The Journal of Open Source Software

ESAT: Environmental Source Apportionment Toolkit Python package

Deron Smith [1], **Michael Cyterski** [1], **John M Johnston** [1], **Kurt Wolfe**[1], **and Rajbir Parmar** [1]

1 United States Environmental Protection Agency, Office of Research and Development, Center for Environmental Measurement and Modeling

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Source apportionment is an important tool in environmental science where sample or sensor data are often the product of many, often unknown, contributing sources. One technique for source apportionment is non-negative matrix factorization (NMF). Using NMF, source apportionment models estimate potential source profiles and contributions providing a cost-efficient method for further strategic data collection or modeling. An important aspect of modeling, especially environmental modeling, is the consideration of input data uncertainty and error quantification.

The EPA's Positive Matrix Factorization version 5 (PMF5)(EPA, 2014) application offers a source apportionment modeling and analysis workflow that has an active international user community. PMF5 was released in 2014 and is no longer supported; additionally the Multilinear Engine v2 (ME2) used in PMF5 is proprietary, with documentation existing only for the prior version ME1 (Paatero, 1999).

Statement of Need

The Environmental Source Apportionment Toolkit (ESAT) has been developed as a replacement to PMF5, and has been designed for increased flexibility, documentation and transparency.ESAT is an open-source Python package for flexible source apportionment workflows. The Python API and CLI of ESAT provides an object-oriented interface that can completely recreate the PMF5 workflow. The matrix factorization algorithms in ESAT have been written in Rust for optimization of the core math functionality. ESAT has two NMF algorithms for updating the profile and contribution matrices of the solution: least-squares NMF (LS-NMF) (Wang et al., 2006) and weighted-semi NMF (WS-NMF) (Ding et al., 2008) (Melo & Wainer, 2012).

ESAT provides a highly flexible API and CLI that can create source apportionment workflows like those found in PMF5, but can also create new workflows that allow for novel environmental research. ESAT was developed for environmental research, though it's not limited to that domain, as matrix factorization is used in many different fields; ESAT places no restriction on the types of input datasets.

Algorithms

The loss function used in ESAT, and PMF5, is a variation of squared-error loss, where data uncertainty is taken into consideration:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{V_{ij} - \sum_{k=1}^{K} W_{ik}H_{kj}}{U_{ij}}\right]^2$$

here V is the input data matrix of features (columns=M) by samples (rows=N), U is the uncertainty matrix of the input data matrix, W is the factor contribution matrix of samples by factors=k, H is the factor profile of factors by features.

The ESAT versions of NMF algorithms convert the uncertainty U into weights defined as $Uw = \frac{1}{U^2}$. The update equations for LS-NMF then become:

$$H_{t+1} = H_t \circ \frac{W_t(V \circ Uw)}{W_t((W_t H_t) \circ Uw)}$$

$$W_{t+1} = W_t \circ \frac{(V \circ Uw)H_{t+1}}{((W_t H_{t+1}) \circ Uw)H_{t+1}}$$

while the update equations for WS-NMF:

$$W_{t+1,i} = (H^T U w_i^d H)^{-1}(H^T U w_i^d V_i)$$

$$H_{t+1,i} = H_{t,i}\sqrt{\frac{((V^T Uw)W_{t+1})_i^+ + [H_t(W_{t+1}^T UwW)^-]_i}{((V^T Uw)W_{t+1})_i^- + [H_t(W_{t+1}^T UwW)^+]_i}}$$

where $W^- = \frac{(|W|-W)}{2.0}$ and $W^+ = \frac{(|W|+W)}{2.0}$.

Error Estimation

An important part of the source apportionment workflow is quantifying potential model error. ESAT offers the same error estimation methods that were available in PMF5 (Brown et al., 2015), but with flexibility for customization.

- Displacement Method (DISP): Quantify the error due to rotational ambiguity by evaluating the amount of change in source profile that correspond to specific changes in the loss.
- Bootstrap Method (BS): Quantify the error due to the order of the samples via block resampling.
- BS-DISP: Calculate the displacement error on a set of bootstrap datasets to quantify the combined error.

Simulator

ESAT contains a data simulator for generating synthetic profiles and contributions which allow for direct model evaluation. The synthetic profiles can either be randomly generated, use a previously defined set of profiles, or a combination of both. The random synthetic contributions can follow specified curves and value ranges. The ESAT model profiles can then be mapped to the known synthetic data for direct comparison and accuracy evaluations.

Acknowledgements

ESAT development has been funded by U.S. EPA. Mention of any trade names, products, or services does not convey, and should not be interpreted as conveying, official EPA approval, endorsement, or recommendation. The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of the US EPA.

References

Brown, S. G., Eberly, S., Paatero, P., & Norris, G. A. (2015). Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. *Science of the Total Environment*, *518*, 626–635.

Ding, C. H., Li, T., & Jordan, M. I. (2008). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55.

EPA, U. S. (2014). *Positive matrix factorization model for environmental data analyses*. https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses

Melo, E. V. de, & Wainer, J. (2012). *Semi-NMF and weighted semi-NMF algorithms comparison*.

Paatero, P. (1999). The multilinear engine—a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, *8*(4), 854–888.

Wang, G., Kossenkov, A. V., & Ochs, M. F. (2006). LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, *7*, 1–10.