Unsupervised Machine Learning Based on Non-Negative Tensor Factorization

Velimir V. Vesselinov, Los Alamos National Laboratory, USA
Daniel O'Malley, Los Alamos National Laboratory, USA
Boian S. Alexandrov, Los Alamos National Laboratory, USA

Key words: feature extraction; species mixing

Introduction

Unsupervised machine learning (ML) methods are powerful tools for data analyses to extract essential features hidden in data. The integration of large datasets, powerful computational capabilities, and affordable data storage has resulted in the widespread use of ML in science, technology, and industry. Here we present applications of ML to characterize physical processes related reactive transport in porous media. Our ML method is based on Sparse Non-Negative Tensor Factorization (SNTF) and is applied to reveal the temporal and spatial features in reactants and product concentrations.

Methodology

The factorization of tensor **X** is typically performed by minimization of the norm $\frac{1}{2}||\mathbf{X} - \mathbf{W} \otimes_{\mathbf{1}} \mathbf{A_1} \ldots \otimes_{\mathbf{N}} \mathbf{A_N}||_F^2$, where **W** is a low-rank tensor (with a rank lower than the rank of **X**), $\mathbf{A_1}, \mathbf{A_2} \ldots \mathbf{A_N}$ are mixing factors, and $\mathbf{W} \otimes_{\mathbf{1}} \mathbf{A_1} \ldots \otimes_{\mathbf{N}} \mathbf{A_N} \equiv \mathbf{f}(\alpha, \beta, \ldots)$ is a factorization model (e.g., Candecomp/Parafac (CP), Tucker, etc.) decomposing the tensor **X** (Fig.1). Note that different models will have different number of free parameters: α, β, ... The reconstruction of **X** is $\mathbf{X} = \mathbf{f}(\alpha, \beta, \ldots) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a tensor of residual errors. Some of the factorization models can theoretically lead to unique solutions under specific, albeit rarely satisfied, noiseless conditions[1-3]. When these conditions are not satisfied, additional constraints can assist the factorization. A popular approach is to add nonnegative constraints leading to Nonnegative Tensor Factorization (NTF)[4]. Nonnegativity enforces parts-based representation of the original data which also allows the NTF results for **W** and $\mathbf{A_1}, \mathbf{A_2} \ldots \mathbf{A_N}$ to be easily interrelated[4]. The NTF results frequently represent hidden features extracted from the original data. The NTF method applied here (SNTF) allows also for sparsity constraints in the minimization process[5,6]; in this way, the solutions for **W** and $\mathbf{A_1}, \mathbf{A_2} \ldots \mathbf{A_N}$ (Fig.1) would have as many zero entries as possible while reproducing **X** with sufficient accuracy.

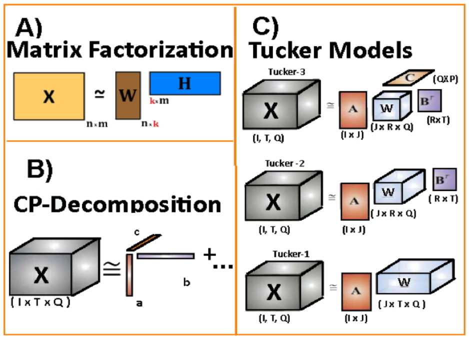

Fig.1: A) *Matrix factorization;* **B & C)** *Example tensor factorization models.*

Data & Results

LANL site: Here we explore the spatial and temporal evolution of groundwater contaminants at the LANL site[7]. A small subset of the site data is shown in Fig.2. The data describe site physical/biogeochemical governing processes that are challenging to conceptualize and simulate with physics models[8,9]. We apply SNTF to analyze the data and extract groundwater types and contaminant sources manifested in the data. The obtained results are presented in Fig.2.

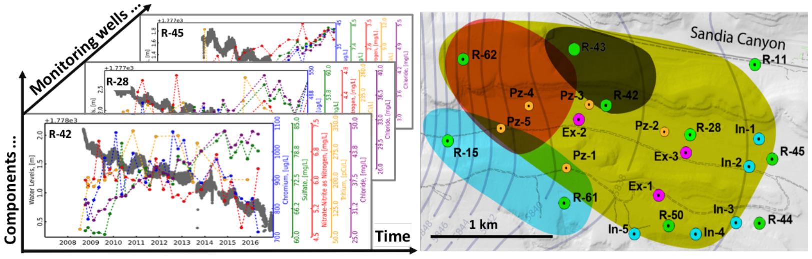

Fig.2: ***Left:*** *A subset of the tensorial LANL site dataset; full dataset includes >100 physical/ biogeochemical components observed at >100 wells over 50 years.*
Right: *Map of identified mixed contaminant plumes (shown with different colors).*

Bimolecular reactions: High-resolution datasets (with dimensions in X, Y and Time: 81 x 81 x 1000) are generated by solving anisotropic reaction-diffusion equations using a non-negative finite element formulation for different input parameters for perturbed vortex-based velocity fields. The input parameters are (1) a time-scale associated with flipping of the velocity, (2) a spatial-scale controlling small/large vortex structures of velocity, (3) a perturbation parameter of the vortex-based velocity, (4) anisotropic dispersion strength/contrast, and (5) molecular diffusion. The simulated reaction is a fast, irreversible bimolecular reaction A + B = C, where two species A and B react to form species C. More than 2000 model runs are performed varying the input model parameters. Without prior knowledge of the simulated processes, we apply SNTF to analyze all these simulation datasets to extract meaningful deconstruction of model outputs to discriminate between different physical processes impacting the reactants, their mixing, and the spatial distribution of the product C. The ML analysis allowed us to identify a series of additive temporal and spatial features that characterize mixing behavior. These features have physical meaning. An example result is presented in Fig.3. Here the model predicted concentration of C (left) are deconstructed into two temporal components (center and right) using STNF. The first temporal component (Fig.3) influence of anisotropy at the late stages of mixing. It defines how deviant is the anisotropic system with respect to that of pure isotropic diffusion case. It also describes how different is the anisotropic system from "the algebraic law of chemical kinetics"[10] at longer times. The second temporal component (Fig.3) is related to Finite-Time Lyapunov Exponent (FTLE). This component defines how fast the reactants are decaying over time. The average of C concentrations represented by the first component decline with a slope that gives the FTLE, which is also related to the exponential concentration decay parameter.

Conclusions

Our analyses demonstrate the applicability of our SNTF ML approach for identification of features in large datasets without prior knowledge about the underlying processes and mechanisms.

Acknowledgments

This research was funded by the Environmental Programs Directorate of the Los Alamos National Laboratory. VVV and DO were supported by the DiaMonD project (U.S. Department of Energy Office of Science, Grant #11145687). MKK and SK also thank the support of the LANL Early Career Award 20150693ECR. MKK also thanks the support of LANL Chick-Keller Postdoctoral Fellowship. VVV, BSA and DO were also supported by LANL LDRD grant 20180060.

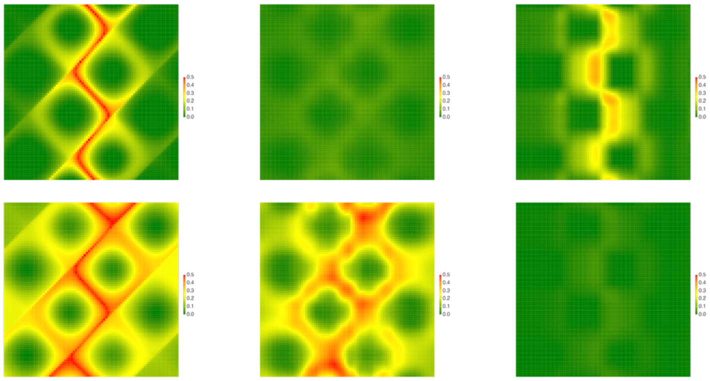

Fig.2: *Example deconstruction of the model predicted concentrations of C at dimensionless times 0.02 (top row) and 0.15 (bottom row). The model predictions (left) is decomposed into two temporal components (center and right) which when added approximately reproduce the model output. The core tensor* **W** *in this Tucker-3 reconstruction has dimensions (3 x 8 x 9).*

References

1 Kruskal, J. B. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications* **18**, 95-138 (1977).

2 Sidiropoulos, N. D. & Bro, R. On the uniqueness of multilinear decomposition of N-way arrays. *Journal of chemometrics* **14**, 229-239 (2000).

3 Zhang, Y., Zhou, G., Zhao, Q., Cichocki, A. & Wang, X. Fast nonnegative tensor factorization based on accelerated proximal gradient and low-rank approximation. *Neurocomputing* **198**, 148-154 (2016).

4 Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S.-i. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. (John Wiley & Sons, 2009).

5 Xu, Y. Alternating proximal gradient method for sparse nonnegative Tucker decomposition. *Mathematical Programming Computation* **7**, 39-70, doi:10.1007/s12532-014-0074-y (2015).

6 Xu, Y. & Yin, W. A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion. *SIAM Journal on Imaging Sciences* **6**, 1758-1789, doi:10.1137/120887795 (2013).

7 Vesselinov, V. V. *et al.* Data and Model-Driven Decision Support for Environmental Management of a Chromium Plume at Los Alamos National Laboratory (LANL). (2013).

8 Appelo, C. A. J. & Postma, D. *Geochemistry, groundwater and pollution*. (CRC press, 2004).

9 He, J., Hansen, S. & Vesselinov, V. V. Analysis of Hydrologic Time Series Reconstruction UncertaintyDue to Inverse Model InadequacyUsing the Laguerre Expansion Method. (Los Alamos National Laboratory (LANL), 2017).

10 Kotomin, E. & Kuzovkov, V. *Modern aspects of diffusion-controlled reactions: Cooperative phenomena in bimolecular processes*. Vol. 34 (Elsevier, 1996).