Unsupervised Machine Learning based on Nonnegative Factorization

Velimir V. Vesselinov (monty), Boian S. Alexandrov, Daniel O’Malley, Maruti K. Mudunuru, Satish Karra

Earth and Environmental Sciences Division — Theoretical Division
Los Alamos National Laboratory, NM 87545, USA
Supervised Machine Learning: requires prior categorization of the processed data (can introduce subjectivity)

Unsupervised Machine Learning: discovers hidden features in the processed data without any prior information

Deep Machine Learning: ... coupled supervised and unsupervised techniques
Data analytics:
- Feature extraction (FE)
- Blind source separation (BSS)
- Image recognition
- Detection of disruptions / anomalies
- Guide development of physics / reduced-order models representing the data

Analyses of model outputs:
- Identify dominant processes (features) in the model outputs
- Guide development of reduced-order models
We have developed a series of novel Unsupervised Machine Learning methods based on Nonnegative Factorization + custom clustering:

- identify the number of robust features in the data
- extract robust features representing the data
- extracted features are parts of the data allowing for intuitive interpretations

Selected publications:

- Alexandrov, Vesselinov, Blind source separation for groundwater level analysis based on nonnegative matrix factorization, Water Resources Research, 10.1002/2013WR015037, 2014

LANL Unsupervised Machine Learning (ML) Patent:
Nonnegative Factorization + custom clustering

- **NMF\(_k\)**: matrix-based (two-dimensional) algorithm (well-tested; widely used)
  - Extract barometric and pumping effects in pressure data
  - Identify and predict processes for optimal control of the LANSCE particle accelerator
  - Characterize materials using X-ray
  - Analyze model predictions of molecular dynamics trajectories
  - Characterize influenza epidemics
  - Extract image features using Quantum Computing (D-Wave)
  - Identify cancer signatures in human genomes (30+ papers in Nature/Science/Cell)

- **NTF\(_k\)**: tensor-based (high-dimensional) algorithm (actively developed at the moment)
  - Here, we present **NTF\(_k\)** applications related to contaminant transport characterization
NMF: Nonnegative Matrix Factorization

X = W × H

- **I**: number of observation points (wells)
- **Q**: number of geochemical species observed (e.g., $Cr^{6+}$, $SO_4^{2+}$, $NO_3^-$, etc.)
- **K**: number of unknown groundwater types mixed at each well
- **Constraints**: all matrix elements $\geq 0$

\[ \sum_{k=1}^{K} W_{i,k} = 1 \quad \forall i \]

- **X**: data matrix
- **W**: mixing matrix (unknown)
- **H**: source matrix (unknown)
Tucker tensor factorization (3D case)

\[ Y = A \times G \times E^T \]

Factorizing all 3 dimensions \((I \rightarrow J, T \rightarrow R, Q \rightarrow P)\)
Tucker-1 tensor factorization (3D case)

Factorizing only 1 of the dimensions ($Q \rightarrow K$)
NTF: Nonnegative Tensor Factorization based on Tucker-1 decomposition

- **I**: number of observation points (wells)
- **Q**: number of geochemical species observed (e.g., $Cr^{6+}$, $SO_4^{2+}$, $NO_3^-$, etc.)
- **T**: number of observation times (e.g., 2001, 2002, ..., 2017)
- **K**: number of unknown groundwater types mixed at each well

**Constraints:**

\[
\sum_{k=1}^{K} G_{i,k,t} = 1 \quad \forall i, t
\]

- **Y**: data tensor
- **A**: source (groundwater type) matrix (unknown)
- **G**: mixing tensor (unknown)
Nonnegative Factorization (NTF) Analyses

- Major challenges for both $\text{NMF}_k$ and $\text{NTF}_k$:
  - identifying the number of unknown features (groundwater types) $K$ (in $\text{NMF}_k$, resolved using custom clustering; based on the Frobenius norm and cluster Silhouettes; identification under $\text{NTF}_k$ is much more challenging)
  - solving the constraint optimization problem to estimate matrix/tensor elements
  - dealing with large high-dimensional datasets (high-performance computing)
  - ...

- We apply (demonstrate) $\text{NTF}_k$ to two datasets
  - **Field Data**: time-dependent mixing of groundwater types (contaminant sources)
  - **Simulation Data**: fluid mixing impacts on a geochemical reaction $A + B \rightarrow C$
LANL site

Unsupervised Machine Learning
NMF/NTF
ML geochemistry
ML fluid mixing
Summary
LANL hydrogeochemical datasets (high-dimensional)

Unsupervised Machine Learning
- NMF/NTF
- ML geochemistry
- ML fluid mixing

Summary
NTFk: estimated time-dependent mixing of 4 groundwater types at various wells

Unsupervised Machine Learning
NMF/NTF
ML geochemistry
ML fluid mixing
Summary
NTF\(_k\): estimated time-dependent mixing of 4 groundwater types at various wells

**R-11**

**R-50#1**

**R-1**

**R-43#1**

Unsupervised Machine Learning

NMF/NTF

ML geochemistry

ML fluid mixing

Summary
We want to find how time/space behavior of $C'$ concentrations is controlled by the simulated physics processes.

$> 2000$ simulations of $C'$ concentrations in time/space for a series of model parameters impacting fluid mixing; 3 example predictions:
NTF<sub>k</sub> results

- > 200 GB simulation data compressed to ≈ 70 MB (compression ≈ 4 × 10^{-4})
  Here, (1000 × 81 × 81) → (3 × 8 × 9)
- NTF<sub>k</sub> processed all the data and extracted the dominant time/space features (processes / vortices)
We have developed a series of novel unsupervised ML methods based on Nonnegative Factorization (Matrices/Tensors).

These ML methods have been used to solve various real-world problems.

Some of our ML analyses brought breakthrough discoveries (especially related to human cancer research).

We have developed a series of ML computational tools for solving big-data problems using high-performance computing (HPC).
Machine Learning (ML) Algorithms / Codes developed by our team

- NMF$_k$ + ShiftNMF$_k$ + GreenNMF$_k$ (patent)
- NTF$_k$ (prototype; work in progress)
- NBMF: Quantum machine learning using D-Wave
- SVR/SVM: Support Vector Regression/Machine
  http://github.com/madsjulia/SVR.jl
- MADS: Model-Analyses & Decision Support
  open-source, version-controlled, high-performance computational framework
  http://madsjulia.github.io/Mads.jl/Examples/blind_source_separation

Unsupervised Machine Learning
- NMF/NTF
- ML geochemistry
- ML fluid mixing

Summary