Intuitive Data-Driven Visualization of Food Relatedness via t-Distributed Stochastic Neighbor Embedding

Kaylen J. Pfisterer University of Waterloo, ON, Canada
Robert Amelard Schlegel-UW Research Institute for Aging, ON, Canada
Alexander Wong University of Waterloo, ON, Canada

Abstract

The relationship between diet and health is important, yet difficult to study in practice. Dietary pattern analysis is one method for investigating this link; having more variety in diet tends to be beneficial and a score can be generated based on a heuristic approach to food intake habits. We aim to enhance the intuition behind these food scores by creating an intuitive data-driven visualization of food relatedness by leveraging t-distributed stochastic neighbor embedding (t-SNE). More specifically, by performing t-SNE analysis in a controlled manner to project the high-dimensional nutritional information of food items into a lower dimensional food similarity space, the natural clustering of foods based on the underlying nutritional composition becomes visually observable. The efficacy of this data-driven approach for visualizing food relatedness was investigated on a total of 8549 food item entries in the USDA food composition database, with the results showing considerable promise as a tool for gaining important nutritional insights. This is the first step toward providing a novel method to enhance dietary pattern analysis with additional context and insight into food intake habits based on the inherent nutritional content of the foods consumed.

1 Introduction

Studying the relationship between diet and health is known as nutritional epidemiology and is difficult to accurately and cost-effectively study in practice; the main barrier to advances in nutritional epidemiology is the ability to measure diet [1]. While methods exist for measuring nutritional intake (e.g., food frequency questionnaires, food diaries, 24-hour recall, dietary pattern analysis [2, 3], and biochemical analyses [3]), these methods fundamentally rely on self-reporting [4, 5] or are too expensive to be used in practice [3]. There has been a shift towards dietary pattern analysis based on food intake in an effort to capture the degree to which a certain diet type (e.g., Mediterranean) is followed. The rationale is that people eat a variety of foods and form food habits. These food habits consist of higher or lower intake frequency of certain foods which in turn contain nutrients with health benefits [6]. For example, the Mediterranean diet emphasizes plant based foods high in unsaturated fats and polyphenols with growing evidence touting its protective effects against cardiovascular disease [7].

Dietary pattern analysis is typically conducted in one of two ways: 1) a-priori through experts' assessment of healthy foods based on the present state of evidence, or 2) a-posteriori by identifying dietary habits (food similarities) through k-means clustering, principal component analysis (PCA), or cluster analysis [6]. Despite their general utility, there are a number of limitations and challenges with existing a-priori and a-posteriori methods that need to be addressed to further enhance dietary pattern analysis. Challenges with current a-priori methods revolve mainly around the current state of available knowledge. In terms of current a-posteriori methods, one challenge that arises is the need to define the number of desired clusters [8], which introduces bias to the analysis process through the assumption that the correct number of clusters was known. Furthermore, PCA and cluster analysis rely on the singular value decomposition of each principal component in which each point is projected onto a line of best fit passing through the origin determined by the maximised sum of squared distances of the projected points to the origin. The scaled x and y components of the lines (i.e., the linear combinations of the variables) are used to determine which variable(s) contribute the most to the total variance (i.e., the degree to which each variable explains the spread of points). From here, dimensionality reduction is accomplished by selecting a subset of principal components to retain the majority of variance using a simpler model. However, these approaches inherently rely on estimates (through the singular value decomposition) of the data which may not capture higher-order interactions since not all variability is retained. These limitations may result in a reduced likelihood of observing potential synergistic effects of nutritional interactions [6]).

One approach that is gaining increasing popularity in a number of different fields is t-distributed stochastic neighbor embedding (t-SNE) [9]. No assumptions about underlying data are made and, instead of linear combination of the variables, t-SNE relies on pairwise comparisons of points and using pairwise similarities to map the relatedness of each point to all other points in high-dimensional space into a low-dimensional space in which the similarity is preserved. As a result, more complex interaction can be retained via t-SNE as no approximations of the distribution are used. This may allow for more natural groupings based solely on the underlying nutritional composition used as features to describe a food. An additional benefit of t-SNE is for data visualization which, in this case depicts food relatedness; when used on food intake data this may also be hypothesis generating to inform future directions for nutritional epidemiological study.

Motivated by this, the focus of this paper is to explore the potential for enhancing dietary pattern analysis by creating an intuitive data-driven visualization of food relatedness via t-distributed stochastic neighbor embedding (t-SNE). More specifically, by performing t-SNE analysis in a controlled manner to project the high-dimensional nutritional information of food items into a lower dimensional food similarity space, the natural clustering of foods based on the underlying nutritional composition becomes visually observable, thus empirically optimizing interpretability of food relatedness visualizations. To the best of the authors' knowledge, this is the first study to investigate the efficacy of t-SNE for enhancing dietary pattern analysis.

The paper is organized as follows. Section 2 provides a detailed description of the proposed approach for intuitive data-driven visualization of food relatedness. Section 3 presents illustrative examples of the visualizations of food relatedness that can be produced using the proposed method, as well as a discussion of the results along with future directions to explore. Conclusions are drawn in Section 4.

2 Methods

The proposed method for enhancing dietary pattern analysis via intuitive data-driven visualization of food relatedness can be broken down into two main components. First, t-SNE analysis is performed in a controlled manner to project the high-dimensional nutritional information of food items into a lower dimensional space. Second, the low-dimensional projection characterizing the clustering of foods based on the underlying nutritional composition is visualized. These two components are described in detail below.

High-dimensional nutritional information space projection via t-SNE. To achieve the goal of enabling intuitive data-driven visualization of food relatedness, we aim to build a low-dimensional food similarity space based on high-dimensional standardized nutrient information features. Importantly, an unsupervised method was developed so that the learned similarity space was built independently of any known food labels, using only the known nutrition content of the food items. Specifically, given a set of N high-dimensional nutrient representations of food items $X = \{x_1, \ldots, x_N \mid x_i \in \mathbb{R}^D\}$, a mapping was learned for a lower d-dimensional food similar space with $Y = \{y_1, \ldots, y_N \mid y_i \in \mathbb{R}^d\}$ such that $d \ll D$. X was built using the abbreviated United States Department of Agriculture (USDA) food composition database version SR28 from 2015 [10]. Baby food was omitted yielding a dataset of 8549 unique food items. Each of the 47 nutrient information was used as a feature to describe that food item, yielding $X \in \mathbb{R}^{47}$. A low-dimensional food similarity space was learned via t-distributed stochastic neighbor embedding (t-SNE) [9], a statistical method for projecting high-dimensional data in low-dimensional space through the measurement of pairwise similarities of the inputs and under a distribution that measures pairwise similarities of the corresponding low-dimensional points

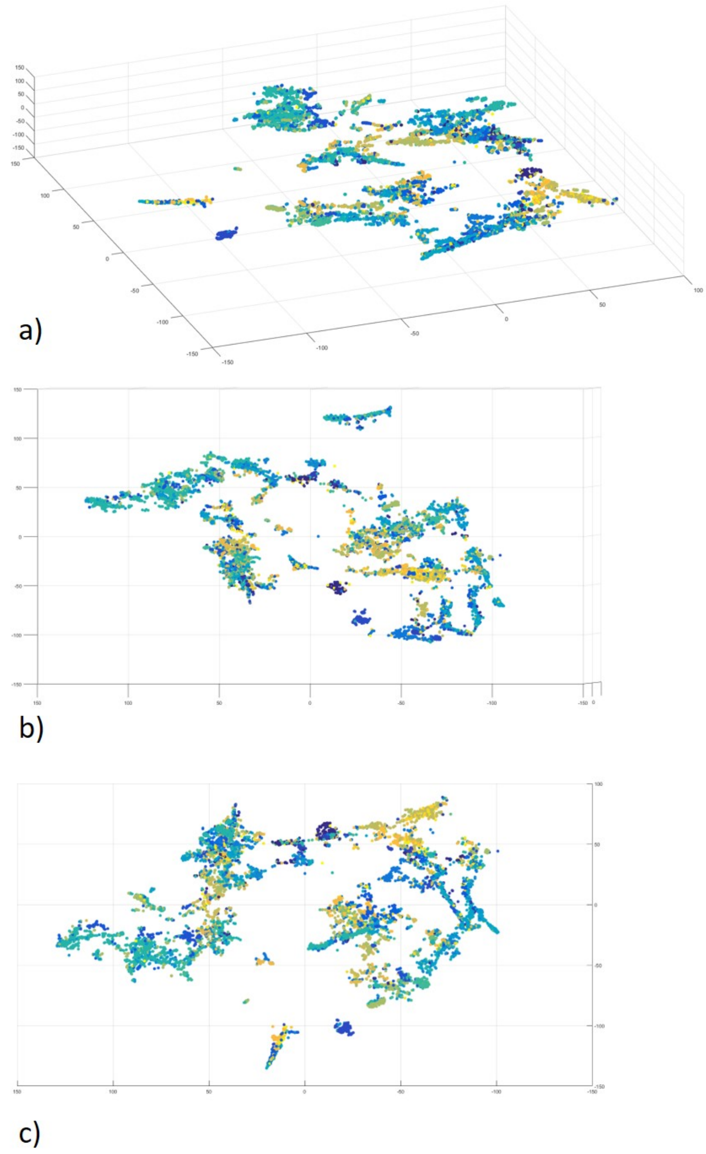

Fig. 1: Intuitive data-driven visualization of the food items in the USDA food database via t-SNE in a controlled manner (perplexity=50, 2000 iterations) using each nutrient value as a feature. Natural clustering exists as shown in three rotations of the clustered cloud in a), b), and c). No parameters were required to achieve this natural clustering. Colours applied after clustering for visualization purposes only.

in the embedding. Specifically, the mathematical similarity of two food items x_i and x_j was modeled according to a conditional neighborhood probability of their inherent nutrient composition under a Gaussian relationship assumption:

$$p_{j|i} = \frac{\exp(||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{1}$$

where σ_i is the standard deviation of the Gaussian kernel centered at the data point x_i, and is determined such that σ_i produces a probability distribution with a fixed perplexity [9]. The low-dimensional space was learned through an iterative gradient descent solution to minimize the Kullback-Leibler divergence between high-dimensional and low-dimensional similarity [9]. To enable the projection of high-dimensional nutritional information into a observable low-dimensional space in a controlled manner, a subspace of perplexity values was searched to empirically determine the optimal perplexity value yielding the greatest separability between clusters with minimal variance within clusters. Specifically, a linear grid search was performed across perplexity values $\{10+20r \mid r \in \mathbb{N}, r \leq 4\}$.

Visualization of food clustering in low-dimensional nutritional information projection. For visualization purposes, each food item in the database was systematically assigned a master category based on the text preceding the first comma in the short description provided in the database. For example, the master category for the food item "BUTTER,WHIPPED,W/ SALT" was "BUTTER". To run the analysis to see the foods clustering, the food group labels were not used as the above method is an unsupervised method; instead, the clustering was conducted using only the nutritional composition. After clustering, colour was applied based on the master category from the short description to demonstrate groups of foods within the same master category.

3 Results and Discussion

Using each of the nutritional components to describe each food type, a natural clustering of foods was observed using the proposed intuitive data-driven approach to visualizing food relatedness. The perplexity value which empirically optimized separability for enhanced interpretability was determined to be 50 for this dataset. An example of the global structure of food relatedness with separability is shown with perplexity 50 in Figure 1. Upon further exploration, these clusters represent semantic relevance as shown in Figure 2. Several examples of inherently semantically grouped food items have been clustered based on their underlying nutritional similarities. For example in Figure 2 clusters comprised of: a meat cluster (e.g., veal, beef, bison, turkey, chicken) as indicated in teal; a high fat oil cluster (e.g., fish oil, sunflower oil, shortening etc) as indicated in light blue; a cheese cluster (e.g., ricotta, mozzarella, processed cheese, American cheese) as indicated in navy; leafy greens (e.g., spinach, collard greens, lettuce, broccoli, cabbage) indicated in turquoise; and a mixed dish consisting of meat, a carbohydrate and tomato (e.g., chili con carne, hamburger, pizza, pasta with red sauce, meat lasagna). Within the cheese cluster, certain ice creams were also clustered which may be explained by the high fat dairy composition.

Of note, relative similarities in colour do not reflect nutritional similarities. For example a light blue and a darker blue are just as similar as a light blue and an orange dot. Encoding semantic similarity not only through clustering, but also through colour similarities would be interesting and enhance utility of this method in practice. As such this is part of future work. Future directions will also include enhanced gerneralizability of nutritional information captured in the food database through the inclusion of additional food databases such as the Canadian Nutrient File [11]. Finally, for improving comparative analysis of the accuracy and precision of t-SNE for the purpose of visualizing food relatedness for enhanced dietary pattern analysis, a parametric manifestation of t-SNE will be leveraged. This will enable a mapping of dietary patterns into the t-SNE space of the food databases to directly compare accuracy of a-priori and a-posteriori methods for dietary pattern analysis within the context of food relatedness more broadly.

4 Conclusions

The proposed approach to intuitive data-driven visualization of food relatedness via tSNE provides a novel method in the domain of nutrition for exploring food-relatedness and may enhance the interpretability of food relatedness through this visualization technique which, with further refinement, may provide dietitians with a novel method to more easily interpret an individual's food variety score.

Acknowledgments

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs program. The authors would also like to thank Professor Heather Keller (Research Chair in Nutrition and Aging, Schlegel-UW Research Institute for Aging) for her input into the application context for this approach.

References

[1] D. R. Thomas, W. Ashmen, J. E. Morley, and W. J. Evans, “Nutritional management in long-term care: development of a clinical guideline,” *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **55**(12), pp. M725–M734, 2000.

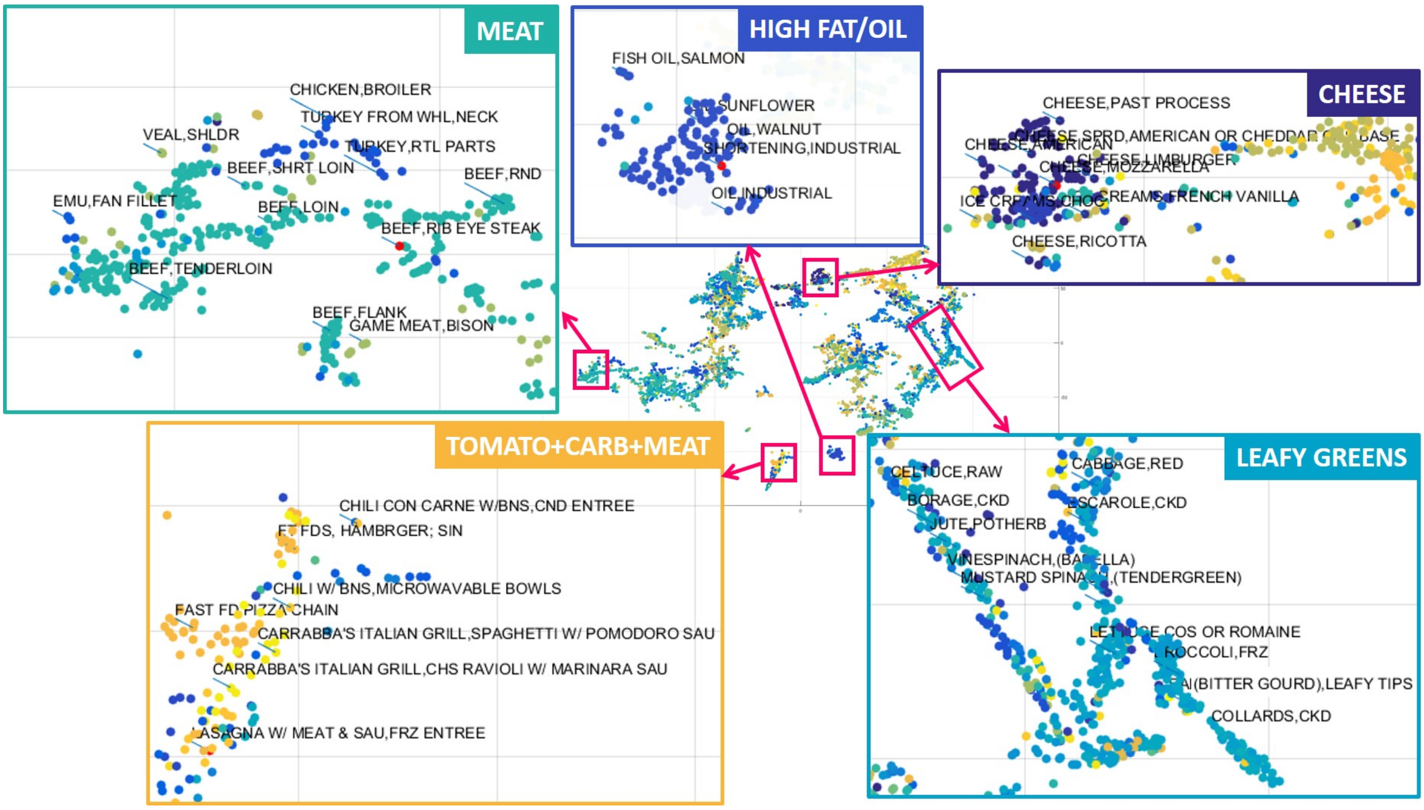

Fig. 2: Examples of the clustering observed in the proposed approach for intuitive data-driven visualization of food relatedness. The centre is a replica of Figure 1c with enlarged clusters to show sample labels. Clustering has semantic meaning. Highlighted here are clusters of meat, high fat/oil, cheese, leafy greens, and mixed dishes comprised of tomato, carbohydrates (e.g., pasta, bread, beans) with meat.

[2] L. Penn, H. Boeing, C. J. Boushey, L. O. Dragsted, J. Kaput, A. Scalbert, A. A. Welch, and J. C. Mathers, "Assessment of dietary intake: Nugo symposium report," *Genes & nutrition* **5**(3), p. 205, 2010.

[3] W. Willett, *Nutritional epidemiology*, Oxford University Press, 2012.

[4] C. K. Martin, H. Han, S. M. Coulon, H. R. Allen, C. M. Champagne, and S. D. Anton, "A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method," *British Journal of Nutrition* **101**(3), pp. 446–456, 2008.

[5] D. A. Williamson, H. R. Allen, P. D. Martin, A. J. Alfonso, B. Gerald, and A. Hunt, "Comparison of digital photography to weighed and visual estimation of portion sizes," *Journal of the American Dietetic Association* **103**(9), pp. 1139–1145, 2003.

[6] D. Panagiotakos, "α-priori versus α-posterior methods in dietary pattern analysis: a review in nutrition epidemiology," *Nutrition bulletin* **33**(4), pp. 311–315, 2008.

[7] M. A. Martínez-González, J. Salas-Salvadó, R. Estruch, D. Corella, M. Fitó, E. Ros, P. Investigators, *et al.*, "Benefits of the mediterranean diet: insights from the predimed study," *Progress in cardiovascular diseases* **58**(1), pp. 50–60, 2015.

[8] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[9] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research* **9**(Nov), pp. 2579–2605, 2008.

[10] N. D. L. US Department of Agriculture, Agricultural Research Service, *USDA National Nutrient Database for Standard Reference, Release 28 (Slightly revised). Version Current: May 2016*, 2018. Internet: `http://www.ars.usda.gov/ba/bhnrc/ndl`.

[11] H. Canada, *Canadian Nutrient File*, 2015. Internet: `http://www.healthcanada.gc.ca/cnf`.