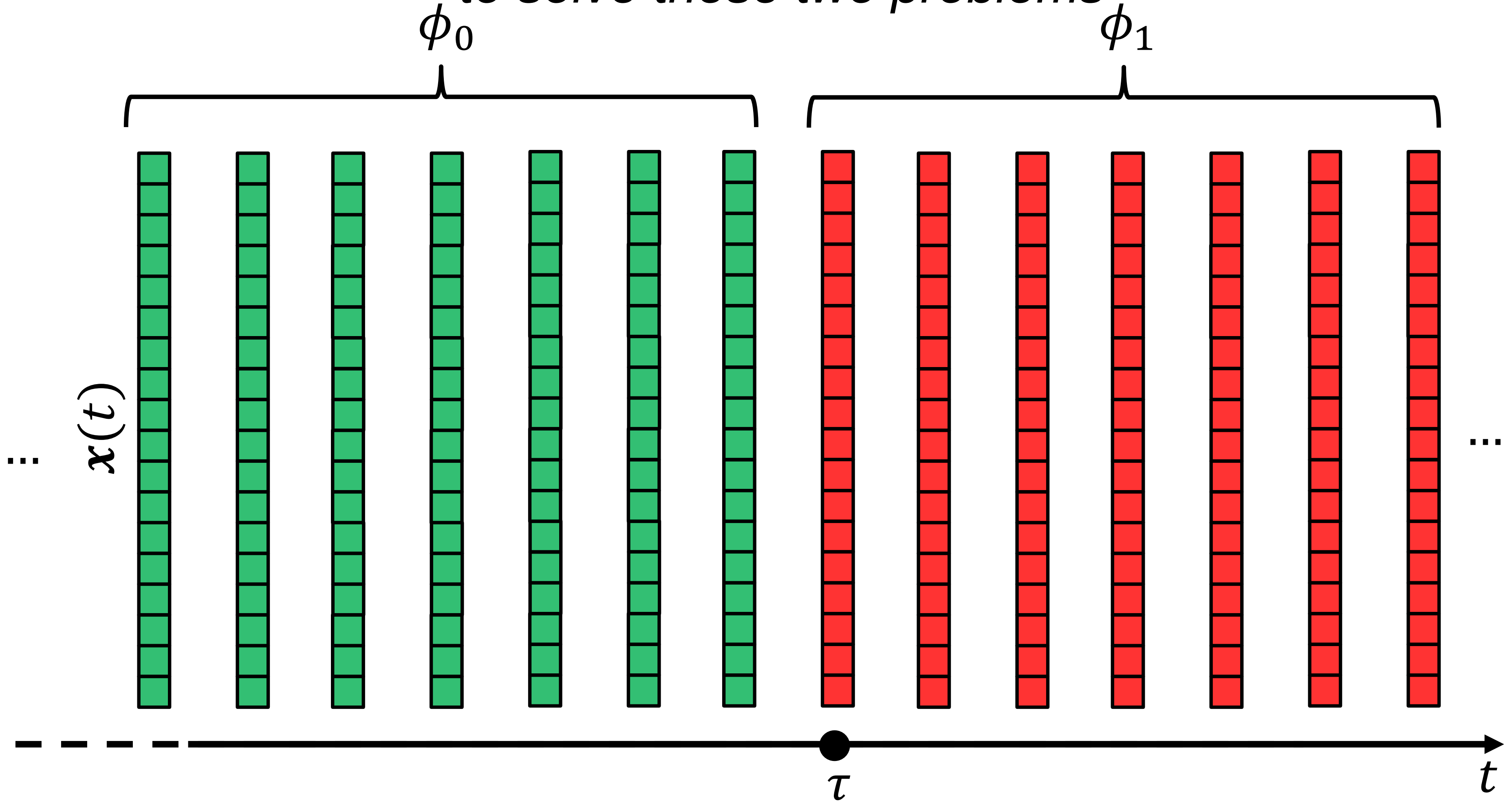
Our Goal

Study how the ***data dimension*** d ***influences*** *the* ***change detectability****, i.e., how difficult is to solve these two problems*



Our Goal

Study how the ***data dimension*** d ***influences*** *the* ***change detectability****, i.e., how difficult is to solve these two problems*



Our Approach

To study the impact of the **sole data dimension** d in **change-detection problems** we need to:

1. Consider a **change-detection approach**
2. Define a measure of **change detectability** that well correlates with traditional performance measures
3. Define a measure of **change magnitude** that refers only to differences between ϕ_0 and ϕ_1

Our Approach

To study the impact of the **sole data dimension** d in **change-detection problems** we need to:

1. Consider a **change-detection approach**
2. Define a measure of **change detectability** that well correlates with traditional performance measures
3. Define a measure of **change magnitude** that refers only to differences between ϕ_0 and ϕ_1

Our goal (reformulated):
Studing how the **change detectability varies** in **change-detection problems** that have

- **different data dimensions** d
- **constant change magnitude**

Our Result

We show there is a **detectability loss** problem, i.e. that change **detectability** steadily **decreases** when d increases.

Detectability loss is shown by:

- Analytical derivations: when ϕ_0 and ϕ_1 are **Gaussians**
- Empirical analysis: measuring the the **power of hypothesis tests** in change-detection problems on real data

Presentation Outline

- Preliminaries:
  - Assumptions
  - The change-detection approach
  - The change magnitude
  - The measure of change detectability
- The *detectability loss*
- Detectability loss and anomaly detection in images

Presentation Outline

- Preliminaries:
  - Assumptions
  - The change-detection approach
  - The change magnitude
  - The measure of change detectability
- The *detectability loss*
- Detectability loss and anomaly detection in images

Our Assumptions

To detect the change $\phi_0 \to \phi_1$ **we assume** that

- ϕ_0 is **unknown**, can be **estimated** from a training set $$TR = \{x(t), t < t_0, x \sim \phi_0\}$$
- ϕ_1 is **unknown, no training data** are provided

We refer to

- ϕ_0 as stationary / normal / pre-change distribution
- $\hat{\phi}_0$ as the estimate of ϕ_0 from a training set
- ϕ_1 as nonstationary / anomalous / post-change distribution

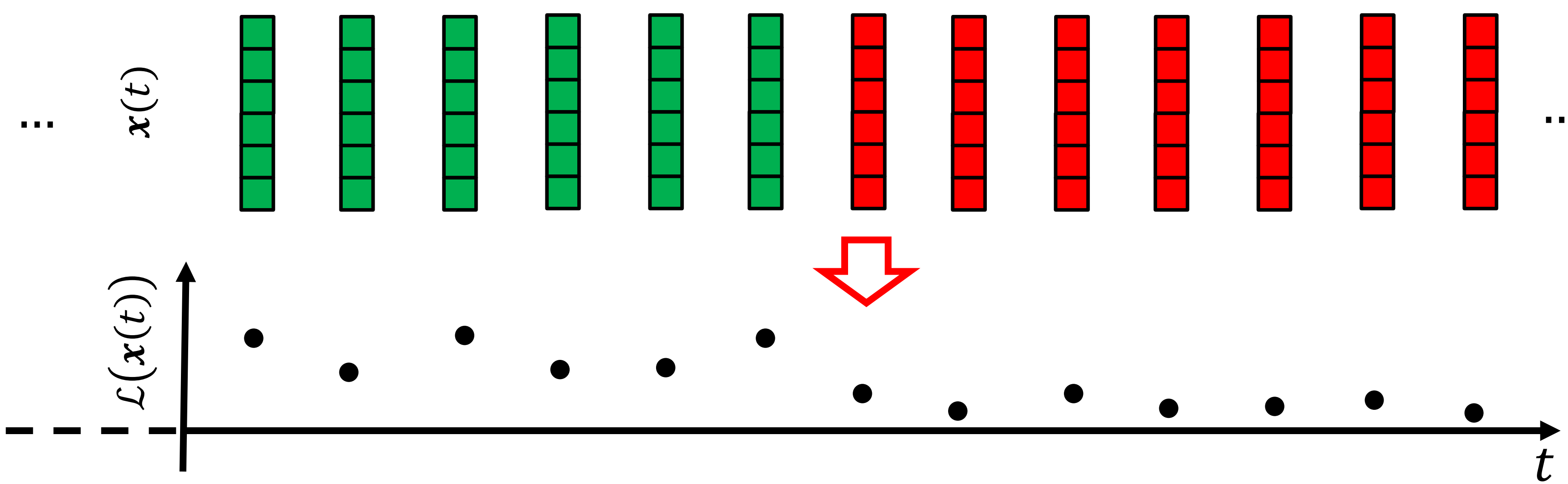
Presentation Outline

- Preliminaries:
  - Assumptions
  - The change-detection approach
  - The change magnitude
  - The measure of change detectability
- The *detectability loss*
- Detectability loss and anomaly detection in images

How? Monitoring the Log-likelihood

A typical approach to monitor the log-likelohood

1. During training, estimate $\hat{\phi}_0$ from TR
2. During testing, compute
$$\mathcal{L}\big(\boldsymbol{x}(t)\big) = \log(\hat{\phi}_0(\boldsymbol{x}(t)))$$
3. Monitor $\{\mathcal{L}\big(\boldsymbol{x}(t)\big),\ t = 1, \dots\}$



How? Monitoring the Log-likelihood

A typical approach to monitor the log-likelohood

1. During training, estimate $\hat{\phi}_0$ from TR
2. During testing, compute
$$\mathcal{L}\big(\boldsymbol{x}(t)\big) = \log(\hat{\phi}_0(\boldsymbol{x}(t)))$$
3. Monitor $\{\mathcal{L}\big(\boldsymbol{x}(t)\big),\ t = 1, \dots\}$

This is quite a popular approach in sequential monitoring and in anomaly detection

L. I. Kuncheva, "*Change detection in streaming multivariate data using likelihood detectors*," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, 2013.

X. Song, M. Wu, C. Jermaine, and S. Ranka, "*Statistical change detection for multidimensional data*," in Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD), 2007.

J. H. Sullivan and W. H. Woodall, "*Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations*," IIE transactions, vol. 32, no. 6, 2000.

Our Goal / Presentation Outline

- Preliminaries:
  - Assumptions
  - $\boxed{\text{The change-detection approach}}$
  - The change magnitude
  - The measure of change detectability
- The *detectability loss*
- Detectability loss and anomaly detection in images

The Change Magnitude

We measure the **magnitude of a change** $\phi_0 \to \phi_1$ by the *symmetric Kullback-Leibler divergence*

$$\mathrm{sKL}(\phi_0, \phi_1) = \mathrm{KL}(\phi_0, \phi_1) + \mathrm{KL}(\phi_1, \phi_0) =$$
$$= \int \log\left(\frac{\phi_0(\boldsymbol{x})}{\phi_1(\boldsymbol{x})}\right)\phi_0(\boldsymbol{x})d\boldsymbol{x} + \int \log\left(\frac{\phi_1(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right)\phi_1(\boldsymbol{x})d\boldsymbol{x}$$

In practice, **large values** of $s\mathrm{KL}(\phi_0, \phi_1)$ correspond to **changes** $\phi_0 \to \phi_1$that are very apparent, since $s\mathrm{KL}(\phi_0, \phi_1)$ is related to the power of hypothesis tests designed to detect either $\phi_0 \to \phi_1$ or $\phi_1 \to \phi_0$ (Stein Lemma)

T. Dasu, K. Shankar, S. Venkatasubramanian, K. Yi, "*An information-theoretic approach to detecting changes in multi-dimensional data streams*" In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications, 2006

Our Goal / Presentation Outline

- Preliminaries:
  - The change-detection approach
  - The change magnitude
  - $\boxed{\text{The measure of change detectability}}$
- The *detectability loss*
- Concluding remarks

The Change Detectability

The *Signal to Noise Ratio of the change*

$$\mathrm{SNR}(\phi_0 \to \phi_1) = \frac{\left(\underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})] - \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})]\right)^2}{\underset{\boldsymbol{x}\sim\phi_0}{\mathrm{var}}[\mathcal{L}(\boldsymbol{x})] + \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{var}}[\mathcal{L}(\boldsymbol{x})]}$$

The $\mathrm{SNR}(\phi_0 \to \phi_1)$

- Measures the extent to which $\phi_0 \to \phi_1$ is detectable by monitoring $\mathrm{E}[\mathcal{L}(\boldsymbol{x})]$
- If we replace $\mathrm{E}[\cdot]$ and $\mathrm{var}[\cdot]$ by the sample estimators we get the *t*-test statistic

DETECTABILITY LOSS

The Detectability Loss

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d \times d}$ *and orthogonal ,* $\boldsymbol{v} \in \mathbb{R}^d$*, then*

$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

The Detectability Loss: Remarks

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d \times d}$ *and orthogonal ,* $\boldsymbol{v} \in \mathbb{R}^d$*, then*

$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $s\mathrm{KL}(\phi_0, \phi_1)$

Remarks:

- Changes of a given magnitude, $s\mathrm{KL}(\phi_0, \phi_1)$, become more difficult to detect when d increases
- DL does not depend on how ϕ_0 changes
- DL does not depend on the specific detection rule
- DL does not depend on estimation errors on $\hat{\phi}_0$

The Detectability Loss: The Change Model

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\boxed{\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})}$ *where* $Q \in \mathbb{R}^{d \times d}$ *and orthogonal ,* $\boldsymbol{v} \in \mathbb{R}^d$*, then*

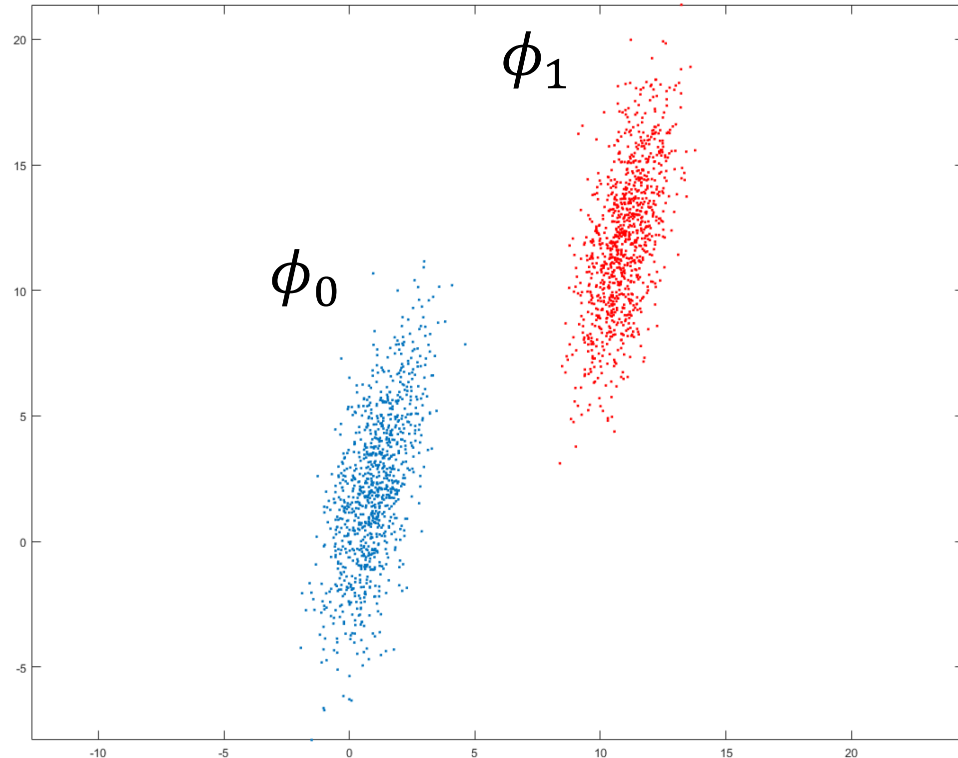
$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

The Detectability Loss: The Change Model

The change model $\phi_1\,(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x}\, + \boldsymbol{v})$ includes:

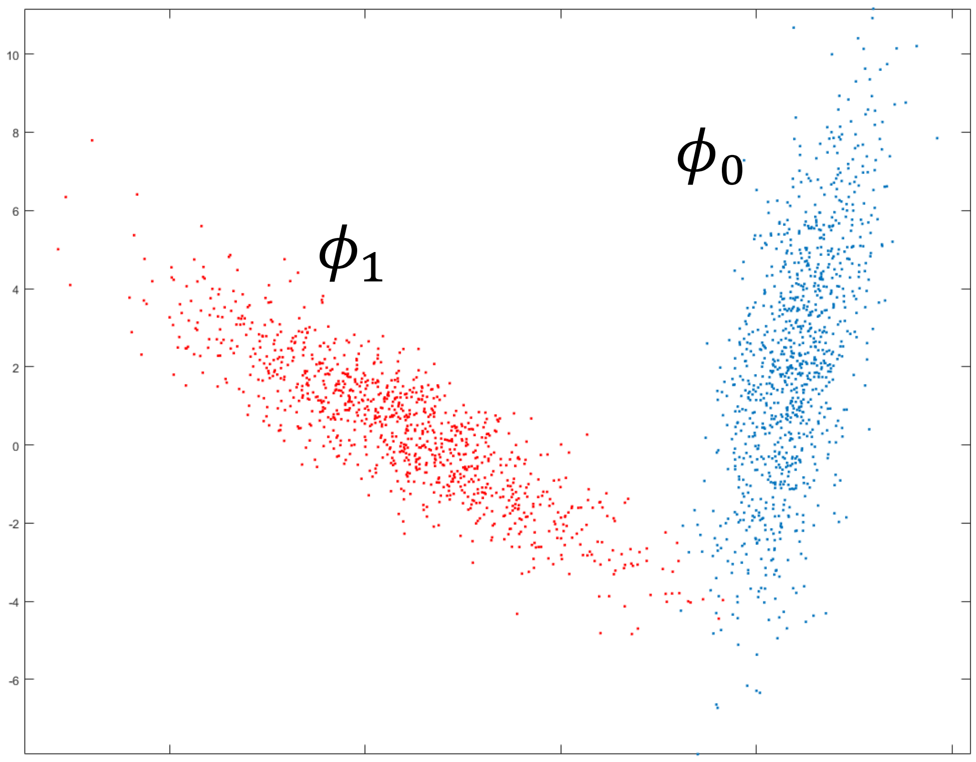
- Changes in the location of ϕ_0 (i.e, $+\boldsymbol{v}$)



The Detectability Loss: The Change Model

The change model $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ includes:

- Changes in the location of ϕ_0 (i.e, $+\boldsymbol{v}$)
- Changes in the correlation of $\boldsymbol{x}$ (i.e, $Q\boldsymbol{x}$)

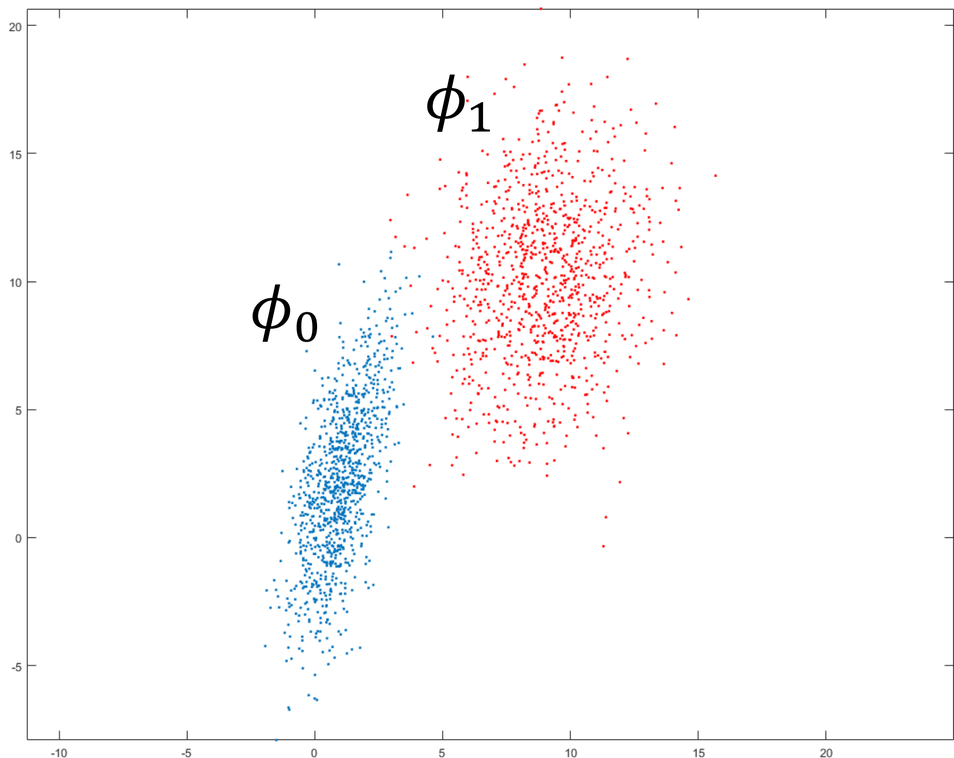


The Detectability Loss: The Change Model

The change model $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ includes:

- Changes in the location of ϕ_0 (i.e, $+\boldsymbol{v}$)
- Changes in the correlation of $\boldsymbol{x}$ (i.e, $Q\boldsymbol{x}$)

It does not include changes in the scale of ϕ_0 that can be however detected monitoring $||\boldsymbol{x}||$



The Detectability Loss: The Gaussian Assumption

Theorem

Let $\boxed{\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)}$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d \times d}$ *and orthogonal ,* $\boldsymbol{v} \in \mathbb{R}^d$*, then*

$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

The Detectability Loss: The Gaussian Assumption

Assuming $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ looks like a severe limitation.

- Other distributions are not easy to handle analytically
- We can prove that DL occurs also in random variables having independent components
- The result can be empirically extended to the apprimations of $\mathcal{L}(\cdot)$ typically used for Gaussian mixtures

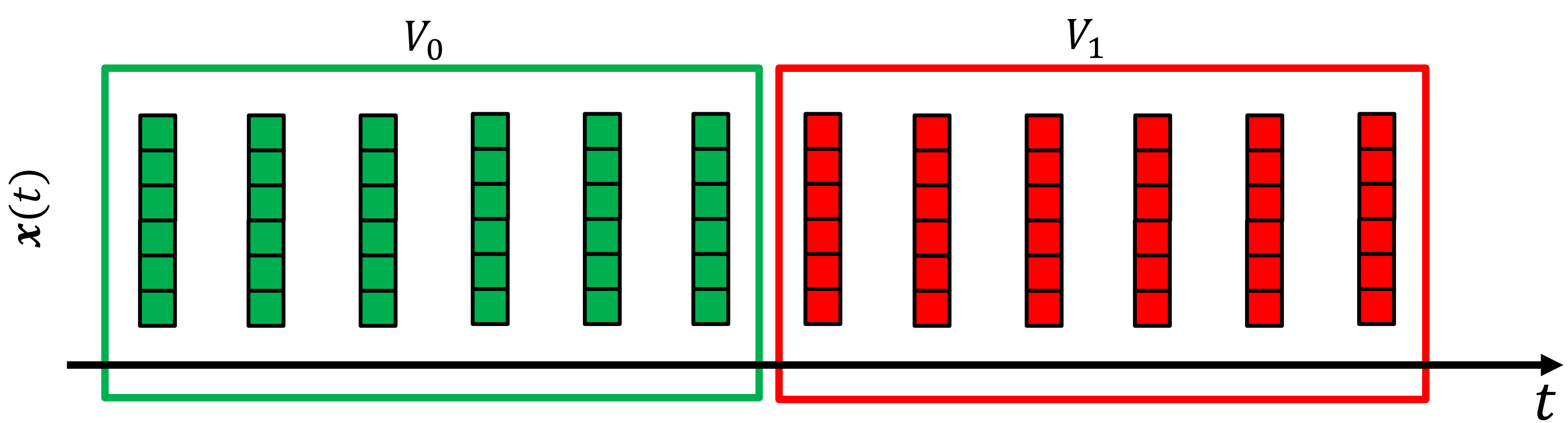
The Detectability Loss: Empirical Analysis

The data

- Two datsets from UCI database (Particle, Wine)
- Synthetically generate streams of different dimension d
- Estimate $\hat{\phi}_0$ by GM from a **stationary training set**
- In each stream we introduce $\phi_0 \to \phi_1$ such that

$$\phi_1\,(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x}\ + \boldsymbol{v})\text{ and }\ \mathrm{sKL}(\phi_0, \phi_1) = 1$$

- **Test data: two windows** V_0 and V_1 (500 samples each) selected before and after the change.



The Detectability Loss: Empirical Analysis

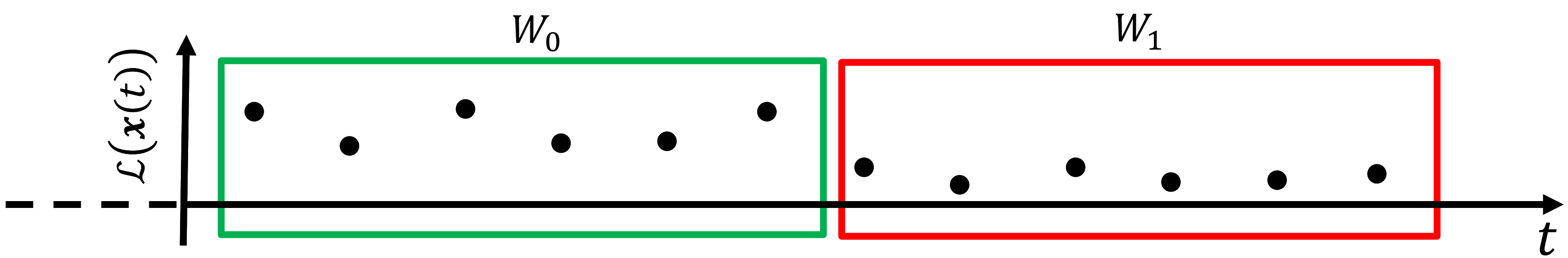
The change-detectabiltity measure:

- Compute $\mathcal{L}\left(\hat{\phi}_0(\boldsymbol{x})\right)$ from V_0 and V_1, obtaining W_0 and W_1
- Compute a test statistic $\mathcal{T}(W_0, W_1)$ to compare the two
- Detect a change by an hypothesis test

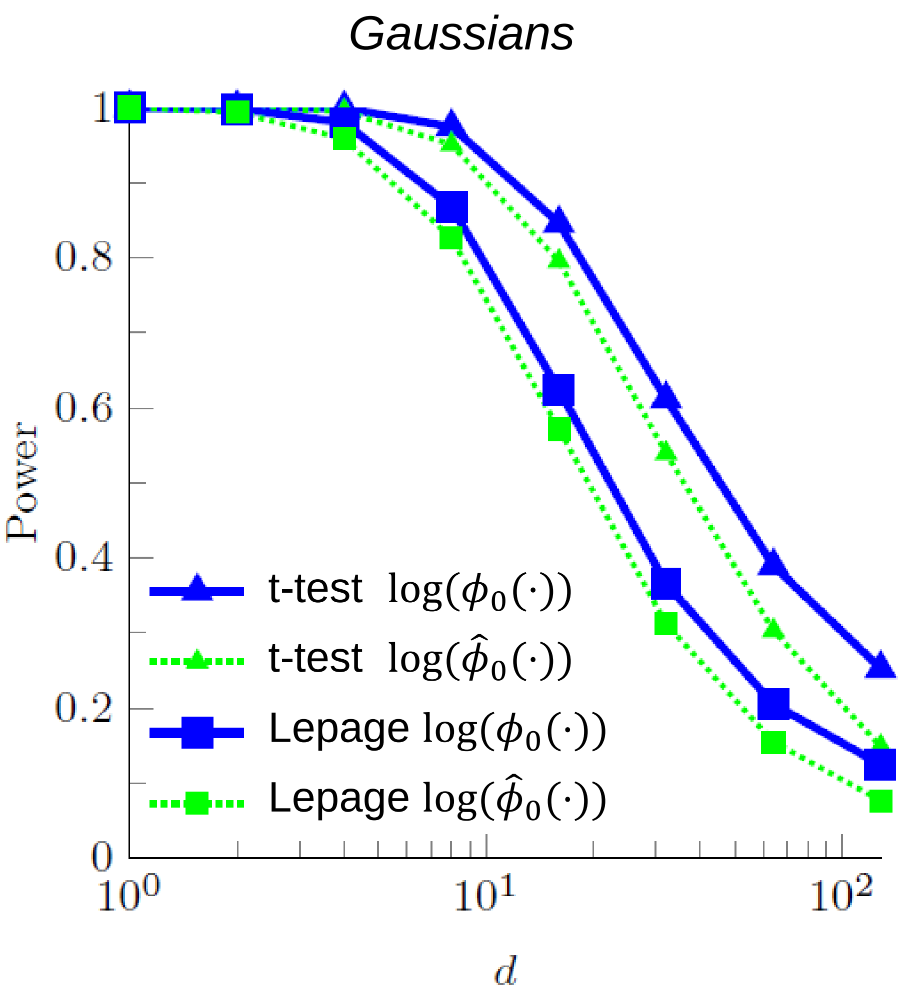
$$\mathcal{T}(W_0, W_1) \lesseqgtr h$$

where h controls the amount of false positives

- Use the **power** of this **test** to assess change detectability



DL: the Power of HTs on Gaussian Streams



Remarks:

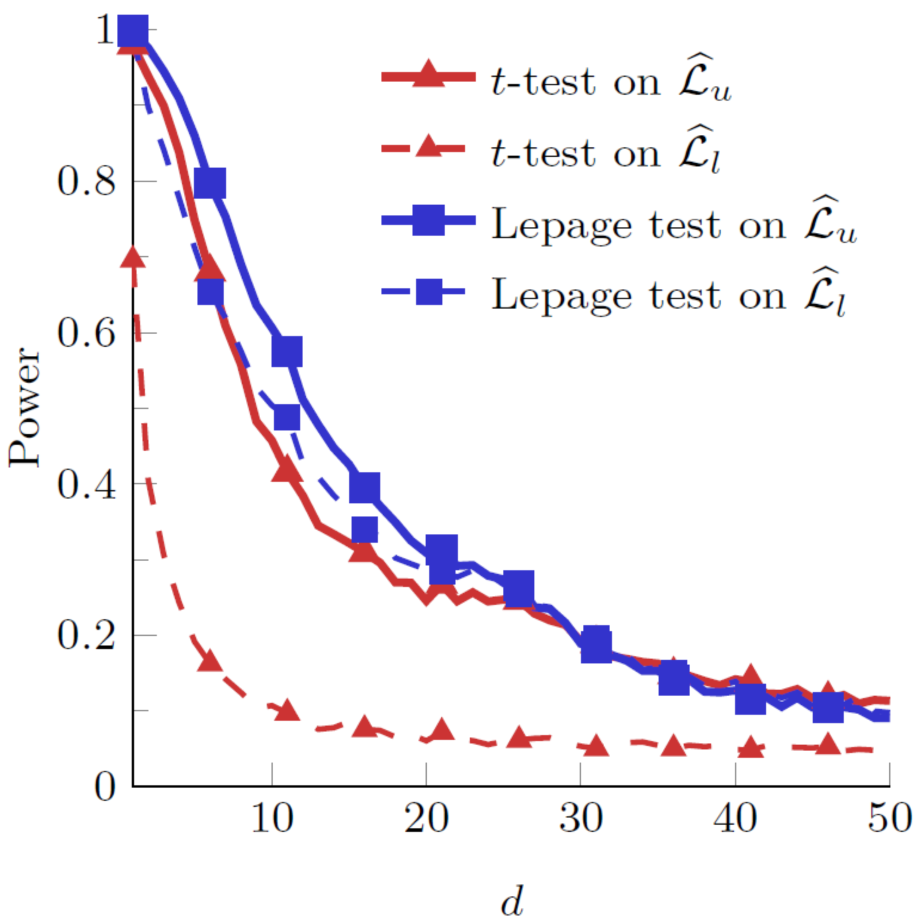
- ϕ_1 is defined analytically
- The t-test detects changes in expectation
- The Lepage test detects changes in the location and scale

Results

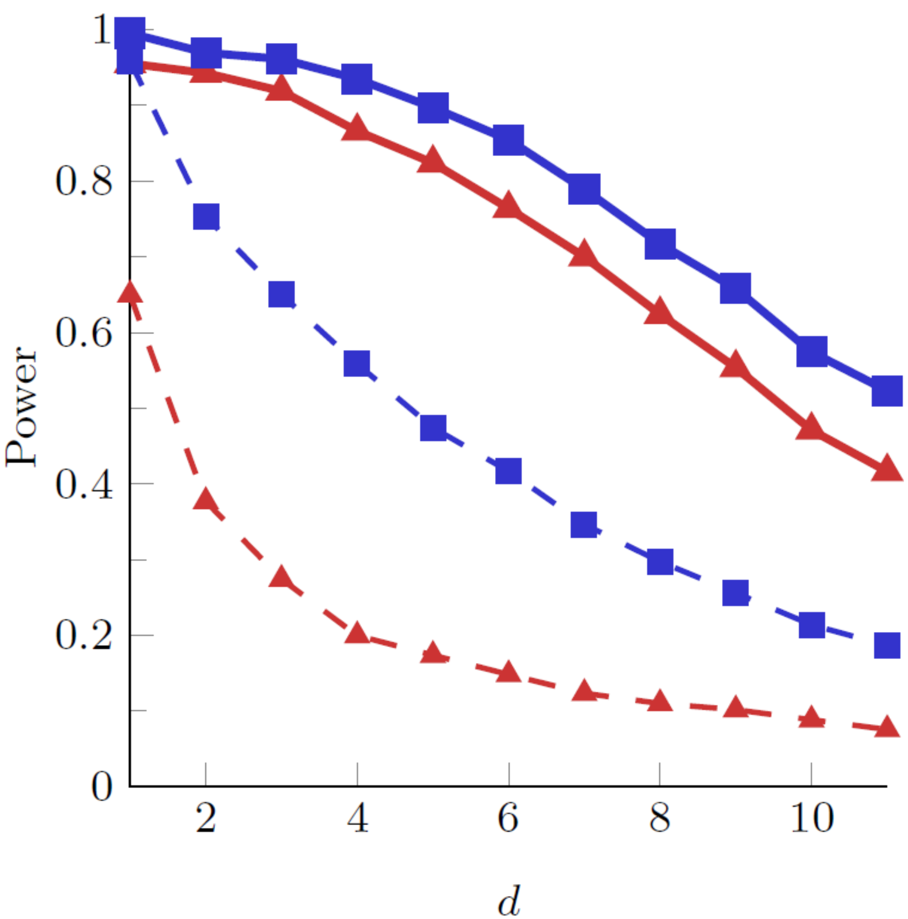
- The HT power decays with d: DL does not only concern the upperbound of SNR.
- DL is not due to estimation errors, but these make things worst.
- The power of the Lepage HT also decreases, which indicates that the change is more difficult to detect also monitoring the variance

Results: the Power of the Hypothesis Tests

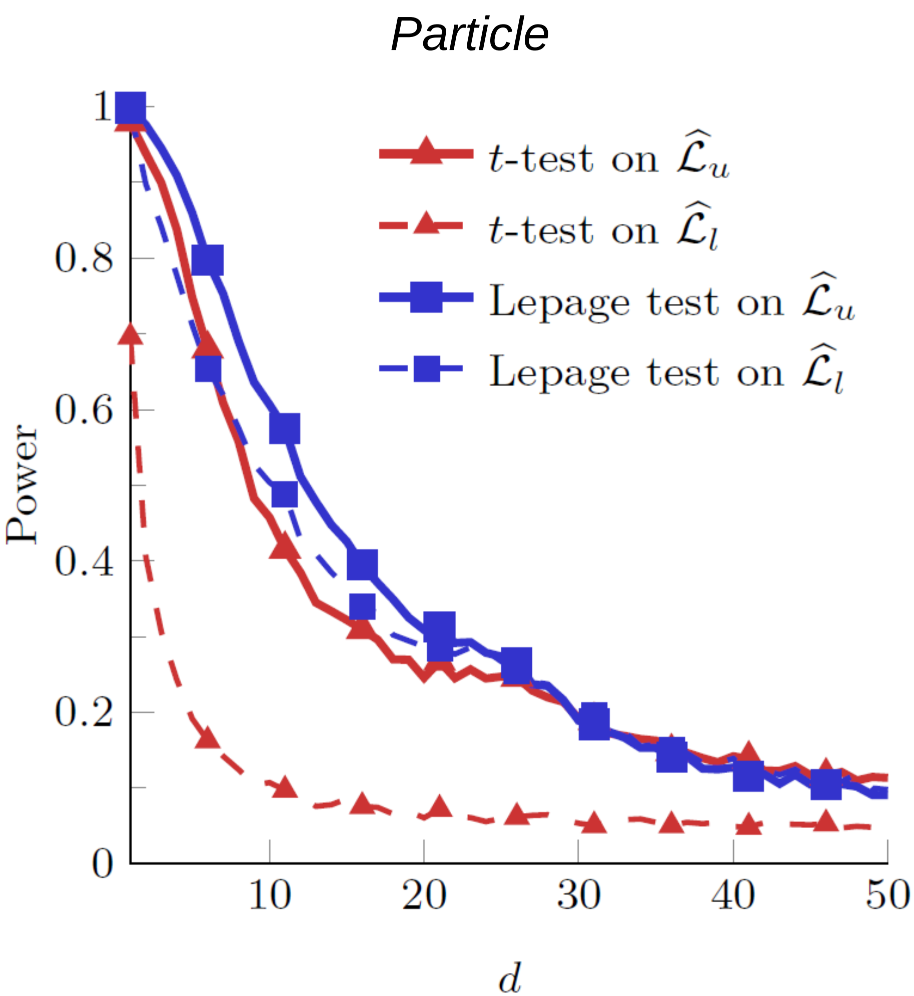
Particle



Wine



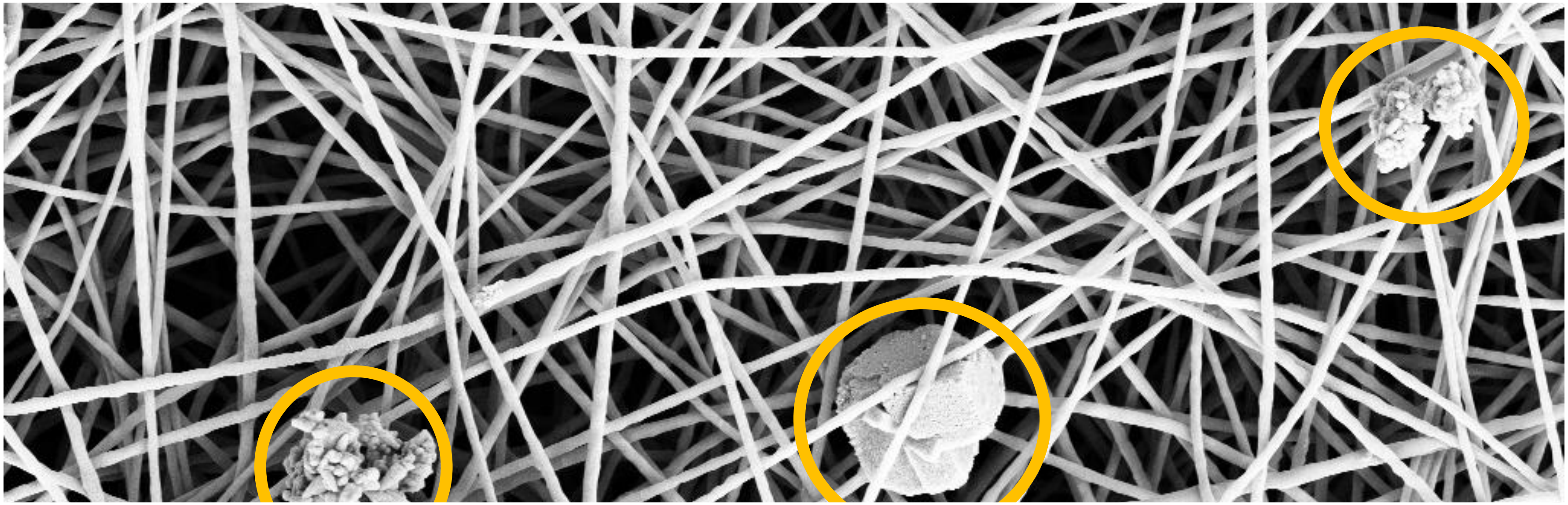
Results: the Power of the Hypothesis Tests

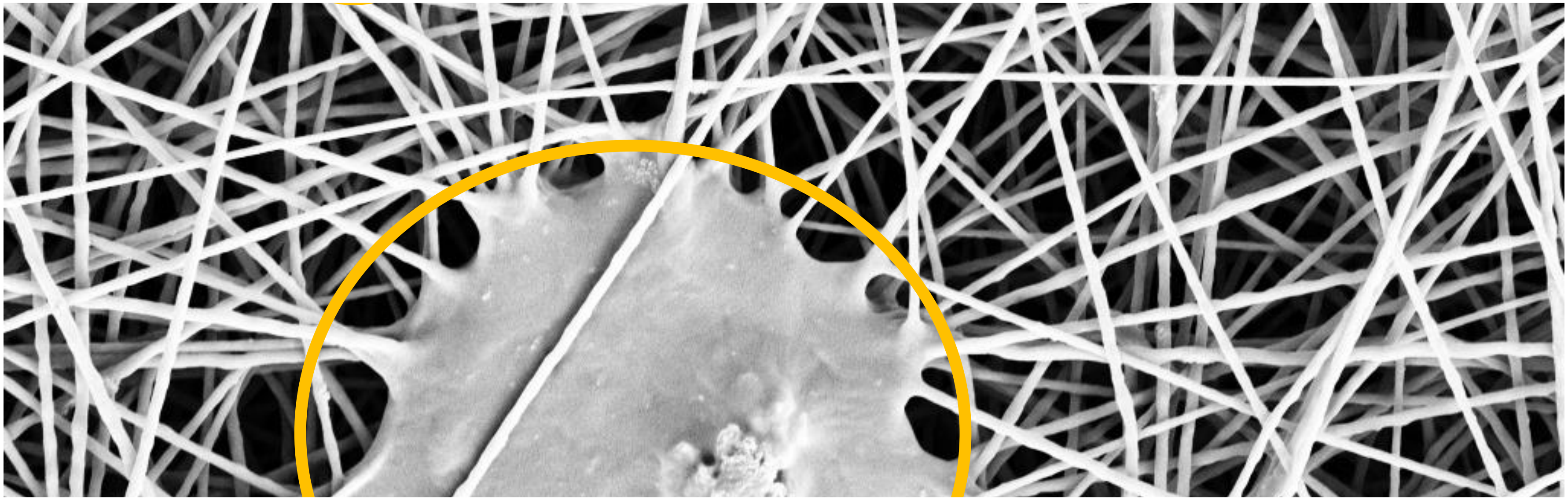


- DL: the power of Hypothesis Tests also decays with d, not just the upperbound of SNR.
- DL occurs also in non-Gaussian data
- The Lepage statistic also decreases, which indicates that the change is more difficult to detect also monitoring the variance
- Experiments on synthetic datasets confirms that DL is not due to estimation errors of $\hat{\phi}_0$

DETECTABILITY LOSS AND ANOMALY DETECTION IN IMAGES

The Considered Problem



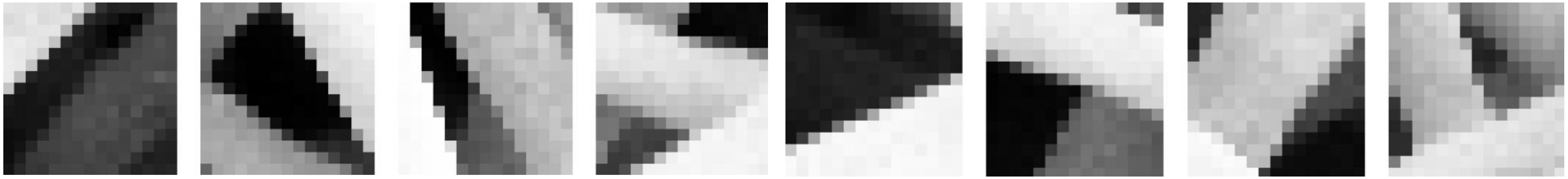


Patch-based processing of nanofibers

Analyze each patch of an image s

$$\mathbf{s}_c = \{s(c+u), u \in \mathcal{U}\}$$

and determine whether it is normal or anomalous



Patches $\mathbf{s}_c \in \mathbb{R}^p$ are **too high-dimensional** ($p \gg 0$) for modeling the distribution ϕ_0 generating normal paches

We need to **extract** suitable **features** to **reduce** the **dimensionality** of our anomaly-detection problem.

Feature Extraction

Expert-driven features: On each patch, compute

- the average,
- the variance,
- the total variation.

These are expected to **distinguish normal** and **anomalous** patches

Data-driven features: our approach consists in

1. Learning a model $\mathcal{D}$ that describes normal patches
2. Assessing the conformance of each patch $\boldsymbol{s}_c$ to $\mathcal{D}$

$\mathcal{D}$: Dictionary of patches

Sparse representations have shown to be a very useful method for **constructing signal models**

The underlying assumption is that

$$\boldsymbol{s} \approx D\boldsymbol{\alpha} \quad \text{i.e.,} \qquad \|\boldsymbol{s} - D\boldsymbol{\alpha}\|^2 \approx 0$$

and $\boldsymbol{\alpha} \in \mathbb{R}^n$ where:

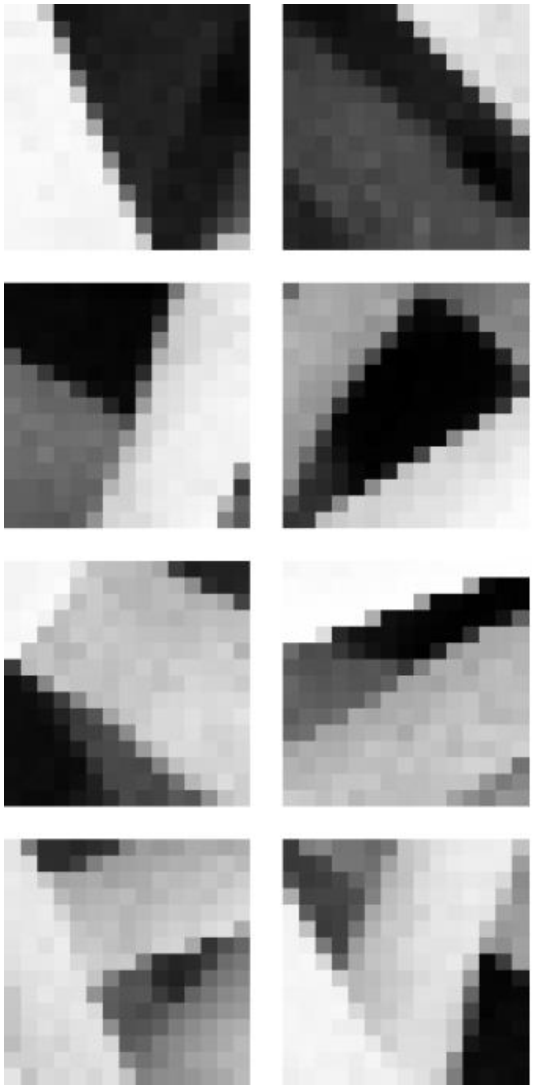
- $D \in \mathbb{R}^{p \times n}$ is the **dictionary**, columns are called **atoms**
- the coefficient vector $\mathbf{x}$ is sparse
  - $\|\boldsymbol{\alpha}\|_0 = L \ll n$ or
  - $\|\boldsymbol{\alpha}\|_1$ is small

The **dictionary** is learned a training set of **normal patches**.

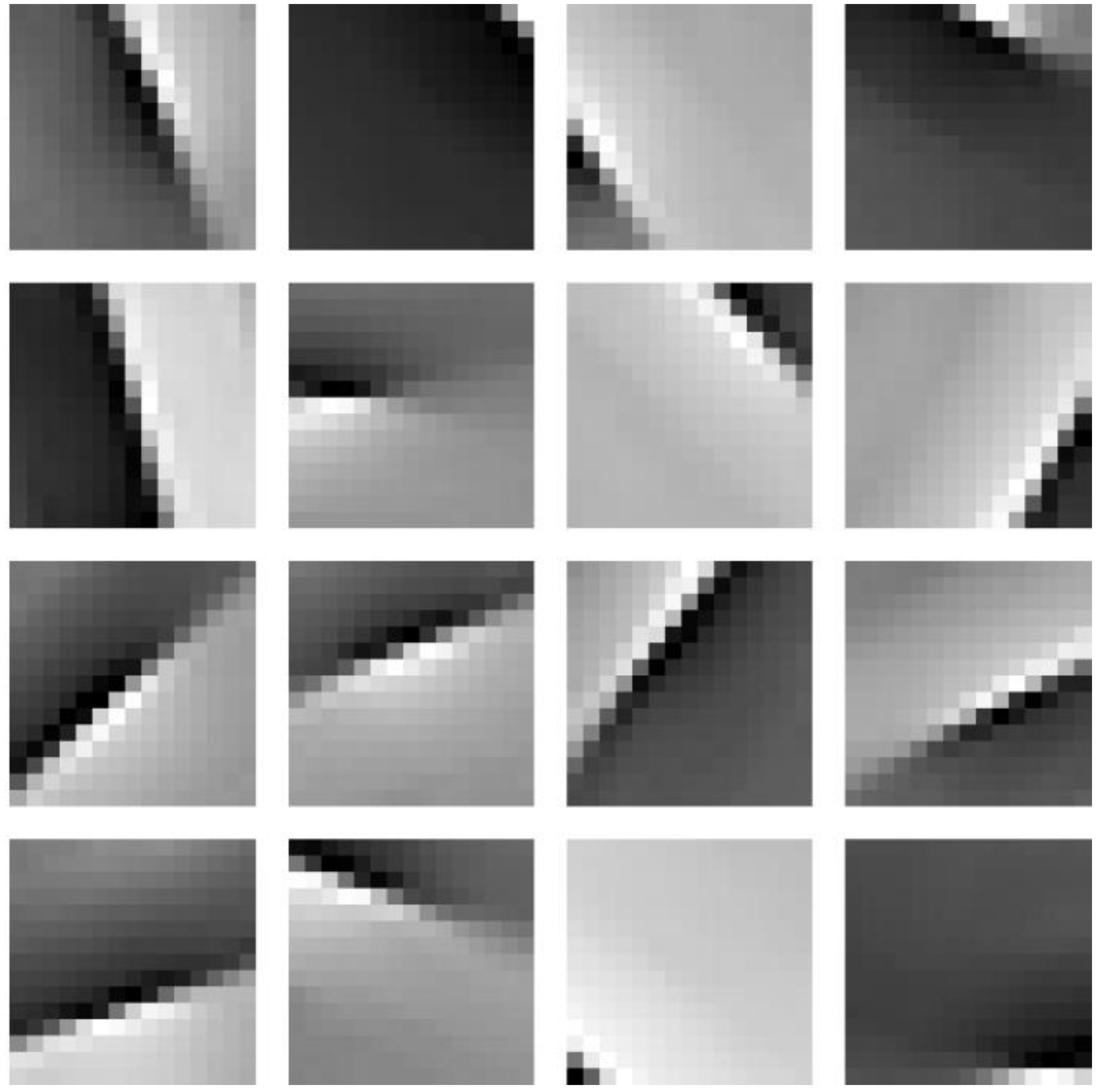
We learn a **union of low-dimensional sub-spaces** where **normal patches** live

The dictionary of normal patches

Example of training patches



Few learned atoms (BPDN-based learning)



Data-Driven Features

To assess the confrmance of $\boldsymbol{s}_c$ with $\mathcal{D}$ we perform the

Sparse coding:

$$\boldsymbol{\alpha} = \underset{\widetilde{\boldsymbol{\alpha}} \in \mathbb{R}^n}{\operatorname{argmin}} \|D\widetilde{\boldsymbol{\alpha}} - \mathbf{s}\|_2^2 + \lambda \|\widetilde{\boldsymbol{\alpha}}\|_1, \qquad \lambda > 0$$

which we solve using the BPDN problem (using ADMM).

We then measure

$$\|D\boldsymbol{\alpha} - \mathbf{s}\|_2^2$$

and

$$\|\boldsymbol{\alpha}\|_1$$

Data-driven features are $\mathbf{x} = \begin{bmatrix} \|D\boldsymbol{\alpha} - \mathbf{s}\|_2^2 \\ \|\boldsymbol{\alpha}\|_1 \end{bmatrix}$

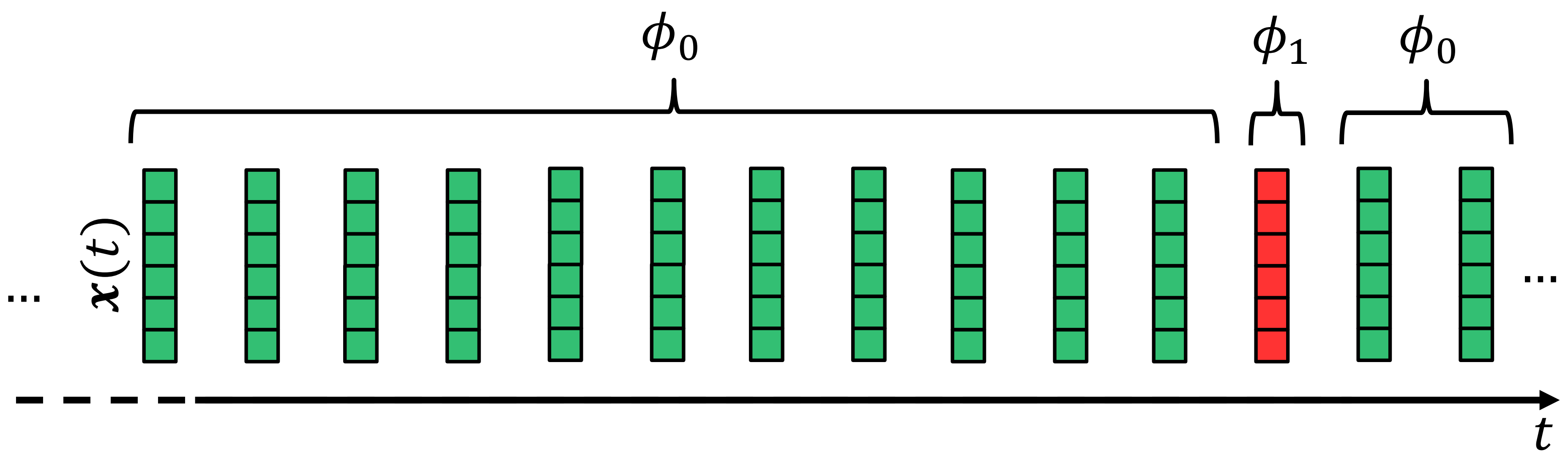
Detecting Anomalies

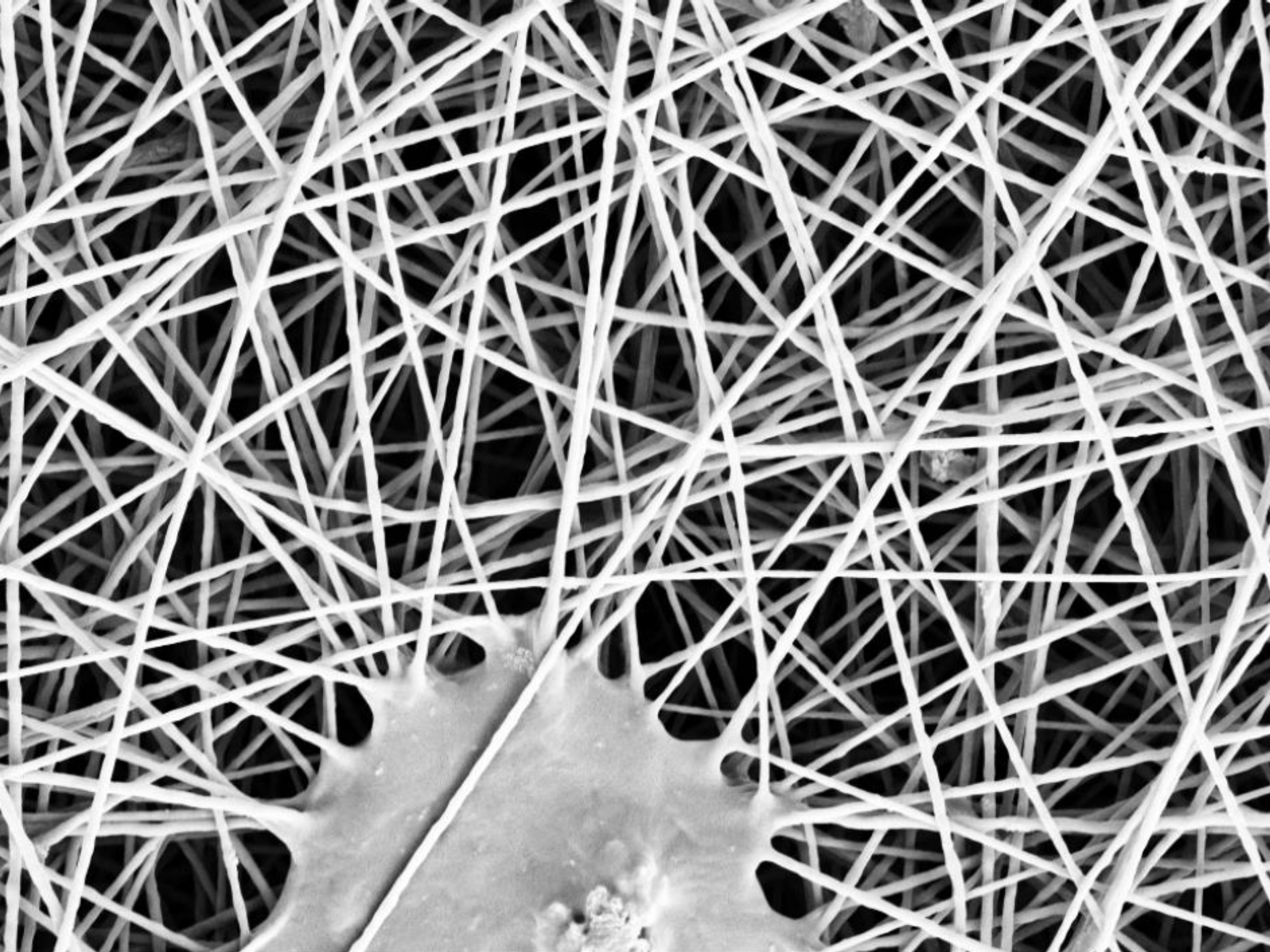
Normal patches are expected to yield features $\mathbf{x}$ that are **i.i.d.** and that **follow a** (unknown) **distribution** ϕ_0, **anomalous patches do not,** as they follow $\phi_1 \neq \phi_0$

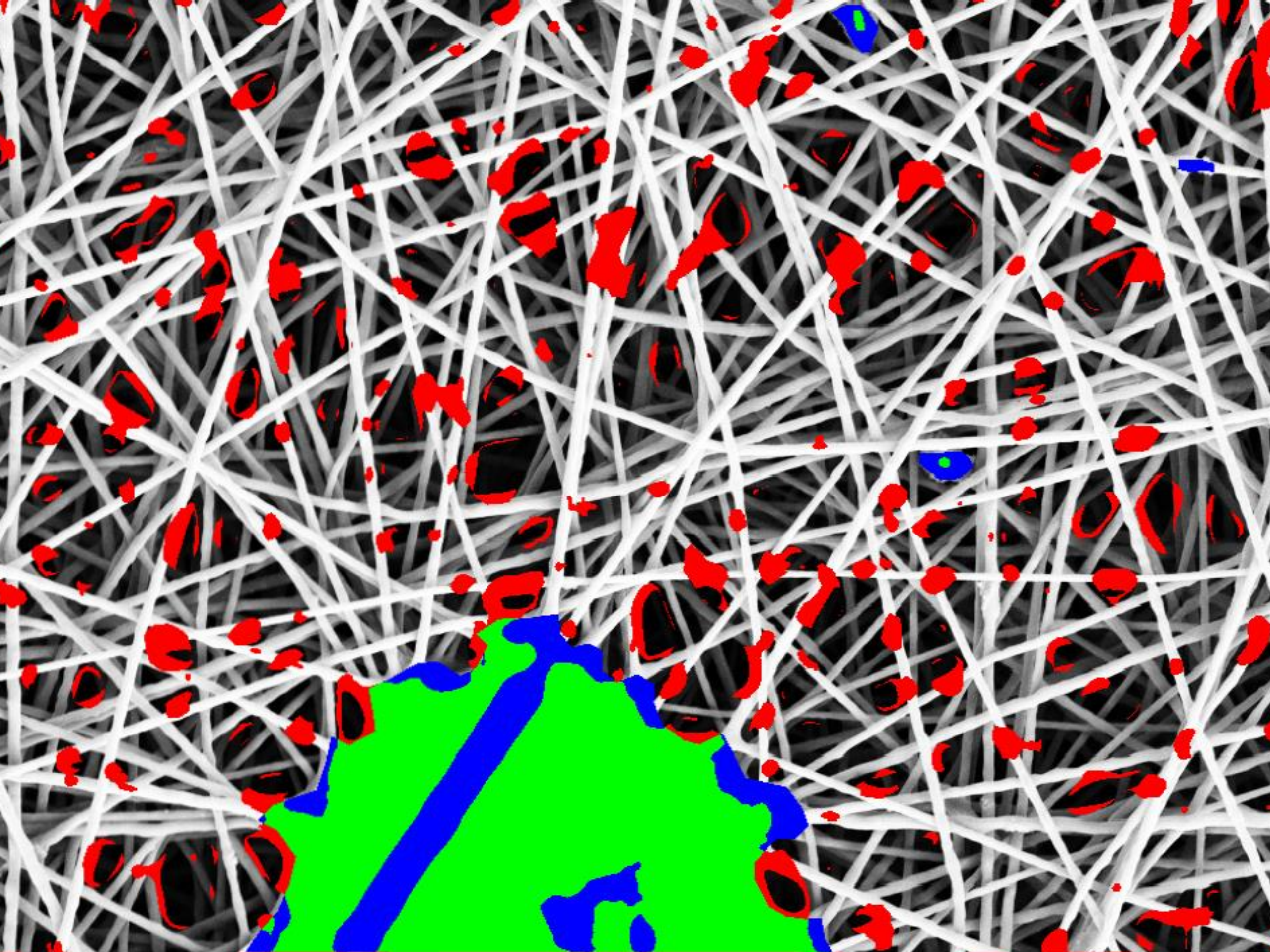
We are back to the original problem

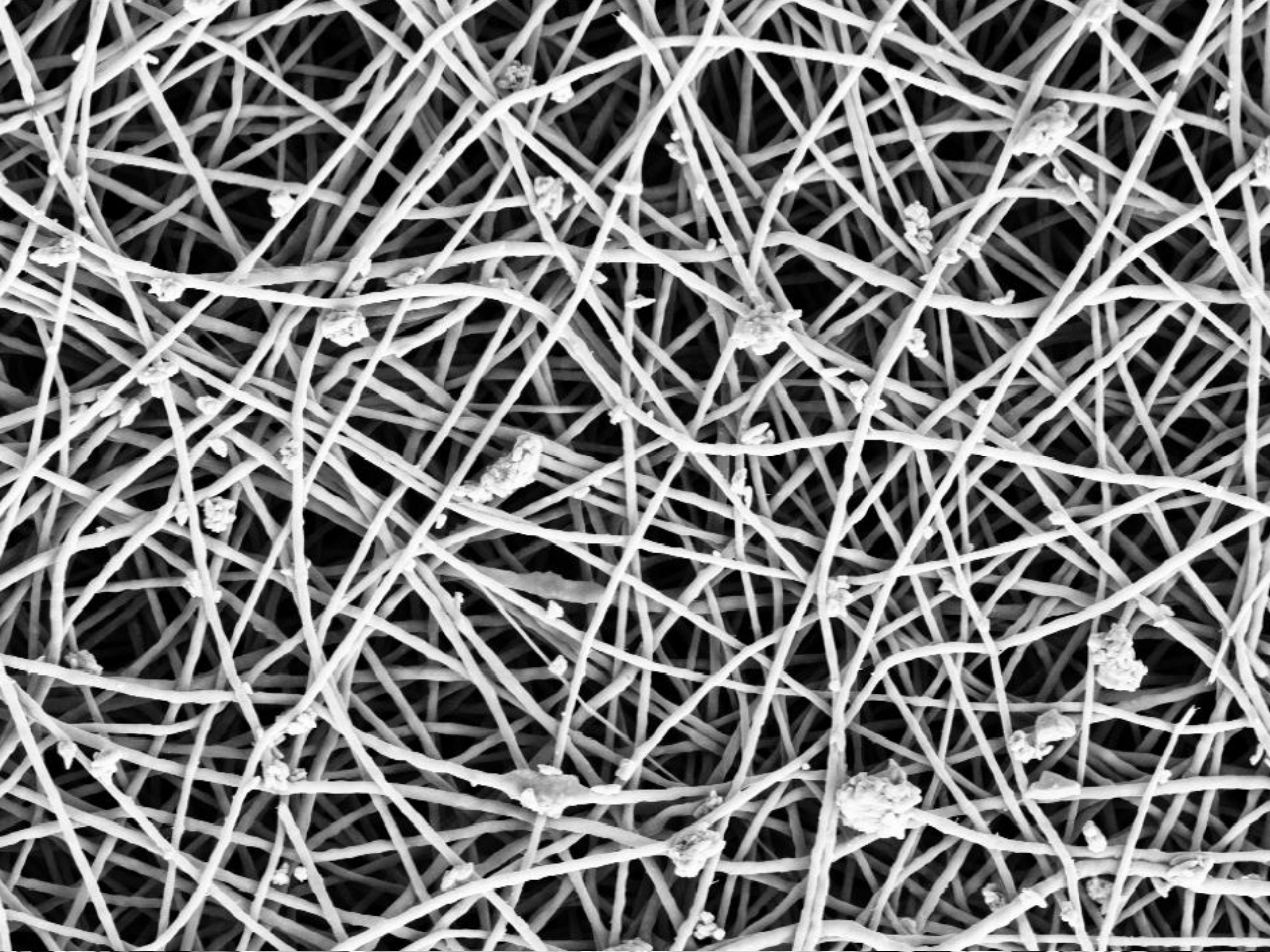
"Determining whether a set of data $\{\boldsymbol{x}_c,\ c = 1, \dots\}$ ***is generated from*** $\boldsymbol{\phi_o}$ *and detect possible* ***outliers"***

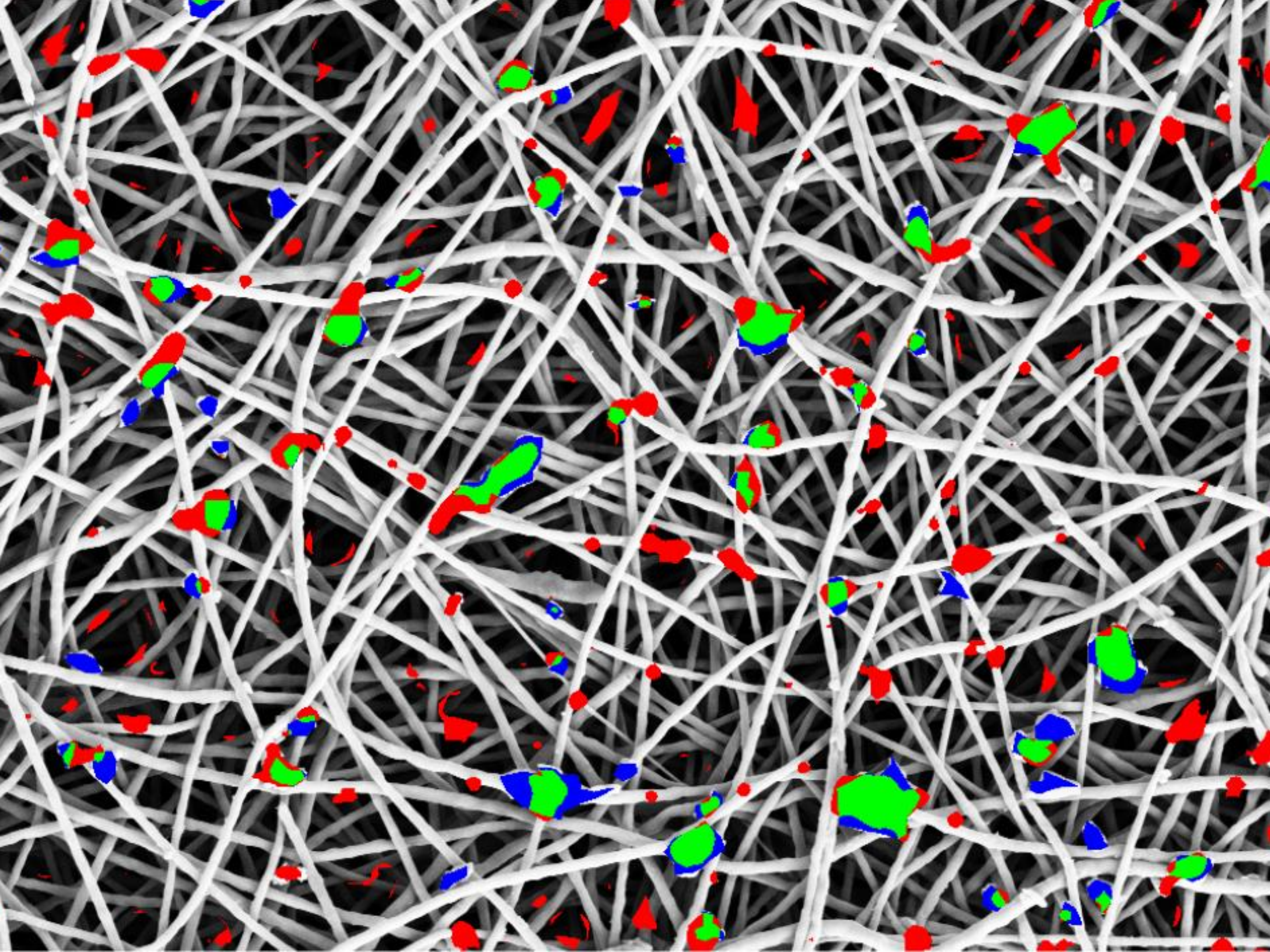
Anomaly Detection Problem







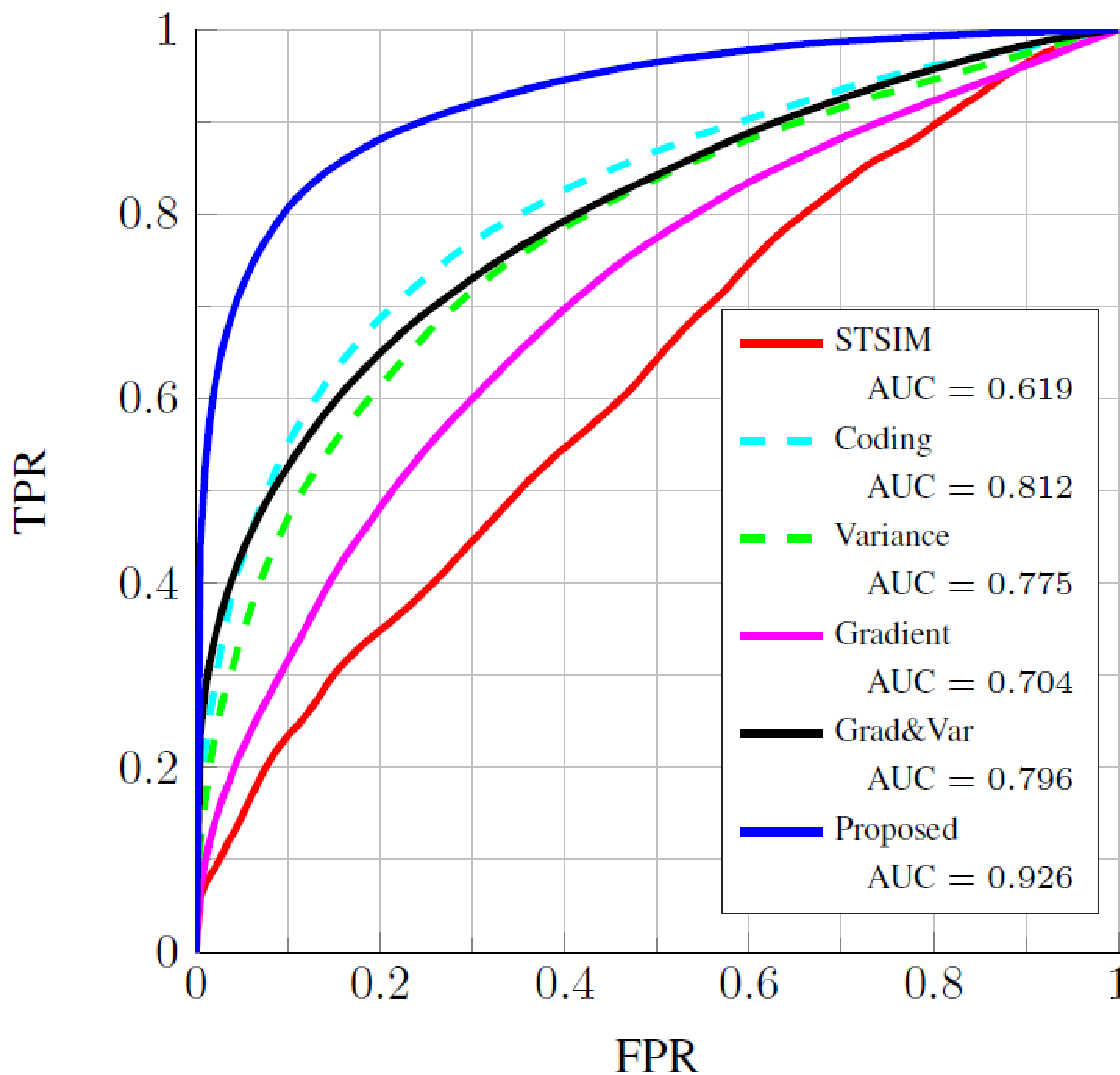




The ROC curves

Tests on 40 images with anomalies manually annotated by an expert

The proposed anomaly detection algorithm outperforms expert-driven features and other methods based on sparse representations



Detectability Loss on these nanofibers

Selecting the good features is obviously important.

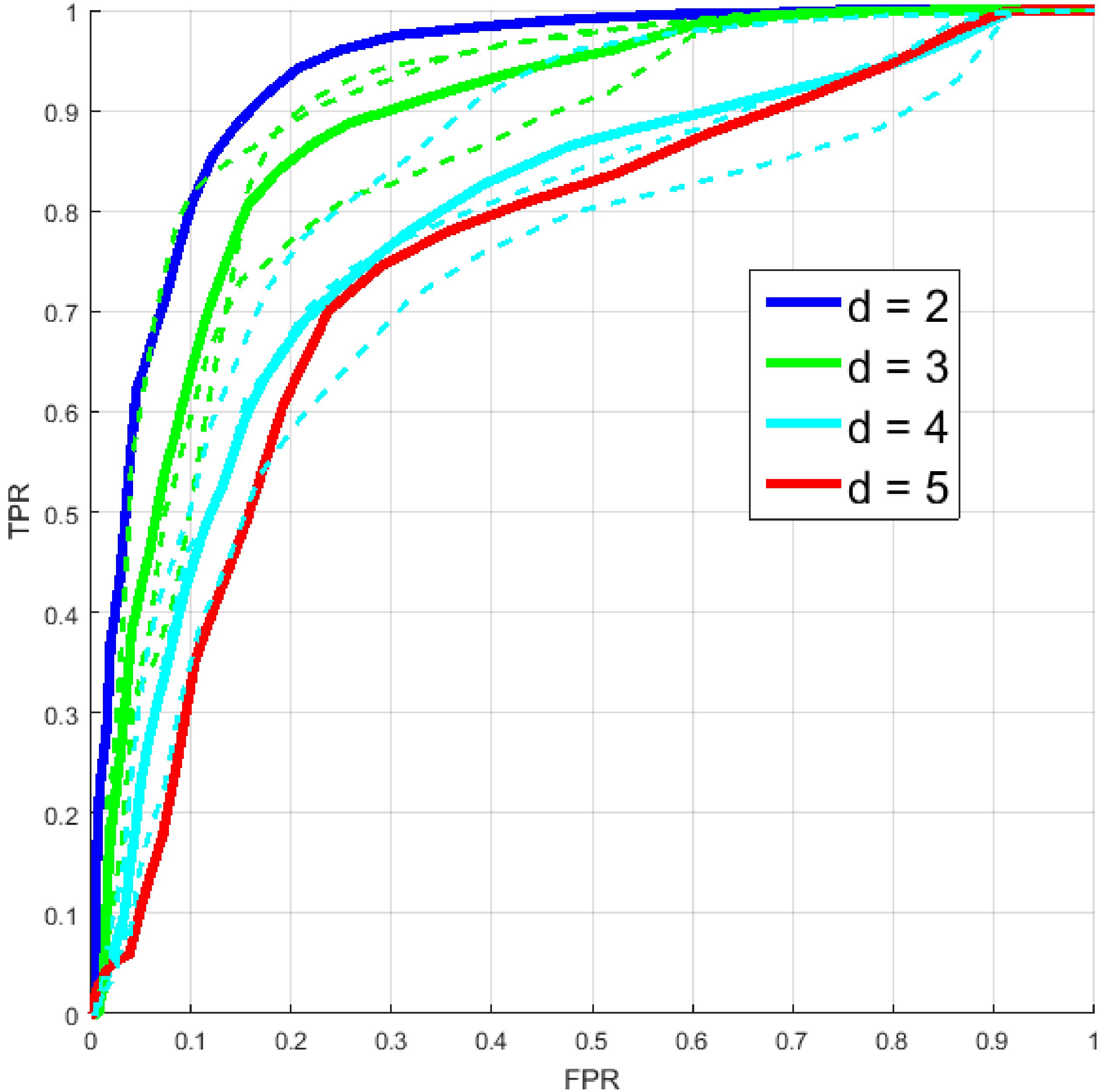
Why not stacking data-driven and expert-driven features?

Consider $d = 3, 4, 5$ dimensional features

- We selectively add the three expert-driven features to the two data-driven ones
- We always fit a GM model to a large-enough number of training data

Detectability Loss on these nanofibers

Anomaly detection performance progressively decay when d increases



Detectability Loss and Irrelevant Features

Irrelevant features, namely features that:

- are not directly affected by the change
- do not provide any additional information for change detection purposes (i.e. leave $s\mathrm{KL}(\phi_0, \phi_1)$ constant)

Adding irrelevant feature yields detectability loss.

Other issues might cause the performance decay

- A biased denisty function for $\hat{\phi}_0$
- Scarcity of training samples when d increases

However, we are inclined to conclude that

- These expert-driven features do not add enough relevant information on top of the data-driven ones (for anomaly-detection purposes).

Obviously is not always the case

We developed data-driven features based on c**onvolutional sparse models**

$$\boldsymbol{s} \approx \sum_{i=1}^{n} \boldsymbol{d}_i \circledast \boldsymbol{\alpha}_i \,, \qquad \text{s.t. } \boldsymbol{\alpha}_i \text{ is sparse}$$

where a signal $\boldsymbol{s}$ is **entirely encoded** as the sum of n convolutions between a filter $\boldsymbol{d}_i$ and a coefficient map $\boldsymbol{\alpha}_i$

Pros:

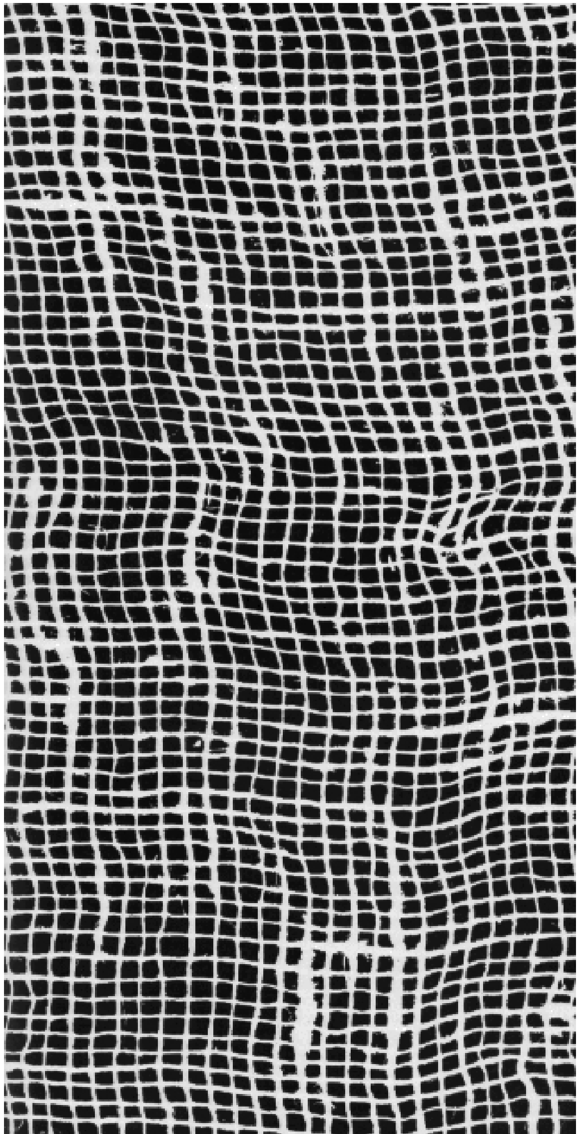
- Translation invariant representation
- Few small filters are typically required
- Filters exhibit very specific image structures
- Easy to use filters having different size

Collaboration with Los Alamos National Laboratory, NM, USA

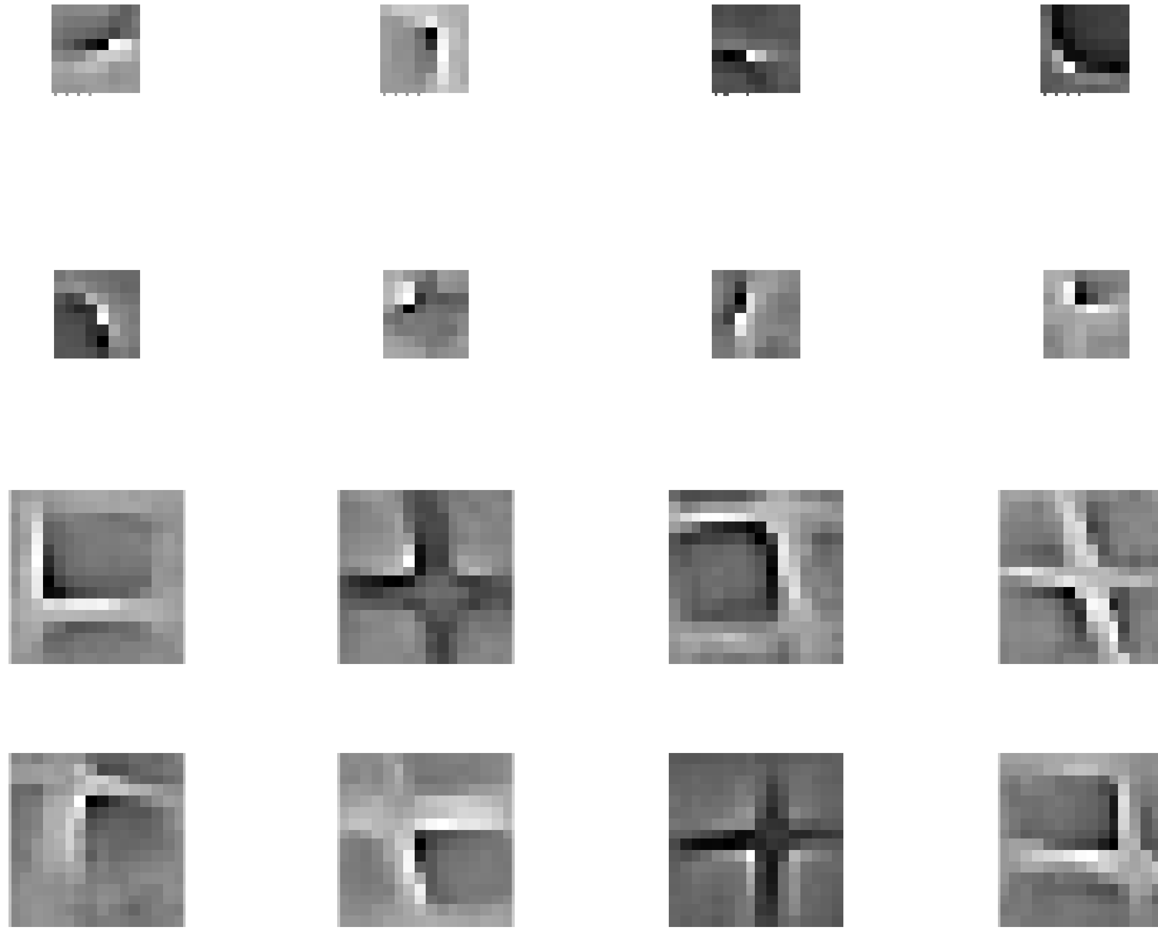
Los Alamos NATIONAL LABORATORY

Example of Learned Filters

Training Image



Learned Filters



Convolutional Sparsity for Anomaly Detection

If we consider the convolutional sparse coding

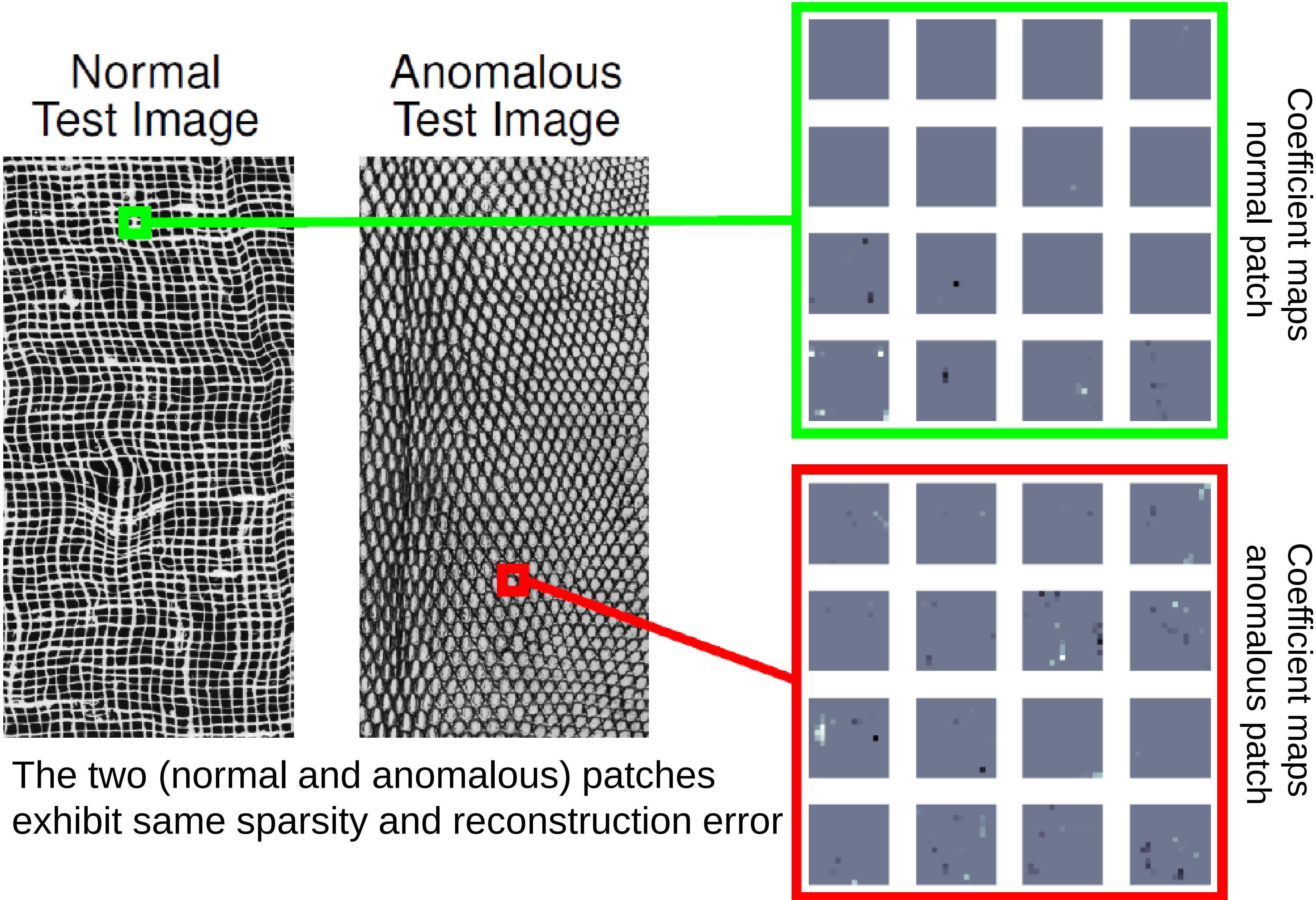
$$\{\hat{\boldsymbol{\alpha}}\} = \underset{\{\boldsymbol{\alpha}\}_n}{\operatorname{argmin}} \underline{\left\| \sum_{i=1}^{n} \boldsymbol{d}_i \circledast \boldsymbol{\alpha}_i - \mathbf{s} \right\|_2^2} + \lambda \underline{\sum_{i=1}^{n} \|\boldsymbol{\alpha}\|_1}$$

we can build the feature vector as:

$$\boldsymbol{x}_c = \begin{bmatrix} \left\| \prod_c \left(\sum_{i=1}^{n} \boldsymbol{d}_i \circledast \hat{\boldsymbol{\alpha}}_i - \mathbf{s} \right) \right\|_2^2 \\ \sum_{i=1}^{n} \left\| \prod_c \hat{\boldsymbol{\alpha}} \right\|_1 \end{bmatrix}$$

…but unfortunately, detection performance are rather poor

Sparsity is too loose a criterion for detection



The two (normal and anomalous) patches exhibit same sparsity and reconstruction error

Convolutional Sparsity for Anomaly Detection

Add the **group sparsity** of the maps on the patch support as an **additional feature**

$$x_c = \begin{bmatrix} \left\| \prod_c \left(\sum_{i=1}^{m} \boldsymbol{d}_i \circledast \widehat{\boldsymbol{\alpha}}_i - \boldsymbol{s} \right) \right\|_2^2 \\ \sum_{i=1}^{m} \left\| \prod_c \widehat{\boldsymbol{\alpha}} \right\|_1 \\ \boxed{\sum_{i=1}^{m} \left\| \prod_c \widehat{\boldsymbol{\alpha}} \right\|_2} \end{bmatrix}$$

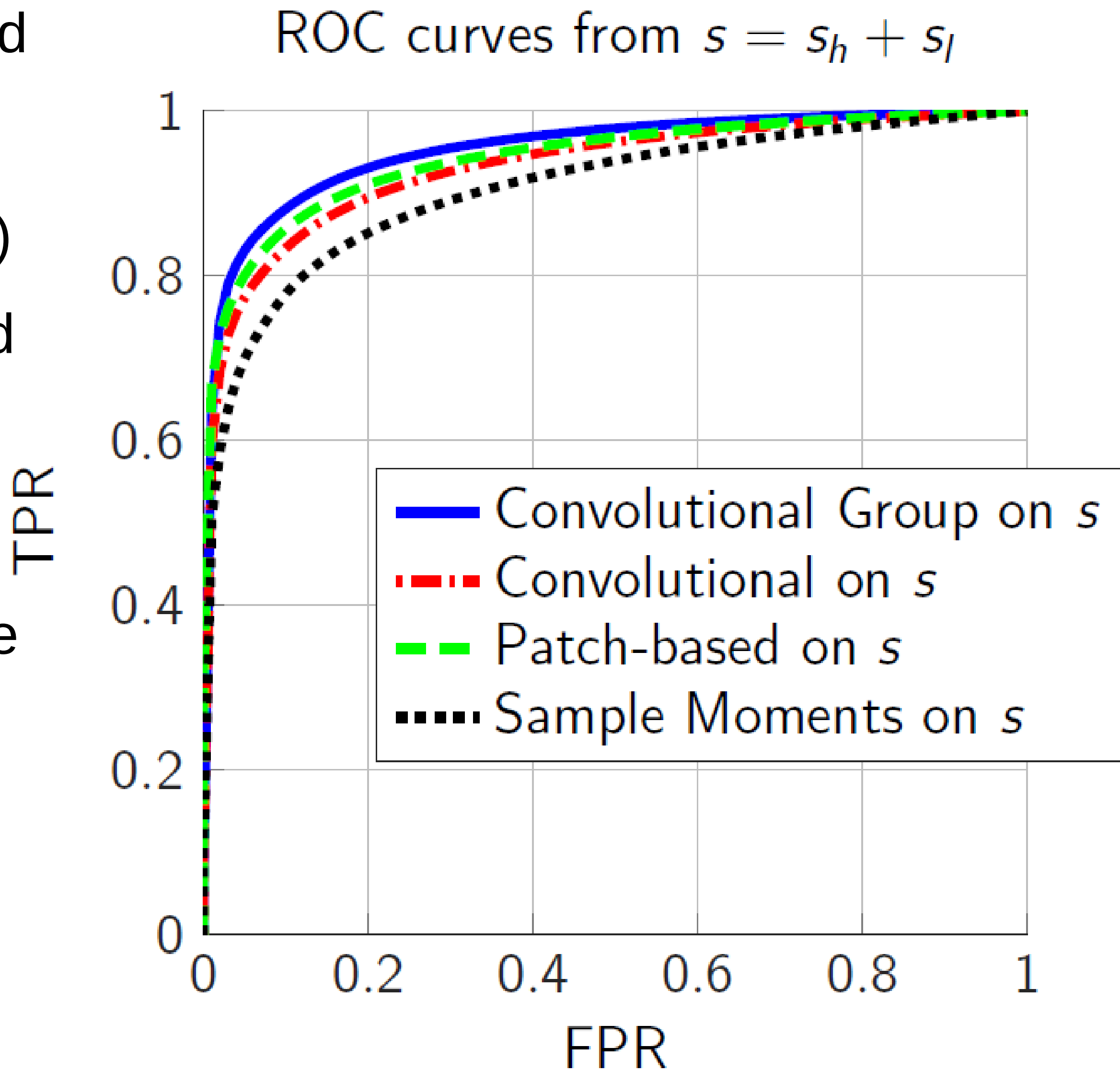
D. Carrera, G. Boracchi, A. Foi and B. Wohlberg , *"Detecting Anomalous Structures by Convolutional Sparse Models "* IEEE IJCNN 2015

Anomaly-Detection Performance

On 25 different textures and 600 test images (pair of textures to mimic normal/anomalous regions)

Best performance achieved by the 3-dimensional feature indicators

Achieve similar performance than steerable pyramid specifically designed for texture classification

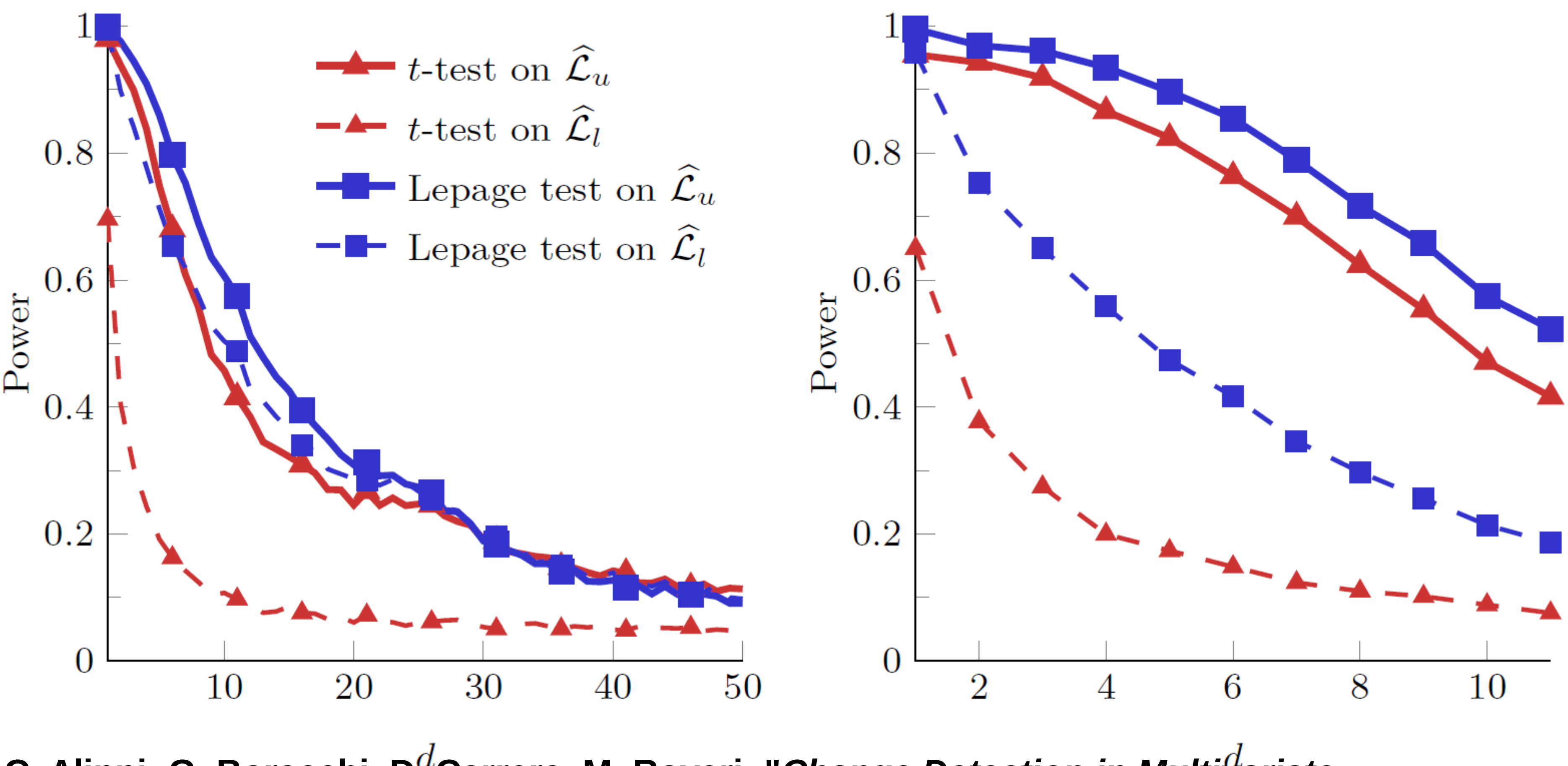


CONCLUDING REMARKS

Comments on Detectability Loss

- Detectability loss occurs:
  - independently on the specific statistical tool used to monitor the log-likelihood
  - does not depend on *how* the change affects ϕ_0, e.g. the number of affected components.
- Empirical analysis confirms DL on real-world datastreams.
  - It is important to keep the change-magnitude constant when changing d (or the dataset)
- Irrelevant components in $\boldsymbol{x}$ are harmful! Consider this in feature-based anomaly-detection methods.
- Ongoing works: extending this study to other change-detection approaches and to other families of distributions.
- Further details http://arxiv.org/pdf/1510.04850v2

Thanks, Questions?



C. Alippi, G. Boracchi, D. Carrera, M. Roveri, "*Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss*" IJCAI 2016, New York, USA, July 9 - 13

D. Carrera, G. Boracchi, A. Foi and B. Wohlberg "*Detecting Anomalous Structures by Convolutional Sparse Models*" IJCNN 2015 Killarney, Ireland, July 12

D. Carrera, F. Manganini, G. Boracchi, E. Lanzarone *"Defect Detection in Nanostructures"*, IEEE Transactions on Industrial Informatics -- Submitted, 11 pages.

BACKUP SLIDES

Sketch of the proof

Theorem

Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and let* $\phi_1(\boldsymbol{x}) = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$ *where* $Q \in \mathbb{R}^{d\times d}$ *and orthogonal ,* $\boldsymbol{v} \in \mathbb{R}^d$***,*** *then*

$$\mathrm{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}$$

Where C *is a constant that depends only on* $\mathrm{sKL}(\phi_0, \phi_1)$

Sketch of the proof: recall

$$\mathrm{SNR}(\phi_0 \to \phi_1) = \frac{\left(\mathop{\mathrm{E}}_{\boldsymbol{x}\sim\phi_0}[\mathcal{L}(\boldsymbol{x})] - \mathop{\mathrm{E}}_{\boldsymbol{x}\sim\phi_1}[\mathcal{L}(\boldsymbol{x})]\right)^2}{\mathop{\mathrm{var}}_{\boldsymbol{x}\sim\phi_0}[\mathcal{L}(\boldsymbol{x})] + \mathop{\mathrm{var}}_{\boldsymbol{x}\sim\phi_1}[\mathcal{L}(\boldsymbol{x})]}$$

We compute an upper bound of the numerator and a lower bound of the denominator

Sketch of the proof

We now show that

$$\mathrm{sKL}(\phi_0, \phi_1) \geq \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})] - \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})] \qquad (*)$$

From $\mathcal{L}(\boldsymbol{x}) = \log(\phi_0(\boldsymbol{x}))$ and the definition of sKL it follows

$$\mathrm{sKL}(\phi_0, \phi_1) = \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\log(\phi_0(\boldsymbol{x})] - \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x})]] + \\ + \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x})]] - \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\log(\phi_0(\boldsymbol{x})]$$

Thus

$$(*) \quad \Leftrightarrow \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x})]] - \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x})]] \geq 0$$

Sketch of the proof

$$\underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x}))] - \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x}))] =$$
$$= \int \log\big(\phi_1(\boldsymbol{x})\big)\phi_1(\boldsymbol{x})d\boldsymbol{x} - \int \log\big(\phi_1(\boldsymbol{x})\big)\phi_0(\boldsymbol{x})d\boldsymbol{x}$$

We denote

$$\boldsymbol{y} = Q'(\boldsymbol{x}-\boldsymbol{v}), \qquad \boldsymbol{x} = Q\boldsymbol{y}+\boldsymbol{v}$$
$$d\boldsymbol{y} = |\det(Q')|d\boldsymbol{x} = d\boldsymbol{x}$$
$$\phi_0(\boldsymbol{x}) = \phi_1\big(Q'(\boldsymbol{x}-\boldsymbol{v})\big) = \phi_1(\boldsymbol{y})$$
$$\phi_1(\boldsymbol{x}) = \phi_1(Q\boldsymbol{y}+\boldsymbol{v}) =: \phi_2(\boldsymbol{y})$$

then

$$\underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x}))] - \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\log(\phi_1(\boldsymbol{x}))] =$$
$$= \int \log\big(\phi_1(\boldsymbol{x})\big)\phi_1(\boldsymbol{x})d\boldsymbol{x} - \int \log\big(\phi_2(\boldsymbol{y})\big)\phi_1(\boldsymbol{y})d\boldsymbol{x} =$$
$$= \mathrm{KL}(\phi_1, \phi_2) \geq 0$$

Sketch of the proof

Thus

$$\mathrm{sKL}(\phi_0, \phi_1) \geq \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})] - \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})]$$

Moreover

$$\underset{\boldsymbol{x}\sim\phi_0}{\mathrm{var}}[\mathcal{L}(\boldsymbol{x})] = \underset{\boldsymbol{x}\sim\phi_0}{\mathrm{var}}\left[-\frac{1}{2}\chi^2\right] = \frac{d}{2}$$

It follows

$$\mathrm{SNR}(\phi_0 \to \phi_1) = \frac{\left(\underset{\boldsymbol{x}\sim\phi_0}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})] - \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{E}}[\mathcal{L}(\boldsymbol{x})]\right)^2}{\underset{\boldsymbol{x}\sim\phi_0}{\mathrm{var}}[\mathcal{L}(\boldsymbol{x})] + \underset{\boldsymbol{x}\sim\phi_1}{\mathrm{var}}[\mathcal{L}(\boldsymbol{x})]} \leq \frac{\mathrm{sKL}(\phi_0, \phi_1)^2}{d/2}$$