Received: 12 June 2018 | Revised: 28 December 2018 | Accepted: 5 January 2019 | Published on: 19 February 2019

DOI: 10.1002/sam.11407

RESEARCH ARTICLE

WILEY

Nonnegative tensor decomposition with custom clustering for microphase separation of block copolymers

Boian S. Alexandrov[1] | **Valentin G. Stanev**[2] | **Velimir V. Vesselinov**[3] | **Kim Ø. Rasmussen**[1]

[1]Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico
[2]Department of Materials Science and Engineering, University of Maryland, College Park, Maryland
[3]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico

Correspondence
Boian S. Alexandrov, Theoretical Division, MS-B221, Los Alamos National Laboratory, PO Box 1663, Los Alamos NM 87545.
Email: boian@lanl.gov

Funding Information
LANL LDRD, 20180060.

High-dimensional datasets are becoming ubiquitous in many applications and therefore unsupervised tensor methods to interrogate them are needed. Here, we report a new unsupervised machine learning (ML) approach (NTFk) based on nonnegative tensor factorization integrated with a custom k-means clustering. We demonstrate the ability of NTFk to extracting temporal and spatial features of phase separation of copolymers as they are modeled by self-consistent field theory. Microphase separation of block copolymers has been extensively studied both experimentally and theoretically. However, the interpretation of computer simulations and/or experimental data, representing temporal and spatial changes of molecular species concentration is still a challenging task. Thus, extracting the phase diagram from simulations or experimental data as well as the interpretation of data requires discernment of the model/experimental parameters (such as, temperature, concentrations, the number of molecular species and the interaction between species) impact on the microphase separation process. An attractive and unique aspect of the introduced ML method is that it ensures the nonnegativity of the extracted latent features. Nonnegativity is an essential constraint needed to obtain interpretable and sparse latent features that are parts-based representation of the data. The custom clustering in NTFk serves to estimate the number of latent features in the data.

KEYWORDS

dimension reduction, feature extraction, nonnegative tensor factorization, phase separation, unsupervised learning

1 | INTRODUCTION

The emerging collections of large and distinct high-dimensional datasets increase the interest in factor analysis based on tensor decomposition [1]. These collected datasets include only directly observable quantities, while the underlying processes are either too complex and cannot be observed directly or are completely unknown. These processes are called latent variables or latent features [2]. Extracting latent variables permits reduction of the large number of directly observable quantities to a smaller set of latent features, where each observable quantity in the data is expressed as a linear or multilinear combination of the extracted latent features. Tensor decompositions are leveraged for estimating latent variables [3], pattern recognition [4] subspace learning, unsupervised separation of unknown mixtures of signals [5] and many other applications [6].

Here we introduce a new approach based on nonnegative tensor factorization (NTF) that integrates NTF with custom k-means clustering to estimate the number of nonnegative latent features. We demonstrate the application of this approach, called NTFk, for a successful identification of the morphologies and phase transitions in a molecular self-assembly.

Molecular self-assembly is a spontaneous association of molecules under equilibrium conditions into stable,

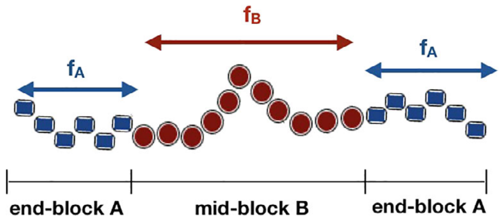

FIGURE 1 Illustration of symmetric ABA triblock copolymer: a macromolecule composed of three strongly bonded alternating groups (blocks) of molecular units (copolymers) of two different molecular species. f_A and f_B denote the contour length of the A and B block, respectively

structurally well-defined aggregates joined by weak bonds. Molecular self-assembly is ubiquitous throughout soft condensed matter and biological materials and it is becoming an important tool in nanoscience and the design and manufacturing of complex functional materials. A significant challenge working with these materials and techniques is to properly identify and characterize the transition points and conditions as well as the emerging morphologies. This is true for experimental situation but also for modeling approaches based on molecular dynamics and density functional techniques. Block copolymers, which are macromolecules composed of strongly bonded alternating groups (blocks) of repeated molecular units (copolymers) of different molecular species, form a simple and prototypical self-assembling system for which continuum modeling techniques such as field theory are well developed.

To demonstrate the applicability of NTFk in this field we are applying our technique to modeling data arising from well-established field theoretic methods [7] and a simple system of triblock copolymers (Figure 1).

2 | NONNEGATIVE TENSOR FACTORIZATION

The combined procedure of dimension reduction and latent features extraction is crucial for data mining and is a subject of factor analysis [8]. Factor analysis includes a group of unsupervised machine learning (ML) techniques based on blind source separation (BSS) [9]. Classical BSS utilizes matrix factorizations, including: principle component analysis (PCA) [10], independent component analysis (ICA) [11], singular value decomposition (SVD) [12] and nonnegative matrix factorization (NMF) [13], which form a class of unsupervised ML methods that are instrumental for model-free latent features extraction.

The matrix factorization methods, although extremely efficient, are inherently deficient for examining high-dimensional datasets, that is, tensor datasets. The tensor datasets are natural extensions of the matrix datasets needed when there are more than two dimensions in the data because any attempts to represent and analyze a high-dimensional dataset as a matrix will neglect the cross-correlations among the different dimensions.

Two main classical tensor factorization methods are: canonical polyadic decomposition (CPD) [14–16] and Tucker decomposition (TD) [17,18] (Figure 2). Nonnegative CPD/TD of a three-dimensional tensor $\boldsymbol{X}$ can be derived by nonconvex constraint minimization of the following objective function, $\boldsymbol{O}$,

$$\boldsymbol{O} = \left\| \boldsymbol{X}_{n,m,l} - \sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3} \boldsymbol{G}_{r_1,r_2,r_3} \boldsymbol{A}_{r_1,n} \boldsymbol{B}_{r_2,m} \boldsymbol{C}_{r_3,l} \right\|_F^2,$$
$$\boldsymbol{G}_{r_1,r_2,r_3} \geq \boldsymbol{0}; \boldsymbol{A}_{r_1,n} \geq \boldsymbol{0}; \boldsymbol{B}_{r_2,m} \geq \boldsymbol{0}; \boldsymbol{C}_{r_3,l} \geq \boldsymbol{0}, \qquad (1)$$

where $\| \ldots \|_F$ is the Frobenius norm. In the case of CPD (also called CANDECOMP/PARAFAC model [19]) the core-tensor $\boldsymbol{G}$ in Equation (1) is superdiagonal, that is, $\boldsymbol{G}_{r_1,r_2,r_3} = 1$; if $\boldsymbol{r_1} = \boldsymbol{r_2} = \boldsymbol{r_3}$ and $\boldsymbol{G}_{r_1,r_2,r_3} = 0$ in all other cases. For CPD, the objective function, $\boldsymbol{O}$, can be written as,

$$\boldsymbol{O} = \left\| \boldsymbol{X}_{n,m,l} - \sum_{k=1}^{R} \boldsymbol{A}_{k,n} \boldsymbol{B}_{k,m} \boldsymbol{C}_{k,l} \right\|_F^2,$$
$$\boldsymbol{A}_{\mathrm{k},n} \geq \boldsymbol{0}; \boldsymbol{B}_{\mathrm{k},m} \geq \boldsymbol{0}; \boldsymbol{C}_{\mathrm{k},l} \geq \boldsymbol{0}, \qquad (2)$$

where $R_1 = R_2 = R_3 = R$ (R is called the rank of $\boldsymbol{X}$), while for TD, in general, $R_1 \neq R_2 \neq R_3$ and (R_1, R_2, R_3) is called multilinear rank of $\boldsymbol{X}$.

When the considered data $\boldsymbol{X}$ is inherently nonnegative the choice of the nonnegative constraints is natural. Many types of data, for example, density, energy, spectral power, population, etc., are naturally nonnegative and the extracted features will lose the physical meaning if the nonnegativity is not preserved. Importantly, the nonnegativity is crucial for the extraction of interpretable and sparse latent features. Indeed, when the factorization produces only nonnegative values the extracted features can only be added and no subtractions are allowed. Hence, reproducing the data only by combinations of nonnegative latent features requires these features to be (a) sparse and (b) parts-based representations of the original data, which is making the extracted features easy to interpret [20]. Thus, the nonnegative factorization produces readily explainable features, which in turn facilitate discoveries of new causal structures and mechanisms hidden in the data without prior assumptions. Additional advantage of the nonnegative factorization is that it also works with data where the latent features are not statistically independent but can be even partially correlated [6]. One of the limitations of nonnegative factorization is that, unlike PCA, SVD, or ICA, it requires *prior* knowledge of the number of the latent features. Thus, the nonnegative CPD requires *prior* estimation of the nonnegative rank, R, of the tensor, that is, the minimal number of rank-1 nonnegative tensors whose sum accurately reproduces the data. It is well known that CPD is unique [21] up to a scaling and permutation of the factors, and that CPD represents the tensor-data, $\boldsymbol{X}$, with the smallest number of parameters. However, CPD is difficult to achieve because, in the general case, computing the rank of a tensor (ie, the number of the latent features) is known to be a

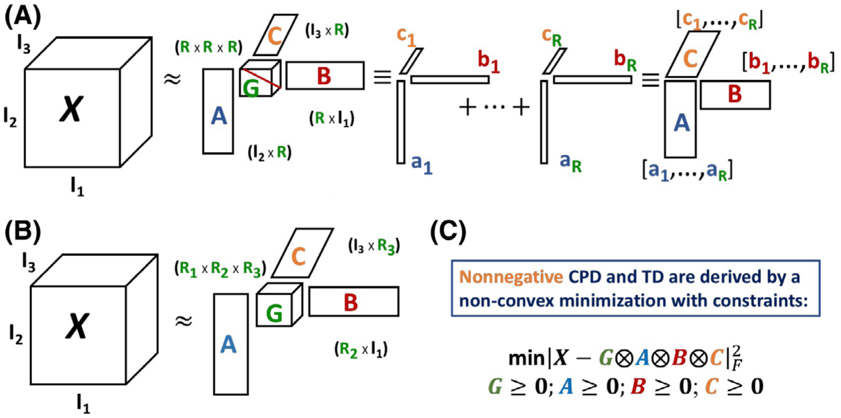

FIGURE 2 Basic tensor factorizations illustrated for a three-dimensional tensor $\boldsymbol{X}$: (A) canonic polyadic decomposition (CPD) and (B) Tucker decomposition (TD). Both factorizations decomposed the three-dimensional data-tensor $\boldsymbol{X}$ with size (I_1, I_2, I_3) into three matrix factors $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and a small core-tensor $\boldsymbol{G}$. In the CPD case, the core-tensor $\boldsymbol{G}$ is a superdiagonal cube (it has zeroes everywhere except on the main diagonal), denoted by the red line in $\boldsymbol{G}$ with size (R, R, R), and R is the rank of the tensor $\boldsymbol{X}$. In the TD case, the core-tensor $\boldsymbol{G}$ is of size (R_1, R_2, R_3), which is the multirank of $\boldsymbol{X}$. Panel C) represents the nonconvex minimization with nonnegative constraints, needed to obtain CPD/TD

nondeterministic polynomial-time (NP) hard problem [22]. Importantly, the nonnegative tensor decomposition is almost always unique [23] and Lim and Comon [24] proved that the best approximation of the nonnegative rank of a nonnegative tensor always exists, which facilitates the nonnegative CP decompositions.

The main difference between CPD and TD is the presence of nondiagonal elements and the different dimensions (R_1, R_2, R_3) of the core-tensor in TD, which allows column vectors of factor matrices to interact with each other in order to reconstruct the original data, $\boldsymbol{X}$. In TD, R_1 is the dimension of the subspace spanned by mode-1 fibers (fiber is a subtensor obtained by fixing all but one index/mode in the array X_{ijk}), R_2 is the dimension of the subspace spanned by mode-2, and R_3 is the dimension of the subspace spanned by mode-3 fibers. Unlike CPD, TD is not unique and offers a poorer compression of the data; however, TD is easier to achieve (no need to know the rank of the tensor) even with additional sparsity constraints in the minimization. In general, the Tucker core-tensor cannot be diagonalized, and the rank of the tensor is not bounded by TD's multilinear rank. There are also other types of tensor factorizations but they all are related to the CPD and the Tucker factorizations [1].

3 | NTF WITH CUSTOM K-MEANS CLUSTERING: NTFK

If the rank R of the data-tensor (ie, the number of the latent features in the data) is known, then the best solution of the minimization of the objective function $\boldsymbol{O}$ in Equation (1) is the solution of the factorization. Unfortunately, the rank of the tensor is typically unknown and it has to be estimated solely from the data. Since it is known that the best approximation of the rank of the nonnegative tensor $\boldsymbol{X}$ exists [24], a naive approach would be to explore all possible solutions applying the nonlinear minimization (2) for a range of possible ranks, and then to use the most accurate solution (ie, the solution with the smallest reconstruction error) as an estimate of the rank R. However, this is obviously a flawed approach—over-fitting will certainly lead to overestimation of the number of latent features; more free parameters will generally lead to a better reconstruction, irrespective of how close the estimated number of features is to the real one. Instead, to determine the optimal approximation of the rank of a nonnegative tensor, we utilize a custom clustering algorithm, following the idea used to estimate the unknown number of latent features in NMF [25].

The original NMF algorithm also requires *prior* knowledge of the number of the latent features. It was demonstrated that the number of the latent features can be estimated based on the reproducibility and robustness of the solution of NMF minimization. This approach has been introduced to decompose the largest available dataset of human cancer genomes [26], and then extended for decomposition of physical pressure transients [27]. Although heuristic in nature, the method, called NMFk, has proven to be accurate and reliable in various problems [28–33], and also has the important practical advantage that it is relatively easy to implement and use.

In NTFk-CPD, which we introduce here, we use the robustness of the extracted factor matrices and the accuracy of the decomposition to determine the optimal number of the nonnegative latent features in the data $\boldsymbol{X}$. Specifically, to estimate the rank of $\boldsymbol{X}$, we perform M sets of CPD-minimizations, Equation (2). Each solution set contains $N \sim 100$ CPD minimizations with random initial guesses for the unknown parameters. Each solution set is with a fixed rank R, and for the different sets we have, $R = 1, 2, \ldots, M$. We concatenated the columns of the extracted factors $\boldsymbol{A^R}$, $\boldsymbol{B^R}$, and $\boldsymbol{C^R}$ (Figure 2) in each set (with same rank R) and obtain, $\boldsymbol{H^R} = ([\boldsymbol{A_1^R}, \boldsymbol{B_1^R}, \boldsymbol{C_1^R}]; [\boldsymbol{A_2^R}, \boldsymbol{B_2^R}, \boldsymbol{C_2^R}]; \ldots ; [\boldsymbol{A_N^R}, \boldsymbol{B_N^R}, \boldsymbol{C_N^R}];)$, where R is the selected rank. Next, we apply a custom k-means clustering algorithm to cluster the columns of $\boldsymbol{H^R}$.

This custom clustering is based on k-means clustering with R clusters (each solution contains R features) but with the following constraint: the number of the elements in each of the clusters needs to be the same. For example, with $N = 100$ minimizations each one of the R clusters has to contain exactly 100 elements. This constraint has to be enforced since each solution of the 100 nonnegative minimizations contains exactly the same number, R, of different features. Thus, a fixed-size k-means clustering is needed since each of the 100 minimizations (with the same number of features) contributes exactly one element to each of the R clusters with features. During the clustering, the similarity is measured by the cosine distance, which is naturally to use in positive spaces where the angle between the feature vectors is bounded in the interval $(0, \pi/2)$.

Finally, the optimal number of latent features (ie, the nonnegative rank of the tensor, $\boldsymbol{X}$) is evaluated by comparing the quality of the derived clusters (obtained for different ranks) with the accuracy of the minimization. The quality of the clusters is estimated by their average Silhouettes values that measure the similarity between an element and the other elements of its own cluster, compared to the similarity to the elements of the other clusters [34]. The accuracy of the minimization, $\boldsymbol{r}$, is calculated based on the relative Frobenius norm: $\boldsymbol{r} = \boldsymbol{X} - \widetilde{\boldsymbol{X}}_{\boldsymbol{F}} / \boldsymbol{X}_{\boldsymbol{F}}$, where $\widetilde{\boldsymbol{X}}$ is the reconstruction of the data. The combination of these two criteria is easy to understand intuitively: The optimal number of clusters means that the Silhouette value has to be close to one, that is, the clusters containing the optimal number of features have to be well separated and with a good cohesion, while the set of these optimal features has to reconstruct the initial data well. Note that, for solutions with number of clusters, R, less than the actual number of latent features we expect the clustering to be good (ie, with an average Silhouette width close to 1), because several of the actual features could be combined to produce one reproducible 'super-cluster'; however, the reconstruction error will be high, due to the model being too constrained (with too few degrees of freedom), and thus on the under-fitting side. In the opposite limit of over-fitting, when the number of clusters R exceeds the actual number of patterns, the average reconstruction error could be small—each solution reconstructs the observation matrix very well, but the solutions will not be well-clustered (with an average Silhouette substantially less than 1), since there is no unique way to reconstruct $\boldsymbol{X}$ with more than the actual number of features, and no well-separated clusters will be formed. Thus, our best estimate for the optimal number of latent features R is given by the value of R that optimizes both these metrics simultaneously. Finally, after determining R, we use the centroids of the final R clusters to represent the final robust latent features, that is, the columns of the factor matrices. It is clear, that the same protocol can be applied to high-dimensional data with $d > 3$.

In NTFk-TD, to identify the optimal number of features, we also perform M TD runs with random initial guesses for the unknown parameters, and then cluster (as for NTFk-CPD) the resulting set of concatenated columns of the factor matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, with different sizes R_1, R_2, R_3, and determine the optimal multirank, that is the dimension of the corresponding subspaces by comparing the quality of the reconstruction and the average Silhouettes of the derived clusters.

4 | SELF-CONSISTENT FIELD THEORY

Copolymer melts are made of chemically complicated, large molecules and are characterized by a variety of competing length scales—size of the individual molecular units vs the extension of entire macromolecule molecules (Figure 1. A full treatment in atomistic detail is in most cases out of computational reach. On the other hand, the apparent complexity of the systems in fact contributes to a simplification of the physics on a coarser than atomistic level. Because of the large number of possible interactions between molecules, microscopic details average out to a large extent. A few characteristic attributes of the molecules are often responsible for the main features of a substance. This has motivated the study of idealized simplified models, which account only for the main properties of the molecules and absorb the microscopic details in a few, effective parameters. A second important point is that dense macromolecular systems are often unusually well described by mean-field approximations. As large molecules interact with many others, the effective interaction range in the limit of large molecules is very long, and the critical region in which concentration fluctuations become important is very small as a result. Self-consistent field theory (SCFT) [35–38] is based on these two observations and has proven to be a very successful description of block copolymer melts. Here we used the implementation of SCFT for symmetric triblock copolymers ABA (see Figure 1) as described in Ref. 39. The morphological phase diagram for this type of triblock copolymer can be fully parameterized by the normalized block contour length$\boldsymbol{f_A} = (1 - \boldsymbol{f_B})/2$ and the incompatibility, χ, between A and B blocks. The incompatibility χ is the strength of the interaction between A and B blocks, it depends on the polymer species and scales with temperature roughly as, $\chi \sim 1/T$.

Block copolymers provide a paradigm for molecular self-assembly in soft condensed matter. Unfavorable interactions between distinct blocks may lead—above a threshold value of the parameter quantifying blocks' incompatibility—to segregation into nano- and micron-scale domains, producing well-ordered periodic structures [40–42]. The morphology of the resulting self-assembly arises from the interplay between the incompatibility of unlike blocks and forces resulting from local configurational entropy considerations. The subtle counterbalance of these competing physical mechanisms is sensitive to the composition of the macromolecules. By changing, for example, the faction of

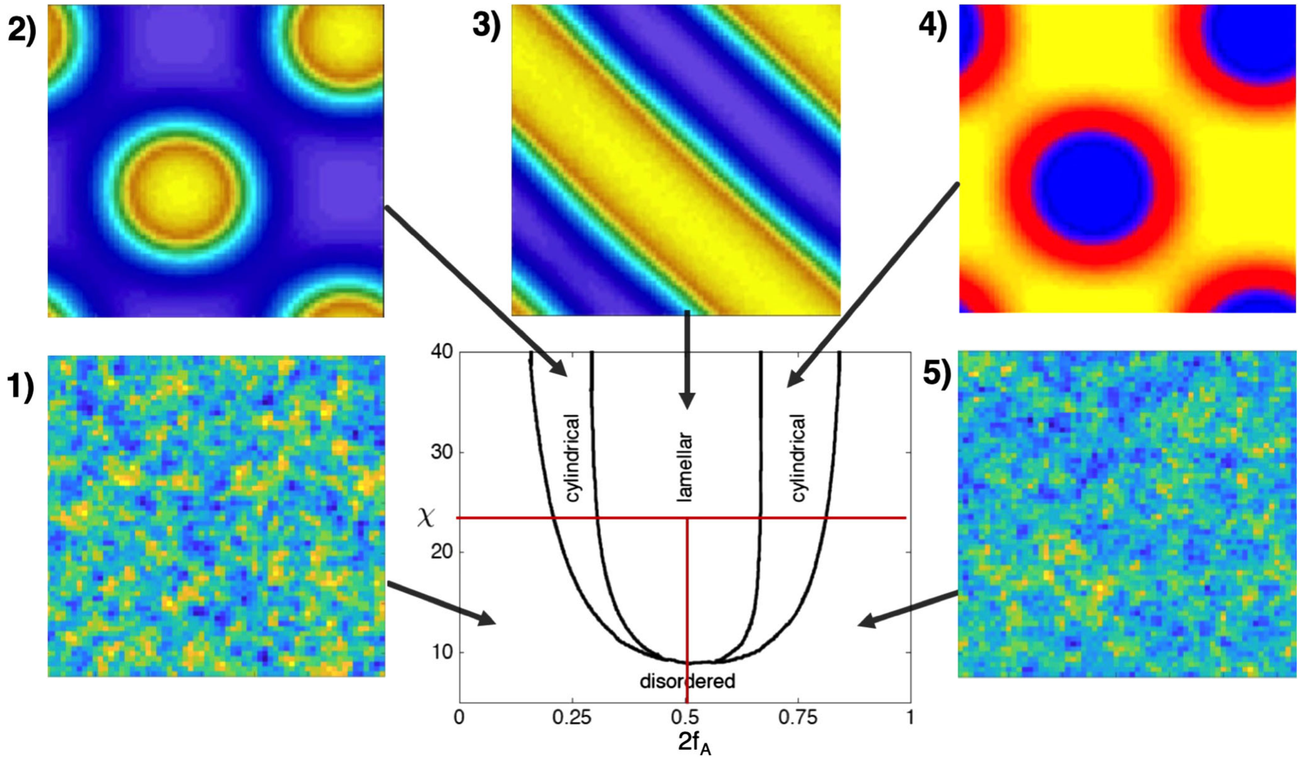

FIGURE 3 Sketch of morphological phase diagram for an ABA copolymer in two spatial dimensions (in the middle), x-axis represents the parameter $2f_A$, while y-axis represents the parameter χ. The vertical red line denotes the cross section at $2f_A = 0.50$, while the horizontal red line denotes the cross section at $\chi = 25$. The panels (1), (2), (3), (4), and (5) show the specific phases, that is: the disorder, cylindrical, lamellar, inverted cylindrical, and again disorder, in the phase diagram. The colors in these panels show the spatial density distribution $\boldsymbol{\phi}_A$ of the A block. Yellow color indicates high density while blue color indicates low density

molecular repeats forming one type of block in the chain, phase transitions (defined by nonanalyticities in the main thermodynamic functions) may take place among different ordered structures. In particular, the system transitions from one phase to another, and for example, it can pass from a disordered to an ordered phase. Any quantity that characterizes the symmetry of the system (or simply the system) that exists in one of the phases but vanishes in the other phase can be used for description of the phase transition and then it is called order parameter.

A sketch of the phase diagram as produced by SCFT in two spatial dimensions is given in Figure 2 where three morphological phases are displayed: (a) *disordered*, which occurs when the incompatibility, χ, between blocks is insufficient to cause segregation, (b) *cylindrical*, where one of the molecular blocks aggregate into rod like structures embedded in a matrix formed by the other molecular blocks—this transition occurs when one block is much shorter than the other, and finally (c) *lamellar*, where each of the molecular blocks aggregate into sheet-like structures, this transition occurs when the blocks are of similar length. Examples produced by SCFT of the lamellar and cylindrical phases are given in the insets of Figure 2. It is worth noticing that because of the structure of the triblock copolymer the phase diagram is not quite symmetric around $2f_A = 0.5$. When B is the minority component, the cylindrical phase occupies a larger portion of the phase space than A component. Also, the phase boundaries of the lamellar phase are slightly shifted toward small $\boldsymbol{f}_A$.

As the order parameter we choose the density of the A block. The A block density is obtained as output of SCFT simulations, as it is represented in in the panels (1), (2), (3), (4) and (5) of Figure 3, and this representation is what we used as an input for the Nonnegative Tensor Factor analysis.

5 | NTFk ANALYSIS OF SCFT SIMULATION DATA

5.1 | SCFT input data

We apply NTFk to two data-tensors representing the order parameter, $\boldsymbol{\Delta}(\boldsymbol{f}_A, \boldsymbol{x}, \boldsymbol{y}, \chi)$ of a system of copolymers as a function of: (a) χ, at a fixed length $\boldsymbol{f}_A$, and (b) as a function of $\boldsymbol{f}_A$, at a fixed χ. As order parameter, $\boldsymbol{\Delta}(\boldsymbol{f}_A, \boldsymbol{x}, \boldsymbol{y}, \chi)$, we choose the density of the A block. Thus, $\boldsymbol{\Delta}(\boldsymbol{f}_A, \boldsymbol{x}, \boldsymbol{y}, \chi) \equiv \boldsymbol{\phi}_A(\boldsymbol{f}_A, \boldsymbol{x}, \boldsymbol{y}, \chi)$, where $\boldsymbol{\phi}_A(\boldsymbol{f}_A, \boldsymbol{x}, \boldsymbol{y}, \chi)$ is the density distribution of block A polymers and it is generated by the SCFT simulations described in the previous section. For a fixed length $\boldsymbol{f}_A$, the output of SCFT simulations is a data-tensor $\boldsymbol{\Delta}_{nml}$ with three dimensions: $\boldsymbol{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$ with size $(11 \times 64 \times 64)$, while for a fixed χ, the output is, $\widetilde{\boldsymbol{\Delta}}(\boldsymbol{f}_A, \boldsymbol{x}, \boldsymbol{y}, \chi)$ with size $(201 \times 64 \times 64)$. Each cell of these three-dimensional tensors contains the value of the order parameter, $\boldsymbol{\Delta}$, at a specific $\boldsymbol{f}_A$, at a specific χ, and at a point with coordinates $(\boldsymbol{x}_m, y_l)$.

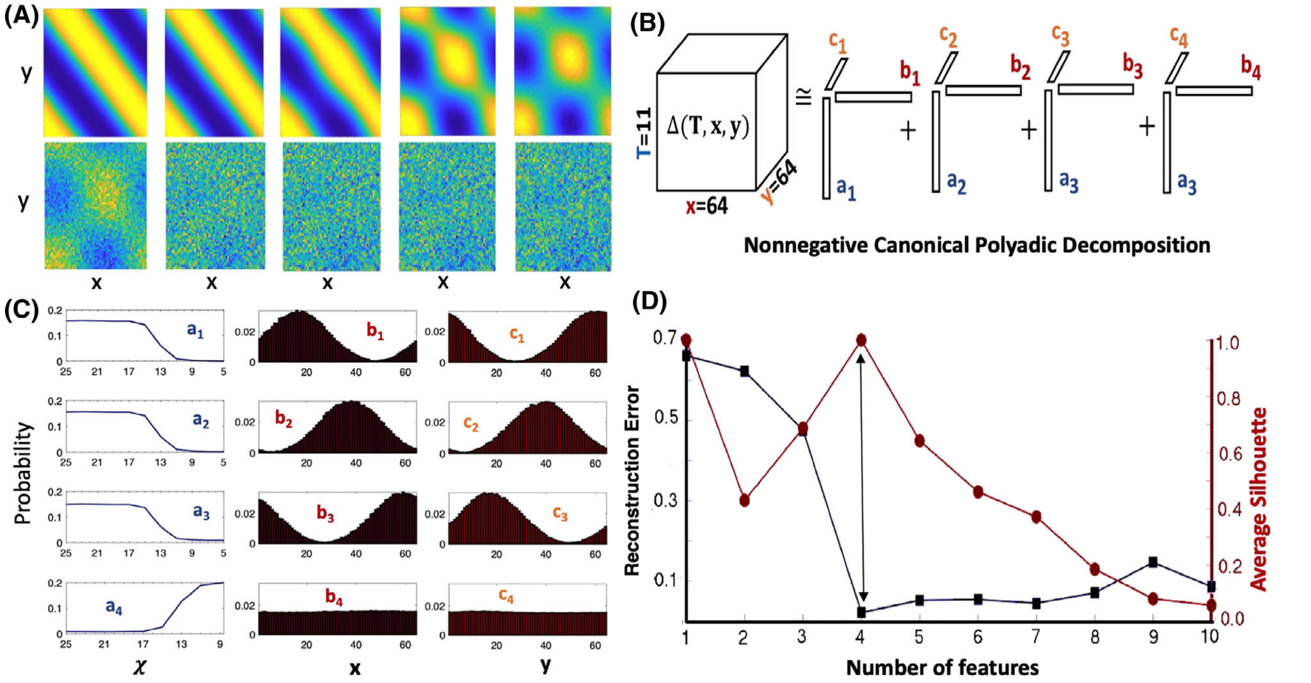

FIGURE 4 The results of NTFk-CPD: (A) 10 2D slices ($x = 1, ..., 64$; $y = 1, ..., 64$) representing the self-consistent field theory simulations with decreasing value of χ and fixed $\boldsymbol{f_A} = 0.5$, showing the evolution of the microphase separation of triblock copolymers as a function of inverse temperature (along the vertical line in Figure 3). (B) CPD of the data $\boldsymbol{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$ with rank $R = 4$. (C) The plot represents the four latent features, that is, the columns of each of the factor matrices $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$ (color coded as in (B)). (D) Silhouette-Reconstruction criterium for determination the number of the latent features in $\boldsymbol{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$. The double arrow depicts that the optimal number of the latent features is $R = 4$

5.2 | NTFk-CPD feature extraction of the order parameter $\boldsymbol{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$

The nonnegative CPD of the three-dimensional data-tensor representing the evolution of the order parameter $\boldsymbol{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$, with χ on a two-dimensional (x, y) lattice, and it is given by,

$$\boldsymbol{\Delta}_{n,m,l} = \sum_{k=1}^{R} \boldsymbol{A}_{k,n} \boldsymbol{B}_{k,m} \boldsymbol{C}_{k,l} + \boldsymbol{\varepsilon}_{n,m,l} \tag{3}$$

where $\boldsymbol{A} = [a_1, ..., a_R]$; $\boldsymbol{B} = [b_1, ..., b_R]$; $\boldsymbol{C} = [c_1, ..., c_R]$ are the nonnegative CPD factor matrices (Figure 2A, while $\boldsymbol{\varepsilon}_{n,m,l}$ is normally distributed unbiased error of the approximation. Equation (3) can be also represented as:

$$\boldsymbol{\Delta}_{n,m,l} = \widetilde{\boldsymbol{\Delta}}_{n,m,l} + \boldsymbol{\varepsilon}_{n,m,l} \tag{4}$$

where $\widetilde{\boldsymbol{\Delta}}$ ($\widetilde{\boldsymbol{\Delta}} \in \mathbb{R}_{\geq 0}^{N \times M \times L}$) is the CPD estimate of $\boldsymbol{\Delta}$.

CPD includes:

- An unknown matrix $\mathbf{A}$ ($\mathbf{A} \in \mathbb{R}_{\geq 0}^{N \times R}$) representing the changes of the order parameter $\boldsymbol{\Delta}$ with χ (the χ-component);
- An unknown matrix $\mathbf{B}$ ($\mathbf{B} \in \mathbb{R}_{\geq 0}^{M \times R}$) representing the changes of the order parameter $\boldsymbol{\Delta}$ in $\boldsymbol{x}$-direction (the $\boldsymbol{x}$-component);
- An unknown matrix $\mathbf{C}$ ($\mathbf{C} \in \mathbb{R}_{\geq 0}^{L \times R}$) representing the changes of the order parameter $\boldsymbol{\Delta}$ in $\boldsymbol{y}$-direction (the $\boldsymbol{y}$-component);

Here, $\mathbb{R}_{\geq 0}$ denotes the set of nonnegative real numbers $\mathbb{R}_{\geq 0}$ $\{x \in \mathbb{R}; x \geq 0\}$. Additionally, $\boldsymbol{\varepsilon}$ ($\boldsymbol{\varepsilon} \in \mathbb{R}_{\geq 0}^{N \times M \times L}$) denotes the unknown discrepancy between the original order parameter $\boldsymbol{\Delta}$ and the estimate $\widetilde{\boldsymbol{\Delta}}$.

To extract the unknown factor matrices **A**, **B**, and **C**, we utilize the block coordinate descent method for regularized multiconvex optimization algorithm introduced in [43,44]. Mathematically, the solution of the nonnegative CPD is given by the minimization with nonnegative constraints,

$$\boldsymbol{O} = \left\| \Delta - \sum_{k=1}^{R} \lambda_k \boldsymbol{a}_k(\chi) \otimes \boldsymbol{b}_k(\boldsymbol{x}) \otimes \boldsymbol{c}_k(\boldsymbol{y}) \right\|_F^2,$$
$$\lambda_k \geq \mathbf{0}; \boldsymbol{a}_k(\chi) \geq \mathbf{0}; \boldsymbol{b}_k(\boldsymbol{x}) \geq \mathbf{0}; \boldsymbol{c}_k(\boldsymbol{y}) \geq \mathbf{0}; \tag{5}$$

where λ_k are the weights of the normalized rank-1 tensors $\boldsymbol{a}_k(\chi) \otimes \boldsymbol{b}_k(\boldsymbol{x}) \otimes \boldsymbol{c}_k(\boldsymbol{y})$. The results are presented in Figure 4. Figure 4, panel A shows the data that the NTFk-CPD algorithm is applied to. The data consist of 11 SCFT simulation outputs for (χ = 25, 23, 21, …, 5) each consisting of spatial density distribution, $\boldsymbol{\phi}_A(\chi, \boldsymbol{x}, \boldsymbol{y})$, of the polymer species A discretized in a (64 × 64) grid. Panel A clearly shows that the polymer melt is transitioning from a disordered state at low χ to a lamellar ordered state at larger χ. Panel B illustrates the results of NTFk-CP decomposition as it is applied to the data. NTFk-CPD extracts four (R = 4) rank-1 tensors each given as a tensor product of 3 vectors, where each of these vector-columns is of size (1 × 64). These 12 vectors are given in panel C. Finally, panel D shows the Silhouette-Reconstruction criterium that explains how NTFk determines that the rank of $\boldsymbol{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$ is R = 4. The results given in Figure 4, panel C demonstrate that the methodology readily identifies the existence of a phase transition. This is

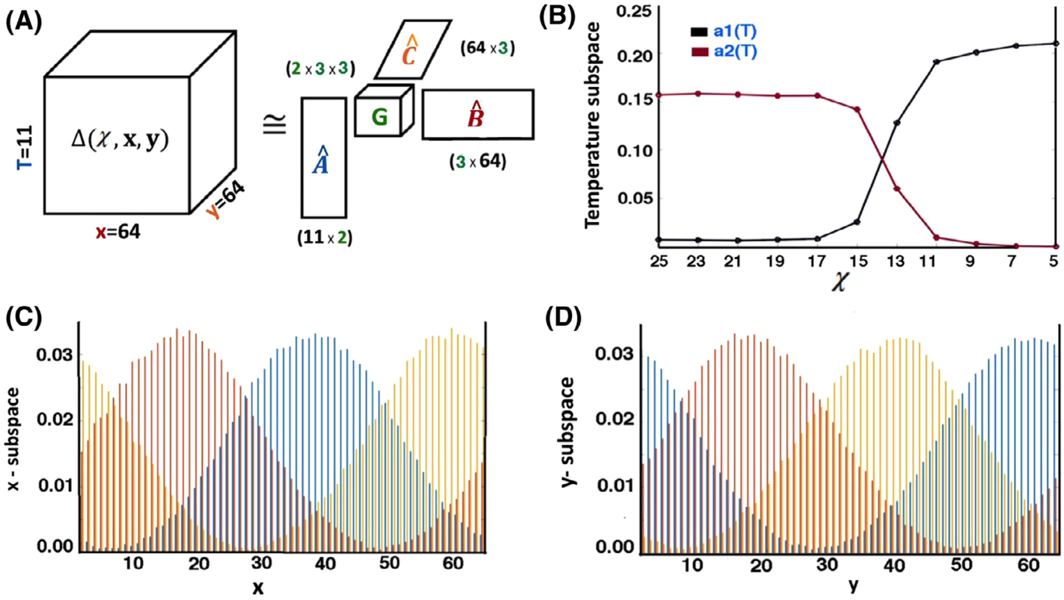

FIGURE 5 The results of NTFk-TD: (A) Tucker decomposition of the data $\mathbf{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$ with multirank $R = (2, 3, 3)$. (B) $\widehat{\mathbf{A}}$, the χ subspace with two columns. (C) $\widehat{\mathbf{B}}$, the x-subspace with three columns. (D) $\widehat{\mathbf{C}}$, the y-subspace with three columns

most clearly seen by the behavior of the columns of the factor matrix $\mathbf{A}(\chi)$ as a function of χ. Panel C shows that three vector sets, $\boldsymbol{S}_i = (\boldsymbol{a}_i(\chi), \boldsymbol{b}_i(\boldsymbol{x}), \boldsymbol{c}_i(\boldsymbol{y}))$; $i = 1$, 2, 3 are required to describe the phase segregated state. The remaining set, $\boldsymbol{D} = (\boldsymbol{a}_4(\chi), \boldsymbol{b}_4(\boldsymbol{x}), \boldsymbol{c}_4(\boldsymbol{y}))$, represents the disordered state. It is also clear that the three vector sets, $\boldsymbol{S}_i$, representing the segregated state, are degenerate in the sense that they only differ by a spatial shift on the vectors $\boldsymbol{b}_i(\boldsymbol{x})$ and $\boldsymbol{c}_i(\boldsymbol{y})$. The degeneration is caused by the orientation of the lamella along the diagonal simulation box. Note that panel A can give the impression that two phase transitions occur—one from disorder to cylindrical phase and the second from cylindrical to lamella phase. However, panel C clearly demonstrates that there is only a single transition. Therefore, the cylindrical-like structures are intermediate states but not a true phase.

5.3 | NTFk-TD subspace-analysis of the order parameter $\mathbf{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$

The nonnegative TD of the three-dimensional data-tensor represents the subspaces of the parameter $\mathbf{\Delta}(\chi, \boldsymbol{x}, \boldsymbol{y})$,

$$\Delta_{n,m,l} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} G_{r_1,r_2,r_3} \widehat{A}_{r_1,n} \widehat{B}_{r_2,m} \widehat{C}_{r_3,l} + \widehat{\varepsilon}_{n,m,l} \quad (6)$$

where all elements of $\mathbf{\Delta}$, $\boldsymbol{G}$, $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{B}}$, and $\widehat{\boldsymbol{C}}$ are nonnegative, Equation (6) can be also represented as:

$$\Delta_{n,m,l} = \widetilde{\Delta}_{n,m,l} + \widehat{\varepsilon}_{n,m,l} \quad (7)$$

where $\widetilde{\mathbf{\Delta}}$:($\widetilde{\mathbf{\Delta}} \in \mathbb{R}_{\geq 0}^{N \times M \times L}$) is the TD estimate of $\mathbf{\Delta}$.

TD includes:

- An unknown core-tensor $\boldsymbol{G}$ ($\boldsymbol{G} \in \mathbb{R}_{\geq \mathbf{0}}^{R_1 \times R_2 \times R_3}$) that represents the mixing between the χ, $\boldsymbol{x}$, and $\boldsymbol{y}$ components.
- An unknown matrix $\widehat{\mathbf{A}}$ ($\widehat{\mathbf{A}} \in \mathbb{R}_{\geq \mathbf{0}}^{N \times R_1}$) representing the changes of the order parameter $\mathbf{\Delta}$ with χ (the χ-component);
- An unknown matrix $\widehat{\mathbf{B}}$ ($\widehat{\mathbf{B}} \in \mathbb{R}_{\geq \mathbf{0}}^{M \times R_2}$) representing the changes of the order parameter $\mathbf{\Delta}$ in $\boldsymbol{x}$-direction (the $\boldsymbol{x}$-component);
- An unknown matrix $\widehat{\mathbf{C}}$ ($\widehat{\mathbf{C}} \in \mathbb{R}_{\geq \mathbf{0}}^{L \times R_3}$) representing the changes of the order parameter $\mathbf{\Delta}$ in $\boldsymbol{y}$-direction (the $\boldsymbol{y}$-component);

Mathematically, the solution of the nonnegative TD decomposition is a solution of a nonconvex optimization with nonnegative constraints. To extract the unknown core-tensor $\boldsymbol{G}$, and the factor matrices, $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, and $\widehat{\mathbf{C}}$, we utilized the block coordinate descent method for regularized multi-convex optimization algorithms introduced in Refs. [43,44].

Figure 5 shows the results of the TD factorization. This factorization results in three matrices ($\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, and $\widehat{\mathbf{C}}$) and a single $(2 \times 3 \times 3)$ core-tensor $\boldsymbol{G}$. The optimal value of the multirank R was determined through application of the clustering procedure described in Section 3. The matrices are given in panels B to D, where different colors represent different columns of the matrices. Again, the columns of matrix $\boldsymbol{A}$ clearly identify the phase transition. However, in this factorization the disordered state is not readily identifiable as it is achieved by a combination of the columns of $\boldsymbol{B}$ and $\boldsymbol{C}$ through the components of the tensor $\boldsymbol{G}$.

5.4 | NTFk-TD subspace-analysis of the order parameter $\widehat{\Delta}(f_A, x, y)$

In the previous section we used SCFT data generated along the vertical line through the polymer phase diagram given in Figure 3. Here we show the TD applied to 201 SCFT

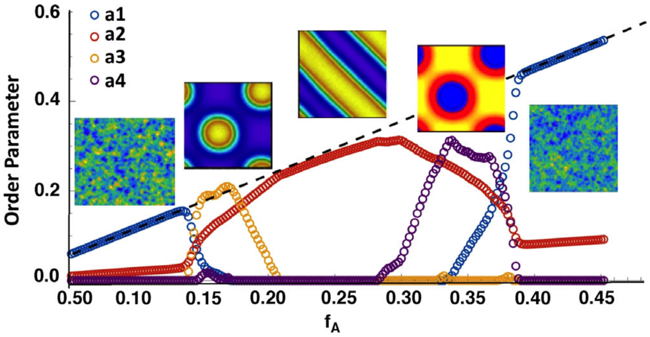

FIGURE 6 The results of NTFk-TD of $\hat{\mathbf{\Delta}}(f_A, x, y)$. This dataset is generated along the horizontal line on the phase diagram presented in Figure 3. It consists of 201 simulations for various f_A and for fixed $\chi = 25$. We present the four density features extracted by NTFk-TD from the data, that is, the columns of the first factor matrix $\hat{\mathbf{A}}$. The core-tensor has multirank $R = (4, 12, 12)$. The four insets illustrate the microphase separation patterns formed by the $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ factors corresponding to these features that follow the changes of f_A. The dashed line is the average density of block A

datasets generated along the horizontal line in the phase diagram presented in Figure 3. These 201 datasets are for $\chi = 25$ and $f_A = 0.05, 0.052, 0.054, \ldots, 0.45$. As seen from the phase diagram this data contains 4 phase transitions between disordered, cylindrical and lamellar states. Similarly, to the case presented in Figure 5, here the TD factorization results in three matrices ($\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$) and one core-tensor $\boldsymbol{G}$ with size $(4 \times 12 \times 12)$. The size of $\boldsymbol{G}$ was determined via the clustering procedure described in Section 3. In Figure 6 the four columns of the matrix $\hat{\mathbf{A}}$ are shown. From Figure 6 it is clear that the factorization identifies all four phase transitions. The column a_1 (in blue) represents the disordered state, a_3 (in yellow) and a_4 (in purple) represent cylindrical states, and finally a_2 (in red) represents the lamellar state, as is illustrated by the five inset figures. It can be noticed that the magnitudes of the components increase linearly with f_A as expected from the fact that average value of $\hat{\mathbf{\Delta}}(f_A, x, y) \sim f_A$ as illustrated by the black dashed line in Figure 6. It is also noteworthy that $\hat{\mathbf{A}}$, components are not symmetric around $f_A = 0.25$ but possess the asymmetry of the phase diagram as shown in Figure 3. The $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ factor matrices reproduce the four microphases of the phase diagram at each f_A point, as illustrated in the five insets of Figure 6: disordered, cylindrical, lamellar, inverted cylindrical, and again disordered phase.

6 | CONCLUSIONS

Here we demonstrated the applicability of our approach NTFk to analyze complex model outputs and determine the number of latent features in the case of block copolymer microphase segregation as modeled by SCFT. NTFk is based on NTF combined with custom k-means clustering. We demonstrated two types of nonnegative decompositions combined with our clustering: NTFk-CPD and NTFk-TD. NTFk-CPD allows for deconstruction of tensor datasets (multidimensional arrays) into sum of tensors with rank-1 and NTFk-TD—for a product of small core-tensor and factor matrices. We have shown that our approach can identify the unknown number of latent features and characterize the phase transitions based on the data alone. While our particular polymer system and the SCFT model are well studied and well understood, this is not true for most phase separating systems of interest in materials science and engineering. Therefore, a systematic methodology to characterize phase transitions solely from data, experimental or simulated, based on NTFk will be of importance for future materials development. We demonstrate that the latent features identified by NTFk have a clear physical meaning and enable easy interpretation of the processes governing phase separation. NTFk also acts as a data compression of the numerical simulations and may streamline the development of reduced-order models that can be applied to predict the system behavior in a more computationally efficient manner. NTFk can readily be applied to any observed or simulated dataset that is represented as a tensor (a multidimensional array) and possess latent features that span subspaces in the space of the observable quantities. We illustrated the different strengths of CPD and TD: CPD extracts straightforwardly the latent features, but it can result in a rank deficiency (linearly dependent features), which may obscure the final results and make the calculations difficult. TD does not suffer by rank deficiency but certain latent features are not clearly manifested in the extracted subspaces, for example, there is no spatial representation of the disorder phase in Figure 5, although the transition can be still clearly observed. In general, it therefore seems that a combined application of both approaches, NTFk-CPD and NTFk-TD, is preferred. Extracting latent features and subspaces can help the development of conceptual models, reduced-order models, and simplified closed-form mathematical expressions, which can then be used to predict the system behavior. NTFk can also be applied to detect anomalies or to complete gaps in the data. Lastly, NTFk-extracted latent features can be coupled with a suitable supervised ML method to produce predictions.

ACKNOWLEDGMENTS

This research was funded by LANL LDRD grant 20180060 and resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396.

CONFLICTS OF INTEREST

The authors declare no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

All authors conceived the concept and designed the research. B.S.A., V.V.V., and K.Ø.R. performed the simulations.

All authors analyzed the simulations, wrote the paper, and reviewed the manuscript.

ORCID

Boian S. Alexandrov https://orcid.org/0000-0001-8636-4603

REFERENCES

1. T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev. 51(3) (2009), 455–500. MR2535056
2. B. Everett, *An introduction to latent variable models*, Springer Science & Business Media, London, 2013. MR0769300
3. M. Ishteva, *Tensors and latent variable models*, International Conference on Latent Variable Analysis and Signal Separation, Springer, Berlin, 2015, pp. 49–55.
4. A. H. Phan and A. Cichocki, *Tensor decompositions for feature extraction and classification of high dimensional datasets*, Nonlinear Theory Appl. IEICE 1(1) (2010), 37–68.
5. N. D. Sidiropoulos et al., *Tensor decomposition for signal processing and machine learning*, IEEE Trans. Signal Process. 65(13) (2017), 3551–3582. MR3666587
6. A. Cichocki et al., *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*, Vol 500, John Wiley & Sons, West Sussex, UK, 2009.
7. G. Fredrickson, *The equilibrium theory of inhomogeneous polymers*, Vol 134, Oxford University Press on Demand, Oxford, 2005. MR2261726
8. C. Spearman, *General Intelligence," objectively determined and measured*, Amer. J. Psychol. 15(2) (1904), 201–292.
9. S. Haykin and Z. Chen, *The cocktail party problem*, Neural Comput. 1 (2005), 1875–1902.
10. I. T. Jolliffe, *Principal component analysis and factor analysis*, in *Principal component analysis*, Springer Series in Statistics, Springer Science & Business Media, Berlin, 1986, 115–128. MR203608
11. S. Amari, A. Cichocki, and A. H. Yang, *A new learning algorithm for blind signal separation*, Adv. Neural Inform. Process. Syst. 8 (1996), 757–763.
12. G. H. Golub and C. Reinsch, *Singular value decomposition and least squares solutions*, Numer. Math. 14(5) (1970), 403–420.
13. P. Paatero and U. Tapper, *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics 5(2) (1994), 111–126.
14. F. L. Hitchcock, *The expression of a tensor or a polyadic as a sum of products*, Stud. Appl. Math. 6(1–4) (1927), 164–189.
15. R. A. Harshman and M. E. Lundy, *Parafac: Parallel factor analysis*, Comput. Stat. Data Anal. 18(1) (1994), 39–72.
16. L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl. 21(4) (2000), 1253–1278. MR1780272
17. L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika 31(3) (1966), 279–311. MR0205395
18. C. A. Andersson and R. Bro, *The N-way toolbox for MATLAB*, Chemometrics Intell. Lab. Syst. 52(1) (2000), 1–4.
19. N. K. Faber, R. Bro, and P. K. Hopke, *Recent developments in CANDECOMP/PARAFAC algorithms: A critical review*, Chemometrics Intell. Lab. Syst. 65(1) (2003), 119–137.
20. D. D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature 401(6755) (1999), 788–791.
21. J. M. F. ten Berge and N. D. Sidiropoulos, *On uniqueness in CANDECOMP/PARAFAC*, Psychometrika 67(3) (2002), 399–409. MR1961325
22. J. Håstad, *Tensor rank is NP-complete*, J. Algorithms 11(4) (1990), 644–654. MR107945
23. Y. Qi, P. Comon, and L. H. Lim, *Uniqueness of nonnegative tensor approximations*, IEEE Trans. Inform. Theory 62(4) (2016), 2170–2183.
24. L. H. Lim and P. Comon, *Nonnegative approximations of nonnegative tensors*, J. Chemometrics 23(7–8) (2009), 432–441.
25. L. B. Alexandrov et al., *Deciphering signatures of mutational processes operative in human cancer*, Cell Rep. 3 (2013), 246–259.
26. L. B. Alexandrov et al., *Signatures of mutational processes in human cancer*, Nature 500 (2013), 415–421.
27. B. S. Alexandrov and V. V. Vesselinov, *Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization*, Water Resour. Res. 50(9) (2014), 7332–7347.
28. L. B. Alexandrov et al., *Clock-like mutational processes in human somatic cells*, Nat. Genet. 47 (2015), 1402–1407.
29. L. B. Alexandrov, *Understanding the origins of human cancer*, Science 350 (2015), 1175–1177.
30. V. V. Vesselinov, B. S. Alexandrov, and D. O'Malley, *Contaminant source identification using semi-supervised machine learning*, J. Contam. Hydrol. 212(2018) (2017), 134–142.
31. V. Stanev et al., *Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering*, NPJ Comput. Mater. 4 (2018), 1–10.
32. V. G. Stanev et al., *Identification of release sources in advection-diffusion system by machine learning combined with Green's function inverse method*, Appl. Math. Model. 60(2018) (2018), 64–76. MR3802617
33. F. L. Iliev et al., *Nonnegative matrix factorization for identification of unknown number of sources emitting delayed signals*, PLOS ONE 13 (2018), e0193974.
34. P. J. Rousseeuw, *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, J. Comput. Appl. Math. 20 (1987), 53–65.
35. E. Helfand, *Theory of inhomogeneous polymers: Fundamentals of the Gaussian random walk model*, J. Chem. Phys. 62 (1975), 999–1005.
36. E. Helfand and Z. R. Wasserman, *Block copolymer theory. 4. Narrow interphase approximation*, Macromolecules 9 (1976), 879–888.
37. K. M. Hong and J. Noolandi, *Theory of inhomogeneous multicomponent polymer systems*, Macromolecules 14 (1981), 727–736.
38. J. D. Vavasour and M. D. Whitmore, *Self-consistent mean field theory of the microphases of diblock copolymers*, Macromolecules 25 (1992), 5477–5486.
39. K. Ø. Rasmussen and G. Kalosakas, *Improved numerical algorithm for exploring block copolymer mesophases*, J. Polym. Sci. Part B: Polym. Phys. 40(16) (2002), 1777–1783.
40. F. S. Bates and G. H. Fredrickson, *Block copolymers-designer soft materials*, Phys. Today 52 (2000), 32–38.
41. I. W. Hamley, *The physics of block copolymers*, Oxford University Press, New York, 1998, 24–108.
42. M. W. Matsen, *The standard Gaussian model for block copolymer melts*, J. Phys.: Condens. Matter 14 (2001), R21.
43. Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci. 6(3) (2013), 1758–1789.
44. Y. Xu, *Alternating proximal gradient method for sparse nonnegative Tucker decomposition*, Math. Program. Comput. 7(1) (2015), 39–70.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Alexandrov BS, Stanev VG, Vesselinov VV, Rasmussen KØ. Nonnegative tensor decomposition with custom clustering for microphase separation of block copolymers. *Stat Anal Data Min: The ASA Data Sci Journal* 2019;12:302–310. https://doi.org/10.1002/sam.11407