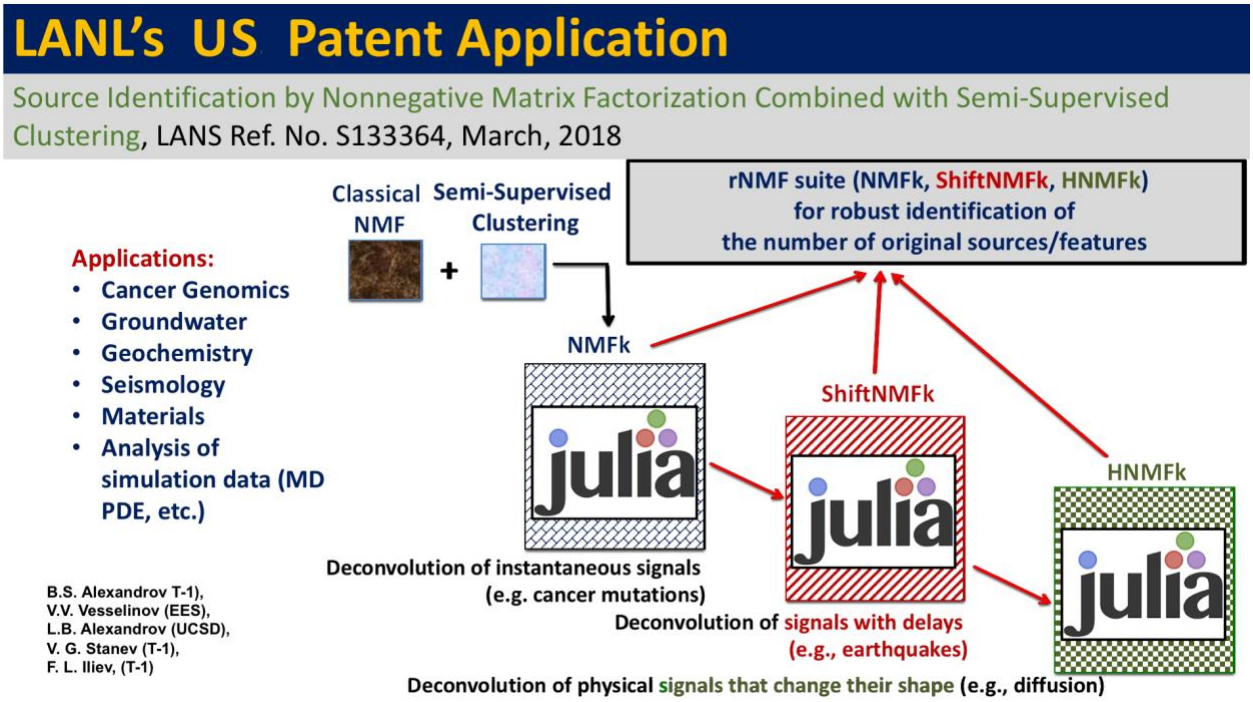
March 2018

New Machine Learning Algorithms: A Suite of One Patent and Two Publications



rNMF Schematic: A suit of three new methods and algorithms for data exploratory analysis, features extraction, data mining and dimension reduction based on subspace learning via robust Nonnegative Matrix Factorization (NMF) with a custom estimation of the unknown number of factors (features), and integrated Green's function, time delays, and other physical constraints.

The Algorithms and Mathematics

The novelty: The unknown number of sources/features are identified by multiple trials of nonnegative matrix factorization performed for a fixed number of sources/features, with selection criteria applied to determine successful trials. A custom semi-supervised clustering procedure is applied to the trial results, and the clustering results are evaluated for robustness using measures for reconstruction quality and cluster separation. The number of sources/features is determined by comparing the quality and cluster separation measures, the accuracy of the reconstruction and leveraging AIC statistical criterion for the trials with different numbers of sources. Source locations and parameters of the signal propagation also are determined.

The Impact

Disclosed methods are applicable to a wide range of time transients and spatial problems including chemical dispersal, pressure transients, heat conduction, electromagnetic signals, and also to non-spatial

problems such as cancer mutations, data and text mining, and X-ray data of combinatorial libraries. Our methods have been also applied for analysis of computer simulation data.

Summary

Unsupervised Machine Learning (ML) methods aim to extract sets of latent (and often previously unknown) features from uncategorized datasets. Nowadays, integration of big multicomponent datasets, powerful computational capabilities, and affordable data storage has resulted in active use of advanced ML algorithms. However, ML tools for efficient and robust extraction of latent features buried in petabytes of multicomponent big datasets are still lacking. Identification of the different manifestations of these latent processes in the data is part of exploratory data science that allows discoveries of new mechanisms and causalities hidden in the datasets. Unsupervised ML methods based on factor analysis, such as, Principle Component Analysis (PCA), Independent Component Analysis (ICA) and Nonnegative Matrix Factorization (NMF) are widely used to accomplish this kind of tasks. NMF is superior in many cases because the nonnegativity constraints in guarantee that the extracted latent features will be physically interpretable. Indeed, if only addition but not subtractions are permissible, reproducing a dataset requires the identified features to be parts of the original data, thus, making these processes easy to understand and interpret. Many types of state variables (e.g., density, energy, spectra, etc.) are naturally nonnegative and many others can be examined as nonnegative via suitable transformations. However, the classical NMF algorithm requires *prior* knowledge of the number of the original features. Our patent proposes a novel method for estimation the unknown number of latent features based on the solution robustness. The methods from our patent were actively used to perform the world's largest analysis of human cancer genomics data and to extract unique hidden mutagenic signatures buried in petabytes of data. Our methods have been also applied for identification of pressure transients, contaminant sources, and crystal structures in materials combinatory libraries.

Contact

Boian S. Alexandrov, boian@lanl.gov
Los Alamos National Laboratory

Funding

This research was funded by the Environmental Programs Directorate of the Los Alamos National Laboratory, the DiaMonD project (An Integrated Multifaceted Approach to Mathematics at the Interfaces of Data, Models, and Decisions, U.S. Department of Energy Office of Science, Grant #11145687, and LANL LDRD office, reserve DR Grant#20180060.

Publications

Alexandrov BS, Alexandrov LB, Iliev FL, Stanev VG, Vesselinov VV, inventors; Los Alamos National Security LLC, assignee. *Source identification by non-negative matrix factorization combined with semi-supervised clustering*. United States Patent Application US 15/690,176. 2018 Mar 1.

Iliev FL, Stanev VG, Vesselinov VV, Alexandrov BS. *Nonnegative Matrix Factorization for identification of unknown number of sources emitting delayed signals*. PloS one. 2018 Mar 8;13(3):e0193974, https://doi.org/10.1016/j.apm.2018.03.006.

Stanev, V.G., Iliev, F.L., Hansen, S., Vesselinov, V.V. and Alexandrov, B.S., 2018. Identification of release sources in advection-diffusion system by machine learning combined with Green's function inverse method. *Applied Mathematical Modelling*. https://doi.org/10.1016/j.apm.2018.03.006

Related Links

https://patents.google.com/patent/US20180060758A1/en

https://www.sciencedirect.com/science/article/pii/S0307904X18301227

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193974