Publishing Nutrition Research: A Review of Multivariate Techniques—Part 3: Data Reduction Methods

Philip M. Gleason, PhD; Carol J. Boushey, PhD, MPH, RD; Jeffrey E. Harris, DrPH, MPH, RD, LDN; Jamie Zoellner, PhD, RD

ARTICLE INFORMATION

Article history:
Submitted 1 September 2014
Accepted 6 March 2015
Available online 30 April 2015

Keywords:
Principal component analysis
Factor analysis
Reduced rank regression
Cluster analysis
Eigenvalue

2212-2672/Copyright © 2015 by the Academy of Nutrition and Dietetics.
http://dx.doi.org/10.1016/j.jand.2015.03.011

ABSTRACT

This is the ninth in a series of monographs on research design and analysis, and the third in a set of these monographs devoted to multivariate methods. The purpose of this article is to provide an overview of data reduction methods, including principal components analysis, factor analysis, reduced rank regression, and cluster analysis. In the field of nutrition, data reduction methods can be used for three general purposes: for descriptive analysis in which large sets of variables are efficiently summarized, to create variables to be used in subsequent analysis and hypothesis testing, and in questionnaire development. The article describes the situations in which these data reduction methods can be most useful, briefly describes how the underlying statistical analyses are performed, and summarizes how the results of these data reduction methods should be interpreted.
J Acad Nutr Diet. 2015;115:1072-1082.

THIS ARTICLE REPRESENTS THE NINTH IN A SERIES exploring the importance of research design, statistical analysis, and epidemiologic methods in nutrition and dietetics research. The purpose of this series is to help nutrition and dietetics practitioners apply and interpret analytic and scientific principles consistent with high-quality nutrition research in their own work. Research is the foundation of the dietetics profession, providing the basis for decisions in practice and policy. This series uses examples relevant to the field of nutrition and dietetics. An effort is made to appeal to the seasoned researcher as well as the nutrition research novice.

The purpose of this monograph is to provide an overview of data reduction methods, and it represents the third installment in a set of three articles on multivariate statistical techniques.[1,2] Nutrition researchers are often faced with the challenge of summarizing with a few simple variables complex concepts, such as diet quality, for which there are a large number of individual measures. Data reduction methods help researchers face this challenge by creating new variables that more efficiently summarize the large quantity of information originally available or use that information efficiently in subsequent analysis. One example outside of nutrition where these techniques are used is in the analysis of clusters of genes. There are thousands of genes that could be studied; data reduction methods have enabled researchers to reduce this number to smaller groups for further focus or analysis. Four major data reduction techniques will be covered in this article: principal components analysis (PCA), factor analysis (FA), reduced rank regression, and cluster analysis. Figure 1 provides a glossary of the relevant terminology.

Data reduction methods are set apart from other quantitative statistical methods covered in this series in that these methods themselves do not fall within the category of inferential statistics that would include statistical significance testing. Rather, these methods represent techniques researchers use to manipulate data they have collected for further analysis. The techniques include different algorithms that group variables of interest or sample members into underlying correlated or clustered groups according to well-defined rules set a priori by the investigator. In each of these analyses, the researcher plays an important role in guiding the selection and manipulation of variables and/or grouping of sample members.

In the field of nutrition, data reduction methods can be used for three general purposes. First, researchers often use these techniques for descriptive purposes alone; to summarize dietary patterns; or analyze the relationships among multiple foods, nutrients, or combinations of foods and nutrients. Research using the concept of dietary patterns rather than the analysis of individual nutrients is useful as people consume combinations of foods containing multiple nutrients vs individual nutrients alone. The dietary patterns paradigm represents a more comprehensive characterization of the diets of individuals or groups. Examples of research using data reduction methods to identify and summarize dietary patterns are provided by Nettleton and colleagues[4] and Reedy and colleagues.[5]

Second, data reduction methods can be used to create variables to be used in subsequent analysis and hypothesis testing. Although it might be impractical to test hypotheses involving dozens of measures of dietary intake, for example, it is more feasible to test hypotheses involving a single variable representing an overall dietary pattern created using data reduction methods.

A third common use of these methods is in questionnaire development. PCA and cluster analysis can be applied when developing a questionnaire to reduce the number of questionnaire items (variables) to a subset of items that best captures variation within the target population in the underlying concepts of interest. When considering questionnaire development, there may be many salient items of interest that could be included in the questionnaire; for example, those ascertained from focus groups. PCA and cluster analysis can help identify those items that correlate well with each other and with the underlying concepts of interest and guide elimination of those that do not. The factors that remain make good candidates for a questionnaire. For example, Glanz and Steffen[6] used cluster analysis as part of the development process of a questionnaire to assess psychosocial constructs related to calcium consumption among adolescents.

UNDERSTANDING AND PERFORMING THE TECHNIQUES

Principal Components Analysis and Factor Analysis

When researchers have a large number of potential variables to analyze and would like to summarize the information contained in those variables as efficiently as possible, PCA and FA are two closely related options for doing so. In each case, the method begins with a large number of "input variables" and ends with a much smaller number of variables—referred to as "principal components" or "factors"—that summarize the information in the input variables. This section describes PCA and briefly summarizes FA and its similarities to and differences with PCA.

PCA and FA might be used for a number of different reasons. One key distinction to make between different uses of these techniques involves whether the method is applied before or after the study's main data collection effort. A researcher or research team might use a data reduction technique like PCA or FA before collecting information from the full study sample in order to determine which questions to include in a survey instrument to best capture a particular construct of interest. In a situation like this, there may be a large number of candidate questions that capture some key aspect of the underlying concept, and the researcher would collect data on all candidate variables from a small subsample and conduct PCA or FA to identify the candidate questions that best "hang together" and reflect the underlying construct. Then, these questions would be included in a final survey instrument that could be administered to the full sample of interest.

In other situations, the researcher may be working with data that have already been collected on the full sample of interest, but wishes to whittle down a large group of input variables so that any subsequent analysis can be conducted more efficiently. Often, a small number of summary measures can provide more useful and easy-to-digest descriptive information about some underlying construct than a large number of input variables. If the constructs are intended to be used as covariates in a statistical technique such as a regression model, reducing the number of covariates can minimize the likelihood of a statistical problem known as multicollinearity, whereby high correlations among covariates make it difficult to identify their true relationship with the model's dependent variable.

Nettleton and colleagues[4] provide an example of a study that used PCA on already collected data in a sample of adults. The input variables in their analysis were dietary intake measures that were systematically condensed to 47 different food groups. Using PCA, the researchers created four principal components that they labeled as representing different dietary patterns. For example, the first principal component represented diets high in "fats and processed meats." They used these principal components both for descriptive purposes and in an analysis of the relationship between dietary patterns and markers of subclinical atherosclerosis.

Performing Principal Components Analysis. PCA can be performed with many statistical software packages. This section first describes the basic process for performing PCA and then provides several examples that illustrate these procedures. The starting point for researchers is a dataset that includes a large number of input variables that are distinct yet related in some way, such as numerous variables reflecting individuals' nutritional attitudes. The software will produce output that can be used to create a set of principal components, or summary variables that reflect some combination of the input variables included in the PCA. The researcher must interpret this output in order to determine the number of principal components to be created and determine how these principal components will be created and interpret the meaning of each, if possible.

Suppose the researcher begins with eight input variables and the goal is to summarize these variables with a smaller number of summary variables, or principal components. A useful preliminary step is to generate a correlation matrix that shows correlations between all of the input variables. This will provide a general sense of whether particular subsets of input variables tend to "hang together"; that is, to be highly correlated with one another. These are variables that will likely contribute most importantly to the same principal components.

A key step in the analysis involves the researcher determining the number of principal components to "retain." In other words, he or she must determine how many principal components will be used to summarize the full set of input variables. Technically, PCA will generate the same number of principal components as there are input variables in the analysis (eg, eight in the example given). However, each additional component identified will explain a smaller amount of the variation, and thus will become increasingly less useful. As shown here, researchers will typically "retain" the principal components that explain the most variation and drop from the analysis those that explain the least. Each principal component is a linear combination of the input

Algorithm: Series of steps to be taken to carry out a particular task or calculation.

Centroid: The geometric center of a two-dimensional area.

Centroid-Based Clustering: A type of algorithm used to conduct cluster analysis, also known as iterative partitioning. Under centroid-based clustering, an iterative process is used after the researcher initially determines the number of clusters to be constructed and defines an initial set of centroids. K-means clustering is one example of centroid-based clustering.[a]

Cluster Analysis: Set of algorithms or methods used to group a set of observations or cases into a set of clusters (categories, groups, trees, structures) where the cases in a given cluster are similar to one another and different from cases in other clusters with respect to some meaningful and predetermined set of characteristics or attributes.

Connectivity-Based Clustering: A type of algorithm used to conduct cluster analysis. Under connectivity-based clustering, clusters include data points that are all "connected" with one another based on having a sufficiently high degree of similarity with one another. It is also referred to as hierarchical clustering.[a]

Eigenvalue: A mathematical term that in the context of principal components analysis (or factor analysis or reduced rank regression) represents the total amount of variance in a set of input variables that is captured by a given principal component. Eigenvalues are standardized so that their average value is 1 and the sum of eigenvalues for a set of principal components is equal to the number of principal components created.

Eigenvector: A mathematical term that in the context of principal components analysis (or factor analysis or reduced rank regression) represents a set of "factor loadings" associated with a given principal component (factor). An eigenvector contains a separate factor loading for each input variable, and the factor loading represents the weight or importance of the input variable for that principal component (factor); input variables with the largest factor loadings (in absolute value) influence the principal component most strongly.

Factor Analysis: Set of procedures used to summarize the information contained in a group of input variables with a smaller number of variables, referred to as factors. Factor analysis results in the creation of eigenvalues and eigenvectors that can be used to construct the factors. Factor analysis is closely related to principal components analysis.

Factor Loading: Weight given to a particular input variable in constructing a principal component or factor. Input variables with positive factor loadings are positively correlated with the principal component; those with negative values are negatively correlated with the principal component. Factor loadings for each of the input variables are contained in the eigenvector of a particular principal component or factor.

Hierarchical Clustering: A type of algorithm used to conduct cluster analysis. Under hierarchical clustering, clusters include data points that are all "connected" with one another based on having a sufficiently high degree of similarity with one another. It is also referred to as connectivity-based clustering.[a]

Iterative Partitioning: A type of algorithm used to conduct cluster analysis, also known as centroid-based clustering. Under this algorithm, an iterative process is used after the researcher initially determines the number of clusters to be constructed and defines an initial set of centroids. K-means clustering is one example of iterative partitioning.[a]

k-Means Clustering: An algorithm used to conduct cluster analysis. Under k-means clustering, the researcher determines the number of clusters to be constructed (where k is equal to this number), and also chooses a starting point for each of the clusters to be created by selecting an initial set of k centroids. The algorithm then proceeds to group observations being analyzed into clusters based on their distance to the nearest centroid.[a]

Monte Carlo Analysis: A mathematical technique involving repeatedly conducting a specific analysis based on artificially constructed random samples, allowing researchers to determine the expected result of the analysis through simulation when it is not possible to determine the expected result theoretically.

(continued on next page)

Figure 1. Definitions of commonly used terms in data reduction methods. [a]Definition based on material in Aldenderfer and Blashfield.[3]

Principal Component: Variable that is produced by principal components analysis and summarizes the information contained in a larger number of input variables. Each principal component produced by principal components analysis is a weighted average of the input variables included in the analysis.

Principal Components Analysis: Set of procedures used to summarize the information contained in a group of input variables with a smaller number of variables, referred to as principal components. Principal components analysis results in the creation of eigenvalues and eigenvectors that can be used to construct the principal components. Principal components analysis is closely related to factor analysis.

Reduced Rank Regression: Set of procedures also known as maximum redundancy analysis that is used to summarize the information contained in a group of input variables with a smaller number of variables, referred to as factors or principal components. Reduced rank regression is closely related to principal components analysis and factor analysis, except that it derives factors by accounting for as much variation as possible in researcher-determined response variable(s).

Scree plot: Graphical representation of the eigenvalues resulting from principal components (or related) analysis. Each eigenvalue is represented sequentially on the x axis of the plot, with the height of the plot showing its value. A scree plot is used by researchers in a scree test to determine the number of principal components to retain from the analysis.

Scree test: Test used in principal components (and related) analysis to determine the number of principal components to retain from the analysis. In a scree test, a researcher examines a scree plot to identify the point at which there is a large drop-off in eigenvalues (that is, where there is a large drop-off in the height of the plot). Principal components with larger eigenvalues before (to the left of) the drop-off are retained; others are dropped.

Figure 1. *(continued)* Definitions of commonly used terms in data reduction methods. [a]Definition based on material in Aldenderfer and Blashfield.[3]

variables, and the first principal component is the linear combination that explains the maximum amount of the total variance across all input variables. In other words, it is the combination that summarizes the full set of input variables better than any other possible linear combination. The second principal component explains the maximum amount of any remaining variance in the input variables not explained by the first principal component. The remaining principal components are defined accordingly.

For each principal component, the output of PCA will include a set of eigenvalues, including one eigenvalue for each principal component that has been created. Each eigenvalue represents the total amount of variance explained by its principal component, standardized so that the mean eigenvalue is 1 and the sum of eigenvalues is equal to the total number of principal components created (which in turn is equal to the total number of input variables). Thus, an eigenvalue >1 implies that the principal component explains more of the total variance than the typical input variable, or is more useful in the analysis than the mean input variable. Conversely, an eigenvalue <1 implies that its principal component explains less of the total variance than the typical input variable. A researcher can determine the proportion of the total variation across all input variables that a given principal component explains by dividing its eigenvalue by the total number of input variables. The example shown in Table 1 provides a hypothetical example of eigenvalues from a PCA. The first principal component has an eigenvalue of 2, so it explains, or summarizes, 25% of the total variation represented by the full set of input variables. The first three principal components cumulatively summarize 71% of the total variance.

There are various possible approaches for determining the number of principal components to retain. One approach is to retain any principal component with an eigenvalue >1 and drop those with eigenvalues <1. Alternatively, the researcher could examine the overall pattern to look for a large drop-off in eigenvalues across all principal components. In the example shown in Table 1, since the first three eigenvalues are 2.0, 1.9, and 1.8, and the fourth is 1.1, then only the first three principal components would be retained. A figure called a "scree plot" can be used to help identify this drop-off, and this approach is sometimes called a scree test or the broken stick method. Figure 2 shows a screen plot from a different hypothetical example. In this case, the researchers might retain only a single principal component.

Table 1. Examples of eigenvalues from principal components analysis

Principal component	Eigenvalue	Percentage of variance explained	Cumulative percentage of variance explained
1	2.0	25.0	25.0
2	1.9	23.8	48.8
3	1.8	22.5	71.3
4	1.1	13.8	85.0
5	0.7	8.8	93.8
6	0.3	3.8	97.5
7	0.1	1.3	98.8
8	0.1	1.3	100.0

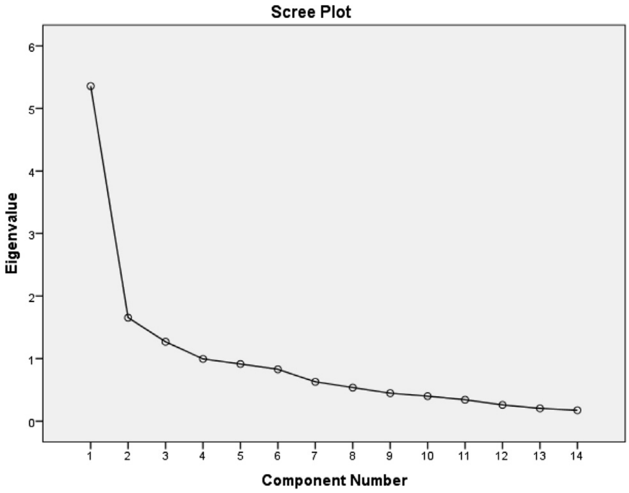

Figure 2. Hypothetical scree plot.

Finally, researchers sometimes make the decision as to which principal components to retain based on whether the principal components under consideration can be easily interpreted. Nettleton and colleagues[4] used a combination of a scree plot and interpretability considerations (ie, their ability to interpret the retained principal components) in making their decision to retain four principal components to summarize the original 47 input variables. In a different study, Panagiotakos and colleagues[7] also summarized dietary patterns using PCA. The authors chose to retain six principal components summarizing information from 22 input variables. In some cases, there may be a tradeoff between the number of input variables in the original analysis and the ease of interpreting the resulting principal components—it may be more challenging to interpret the principal components when there are a larger number of input variables. As an example, Bountziouka and colleagues[8] explored dietary patterns using two approaches—a large number of food items and a more limited number of food groups. The food groups, which represent aggregated food items, explained more variation in dietary intake than the food items and led to easier interpretations. Thus, the reduced number of input variables was advantageous and the resulting patterns were easier to explain.

A second key aspect of PCA output is the eigenvector, or set of "factor loadings," associated with each principal component. The final eigenvector is often achieved through the use of rotation methods, either orthogonal or oblique. When orthogonal rotation methods are applied, this ensures that the resulting eigenvectors are orthogonal—or uncorrelated—with one another. Each eigenvector includes a separate factor loading for each input variable, and the factor loading represents the weight or importance of the input variable on that principal component. The larger the factor loading, the more highly correlated that input variable is with the principal component or, from a different perspective, the more strongly the principal component will reflect the influence of that input variable.

The factor loadings in a given eigenvector can be used to interpret the meaning of the principal component. In particular, the researcher can examine input variables that have large positive factor loadings and/or large negative factor loadings to look for common themes that are not present among input variables with factor loadings close to zero. The issue is whether the input variables with large factor loadings seem to represent some underlying construct. The interpretation researchers give to a particular principal component is typically reflected in the name or label it is given. For example, Nettleton and colleagues[4] examined the eigenvector associated with the first principal component and found that input variables representing food groups high in fats and oils, processed meats, fried potatoes, and desserts had large positive factor loadings, while input variables representing other food groups had factor loadings that were either close to zero or negative. They labeled this principal component the "fats and processed meats" dietary pattern.

Table 2 provides a hypothetical example of a table showing eigenvectors. In this example, the eigenvectors for three retained principal components are shown, with factor loadings for each of the input variables listed under the eigenvector and the factor loadings that are relatively large in magnitude ($>|0.40|$) shown in bold. In the case of the first principal component, the input variables with the largest factor loadings are fruits, vegetables, and low-fat dairy products, and this principal component was labeled the "fruits and vegetables" dietary pattern. The other two principal components were labeled as the "meats" and "sweets" dietary patterns.

In addition to simply knowing the factor loadings for each retained principal component, researchers typically wish to create actual variables representing the principal components that can be used in subsequent analysis. They may want to describe values of the underlying construct among a given population, such as the extent to which different groups of adults have diets low or high in the principal component in the example from Nettleton and colleagues[4] and described here; that is, oils, fats, and processed meats. Alternatively, the principal component may be a key mediating or moderating variable in an analysis.[9,10] For example, one might do this by examining the role of dietary patterns in mediating the relationships between various individual characteristics and cardiovascular disease. Finally, the principal components might be used as control variables to adjust for potential confounders in a regression model. In the case of the work done by Nettleton and colleagues, this would mean that a construct like an individual's dietary patterns could be represented in a regression model by just four covariates rather than the original 47 input variables.

Two approaches for using factor loadings to create principal component variables for subsequent analysis are common. The first approach is to define the principal component as a factor score, or weighted average of all the input variables included in the analysis. Specifically, the factor score would be calculated by multiplying each input variable by its associated factor loading in the eigenvector for that principal component, and then summing all of these multiplied values. In a principal component calculated as a factor score, all of the input variables contribute to its value, although the input variables with the highest factor loading contribute most strongly.

The second approach is to define the principal component as a factor-based score using a subset of the input variables. The researcher would identify the input variables that seem

Table 2. Examples of eigenvectors for retained principal components[a]

Input variables	PC[b] 1 (fruits and vegetables)[c]	PC 2 (meats)	PC 3 (sweets)
Fruits	**0.537**	0.015	−0.322
Vegetables	**0.611**	0.189	−0.077
Dairy, high fat	−0.103	0.183	**0.488**
Dairy, low fat	**0.401**	−0.129	−0.007
Meats	0.094	**0.727**	0.140
Grains, whole	0.233	0.144	−0.048
Grains, processed	−0.175	**0.443**	0.259
Sweets	0.131	0.182	**0.649**

[a] Numbers in table are the factor loadings for each principal component. Numbers in bold type indicate factor loadings >|0.40|.
[b] PC=principal component.
[c] Names of each PC were assigned based on variables with highest factor loadings.

to be most important for that principal component in the sense of having the factor loadings that are largest in absolute value, keeping those input variables in the calculation and dropping the others. Often, a threshold such as 0.40 (in absolute value) is used to determine whether the factor loading is "large enough" to keep the input variable in the calculation of the factor-based score. The score would then be calculated as a simple mean of the values of all these retained input variables. Input variables with factor loadings that are small in absolute value would be ignored, meaning that they would contribute nothing to that principal component.

Once a principal component has been created, regardless of which method was used, it can be used in subsequent analyses like any other variable. By conducting PCA, however, the researcher has summarized a much larger set of input variables and can conduct these subsequent analyses much more efficiently.

Performing Factor Analysis. FA—sometimes referred to as exploratory FA—is similar to PCA, both conceptually and in the details of its calculation, and the distinctions between the two methods are subtle. Like PCA, FA produces a small number of factors that summarize the information from a larger number of input variables. FA results in output that includes eigenvalues and eigenvectors, and the calculation of the summary variables based on FA is typically conducted in much the same way as in PCA. Under certain circumstances, the results of FA and PCA will be very similar.

As described in the previous section, PCA is sometimes viewed as being fully data driven, where a researcher simply wants to summarize the information contained in a set of input variables with a few principal components. FA is more theoretically driven, typically conducted when a researcher has a set of input variables and a hypothesis about an underlying factor structure that determines the values of these input variables. In other words, the researcher believes that the values of these observed variables are determined based on the values of an underlying set of constructs, or factors that are not directly observable. FA is used to develop proxy measures of these underlying constructs.

For example, Reedy and colleagues[5] used FA to measure dietary patterns, implying a hypothesis that a small number of underlying dietary patterns collectively influenced sample members' observed intakes of a large number of individual foods. This distinction between the purpose of FA vs PCA is somewhat philosophical, as evidenced by the observation that Nettleton and colleagues[4] also summarized a large set of food intake variables with a small number of summary variables that they called dietary patterns, only they did so using PCA rather than FA.

Although the statistical methods that FA and PCA are based upon are similar, one methodological difference is that while the principal components generated by PCA take into account all of the variation across the full set of input variables, the factors generated by FA take into account only the common variation across those input variables. In other words, if a given input variable has some variation that is not related to variation in any of the other input variables, that variation is ignored by FA but taken into consideration by PCA. A second technical distinction is that eigenvalues can be negative in FA, but not in PCA. Otherwise, the interpretation of eigenvalues, eigenvectors, and factor loadings in FA is similar to their interpretation in PCA, and factor scores or factor-based scores can be calculated in the same way across the two methods.

Fernandez and colleagues[11] provide another example of a study that used FA. The authors conducted exploratory FA to develop a measure of dietary behaviors among a population of Latino adults with diabetes. They examined 13 input variables based on a questionnaire of dietary behaviors administered to a sample drawn from this population and came up with a four-factor structure that explained 47% of the common variance of these input variables. In interpreting the factor loadings from the eigenvectors for these four factors, the authors defined factor loadings of at least 0.32 in absolute value as being "acceptable" and factor loadings of at least 0.50 in absolute value as being "strong." Based on the input variables that met these criteria, they generated interpretations and labels for each of the four factors: healthy dietary changes, artificial sweeteners in drinks, number of meals per day, and fat consumption.

For a general description of FA, see Child.[12] For a discussion of the similarities and differences of FA and PCA, see Suhr.[13]

Reduced Rank Regression

Another closely related data reduction method is reduced rank regression, otherwise known as the maximum redundancy analysis. The statistics underlying reduced rank regression are very similar to those upon which PCA and FA are based. Each of these methods involves the calculation of principal components, or factors, using a similar approach involving eigenvalues and factor loadings based on eigenvectors. However, while the two methods described here determine the factors (principal components) by maximizing the explained variation of a set of predictor or input variables (eg, food groups), reduced rank regression derives the factors by accounting for as much of the variation as possible in a researcher-determined response variable or variables (eg, body mass index, nutrients, biomarkers). Reduced rank regression is particularly useful if a researcher wants to efficiently summarize a large set of input variables with the ultimate purpose of explaining or predicting some ultimate outcome of interest. In that case, the researcher could select as response variables a set of intermediate outcomes thought to be associated with the ultimate outcome of interest, based on prior research.

Performing Reduced Rank Regression. Hoffman and colleagues[14] are considered early adopters of using reduced rank regression to define dietary patterns, with the ultimate goal of identifying dietary patterns associated with diabetes. They used data on individuals' intakes of 49 food groups, obtained from study participants' responses on a food frequency questionnaire. The predictor variables included four nutrient intake measures that had been identified in prior research as being associated with the development of type 2 diabetes. In the study, reduced rank regression was used to define four factors representing dietary patterns that were then associated with diabetes onset.

As noted here, the key distinction between reduced rank regression and both PCA and FA involves the criteria for selecting and calculating the factors. Reduced rank regression defines factors to be linear combinations of input variables that best explain the total variance in the set of response variables. By contrast, PCA and FA define factors to be linear combinations of input variables that best explain the total variance in the set of input variables themselves. The researcher's goal in reduced rank regression is to represent to as great a degree as possible the predetermined response variables thought to be related to an ultimate outcome of interest.

This distinction leads to some differences between reduced rank regression and PCA/FA in the way that the output is generated and interpreted. Each of the three methods can be conducted using procedures available in the Statistical Analysis Software package (version 9.3, 2011, SAS Institute Inc). Unlike PCA/FA, however, reduced rank regression does not produce a number of factors equal to the original number of input variables. Instead, reduced rank regression produces a number of factors equal to the number of response or predictor variables. Typically, researchers use a fairly small number of response variables and retain all of the factors produced by the procedure. Because of the way the factors are created, those created by reduced rank regression typically explain substantially less of the total variance of the full set of input variables compared with factors created by PCA or FA, but substantially more of the total variance of the response variables (and, it is hoped, of the variance of the ultimate outcome of interest).

Fialkowski and collegues[15] provide another example of the use of reduced rank regression to summarize dietary patterns. They examined a sample of individuals from Pacific Northwest Tribal Nations to determine whether dietary patterns could be derived using reduced rank regression. In this study, the ultimate outcome of interest was weight status, and the researchers selected three response variables indicated by research to be associated with weight status—nutrient densities of total fat, total carbohydrate, and fiber. The input variables for this analysis were a set of 42 food components obtained through up to 4 days of dietary records.

With three response variables, reduced rank regression produced three factors, or principal components, and factor loadings for and proportion of variance explained by each of the factors are shown in Table 3. For simplicity, only the factor loadings above a certain threshold (|0.20|) are shown. Overall, the factors explained 82.3% of the variance in response variables. Of this total, a substantial proportion of the variance of the response variables were explained by two of the three factors, with the third factor explaining relatively little of the variance.

As in most analyses using a data reduction method, the reduced rank regression conducted by Fialkowski and colleagues[15] only sets the stage for further analysis—determining whether the factors created by reduced rank regression were related to an outcome of interest. In their case, the authors conducted inferential statistical analyses to determine whether the dietary patterns identified through reduced rank regression were associated with body mass index and whether dietary reference intakes were met.

The study conducted by Fialkowski and colleagues[15] also illustrates a common approach to the challenge of naming factors, or principal components, given that they may capture information from many different input variables. They based the factor name on the input variables with the largest positive factor loadings. For example, as the first factor heavily and positively loaded sweetened drinks, legumes, tomatoes, unsweetened cereals, and pasta, this factor was labeled "vegetarian and grains." This is a straightforward and useful approach for naming factors that arise from these methods, and Hoffman and colleagues[14] used a similar approach. A limitation is that the input variables identified in the factor label may end up being given disproportionate attention when readers interpret results based on that factor.

Cluster Analysis

Cluster analysis refers to a variety of algorithms and methods for creating a classification that summarizes data from a large set of related variables in an efficient but meaningful way.[3] In particular, cluster analysis is designed to group observations (or sample members) that share similar characteristics into meaningful groups, clusters, trees, or structures. These groups/clusters are mutually exclusive, and sample members within each group/cluster are similar to one another and different from sample members in other groups/clusters. While similar to the other data reduction methods in many ways, cluster analysis differs from PCA, FA, reduced rank regression in that it is a set of procedures used to classify or group similar or related sample members together into homogeneous groups, whereas the other procedures are used to group related variables together. A practical manifestation of this difference in that sample members typically fall into separate and distinct groups, or clusters, as a result of cluster analysis, but in the case of PCA, FA, and reduced rank regression every sample member is represented with a value on each factor or principal component.

Researchers may wish to use cluster analysis when it is important to clearly identify distinct groups of sample members within their population of interest. This may be the case if they wish to perform additional analysis using these groups. Thus, cluster analysis methods are particularly useful in the exploratory period of research. Once satisfactory clusters are formed, the clusters can be prepared for subsequent inferential statistics if a suitable a priori hypothesis has been established, such as by comparing mean outcomes among the newly defined groups of sample members.

As with the previous methods, nutrition researchers have used cluster analysis to examine individuals' dietary patterns,

Table 3. Factor loading matrix and percentage of variance explained using dietary records from adults (n=236) representing plausible reporters in the Communities Advancing the Studies of Tribal Nations Across the Lifespan (CoASTAL) study[a]

Food group	PC[b] 1 (vegetarian and grains)	PC 2 (healthy)	PC 3 (sweet drinks)	% of Variance explained
Fish (other than salmon)			−0.21	5.4
Alcohol			−0.66	56.8
Salmon			−0.28	9.8
Sweetened drinks	0.30	−0.41	0.23	44.0
Unsweetened drinks		0.21		7.1
Butter	−0.21			8.5
Fruit juices		−0.21		7.7
Fruit		0.35		23.5
Legumes, beans, soybeans	0.24	0.34		28.7
Tomato (including juice)	0.24			9.1
Nuts, seeds, peanut butter		0.29		18.1
Vegetables		0.29		14.1
Unsweetened cereals	0.29			15.4
Refined grains		−0.23		9.6
Pasta	0.29			11.6
Red meat	−0.38		−0.23	24.0
Processed meats	−0.21			6.8
Eggs	−0.33			13.8
% Variance explained	52.4	24.1	5.9	Sum=82.3

[a] Adapted from Fialkowski and colleagues,[15] with permission.

[b] PC=principal component. Factor loadings <|0.20| are not shown.

classifying individuals into groups based on the nature of their diets. For example, Reedy and colleagues[5] classified a group of adult and elderly residents of six states into separate clusters based on their dietary patterns with respect to a set of food groups thought to be related to colorectal cancer. These groups were then used to ascertain whether having a particular diet was related to an individual's risk of getting colorectal cancer. Wirfält and colleagues[16] also grouped study participants—a set of adult men and a set of women—into clusters as a way of summarizing their dietary patterns.

Cluster analysis has also been used for other purposes in nutrition research. Reicks and colleagues[17] used cluster analysis to classify parents of early adolescents into groups based on their practices and perceptions regarding the calcium intakes of their early adolescent children. The authors then correlated membership in these clusters with the calcium intakes and beverage consumption of both the parents and their children to better understand the association between practices/perceptions and dietary behavior.

Performing Cluster Analysis. Unlike the three methods described here, there is not a single, well-defined algorithm or procedure for conducting cluster analysis. Instead, there are many different possible approaches for doing this type of analysis. Moreover, different approaches may result in different solutions; that is, different arrangements of sample members into clusters. Therefore, in describing cluster analysis, researchers should be very clear about which algorithms were used in conducting the analysis and they should also provide some evidence as to the validity of their choices.

Once a researcher has selected the sample of cases to be clustered, he or she must select the set of variables upon which to base this clustering. These variables will be used as a basis for assessing the extent to which sample members are similar to or distinct from one another. Cluster analysis then involves:

- specifying a measure of similarity; that is, how the similarity between different sample members will be measured;
- selecting a cluster analysis algorithm or method to be used;
- determining the number of clusters to be formed; and
- validating the cluster solution.

In most applications of cluster analysis in nutrition research, Euclidean distance—the distance of a straight line between points on a plane—is used as the measure of similarity between two observations. This measure captures differences between observations in the values of each of the variables included in the analysis. Euclidean distance was used as the measure of similarity in each of the three studies employing cluster analysis noted here.

Table 4. Calcium intakes among early adolescent children (n=487) and their parents (n=498) separated by parents' calcium-rich foods practices/perceptions clusters[a]

Calcium intake from different food/beverage sources	Sweet-drink permissive parents	Dedicated-milk providers/ drinkers	Water regulars	*P* value (ANOVA[b])
	←———— *mean (mg/day)* ————→			
Early adolescents				
All food sources	873	1,273	1,001	<0.0001
Dairy foods	515	919	630	<0.0001
Milk	351	711	474	<0.0001
Parents				
All food sources	744	1055	446	<0.0001
Dairy foods	350	663	361	<0.0001
Milk	205	452	248	<0.0001

[a] Adapted from Reicks and colleagues,[17] with permission.

[b] ANOVA=analysis of variance.

The cluster analysis algorithm is the method for determining which cases should be classified into the same clusters on the basis of their similarity with one another. Two major categories of clustering algorithms are hierarchical or connectivity-based methods and iterative partitioning or centroid-based clustering (of which k-means clustering is one example). Aldenderfer and Blashfield[3] describe these methods along with several other less common approaches, including factor analytic methods, density search, clumping, and graph theoretic approaches.

With so many different possible approaches to clustering, there is no single correct way of conducting cluster analysis. The algorithms vary according to how they determine the level of similarity between different sample members, whether they have a single solution or may have different solutions depending on the starting point, and the size and shape of the clusters they are likely to create.

Hierarchical clustering includes in each cluster all of the data points that are "connected" with one another based on having a sufficiently high degree of similarity. How the degree of similarity is measured differs based on how connectivity is defined. In single linkage models, a given observation is included in a cluster if it is sufficiently similar to at least one observation already in the cluster. A single linkage hierarchical algorithm tends to produce large, elongated clusters.[3] With complete linkage, a given observation is included in the cluster only if it is sufficiently similar to all of the observations already in the cluster. This approach tends to produce smaller, more compacted clusters. In either case, the decision about how similar observations need to be in order to be included in the same cluster implicitly determines the number of clusters created by this approach.

Reicks and colleagues,[17] Reedy and colleagues,[5] and Wirfält and colleagues[16] each used k-means clustering, a type of centroid-based clustering. Under this approach, the researcher needs to specify both the number of clusters to be identified and a starting point for the iterative process of determining clusters. Under k-means cluster analysis, for example, the researcher sets an initial set of k centroids—the geometric center of an initial partitioning of the data. In the next step of the process, each observation of the dataset is assigned to the nearest centroid. This process forms a new set of clusters and a new set of centroids can be calculated in turn. The process is then repeated—observations are assigned to the new set of centroids. The iteration of this process is continually repeated until no observations change clusters after new centroids are defined. The clusters identified using the k-means algorithm tend to be spherical and to have similar numbers of observations. A drawback of k-means clustering is that if different starting points (or starting partitions of the data) are used, it is possible for the algorithm to lead to different final solutions of clusters with the same data.

As noted, the number of clusters to be identified must be specified by the researcher when using k-means clustering. There is no single universally accepted approach for selecting the number of clusters. One possibility is to use PCA or FA to guide the decision on the number of clusters. Another common approach is to conduct the cluster analysis repeatedly, each time specifying a different number of clusters. For example, Reicks and colleagues[17] produced separate cluster solutions based on creating 3, 4, 5, and 6 clusters. They ultimately settled on the analysis that produced three clusters, because they determined that it produced clusters that were reasonably distinct from one another, that differed from one another in the most meaningful way in terms of interpreting what being in each cluster signified, and that resulted in cluster sample sizes that were large enough for subsequent analysis. Wirfält and colleagues[16] produced cluster solutions for 2 through 20 clusters. They then formed a scree plot with the number of clusters plotted against the ratio of the degree of between-cluster heterogeneity to the degree of within-cluster heterogeneity. The value of this ratio generally increases as the number of clusters increases, but when the size of this increase is small then the increasing complexity of adding clusters likely outweighs the small gains in the ability to distinguish between clusters. Ultimately, the authors selected a 6-cluster solution.

Because different cluster analysis algorithms (or a single algorithm with different starting points) may produce different clustering solutions, it is important to assess the validity of the final set of clusters. Aldenderfer and Blashfield[3] highlight several validation approaches, indicating that the most useful approaches include replicating the cluster analysis on independent samples, conducting significance tests between clusters on the values of external variables (ie, variables not used to form the clusters), and conducting Monte Carlo analysis (a simulation approach in which the analysis is conducted repeatedly on artificially constructed samples). For example, Wirfält and colleagues[16] conducted cluster analysis on halves of the overall sample selected at random (as well as for several subgroups of their overall sample). Based on the observation that the cluster solutions produced by different samples were similar, they concluded that their clusters were valid. Reicks and colleagues[17] assessed the validity of clusters by examining whether the demographic characteristics and calcium intakes of the clusters were significantly (and substantively) different from one another. For example, Table 4 shows the results for calcium intakes among children and parents based on the three clusters of parents. There were statistically significant differences between clusters for each of the measures of calcium intakes shown in the table.

Ultimately, cluster analysis should also produce clusters of sample members that are different from one another in ways that can be easily interpreted and are relevant to the ultimate objectives of the research. Based on the values of the variables included in the analysis that produced three clusters, for example, Reicks and colleagues[17] defined one of the clusters of parents as “sweet-drink-permissive parents” because their reported attitudes consistently indicated that they allowed their children to drink fruit drinks and soda pop, although they did not regularly keep these beverages in their house. The other two clusters were “dedicated-milk providers/drinkers” and “water regulars.” These distinctions proved useful for the subsequent analysis of how the type of parent was related to the intake of calcium and beverage choices of the parent as well as the child.

CONCLUSIONS

This monograph provided an overview of data reduction methods as applied to nutrition and dietetic research. These methods allow nutrition researchers to summarize large quantities of data for three purposes: to succinctly describe potentially complex dietary phenomena, create analytic variables that can be used in subsequent analyses and hypothesis testing, and assist researchers in survey instrument development. Researchers face several challenges in conducting and interpreting data reduction methods. An initial challenge is in choosing among competing methods. This monograph presented four different classes of data reduction methods, and even within each class there are many different analytic decisions to be made. Another common challenge across methods involves the naming or labeling of principal components or factors, because they are created as linear combinations of a sometimes large number of input variables. One approach described in the article involved naming the factor on the basis of the input variables with the largest positive factor loadings. It is useful to keep in mind, however, that a simplified name for a factor does not always accurately capture the full range of input variables that contribute in important ways to the factor’s values. Despite these challenges, data reduction methods can be used in a variety of ways to better explain complex dietary patterns, behaviors, or perceptions. When conducting research that would benefit from these methods, investigators can be guided by the principles described in this article about the four major data reduction techniques: PCA, FA, reduced rank regression, and cluster analysis. These same principles can aid the reader’s comprehension and appreciation of studies using these methods to better evaluate the application of the results to practice.

References

1. Sheean PM, Bruemmer B, Gleason P, Harris J, Boushey C, Van Horn L. Publishing nutrition research: A review of multivariate techniques—Part 1. *J Am Diet Assoc*. 2011;111(1):103-110.
2. Harris JE, Sheean PM, Gleason PM, Bruemmer B, Boushey C. Publishing nutrition research: A review of multivariate techniques—Part 2: Analysis of variance. *J Acad Nutr Diet*. 2012;112(1):90-98.
3. Aldenderfer MS, Blashfield RK. *Cluster Analysis*. Thousand Oaks, CA: Sage Publications; 1984.
4. Nettleton JA, Steffen LM, Schulze MB, et al. Associations between markers of subclinical atherosclerosis and dietary patterns derived by principal components analysis and reduced rank regression in the Multi-Ethnic Study of Atherosclerosis (MESA). *Am J Clin Nutr*. 2007;85(6):1615-1625.
5. Reedy J, Wirfält E, Flood A, et al. Comparing 3 dietary pattern methods—cluster analysis, factor analysis, and index analysis—with colorectal cancer risk. *Am J Epidemiol*. 2009;171(4):479-487.
6. Glanz K, Steffen A. Development and reliability testing for measures of psychosocial constructs associated with adolescent girls’ calcium intake. *J Am Diet Assoc*. 2008;108(5):857-861.
7. Panagiotakos DB, Pitsavos C, Skoumas Y, Stefanadis C. The association between food patterns and the metabolic syndrome using principal components analysis: The ATTICA Study. *J Am Diet Assoc*. 2007;107(6):979-987.
8. Bountziouka V, Constantinidis TC, Polychronopoulos E, Panagiotakos DB. Short-term stability of dietary patterns defined a priori or a posterior. *Maturitas*. 2011;68(3):272-278.
9. Baranowski T. Advances in basic behavioral research will make the most important contributions to effective dietary change programs at this time. *J Am Diet Assoc*. 2006;106(6):808-811.
10. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173-1182.
11. Fernandez S, Olendzki B, Rosal MC. A dietary behaviors measure for use with low-income, Spanish-speaking Caribbean Latinos with type 2 diabetes: The Latino Dietary Behaviors Questionnaire. *J Am Diet Assoc*. 2011;111(4):589-599.
12. Child D. *The Essentials of Factor Analysis*. London and New York: Continuum International Publishing Group; 2006.
13. Suhr DD. Principle Component Analysis vs Exploratory Factor Analysis. Proceedings of the SAS Users Group International 30 (SUGI 30), Philadelphia, PA, 2005. http://www2.sas.com/proceedings/sugi30/203-30.pdf. Accessed February 25, 2014.
14. Hoffmann K, Schulze MB, Schienkiewitz A, Nothlings U, Boeing H. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol*. 2004;159(10):935-944.
15. Fialkowski MK, McCrory MA, Roberts SM, Tracy JK, Grattan LM, Boushey CJ. Dietary patterns are associated with dietary recommendations but have limited relationship to BMI in the Communities Advancing the Studies of Tribal Nations Across the Lifespan (CoASTAL) cohort. *Public Health Nutr*. 2012;15(10):1948-1958.
16. Wirfält E, Midthune D, Reedy J, et al. Associations between food patterns defined by cluster analysis and colorectal cancer incidence in the NIH-AARP diet and health study. *Eur J Clin Nutr*. 2009;63(6):707-717.
17. Reicks M, Degeneffe D, Ghosh K, et al. Parent calcium-rich-food practices/perceptions are associated with calcium intake among parents and their early adolescent children. *Public Health Nutr*. 2012;15(2):331-340.

AUTHOR INFORMATION

P. M. Gleason is a senior fellow, Mathematica Policy Research, Geneva, NY. C. J. Boushey is an associate research professor, University of Hawaii Cancer Center, Honolulu. J. E. Harris is professor and chair of the department of nutrition, West Chester University of Pennsylvania, West Chester. J. Zoellner is an associate professor, Department of Human Nutrition, Foods and Exercise, Virginia Tech University, Blacksburg.

Address correspondence to: Philip M. Gleason, PhD, Mathematica Policy Research, 331 Washington St, Geneva, NY 14456. E-mail: pgleason@mathematica-mpr.com

STATEMENT OF POTENTIAL CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

FUNDING/SUPPORT

There is no funding to disclose.

ACKNOWLEDGEMENTS

The authors are grateful for contributions from all members of the JAND STATS Committee.