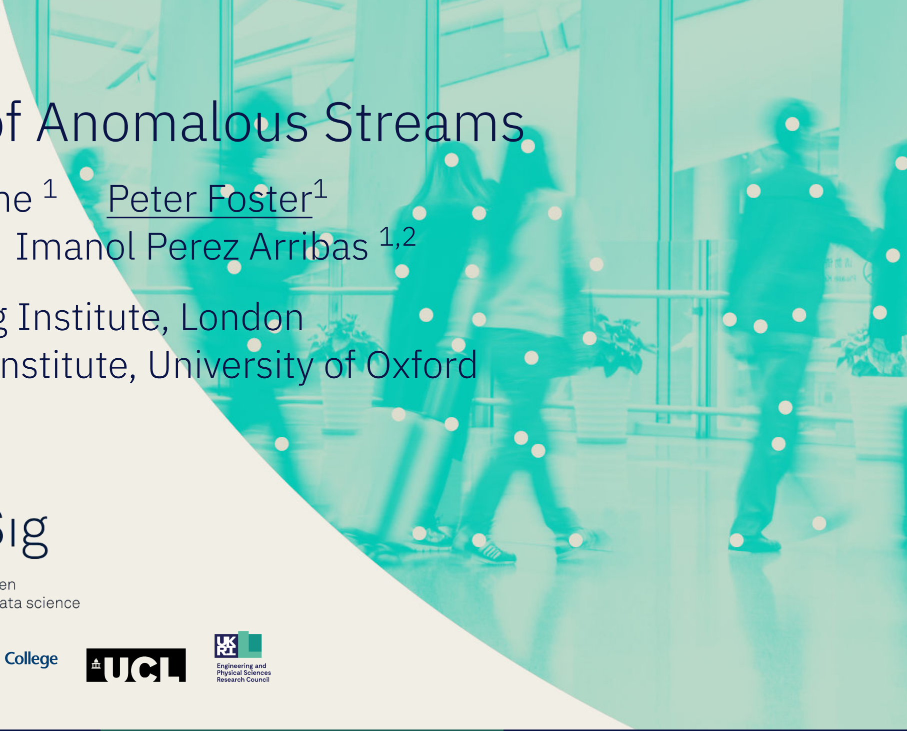
Detection of Anomalous Streams

Thomas Cochrane [1] Peter Foster[1]
Terry Lyons [1,2] Imanol Perez Arribas [1,2]

[1]The Alan Turing Institute, London
[2]Mathematical Institute, University of Oxford



DataSig

A rough path between
mathematics and data science

Overview

- **Task**: Given a corpus of observations deemed to be *normal*, determine if a new observation is *normal* or *anomalous*.
- We propose a powerful but simple method for detecting anomalies among vector-valued observations
- Using path signatures as features, our method addresses the task of detecting anomalous time series and other types of *streamed data*
- We demonstrate the effectiveness of our method experimentally using both univariate and multivariate stream datasets

Anomaly Detection

- Anomaly detection has been considered widely (Chandola et al., 2009), with applications in numerous areas e.g. medicine, financial fraud, cyber-security
- Relatively little prior work on deciding whether an entire time series (or other type of stream) is anomalous (Gupta et al, 2013; Blázquez-García, 2020)
- Natural and widespread approach involves using a distance metric to quantify unusualness
- **What constitutes a useful metric?**

The Variance Norm

Let μ be a probability measure on a vector space V. The covariance quadratic form $\mathrm{Cov}(\psi,\phi) := \mathbb{E}^{\mu}[\psi(\mathbf{x})\phi(\mathbf{x})]$, defined on the dual of V, induces a dual norm defined for $\mathbf{x} \in V$ by

$$\|\mathbf{x}\|_{\mu} := \sup_{\mathrm{Cov}(\phi,\phi)\leq 1} \phi(\mathbf{x}). \tag{1}$$

We refer to this norm, computed for the measure μ re-centered to have mean zero, as the **variance norm** $\|\cdot\|_{\mu}$ (a.k.a. Mahalanobis norm) associated to μ.

- $\|\cdot\|_{\mu}$ is finite on the linear span of the support of μ, and infinite outside of it.
- $\|\cdot\|_{\mu}$ is well-defined whenever the measure has finite second moments and, in particular, for the empirical measure associated to a finite set of observations.

The Conformance Distance

- One possible approach involves using the *variance norm* directly to quantify unusualness
- This approach has a significant drawback: In the high-dimensional case, the norm is huge; in the infinite-dimensional case, the norm is infinite
- Instead, we define the *conformance* of $\mathbf{x}$ to μ to be the distance

$$\operatorname{dist}(\mathbf{x};\mu) := \inf_{\mathbf{y}\in\operatorname{supp}(\mu)} \|\mathbf{x}-\mathbf{y}\|_{\mu}. \tag{2}$$

- We motivate this approach based on the Tsirelson-Sudakov-Borell isoperimetric inequality
- A new member of the corpus will be far away from most members of the corpus, but with high probability there will be some members of the corpus to which it is close

Streams of Data

- We define the space of streams of data in a set $\mathcal{Z}$ as

$$\mathcal{S}(\mathcal{Z}) := \{\mathbf{z} = (z_1, \ldots, z_m) \, : \, z_i \in \mathcal{Z}, m \in \mathbb{N}\}. \tag{3}$$

- For example, when a person writes a character by hand, the stroke of the pen naturally determines a path.
- If we record the trajectory we obtain a two-dimensional stream of data $\mathbf{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in \mathcal{S}(\mathbb{R}^2)$.
- Number of points may vary across streams
- Points need not be evenly spaced; point spacing may vary across streams
- To obtain a path from stream (and compute its signature), we interpolate between successive points in the stream

Expected Signature and the Variance Norm

- Let $\mathcal{C}$ be a finite corpus (or empirical measure) of streams of data.
- Let Sig^N be the signature of order N.
- Then $\|\cdot\|_{\mathrm{Sig}^N(\mathcal{C})}$ is the variance norm associated with the empirical measure of $\mathrm{Sig}^N(\mathcal{C})$.
- Let $\mathbf{w} \in \mathbb{R}^{d_N}$. Using the shuffle product ш, we define $\mathbf{A}_{i,j} := \langle e_i ш e_j, \mathbb{E}[\mathrm{Sig}^{2N}(\mathbf{x})]\rangle$ for $i, j = 1, \ldots, d_N$. We may express the variance norm as

$$\|\mathbf{w}\|^2_{\mathrm{Sig}^N(\mathcal{C})} = \langle \mathbf{w}, \mathbf{A}^{-1}\mathbf{w}\rangle. \tag{4}$$

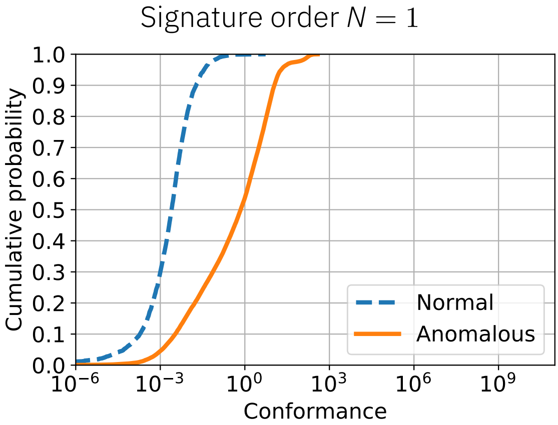
Evaluation

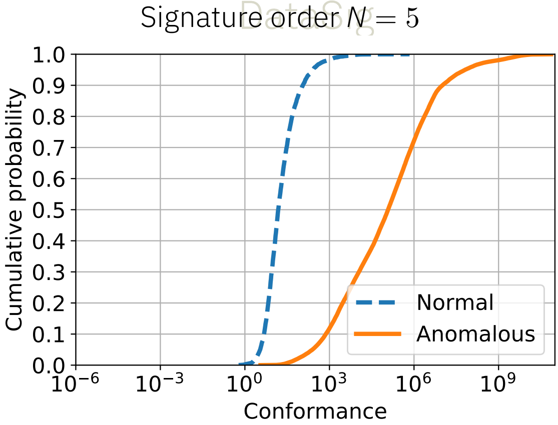
- We have a data set $\mathcal{I}$ partitioned into those data deemed to be normal $\mathcal{I}_{\text{normal}}$ and those data deemed to be anomalous $\mathcal{I}_{\text{anomaly}}$
- By further partitioning, we obtain the corpus $\mathcal{C} \subset \mathcal{I}_{\text{normal}}$ which we use for training
- As our testing data $\mathcal{Y}$ we use $\mathcal{Y} := \mathcal{I} \setminus \mathcal{C}$
- Evaluate performance against three separate datasets
  - *PenDigits* handwritten digit
  - Marine vessel traffic data
  - UEA & UCR univariate time series

PenDigits Dataset

- 10 992 instances of hand-written digits captured from 44 subjects
- Each instance represented as a 2-dimensional stream
- Define $\mathcal{I}_{\text{normal}}$ as the set of instances representing digit $m \in [0..9]$, assign remaining instances to $\mathcal{I}_{\text{anomaly}}$
- Results based on aggregating conformance scores across digits and computing receiver operating characteristic area under curve (ROC-AUC)

PenDigits Dataset – Results


Signature order $N = 1$

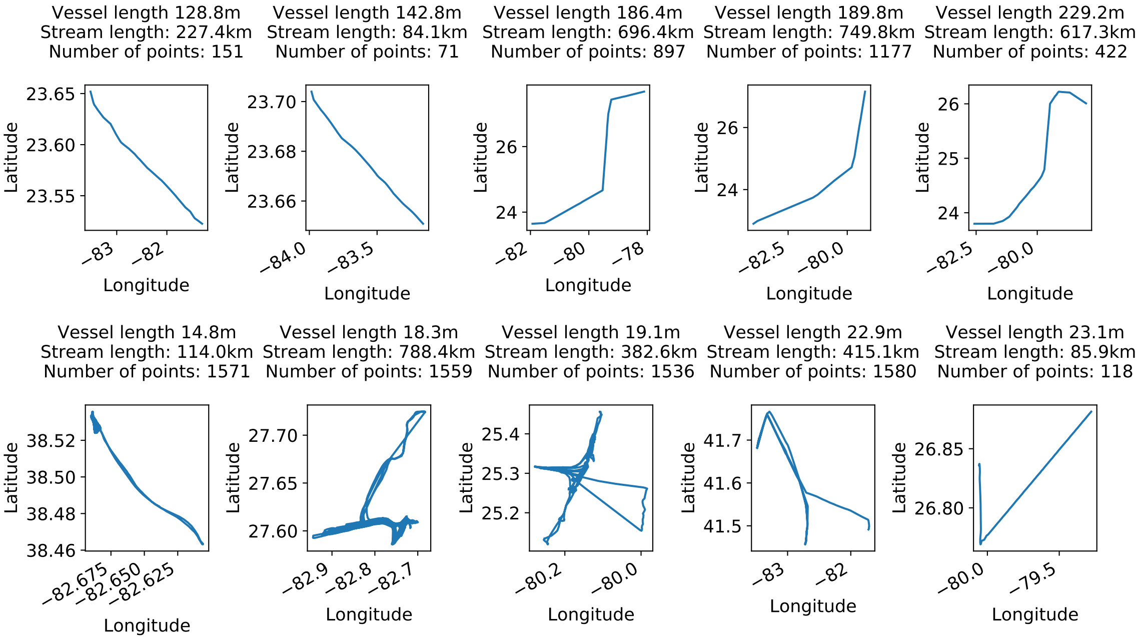

Signature order $N = 5$

Signature order	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
ROC-AUC	0.901	0.965	0.983	0.987	0.989

Marine Vessel Traffic Dataset

- Automatic identification system (AIS) data collected by the US Coast Guard
- Total of 31 884 021 timestamped geographical (lat/lon) positions recorded for 6 282 distinct vessel identifiers in January 2017
- To evaluate effect of stream length on performance, disintegrate streams into sub-streams of length $D \in \{4\text{km}, 8\text{km}, 16\text{km}, 32\text{km}\}$ between initial and final points
- Deem sub-stream normal if it belongs to vessel with length greater than 100m, anomalous if length less than or equal to 50m
- Evaluate combinations of stream transformation, with signature order $N = 3$
- **Baseline**: Summarise sub-stream using its component-wise mean and covariance, use as features for isolation forest

Marine Vessel Traffic Dataset – Example Paths



Vessel length 128.8m
Stream length: 227.4km
Number of points: 151

Vessel length 142.8m
Stream length: 84.1km
Number of points: 71

Vessel length 186.4m
Stream length: 696.4km
Number of points: 897

Vessel length 189.8m
Stream length: 749.8km
Number of points: 1177

Vessel length 229.2m
Stream length: 617.3km
Number of points: 422

Vessel length 14.8m
Stream length: 114.0km
Number of points: 1571

Vessel length 18.3m
Stream length: 788.4km
Number of points: 1559

Vessel length 19.1m
Stream length: 382.6km
Number of points: 1536

Vessel length 22.9m
Stream length: 415.1km
Number of points: 1580

Vessel length 23.1m
Stream length: 85.9km
Number of points: 118

Marine Vessel Traffic Dataset – Results

Signature conformance – performance quantified using ROC-AUC

Transformation			Sub-stream length *D*			
Lead-lag	Time-Diff	Inv. Reset	4km	8km	16km	32km
No	No	No	**0.723**	0.706	0.705	**0.740**
No	No	Yes	**0.776**	**0.789**	**0.785**	**0.805**
No	Yes	No	**0.810**	**0.813**	**0.818**	**0.848**
No	Yes	Yes	**0.839**	**0.860**	**0.863**	**0.879**
Yes	No	No	**0.811**	**0.835**	**0.824**	**0.837**
Yes	No	Yes	**0.812**	**0.835**	**0.833**	**0.855**
Yes	Yes	No	**0.845**	**0.861**	**0.862**	**0.877**
Yes	Yes	Yes	**0.848**	**0.863**	**0.870**	***0.891***

Marine Vessel Traffic Dataset – Baseline Comparison

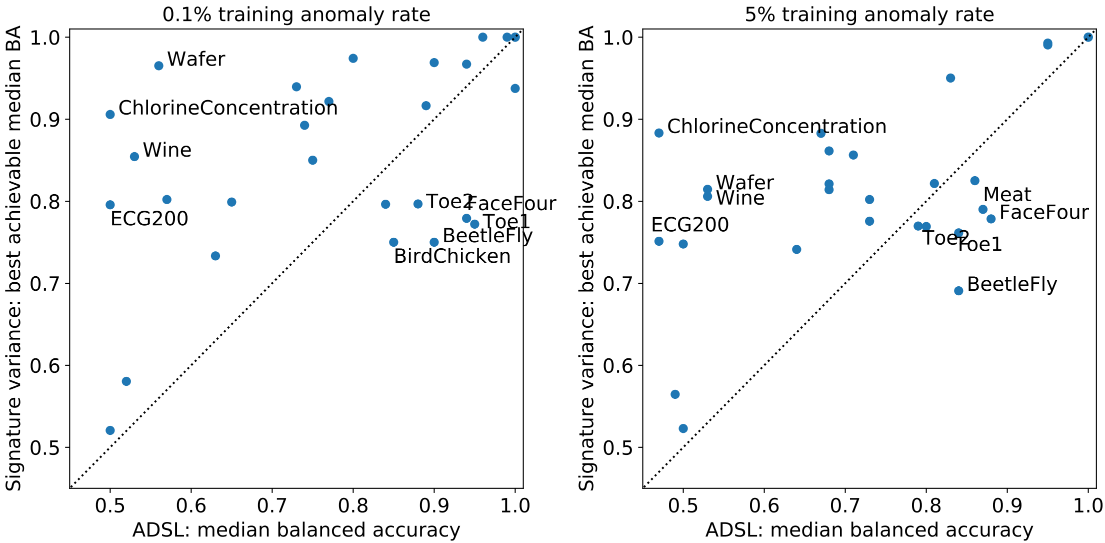
Baseline approach – performance quantified using ROC-AUC

Transformation			Sub-stream length D			
Lead-lag	Time-Diff	Inv. Reset	4km	8km	16km	32km
No	No	No	0.690	**0.718**	**0.717**	0.733
No	No	Yes	0.682	0.698	0.714	0.716
No	Yes	No	0.771	0.779	0.779	0.803
No	Yes	Yes	0.745	0.751	0.761	0.797
Yes	No	No	0.759	0.765	0.766	0.763
Yes	No	Yes	0.755	0.761	0.763	0.762
Yes	Yes	No	0.820	0.815	0.823	0.817
Yes	Yes	Yes	0.810	0.795	0.816	0.815

UEA & UCR Dataset

- Collection of 28 individual univariate datasets
- Each dataset comprises a set of time series of equal length, together with class labels
- One class designated as the normal class, with all other classes designated as anomalies
- Include a small amount of anomalous observations in the training corpus (contamination rates 0.1%, 5%)
- Convert to 2-dimensional stream by applying time-integrated transformation
- Take signatures of order $N = 5$
- Quantify performance using balanced accuracy with optimal decision threshold
- **Baseline**: Shapelet method proposed by Beggel et al. (2019)

UEA & UCR Dataset – Results



Conclusion

- Motivated by the Tsirelson-Sudakov-Borell isoperimetric inequality, we introduce the notion of conformance as a method for anomaly detection
- Using signatures as vector-valued features, our approach aims at detecting anomalous streams
- Approach applicable generally in both univariate and multivariate settings; points in stream need not be evenly spaced; point spacing may vary across streams; streams need not be of equal length
- Steam transformations allow us to modify representational properties as required for the task
- Experimental results suggest that our approach performs strongly in several application domains, with favourable results obtained compared to baseline approaches, especially more specialised univariate baseline
- We are currently focussing on a more detailed evaluation of our method applied to alternative anomaly detection tasks, as well comparing it to other existing approaches

References

- L. Beggel, B.X. Kausler, M. Schiegg, M. Pfeiffer, and B. Bischl. Time series anomaly detection based on shapelet learning. *Computational Statistics*, 34(3):945–976, 2019.
- A. Blázquez-García, A. Conde, U. Mori, and J.A. Lozano. A review on outlier/anomaly detection in time series data. *arXiv preprint* arXiv:2002.04236, 2020.
- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58, 2009.
- T. Cochrane, P. Foster, T. Lyons, and I. P. Arribas. Anomaly detection on streamed data. *arXiv preprint* arXiv:2006.03487, 2020.
- M. Gupta, J. Gao, C.C. Aggarwal, and J. Han. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267, 2013.