HHS Public Access

Author manuscript

Curr Opin Environ Sci Health. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

Curr Opin Environ Sci Health. 2020 June ; 15: 32–38. doi:10.1016/j.coesh.2020.05.001.

Unsupervised dimensionality reduction for exposome research

Vrinda Kalia[1], **Douglas I. Walker**[2], **Katherine M. Krasnodemski**[3], **Dean P. Jones**[3], **Gary W. Miller**[1], **Marianthi-Anna Kioumourtzoglou**[1]

[1]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY 10032

[2]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[3]Division of Pulmonary, Allergy and Critical Medicine, Department of Medicine, School of Medicine, Emory University, Atlanta, GA 30322

Abstract

Understanding the effect of the environment on human health has benefited from progress made in measuring the exposome. High resolution mass spectrometry (HRMS) has made it possible to measure small molecules across a large dynamic range, allowing researchers to study the role of low abundance environmental toxicants in causing human disease. HRMS data have a high dimensional structure (number of predictors >> number of observations), generating information on the abundance of many chemical features (predictors) which may be highly correlated. Unsupervised dimension reduction techniques can allow dimensionality reduction of the various features into components that capture the essence of the variability in the exposome dataset. We illustrate and discuss the relevance of three different unsupervised dimension reduction techniques: principal component analysis, factor analysis, and non-negative matrix factorization. We focus on the utility of each method in understanding the relationship between the exposome and a disease outcome and describe their strengths and limitations. While the utility of these methods is context specific, it remains important to focus on the interpretability of results from each method.

Environmental determinants of health are receiving increased attention since genome and inheritance studies have revealed that genetic variation only explains about 20% of human disease risk [1,2]. There are various environmental factors that affect human health, including environmental chemicals, dietary factors, lifestyle, the built environment, exposure to microbes, as well as structural policies that influence healthcare and healthy behaviors. To this effect, many studies have evaluated the effect of single or a small number of exposures

Corresponding author Vrinda Kalia: vk2316@cumc.columbia.edu.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

in isolation on various health outcomes. However, environmental exposures rarely happen in isolation and are accompanied with other exposures and context-specific factors. Along with the complexity of co-exposures, humans are constantly interacting with their environment over their life course, and exposures are spatiotemporally dynamic. The exposome concept emerged from this realization. In 2005, Christopher Wild formally coined the term and defined it as the "life-course environmental exposures (including lifestyle factors), from the prenatal period onwards" [3]. This definition has been modified over the last 15 years, but most revisions agree that the exposome comprises the entire set of lifelong environmental exposures and their associated biological response [1,2,4–6].

Measuring the exposome has benefited from technological advances in mass spectrometers. The arrival of high-resolution mass spectrometers (HRMS) has made it possible to measure small molecules (85–1250 Da) in a biological matrix with a large dynamic range, facilitating measurement of small molecules arising from endogenous metabolism—which are usually highly abundant (milli to picomoles)—and small molecules from exogenous sources that have a lower abundance in biological matrices (nano to picomoles) [7]. In this paper, we define the exposome as the sum of environmental exposures and the biological response induced by these environmental exposures. This definition assumes that meaningful environmental exposures that affect human health will produce a biological response which will be reflected in alterations to endogenous metabolic processes. Thus, measurements of small molecules in a biological matrix will allow us to measure this exposomic profile and assess biochemical changes that accompany exposure on a systems level. Using techniques from metabolomics, we can understand the biological effect associated with the exposome. For example, Walker et al. measured the exposomic profile of trichloroethylene (TCE) exposure among factory workers and found chlorinated metabolites in the plasma of factory workers that were correlated with markers of immune and kidney function along different parts of the pathway of TCE toxicity [8]. HRMS can also be leveraged to measure the sum of environmental chemicals in a matrix. By coupling gas chromatography to an HRMS, researchers have measured numerous chemicals of interest in a single sample simultaneously, e.g., the US Environmental Protection Agency's ring trial, wrist band studies, etc. [9,10].

Exposomic data from HRMS are high-dimensional and complex. When untargeted methods are used to characterize the exposome, the number of chemical signals can be greater than 100,000 [11]. These data tables represent the primary input for bioinformatic and biostatistical analysis to evaluate biological meaning and to determine exogenously derived chemicals that may be of relevance to an outcome of interest. Since detected analytes are not annotated *a priori,* the results from HRMS contain information on each detected ion (a feature). Each feature is characterized by its mass-to-charge ratio, the retention time at which the compound eluted from the chromatographic column, and its abundance in the sample. Some features may arise from the same parent compound but during ionization may lead to formation of multiple ions with different masses with the same retention times. Therefore, the feature table may contain degenerate features which will be highly correlated since they arise from the same parent compound. Thus, correlation is present not only due to exposure sources and biological processes, but also arises from the analysis itself [12]. This forces researchers to deal with the "curse of dimensionality" and to capture the essence of

the data generated [13]. Given the high correlations across chemical exposures, single chemical association studies are not appropriate, with implications for health that are missed using a “candidate” approach. Furthermore, high correlation compounds correction for multiple testing, decreasing study power and inflating the potential for false negatives.

Data dimensionality reduction approaches produce a manageable number of variables, allow for better visualization, remove redundant and uninformative variables, and reduce computational burden [14]. Several techniques have been described that reduce the size of the data table while minimizing loss of information, describing the essence of the data generated. These include unsupervised and supervised methods. In supervised methods, the outcome of interest informs the dimensionality reduction solution. These methods are used for feature selection and include—but are not limited to—different forms of penalized regression: least absolute shrinkage and selection operator (LASSO) [15], ridge regression [16], and elastic net regression [17], and different modifications of partial least squares regression [18,19]. Unsupervised methods do not take the outcome of interest under consideration during feature extraction. The goal in unsupervised dimensionality reduction is to discover the underlying structure in the data. These methods are tuned for pattern recognition that can aid in data visualization, data exploration, and uncovering latent variables. A popular method for unsupervised linear transformation of data is principal components analysis. Different unsupervised methods also exist for non-linear data transformations [20] like the kernel PCA [21], isomap [22] and autoencoders [23]. In this paper, we focus on unsupervised methods aiming to reduce the number of variables under study and obtain a smaller set of principal variables through feature extraction. We focus on the use of three methods for exposome research: factor analysis (FA), principal components analysis (PCA), and non-negative matrix factorization (NMF) (Figure 1). While NMF and FA have not yet been widely used with high-resolution mass spectrometry data, we think they offer useful applications.

1. Principal components analysis (PCA)

Principal components analysis is one of the most commonly used approaches for dimensionality reduction. The method uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principal components. The first component explains the most variance in the data and each succeeding component has the highest variance possible under the constraint that it is orthogonal, i.e., independent, to the preceding components [24]. The method does not reduce the number of variables, m variables produce m components. The analyst chooses the number of components to include in analyses based on some *a priori* defined criterion/a, e.g., looking at the scree plot, selecting components with eigenvalues above one, or selecting the number of components that explain a pre-specified proportion of the variance in the data, e.g., at least 75%. Since PCA forces orthogonality between components, it imposes a rigid structure [25]. The alternating least squares variant of PCA, independent component analysis, is more successful in dealing with this rigidity but has less compression in the first components [26].

In a study of 1301 European mother-child pairs, Tamayo-Uria et al. measured 87 environmental exposures during pregnancy and 122 exposures in early childhood. These included atmospheric, GIS, meteorological, built environment, SES, lifestyle, and biomarker data. The authors used PCA in two different ways. First, in the 19 pre-defined exposure groups, the first PC from each group was used as a composite index variable. Second, they used all exposure variables in the PCA and found that 65 and 90 PCs were needed to explain 95% of the variance in the pregnancy and early childhood exposome respectively [27]. In another study of 397 pregnant women in the Child Health and Development Studies in California, Li and colleagues investigated the relationship between 39 environmental chemicals and the serum metabolome. To reduce the dimensions and study the variance of the metabolome, they conducted a PCA on the high-resolution metabolomics data. The first PC explained 58.5% of the variance. They assessed the influence of the environmental chemicals by regressing the levels of contaminants measured on the first 10 PCs. They found the chlorinated pesticide metabolite DDE to be associated with the first four PCs and the parent pesticide DDT to be associated with PC 9 [28].

2. Factor analysis (FA)

Factor analysis aims to explain the covariance among variables and assumes that the underlying correlation patterns between variables arise from a few common latent variables called factors. The observed variables are defined as linear combinations of the factors, revealing the underlying constructs that give rise to the observed phenomenon. Then, factor loadings can be used to discern a pattern in the original variables, like exposomic features. Thus, FA finds a new set of variables, fewer in number than the original variables while expressing that which is common among the original variables. The analysis can be exploratory, where no assumptions are made about relationships among factors. It can also be confirmatory to test the hypothesis that the variables are associated with specific factors [29,30]. Most studies of the exposome take an exploratory approach, thus scientific concept or hypothesis underlying the factors is secondary to the analysis. Different models and methods of calculation can be used for the analyses, contributing to diversity in the methodology and results from factor analysis [31]. The analyst is required to choose the optimal number of factors that explain the data using model diagnostics, like the Bayesian information criterion (BIC). Confirmatory factor analysis relies on sophisticated math to confirm or test for generalizations [32]. Labelling interpretable factors identified as part of the underlying structure is a difficulty in factor analysis. It is a struggle to “see what there is and not what we want to see” [31].

Juarez and colleagues (2017) investigated the association between 2162 environmental exposures and lung cancer mortality rates in more than 2000 US counties using graph theoretical algorithms and factor analysis. They first computed correlation coefficients between each exposure-outcome pair, then used a clique doubling technique [33] to determine a threshold for significant correlation ($r > |0.14|$). They extracted paracliques [34,35] from graphs created using the significant threshold. To eliminate redundant information contained in the paracliques, the authors used factor analysis with varimax (orthogonal) rotation to obtain 172 factors that were subsequently used in stepwise regression on each of their outcomes of interest [36].

3. Non-negative matrix factorization (NMF)

In FA and PCA, latent variables and their errors are assumed normally distributed, and are thus not really suitable for non-negative data, such as exposome data. Non-negative matrix factorization was proposed by Lee and Seung as a solution [37,38]. This method works well with non-negative data that have excess zeros and measurement error, like measurement of chemicals and metabolites through HRMS [39]. NMF works to factorize a data matrix into two matrices, a basis matrix and a coefficient matrix—both are constrained to have non-negative values. The dimensions of the new matrices after factorization, i.e. number of underlying factors, has to be set by the analyst. No orthogonality constraints are placed on the basis components, allowing them to overlap. This overlap can be used to determine molecules that belong to multiple pathways or processes. The factorization produces sparse results with only a few non-zero entries [40]. The method has gained traction in systems biology and has been applied to transcriptomics data. According to Stein-O'Brien and colleagues, NMF learns two matrices—one that describes the structure between the molecules, genes or metabolites, and the other describes the structure between the different samples, i.e., each observation. They describe the first matrix as the amplitude matrix and the latter as the pattern matrix. The value in each column of the amplitude matrix can be thought to represent the relative contribution of a molecule in each inferred factor, which may be used to distinguish different complex biological processes or function pathways [41].

Béchaux and colleagues used NMF to extract patterns in the mixture of pesticides that derive from the major food consumption systems found in the French diet. They subsequently used the data to show clusters of individuals with similar consumption habits and exposure to pesticides. In order to determine the number of underlying latent structures (k), they ran NMF using different values of k and used residual sum of squares and the Bayesian information criterion (BIC) to choose the appropriate number, k=10 [39].

PCA, FA or NMF? The choice depends on the research question

When to use one of these methods depends mainly on the research question of interest. If the goal is to capture the total variance or a proportion of the variance in the data, PCA is the appropriate technique to employ. If the goal is to uncover a latent structure, an analyst can employ either NMF or factor analysis. NMF allows for better interpretation of the reduced dimensions because of the non-negative constraint; however, one can run into identifiability problems. To the best of our knowledge, NMF has not yet been applied in an exposomic context. In PCA, the analyst would have to decide on the appropriate number of components to extract from the solution based on *a priori* decided criteria. In factor analysis and NMF the number of factors is an analyst-defined input and different runs for different numbers of factors should be conducted before deciding on the optimal one, either in terms of data fit or/and interpretability of the solution. Nguyen and Holmes offer useful tips for dimensionality reduction [42]

A PCA Application: An HRMS Example

To illustrate an example of exposomic data dimensionality reduction, we analyzed respiratory exposures present in air using passive silicone badge samples and an untargeted GC-orbitrap assay developed for measuring the human exposome. Silicone badge samplers were placed in groups of four throughout different locations in two houses and a laboratory environment to evaluate the room and residence specific exposome near Atlanta, GA. The wristbands passively sampled chemicals present in the air of these locations for 7 days. These low-cost passive samplers allow access to micro- and personal-exposure from a large population and have potential as a key, exposome sampler for population research. Following deployment, samplers were weighed, placed in amber vials and extracted with 1mL ethyl acetate at room temperature with gentle shaking for 24 hours. The extracts were transferred to a GC autosampler vial, and analyzed in duplicate using full scan mode over *m/z* range 85-850 and a 15m Agilent DB-5MSUI capillary column with the following temperature gradient: 100°C for 1 min, increased to 180°C at 25°C/min; followed by a temperature ramp to 215°C at 5°C/min, and finally increased to 300°C at 25°C/min and held for 10 min, resulting in a total run time of 26.6 min. Detected ion signals were extracted using XCMS [43] at two different parameter settings and merged using xMSanalyzer [44].

A total of 49,585 features were detected in the 78 samplers from three different locations. In this example, we present results from one location, a house where samplers were placed in 6 different areas throughout the residence, a total of 23 samplers. Using a blank badge sampler as background, we retained a feature in the feature table if it was present at twice the blank intensity in at least 80% of the badge samplers. This reduced the number of detected features to 1347. The dimensions, thus, of the feature table for PCA was 23×1347. Missing data were imputed with half the lowest abundance of that feature. Data were $\log_{10}$-transformed and used as input to the prcomp() function in R (version 3.6.0). The data were centered and scaled. The first five PCs explained 81.2% of the total variance. The biplot (Figure 2) shows grouping of exposomic features derived from wristbands placed in the same location.

Discussion and conclusion

We discussed the application of three different unsupervised dimension reduction techniques that can be used in exposome research. While unsupervised, FA and NMF require analyst input to decide the optimal number of underlying factors that describe the observed phenomenon, but they reduce the number of independent variables. While PCA does not require analyst input during orthogonal transformation, input is required when selecting the number of PCs to extract and include in further analyses, since the total number of PCs equal the number of observations in original dataset. All three methods are sensitive to scale and data should be standardized before dimension reduction.

While dimensionality reduction removes redundant information, such as multiple ions present from one chemicals or highly correlated metabolites from the same biological pathway, these redundant signals are informative when evaluating chemical identifications and evaluating systemic biological response to exposure [45]. Further, simply creating new variables that are linear combinations of other variables can hinder interpretability of results.

Chemical risk assessment is difficult to accomplish with linear combinations of multiple chemicals and researchers must attempt to report interpretable results from high-dimensional data. Lastly, generalizing latent structures from observed data should be performed cautiously. As noted by Alexandre Dumas (Junior), “all generalizations are dangerous, even this one” [32]. Here, we highlighted a few unsupervised methods for dimensionality reduction; however, to identify exposomic patterns specific to outcomes of interest, supervised extensions will likely need to be incorporated into the informatic workflow, e.g., the supervised PCA [46]. This paper illustrates that exposome research requires approaches to analyze the data that are as sophisticated as the approaches used to generate the data. In the future, a detailed simulation study would provide an opportunity to assess the performance of different supervised and unsupervised dimensionality reduction methods in exposomic research, given potential different research questions, to best inform the choice of method for data analysis.

Acknowledgments

Funding Sources: NIH P30 ES009089, P30 ES023515, R01 ES028805, P30 ES019776, U2C ES030859, RC2DK118619.

REFERENCES

[1]. Rappaport SM, Implications of the exposome for exposure science, J Expo Sci Environ Epidemiol. 21 (2011) 5–9. 10.1038/jes.2010.50. [PubMed: 21081972]

[2]. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW, The Exposome: Molecules to Populations, Annual Review of Pharmacology and Toxicology. 59 (2019) 107–127. 10.1146/annurev-pharmtox-010818-021315.

[3]. Wild CP, Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology, Cancer Epidemiol Biomarkers Prev. 14 (2005) 1847–1850. 10.1158/1055-9965.EPI-05-0456. [PubMed: 16103423]

[4]. Wild CP, The exposome: from concept to utility, Int J Epidemiol. 41 (2012) 24–32. 10.1093/ije/dyr236. [PubMed: 22296988]

[5]. Miller GW, Jones DP, The Nature of Nurture: Refining the Definition of the Exposome, Toxicol Sci. 137 (2014) 1–2. 10.1093/toxsci/kft251. [PubMed: 24213143]

[6]. Miller G, The Exposome - 1st Edition, (2013). https://www.elsevier.com/books/the-exposome/miller/978-0-12-417217-3 (accessed January 11, 2019).

[7]. Vermeulen R, Schymanski EL, Barabási A-L, Miller GW, The exposome and health: Where chemistry meets biology, Science. 367 (2020) 392–396. 10.1126/science.aay3164. [PubMed: 31974245]

[8]. Walker DI, Uppal K, Zhang L, Vermeulen R, Smith M, Hu W, Purdue MP, Tang X, Reiss B, Kim S, Li L, Huang H, Pennell KD, Jones DP, Rothman N, Lan Q, High-resolution metabolomics of occupational exposure to trichloroethylene, Int J Epidemiol. 45 (2016) 1517–1527. 10.1093/ije/dyw218. [PubMed: 27707868]

[9]. Sobus JR, Grossman JN, Chao A, Singh R, Williams AJ, Grulke CM, Richard AM, Newton SR, McEachran AD, Ulrich EM, Using prepared mixtures of ToxCast chemicals to evaluate non-targeted analysis (NTA) method performance, Analytical and Bioanalytical Chemistry. 411 (2019) 835–851. [PubMed: 30612177]

[10]. O’Connell SG, Kincl LD, Anderson KA, Silicone Wristbands as Personal Passive Samplers, Environ. Sci. Technol 48 (2014) 3327–3335. 10.1021/es405022f. [PubMed: 24548134]

[11]. Vermeulen R, Schymanski EL, Albert-Laszlo B, Miller GW, The exposome and health: where chemistry meets biology., Science. In press (n.d.).

[12]. Alonso A, Marsal S, Julià A, Analytical Methods in Untargeted Metabolomics: State of the Art in 2015, Front. Bioeng. Biotechnol 3 (2015). 10.3389/fbioe.2015.00023.

[13]. Bellman R, Dynamic Programming, sixth, Princeton University Press, 1972 https://press.princeton.edu/books/paperback/9780691146683/dynamic-programming (accessed January 10, 2020).

[14]. Xue L, Stahura FL, Bajorath J, Cell-Based Partitioning, in: Bajorath J (Ed.), Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery, Humana Press, Totowa, NJ, 2004: pp. 279–289. 10.1385/1-59259-802-1:279.

[15]. Tibshirani R, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological). 58 (1996) 267–288.

[16]. Hoerl AE, Kennard RW, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics. 12 (1970) 55–67.

[17]. Zou H, Hastie T, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology). 67 (2005) 301–320.

[18]. Lê Cao K-A, Boitard S, Besse P, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems, BMC Bioinformatics. 12 (2011) 253 10.1186/1471-2105-12-253. [PubMed: 21693065]

[19]. Nguyen DV, Rocke DM, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics. 18 (2002) 39–50. 10.1093/bioinformatics/18.1.39. [PubMed: 11836210]

[20]. Bartenhagen C, Klein H-U, Ruckert C, Jiang X, Dugas M, Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data, BMC Bioinformatics. 11 (2010) 567 10.1186/1471-2105-11-567. [PubMed: 21087509]

[21]. Schölkopf B, Smola A, Müller K-R, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation. 10 (1998) 1299–1319.

[22]. Tenenbaum JB, De Silva V, Langford JC, A global geometric framework for nonlinear dimensionality reduction, Science. 290 (2000) 2319–2323. [PubMed: 11125149]

[23]. Kingma D, Welling M, Auto-Encoding Variational Bayes, in: 2014.

[24]. Jolliffe IT, Principal Component Analysis, 2nd ed., Springer-Verlag, New York, 2002 10.1007/b98835.

[25]. Liland KH, Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis, TrAC Trends in Analytical Chemistry. 30 (2011) 827–841. 10.1016/j.trac.2011.02.007.

[26]. Comon P, Independent component analysis, A new concept?, Signal Processing. 36 (1994) 287–314. 10.1016/0165-1684(94)90029-9.

[27]. Tamayo-Uria I, Maitre L, Thomsen C, Nieuwenhuijsen MJ, Chatzi L, Siroux V, Aasvang GM, Agier L, Andrusaityte S, Casas M, de Castro M, Dedele A, Haug LS, Heude B, Grazuleviciene R, Gutzkow KB, Krog NH, Mason D, McEachan RRC, Meltzer HM, Petraviciene I, Robinson O, Roumeliotaki T, Sakhi AK, Urquiza J, Vafeiadi M, Waiblinger D, Warembourg C, Wright J, Slama R, Vrijheid M, Basagaña X, The early-life exposome: Description and patterns in six European countries, Environment International. 123 (2019) 189–200. 10.1016/j.envint.2018.11.067. [PubMed: 30530161]

[28]. Li S, Cirillo P, Hu X, Tran V, Krigbaum N, Yu S, Jones DP, Cohn B, Understanding mixed environmental exposures using metabolomics via a hierarchical community network model in a cohort of California women in 1960's, Reproductive Toxicology. (2019). 10.1016/j.reprotox.2019.06.013.

[29]. Thompson B, Exploratory and confirmatory factor analysis: Understanding concepts and applications., American Psychological Association, 2004.

[30]. Mulaik SA, A Brief History of the Philosophical Foundations of Exploratory Factor Analysis, Multivariate Behav Res. 22 (1987) 267–305. 10.1207/s15327906mbr2203_3. [PubMed: 26776378]

[31]. Cattell RB, Extracting Factors: The Algebraic Picture, in: Cattell RB (Ed.), The Scientific Use of Factor Analysis in Behavioral and Life Sciences, Springer US, Boston, MA, 1978: pp. 15–39. 10.1007/978-1-4684-2262-7_2.

[32]. Child D, The Essentials of Factor Analysis, A&C Black, 2006.

[33]. Borate BR, Chesler EJ, Langston MA, Saxton AM, Voy BH, Comparison of threshold selection methods for microarray gene co-expression matrices, BMC Research Notes. 2 (2009) 240. [PubMed: 19954523]

[34]. Hagan RD, Langston MA, Wang K, Lower bounds on paraclique density, Discrete Applied Mathematics. 204 (2016) 208–212. [PubMed: 27057077]

[35]. Chesler EJ, Langston MA, Eskin E, Ideker T, Raphael B, Workman C, Systems biology and regulatory genomics, (2006).

[36]. Juarez PD, Hood DB, Rogers GL, Baktash SH, Saxton AM, Matthews-Juarez P, Im W, Cifuentes MP, Phillips CA, Lichtveld MY, Langston MA, A novel approach to analyzing lung cancer mortality disparities: Using the exposome and a graph-theoretical toolchain, Environ Dis. 2 (2017) 33–44. [PubMed: 29152601]

[37]. Lee DD, Seung HS, Learning the parts of objects by non-negative matrix factorization, Nature. 401 (1999) 788–791. 10.1038/44565. [PubMed: 10548103]

[38]. Saul LK, Lee DD, Multiplicative Updates for Classification by Mixture Models, in: Dietterich TG, Becker S, Ghahramani Z (Eds.), Advances in Neural Information Processing Systems 14, MIT Press, 2002: pp. 897–904. http://papers.nips.cc/paper/2085-multiplicative-updates-for-classification-by-mixture-models.pdf.

[39]. Béchaux C, Zetlaoui M, Tressou J, Leblanc J-C, Héraud F, Crépet A, Identification of pesticide mixtures and connection between combined exposure and diet, Food and Chemical Toxicology. 59 (2013) 191–198. 10.1016/j.fct.2013.06.006. [PubMed: 23774259]

[40]. Gaujoux R, Seoighe C, A flexible R package for nonnegative matrix factorization, BMC Bioinformatics. 11 (2010) 367 10.1186/1471-2105-11-367. [PubMed: 20598126]

[41]. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, Xu Y, Fertig EJ, Enter the Matrix: Factorization Uncovers Knowledge from Omics, Trends in Genetics. 34 (2018) 790–805. 10.1016/j.tig.2018.07.003. [PubMed: 30143323]

[42]. Nguyen LH, Holmes S, Ten quick tips for effective dimensionality reduction, PLOS Computational Biology. 15 (2019) e1006907 10.1371/journal.pcbi.1006907. [PubMed: 31220072]

[43]. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem 78 (2006) 779–787. 10.1021/ac051437y. [PubMed: 16448051]

[44]. Uppal K, Soltow QA, Strobel FH, Pittard WS, Gernert KM, Yu T, Jones DP, xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data, BMC Bioinformatics. 14 (2013) 15 10.1186/1471-2105-14-15. [PubMed: 23323971]

[45]. Boccard J, Rudaz S, Harnessing the complexity of metabolomic data with chemometrics, Journal of Chemometrics. 28 (2014) 1–9. 10.1002/cem.2567.

[46]. Barshan E, Ghodsi A, Azimifar Z, Zolghadri Jahromi M, Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds, Pattern Recognition. 44 (2011) 1357–1371. 10.1016/j.patcog.2010.12.015.

Highlights

- High-resolution mass spectrometry (HRMS) has enabled measurement of exposomic features in biomatrices
- HRMS poses data complexities that require dimensionality reduction
- Principal components analysis, factor analysis, and non-negative matrix factorization have been successfully applied in exposomic data analysis
- Interpretability of results after dimensionality reduction is essential for environmental policy and risk assessment

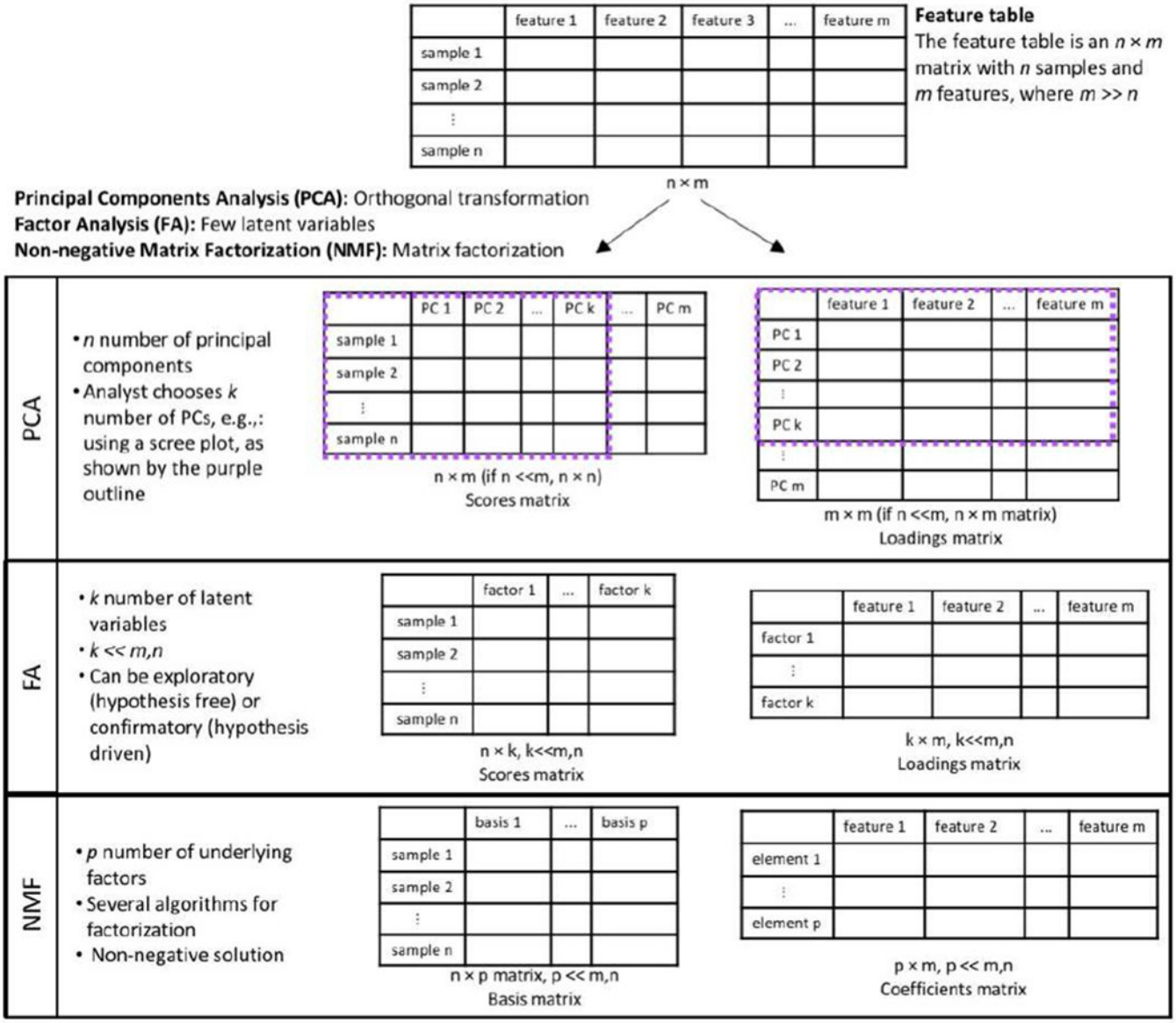

Figure 1.
Visual representation of dimension reduction using PCA, FA, and NMF.

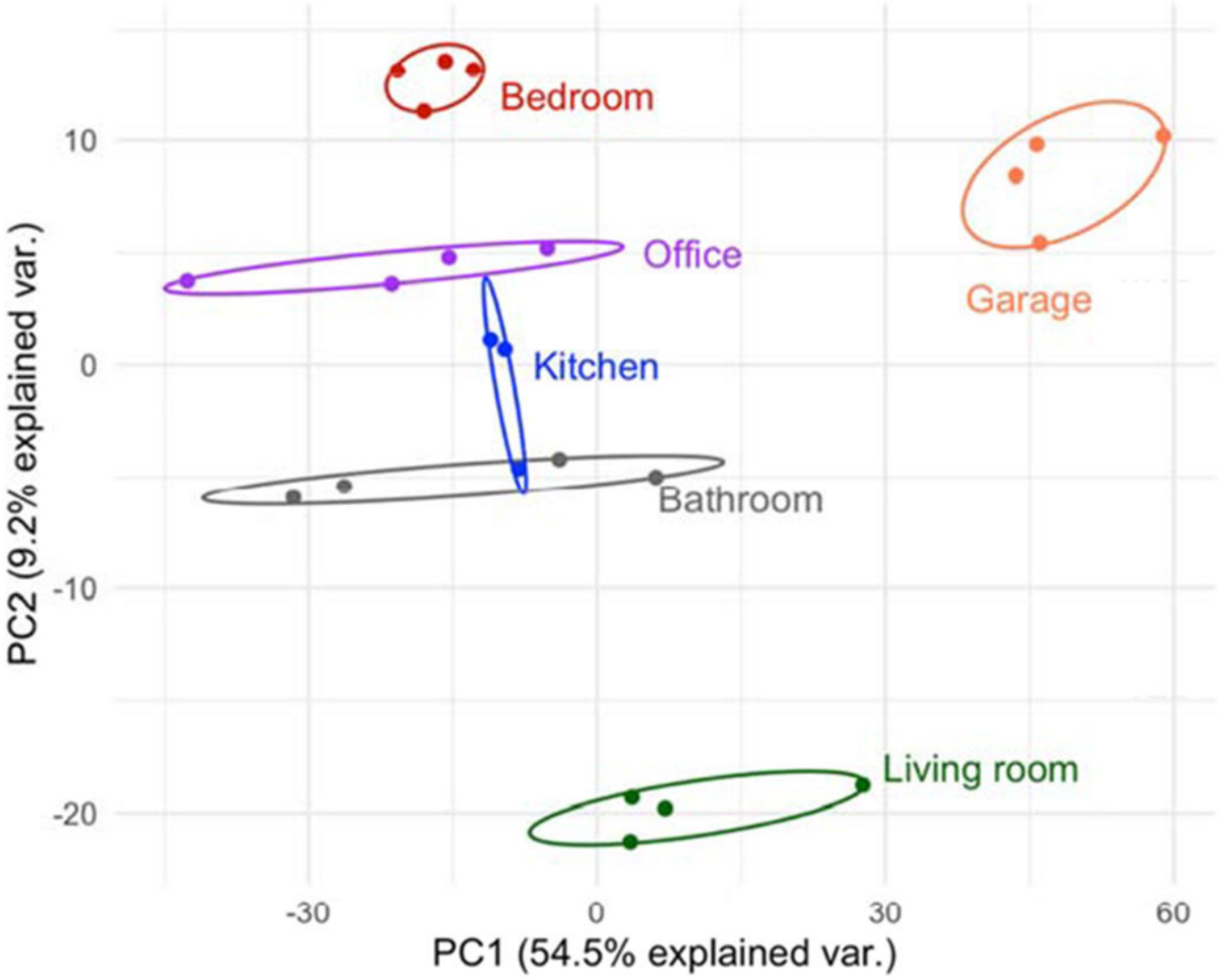

Figure 2.
Biplot: Principal Component 1 (PC1; explained 54.5% of the total variance in the data) plotted against PC2 (explained 9.2% of the variance) showing clustering of samplers placed in the same location of house 2.

Table 1.

Glossary of terms.

Term	Definition
Variance	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$; sample variance is usually interpreted as the average squared deviation from the mean
Covariance	$cov(x,y)=\sum_{i=1}^{N}\frac{(x_i-\bar{x})(y_i-\bar{y})}{N}$; a measure used to quantify the relationship between two random variables
Correlation	$cor(x,y)=\frac{cov(x,y)}{\sigma_x\sigma_y}$; *where σ denotes standard deviation;* unlike covariance, correlation is a unitless measure of the relationship between two random variables
Orthogonal transformation	A linear transformation is called orthogonal if it preserves the length of the vectors; $\| T(\overrightarrow{x}) \| = \| \overrightarrow{x} \|$. In PCA, the solution forces each component to be orthogonal to the previous, i.e., independent.
Matrix factorization	Decomposition of a matrix into the product of two or more lower-dimension rectangular matrices
Latent variable	An unobserved (“hidden”) variable that is inferred through observed variables
Bayesian Information Criterion (BIC)	Determines model fit by considering the likelihood function of a model, number of data points, and number of free parameters to be estimated. It is a criterion used for model selection; the model with lowest BIC is preferred.
Clique	In graph theory, a complete subgraph is called a clique. A graph is complete when every pair of distinct vertices—a corner or a point where lines meet—is connected by a unique edge, i.e., every vertex has an edge to every other vertex.
Paraclique	It consists of a clique and all vertices with at least some proportion of edges to the clique. It is considered a relaxation of clique.