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

Examples of heartbeat acquired from Pulse Sensor
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 \( \{ x(t), t = 1, \ldots \} \), \( x(t) \in \mathbb{R}^d \) of realizations of a random variable, and detect the change-point \( \tau \),

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

where \( \{ x(t), t < \tau \} \) are i.i.d. and \( \phi_0 \neq \phi_1 \), \( \phi_1 \) is unknown and \( \phi_0 \) can be possibly estimated from training data.
The Change-Detection Problem

Often, these problems boil down to:

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

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

where \( \{x(t), t < \tau\} \) are i.i.d. and \( \phi_0 \neq \phi_1 \), \( \phi_1 \) is unknown and \( \phi_0 \) can be possibly estimated from training data.
The Change-Detection Problem

Often, these problems boil down to:

b) Determining whether a set of data \( \{x(t), \ t = t_0, ..., t_1\} \) is generated from \( \phi_0 \) and detect possible outliers.

We refer to:

- \( \phi_0 \) pre-change distribution / normal (can be estimated)
- \( \phi_1 \) post-change distribution / anomalous (unknown)
The Change-Detection Problem

Often, these problems boil down to:

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

We refer to

- \( \phi_0 \) pre-change distribution / normal (can be estimated)
- \( \phi_1 \) 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.
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 $\phi_0$ and $\phi_1$
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 \( \phi_0 \) and \( \phi_1 \)

**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 \textbf{detectability loss} problem, i.e. that change \textbf{detectability} steadily \textbf{decreases} when $d$ increases.

Detectability loss is shown by:

- Analytical derivations: when $\phi_0$ and $\phi_1$ are \textbf{Gaussians}
- Empirical analysis: measuring the the \textbf{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
To detect the change $\phi_0 \rightarrow \phi_1$ we assume that

- $\phi_0$ is unknown, can be estimated from a training set $TR = \{x(t), t < t_0, x \sim \phi_0\}$
- $\phi_1$ is unknown, no training data are provided

We refer to

- $\phi_0$ as stationary / normal / pre-change distribution
- $\hat{\phi}_0$ as the estimate of $\phi_0$ from a training set
- $\phi_1$ 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-likelihood

1. During training, estimate $\hat{\phi}_0$ from $TR$
2. During testing, compute
   \[
   \mathcal{L}(x(t)) = \log(\hat{\phi}_0(x(t)))
   \]
3. Monitor $\{\mathcal{L}(x(t)), t = 1, \ldots\}$
A typical approach to monitor the log-likelihood

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

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


Our Goal / 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
The Change Magnitude

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

$$sKL(\phi_0, \phi_1) = KL(\phi_0, \phi_1) + KL(\phi_1, \phi_0) =$$

$$= \int \log \left( \frac{\phi_0(x)}{\phi_1(x)} \right) \phi_0(x) dx + \int \log \left( \frac{\phi_1(x)}{\phi_0(x)} \right) \phi_1(x) dx$$

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

Our Goal / Presentation Outline

- Preliminaries:
  - The change-detection approach
  - The change magnitude
  - The measure of change detectability

- The detectability loss

- Concluding remarks
The Change Detectability

The *Signal to Noise Ratio of the change*

\[
\text{SNR}(\phi_0 \rightarrow \phi_1) = \frac{\left( E_{x \sim \phi_0} [\mathcal{L}(x)] - E_{x \sim \phi_1} [\mathcal{L}(x)] \right)^2}{\text{var}_{x \sim \phi_0}[\mathcal{L}(x)] + \text{var}_{x \sim \phi_1}[\mathcal{L}(x)]}
\]

The SNR(\(\phi_0 \rightarrow \phi_1\))

- Measures the extent to which \(\phi_0 \rightarrow \phi_1\) is detectable by monitoring \(E[\mathcal{L}(x)]\)

- If we replace \(E[\cdot]\) and \(\text{var}[\cdot]\) 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 \( \phi_1(x) = \phi_0(Qx + v) \) where \( Q \in \mathbb{R}^{d \times d} \) and orthogonal, \( v \in \mathbb{R}^d \), then

\[
\text{SNR}(\phi_0 \rightarrow \phi_1) < \frac{C}{d}
\]

Where \( C \) is a constant that depends only on \( s\text{KL}(\phi_0, \phi_1) \)
The Detectability Loss: Remarks

Theorem

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

\[
\text{SNR}(\phi_0 \rightarrow \phi_1) < \frac{C}{d}
\]

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

Remarks:

- Changes of a given magnitude, \( s\text{KL}(\phi_0, \phi_1) \), become more difficult to detect when \( d \) increases
- DL does not depend on how \( \phi_0 \) 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 $\phi_1(x) = \phi_0(Qx + \nu)$ where $Q \in \mathbb{R}^{d \times d}$ and orthogonal, $\nu \in \mathbb{R}^d$, then

$$\text{SNR}(\phi_0 \rightarrow \phi_1) < \frac{C}{d}$$

Where $C$ is a constant that depends only on $\text{sKL}(\phi_0, \phi_1)$
The change model \( \phi_1(x) = \phi_0(Qx + v) \) includes:

- Changes in the location of \( \phi_0 \) (i.e., \( +v \))
The change model $\phi_1(x) = \phi_0(Qx + v)$ includes:

- Changes in the location of $\phi_0$ (i.e., $+v$)
- Changes in the correlation of $x$ (i.e., $Qx$)
The change model $\phi_1 (x) = \phi_0 (Qx + v)$ includes:

- Changes in the location of $\phi_0$ (i.e., $+v$)
- Changes in the correlation of $x$ (i.e., $Qx$)

It does not include changes in the scale of $\phi_0$ that can be however detected monitoring $||x||$
The Detectability Loss: The Gaussian Assumption

Theorem

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

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

Where $C$ is a constant that depends only on $\text{sKL}(\phi_0, \phi_1)$.
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 approximations of $\mathcal{L}(\cdot)$ typically used for Gaussian mixtures.
The data

- Two datasets 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 $\phi_0 \rightarrow \phi_1$ such that
  $$\phi_1(x) = \phi_0(Qx + \nu)$$ and $sKL(\phi_0, \phi_1) = 1$
- Test data: two windows $V_0$ and $V_1$ (500 samples each) selected before and after the change.

![Graph showing two windows $V_0$ and $V_1$ selected before and after the change](image)
The change-detectability measure:

- Compute $\mathcal{L}(\hat{\phi}_0(x))$ 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 
  \[ \mathcal{T}(W_0, W_1) \leq h \]
  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:**
- $\phi_1$ 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

- $t$-test on $\hat{L}_u$
- $t$-test on $\hat{L}_l$
- Lepage test on $\hat{L}_u$
- Lepage test on $\hat{L}_l$

Power

$d$

$\begin{array}{c}
\text{Particle} \\
\text{Wine}
\end{array}$
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 $s$

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

and determine whether it is normal or anomalous

Patches $s_c \in \mathbb{R}^p$ are too high-dimensional ($p \gg 0$) for modeling the distribution $\phi_0$ generating normal patches.

We need to extract suitable features to reduce the dimensionality of our anomaly-detection problem.
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 $s_c$ to $\mathcal{D}$
**Sparse representations** have shown to be a very useful method for **constructing signal models**

The underlying assumption is that

\[ s \approx D\alpha \quad \text{i.e.,} \quad \|s - D\alpha\|^2 \approx 0 \]

and \( \alpha \in \mathbb{R}^n \) where:

- \( D \in \mathbb{R}^{p \times n} \) is the **dictionary**, columns are called **atoms**
- the coefficient vector \( x \) is sparse
  - \( \|\alpha\|_0 = L \ll n \) or
  - \( \|\alpha\|_1 \) 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 $s_c$ with $D$ we perform the Sparse coding:

$$\alpha = \arg\min_{\tilde{\alpha} \in \mathbb{R}^n} \| D\tilde{\alpha} - s \|_2^2 + \lambda \| \tilde{\alpha} \|_1, \quad \lambda > 0$$

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

We then measure

$$\| D\alpha - s \|_2^2$$

and

$$\| \alpha \|_1$$

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

Normal patches are expected to yield features $x$ that are i.i.d. and that follow a (unknown) distribution $\phi_0$, 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 $\{x_c, \ c = 1, \ldots\}$ is generated from $\phi_0$ 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 $sKL(\phi_0, \phi_1)$ 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).
We developed data-driven features based on convolutional sparse models

\[ s \approx \sum_{i=1}^{n} d_i \ast \alpha_i , \quad \text{s.t. } \alpha_i \text{ is sparse} \]

where a signal \( s \) is entirely encoded as the sum of \( n \) convolutions between a filter \( d_i \) and a coefficient map \( \alpha_i \).

Pros:

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

Training Image

Learned Filters
If we consider the convolutional sparse coding

$$\{\hat{\alpha}\} = \arg\min_{\{\alpha\}_n} \left\| \sum_{i=1}^{n} d_i \odot \alpha_i - s \right\|_2^2 + \lambda \sum_{i=1}^{n} \|\alpha\|_1$$

we can build the feature vector as:

$$x_c = \left[ \left\| \prod_{c} \left( \sum_{i=1}^{n} d_i \odot \hat{\alpha}_i - s \right) \right\|_2^2 \sum_{i=1}^{n} \left\| \prod_{c} \hat{\alpha} \right\|_1 \right]$$

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

\[
 x_c = \left( \sum_{i=1}^{m} \left( \prod_c d_i \ast \hat{\alpha}_i - s \right) \right)^2
\]

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
Detectability loss occurs:
- independently on the specific statistical tool used to monitor the log-likelihood
- does not depend on how the change affects $\phi_0$, 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 $x$ 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.

Thanks, Questions?


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

Sketch of the proof

**Theorem**

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

\[
\text{SNR}(\phi_0 \to \phi_1) < \frac{C}{d}
\]

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

**Sketch of the proof:** recall

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

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

We now show that

\[ sKL(\phi_0, \phi_1) \geq E_{x \sim \phi_0} [L(x)] - E_{x \sim \phi_1} [L(x)] \quad (*) \]

From \( L(x) = \log(\phi_0(x)) \) and the definition of sKL it follows

\[ sKL(\phi_0, \phi_1) = E_{x \sim \phi_0} [\log(\phi_0(x))] - E_{x \sim \phi_0} [\log(\phi_1(x))] + \]

\[ + E_{x \sim \phi_1} [\log(\phi_1(x))] - E_{x \sim \phi_1} [\log(\phi_0(x))] \]

Thus

\[ (*) \iff E_{x \sim \phi_1} [\log(\phi_1(x))] - E_{x \sim \phi_0} [\log(\phi_1(x))] \geq 0 \]
Sketch of the proof

\[
E_{x \sim \phi_1} \left[ \log(\phi_1(x)) \right] - E_{x \sim \phi_0} \left[ \log(\phi_1(x)) \right] = \\
= \int \log(\phi_1(x))\phi_1(x)dx - \int \log(\phi_1(x))\phi_0(x)dx
\]

We denote

\[
y = Q'(x - v), \quad x = Qy + v
\]

\[
dy = |\det(Q')|dx = dx
\]

\[
\phi_0(x) = \phi_1(Q'(x - v)) = \phi_1(y)
\]

\[
\phi_1(x) = \phi_1(Qy + v) =: \phi_2(y)
\]

then

\[
E_{x \sim \phi_1} \left[ \log(\phi_1(x)) \right] - E_{x \sim \phi_0} \left[ \log(\phi_1(x)) \right] = \\
= \int \log(\phi_1(x))\phi_1(x)dx - \int \log(\phi_2(y))\phi_1(y)dy = \\
= KL(\phi_1, \phi_2) \geq 0
\]
Sketch of the proof

Thus

$$sKL(\phi_0, \phi_1) \geq E_{x \sim \phi_0} [\mathcal{L}(x)] - E_{x \sim \phi_1} [\mathcal{L}(x)]$$

Moreover

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

It follows

$$\text{SNR}(\phi_0 \rightarrow \phi_1) = \frac{\left( E_{x \sim \phi_0} [\mathcal{L}(x)] - E_{x \sim \phi_1} [\mathcal{L}(x)] \right)^2}{\text{var}_{x \sim \phi_0} [\mathcal{L}(x)] + \text{var}_{x \sim \phi_1} [\mathcal{L}(x)]} \leq \frac{sKL(\phi_0, \phi_1)^2}{d/2}$$