DATA FUSION BY JOINT NON-NEGATIVE MATRIX FACTORIZATION FOR HYPOTHESIZING PSEUDO-CHEMISTRY USING BAYESIAN NETWORKS

Anjana Puliyanda*, Arno De Klerk, Zukui Li, Vinay Prasad
University of Alberta - Edmonton
114 Street 89 Avenue NW Edmonton AB T6G 2M7

Abstract Overview

A variety of spectroscopic measurement techniques are popularly used to obtain molecular level information of complex reacting systems. In this work FTIR and ^{1}HNMR spectroscopic measurements consisting of absorbances across wavenumbers/chemical-shifts are obtained over varying process conditions (temperature and residence time) during the vis-breaking of Cold Lake bitumen. They are jointly factorized using a weighted, robust non-negative matrix factorization (NMF) algorithm for data fusion. The missing data in spectral measurements are handled by imputing them with a weighting matrix in the objective function that is formulated to minimize the L_{21} norm between a matrix of spectral measurements and its factors, making it robust to outliers in data that would otherwise dominate the objective function because of squared errors if an L_2 norm was used instead. Additionally, the data fusion framework constrains the factors to be non-negative so that the decomposition is physically meaningful by complying with the Beer Lambert law for spectral data. Hence the factors can be physically interpreted as representing the spectral signatures and concentrations of a class of chemical compounds (pseudo-component). Unique information in the spectral signatures is obtained by incorporating a regularization term in the NMF objective that penalizes redundancy in the pseudo-component spectra. The NMF objective also incorporates another regularization term that penalizes overfitting of the spectral signatures that contain unique information about the pseudo-components. Hence, they are used to develop inferential models for monitoring the complex process of vis-breaking by developing pseudo-reaction networks that hypothesize chemical pathways. This is done using the probabilistic graphical framework of Bayesian networks that encode directed acyclic causal pathways among the nodes of random variables which are represented by the spectral signatures. This facilitates building causal inferential models to generate reaction network hypotheses from spectral measurements, to demystify the underlying chemical reaction pathways in complex reacting mixtures.

Keywords

Data fusion, Bayesian networks, Regularized joint non-negative matrix factorization, Pseudo-chemistry, Reaction pathway hypotheses

Introduction

In-line spectral analyzers are popularly used to obtain molecular-level information as they are fast, non-invasive, non-destructive, inexpensive and do not require sample preparation [1]. The process data from spectral analyzers are high dimensional, non-causal, non-full rank, noisy and have missing values [2]. Hence, this work focuses on using

* To whom all correspondence should be addressed, puliyand@ualberta.ca

machine learning models on process data to develop causal inferential models for monitoring complex processes; with an application to developing pseudo-reaction networks that hypothesize chemical pathways[3]. In this paper the spectral datasets from FTIR and ^{1}HNMR during the thermal upgrading of Cold Lake bitumen are mined to develop reaction pathways.

Non-negative matrix factorization (NMF) is a workhorse in signal and data analytics for composition deconvolution from spectral data, text, image or audio signals [4]. NMF identifies latent factors to a level of limited ambiguities thereby increasing interpretability as compared to alternate factorization methods like Singular Value Decomposition (SVD) and Independent Component Analysis (ICA) based on orthogonal and independent factor decompositions that are unconstrained [4]. As a parts-based representation of latent factors NMF has been a good algorithm for soft clustering [5].

Spectroscopic techniques like Electron Spectroscopy, Mass Spectroscopy, Raman, FTIR, ^{1}HNMR spectroscopy provide multi-dimensional information of chemical samples. These multi-dimensional datasets can be viewed as a linear mixing of weights (interpreted as concentrations) and reduced number of basis factors (interpreted as spectra of pseudo-components); the linear unmixing of which is done using NMF[6]. The absence of an NMF algorithm which integrates multi-view information i.e. JNMF of spectral data from different sources, is the prime motivation of this work.

Data fusion methods are often classified based on the stage at which the fusion is performed [7],[8]: (a) Early fusion: Sequential concatenation of data by neglecting modularity (b) Late fusion: Fusing the prediction model results obtained from each data source separately. (c) Intermediate fusion: Fusion propagated by features of each independent data source [9],[10], making the structure of the predictive model robust. A popular algorithm to implement intermediate fusion is the constrained simultaneous matrix factorization [7], which is tantamount to multi-view non-negative matrix factorization or the joint non-negative matrix factorization (JNMF).

JNMF was used for data fusion of multi-view gene interaction network data with sparse penalty regularization constraints [10]. Diverse-JNMF was used to penalize redundancy in the fusion of multi-view data by using an orthogonality regularizer between the multi-view basis factors [11]. Weighted-NMF where missing values are imputed by zero [12] and Robust-NMF where the objective function is based on minimizing the L_{21} loss function to robustly deal with outliers and noise [13] are other variants of NMF that are proposed to be extended to multi-view data for JNMF in this work.

Evidence of stage-based fusion of FTIR, ^{1}HNMR and Raman spectroscopic data having resulted in better crude characterization [14]; has motivated us to develop a more robust intermediate fusion algorithm for integrating multi-view spectral data to build models for hypotheses generation of chemical pathways. Since NMF has the advantage of being an interpretable factor decomposition method, utilizes optimization based matrix computation routines for its solution and has a scalable formulation for large-scale problems; this work focuses on using it as an unsupervised technique for the soft clustering of multiview spectral data into basis factors of the underlying latent objects weighted by a common parts-based matrix across all views. A projected optimal step gradient algorithm is developed to solve the Robust-Weighted-Joint-Non-negative matrix decomposition with regularizers that penalize redundancy and overfitting of basis factors from different views. The interaction among the associated basis factors of the latent objects across the views obtained from the above formulation is encoded in causal Bayesian Networks that hypothesize the chemical pathways among the latent factors, which in the physical sense correspond to chemically similar compounds i.e. pseudo-components and mathematically correspond to the rank of the matrix from a spectral data view.

Methods

The data matrix from FTIR and ^{1}HNMR measurements are denoted as X_1 and X_2, respectively. whereby the rows consist of the samples (m), while the columns are the spectral channels viz. wavenumbers (n_1) and chemical shifts (n_2). The objective of fusion is to combine both the spectral data matrices such that the following objective is minimized:

$$\min_{W,H_i \geq 0} F(W,H_i) = \sum_{i=1,2} P_i \circ ||X_i - WH_i||_{21} + \alpha||H_1 R_{12} H_2{}^T||_{21} + \sum_{i=1,2} \beta|| H_i R_i H_i{}^T||_{21} \quad (1)$$

P_i is a weighting matrix which imputes missing samples in X_i by a zero element to discount it in the JNMF factorization. R_{12} and R_i are correlation matrices to penalize redundancy and can be set to identity matrices to enforce orthogonal factorization. To make NMF robust to outliers a L_{21} norm is used instead of an L_2 norm [13].

The number of latent factors, the level at which the JNMF is implemented is determined as the minimum of the ranks of each spectral matrix i.e. r=min(rank(X_1), rank(X_2)). The mathematical matrix rank can be interpreted as the number of compound classes whose concentrations change. Rank for each X_i is determined using SVD of X_i to obtain as many principal components as the number of variables (n_i) [15].

The Robust Weighted Joint NMF algorithm is used to solve the non-convex objective function (eqn. 1) using the projected optimal gradient approach. The Non-negative Double SVD (NNDSVD) is used to initialize the decision variables W(mxr), H_1 (rxn_1), H_2(mxn_2) of the said dimensions. The gradients of the objective function with respect to the decision variables are given below:

$$\nabla F_W = \sum_{i=1,2} P_i \circ (WH_i - X_i)\, \mathrm{Diag}(X_i - WH_i) H_i{}^T \quad (2)$$

$$\nabla F_{H_i} = W^T (P_i \circ (WH_i - X_i)) \,\mathrm{Diag}(X_i - WH_i) + \alpha H_1 R_{12} H_2^T \,\mathrm{Diag}(H_1 R_{12} H_2^T) H_{j \neq i} R_{ji} + \beta H_i R_i H_i^T \,\mathrm{Diag}(H_i R_i H_i^T) H_i R_i \quad (3)$$

The spectral signatures obtained from this method are used to construct Bayesian Networks by designating them as random variables with a multinomial distribution (the hyperparameters have a Dirichlet distribution). A directed path exists between nodes if it maximizes the log likelihood, which is a function of the mutual information and entropy. This amounts to maximizing the Bayesian information criterion (BIC), which is the log likelihood of the entire network (pairwise directed edges between nodes) penalized by the complexity of the network (number of edges between nodes). Heuristic greedy search score-based methods are used to obtain the Bayesian networks i.e. the directed acyclic graph (DAG) encoding causal relationships among the factors obtained from JNMF.

Results and Discussion

The NMF objective function in eqn. 1 was solved using the method of projected optimal gradients over a range of α,β values [10^{-3},10^{-2},10^{-1},1,0,10,10^{2},10^{3}]. The values of α=1,β=10^{-1} resulted in lowest reconstruction error of the matrices X_i from their factors W,H_i. The spectra H_i where i=1 corresponds to the FTIR spectra; i=2 corresponds to the ^{1}HNMR spectra of the pseudo-components, are shown in Fig.1. It can be seen that the ^{1}HNMR profiles of pseudo-components 2 and 3 are insignificant owing to the regularization constraint that penalizes redundancy between the corresponding FTIR spectra.

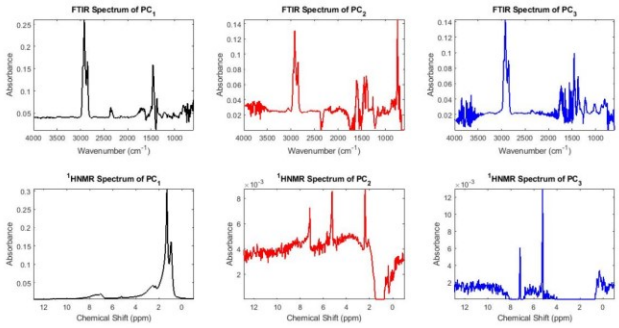

Figure 1. FTIR and ^{1}HNMR spectra

The spectral signatures reveal the following: pseudo-component 1 (PC_1) mainly consists of carbonyl groups and cycloalkanes; PC_2 consists of polyaromatics, alkoxy groups, phenols, alkenes; PC_3 consists of aromatics, alkanes and condensed products; and PC_4 consists of phenols, acyls and condensed aromatics.

It can therefore be hypothesized from the Bayesian network structure in Fig. 2 that the underlying chemical reaction pathways during the vis-breaking of bitumen aim at obtaining more saturated end products through the free radical mechanism of hydrogen radical addition. However, the longer chain aliphatics crack to give more condensed polyaromatic products which are undesirable even as the end products are more aliphatic (alkanes and olefins).

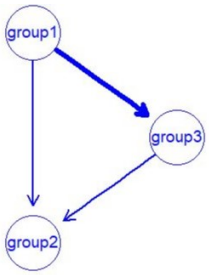

Figure 2. Bayesian Network

This work facilitates jointly mining spectral measurements in the framework of constrained data fusion to make the factors physically interpretable and representative of unique information, to implement a first pass at building causal inferential models to generate reaction network hypotheses from process data (spectral measurements).

References

[1] J. Felten, H. Hall, J. Jaumot, R. Tauler, A. De Juan, A. Gorzsás, *Nat. Protoc.*, 10(2), 217–240, 2015.

[2] M.A. Nemeth, *Technometrics*, 45(4), 362–362, 2003.

[3] D. T. Tefera, L. M. Yañez Jaramillo, R. Ranjan, C. Li, A. De Klerk, V. Prasad, *Ind. Eng. Chem. Res.*, 56(8), 1961–1970, 2017.

[4] X. Fu, K. Huang, N. D. Sidiropoulos, W. Ma, *arXiv preprint arXiv:1803.01257*, (2018).

[5] J. Wang, F. Tian, W. Liu, X. Wang, W. Zhang, K. Yamanishi, *Proc. Int. Joint Conf. Artificial Intelligence*, 2776–2782, 2018.

[6] R. Kannan, A.V. Ievlev, N. Laanait, M.A. Ziatdinov, R.K. Vasudevan, S. Jesse, S.V. Kalinin, *Adv. Struct. Chem. Imag.*, 4(1), 6, 2018.

[7] M. Žitnik, B. Zupan, *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(1), 41–53, 2015.

[8] Y. Zheng, *IEEE Trans. Big Data*, 1(1), 16–34, 2015.

[9] A.K. Smilde, I. Måge, T. Naes, T. Hankemeier, M.A. Lips, H.A. Kiers, E. Acar, R. Bro, *J. Chemom.*, 31(7), 1–20, 2017.

[10] L. Zhang, S. Zhang, *arXiv preprint arXiv:1707.08183*, 2017.

[11] J. Wang, F. Tian, H. Yu, C. H. Liu, K. Zhan, X. Wang, *IEEE Trans. Cybern.*, 48(9), 2620–2632, 2018.

[12] Y. Kim S. Choi, *IEEE Int. Conf. Acoustics, Speech Signal Proc.*, 1541–1544, 2009.

[13] D. Kong, C. Ding, H. Huang, Proc. ACM Int. Conf. Info. Know. Mgmt., 673-682, 2011.

[14] T. I. Dearing, W. J. Thompson, C. E. Rechsteiner, B. J. Marquardt, *Appl. Spectrosc.*, 65(2), 181–186, 2011.

[15] D. T. Tefera, A. Agrawal, L. M. Yañez Jaramillo, A. de Klerk, V. Prasad, *Ind. Eng. Chem. Res.*, 56(38), 10756–10769, 2017.