Streaming multiscale anomaly detection

DATA-ENS Paris and ThalesAlenia Space

B Ravi Kiran,
Université Lille 3, CRISTaL
Joint work with Mathieu Andreux

beedotkiran@gmail.com

June 20, 2017

Overview

1. Anomaly detection
2. Time series representation
3. Streaming Subspace Tracking
4. Results

Motivation : Anomaly detection

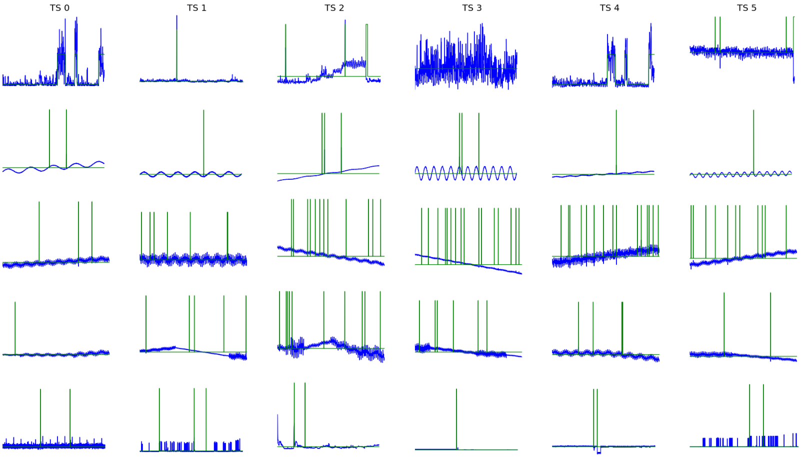
- Areas : Industrial processes, medical and satellite telemetry, Finance
- Anomaly Detection : $x(t), t$ where signal "deviates" from the local mean value.
- $x(t)$ are observations over time t where new data arrives over time t.
- High volumes of data are generated per day (in the GBs).

Requirements :

- Time series representation is robust to variation in scale of pseudo-periodicities (window size).
- Streaming time series anomaly detection to handle large abouts of data.

Require an online multiscale anomaly detection algorithm.

Yahoo! and Numenta Datasets



- Yahoo! unsupervised anomaly detection Benchmark [8] provides datasets with annotated anomalies and changepoints.
- Numenta Anomaly Benchmark (NAB) [9] provides an evaluation of streaming time series anomaly detection algorithms.
- The datasets contain various types of anomalies : level shifts/change-points, point anomalies, change in periodicities, value drifts, change in envelopes, linear trends.

Anomaly detection problem

- Formulation :
  - Track the principal direction given a scale/lag p for the design matrix of time series.
  - Evaluate the reconstruction error to measure deviation from the rest of the windows.
  - Evaluate across multiple lags (p)
- Characterize anomalies by their variation in reconstruction error across scale of lag-window size.

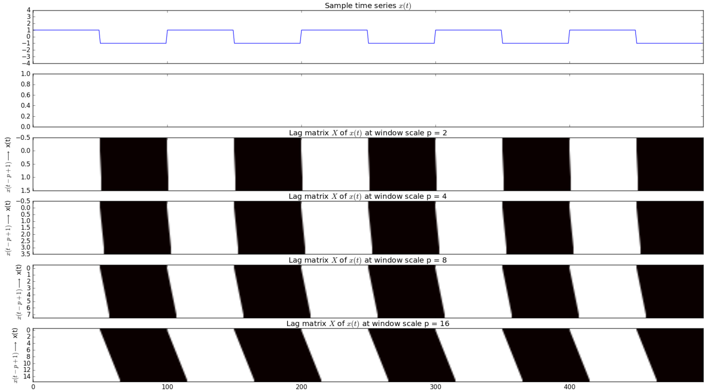
Related work :

- Streaming anomaly detection by subspace tracking [5]
- Tracking correlations over multi-scale windows for frequent motif extraction [12]
- Multi-scale anomaly detection offline [3]

Time series Embedding

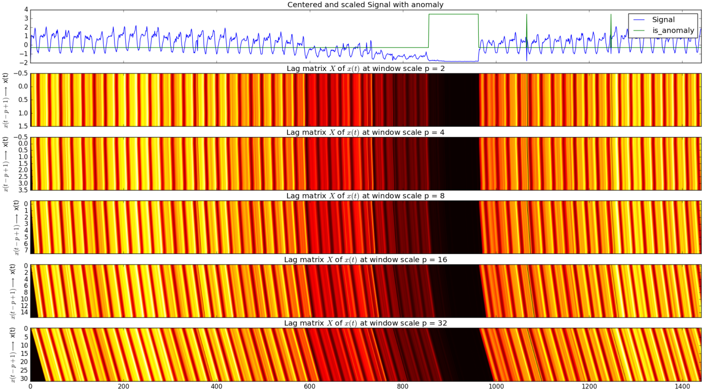
- We build a lag matrix over a window of size p

$$X_t^p = [x_t, x_{t-1}, \ldots, x_{t-p+1}]^T \in \mathbb{R}^p$$



Multiscale Lagmatrix

$$X_t^p = [x_t, x_{t-1}, \ldots, x_{t-p+1}]^T \in \mathbb{R}^p$$



Principal Subspace Tracking

Streaming PCA :

- Dimensionality reduction for time series lag embedding
- Recursive update for principal subspace

Linear Principal Component Analysis criterion :

$$J(\mathbf{w}_t) = E\left[\|X_t - \mathbf{w}_t\mathbf{w}_t^T X_t\|^2\right]$$

$\mathbf{w}_t \in \mathbb{R}^{p\times r}$ At the global minimum for $\mathbf{w}_t$ shall contain the r dominant eigen-vectors.

- Online principal subspace tracking of the lagmatrix to track correlations : SPIRIT algorithm [11]
- Given $X^p \in \mathbb{R}^{T\times p}$, $\mathbf{w}_p$ is defined as the 1-D projection capturing most of the energy of the data samples :

$$\mathbf{w}_p = \arg\min_{\|\mathbf{w}\|=1} \sum_{t=1}^{T} \|X_t^p - (\mathbf{w}_p\mathbf{w}_p^T)X_t^p\|^2$$

Streaming PCA

Algorithm 1 Streaming PCA

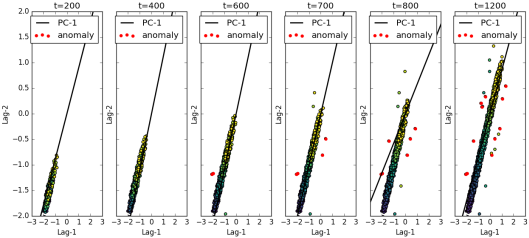
Initialization: $\mathbf{w}_j \leftarrow \mathbf{0}$, $\sigma_j^2 \leftarrow \epsilon$ with $\epsilon \ll 1$
for $t = 1, \ldots, T$ **do**
  for $j = 1, \ldots, J$ **do**
    $Z_t^j \leftarrow H_{2^j}^T X_t^j$
    $y_t^j \leftarrow \mathbf{w}_j^T Z_t^j$
    $\sigma_j^2 \leftarrow \sigma_j^2 + (y_t^j)^2$
    $\mathbf{e}_t^j \leftarrow Z_t^j - y_t^j \mathbf{w}_j$
    $\mathbf{w}_j \leftarrow \mathbf{w}_j + \sigma_j^{-2} y_t^j \mathbf{e}_t^j$
    $\pi_t^j \leftarrow \mathbf{w}_j^T Z_t^j$
    $\widetilde{Z}_t^j \leftarrow \pi_t^j \mathbf{w}_j$
    $\alpha_t^j \leftarrow \|\widetilde{Z}_t^j - Z_t^j\|^2$
  end for
end for
return $\boldsymbol{\alpha} \in \mathbb{R}^{T \times J}$

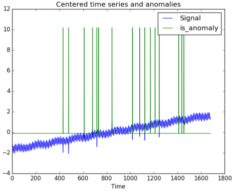
Given $x(t) \in \mathbb{R}$ for $t = 1 : T$

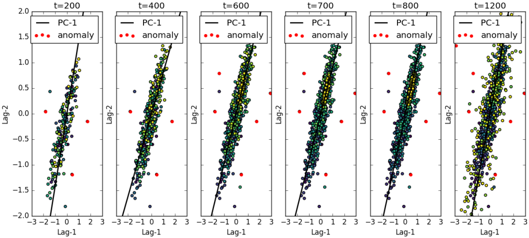
- We evaluate the lag-matrix $X \in \mathbb{R}^{T \times p}$ where $p = 2^j$.
- For each vector $X_t \in \mathbb{R}^p$ we perform a change of basis $Z_t := \Phi^T X_t$
- We require a unitary transform to
  - localize a deviation from the local mean and variations.
  - Preserve the variance.
- Haar transform $\Phi = H$:

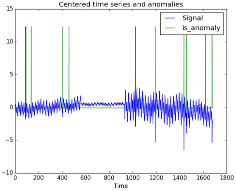
$$H_{2N} = \frac{1}{\sqrt{2}} \begin{bmatrix} H_N \otimes [1, 1] \\ I_N \otimes [1, -1] \end{bmatrix}$$

Streaming PCA point cloud (2d embedding only)



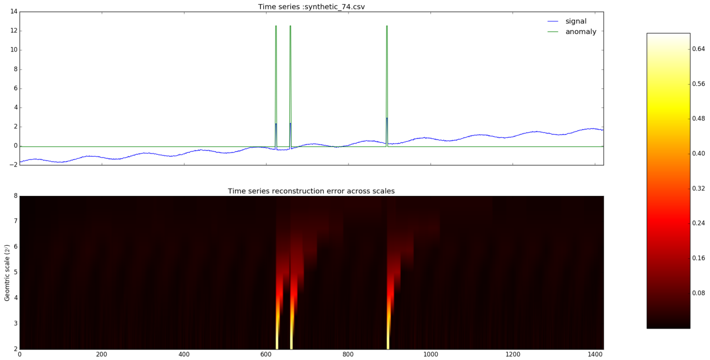






Error Spectrogram

Reconstruction error of the lag-matrix calculated in logarithmic scales :



Hierarchical Approximation for multi-scale windows [12]

- When passing from one scale p_j to the next $p_{j+1} = 2p_j$, instead of rebuilding a lag matrix X_t^{j+1} whose size doubles, it builds a reduced lag matrix Z_t^{j+1} by considering the projection of each component of size p_j on the principal direction obtained at this scale, *i.e.* $Z_t^{j+1} = [\mathbf{w}_j^T Z_t^j, \mathbf{w}_j^T Z_{t-2^j}^j]^T$ with $Z_t^1 = X_t^1$.
- The principal direction at scale p_{j+1} is then obtained by applying the streaming PCA algorithm on this reduced representation.

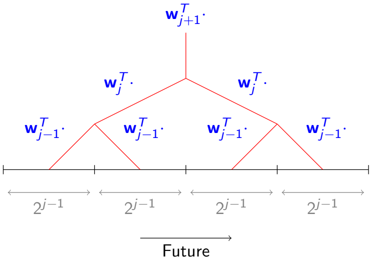


Figure : Hierarchical PCA.

Aggregating Multi-scale Anomaly score

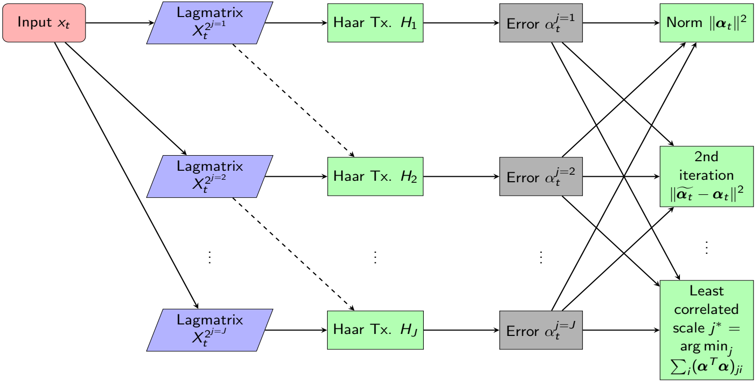
At time t, we denote by $\widetilde{X}_t^p$ the projection of X_t^p upon $\mathbf{w}^p$ (at this time step), *i.e.* $\widetilde{X}_t^p = \mathbf{w}_p^T X_t^p$. We obtain $\boldsymbol{\alpha}_t \in \mathbb{R}^{T \times J}$, we propose the following ways to aggregated the J scales :

1. $\|\boldsymbol{\alpha}_t\|^2$: Norm of multiscale anomaly score
2. $\|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2$: Streaming reconstruction error on anomaly score, obtained *via* a 2nd iteration of the streaming PCA algorithm on the multiscale anomaly score instead of the lag-matrix.
3. $\alpha_t^{j^*}$ where $j^* = \arg\min_j \sum_i (\boldsymbol{\alpha}^T \boldsymbol{\alpha})_{ji}$: the anomaly score corresponding to the scale which is least correlated with others.

Performance Evaluation :

- Area under the receiver operators characteristics curve (AUC)
- integrating the curve of the False positive rate(FPR) *vs* the True positive rate (TPR) obtained for all possible thresholds.
- 0 (worst value) and 1 (perfect detector)

Global-flow



- Representation $\Phi^T X_t$: Localize the anomaly in a basis
- Multiscale Anomaly Score : Compose anomaly scores

Results

Multi-scale score-Norm $\|\boldsymbol{\alpha}_t\|^2$ (PC=1)

Method / AUCs	Bench 1	Bench 2	Bench 3	Bench 4	NAB
fixed-scale	0.828 ± 0.240	0.835 ± 0.180	0.614 ± 0.108	0.568 ± 0.160	0.815 ± 0.238
fixed-scale-haar	0.826 ± 0.238	0.878 ± 0.143	0.617 ± 0.115	0.576 ± 0.157	0.812 ± 0.232
multiscale-lagmatrix	0.884 ± 0.232	0.978 ± 0.057	0.816 ± 0.092	0.696 ± 0.157	0.879 ± 0.199
hierarchical-approx	0.871 ± 0.236	0.997 ± 0.002	0.980 ± 0.025	0.897 ± 0.104	0.900 ± 0.189
multiscale-haar	0.906 ± 0.231	0.989 ± 0.019	0.992 ± 0.019	0.892 ± 0.126	0.892 ± 0.198

PCA on multi-scale score $\|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2$ (PC=1)

Method / AUCs	Bench 1	Bench 2	Bench 3	Bench 4	NAB
fixed-scale	0.632 ± 0.264	0.754 ± 0.206	0.533 ± 0.124	0.525 ± 0.133	0.700 ± 0.247
fixed-scale-haar	0.649 ± 0.251	0.723 ± 0.194	0.514 ± 0.110	0.522 ± 0.129	0.699 ± 0.244
multiscale-lagmatrix	0.895 ± 0.218	0.997 ± 0.006	0.993 ± 0.017	0.959 ± 0.063	0.891 ± 0.194
hierarchical-approx	0.859 ± 0.233	0.997 ± 0.002	0.961 ± 0.071	0.895 ± 0.108	0.884 ± 0.204
multiscale-haar	0.888 ± 0.219	0.988 ± 0.031	0.956 ± 0.059	0.898 ± 0.106	0.886 ± 0.178

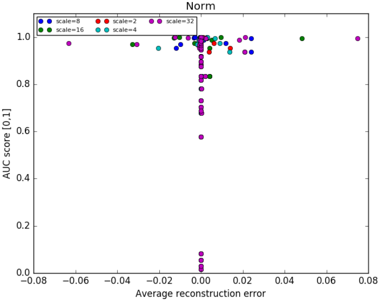
Least correlated scale $\alpha_t^{j^*}$ where $j^* = \arg\min_j \sum_i (\boldsymbol{\alpha}^T\boldsymbol{\alpha})_{ji}$ (PC=1)

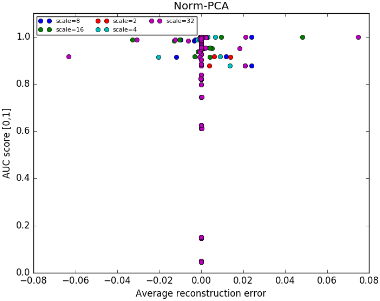
Method / AUCs	Bench 1	Bench 2	Bench 3	Bench 4	NAB
fixed-scale	0.828 ± 0.240	0.835 ± 0.180	0.614 ± 0.108	0.568 ± 0.160	0.815 ± 0.238
fixed-scale-haar	0.826 ± 0.238	0.878 ± 0.143	0.617 ± 0.115	0.576 ± 0.157	0.812 ± 0.232
multiscale-lagmatrix	0.816 ± 0.238	0.773 ± 0.236	0.993 ± 0.017	0.964 ± 0.055	0.885 ± 0.196
hierarchical-approx	0.816 ± 0.238	0.773 ± 0.236	0.993 ± 0.017	0.964 ± 0.055	0.885 ± 0.196
multiscale-haar	0.832 ± 0.238	0.997 ± 0.007	0.799 ± 0.120	0.817 ± 0.123	0.886 ± 0.183

Effect of the Iterated Streaming PCA

- The error here should decorrelate the scores at different scales.
- Plotting Mean Recon. Error (Approximation) Vs. AUC (Detection)

$\|\boldsymbol{\alpha}_t\|^2$ Vs. $\|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2$:

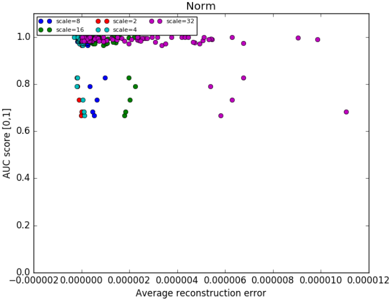


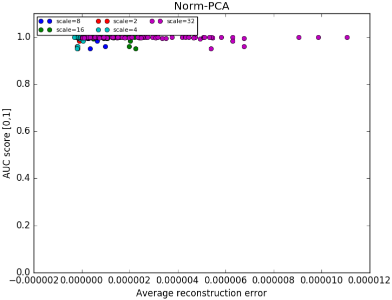


Effect of the Iterated Streaming PCA

- The error here should decorrelate the scores at different scales.
- Plotting Mean Recon. Error (Approximation) Vs. AUC (Detection)

$$\|\boldsymbol{\alpha}_t\|^2 \quad \text{Vs.} \quad \|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2:$$

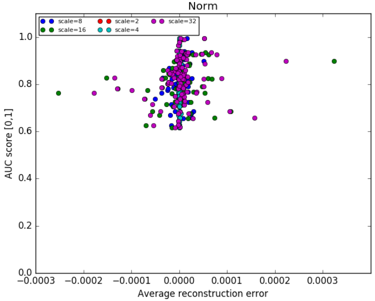


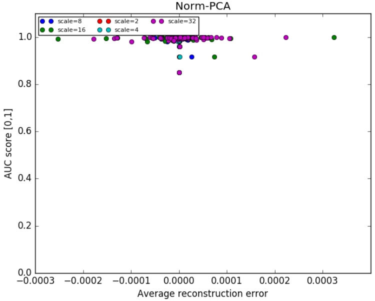


Effect of the Iterated Streaming PCA

- The error here should decorrelate the scores at different scales.
- Plotting Mean Recon. Error (Approximation) Vs. AUC (Detection)

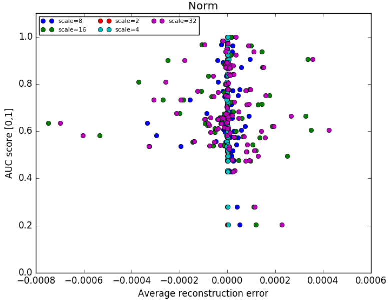
$$\|\boldsymbol{\alpha}_t\|^2 \quad \text{Vs.} \quad \|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2:$$

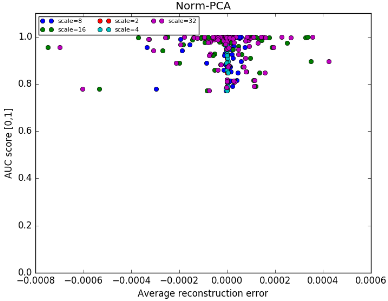




- The error here should decorrelate the scores at different scales.
- Plotting Mean Recon. Error (Approximation) Vs. AUC (Detection)

$$\|\boldsymbol{\alpha}_t\|^2 \quad \text{Vs.} \quad \|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2:$$

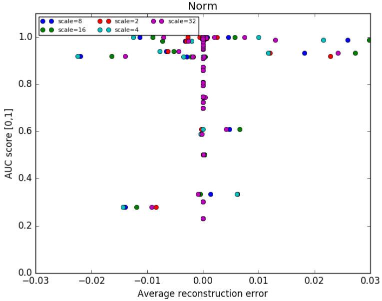


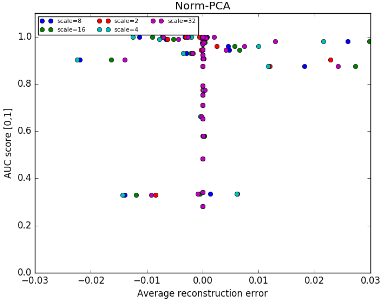


Effect of the Iterated Streaming PCA

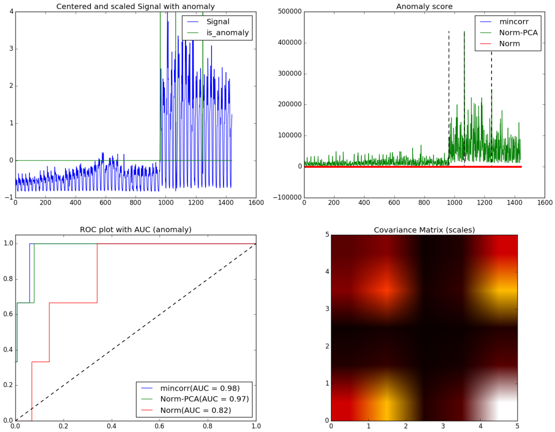
- The error here should decorrelate the scores at different scales.
- Plotting Mean Recon. Error (Approximation) Vs. AUC (Detection)

$$\|\boldsymbol{\alpha}_t\|^2 \quad \text{Vs.} \quad \|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2:$$





Effect of Iterated Streaming PCA



Failure Cases

When the errors of reconstruction across scales remain correlated :

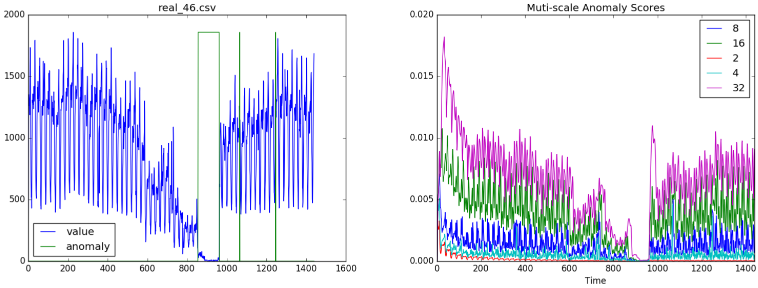


Figure : Scale correlation.

A larger scale of lag-window provides a least correlated scale.

Failure Cases

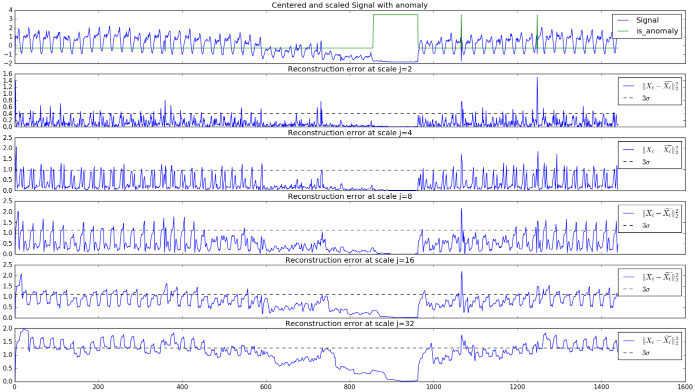


Figure : Near zero AUC score.

Future work

Improvements on current model

- Understand the bounds on the reconstruction error $\alpha(t)$ for Streaming PCA.
- Better base-line with the multivariate zscore by calculating covariance matrix online.
- Add anomaly-score likelihood to filter the anomaly score by using a moving window gaussian neg-log score.
- Use a streaming recursively calculable multi-scale time series representation $\Phi^T X_t$: This should make use of coefficients that are calculated in the past. For now the Haar transformation HX_t operates on a single vector. [4]

Future work

Other Tasks

- Anomolous time series ranking [8]
- Online Change-point evaluation [8]

Other applications :

- Unsupervised unsual action recognition in videos
- Change detection in areal/remote sensing data : hyperspectral video.

The End.

Hierarchical PCA

Initialization: $\mathbf{w}_j \leftarrow \mathbf{0}$, $\sigma_j^2 \leftarrow \epsilon$ with $\epsilon \ll 1$
for $t = 1, \ldots, T$ **do**
  for $j = 2, \ldots, J$ **do**
    if $j = 1$ **then**
      $Z_t^j \leftarrow X_t^j$
    else
      $Z_t^j \leftarrow [\pi_t^{j-1}, (X_t^j)^T]$
    end if
    $y_t^j \leftarrow \mathbf{w}_j^T Z_t^j$
    $\sigma_j^2 \leftarrow \sigma_j^2 + (y_t^j)^2$
    $\mathbf{e}_t^j \leftarrow Z_t^j - y_t^j \mathbf{w}_j$
    $\mathbf{w}_j \leftarrow \mathbf{w}_j + \sigma_j^{-2} y_t^j \mathbf{e}_t^j$
    $\pi_t^j \leftarrow \mathbf{w}_j^T Z_t^j$
    $\widetilde{Z}_t^j \leftarrow \pi_t^j \mathbf{w}_j$
    $\alpha_t^j \leftarrow \|\widetilde{Z}_t^j - Z_t^j\|^2$
  end for
end for
return $\alpha \in \mathbb{R}^{T \times J}$

Results PC=2

Multi-scale score-Norm $\|\boldsymbol{\alpha}_t\|^2$ (PC=2)					
fixed-scale	0.783 ± 0.269	0.918 ± 0.065	0.616 ± 0.142	0.569 ± 0.154	0.815 ± 0.231
fixed-scale-haar	0.808 ± 0.259	0.925 ± 0.074	0.627 ± 0.146	0.586 ± 0.144	0.811 ± 0.232
multiscale-lagmatrix	0.850 ± 0.242	0.969 ± 0.031	0.803 ± 0.116	0.686 ± 0.163	0.862 ± 0.210
hierarchical-approx	0.848 ± 0.240	0.985 ± 0.056	0.982 ± 0.021	0.941 ± 0.079	0.876 ± 0.213
multiscale-haar	0.862 ± 0.245	0.976 ± 0.021	0.805 ± 0.150	0.710 ± 0.166	0.873 ± 0.195

PCA on multi-scale score $\|\widetilde{\boldsymbol{\alpha}_t} - \boldsymbol{\alpha}_t\|^2$ (PC=2)					
fixed-scale	0.778 ± 0.270	0.908 ± 0.091	0.609 ± 0.133	0.573 ± 0.154	0.813 ± 0.232
fixed-scale-haar	0.804 ± 0.261	0.922 ± 0.079	0.625 ± 0.148	0.584 ± 0.143	0.811 ± 0.232
multiscale-lagmatrix	0.828 ± 0.237	0.872 ± 0.134	0.834 ± 0.172	0.793 ± 0.181	0.829 ± 0.207
hierarchical-approx	0.831 ± 0.248	0.978 ± 0.084	0.976 ± 0.031	0.935 ± 0.084	0.841 ± 0.231
multiscale-haar	0.816 ± 0.239	0.933 ± 0.088	0.859 ± 0.161	0.799 ± 0.171	0.807 ± 0.226

Least correlated scale $\alpha_t^{j^*}$ where $j^* = \arg\min_j \sum_i (\boldsymbol{\alpha}^T\boldsymbol{\alpha})_{ji}$ (PC=2)					
fixed-scale	0.783 ± 0.269	0.918 ± 0.065	0.616 ± 0.142	0.569 ± 0.154	0.815 ± 0.231
fixed-scale-haar	0.808 ± 0.259	0.925 ± 0.074	0.627 ± 0.146	0.586 ± 0.144	0.811 ± 0.232
multiscale-lagmatrix	0.685 ± 0.332	0.757 ± 0.225	0.555 ± 0.140	0.597 ± 0.168	0.736 ± 0.327
hierarchical-approx	0.689 ± 0.333	0.757 ± 0.225	0.555 ± 0.140	0.596 ± 0.167	0.736 ± 0.327
multiscale-haar	0.739 ± 0.318	0.765 ± 0.241	0.533 ± 0.200	0.512 ± 0.200	0.736 ± 0.336

References I

[1] V. Alarcon-Aquino and J. A. Barria. Anomaly detection in communication networks using wavelets. *IEE Proceedings - Communications*, 148(6):355–362, dec 2001.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, may 2000.

[3] X.-y. Chen and Y.-y. Zhan. Multi-scale anomaly detection algorithm based on infrequent pattern of time series. *J. Comput. Appl. Math.*, 214(1):227–237, Apr. 2008.

[4] G. Cormode, M. Garofalakis, and D. Sacharidis. Fast approximate wavelet tracking on streams. In *International Conference on Extending Database Technology*, pages 4–22. Springer, 2006.

[5] P. H. dos Santos Teixeira and R. L. Milidiú. Data stream anomaly detection through principal subspace tracking. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1609–1616, New York, NY, USA, 2010. ACM.

[6] R. Ganesan, T. K. Das, and V. Venkataraman. Wavelet-based multiscale statistical process monitoring: A literature review. *IIE transactions*, 36(9):787–806, 2004.

[7] C. T. Huang, S. Thareja, and Y. J. Shin. Wavelet-based Real Time Detection of Network Traffic Anomalies. In *Securecomm and Workshops, 2006*, pages 1–7, aug 2006.

[8] N. Laptev, S. Amizadeh, and I. Flint. Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1939–1947. ACM, 2015.

[9] A. Lavin and S. Ahmad. Evaluating real-time anomaly detection algorithms - the numenta anomaly benchmark. In *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, pages 38–44. IEEE, 2015.

[10] W. Lu, M. Tavallaee, and A. A. Ghorbani. Detecting Network Anomalies Using Different Wavelet Basis Functions. In *Communication Networks and Services Research Conference, 2008. CNSR 2008. 6th Annual*, pages 149–156, may 2008.

[11] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 697–708. VLDB Endowment, 2005.

References II

[12] S. Papadimitriou and P. Yu. Optimal multi-scale patterns in time series streams. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 647–658. ACM, 2006.

[13] S. Pukkawanna and K. Fukuda. Combining sketch and wavelet models for anomaly detection. In *2010 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 313–319, aug 2010.

Previous work

- Standard offline multiscale anomaly detection using wavelet transform [1], [10], [13], [7].
- Wavelet methods introduce a time delay in the computation of the coefficients at non-dyadic locations which worsens geometrically for coarser scales. Furthermore, they suffer from non-causality, *i.e.* they need to see some part of the future to assess the presence of an anomaly at present time [6].
- [8] proposed several linear predictive models (Autoregressive, Kalman filter) followed by an anomaly score filtering (by $k\sigma$ rule, or local outlier factor scores introduced by [2]) to detect anomalies.