Biostatistics Series Module 10: Brief Overview of Multivariate Methods

Avijit Hazra, Nithya Gogtay

Abstract

Multivariate analysis refers to statistical techniques that simultaneously look at three or more variables in relation to the subjects under investigation with the aim of identifying or clarifying the relationships between them. These techniques have been broadly classified as dependence techniques, which explore the relationship between one or more dependent variables and their independent predictors, and interdependence techniques, that make no such distinction but treat all variables equally in a search for underlying relationships. Multiple linear regression models a situation where a single numerical dependent variable is to be predicted from multiple numerical independent variables. Logistic regression is used when the outcome variable is dichotomous in nature. The log-linear technique models count type of data and can be used to analyze cross-tabulations where more than two variables are included. Analysis of covariance is an extension of analysis of variance (ANOVA), in which an additional independent variable of interest, the covariate, is brought into the analysis. It tries to examine whether a difference persists after “controlling” for the effect of the covariate that can impact the numerical dependent variable of interest. Multivariate analysis of variance (MANOVA) is a multivariate extension of ANOVA used when multiple numerical dependent variables have to be incorporated in the analysis. Interdependence techniques are more commonly applied to psychometrics, social sciences and market research. Exploratory factor analysis and principal component analysis are related techniques that seek to extract from a larger number of metric variables, a smaller number of composite factors or components, which are linearly related to the original variables. Cluster analysis aims to identify, in a large number of cases, relatively homogeneous groups called clusters, without prior information about the groups. The calculation intensive nature of multivariate analysis has so far precluded most researchers from using these techniques routinely. The situation is now changing with wider availability, and increasing sophistication of statistical software and researchers should no longer shy away from exploring the applications of multivariate methods to real-life data sets.

Key Words: Cluster analysis, discriminant function analysis, factor analysis, logistic regression, multiple linear regression, multivariate analysis, principal components analysis

Introduction

Multivariate analysis refers to statistical techniques that simultaneously look at three or more variables in relation to the subject under investigation with the aim of identifying or clarifying the relationships between them. The real world is always multivariate. Anything happening is the result of many different inputs and influences. However, multivariate methods are calculation intensive and hence have not been applied to research problems with the frequency that they should have been. Fortunately, in recent years, increasing computing power and the availability and user-friendliness of statistical software are leading to increased interest in and use of multivariate techniques. The aim of this module is to demystify multivariate methods, many of which are the basis for statistical modeling, and take a closer look at some of these methods. However, it is to be borne in mind that merely familiarizing oneself with the meaning of a technique does not mean that one has gained an understanding of what the analytic procedure does, what are its limitations, how to interpret the output that is generated and what the results signify. Mastery of a technique can only come through actually working with it a number of times using real data sets.
**Classification of Multivariate Methods**

This in itself is a complex issue. Most often, multivariate methods are classified as dependence or interdependence techniques and selection of the appropriate technique hinges on an understanding of this distinction. Table 1 lists techniques on the basis of this categorization. If in the context of the research question of interest, we can identify dependent or response variables whose value is influenced by independent explanatory or predictor variables, then we are dealing with a dependency situation. On the other hand, if variables are interrelated without a clear distinction of dependent and interdependent variables, or without the need for such a distinction, then we are dealing with an interdependency situation. Thus in the latter case, there are no dependent and independent variable designations, all variables are treated equally in a search for underlying patterns of relationships.

**Multiple Linear Regression**

Multiple linear regression attempts to model the relationship between two or more metric explanatory variables and a single metric response variable by fitting a linear equation to observed data.

Thus if we have a question like “do age, height and weight explain the variation in fasting blood glucose level?” or can “fasting blood glucose level be predicted from age, height and weight?,” we may use multiple regression after capturing data on all four variables for a series of subjects. Actually, there are three uses for multiple linear regression analysis. First, it may be used to assess the strength of the influence that individual predictors have on the single dependent variable. Second, it helps us to understand how much the dependent variable will change when we vary the predictors. Finally, a good model may be used to predict trends and future values.

### Table 1: Classification of multivariate statistical techniques

| Dependence techniques | Interdependence techniques |
|-----------------------|---------------------------|
| **One dependent variable** | Variables are metric |
| Dependent variable metric, independent variables also metric | Factor analysis and principal components analysis |
| Multiple (linear) regression | Cluster analysis |
| Multiple classification analysis | Multidimensional scaling |
| Automatic interaction detection | |
| Dependent variable metric, independent variables nonmetric | Variables are nonmetric |
| Multiple regression with dummy variables | Nonmetric multidimensional scaling |
| Log-linear analysis | Correspondence analysis |
| Dependent variable nonmetric, independent variables metric | |
| Discriminant (function) analysis | |
| Logit and probit analysis | |
| Dependent variable nonmetric, independent variables also nonmetric | |
| Canonical analysis with dummy variables | |
| Dependent variable nonmetric, independent variables metric or nonmetric or any combination of the two | |
| Logistic regression (binary or polychotomous) | |
| **More than one dependent variable** | |
| Dependent variables metric, independent variables also metric | |
| CCA | |
| SEM | |
| Dependent variables metric, independent variables nonmetric | |
| MANOVA | |
| Dependent variables nonmetric | |
| Conjoint analysis | |

*The terms “metric” and “nonmetric” variables in this table refer to numerical and categorical variables respectively. In logit and probit analysis, the independent variable is metric only in the sense that a transformed proportion is used, in log-linear analysis, the independent variable is metric only when we consider that the counts in the cross-tabulation cells are used to calculate the logs, in MANOVA, factors may be considered nonmetric independent variables in that they organize the data into groups, SEM for latent variables actually represents a family of advanced modeling techniques. Since it may handle dependence and interdependence, metric and nonmetric variables, its placement here (reflecting most common situation of its application) is somewhat arbitrary. MANOVA: Multivariate analysis of variance, CCA: Canonical correlation analysis, SEM: Structural equation modeling*
Mathematically the model for multiple regression, given n predictors, is denoted as
\[ y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \]
Where \( \beta_i \) represent a constant and \( \epsilon \) represents a random error term, while \( \beta_1, \beta_2, \beta_3, \text{ etc.} \) denote the regression coefficients (called partial regression coefficients in the multiple regression situation) associated with each predictor. A partial regression coefficient denotes the amount by which Y changes, when the particular X value changes by one unit, given that all other X values remain constant. Calculation of the regression coefficients makes use of the method of least squares technique as in simple linear regression.

We can refine and simplify a multiple linear regression model by judging which predictors substantially influence the dependent variable, and then excluding those that have only trivial effect. Software packages implement this as a series of hypothesis test for the predictors. For each predictor variable \( X_i \), we may test the null hypothesis \( \beta_i = 0 \) against the alternative \( \beta_i \neq 0 \), and obtain series of p values. Predictors returning statistically significant \( p \) values may be retained in the model, and others discarded. The software also offers various diagnostic tools to judge model fit and adequacy.

Multiple linear regression analysis makes certain assumptions. First and foremost linear relationship is assumed between each independent variable and the dependent variable. These can be verified through inspection of scatter plots. The regression residuals are assumed to be normally distributed and homoscedastic, that is show homogeneity of variance. The absence of multicollinearity is assumed in the model, meaning that the independent variables are not too highly correlated with one another. Finally, when modeling a dependent variable by the multiple linear regression procedure, an important consideration is the model fit. Adding independent variables to the model will always increase the amount of explained variance in the dependent variable (typically expressed as \( R^2 \) in contrast to \( r^2 \) used in the context of simple linear regression). However, adding too many independent variables without any theoretical justification may result in an overfit model.

**Logistic Regression**

Logistic regression analysis, in the simplest variant called binary logistic regression, models the value of one dependent variable, that is binary in nature, from two or more independent predictor variables.

It can be used to model any situation when the relationship between multiple predictor variables and one dependent variable is to be explored, provided the latter is dichotomous or binary in nature. The predictor variables can be any mix of numerical, nominal and ordinal variables. The technique addresses the same questions that discriminant function analysis or multiple linear regression does but with no distributional assumptions on the predictors (the predictors do not have to be normally distributed, linearly related or have equal variance in each group). There are more complex forms of logistic regression, called polychotomous logistic regression (also called polytomous or multinomial logistic regression), that can handle dependent variables that are polychotomous, rather than binary, in nature. If the outcome categories are ordered, the analysis is referred to as ordinal logistic regression. If predictors are all continuous, normally distributed and show homogeneity of variance, one may use discriminant analysis instead of logistic regression. If predictors are all categorical, one may use logit analysis instead.

Many questions may be addressed through logistic regression:
- What is the relative importance of each predictor or, in other words, what is the strength of association between the outcome variable and a predictor?
- Can the outcome variable be correctly predicted given a set of predictors?
- Can the solution generalize to predicting new cases?
- How does each variable affect the outcome? Does a predictor make the solution better or worse or have no effect?
- Are there interactions among predictors, that is, whether the effect of one predictor differs according to the level of another?
- Does adding interactions among predictors (continuous or categorical) improve the model?
- How good is the model at classifying cases for which the outcome is known?
- Can prediction models be tested for relative fit to the data? (the “goodness of fit” estimation).

Logistic regression utilizes the logit function and one of its few assumptions is that of linearity in the logit – the regression equation should have a linear relationship with the logit form of the predicted variable (there is no assumption about the predictors being linearly related to each other). The problem with probabilities is that they are nonlinear. Thus moving from 0.10 to 0.20 doubles the probability, but going from 0.80 to 0.90 barely increases the probability. Odds, we know, express the ratio of the probability of occurrence of an event to the probability of nonoccurrence of the event. Logit or log odds is the natural logarithm (log to base e or 2.71828) of a proportion expressed as the odds. The logit scale is linear. The unique impact of each predictor can be expressed as an odds ratio that can be tested for statistical significance, against the null hypothesis that the ratio is 1 (meaning no influence of a predictor on the predicted variable). This is done by using a form of Chi-square test that gives a Wald statistic (also called
Wald coefficient) and by looking at the 95% confidence limits of the odds ratio. The odds ratio of an individual predictor in a logistic regression output is called the adjusted odds ratio since it is adjusted against the values of other predictors.

The regression equation, in logistic regression, with \( n \) predictors, takes the form:

\[
\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_nX_n
\]

Where, \( \log\left[ (p/(1-p) \right] \) denotes logit \((p) = p \) being the probability of occurrence of the event in question, \( X_1, X_2, \ldots, X_n \) are the values of the explanatory variables, \( \beta_0 \) is the constant in the equation, and the beta values are the regression coefficients.

Using the logistic transformation in this way overcomes problems that might arise if \( P \) was modeled directly as a linear function of the explanatory variables. The constant and the regression coefficients for the predictors are usually displayed in table form in most computer outputs. A Wald statistic is provided for each predictor in the regression model, with its corresponding significance value.

There are three strategies for executing a logistic regression analysis – direct, sequential and stepwise. In the direct or “enter” method, all the independent variables are forced into the regression equation. This method is generally used when there is no specific hypotheses regarding the relative importance of the predictors. In sequential logistic regression, the investigator decides the order in which the independent variables are entered into the model to get an idea of the relative importance of the predictors. The predictor that is entered first is given priority, and all subsequent independent variables are assessed to see if they add significantly to the predictive value of the first variable. Stepwise logistic regression is an exploratory method concerned with hypothesis generation. The inclusion or exclusion of independent variables in the regression equation is decided on statistical grounds as the model is run. With some statistical packages, it is possible to set the significance level that will decide when to exclude a variable from the model. In a stepwise approach, high correlations between independent variables may mean that a predictor may be discarded, even if it does significantly predict the dependent variable because predictors entered earlier have already accounted for the prediction. Therefore, it is important to choose independent variables that are likely to have a high impact on the predicted variable but have little or no correlations between themselves (avoiding multicollinearity). This is a concern in all regression techniques but is particularly important in stepwise logistic regression because of the manner in which the analysis is run. Therefore, it helps greatly to precede a logistic regression analysis by univariate analysis on individual predictors and correlation analysis for various predictors. Discarding predictors with minimal influence as indicated by univariate analysis, will also simplify the model without decreasing its predictive value. The regression analysis will then indicate the best set of predictors, from those included, and one can use these to predict the outcome for new cases.

The exponential beta value in the logistic regression output denotes the odds ratio of the dependent variable. We can find the probability of the dependent variable from this odds ratio. If the exponential beta value is greater than one, then the probability of higher category (i.e., usually the occurrence of the event) increases, and if the probability of exponential beta is less than one, then the probability of higher category decreases. For categorical variables, the exponential beta value is interpreted against the reference category, where the probability of the dependent variable will increase or decrease. For numerical predictors, it is interpreted as one unit increase in the independent variable, corresponding to the increase or decrease of units of the dependent variable.

In classical regression analysis, the \( R^2 \) (coefficient of determination) value is used to express the proportion of the variability in the model that can be explained by one or more predictors that are included in the regression equation. This measure of the effect size is not acceptable in logistic regression; instead parameters such as Cox and Snell’s \( R^2 \), Nagelkerke’s \( R^2 \), and McFadden’s –2 log-likelihood (−2LL) statistic, are considered more appropriate. The \( R^2 \) values are more readily understood; if multiplied by 100 they indicate the percentage of variability accounted for by the model. Again, a “most likely” fit means −2LL value is minimized.

The overall significance of the regression model can also be tested by a goodness-of-fit test such as the likelihood ratio Chi-square test or the Hosmer–Lemeshow Chi-square test for comparing the goodness-of-fit of the model (including the constant and any number of predictors) with that of the constant-only model. If the difference is nonsignificant, it indicates that there is no difference in predictive power between the model that includes all the predictors and the model that includes none of the predictors. Another way of assessing the validity of the model is to look at the proportion of cases that are correctly classified by the model on the dependent variable. This is presented in a classification table.

Finally, it is important to be aware of the limitations or restrictions of logistic regression analysis. The dependent variable must be binary in nature. It is essentially a large sample analysis, and a satisfactory model cannot be found with too many predictors. One convenient thumb rule of sample size for logistic regression analysis.
is that there should be at least ten times as many cases as predictors. If not, the least influential predictors should be dropped from the analysis. Outliers will also influence the results of the regression analysis, and, therefore, it is good practice to examine the distribution of the predictor variables and exclude outliers from the data set. For this purpose, the values of a particular predictor may be transformed into standardized z scores and values with a z score of ± 3.29 or more are to be regarded as outliers. Finally, even if an independent variable successfully predicts the outcome, this does not imply a causal relationship. With large samples, even trivial correlations can become significant, and it is to be always borne in mind that statistical significance does not necessarily imply clinical significance.

Discriminant Function Analysis

Discriminant function analysis or discriminant analysis refers to a set of statistical techniques for generating rules for classifying cases (subjects or objects) into groups defined a priori, on the basis of observed variable values for each case. The aim is to assess whether or not a set of variables distinguish or discriminate between two (or more) groups of cases.

In a two group situation, the most commonly used method of discriminant analysis is Fisher’s linear discriminant function, in which a linear function of the variables giving maximal separation between the groups is determined. This results in a classification rule or allocation rule that may be used to assign a new case to one of the two groups. The sample of observations from which the discriminant function is derived is often known as the training set. A classification matrix is often used in the discriminant analysis for summarizing the results obtained from the derived classification rule and obtained by cross-tabulating observed against predicted group membership. In addition to finding a classification rule or discriminant model, it is important to assess its performance (prediction accuracy), by finding the error rate, that is the proportion of cases incorrectly classified.

Discriminant analysis makes several assumptions. Logistic regression is the alternative technique that can be used in place of discriminant analysis when data do not meet these assumptions.

Some Other Dependence Techniques

In the analysis of variance (ANOVA) we see whether a metric variable mean is significantly different with respect to a factor which is often the treatment grouping. Analysis of covariance (ANCOVA) is an extension of ANOVA, in which an additional independent metric variable of interest, the covariate, that can influence the metric dependent variable is brought into the analysis. It tries to examine whether a difference persists after “controlling” for the effect of the covariate. For instance, ANCOVA may be used to analyze if different blood pressure lowering drugs show a statistically significant difference in the extent of lowering of systolic or diastolic blood pressure, after adjusting for baseline blood pressure of the subject as covariate.

Multivariate analysis of variance (MANOVA) is an extension of univariate ANOVA. In ANOVA, we examine the relationship of one numerical dependent variable with the independent variable that determines the grouping. However, ANOVA cannot compare groups when more than one numerical dependent variable has to be considered. To account for multiple dependent variables, MANOVA bundles them together into a weighted linear combination or composite variable. These linear combinations are known variously as canonical variates, roots, eigenvalues, vectors, or discriminant functions, although the terms canonical variates or roots are more commonly used. Once the dependent variables combine into a canonical variate, MANOVA explores whether or not the independent variable group differs from the newly created group. In this way, MANOVA essentially tests whether or not the independent grouping variable explains a significant amount of variance in the canonical variate.

Log-linear analysis models count type of data and can be used to analyze contingency tables where more than two variables are included. It can look at three or more categorical variables at the same time to determine associations between them and also show just where these associations lie. The technique is so named because the logarithm of the expected value of a count variable is modeled as a linear function of parameters – the latter represent associations between pairs of variables and higher order interactions between more than two variables.

Probit analysis technique is most commonly employed in toxicological experiments in which sets of animals are subjected to known levels of a toxin, and a model is required to relate the proportion surviving at particular doses, to the dose. In this type of analysis, a special kind of data transformation of proportions, the probit transformation, is modeled as a linear function of the dose, or more commonly, the logarithm of the dose. Estimates of the parameters in the model are found by maximum likelihood estimation.

Closely related to the probit function (and probit analysis) are the logit function and logit analysis. The logit function is defined as

$$\text{Logit} (p) = \log \left( \frac{p}{1- p} \right)$$

Analogously to the probit model, we may assume that such a transformed quantity is related linearly to a set of
predictors, resulting in the logit model, which is the basis in particular of logistic regression, the most prevalent form of regression analysis for categorical response data. Box 1 provides examples of the application of dependence multivariate methods in dermatology from published literature.

---

### Box 1: Examples of dependence multivariate analysis from literature

**Ataş H, Cemil BC, Kurmus GI, Gönül M. Assessment of systemic inflammation with neutrophil-lymphocyte ratio in lichen planus. Postepy Dermatol Alergol** 2016;33:188-92.

Lichen planus (LP) is a papulosquamous disorder of the skin and mucous membranes. Although the pathogenesis remains unclear, it is believed that LP may be an inflammatory disorder. Neutrophil-lymphocyte (N/L) ratio is considered a systemic inflammation marker that may be related to the severity of the disease. The authors measured total leukocyte count (TLC), neutrophil and lymphocyte counts, N/L ratio, erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) in 55 patients and 48 healthy controls. N/L ratio, ESR and CRP were found to be statistically higher in LP patients than in controls upon univariate analysis. To assess if factors other than disease severity affect N/L ratio, the authors did multiple linear regression analysis with N/L ratio as the dependent variable and the other factors found to be significantly higher in LP patients, plus body surface area (BSA), as predictors. BSA and ESR were found to be significant independent predictors in the multiple linear regression model and hence can be confounders in assessment of disease severity on the basis of N/L ratio.

**Kim SH, Choi SM, Kim BC, Choi KH, Nam TS, Kim JT, et al. Risk factors for aseptic meningitis in herpes zoster patients. Ann Dermatol** 2017;29:283-7.

The authors conducted a retrospective study, in a tertiary care center, to identify potential factors associated with aseptic meningitis development in herpes zoster (HZ). Among 578 eligible patients over two years, 24 developed meningitis. On univariate analysis, the distribution of skin rash and gender showed statistically significantly different between the two groups. Patients with craniofacial distribution of HZ accounted for 87.5% as against 54.3% in the uncomplicated HZ group. The aseptic meningitis group had more men (66.7%) than the uncomplicated HZ group (42.8%). Binary logistic regression analysis indicated a weakening of gender difference (P=0.050) as a predictor but craniofacial involvement had adjusted odds ratio of 5.67 for aseptic meningitis (P=0.006).

**Atalay S, van den Reek JM, van Vugt LJ, Otero ME, van de Kerkhof PC, den Broeder AA, et al. Tight controlled dose reduction of biologics in psoriasis patients with low disease activity: A randomized pragmatic non-inferiority trial. BMC Dermatol** 2017;17:6.

Several targeted biologic therapies are now available for psoriasis. For rational use, the lowest effective dose of biologics needs to be determined. The authors want to know whether dose tapering of biologics guided by Psoriasis Area and Severity Index (PASI) and Dermatology Quality of Life Index (DLQI) scores in psoriasis patients with controlled disease activity is non-inferior to usual care. To address this objective they have designed a multicenter, randomized, non-inferiority trial that will recruit 120 patients with stable low disease activity (PASI ≤5 and DLQI ≤5) for at least 6 months, receiving etanercept, adalimumab or ustekinumab. Subjects will be randomized 1:1 to the dose reduction group or usual care. In the former group, the treatment intervals will be prolonged stepwise, aiming for 33% and 50% dose reduction, respectively. Disease activity will be monitored every three months with PASI and DLQI and in case of flare, the treatment will be adjusted to the previous effective dose. PASI will be the primary outcome to be compared between groups at 12 months using analysis of covariance (ANCOVA) in which the baseline PASI will be included as covariate. The study is ongoing.

**Rubegni P, Sbano P, Burrioni M, Cevenini G, Bocchi C, Severi FM, et al. Melanocytic skin lesions and pregnancy: Digital dermoscopy analysis. Skin Res Technol** 2007;13:143-7.

Few studies have tried to clarify how pregnancy influences the morphology of pigmented skin lesions (PSL). The authors aimed to determine, by digital dermoscopy analysis (DDA), any dermoscopic changes of acquired melanocytic nevi during pregnancy and after 1 year from delivery. For this, 35 healthy pregnant women and 35 age-matched female controls were enrolled in the study. Nevi of pregnant women were analyzed by DDA between 5 and 8 weeks of pregnancy, then between 39 and 41 weeks and finally 12 months after delivery. Nevi of control women were analyzed by DDA in a month of the year matching the period of recruitment of pregnant women and 21 months later. Multivariate analysis of variance (MANOVA) for repeated measures revealed that dermoscopic variables SKIN-GREEN-AVERAGE, SKIN-BLUE-AVERAGE and CONTRAST changed during pregnancy but returned to non-significant values after a year from delivery. The variable ENTROPY showed significant differences between initial evaluation and 1 year after delivery. Finally, the variable VARIANCE OF BORDER GRADIENT showed a significant difference between the first and the last evaluations, in both pregnant and control subjects. The conclusions were that pregnancy leads to significant modifications in PSL, especially with regard to pigment network, globules and architectural order or disorder.

**Norton HL, Edwards M, Krithika S, Johnson M, Werren EA, Parra EJ. Quantitative assessment of skin, hair, and iris variation in a diverse sample of individuals and associated genetic variation. Am J Phys Anthropol** 2016;160:570-81.

The authors quantitatively characterized skin, hair, and iris pigmentation using the Melanin (M) Index (skin) and CIELab values (hair) in 1450 individuals who self-identified as African American, East Asian, European, Hispanic, or South Asian. They also quantified iris pigmentation in a subset of these individuals using CIELab values from high-resolution iris photographs. They then compared mean skin M index and hair and iris CIELab values among populations using ANOVA and MANOVA respectively. The results indicated that all five ethnic groups were significantly different for skin and hair color. Iris and hair pigmentation appeared to be continuous, rather than discrete traits.
Factor Analysis and Principal Components Analysis

Factor analysis refers to techniques that seek to reduce a large number of variables to a smaller number of composite factors or components, which are linearly related to the original variables. It is regarded as a data or variable reduction technique that attempts to partition a given set of variables into groups of maximally correlated variables. In the process, meaningful new constructs may be deciphered for the underlying latent variable that is being studied. The variables to which factor analysis is applied are metric in nature. Factor analysis is commonly used in psychometrics, social science studies, and market research.

The term exploratory factor analysis (EFA) is used if the researcher does not have any idea of how many dimensions of a latent variable that is being studied is represented by the manifest variables that have been measured. For instance, a researcher wants to study patient satisfaction with doctors (a latent variable since it is not easily measured) and to do so measures a number of items (e.g., how many times doctor visited in last 1 year, how many other patients referred to the doctor by each patient, how long patient is willing to wait at each clinic visit to see the doctor, and so on) using a suitably framed questionnaire. The researcher may then apply EFA to assess if the different manifest variables group together into a set of composite variables (factor) that may be logically interpreted as representing different dimensions of patient satisfaction.

On the other hand, confirmatory factor analysis (CFA) is used for verification when the researcher already has a priori idea of the different dimensions of a latent variable that may be represented in a set of measured variables. For instance, a researcher may have grouped items in a fatigue rating scale as those representing physical fatigue and mental fatigue, and now wants to verify if this grouping is acceptable, through CFA. It is noteworthy that EFA is used to identify the hypothetical constructs in a set of data, while CFA may be used to confirm the existence of these hypothetical constructs in a fresh set of data. However, CFA is less often used, is mathematically more demanding and has a strong similarity to techniques of structural equation modeling.

Principal components analysis (PCA) is a related technique that transforms the original variables into new variables that are uncorrelated and account for decreasing proportion of the variance in the data. The new variables are called principal components and are linear functions of the original variables. The first principal component, or first factor, is comprised of the best linear function of the original variables so as to maximize the amount of the total variance that can be explained. The second principal component is the best linear combination of variables for explaining the variance not accounted for by the first factor. There may be a third, fourth, fifth principal component and so on, each representing the best linear combination of variables not accounted for by the previous factors. The process continues until all the variance is accounted for, but in practice, is usually stopped after a small number of factors have been extracted. The eigenvalue is a measure of the extent of the variation explained by a particular principal component. A plot of eigenvalue versus component number, the scree plot (Figure 1), is used to decide on the number of principal components to retain.

EFA explores the correlation matrix between a set of observed variables to determine whether the correlations arise from the relationship of these observed or manifest variables to a small number of underlying latent variables, called common factors. The regression coefficients of the observed variables on the common factors are known in this context as factor loadings. After the initial estimation phase, an attempt is generally made to simplify the often difficult task of interpreting the derived factors using a process known as factor rotation. The aim is to produce a solution having what is known as simple structure, where each common factor affects only a small number of the observed variables.

EFA and PCA may appear to be synonymous techniques and are often implemented simultaneously by statistical software packages but are not identical. Although they may give identical information, PCA may be regarded as a more basic version of EFA and was developed earlier on (actually first proposed in 1901 by Karl Pearson). Factor analysis has technical differences with PCA. It is now generally recommended that factor analysis should
Cluster Analysis

Cluster analysis refers to a set of statistical techniques for informative classification of an initially

be used when some theoretical ideas about relationships between variables exist, whereas PCA should be used if the goal of the researcher is to explore patterns in the data.
unclassified set of cases (subjects or objects) using observed variable values for each case. The aim is to identify relatively homogeneous groups (clusters) without prior information about the groups or group membership for any of the cases. It is also called classification analysis or numerical taxonomy. Unlike discriminant function analysis, which works on the basis of decided groups, cluster analysis tries to identify unknown groups.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the cluster profiles and finally, assessing the validity of clustering. The variables on which the cluster analysis is to be done have to be selected carefully on the basis of a hypothesis, past experience and judgment of the researcher. An appropriate measure of distance or similarity between the clusters needs to be selected – the most commonly used one is the Euclidean distance or its square.

The clustering procedure in cluster analysis may be hierarchical, nonhierarchical, or a two-step one. Hierarchical clustering methods do not require preset knowledge of the number of groups and can be agglomerative or divisive in nature. In either case, the results are best described using a diagram with tree-like structure, called tree diagram or dendrogram [Figure 2]. The objects are represented as nodes in the dendrogram, and the branches illustrate how they relate as subgroups or clusters. The length of the branch indicates the distance between the subgroups when they are joined. The nonhierarchical methods in cluster analysis are frequently referred to as $K$ means clustering. The choice of clustering procedure and the choice of distance measure are interrelated. The relative sizes of clusters in cluster analysis should be meaningful.

Cluster analysis has been criticized because a software package will always produce some clustering whether or not such clusters exist in reality. There are many options one may select when doing a cluster analysis using a statistical package. Thus, a researcher may mine the data trying different methods of linking groups until the researcher “discovers” the structure that he or she originally believed was contained in the data. Therefore, the output has to be interpreted with great caution.

Box 2 provides examples of the application of interdependence multivariate methods in dermatology from published literature.

---

**Sample size in Multivariate Analysis**

We will bring this overview on multivariate methods to a close with the vexing question of what is the appropriate sample size for multivariate analysis. Unfortunately, there is no ready answer. It is important to remember that most multivariate analyses are essentially large sample methods and applying these techniques to small datasets may yield unreliable results. There are no easy procedures or algorithms to calculate sample sizes a priori for majority of these techniques.

For regression analyses, a general rule is that 10 events of the outcome of interest are required for each variable in the model including the exposures of interest (Peduzzi P, Concordato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49: 1373-9). Even this rule may be inadequate in the presence of categorical covariates. Therefore, even though the sample size may seem enough, wherever possible, model quality checks should be done to ensure that they are reliable enough to base clinical judgments upon.

**Financial support and sponsorship**
Nil.

**Conflicts of interest**
There are no conflicts of interest.

**Further Reading**

1. Foster J, Barkus E, Yavorsky C. Understanding and using advanced statistics. London: Sage Publications; 2006.
2. Everitt BS. Medical statistics from A to Z: A guide for clinicians and medical students. Cambridge: Cambridge University Press; 2006.
3. Pereira-Maxwell F. A-Z of medical statistics: A companion for critical appraisal. London: Arnold (Hodder Headline Group); 1998.
4. An introduction to the general linear model. In: Hinton PR, Statistics explained. 2nd ed. New York: Routledge (Taylor & Francis); 2005. p. 315-342.