Shafii et al., Int J Appl Exp Math 2017, 2: 120
https://doi.org/10.15344/2456-8155/2017/120

International Journal of
Applied & Experimental Mathematics

Research Article Open Access

Characterizing the Kootenai River Macroinvertebrate Community Structure Using Nonnegative Matrix Factorization

Bahman Shafii[1*], William J. Price[1], Charlie Holderman[2], G. Wayne Minshall[3] and Paul J. Anders[4]

[1] *Statistical Programs, University of Idaho, 875 Perimeter Dr, Moscow, ID 83844, USA*
[2] *Kootenai Tribe of Idaho, USA*
[3] *Stream Ecology Center, Idaho State University, 921 S 8th Ave, Pocatello, ID 83209, USA*
[4] *Cramer Fish Sciences, USA*

Abstract

Nonnegative matrix factorization (NMF), also known as nonnegative matrix approximation, is a group of methods in multivariate statistical analyses, where matrix $\boldsymbol{X}$ is factorized into two matrices $\boldsymbol{W}$ and $\boldsymbol{H}$, with the property that all three matrices have no negative elements. This produces matrices that are easier to inspect and interpret. NMF is often referred to as an unsupervised machine learning technique for pattern recognition due to its clustering capability, with a wide range of applications in engineering, genomics, bioinformatics, and in processing audio spectrograms and textual data. The purpose of this article is to present the essence of the NMF technique and its potential use in characterizing the observed pattern and structure in a benthic macroinvertebrate community. Applications are demonstrated with reference to twelve years of benthic macroinvertebrate survey data collected from an ultra-oligotrophic reach of the Kootenai River in Northern Idaho and Western Montana downstream from a hydro-electric dam.

Publication History:

Received: June 07, 2017
Accepted: August 28, 2017
Published: August 30, 2017

Keywords:

Matrix Decomposition, Pattern Recognition, Community Structure

Introduction

Nonnegative matrix factorization (NMF) is a useful method for decomposition of multivariate data. NMF may be compared to other multivariate decomposition/factorization techniques such as principal component analysis and vector quantization, however, due to different implementation of imposed constraints, it leads to a different representation of data. The original articles by Pentti Paatero [1,2] on “positive matrix factorization” initiated a flurry of research articles in this area. However, it was the article in *Nature* by Daniel Lee and Sebastian Seung [3] that established the use of NMF in its modern formulation for scientific investigations.

Subsequently, it has been shown that different formulations of NMF are related to a more general probabilistic model, multinomial PCA [4], and that when NMF is obtained by minimizing the Kullback–Leibler divergence, it is in fact equivalent to multinomial PCA (probabilistic latent semantic analysis), trained by maximum likelihood estimation [5]. Furthermore, it has been established that NMF with the least-squares objective is equivalent to a relaxed form of K-means clustering [6], hence, providing a foundation for use of NMF in data clustering.

NMF has a wide range of applications in science, engineering and medicine. It has also been quite extensively used in bioinformatics investigations for the purpose of clustering gene expressions and determining the genes most representative of the clusters [7,8,9]. Most recently, NMF has been used in analysis of human milk oligosaccharides, HMOs, across various geographical populations in a nutritional study [10].

In this paper, NMF is utilized to investigate the structure of a benthic macroinvertebrate community in a large regulated river. The clustering of sampling units, based on multiple macroinvertebrate metrics across river zones is explored and the potential impact of nutrient supplementation on a set of specific macroinvertebrate response metrics is evaluated. Applications are demonstrated using twelve years of replicated benthic macroinvertebrate data collected from the Kootenai River in northern Idaho and northwestern Montana, as part of the Kootenai River Bio-monitoring Project conducted by the Kootenai Tribe of Idaho.

Let r > 0 be an integer, and $\boldsymbol{X}$ a matrix with n rows, the measured features, and m columns, the samples with non-negative entries. NMF entails finding an approximation:

$$\boldsymbol{X} \sim \boldsymbol{WH} \qquad (1)$$

where $\boldsymbol{W}, \boldsymbol{H}$ are $n \times r$ and $r \times m$ non-negative matrices, respectively. Since the objective is usually to reduce the dimension of the original data, the factorization rank, r, is often chosen so that $r << \min(n, m)$.

Simply put, equation (1) states that each column of $\boldsymbol{X}$ (i.e. the observed features of each sample) is approximated by a non-negative linear combination of the columns of $\boldsymbol{W}$ (i.e. the basis components), where the coefficients are given by the corresponding column of $\boldsymbol{H}$ (i.e. the mixture coefficients). The NMF algorithm iteratively computes and updates an approximation of (1), commonly by randomly initializing matrices $\boldsymbol{W}$ and $\boldsymbol{H}$, to minimize a divergence functional. Mathematically, the NMF algorithm estimates matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ as a local minimum of the following optimization problem:

$$\min \{\delta(\mathbf{X}, \mathbf{WH}) + \rho(\mathbf{W}, \mathbf{H})\} \qquad (2)$$

where δ is defined as a loss function measuring the quality of the approximation. Loss functions are often based on the Frobenius norm or the Kullback-Leibler divergence [11,12]. ρ is a regularization function,

[*] **Corresponding Author:** Dr. Bahman Shafii, Statistical Programs, University of Idaho, 875 Perimeter Dr, Moscow, ID 83844, USA, E-mail: bshafii@uidaho.edu

Citation: Shafii B, Price WJ, Holderman C, Minshall GW, Anders PJ (2017) Characterizing the Kootenai River Macroinvertebrate Community Structure Using Nonnegative Matrix Factorization. Int J Appl Exp Math 2: 120. doi: https://doi.org/10.15344/2456-8155/2017/120

utilized to ensure desirable properties on matrices W and H, such as smoothness or sparsity. Bayesian formulations of the optimization problem in (2) are also possible, depending on one's prior knowledge concerning the data and the field of application [13].

Several software packages are available to conduct the necessary computations for the NMF algorithm, including those recommended by Lee and Seung [11], Brunet, et al. [14], and Zhang [15]. In this article, we have employed an R® shared package, NMF (version 0.20.6), originally developed by Renaud Gaujoux and Cathal Seoighe in 2010 [16]. All other statistical computations were carried out using SAS version 9.4 (2012).

Empirical Results and Demonstrations

Data description

The Kootenai River is located along the junction of the Idaho, Montana and Canadian borders. The river runs south and west from Canada to Montana and Idaho, then returning back north to Canada. A large hydro-electric facility, Libby Dam, impedes the river flow near the town of Libby, MT and has resulted in oligotrophic conditions downstream. In 2002, fourteen biomonitoring sites were established along the river to monitor water quality, primary production, benthic invertebrates, and fish populations (Figure 1). Of the variables measured, the benthic macroinvertebrate data for ten selected sites will be considered for analysis here. As a means of mitigating the biological impacts due to operation of the dam, a nutrient addition (phosphorus) program was initiated at the ID – MT border in 2005 and has continued during the June-September time frame of each subsequent year [17]. Based on this nutrient addition program, three river zones, encompassing the ten selected sites, were designated as: the Upper River Zone (URZ, sites KR12, KR11, and KR10: an untreated control region above the nutrient addition point); the Nutrient Addition Zone (NAZ, sites KR9, KR7, and KR6: a region immediately adjacent and downstream of the nutrient addition point), and; the Lower River Zone (LRZ, sites KR4, KR3, KR2, and KR1: a river section located further downstream from the nutrient addition point and having different flow and channel conditions.

Benthic macroinvertebrate sampling was typically carried out multiple times per year, although the exact timing and number of samples varied depending on river conditions and project requirements. In order to establish a common sampling timeframe across years, the months of July through November were selected for analysis. These months were the most populated with available data and, because nutrient addition was initiated approximately in June of each year, these months were considered as the most biologically relevant for the benthic invertebrate communities of interest. The years spanning 2003 through 2015 were selected for analysis as they encompassed several years of both pre and post-nutrient addition periods.

At each sampling event, 5-6 replicates (random samples) were taken at each site and date. Each sample was sorted and identified to the species level or the nearest taxonomic grouping of benthic macroinvertebrates. From this taxonomic information, additional responses such as total abundance, total biomass, and community diversity metrics were determined. From a large collection of potential macroinvertebrate variables, including abundance and richness measures, a set of six metrics were determined to be important for the assessment of the effect of nutrient addition on the benthic community [18,19], and will therefore be used in the subsequent NMF analyses. These metrics were: total Biomass, Chironomidae (midges) abundance, Filterer abundance, total taxa abundance excluding sub-taxa of Chironomidae and Oligochaeta (worms) (NCO), Ephemeroptera (mayflies) abundance, and NCO richness. For each metric, the average value was then computed for each year-month-site combination and these data were further classified according to the nutrient addition period (Pre: 2003-2005; or Post: 2006-2015) and river zone as defined previously. Finally, the data for each river zone were arranged in matrix form with columns and rows corresponding to year-month-sites and response metrics, respectively.

NMF analysis

All procedures described here were carried out separately for each nutrient addition period in each river zone, however, for the purposes of demonstration, only the results for the Nutrient Addition Zone (NAZ) will be provided. The corresponding results for the LRZ and

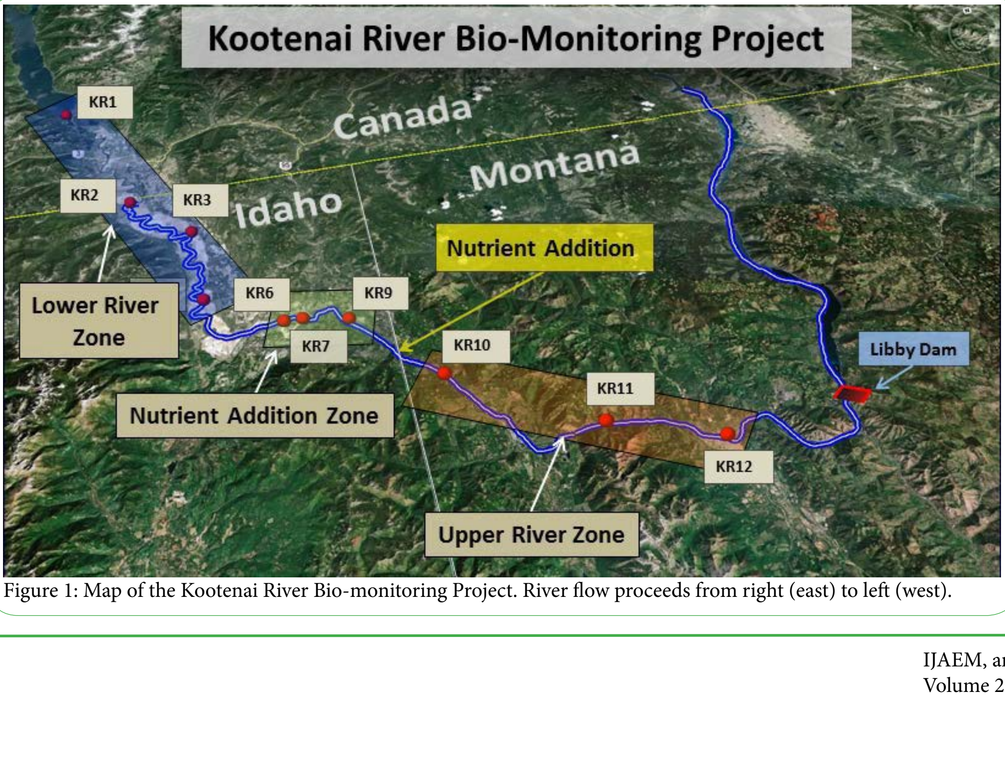

Figure 1: Map of the Kootenai River Bio-monitoring Project. River flow proceeds from right (east) to left (west).

URZ are given in the supplementary materials (supplementary files, 1 and 2).

An initial assessment of the NMF space for each data set was completed to determine the rank, r, or number of basis components sufficient to describe the data matrices. Through sequential NMF estimations from r = 2 to 5, scree plots describing the average model residuals versus the model rank, r, were developed (Figure 2). In both the Pre and Post-nutrient addition periods, the residual value drops substantially after two basis components, hence, three basis components were deemed adequate to describe the data matrices.

Following rank determination, full NMF analyses were conducted for the Pre and Post nutrient addition periods in each river zone. A primary outcome of NMF analysis is the quantification of the r basis component coefficients, which make up the matrix W. While the ordering of the basis components does not reflect their importance, the relative magnitude of the coefficients within each basis component relates to the contribution each of the constituent metrics makes to a basis. Typically, these basis coefficients are visually displayed in the form of a heat map (basis map). In this example, the maps for Pre and Post-nutrient addition for the NAZ are shown in Figures 3 and 4, respectively. Within each basis (columns in the heat map), the darker

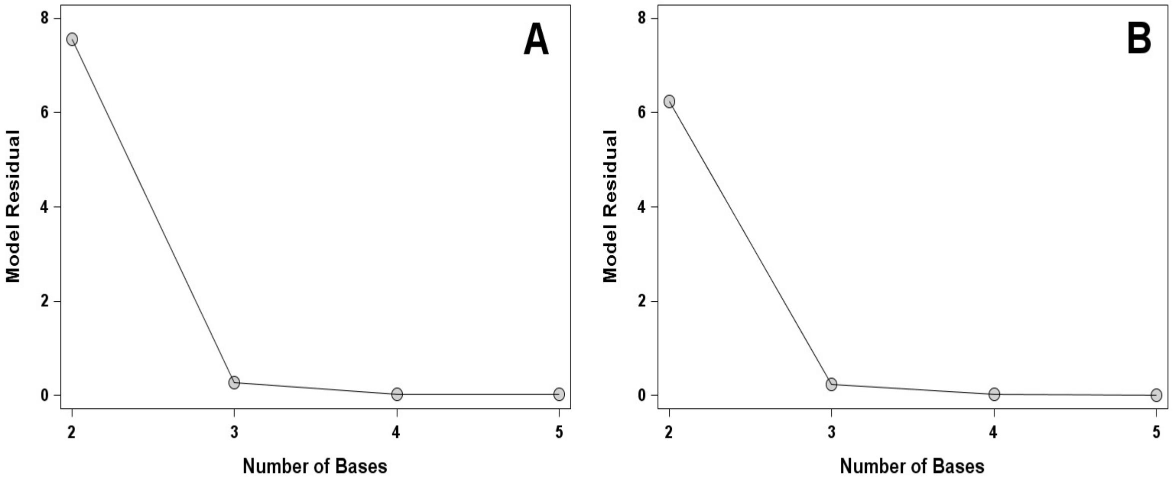

Figure 2: NMF residual scree plots for the Nutrient Addition Zone in the A) Pre and B) Post-nutrient addition periods.

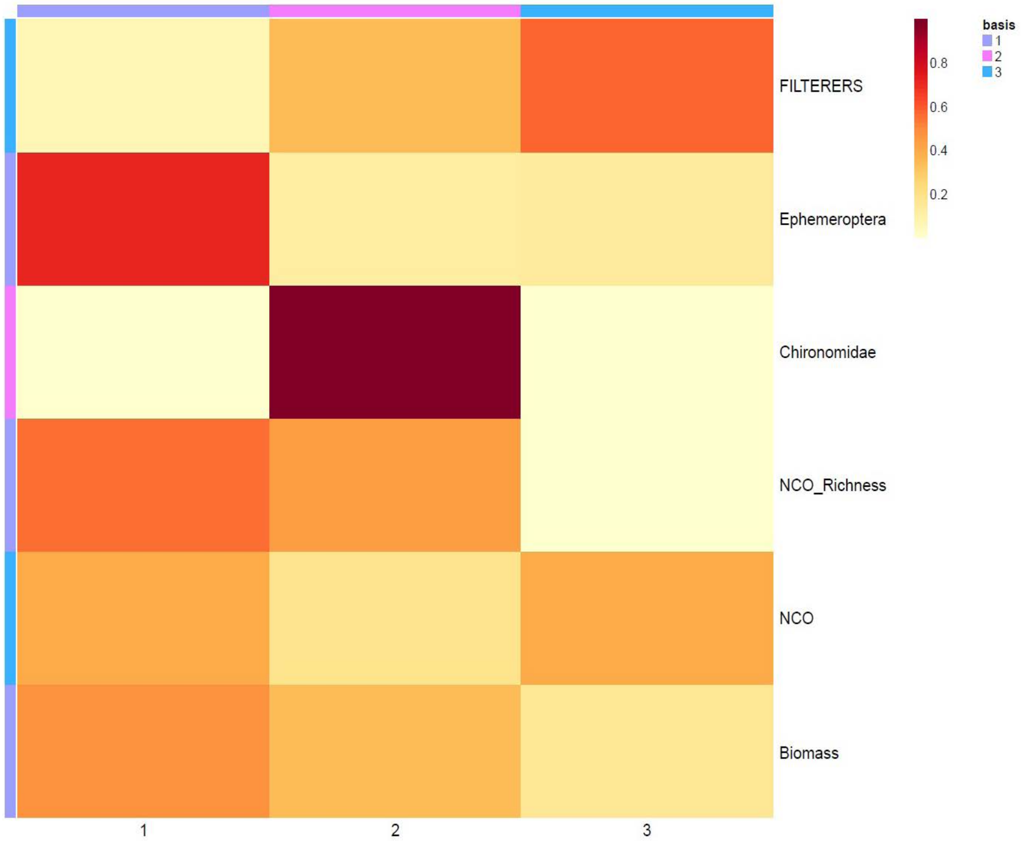

Figure 3: Basis map for NMF analysis of Pre nutrient addition period for the Nutrient Addition River Zone. Darker colors indicate larger loadings for specific metrics (rows) within each of the three basis components (columns). Edge colors represent the linkage between basis components and metrics.

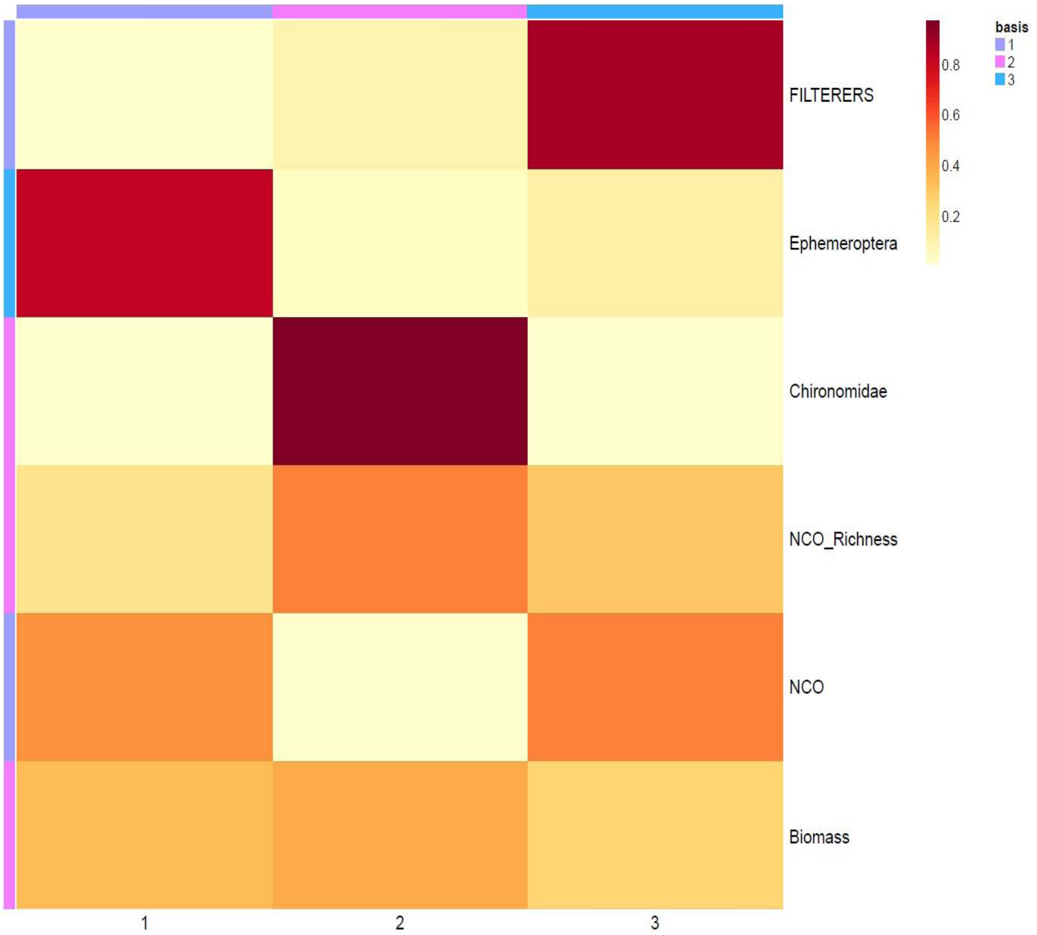

Figure 4: Basis map for NMF analysis of Post nutrient addition period for the Nutrient Addition River Zone. Darker colors indicate larger loadings for specific metrics (rows) within each of the three basis components (columns). Edge colors represent the linkage between basis components and metrics.

colors indicate higher coefficient loadings for the specified metrics (rows in the heat map). For example, in the Pre-nutrient addition heat map (Figure 3), basis 1 has a relatively higher contribution from the Ephemeroptera abundance metric, while the second and third components have high loadings on Chironomidae and Filterers abundances, respectively. Taken jointly, one can conclude that these three classifications are driving metrics in the response patterns in the Pre-nutrient dataset. Of these, the Chironomidae abundance response is dominant.

In the Post-nutrient addition NMF heat map (Figure 4), a similar pattern of loading is seen, however, the relative magnitude of the same metrics are higher, suggesting that nutrient addition had an influence on the response patterns for these categories. Of the three metrics, the Chironomidae and Filterer classifications are the most responsive, which may be expected as nutrient addition enhances the algae growth on which they feed.

One means of quantifying metrics across all basis components is the NMF (gene) score as described by Kim and Park [7]. This value is computed from the individual coefficients of W, and it is a real value ranging from 0.0 to 1.0, proportional to the probability of contribution of a specified metric (factor) across basis components. A higher score indicates a larger contribution for a specific metric. The NMF scores for this scenario are given in Table 1. As expected, the three metrics identified earlier, Filterer abundance, Chironomidae abundance, and Ephemeroptera abundance, show the largest score values within each nutrient addition period, with the Chironomidae metric being dominant. The change in scores across periods is also informative. Here, the scores for Ephemeroptera and Chironomidae abundances show little change after nutrient addition. The score for the filterer abundance, however, shows a substantial increase from 0.16 in the Pre-nutrient addition period to 0.63 in the Post nutrient addition period, suggesting that the filterer taxonomic group is being impacted by the nutrient addition treatment. It is also notable that the metrics for NCO richness and total Biomass have decreased in score values and contribute little to the underlying pattern in the Post nutrient addition period.

Period	Filterers	Ephemeroptera	Chironomidae	NCO_Richness	NCO	Biomass
Pre-NA	0.1625	0.4696	0.9158	0.3711	0.0875	0.0710
Post-NA	0.6269	0.4379	0.8673	0.0641	0.2780	0.0107

Table 1: Pre and Post-nutrient addition (Pre-NA; Post-NA) NMF scores for the Nutrient Addition River Zone.

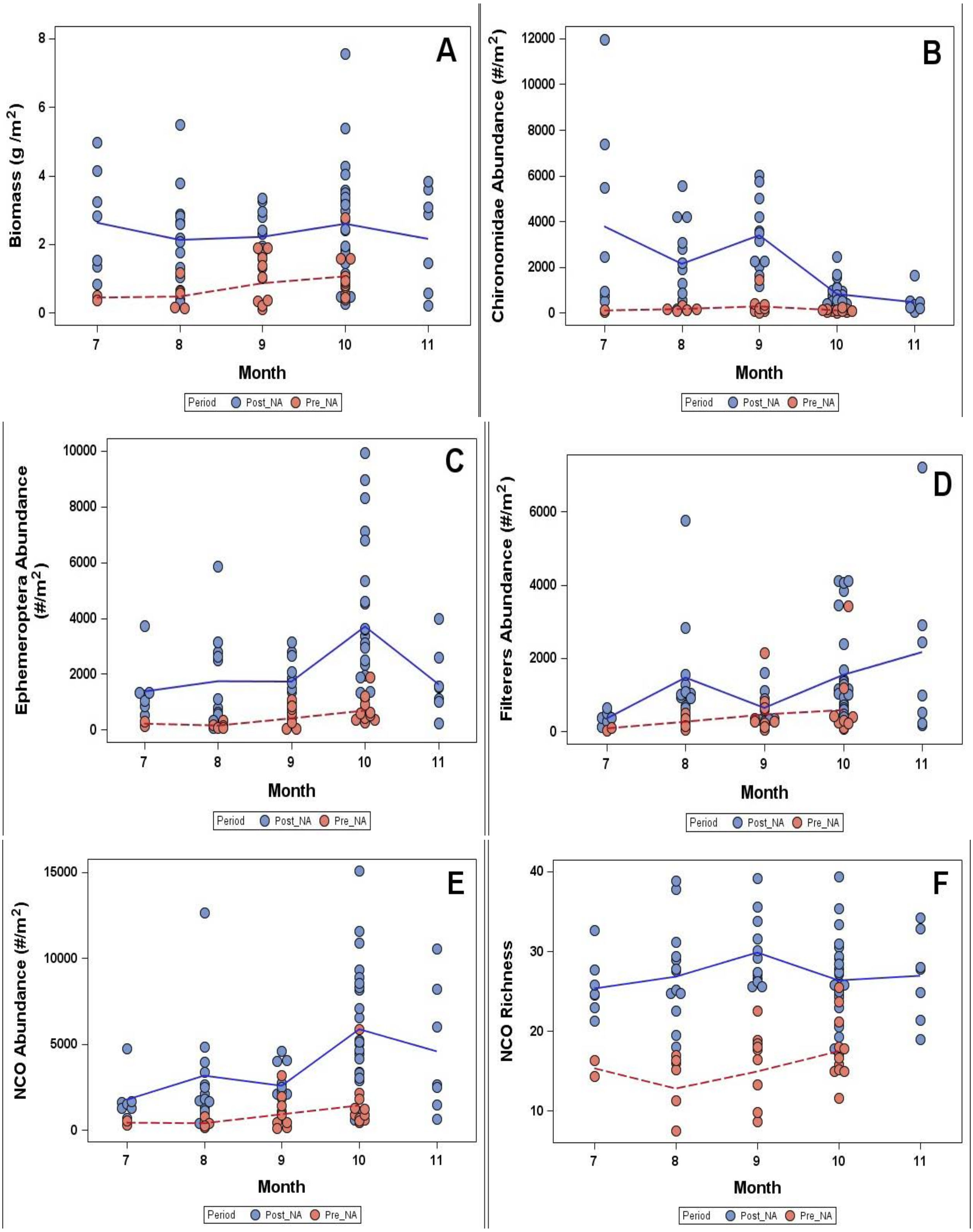

Figure 5: Trend and scatter plots across months in the Nutrient Addition River Zone for the six macroinvertebrate metrics A) Biomass, B) Chironomidae abundance, C) Ephemeroptera abundance, D) Filterer abundance, E) NCO abundance, and F) NCO richness. Red markers and dashed lines indicate the Pre-nutrient addition period and blue colors and solid lines indicate the Post-nutrient addition period.

It is also useful to augment the NMF analyses with an examination of the raw data patterns involved. Here, trend and scatter plots are used to assess data patterns across months for each nutrient addition period within NAZ. These plots are given in Figure 5. In panels 5B, 5C, and 5D, corresponding to Chironomidae abundance, Ephemeroptera abundance, and Filterer abundance, respectively, there are notable changes in the response patterns across months, particularly in the Post-nutrient addition period, all of which correspond to the previous NMF results. Note that, in NMF analysis, attention is drawn to changes in the overall pattern of the data, not necessarily their magnitudes. For example, in panels 5A and 5F, corresponding to total biomass and NCO richness, there is a sizable shift in the magnitude of the responses across periods moving from the Pre to Post-nutrient data sets. Within each period, however, the trends are relatively flat. This corresponds to the limited response of these metrics in the NMF analysis (light colors in Figures 3 and 4; lower score values in Table 1). In contrast, those metrics showing the greatest response in the NMF analyses also showed the greatest variation among months.

Conclusion

Nonnegative matrix factorization is a useful and powerful technique for pattern recognition, particularly in high dimensional data. In many biological and ecological applications, dimension reduction is crucial for efficient representation and interpretation of the data. In this study, we employed NMF to describe changes in a benthic macroinvertebrate community structure at the aggregated level following a nutrient enhancement treatment. It is important to note that each of the identified macroinvertebrate metrics at the aggregated level is comprised of many taxonomic categories, with varying degree of potential contribution to the underlying pattern and variability. For example, the family Chironomidae encompasses 73, 84 and 57 identified taxonomic categories at the genus and species level, for the lower, nutrient addition, and upper river zones, respectively, across all years and sampling dates. It will be very informative to perform NMF at this (higher dimension) level of taxonomy and identify specific taxa within the Chironomidae family (midges) responsible for the observed pattern in the data. The authors are currently conducting the required analyses and intend to report the results in future publications.

Competing Interests

The author declares that they have no competing interests.

References

1. Paatero P, Tapper U (1994) Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5: 111-126.
2. Paatero P (1997) Least-squares formulation of robust non-negative factor analysis. Chemometrics and Intelligent Laboratory Systems 38: 223-242.
3. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788-791.
4. Buntine W (2002) Variational Extensions to EM and Multinomial PCA. Proc. European Conference on Machine Learning (ECML-02). LNAI 2430: 23-34.
5. Gaussier E, Goutte C (2005) Relation between PLSA and NMF and Implications. Proc. 28th international ACM SIGIR conference on research and development in information retrieval (SIGIR-05). pp. 601-602.
6. Ding C, He X, Simon HD (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. Proc. SIAM Int'l Conf. Data Mining, pp. 606-610.
7. Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23: 1495-1502.
8. Devarajan K (2008) Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. PLoS Computational Biology 4: e1000029.
9. Schwalbe EC, Williamson D, Lindsey JC, Hamilton D, Ryan SL, et al. (2013) DNA methylation profiling of medulloblastoma allows robust sub-classification and improved outcome prediction using formalin-fixed biopsies. Acta Neuropathol 125: 359-371.
10. McGuire MK, Meehan CL, McGuire MA, Williams JE, Foster J, et al. (2017) What's normal? Oligosaccharide concentrations and profiles in milk produced by healthy women vary geographically. Am J Clin Nutr 105: 1086-1100.
11. Lee DD, Seung HS (2001) Advances in neural information processing systems, pp 556-562.
12. Cichocki A, Zdunek R, Amari S (2006) New Algorithms for non-negative matrix factorization in applications to blind source separation. Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. IEEE International Conference 5: V-621-V-624.
13. Cemgil AT (2007) Bayesian inference for nonnegative matrix factorization models. Computational Statistics & Data Analysis 52: 155-173.
14. Brunet JP, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 101: 4164-4169.
15. Zhang J, Wei L, Feng X, Ma Z, Wang Y (2008) Pattern expression nonnegative matrix factorization: algorithm and applications to blind source separation. Computational intelligence and neuroscience 2008: 168769.
16. Gaujoux R, Seoighe C (2010) A flexible R package for nonnegative matrix factorization. In: BMC Bioinformatics 11: 367.
17. Holderman C, Hoyle G, Hardy R, Anders P, Ward P, et al. (2009) Libby Dam Hydro-electric Project Mitigation: Efforts for Downstream Ecosystem Restoration. In Section C-4 of 33rd International Association of Hydraulic Engineering and Research Congress, pp. 6214-6222.
18. Shafii B, Price WJ, Minshall WG, Holderman C, Anders PJ, Lester G, et al. (2013) Characterizing benthic macroinvertebrate community responses to nutrient addition using NMDS and BACI analyses. Applied Statistics in Agriculture, W. Song (Ed). Kansas State University, Manhattan, Kansas, pp.64-79.
19. Minshall GW, Shafii B, Holderman C, Price JW, Holderman C (2014) Effects of nutrient replacement on benthic macroinvertebrates in an ultra-oligotrophic reach of the Kootenai River, 2003-2010. Freshwater Science 33:1009-1023.