EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

NOVELTY DETECTION based on LEARNING ENTROPY

Gejza Dohnal, Ivo Bukovský

Fakulta strojní
Centrum pokročilých leteckých technologií

FAKULTA STROJNÍ ČVUT V PRAZE

Change Point Detection

tries to identify times when the probability distribution of a stochastic process or time series changes. In general the problem concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes.

Novelty Detection

the task of recognising that test data differ in some respect from the data that are available during training.

M.A.F. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, 2013

Anomaly Detection

the problem of finding patterns in data that do not conform to expected behavior.

V. Chandola, A. Banerjee, V. Kumar, University of Minnesota, 2009

Concept Drift Detection

Concept drift is an important problem in the context of machine learning and data mining. It can be described as a change in the fundamental concepts underlying the data, or, in its most basic form, as a significant change in the distribution of the data.

A. Dries, Dept. of Computer Science, Katholieke Universiteit Leuven, U. Rückert, International Computer Science Institute, Berkeley, 2009

Classical „Change-point problem“:

we investigate a random process $X(t)$, $t \geqq 0$ with the following assumptions:

there exists an unknown time $T > 0$ such that:
for $0 \leqq t \leqq T$ the process $X(t)$, is stationary, driven by some (usualy known) probability measure P^0,
for $T \leqq t$ the process $X(t)$, is driven by some different probability measure P^1.

A plenty of work has been done in this research field and it is impossible to give an exhaustive overview. The Zentralblatt MATH returns almost 3000 hits, when entering the word ”change-point”

Page, E. (1954). Continuous inspection schemes. Biometrika 41, 100–115.

Shiryaev, A. (1978). Optimal Stopping Rules. Springer, New York.

Bhattacharya, P. (1994). Some aspects of change-point analysis. In Change-point Problems (eds. E. Carlstein, H. Mu¨ller & D. Siegmund), 28–56, Institute of Mathematical Statistics, Hayward.

Zacks, S. (1982). Classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures. Stat. Anal. Donnees 7, 48–81.

The d-dimensional random process $\boldsymbol{X}(t)$, $t \geqq 0$ is $\boldsymbol{h}$-stationary, if there exists a d-dimensional real function $\boldsymbol{h}(t)$ and a d-dimensional centered weak stationary random process $\xi(t)$ such that $\boldsymbol{X}(t) = \boldsymbol{h}(t) + \xi(t)$, $t \geqq 0$.

Consider the following situation:

Novelty Detection

there exists an unknown time $T > 0$ such that:
for $0 \leqq t \leqq T$ the process $X(t)$, is $\boldsymbol{h}$-stationary with some (known or unknown) function $\boldsymbol{h}(t)$ and a stationary random process $\xi(t)$,
for $T \leqq t$ the process $X(t)$, is $\boldsymbol{g}$-stationary with some different (usualy unknown) function $\boldsymbol{g}(t)$ and some random process $\psi(t)$.

$\boldsymbol{h}(t)$ is known => $\boldsymbol{Y}(t) = \boldsymbol{X}(t) - \boldsymbol{h}(t)$ allows use of classical methods

$\boldsymbol{h}(t)$ is unknown => $\boldsymbol{X}(t)$ is non-stationary, we can't use classical methods

To detect the change, we need some learning method to know $\boldsymbol{h}(t)$.

Applied incrementally learning system:

The predictor $\tilde{y}(t+\delta) = f(\mathbf{x}(t), \mathbf{w}(t))$ is a a function $f(\mathbf{x}, \mathbf{w})$ such that for any time $t \geqq 0$ and a given horizon δ minimizes the error

$$|X(t+\delta) - f(\mathbf{x}(t), \mathbf{w}(t))|.$$

As a predictor we use neural network based on HONU (High Order Neural Units)

LNU: $\tilde{y}(t) = \mathbf{x}(t)\mathbf{w}(t)$ (linear)

QNU: $\tilde{y}(t) = \mathbf{x}^{\mathrm{T}}(t)\mathbf{W}(t)\mathbf{x}(t)$ (quadratic)

$\mathbf{w}(t)$ is n-dimensional vector of weights (adaptive parameters):

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \Delta\mathbf{w}(t), \quad \text{where} \quad \triangle\mathbf{w}(t) = -\frac{\mu}{2}\frac{\partial e(t)^2}{\partial \mathbf{w}}$$

static gradient descent algorithm

LNU: $\triangle\mathbf{w}(t) = \mu e(t)\mathbf{x}(t)$

QNU: $\triangle\mathbf{W}(t) = \mu e(t)\mathbf{x}(t)\mathbf{x}(t)^T$

$\{\boldsymbol{X}(t),\ t \geqq 0\}$ is $\boldsymbol{h}$-stationary => $\{\Delta \mathbf{w}(t)\}$ tends to stationary process

$\{\boldsymbol{X}(t),\ t \geqq \mathrm{T}\}$ is no more $\boldsymbol{h}$-stationary => $\{\Delta \mathbf{w}(t)\}$ loses its stationarity

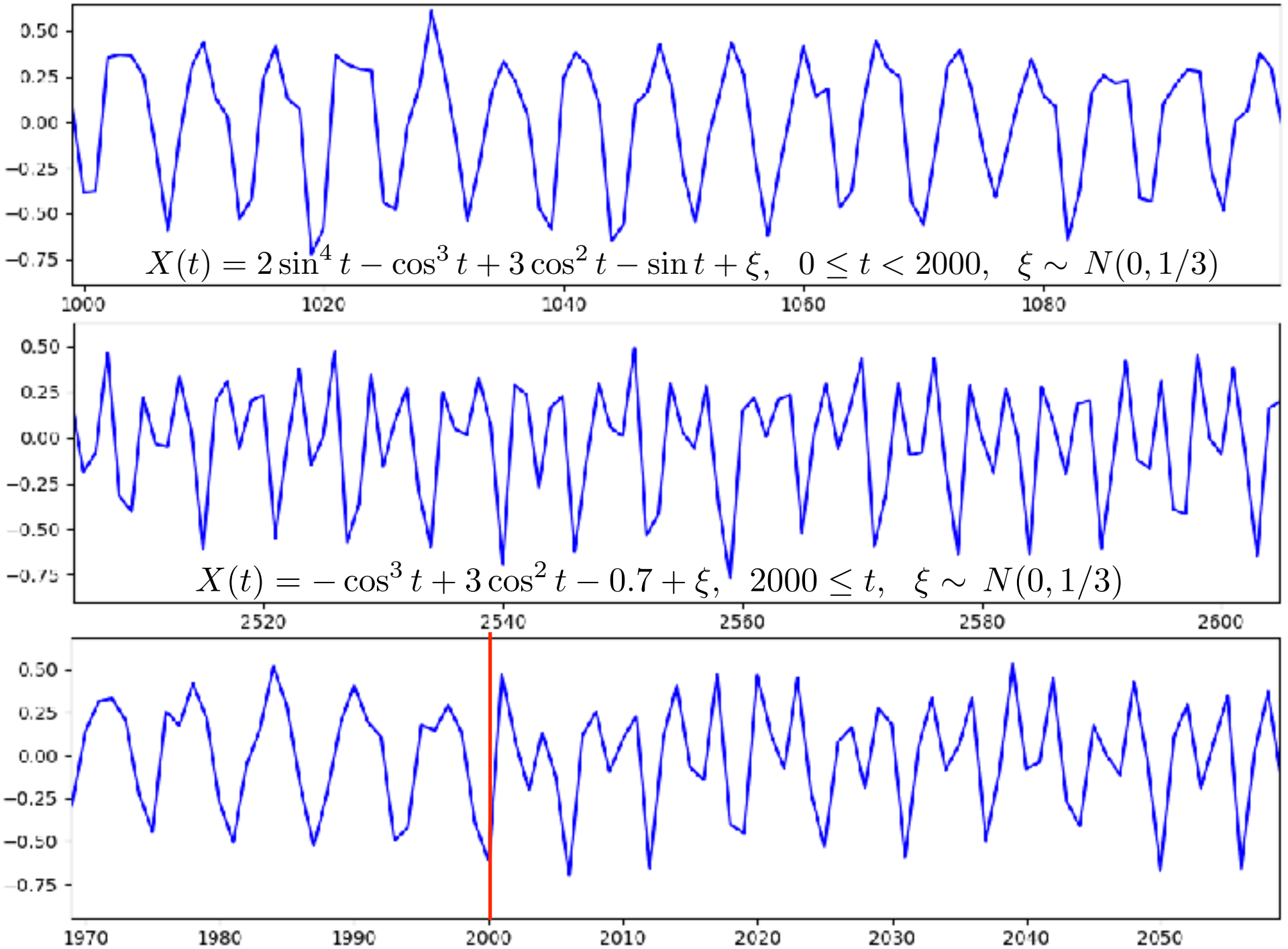
=> the change can be detected by classical detection methods applied to the process $\{\Delta \mathbf{w}(t)\}$

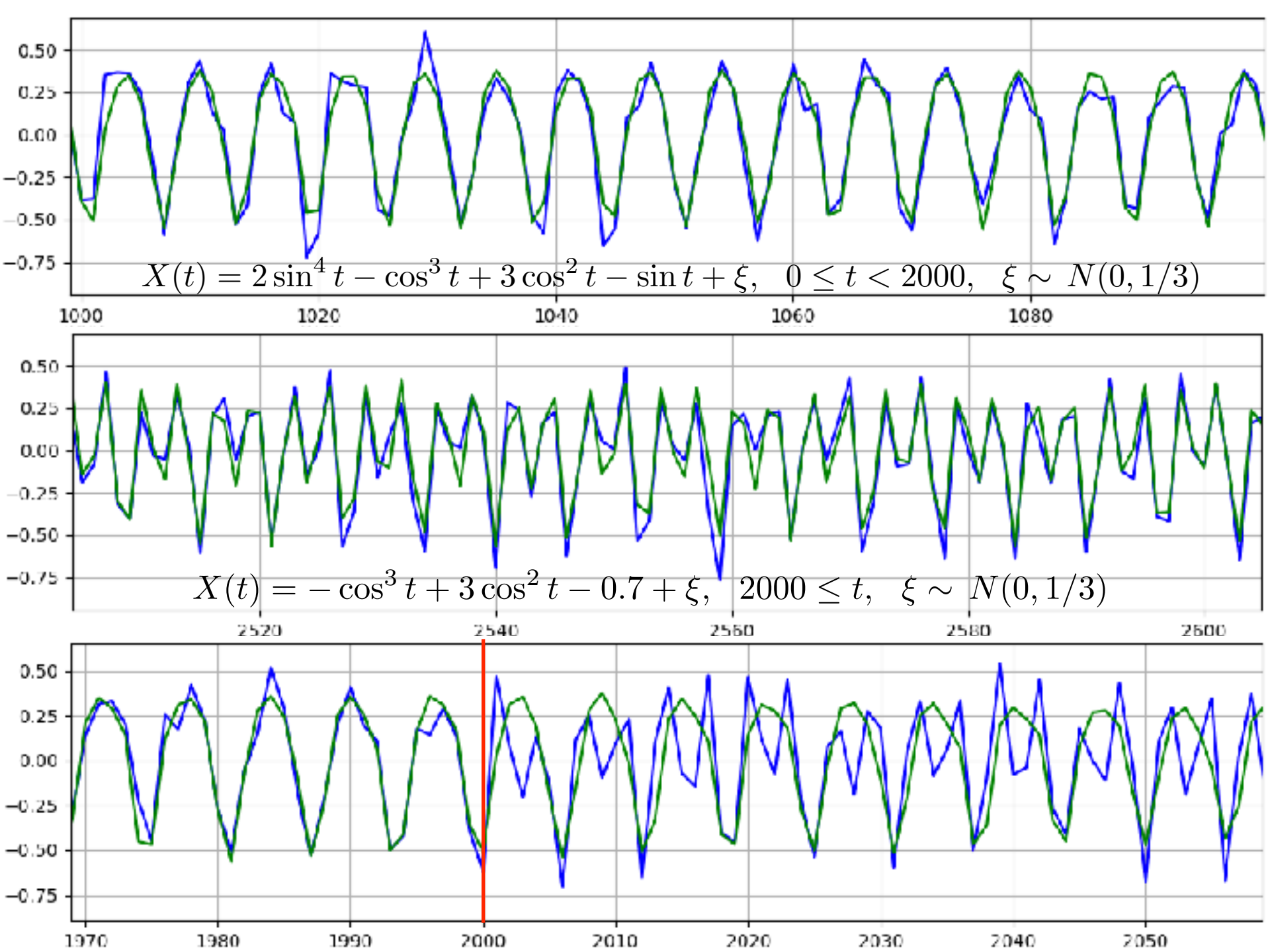
For illustration, we use data generated by the process

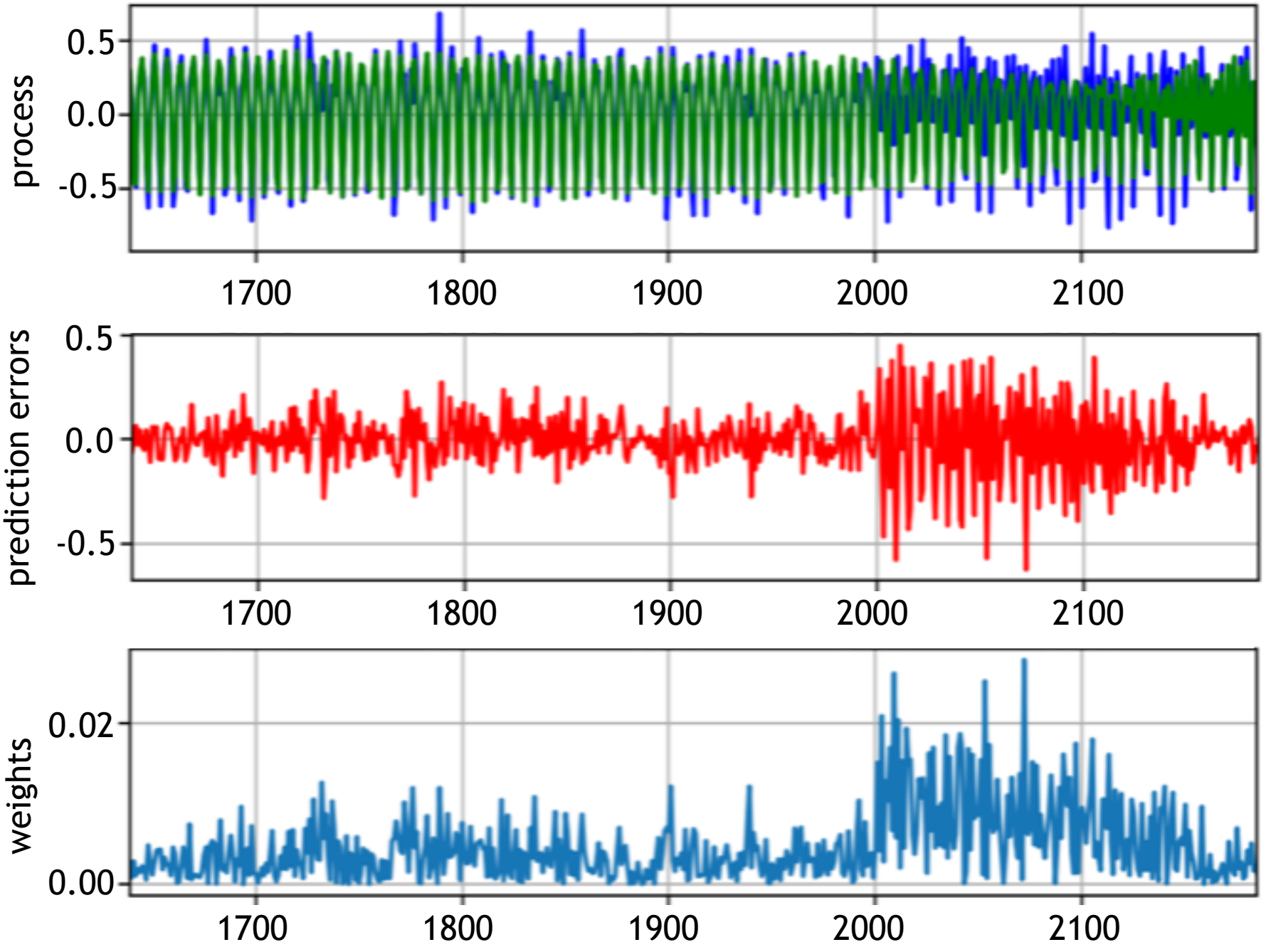
$$X(t) = a\sin^4 t + b\cos^3 t + c\cos^2 t + d\sin t + e + \xi$$

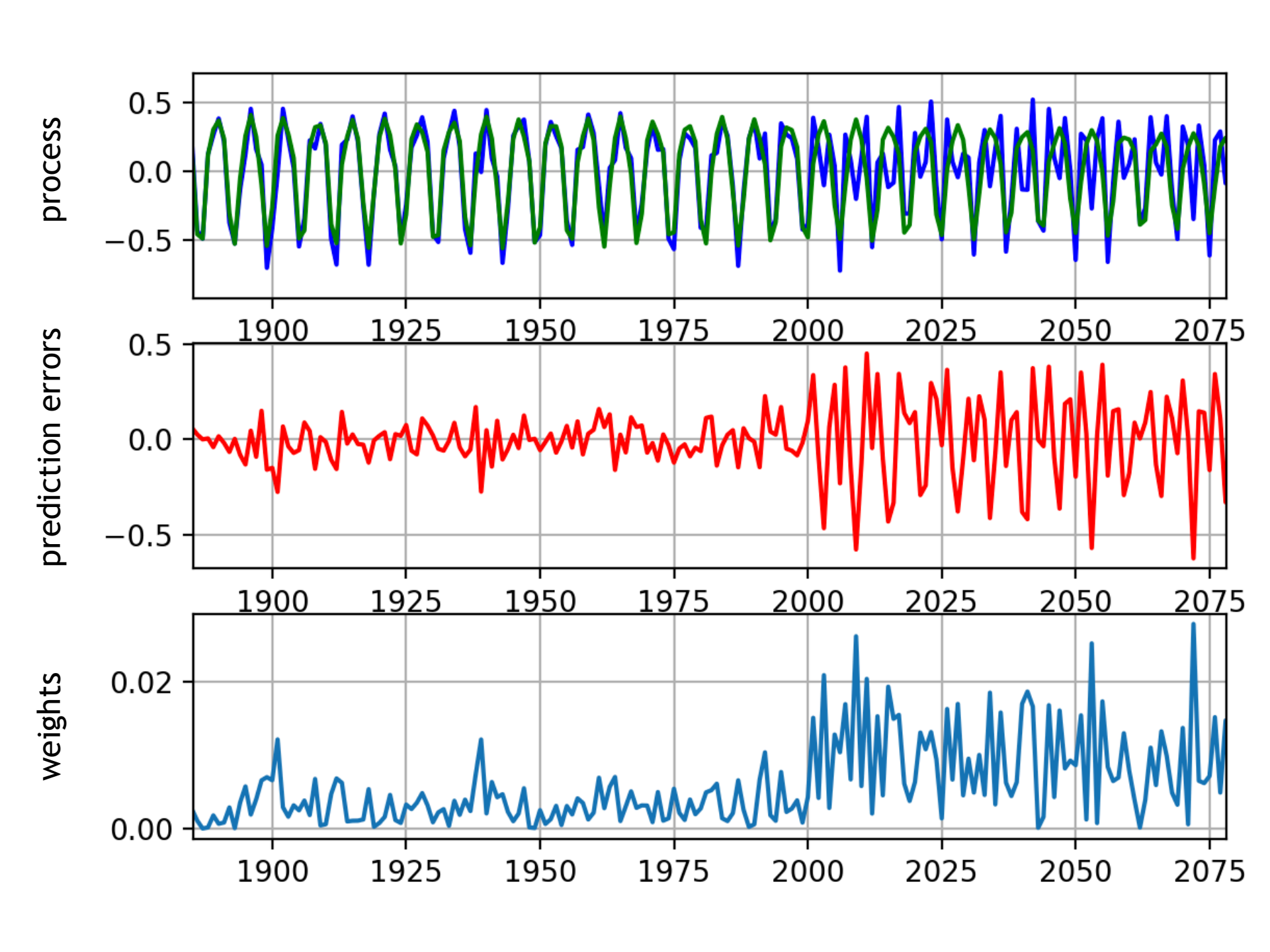
with parameters $a = 2,\ b = -1,\ c = 3,\ d = -1,\ e = 0$
for $t \geqq 0$ and random fluctuations $\xi \backsim N(0, 1/3)$.

The change is simulated in the time t = 2000 from which longer is a = 0, d = 0 and e = -0.7.


$X(t) = 2\sin^4 t - \cos^3 t + 3\cos^2 t - \sin t + \xi, \quad 0 \le t < 2000, \quad \xi \sim N(0, 1/3)$
$X(t) = -\cos^3 t + 3\cos^2 t - 0.7 + \xi, \quad 2000 \le t, \quad \xi \sim N(0, 1/3)$
0.50
0.25
0.00
−0.25
−0.50
−0.75
1000
1020
1040
1060
1080
2520
2540
2560
2580
2600
1970
1980
1990
2000
2010
2020
2030
2040
2050


$X(t) = 2\sin^4 t - \cos^3 t + 3\cos^2 t - \sin t + \xi, \quad 0 \le t < 2000, \quad \xi \sim N(0, 1/3)$
$X(t) = -\cos^3 t + 3\cos^2 t - 0.7 + \xi, \quad 2000 \le t, \quad \xi \sim N(0, 1/3)$
0.50
0.25
0.00
−0.25
−0.50
−0.75
1000
1020
1040
1060
1080
2520
2540
2560
2580
2600
1970
1980
1990
2000
2010
2020
2030
2040
2050


process
0.5
0.0
-0.5
1700
1800
1900
2000
2100
prediction errors
0.5
0.0
-0.5
1700
1800
1900
2000
2100
weights
0.02
0.00
1700
1800
1900
2000
2100


process
0.5
0.0
−0.5
1900
1925
1950
1975
2000
2025
2050
2075
prediction errors
0.5
0.0
−0.5
1900
1925
1950
1975
2000
2025
2050
2075
weights
0.02
0.00
1900
1925
1950
1975
2000
2025
2050
2075

Shanon-based Entropies are data-window and probabilistic based computations that are widely used for time series analyses.

- **Sample Entropy** is a signal complexity evaluation algorithm (floating window based quantification of signal complexity, probability-based approach).

- **Entropy Learning** is a Shannon-inspired neural network learning algorithm based on minimizing complexity (entropy) of neural weights in a network.

- **Learning Entropy** is a non-Shannon based novelty detection algorithm based on observation of unusual learning effort of incrementally learning systems. LE is a relative measure of novelty (information) recognized as unusual learning effort of pre-trained learning system on individual data samples.

$$\mathbf{x}^M(l) = \{\mathbf{x}(l\text{-}M), \ldots, \mathbf{x}(l\text{-}1), \mathbf{x}(l)\} \Rightarrow \{\mathbf{w}(l\text{-}M), \ldots, \mathbf{w}(l\text{-}1), \mathbf{w}(l)\}$$

$$\Delta\mathbf{w}(k) = \mathbf{w}(k+1) - \mathbf{w}(k) \Rightarrow \{\Delta\mathbf{w}(l\text{-}M), \ldots, \Delta\mathbf{w}(l\text{-}1)\}$$

$$|\triangle\bar{\mathbf{w}}(l)| = \frac{1}{M}\sum_{i=1}^{M}|\triangle\mathbf{w}(l-i)|$$

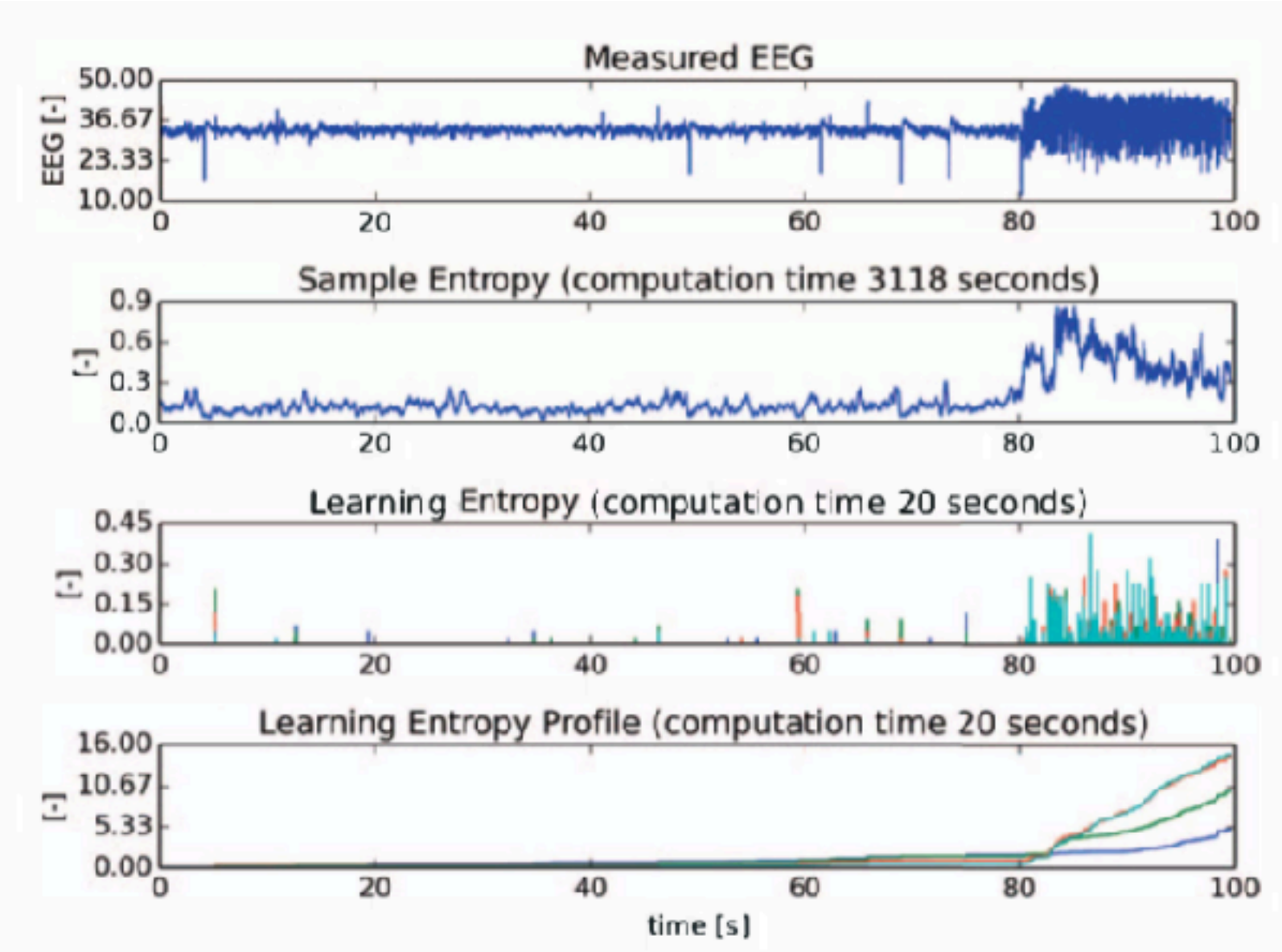
$$E_A(k) = \frac{1}{n_A n_w}\sum_{j=1}^{n_A}\sum_{i=1}^{n_w} I_{(|\triangle w_i(k)|>\alpha_j|\triangle\bar{w}_i(k)|)}$$

Learning Entropy

where $A = (\alpha_1, \ldots, \alpha_{n_A})$ is n_A-dimensional vector of detection senzitivities.

- We use $\Delta\mathbf{w}(t)$ for detection rather than prediction error due to its higher sensitivity.

Onset Detection od Epileptic Seizures in EEG Time Series



Ivo Bukovsky, Matous Cejnek, Jan Vrba, Noriyasu Homma, 2016 Int. Joint Conf. on Neural Networks

$$\mathcal{X}^M(l) = \{\mathbf{x}(l\text{-}M), \ldots, \mathbf{x}(l\text{-}1), \mathbf{x}(l)\} \Rightarrow \{\mathbf{w}(l\text{-}M), \ldots, \mathbf{w}(l\text{-}1), \mathbf{w}(l)\}$$

$$\Delta\mathbf{w}(k) = \mathbf{w}(k+1) - \mathbf{w}(k) \Rightarrow \{\Delta\mathbf{w}(l\text{-}M), \ldots, \Delta\mathbf{w}(l\text{-}1)\}$$

$$|\triangle\bar{\mathbf{w}}(l)| = \frac{1}{M}\sum_{i=1}^{M}|\triangle\mathbf{w}(l-i)|$$

$$\Omega_\alpha(k) = \{\, \mathbf{v} \in \mathbb{R}^n : \ (\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|)^T\Sigma^{-1}(\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|) \leq \omega_\alpha^2 \,\}$$

Hotelling's t-square test:

1) In the Phase I: Observe $\{\mathbf{x}(l\text{-}M), \ldots, \mathbf{x}(l\text{-}1), \mathbf{x}(l)\}$ and compute $\{\mathbf{w}(l\text{-}M), \ldots, \mathbf{w}(l\text{-}1), \mathbf{w}(l)\}$, evaluate the centroid $|\Delta\mathbf{w}(l)|$ and sample correlation matrix $\boldsymbol{\Sigma}$

2) For $k = l+1, \ldots$ evaluate the predictor $y(k) = f(\mathbf{x}(k\text{-}1), \mathbf{w}(k\text{-}1))$ and compute the Hotelling's t^2-statistics

$$t^2 = (|\triangle\bar{\mathbf{w}}(l)| - \triangle\mathbf{w}(k))^T\Sigma^{-1}(|\triangle\bar{\mathbf{w}}(l)| - \triangle\mathbf{w}(k))$$

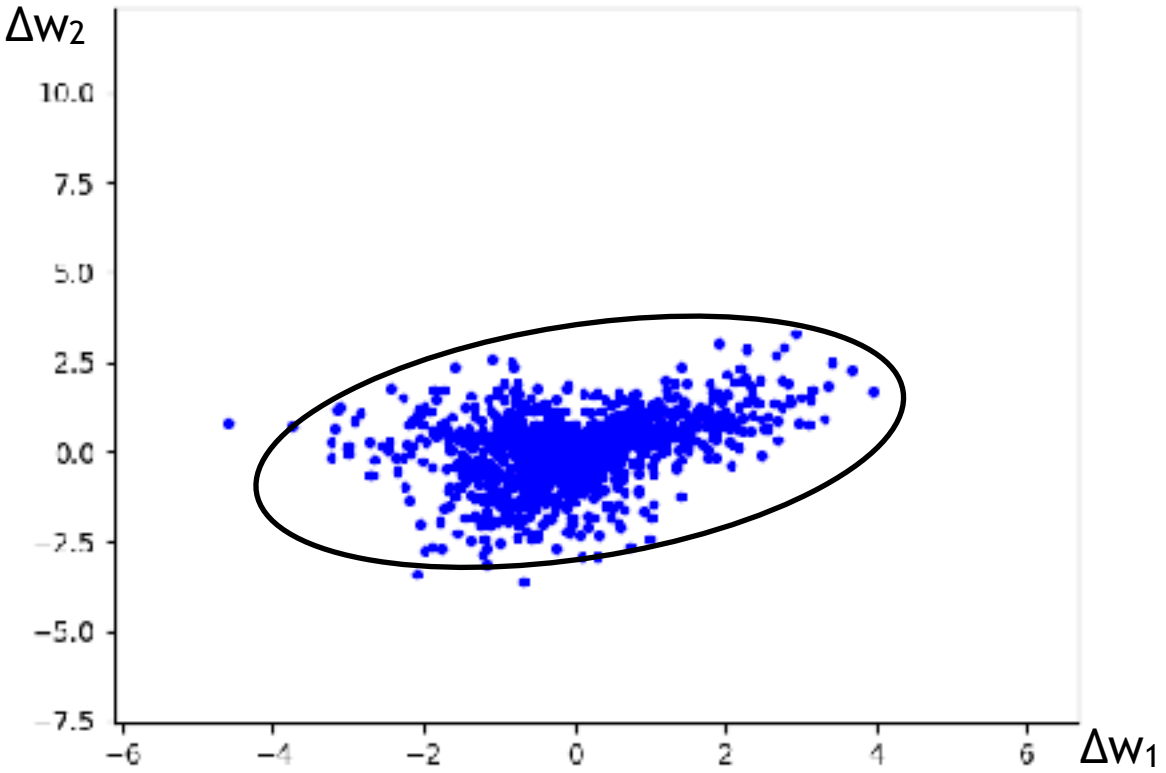
3)
$$t^2 \sim T^2_{n,M-1} = \frac{n(M-1)}{M-n}F_{n,M-n}$$

$$\mathcal{X}^M(l) = \{\mathbf{x}(l\text{-}M), \ldots, \mathbf{x}(l\text{-}1), \mathbf{x}(l)\} \Rightarrow \{\mathbf{w}(l\text{-}M), \ldots, \mathbf{w}(l\text{-}1), \mathbf{w}(l)\}$$

$$\Delta\mathbf{w}(k) = \mathbf{w}(k+1) - \mathbf{w}(k) \Rightarrow \{\Delta\mathbf{w}(l\text{-}M), \ldots, \Delta\mathbf{w}(l\text{-}1)\}$$

$$|\triangle\bar{\mathbf{w}}(l)| = \frac{1}{M}\sum_{i=1}^{M}|\triangle\mathbf{w}(l-i)|$$

$$\Omega_\alpha(k) = \{\, \mathbf{v} \in \mathbb{R}^n : \; (\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|)^T\Sigma^{-1}(\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|) \leq \omega_\alpha^2 \,\}$$

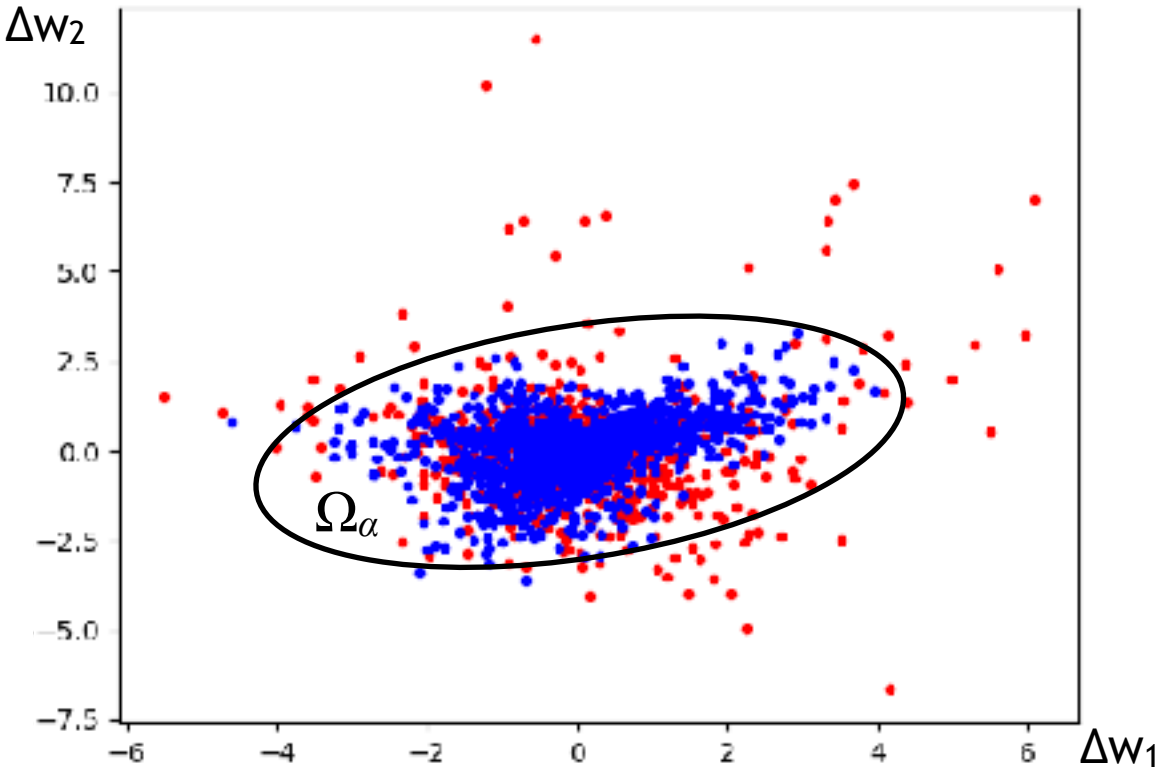


$$\mathbf{x}^M(l) = \{\mathbf{x}(l\text{-}M), \ldots, \mathbf{x}(l\text{-}1), \mathbf{x}(l)\} \Rightarrow \{\mathbf{w}(l\text{-}M), \ldots, \mathbf{w}(l\text{-}1), \mathbf{w}(l)\}$$

$$\Delta\mathbf{w}(k) = \mathbf{w}(k+1) - \mathbf{w}(k) \Rightarrow \{\Delta\mathbf{w}(l\text{-}M), \ldots, \Delta\mathbf{w}(l\text{-}1)\}$$

$$|\triangle\bar{\mathbf{w}}(l)| = \frac{1}{M}\sum_{i=1}^{M}|\triangle\mathbf{w}(l-i)|$$

$$\Omega_\alpha(k) = \{\, \mathbf{v} \in \mathbb{R}^n : \ (\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|)^T\Sigma^{-1}(\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|) \leq \omega_\alpha^2 \,\}$$

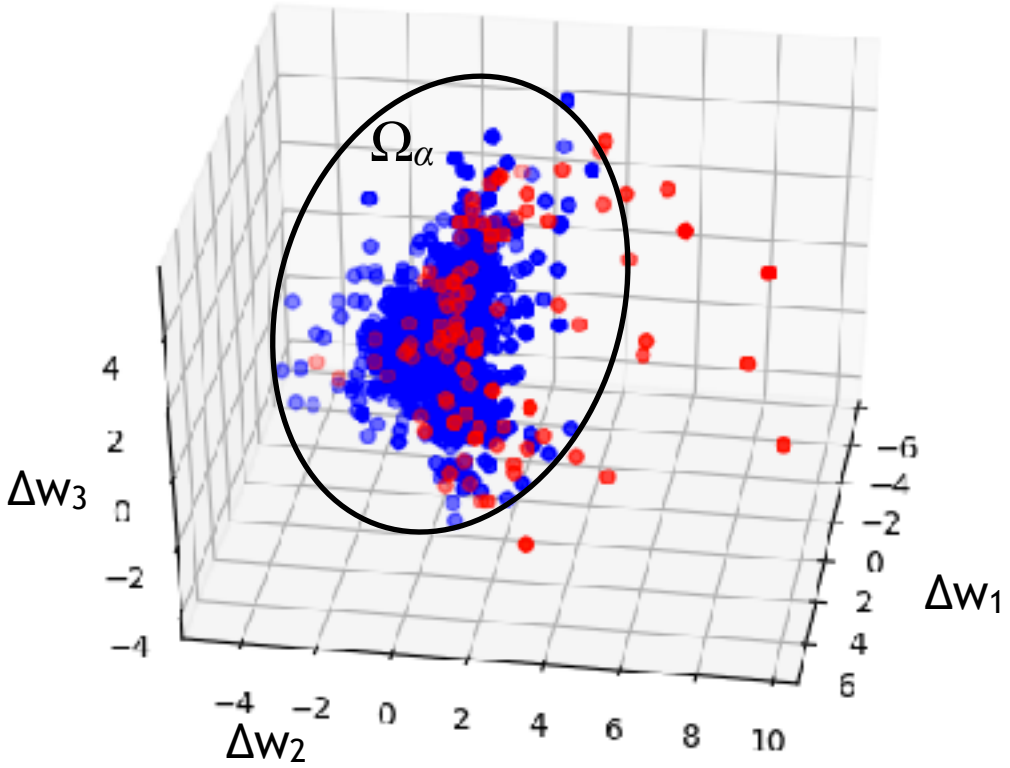


$$\mathcal{X}^M(l) = \{\mathbf{x}(l\text{-}M), \ldots, \mathbf{x}(l\text{-}1), \mathbf{x}(l)\} \Rightarrow \{\mathbf{w}(l\text{-}M), \ldots, \mathbf{w}(l\text{-}1), \mathbf{w}(l)\}$$

$$\Delta\mathbf{w}(k) = \mathbf{w}(k+1) - \mathbf{w}(k) \Rightarrow \{\Delta\mathbf{w}(l\text{-}M), \ldots, \Delta\mathbf{w}(l\text{-}1)\}$$

$$|\triangle\bar{\mathbf{w}}(l)| = \frac{1}{M}\sum_{i=1}^{M}|\triangle\mathbf{w}(l-i)|$$

$$\Omega_\alpha(k) = \{\, \mathbf{v} \in \mathbb{R}^n : \; (\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|)^T\Sigma^{-1}(\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|) \leq \omega_\alpha^2 \,\}$$



$$\mathcal{X}^M(l) = \{\mathbf{x}(l\text{-}M), \ldots, \mathbf{x}(l\text{-}1), \mathbf{x}(l)\} \Rightarrow \{\mathbf{w}(l\text{-}M), \ldots, \mathbf{w}(l\text{-}1), \mathbf{w}(l)\}$$

$$\Delta\mathbf{w}(k) = \mathbf{w}(k+1) - \mathbf{w}(k) \Rightarrow \{\Delta\mathbf{w}(l\text{-}M), \ldots, \Delta\mathbf{w}(l\text{-}1)\}$$

$$|\triangle\bar{\mathbf{w}}(l)| = \frac{1}{M}\sum_{i=1}^{M}|\triangle\mathbf{w}(l-i)|$$

$$\Omega_\alpha(k) = \{\, \mathbf{v} \in \mathbb{R}^n : \ (\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|)^T\Sigma^{-1}(\mathbf{v} - |\triangle\bar{\mathbf{w}}(k)|) \leq \omega_\alpha^2 \,\}$$

Algorithm:

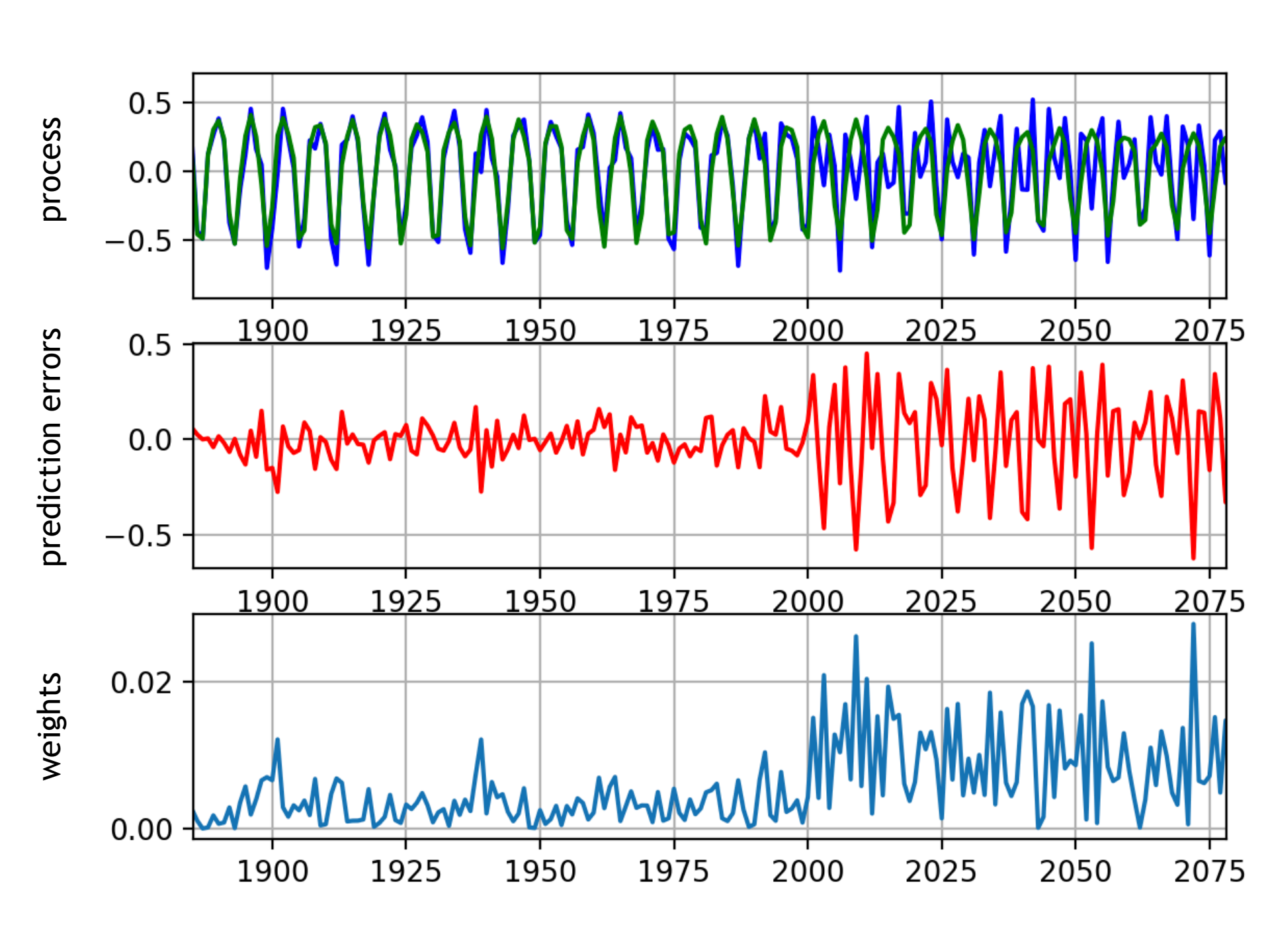
1) Find $\Omega_\alpha(l)$, i.e., evaluate the centroid $|\Delta\bar{\mathbf{w}}(l)|$ and sample correlation matrix $\boldsymbol{\Sigma}$ on $\mathcal{X}(l)$

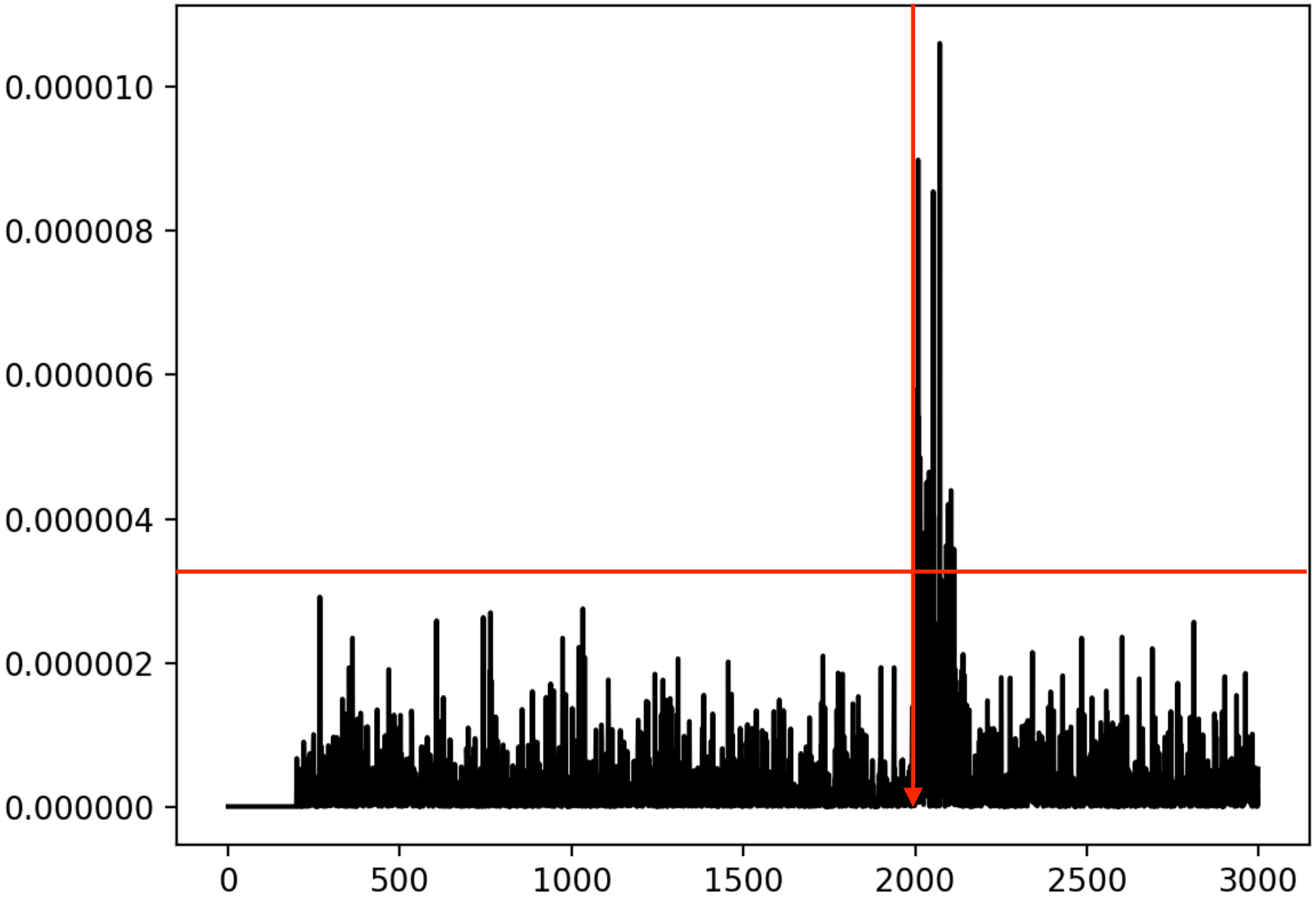
2) For $k = l+1,\ldots$ evaluate the predictor $y(k) = f(\mathbf{x}(k\text{-}1), \mathbf{w}(k\text{-}1))$

3) If $|\Delta\mathbf{w}(k)| \notin \Omega_\alpha(l)$, detect a change.

4) Update $\Omega_\alpha(l)$ for $l = k$.

5) Continue in step 2).


process
0.5
0.0
−0.5
1900
1925
1950
1975
2000
2025
2050
2075
prediction errors
0.5
0.0
−0.5
1900
1925
1950
1975
2000
2025
2050
2075
weights
0.02
0.00
1900
1925
1950
1975
2000
2025
2050
2075


0.000010
0.000008
0.000006
0.000004
0.000002
0.000000
0
500
1000
1500
2000
2500
3000

Summary:

- The novelty detection algorithm for nonstationary process based on observation of unusual learning effort (Learning Entropy) of incrementally learning systems is proposed.
- As a predictor we use neural network based on HONU.
- We use $\Delta \mathbf{w}(t)$ for detection rather than prediction error due to higher sensitivity.
- The novelty is detected using classical methods (e.g. the Hotelling t^2) for change detection applied to the process $\{\Delta \mathbf{w}(t)\}$.
- Using sliding window, we adapt the confidence d-dimensional elipsoid $\Omega_\alpha(t)$
- If the vector of weight increments $\Delta \mathbf{w}(t+1)$ corresponding to next observation of monitored process doesn't belong to $\Omega_\alpha(t)$, a signal is emitted.

Thanks for your attention