NetSciX2026: International School and Conference on Network Science
February 17^{th}-20^{th}, 2026 - Auckland, New Zealand

Surprisal-Based Anomaly Detection in Sparse High-Dimensional Data

Keywords: surprisability, missing commonalities, high-dimensional data, anomaly detection

Extended Abstract

A central challenge lies in analyzing high-dimensional data, where the number of measured variables vastly exceeds the available samples. The volume of the ambient space grows exponentially, whereas the number of observations grows only polynomially, leading to sparsity. That is, points become widely separated, neighborhoods collapse, and classical asymptotic guarantees fail [1]. Quantitative data analysis has operated under a fundamental assumption: meaningful patterns come from correlations between present features alone. This overlooks a critical insight: the structure of complex data emerges not only from what is present, but from patterns of both presence and absence. **When expected patterns are absent from a specific datum, this carries diagnostic information that current methods ignore**. We term these absent population-prevalent elements *missing commonalities (MCs)*.

Learning via Surprisability (LvS) was designed to identify MCs. LvS works on data where each data point is a discrete probability distribution over a finite set of features. (1) LvS creates a population-wide statistical reference model, comprised of all features in the data, that corresponds to a mixture model MLE with equal component weights. This yields a full-support reference model, enabling the well-defined measurement of Jensen–Shannon divergence (JSD) for every datum. (2) LvS then computes the JSD between the reference model and each data point. Since JSD is defined as the average of two Kullback-Leibler Divergence (KLD) measurements relative to a mixture distribution, it remains bounded and finite even for sparse distributions. *LVS is a key innovation because it decomposes divergence to isolate which specific features drive each data point's deviation from the reference model, whether overrepresented or absent*. These features define each datum's surprisal signature [1].

Empirical Results. Three domains validate LvS effectiveness in identifying anomalies and patterns on high-dimensional data: **Cybersecurity and time series anomaly detection.** LvS achieved an 82.3% AUC score and F_1 of 0.7 on the SWaT industrial control system dataset [2] (450,000 records), significantly outperforming established methods like Isolation Forest(54.5% AUC, 0.15 F_1), Local Outlier Factor (LOF) (50.3%, 0.06), One Class SVM (9.1%,0.83), LSTM optimized for anomaly detection (78.5%, 0.74), and fully connected Autoencoder (71%,0.63) (Baseline methods reviews [3, 4]). **Global Health Surveillance**: 30-year worldwide mortality data (30 causes). Results are demonstrated in Figure 1. **Temporal Text Analysis**: Analyzing U.S. Presidential State of the Union speeches for 46 presidents, from 1790 to 2022), LvS detected major war events, as well as topical shifts: 1942 coordinated wartime language shift (increased "war," "fighting"; decreased "government," "law"), 1951 Cold War terminology emergence ("free," "soviet"). Presidential clustering based on surprisability patterns aligned with independent linguistic studies.

Ethical aspects. While this research uses public datasets for beneficial applications, the method's ability to detect anomalous patterns in high-dimensional data might pose ethical concerns for future applications such as surveillance and warrants careful consideration of appropriate use cases and safeguards.

[1] Formal formulation was removed to adhere to abstract size limits.

References

[1] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.

[2] iTrust, Centre for Resilient and Secure Systems. *Secure Water Treatment (SWaT) Testbed Dataset*. `https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/`. 2016.

[3] O. Alghushairy et al. "A review of local outlier factor algorithms for outlier detection in big data streams". In: *Big Data and Cognitive Computing* 5.1 (2020), p. 1.

[4] Haoqi Huang et al. "Deep learning advancements in anomaly detection: A comprehensive survey". In: *IEEE Internet of Things Journal* (2025).

[5] Christopher JL Murray. "The global burden of disease study at 30 years". In: *Nature medicine* 28.10 (2022), pp. 2019–2026.

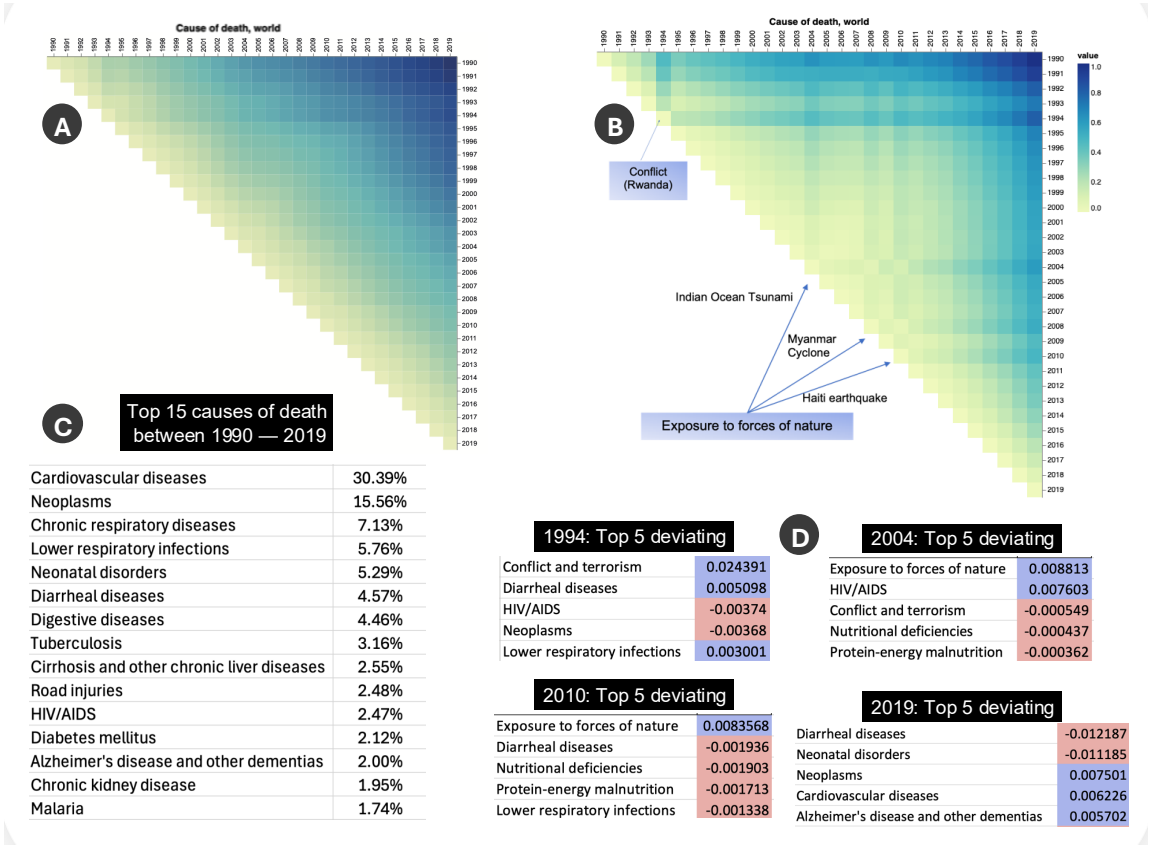

Figure 1: LvS analysis of world main causes of death, data spanning 1990–2019 [5]. (A) Comparison of the original vectors representing causes of death over 30 years. Each square corresponds to the comparison of probability distribution vectors between two years, with lighter colors indicating greater similarity. The figure clearly shows that any two consecutive years exhibit a high degree of similarity in the underlying causes of death worldwide. (B) Comparison of the Surprisal Profile vectors created by LvS of the causes of death over 30 years. Each square corresponds to the comparison of the surprisability signatures' (SP) probability distribution vectors between two years, with lighter colors indicating greater similarity. Notable anomalies are clear in the years 1994, 2004, 2008, and 2010. The values within the surprisability signature vectors enable interpretation of the most surprising elements causing the anomaly in these years. Values were normalized for easier reading. (C) Top 15 causes of death during the period. (D) the most prominent features (causes of death) in the surprisability signatures for the years 1994, 2004, 2010, and 2019. Blue indicates over-represented, while red highlights those that are under-represented or missing relative to the dataset reference model.