REVIEW
On the Application of Multivariate Statistical and Data Mining Analyses to Data in Neuroscience

Paul F. Smith

Department of Pharmacology and Toxicology, School of Biomedical Sciences, and Brain Health Research Centre, University of Otago, Dunedin, New Zealand; Brain Research New Zealand Centre of Research Excellence, and the Eisdell Moore Centre for Hearing and Balance Research, University of Auckland.

Research in neuroscience, whether at the level of genes, proteins, neurons or behavior, almost always involves the interaction of multiple variables, and yet many areas of neuroscience employ univariate statistical analyses almost exclusively. Since multiple variables often work together to produce a neuronal or behavioral effect, the use of univariate statistical procedures, analyzing one variable at a time, limits the ability of studies to reveal how interactions between different variables may determine a particular outcome. Multivariate statistical and data mining methods afford the opportunity to analyze many variables together, in order to understand how they function as a system, and how this system may change as a result of a disease or a drug. The aim of this review is to provide a succinct guide to methods such as linear discriminant analysis, support vector machines, principal component and factor analysis, cluster analysis, multiple linear regression, and random forest regression and classification, which have been used in circumscribed areas of neuroscience research, but which could be used more widely.

Key words: multivariate statistical analyses; data mining; neuroscience; SPSS 24; R

Experimental phenomena in neuroscience usually involve the complex interaction of multiple variables. Nonetheless, historically, statistical analysis has been dominated by the comparison of one variable at a time between treatment groups. In many areas of neuroscience, univariate statistical analyses have been used almost exclusively. This approach may inflate the type 1 error rate due to large numbers of univariate statistical analyses and can neglect the fact that changes may occur in the interactions between variables that cannot be detected in individual variables (Manly, 2005; Stevens, 2009; Liu et al., 2010). Consequently, in areas of neuroscience involving genomics, proteomics and medical diagnostics, multivariate statistical analyses and data mining approaches have been employed in order to understand complex interactions within systems of variables (e.g., Pang et al., 2006; Krafczyk et al., 2006; Dziuda, 2010; Ryan et al., 2011; Brandt et al., 2012; Smith, 2012; Zheng et al., 2012; Smith et al., 2013b; Liu et al., 2017). Despite these applications in neuroscience, most undergraduates receive little exposure to multivariate statistical and data mining methods, unless they do so in a bioinformatics course in a department such as biochemistry.

First, it is necessary to appreciate the broad scope of multivariate statistical analyses (MVAs) and related data mining procedures. Many MVAs can be broadly divided into those that are concerned with associating or predicting qualitative or categorical variables (e.g., linear discriminant analysis, support vector machines, random forest classification, correspondence analysis), and those that are concerned with associating or predicting quantitative variables (e.g., principal component analysis, factor analysis, cluster analysis, multiple linear regression, random forest regression, canonical correlation analysis, multidimensional scaling, ordination, structural equation modelling, neural networks) (Krzanowski, 2005; Stevens, 2009; see Manly, 2005 for an easy-to-understand introductory review). Multiple linear regression (MLR) and other regression methods such as random forest regression, are sometimes not considered to be MVAs, because there is only one dependent variable to be predicted while there are multiple independent variables (e.g., Krzanowski, 2005; Manly, 2005; West and Aiken, 2005). However, such methods are included in this review because of their importance and because what they have in common is that they involve *multiple variables*. For this reason, some textbooks on MVAs include MLR (e.g., Stevens, 2009). Although linear discriminant analysis (LDA), in its simplest form, has only one categorical dependent variable and multiple independent variables, the concept of discriminant analyses can be expanded to include more than one dependent variable (e.g., partial least squares discriminant analysis (PLS-DA) or orthogonal projection to latent structures discriminant analysis (OPLS-DA); He et al., 2017). Similarly, regression methods such as multiple linear regression (MLR) may only involve one dependent variable and multiple independent variables; however, regression methods can be extended to include more than one dependent variable (e.g., canonical correlation analysis, multivariate multiple regression; Manly, 2005; Hartung and Knapp, 2005; Stevens, 2009).

Other MVAs are not focused on specific dependent variables at all but more the degree of association or co-variation amongst multiple variables. For example, cluster analyses (CAs) can be used to investigate the degree to which the different neurochemicals related to aging co-vary with one another, and the results are often shown using dendrograms. CAs have been used extensively in genomics and proteomics. Still other MVAs are more concerned with investigating the way that groups of variables with different weightings, may explain most of the variation in a matrix of variables. Examples of this are Principal Component Analysis (PCA) and Factor Analysis (FA) (see Table 1). A distinction is often made between *supervised* and *unsupervised* MVA and data mining methods. *Supervised* methods are those that are applied to a dependent variable(s), with the objective of determining its (their) relationship with the independent variables, whereas *unsupervised* methods do not require a dependent variable but search for patterns in the independent variables (Questier et al., 2005; Anzanello et al., 2014). *Supervised* methods include linear discriminant analysis, support vector machines, random forest classification and regression, multiple linear regression, canonical correlation analysis, structural equation modelling and neural networks. *Unsupervised* methods include principal component analysis, factor analysis, correspondence analysis and cluster analysis (Questier et al., 2005; Anzanello et al., 2014) (see Table 1).

Data mining procedures are related to MVA but some have arisen out of computer science rather than traditional statistics. Data mining methods include such procedures as random forest regression, random forest classification and support vector machines; however, increasingly, the division between MVA and data mining methods is unclear.

Supervised
Qualitative or Categorical Variables
Linear discriminant analysis Logistic Regression Partial Least Squares Discriminant Analysis Structural Equation Modelling Support Vector Machines (DM) Random Forest Classification (DM)
Quantitative Variables
Multiple Linear Regression Canonical Correlation Analysis Multivariate Multiple Regression Structural Equation Modelling Random Forest Regression (DM) Neural Networks (DM)
Unsupervised
Qualitative or Categorical Variables
Correspondence Analysis
Quantitative Variables
Principal Component Analysis Factor Analysis Cluster Analysis Multidimensional Scaling Ordination

Table 1. Different types of MVA and Dating Mining Methods categorized according to whether they involve a categorical or continuous (quantitative) dependent variable and whether they specify a dependent variable (i.e., Supervised) or not (i.e., Unsupervised). 'DM' denotes those methods that emerged out of dating mining research in computer science.

The aim of this review is to provide a succinct guide to, and an overview of, some MVAs and data mining methods that have been and can be applied to neuroscience data, both at the basic experimental level as well as the clinical levels. In the author's experience, undergraduate and postgraduate neuroscience students often find multivariate statistical analyses surprisingly interesting, because they can reveal differences that were not at all obvious before the analysis. They also often find the procedures less difficult than they anticipate when they realize that many of them can be performed using the same data format as for univariate statistical analyses, depending on the statistical program being used. Because the MVA and data mining fields are vast, this review will focus on linear discriminant analysis, support vector machines, principal component and factor analyses, cluster analysis, MLR and random forest regression and classification, and their potential applications to neuroscience. For other methods listed in Table 1 the reader is referred to specialized textbooks and papers (e.g., Marcoulides and Hershberger, 1997; Gurney, 1997; Latin et al., 2003; Pang et al., 2006; Tabachnick and Fidell, 2007; Blunch, 2008; Marsland, 2009; Kaplan, 2009; Hastie et al., 2009).

CLASSIFICATION METHODS

LINEAR DISCRIMINANT ANALYSIS (LDA)

Linear discriminant analysis (LDA) is a statistical method often used following a multivariate analysis of variance (MANOVA), in which the membership of two or more groups can be predicted from a linear combination of independent variables (Manly, 2005; Stevens 2009). MANOVAs are an extension of univariate analysis of variance (ANOVA) to the case in which there is more than one dependent variable. There are various test statistics, including Wilks's λ, Roy's largest root, Pillai's trace statistic (also known as the Pillai-Bartlett trace statistic) and Lawes-Hotelling trace statistic (Manly, 2005). All four MANOVA statistics appear to be similar in power for small to moderate sample sizes (Field, 2011). According to Seber (1984), simulation studies indicate that Pillai's trace statistic may be more robust against violations of the assumptions of multivariate normality and homogeneity of the covariance matrices (see below) than the other statistics. Field (2011) also draws this conclusion, provided that the sample sizes for the different variables are equal.

A linear discriminant function (LDF) has the general form:

$$Z = a_1X_1 + a_2X_2 +...a_pX_p$$

Where Z refers to the group, X, X_2,….X_p are independent variables, and a_1, a_2,...a_p are coefficients (Manly, 2005).

LDA is similar in aim, but different in approach to logistic regression, in which the dependent variable is binary (0/1) and consists of positive (a 'success') and negative responses (a 'failure') only (Manly, 2005). However, LDA assumes that the independent, explanatory variables, are normally distributed, whereas logistic regression does not

(Field, 2011; Kitbumrungrat, 2012; Liong and Foo, 2013; Pohar et al., 2014).

The statistical significance of the LDF can be assessed using statistics such as Wilk's λ. The success of the LDF in separating the groups can be evaluated using cross-validation (e.g., a leave-one-out or 'LOO' procedure), in which the linear equation is used to classify the data, one observation at a time, without knowledge of the actual group membership. It is possible to use a stepwise LDA. However, some authors (Manly, 2005; Field, 2011) suggest that stepwise methods can result in suppressor effects and an increase in type II error. LDA is readily available in programs such as SPSS, SAS and Minitab.

MULTIVARIATE NORMALITY

MANOVA and LDA make assumptions about the distribution of the data. The first is that, for formal tests of statistical significance to be valid, the data within groups should have a multivariate normal distribution (Manly, 2005). Unlike univariate statistical analyses such as ANOVA, MANOVA and LDA are more sensitive to the violation of the assumption of multivariate normality, which can be a problem especially for very small sample sizes, i.e., < 10 (Marcoulides and Hershberger, 1997; Manly, 2005; Tabachnick and Fidell, 2007; Stevens, 2009). Fortunately, there is a multivariate formulation of the central limit theorem and sample sizes of 10-20 per group appear to be sufficient to afford protection against the consequences of violating multivariate normality (Bock, 1975; Stevens 2009). Furthermore, according to several studies, deviations from multivariate normality appear to have only a small effect on the type I error rate (see Stevens 2009, for a review, p. 222). It is in fact difficult to test for multivariate normality, because most programs such as SPSS and Minitab do not offer such an assumption test (Stevens, 2009). Because univariate normality, i.e., the normality of the individual variables, is necessary but not sufficient for multivariate normality, it is possible for each individual variable to be normally distributed without the multivariate distribution being normally distributed. Stevens (2009) points out that because a multivariate normal distribution entails that all subsets of variables have normal distributions, one way to assess multivariate normality is to determine whether all pairs of variables are bivariate normal. Box's test for the homogeneity of the covariance matrices is sensitive to violation of multivariate normality; therefore, in order to obtain results from that test that are valid, whether the assumption of multivariate normality is fulfilled, is of some concern (Stevens, 2009). It should be noted that LDA may still discriminate between groups even if the assumption of multivariate normality does not hold. On the other hand, multivariate normality does not necessarily mean that LDA will effectively discriminate between the groups.

HOMOGENEITY OF THE COVARIANCE MATRICES

A second assumption is that the population covariance matrices are equal for all groups, usually tested using Box's M test (Marcoulides and Hershberger, 1997; Manly, 2005). If this assumption is violated, a quadratic discriminant analysis (QDA) can be used instead. In a review of several Monte Carlo studies, Stevens (2009) concluded that, provided that the sample sizes are equal, even moderate heterogeneity of the covariances does not substantially affect type I error. Unequal sample sizes, on the other hand, are potentially very problematic if the covariances are unequal.

While Box's M test is often used, its null hypothesis may be rejected only because the multivariate normality assumption is violated (see above) (Stevens, 2009). Therefore, it is important to determine whether this is the reason for a significant Box's M test. Box's M test is also very sensitive to departure from homogeneity of the covariances (Field, 2011). Both Stevens (2009) and Field (2011) suggest that even if the Box's M test is significant, the type I error rate will be only slightly affected provided that there are equal sample sizes, although the power may be somewhat reduced.

NUMBER OF VARIABLES

One of the common problems in many multivariate statistical analyses is the sample size for each variable, n, relative to the number of variables, p. While unequal sample sizes can be problematic, as described above, when p is greater than n, statistical analyses such as MANOVA and LDA can become invalid. Stevens (2009) and Field (2011) suggest that, unless the n is large, p should be ≤ 10. Monte Carlo studies have shown that if the sample size is not large compared to the number of variables, the standardized discriminant function coefficients and correlations obtained in LDA are unstable (Stevens, 2009). By 'large', Stevens (2009) suggests an ideal ratio of n (total sample size) to p (number of variables) of 20:1. He further cautions that a small n:p ratio (i.e., ≤ 5) can be problematic for stepwise LDA in particular, because the significance tests are used to determine which variables are included in the solution (Stevens, 2009).

EXAMPLE OF LDA IN NEUROSCIENCE

Although LDA has not been used extensively in basic neuroscience to predict categorical membership, an example of its application is the prediction of age on the basis of the concentration of specific neurochemicals in different parts of the brain (Liu et al., 2010; Liu et al., 2017). The L-arginine metabolic pathway is a biochemical pathway that is critical for neuronal function (see Figure 1) and involves the neurochemicals: agmatine, putrescine, spermidine, spermine, L-arginine, L-ornithine, L-citrulline, glutamate and γ–aminobutyric acid (GABA). Although Figure 1 presents specific connections between some of these neurochemical variables, the mechanisms through which they interact with one another are not completely understood and additional pathways, specifically feedback pathways, are likely (Liu et al., 2017). It is, therefore, of interest to determine which parts of this complex neurochemical pathway are important in predicting categorical variables such as dementia.

Liu et al. (2010) examined the concentrations of these neurochemicals in two areas of the rat hindbrain concerned with the control of movement and balance: the brainstem

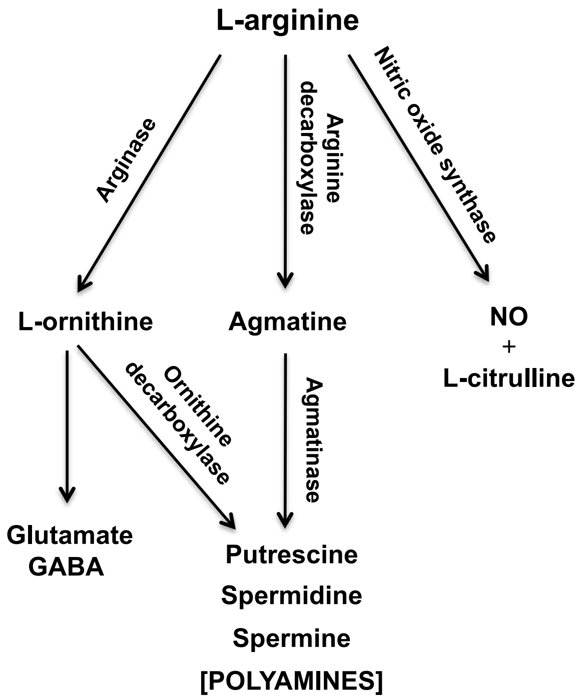

Figure 1. Schematic diagram of the L-arginine metabolic pathways. NO: nitric oxide; GABA: γ–aminobutyric acid. Modified from Liu et al. (2017) with permission.

vestibular nucleus complex (VNC) and the cerebellum (CE), in young (4-month-old; n = 16) and aged (24-month-old; n = 15) rats. Using LDA, a linear equation that could predict whether the animals were young or old, was obtained, based on putrescine, spermine, spermidine, L-citrulline, glutamate and GABA concentrations in the VNC. Cross-validation showed that this LDF could discriminate between young and old animals with 100% accuracy and it was statistically significant ($P \leq 0.0005$, Wilk's λ). The CE results were surprising. An LDF was discovered that could predict the animals' age based on only spermine and spermidine. Cross-validation showed that the LDF had 93% accuracy and was also statistically significant ($P \leq 0.0005$, Wilk's λ). The standardized canonical discriminant function coefficients are shown in Table 2. Both the size and the sign of the coefficients have predictive value.

This method should be applicable to many situations in neuroscience in which multiple variables interact to determine a categorical dependent variable, provided that the sample sizes are sufficient and the cross-validations demonstrate the predictive accuracy of the LDFs. Given that Box's M test of the equality of the covariance matrices assumes multivariate normality, one way to proceed is to determine whether all pairs of variables appear to be bivariate normal. If so, Box's M test can be used as a guide to whether the assumption of the equality of the covariance matrices is fulfilled. However, the cross-validation procedure can be used as the ultimate arbiter of the effectiveness of the LDF.

Field (2011) provides very clear, step-by-step, instructions on how to use SPSS to perform an LDA, including detailed information about the menus for the analysis options and the interpretation of the results that are generated (see pp. 615-622).

Standardised canonical discriminant function coefficients	
Putrescine	0.734
Spermidine	-2.417
Spermine	3.458
L-citrulline	-0.949
Glutamate	2.107
GABA	-1.800

Table 2. Standardised canonical discriminant function coefficients for the linear discriminant function for the VNC, which was 100% successful in predicting the age of the animals. From Liu et al. (2010).

SUPPORT VECTOR MACHINES

Support vector machines (SVMs) are an alternative method for classification, which employ 'support vectors', observations that constitute the spatial boundaries between different classes (Marsland, 2009; Hastie et al., 2009; Williams, 2011). These support vectors are then used to formulate a 'hyperplane' that defines the boundary between the classes (Hastie et al., 2009; Williams, 2011). SVMs can use a variety of kernel functions, for example radial, polynomial, hyperbolic tangent, spline, Bessel and Laplace functions (see Figure 2), in order to remap the data and create new variables that discriminate the different categories (Hastie et al., 2009; Williams, 2011). The data are usually split into training and test data sets (e.g., 70:30) and the difference between the model based on the support vectors in the training data set, and the test data set, is calculated as a measure of the model's success. As with LDA, classification error matrices can be used to evaluate the success of the classification, as well as receiver operating characteristic (ROC) curves, that

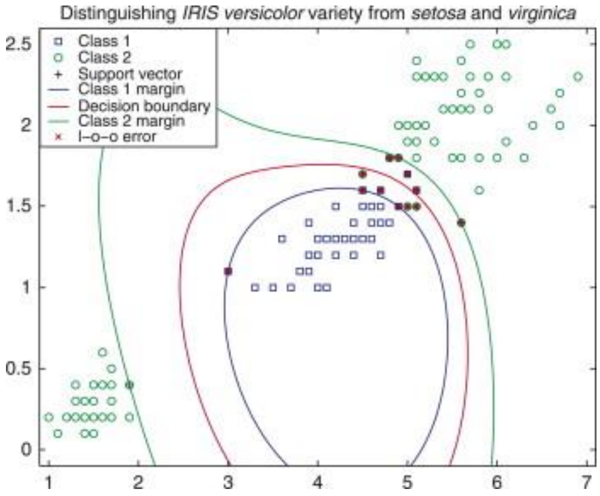

Figure 2. Example of an SVM classification employing a radial basis function to separate iris varieties based on petal width and length. From Wilson, M.D. Support vector machines. In Encyclopedia of Ecology, Elsevier, NY (2008) p. 3436. Reproduced with permission.

quantify the relationship between the true positive rate of classification (‘sensitivity’) and the false positive rate of classification (‘1 – the specificity’) (Hastie et al., 2009).

One of the major advantages of SVMs is that they do not make distributional assumptions like MANOVA and LDA, other than that the data are independent and identically distributed. Wilson (2008) suggests that for this reason, even small sample sizes can provide accurate estimates of prediction error when there is a large number of variables.

EXAMPLE OF SVMS IN NEUROSCIENCE

SVMs have been applied in the context of behavioural neuroscience, to predict whether animals have certain kinds of neurological dysfunction, based on their behavior. For example, rats with bilateral vestibular dysfunction (‘bilateral vestibular deafferentation’ or BVD) exhibit an unusual set of behavioral symptoms that include locomotor hyperactivity in an open field maze, abnormalities in behavior in an elevated T maze, changes in rearing behavior and reduced accuracy in responding in a spatial T maze (Zheng et al., 2012).

Attempts have been made to use multivariate statistical classification methods to determine whether rats can be classified as BVD or sham-operated (control rats), by combining the data from multiple behavioral symptoms, i.e., 12 different behavioral symptoms measured using an automated digital tracking system. LDA could discriminate between the BVD and sham animals with 100% accuracy. SVMs were also investigated, using Gaussian radial basis, polynomial, linear, hyperbolic tangent, Laplace, Bessel, ANOVA radial basis function (RBF) and spline kernels. The success of the predictive models was tested using a test data set, blind to the actual membership, as described above, and all of the kernel functions resulted in 100% accuracy in classifying the BVD animals, except for the Laplace (error rate: 50%), ANOVA RBF (error rate: 17%), and spline kernel functions (error rate: 33%) (Smith et al., 2013b). One of the objectives in investigating these classification methods is the potential application to the early diagnosis of neurological disorders (e.g., Brandt et al., 2012).

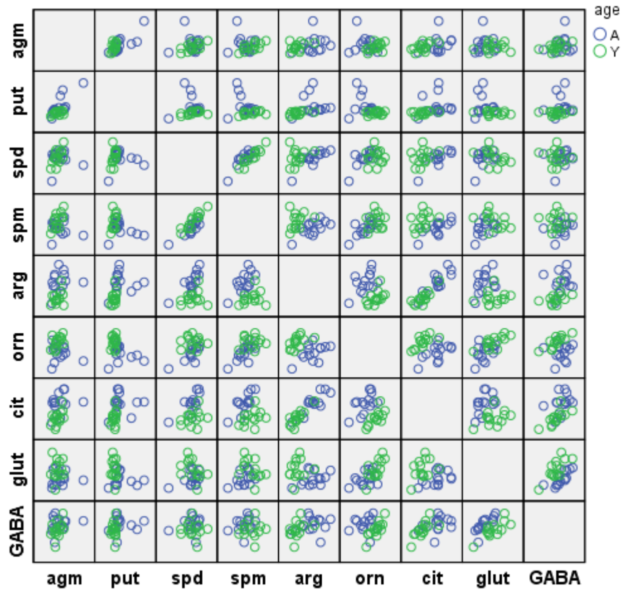

Figure 3. Scatterplot showing the co-variation of the 9 neurochemical variables in the vestibular nucleus with each other, as a function of age (‘A’ = aged, ‘Y’ = young’). Reproduced from Liu et al. (2010) with permission.

Although the use of SVMs may seem daunting, and they are often used with the statistical programming language, R, which many students find challenging to learn, there are simple alternatives to begin using SVMs. ‘R’ is a freely downloadable software program with many specialized packages (Crawley, 2007; Field et al., 2012; Davies, 2016). A particular data mining package, named ‘Rattle’, can be downloaded from the R websites, and while it requires the installation of R, it provides a simple menu interface for performing a variety of multivariate and dating mining analyses, including: decision trees, random forest classification and regression, cluster analyses and SVMs (Williams, 2009; Williams, 2011). Rattle has a very user-friendly graphics user interface (GUI) and data can be imported from Excel as csv files.

QUANTITATIVE METHODS

PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS

Like LDA, Principal Component Analysis (PCA) and Factor Analysis (FA) try to explain variation in the data using linear combinations of multiple variables. However, they look for underlying latent components or factors, representing combinations of variables, without predicting either a categorical or a continuous variable (‘*unsupervised* MVA’). The aim is rather to find a linear combination of variables that explains most of the variation in the data, in the process reducing the number of separate variables in the data (‘reducing dimensionality’) (Kline, 2002; Lattin et al., 2003; Jolliffe, 2004). These components or factors are expressed as ‘eigenvalues’, which in PCA and FA are represented as linear combinations of the original variables, each with a coefficient or ‘eigenvector’ that indicates the ‘direction’ of that particular variable for each component. The different PCs are uncorrelated (Kline, 2002; Lattin et al., 2003; Jolliffe, 2004).

EXAMPLE OF PCA IN NEUROSCIENCE

Data such as those described in the L-arginine data set used as an example previously, can be displayed as a data matrix, showing the co-variation of every variable with every other variable (Fig. 3). In the context of the L-arginine experiment, the general form of these components would be:

PC1 = 0.40 agmatine + 0.42 putrescine + 0.17 spermidine – 0.02 spermine + 0.49 arginine – 0.16 L-ornithine + 0.51 L-citrulline – 0.12 glutamate + 0.31 GABA

PC2 = -0.03 agmatine - 0.23 putrescine + 0.46 spermidine + 0.53 spermine + 0.08 arginine – 0.53 L-ornithine + 0.15

L-citrulline – 0.35 glutamate + 0.17 GABA etc.

For each PC the numbers are the coefficients for the different neurochemical variables in the linear equation. The number of principal components (PCs), which can be large, is usually displayed in decreasing order of importance in explaining the variability in the data matrix. This is often shown graphically in a 'Scree plot' (Fig. 4).

A major decision that has to be made in PCA is whether to use the covariance matrix or the correlation matrix for analysis. If the correlation matrix is used, then the data have to be standardized, i.e., each value subtracted from the mean for that variable and divided by the standard deviation (i.e., 'z scores'). This is done so that extreme differences in variance, e.g., due to different measurement scales, do not disproportionately affect the analysis.

PCA is an exploratory method that does not make many assumptions (the covariance or correlation matrices can be used but often the latter are preferable). For FA, there is a formal statistical model, and assumptions of multivariate normality etc. become important. For FA, the correlation matrix must be used.

The interpretation of the meaning of the components or factors relies on the size of the eigenvalue, i.e., how much variation in the data that it explains, and the contrasts between the eigenvectors for the variables relating to that eigenvalue. There is no clear answer to the question of how many components should be used; however, it is ideal to have a small number of PCs that explain most of the variation in the data (Manly, 2005). Loading plots, which represent the variance or magnitude of the variables within a PC, are often used to compare the different variables for the first two or three PCs (Fig. 5).

Because the interpretation of the PCs relies on the loadings, sometimes 'rotations' are used to maximize the contrasts between them while maintaining the relationship between the variables in the PCs. Examples include 'varimax' and 'quartimax' rotations (Kline, 2002; Lattin et al., 2003; Jolliffe, 2004; Manly, 2005).

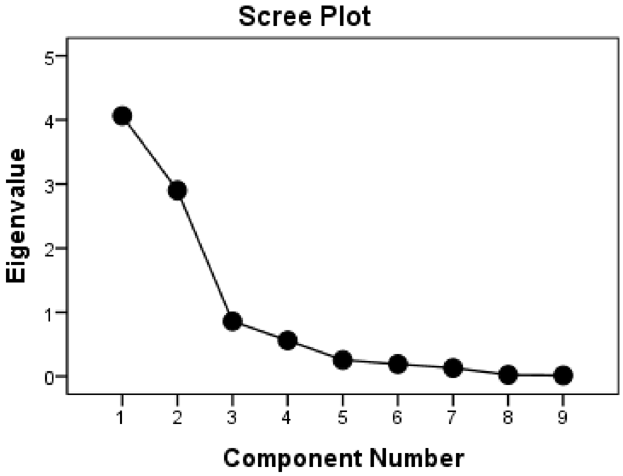

Figure 4. Scree plot for the L-arginine data set showing the size of the eigenvalues for the first 9 PCs. In this case, the first two PCs explain about 77% of the variation in the data. From Liu et al. (2010).

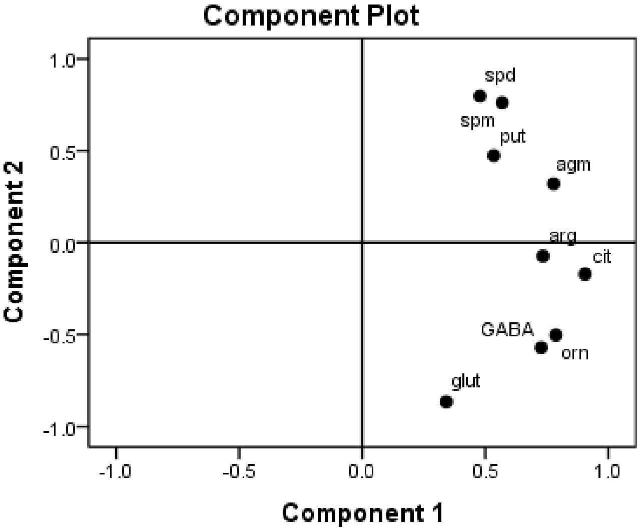

Figure 5. Loading plot showing the weighting or loadings for the first 2 PCs. Note the close co-variation of spermine and spermidine, which are chemically related. From Liu et al. (2010).

Whether methods such as PCA and FA are of any use in the analysis of multiple variables depends very much on whether considering the different variables together, as a component, makes sense in the context of the research question; and also, on what meaning can be attributed to the differences between the loadings. In the analysis of the L-arginine data set from Liu et al. (2010), these methods did prove useful. While the first 2 PCs accounted for approximately 77% of the variation in the data (Fig. 4), the first 3 PCs explained almost 87% (data not shown). Therefore, relatively few PCs were needed to account for the variation in the data. There appeared to be a clear relationship between the loadings for PC1, with agmatine, GABA, L-citrulline, L-ornithine and L-arginine, all exhibiting high, positive values (i.e., > 0.7; see Fig. 5). For PC1, the loading values for spermine and spermidine were very similar, as expected given their chemical relationship. PC1 contrasted sharply with PC2, where all of these variables, except for agmatine, showed negative values (see Fig. 5). It might be expected that these 5 variables would vary together, since agmatine is synthesized from L-arginine, and L-ornithine and GABA are also generated from it via the enzyme, arginase (Fig. 1). Similarly, L-citrulline is produced from L-arginine by the enzyme, nitric oxide synthase (NOS) (Fig. 1).

However, PCA is often useful in neuroscience research in which there are hundreds of variables, for example, metabolomics, where it is useful to determine whether there is a change in the overall pattern of metabolites in different brain regions (e.g., He et al., 2017). Figure 6 shows an example in which PCA was found to be very useful. It shows PC2 plotted against PC1 with loading scores where the PCs represent the combination of 88 metabolites from the auditory cortices of rats either exposed to noise trauma or exposed to a sham condition. It can be seen that the purple dots, representing the acoustic trauma group, are well separated from the red dots, representing the sham controls, and this demonstrates that exposure to noise trauma has caused a metabolite shift in this area of the brain (He et al., 2017). Due to the number of variables, it is advantageous in this case to consider the relationship between all of the variables in single components, and whether these variables, as a 'system', change in the experimental group relative to the control group (He et al., 2017). This statistical procedure has been used extensively in the context of 'metabolomics', the analysis of metabolites, and often the principal components are then used in an orthogonal partial least squares discriminant analysis (OPLS-DA), in which the predictor variables are actually PCs and therefore are independent of one another ('orthogonal') (He et al., 2017).

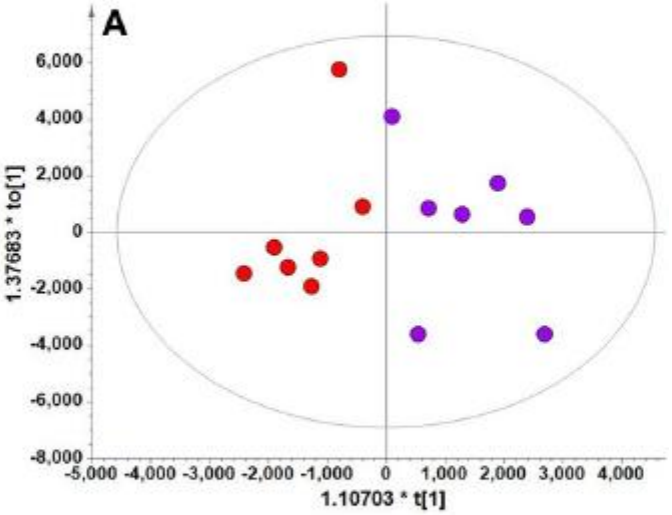

Figure 6. Loading plot showing PC2 (to[1]) against PC1 (t[1], where the purple dots represent the acoustic trauma group and the red dots, the sham controls. Note the clear separation of the two groups. From He et al. (2017) with permission.

PCA and FA can be performed quite easily in SPSS 24 under the 'Dimension Reduction' menu item in the 'Analyze' menu. Once again, Field (2011) provides step-by-step, instructions on how to use SPSS to perform PCA, including explicit interpretations of the menus and the results that are generated (see pp. 627-685).

CLUSTER ANALYSIS

Another multivariate statistical method that has not been used extensively in the context of neuroscience, is cluster analysis. Cluster analyses (CAs) are a type of non-parametric analysis that is used to explore the natural groupings of variables in a data set (Manly, 2005). Therefore, assumptions such as multivariate normality and equality of the variance-covariance matrices are not required (Marcoulides and Hershberger, 1997; Manly, 2005). Different measurements of the distance between the variables, such as Euclidean or Mahalanobis distance, are used to relate them to one another, and specific algorithms (e.g., Ward Minimal Variance Linkage) are used to determine the clusters (Marcoulides and Hershberger, 1997). The standardized data (i.e., z scores) are usually used in order to avoid bias introduced by differences in scales of measurement.

EXAMPLE OF CAs IN NEUROSCIENCE

As an example of the application of this method, Figure 7 shows CAs for the data from Liu et al. (2017). Agglomerative CAs, in which each variable is initially considered its own cluster, were used on the correlation coefficient distance. Some algorithms, such as single linkage, are prone to produce long strings of clusters ('chaining') (Lattin et al., 2003). Comparisons of the different kinds of CAs for the VNC data suggested that the Complete linkage, McQuitty linkage, Average linkage and Ward linkage algorithms for determining clusters, all produced similar results. Ward's method, based on the objective of obtaining the smallest within-cluster sum of squares (the 'minimal variance principle'), was used (Lattin et al., 2003).

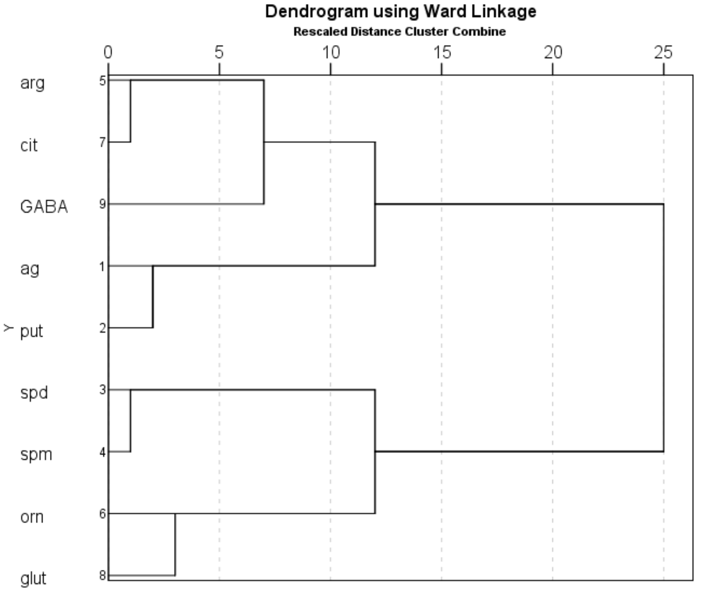

Figure 7. Dendrogram showing the relationship between the expression of the different neurochemical variables in the young and aged VNC. The CA was agglomerative and performed on the data expressed as z scores, using the squared Euclidean distance and the Ward minimal variance method. From Liu et al. (2017).

CAs are not covered in the SPSS book by Field (2011). However, they are relatively simple to do using the same data formatting as for PCA. Under the 'Analyze' menu, choose the 'Classify' menu item and then the 'Hierarchical Cluster' menu option. A menu will appear that requires you to enter the variables you wish to analyze, and then if you are interested in the relationships between the data values for those variables, you choose 'Variables' under 'Cluster'. Under 'Display', the 'Statistics' and 'Plots' options should be selected, and then after selecting 'Statistics' in the right-hand menu, you can choose the 'Agglomeration Schedule.' Under the 'Plots' menu, it is usual to select 'Dendrogram' (see Fig. 7), and then you need to select 'Method.' In the 'Method' menu, you need to select one type of CA. There are 7 options, each with their own strengths and weaknesses (Lattin et al., 2003; Manly, 2005); however, we have found that Ward's method works reliably (Liu et al., 2010; Liu et al., 2017). Under the 'Measure' menu item, you need to select a method of 'Distance' measurement, which could be 'Squared Euclidean Distance' (Lattin et al., 2003; Manly, 2005). Lastly, and this is very important, you need to select 'Z Scores' under the 'Transform Values' and 'Standardize' menus, if your data are not already standardized (see above). Now press 'Continue', and 'OK' in the main menu, and the results will be generated. Dendrograms show the co-variation of the values for the variables as a re-scaled distance value along the y axis, where small values for variables connected close to the x axis represent those that strongly co-vary. Interpreting the dendrogram is a matter of visually inspecting these co-variations (Fig. 7). CAs can also be performed easily in the R package, Rattle (Williams, 2009; Williams, 2011), although there are fewer options than in SPSS 24.

MULTIPLE LINEAR REGRESSION

Yet another statistical method that has been under-employed in neuroscience is multiple linear regression (MLR). Although not strictly a multivariate statistical method, since there is only one dependent variable at a time, MLR is a part of the general linear model (GLM) that is useful for determining whether one variable can be predicted from a combination of other variables. Simple linear regression can be expanded to include more than one predictor variable to become MLR.

MLR has the general form:

$$Y = \beta_0 + = \beta_1 X_1 + = \beta_2 X_2 + \ldots = \beta_p X_p + \varepsilon$$

Where Y = the quantitative dependent variable; $X_1, X_2, \ldots X_p$ are independent variables; $\beta_1, \beta_2, \ldots \beta_p$ are coefficients; β_0 is the intercept and ε is the error term (Brook and Arnold, 1985; Ryan, 2009; Stevens, 2009).

Canonical correlation analysis is an extension of MLR in which multiple Y variables are related to multiple X variables (Manly, 2005).

However, formal statistical tests for MLR, like those for simple linear regression, make assumptions regarding the distribution of the data, which cannot always be fulfilled. These assumptions are the same as those for other methods in the general linear model, such as ANOVA and analysis of covariance (ANCOVA): that the residuals are normally distributed, with homogeneity of variance, and that they are independent of one another (e.g., not autocorrelated) (Brook and Arnold, 1985; Rutherford, 2001; Vittinghoff et al., 2005; Doncaster and Davey, 2007; Gamst et al., 2008; Fig. 8). Furthermore, the predictor variables should be numerical, although indicator variables can be used in order to include nominal variables (e.g., binary coding to represent male and female). The violation of the assumption of normality can sometimes be redressed using data transformation, which may also correct heterogeneity of variance, but other issues such as autocorrelation, are not easily dealt with and methods such as time series regression may be required (Brook and Arnold, 1985; Vittinghoff et al., 2005; Ryan, 2009).

Unlike simple linear regression, MLR is more complicated in terms of avoiding potential artifacts. Because R^2 will increase as more independent variables are incorporated into the regression model, the adjusted R^2 must be used in order to compensate for the number of variables included. For k = 1 variables, the R^2 and adjusted R^2 are approximately equal.

There are various forms of MLR: forward regression, backward regression, stepwise regression and best subsets regression. In forward regression, predictors are added into the model one at a time (if alpha is set to 1.0, then all of them will be included, in ascending order of significance). In backward regression, predictors are taken out one at a time (if alpha is set to 0, all of them will be taken out, in descending order of significance). Backward regression tends to be preferred because it allows examination of the interaction between variables. In stepwise regression, the program stops at each step and checks whether the variables, either in the model or not, are the best combination for that step. The adjusted R^2 will change as different variables are included and an F test can be done at each step to determine whether it has made a significant difference. Best subsets regression, however, computes all possible MLR models from which the researcher must choose the best, based on the adjusted R^2 and various diagnostic information regarding the validity of the regression model. Two of the greatest problems in MLR are 'over-fitting' and 'multicollinearity' (Babayak, 2004). If the regression variables are highly inter-correlated, multicollinearity occurs. This inflates the variance of the least squares estimates and therefore the

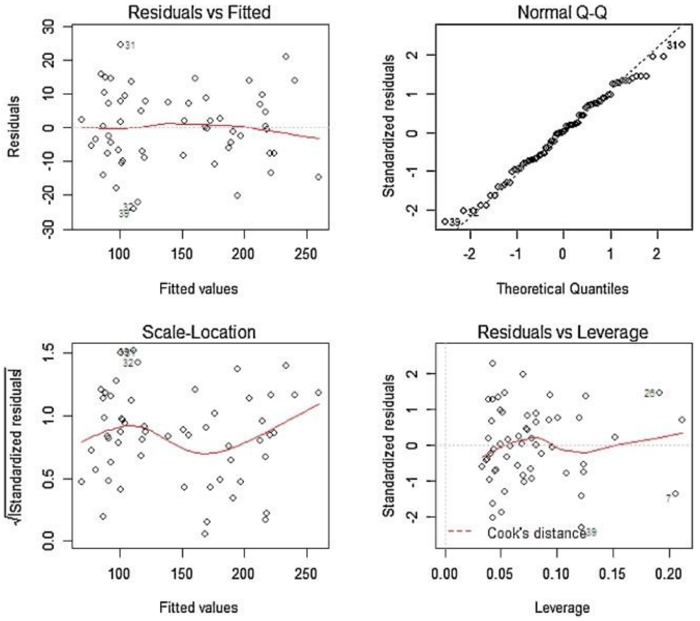

Figure 8. Diagnostic plots for L-citrulline following MLR showing the residuals versus fitted values, normal quantile-quantile (Q-Q), scale location and residuals versus leverage plots. From Smith et al. (2013a) with permission.

coefficients will be inaccurate, which can lead to the situation in which the F test for the regression is significant without any single t test for an individual variable being significant. In this case, one or more of the highly correlated variables should be removed from the regression model. One way of controlling for multicollinearity is using an index such as the Mallow's Cp index. The adjusted R^2 should be high but the Mallow's Cp index (= (the sum of squares for the error at the current step / mean square error for the full regression) - (n - 2p), where n = total number of observations and p = number of estimated coefficients), should be as small as possible. Ideally, it should be one more than the number of parameters in the current step. Other indices of multicollinearity include the variance inflation factor (VIF) and tolerance (1/VIF). Different software packages (e.g., SPSS and Minitab) offer different options.

Autocorrelation in the data can be tested using the Durban-Watson statistic (Fields, 2011). Like most other multivariate statistical procedures, MLR is prone to artifacts and researchers need to be cautious when using it (Babayak, 2004).

MLR is almost as easy to do in SPSS 24 as simple linear regression, in which there is only one explanatory variable. The main difference, in terms of using the GUI, is the number of independent variables entered for the regression model, and the method that must be selected (i.e., backward versus stepwise MLR). Field (2011), once again, provides an excellent step-by-step set of instructions for performing MLR (pp. 197-263). An example of MLR applied to the L-arginine data set, is provided below, following the random forest regression section, which is compared with MLR.

RANDOM FOREST REGRESSION

Modelling using regression trees has been used for decades; nonetheless, its use in neuroscience has been very limited. In regression tree modelling, a flow-like series of questions is formulated for each variable, known as 'recursive partitioning', thereby subdividing a sample into groups with maximal homogeneity by minimizing the within-group variance, with the objective of determining a numerical response variable (Vittinghoff et al., 2005; Hastie et al., 2009). The predictor variables can be continuous variables such as interval or ratio variables, or they can be ordinal or nominal variables. In contrast to MLR, which makes assumptions about the distribution of the data, regression trees make no distributional assumptions. The data are often divided into training and test data sets (e.g., 70:30) and the mean square error (MSE) between the model based on the training data, and the test data, is calculated as a measure of how well the model describes the data. Variables are chosen to divide the data using the reduction in the MSE that is achieved following a split (i.e., the information gained). Interactions between different predictor variables are automatically incorporated into the model and variable selection is not necessary because predictors which are irrelevant are excluded from the model. This means that complex, possibly non-linear interactions, between variables are easier to accommodate compared to linear regression modelling (Hastie et al., 2009). Using the computational power of modern computers, Breiman et al. (1984) extended the concept of regression trees by simultaneously generating hundreds of trees, which became known as *'random forests'*, based on a random selection of a subset of data in the training set. The various regression tree models are then averaged in order to predict the dependent variable with the smallest MSE possible (Marsland, 2009; Hastie et al., 2009; Williams, 2011).

Random forests (RFs) can also be used for classification purposes, in which case the solution is based on the number of 'votes' from different trees for a particular category (Williams, 2009; Williams, 2011). The effect of variable removal on the mean decrease in accuracy, the 'out of bag' (OOB) error, and the overall classification matrix error ('confusion matrix error'), are used to evaluate the success of the classification. The 'out of bag' (OOB) error is the error based on the observations that were excluded from the subset of the training data (the 'bag') used to generate the decision tree (Williams, 2011). Unlike LDA, random forest regression and classification make no distributional assumptions and therefore can be applied to situations in which the sample sizes are small relative to the number of variables (Hastie et al., 2009; Williams, 2011).

Random forest regression and classification, along with SVMs and CAs, can be carried out using specific packages in the statistics program R (Crawley, 2007; Field et al., 2012; Davies, 2016), and are easily done using the R package, 'Rattle', which has a user-friendly GUI and is entirely menu-driven (Williams, 2009; Willams, 2011).

EXAMPLES OF MLR AND RANDOM FOREST REGRESSION IN NEUROSCIENCE

Figure 9 shows a random forest regression for the prediction of spermine from the other variables. The proportion of variance explained was 0.94, which was very high. Figure 10 shows how the error in prediction decreased as the number of trees increased. As a comparison of the application of MLR and random forest regression (RFR) to the data from Liu et al. (2010) described above to illustrate LDA, both forms of regression were applied to the prediction of each of the nine neurochemicals from the other eight. Figure 11 shows the adjusted R^2 (MLR) and variance explained values (RFR), as well as the residual mean square error (RSE) values, for the MLRs and RFRs (respectively).

It was apparent that the adjusted R^2 values for the MLRs were generally higher than the variance explained values for the RFRs: 5/9 of them were ≥ 0.80 compared to 3/9 for the variance explained values (Figure 11). The RSE values were more similar but lower for the MLRs than the RFRs in all but 3 cases (Figure 11). Nonetheless, the general patterns for the adjusted R^2/variance explained values and the RSEs were similar for the MLRs and the RFRs. Although MLR appeared to be more predictive for this particular data set, both forms of regression could potentially be used to predict behavioral, neurophysiological and neurochemical variables in the context of neuroscience research (Smith et al., 2013a,b; Aitken et al., 2017).

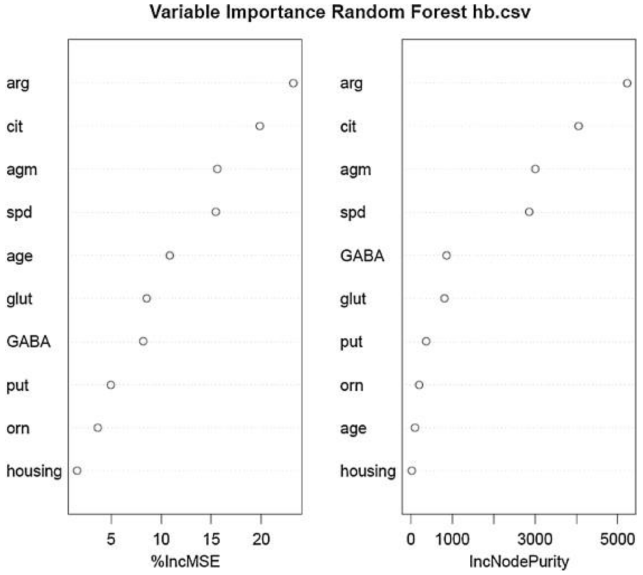

Figure 9. Variables in order of importance for the random forest regression for spermine, which had the highest proportion of variance explained (94%). The mean decrease in Gini coefficient is an indication of the extent to which each variable contributes to the homogeneity of the nodes and leaves in the random forest. From Smith et al. (2013a) with permission.

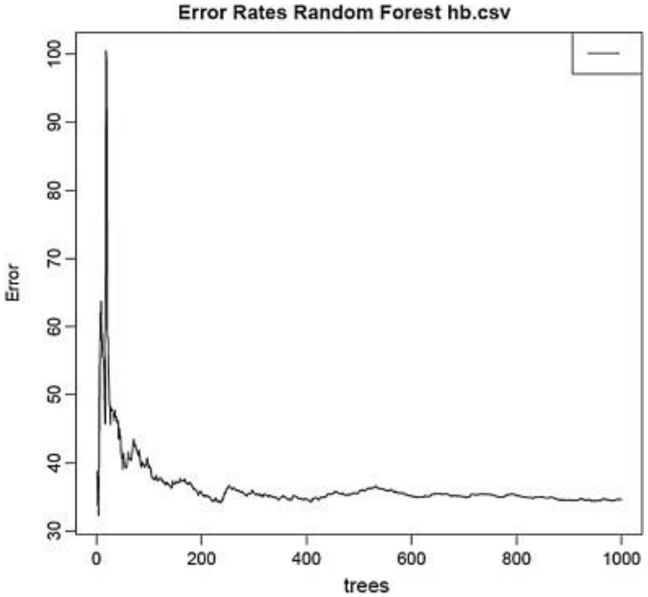

Figure 10. Decrease in error as a function of the number of trees for the random forest regression for spermine. From Smith et al. (2013a) with permission.

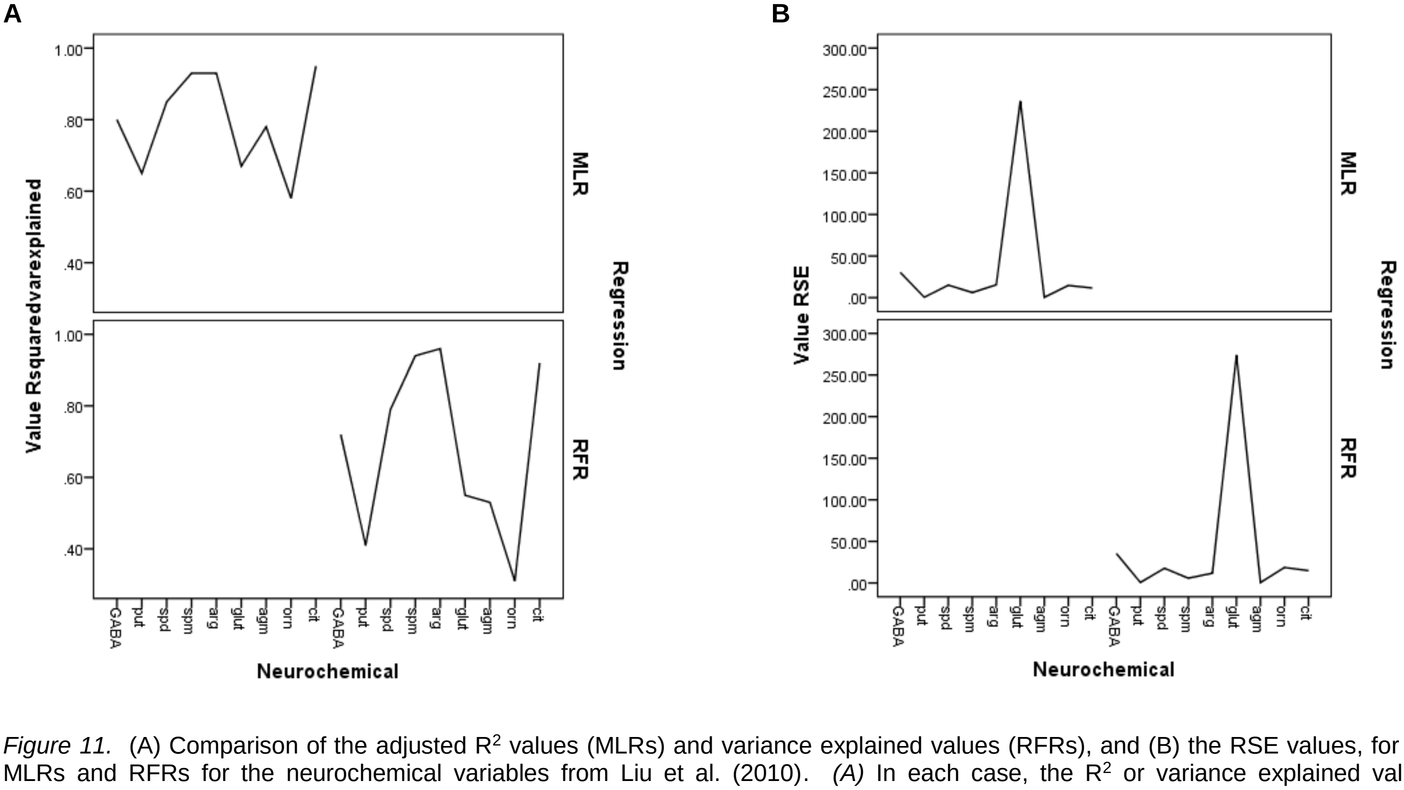

Figure 11. (A) Comparison of the adjusted R^2 values (MLRs) and variance explained values (RFRs), and (B) the RSE values, for the MLRs and RFRs for the neurochemical variables from Liu et al. (2010). *(A)* In each case, the R^2 or variance explained values represent how much of the variance of that variable could be accounted for by the combination of all of the remaining variables, e.g., an R^2 of 0.93 for the MLR means that 93% of the variation in spermine could be predicted by a combination of the other variables. While R^2 values can be used for MLR, RFR uses 'variance explained values' in an analogous way, so that a 'variance explained value' of 0.94 for spermine means that, using RFR, 94% of the variation in that variable could be explained by a combination of the other variables. *(B)* This figure shows the error in the model predictions for the two different kinds of regression, in the case of each prediction for each variable. Note the similarity in the pattern of the errors, despite the use of quite different regression methods. Note that, in both cases, the prediction for glutamate exhibits the most error (from Smith et al., 2013a with permission).

CONCLUDING REMARKS

Phenomena in neuroscience, whether at the level of genes, proteins, neurons or behavior, almost always involve the interaction of multiple variables, and yet many areas of basic neuroscience, in particular, employ univariate statistical analyses almost exclusively. This limits the ability of studies to reveal how the interactions between different variables may determine a particular outcome. For example, for the L-arginine data set that has been used as an example in this review, most of the individual neurochemicals measured were significantly different in the VNC and CE between young and aged animals, but this on its own does not indicate how they may work together to determine the consequences of aging. These neurochemicals are part of a complex system with feedback pathways (Fig. 1) and it is important to understand how this system works as a whole. The fact that age could be predicted from spermine and spermidine levels alone in the CE, with 100% accuracy, suggests that these two polyamines, which are chemically related, have special significance in the neurochemical signature of aging, beyond the other neurochemicals measured, even though many of these other neurochemicals were significantly different between the young and aged animals as well. This does not necessarily mean that they have a causal influence in the aging process; in fact, Fig. 1 shows that changes in spermine and spermidine are likely to be part of the polyamine output of the L-arginine system. However, their predictive significance suggests that they may be some kind of 'biomarker' for aging in the CE. Determining any causal role that they may have in the larger aging process, would require further experimental evidence. Further studies, using multiple age groups, have extended these findings (Liu et al., 2017). Elsewhere, MVAs and data mining methods have been used to explore the way that combinations of variables can account for neurochemical and behavioral changes following the loss of vestibular function (Zheng et al., 2012; Smith et al., 2013b; Zheng et al., 2013; Aitken et al., 2017) and auditory function (He et al., 2017). In clinical neuroscience research, MVAs and data mining methods have been used to predict the progression of patients from one neurological disorder to another (e.g., Krafczyk et al., 2006; Brandt et al., 2012) and the probability that the early adolescent use of *Cannabis* can lead to the development of psychotic symptoms in later life (e.g., Caspi et al., 2005). These methods are now in routine use in areas such as genomics, proteomics, metabolomics (Dziuda, 2010) and the analysis of fMRI data (e.g., Chen et al., 2017). Electrophysiological research in neuroscience is increasingly moving to the use of multi-electrode arrays using 16 or more micro-electrodes simultaneously, and in this situation one of the main objectives is to determine how different brain regions change in relation to one another, which requires MVA (e.g., Staude et al., 2010).

Correlation amongst variables in multivariate data is often a concern. In the case of MLR, it is a significant problem and specific tests of multicollinearity must be undertaken to ensure the validity of the analysis (Brook and Arnold, 1985; Vittinghoff et al., 2005; Ryan, 2009). Similarly, multicollinearity is a problem for LDA (Noes and Mevik, 2001). For PCA, too much correlation amongst the variables can be a problem; however, so can too little (Field, 2011). Since PCA is looking for underlying components or dimensions, it would be expected that the variables comprising those dimensions share a reasonable amount of correlation. In the PCA menu of SPSS 24, Bartlett's test can be used to determine whether the degree of correlation is too low and the correlation matrix resembles what is known as an '*identity matrix*' (Joliffe, 2004; Field, 2011). On the other hand, if the degree of correlation is too high, this can also be a problem. If all of the variables were perfectly correlated, then the correlation matrix would have the property known as *'singularity'* (Joliffe, 2004; Field, 2011). Multicollinearity is not a major problem for PCA as an exploratory technique (although it is for the more formal FA). However, the contribution of highly correlated variables to the PCs may be over-emphasized and Field (2011) suggests that it may be wise to inspect the correlation matrix and remove variables that are highly correlated, e.g., R > 0.8. There are no simple solutions to this issue and the best procedure is to carefully examine the effects of these variables on the PCA results. Removing variables without clear evidence of redundancy could also adversely affect the validity of the analysis. In the case of the non-parametric CA, the dendrogram is intended to reveal the degree of correlation amongst the variables, and in this case the objective is to determine the variables that co-vary, although it is usually of more interest to find co-variation that was not expected. For data mining methods such as SVMs and RFRs, correlation is less of a concern, because there are few distributional assumptions (Breiman et al., 1984; Marsland, 2009; Hastie et al., 2009) and it would be expected that variables that are not useful in discriminating between groups would be excluded or de-emphasized (Pang et al., 2006). For SVMs, the influence of correlation will depend on the specific kernel used, e.g., the linear kernel will be subject to the effects of multicollinearity in a similar way to MLR. A major advantage of methods such as SVMs and RFRs is the use of cross-validation and ROC curves to determine the predictive success of the model (Hastie et al., 2009; Williams, 2011).

Although a good understanding of univariate statistics is necessary in order to use MVAs effectively (see Smith, 2017), it is an investment in time worth making in order to obtain maximal benefit from data that are often difficult to collect and may involve the sacrifice of animal life. The advances that have been made and are currently occurring in technology mean that more often than not in the future, neuroscientists will have data from many variables simultaneously and MVAs and data mining procedures will offer the only way of effectively analyzing such data sets.

REFERENCES

Aitken P, Zheng Y, Smith PF (2017) Ethovision™ analysis of open field behaviour in rats following bilateral vestibular loss. J Vestib Res 27:89-101.

Anzanello MJ, Ortiz RS, Limberger R, Mariotti K (2014) Performance of some supervised and unsupervised multivariate techniques for grouping authentic and unauthentic Viagra and Cialis. Egyptian J Forensic Sci. 4:83-89.

Babayak MA (2004) What you see may not be what you get: a brief, non-technical introduction to over-fitting in regression-type models. Psychosomat Med 66:411-421.

Blunch NJ (2008) Introduction to structural equation modelling using SPSS and AMOS. Los Angeles: Sage.

Bock RD (1975) Multivariate statistical methods in behavioral research. New York: McGraw Hill.

Brandt T, Strupp M, Novozhilov S, Krafczyk S (2012) Artificial neural network posturography detects the transition of vestibular neuritis to phobic postural vertigo. J Neurol 259:182-184.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. 1st Edition. Boca Raton: CRC Press.

Brook RJ, Arnold GC (1985) Applied regression analysis and experimental design. Boca Raton: Chapman and Hall/CRC.

Caspi A, Moffitt TE, Cannon M, McClay J, Murray R, Harrington H, Taylor A, Arseneault L, Williams B, Braithwaite A, Poulton R, Craig IW (2005) Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. Biol Psychiat 57(10):1117-27.

Chen H, Cao G, Cohen RA (2017) Multivariate semiparametric spatial methods for imaging data. Biostat 18(2):386-401.

Crawley MJ (2007) The R Book. Chichester, UK: Wiley.

Davies TM (2016) The book of R. San Francisco, CA: No Starch Press.

Doncaster CP, Davey AJH (2007) Analysis of variance and covariance. Cambridge: Cambridge University Press.

Dziuda DM (2010) Data mining for genomics and proteomics. Hoboken, NJ: Wiley.

Field A (2011) Discovering statistics using SPSS. Los Angeles, CA: Sage.

Field A, Miles J, Field Z (2012) Discovering statistics using R. Los Angeles: Sage.

Gamst G, Meyers LS, Guarino AJ (2008) Analysis of variance designs. A conceptual and computational approach with SPSS and SAS. New York: Cambridge University Press.

Gurney K (1997) An introduction to neural networks. Boca Raton: CRC Press.

Hartung J, Knapp G (2005) Multivariate multiple regression. In Encyclopedia of statistics in behavioral science (Everitt BS, Howell DC, eds) pp 1370-1373. Chichester, NH: John Wiley and Sons.

Hastie T, Tibshirani R, Friedman J (2009) Elements of statistical learning: data mining, inference and prediction. 2nd Edition. Heidelberg, Germany: Springer Verlag.

He J, Zhu Y, Aa J, Smith PF, De Ridder D, Wang G, Zheng Y (2017) Brain metabolic changes in rats following acoustic trauma. Front Neurosci 11:148. doi: 10.3389/fnins.2017.00148. pp 1-13.

Jolliffe IT (2004) Principal component analysis. 2nd Edition. New York: Springer.

Kaplan D (2009) Structural equation modelling. Foundations and extensions. 2nd Ed. Los Angeles, CA: Sage.

Kitbumrungrat K (2012) Comparison of logistic regression and discriminant analysis in classification groups for breast cancer. IJCSNS Int. J. Comp. Sci. Network Security. 12(5):111-115.

Kline P (2002) An easy guide to factor analysis. London: Routledge.

Krafczyk S, Tietze S, Swoboda W, Valkovic P, Brandt T (2006) Artificial neural network: a new diagnostic posturographic tool for disorders of stance. Clin Neurophysiol 117:1692-1698.

Krzanowski WJ. (2005) Multivariate analysis: an overview. In Encyclopedia of statistics in behavioral science (Everitt BS, Howell DC, eds) pp 2-8. Chichester, NH: John Wiley and Sons.

Lattin J, Carroll JD, Green PE (2003) Analyzing multivariate data. Pacific Grove, CA: Duxbury.

Liong C-Y, Foo S-F (2013) Comparison of linear discriminant analysis and logistic regression for data classification. AIP Conf. Proc. 1522:1159.

Liu P, Gupta N, Jing Y, Collie ND, Zhang H, Smith PF (2017) Further studies of age-related changes in arginine metabolites in the rat vestibular nucleus and cerebellum. Neurosci 348:273-287.

Liu P, Zhang H, Devaraj R, Ganesalingam G, Smith PF (2010) A multivariate analysis of the effects of aging on glutamate, GABA and arginine metabolites in the rat vestibular nucleus. Hear Res 269:122-133.

Manly BFJ (2005) Multivariate statistical analyses. A Primer. 3rd Edition. London, UK: Chapman and Hall/CRC.

Marcoulides GA, Hershberger SL (1997) Multivariate statistical methods. A first course. Mahwah, New Jersey: Lawrence Erlbaum Assoc.

Marsland S (2009) Machine learning. An algorithmic perspective. Boca Raton, FL: CRC Press.

Noes T, Mevik B-H (2001) Understanding the collinearity problem in regression and discriminant analysis. J Chemometrics 15:413-426.

Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H (2006) Pathway analysis using random forests classification and regression. Bioinformatics 22:2028-2036.

Pohar M, Blas M, Turk S (2014) Comparison of logistic regression and linear discriminant analysis: a simulation study. Metodoloski Zvezki 1(1):143-161.

Questier F, Put R, Coomans D, Walczak B, Vander-Heyden Y (2005) The use of CART and multivariate regression trees for supervised and unsupervised feature selection. Chemometrics and Intelligent Laboratory Systems. 76(1):45-54.

Rutherford A (2001) Introducing ANOVA and ANCOVA. A GLM approach. London, UK: Sage Publications.

Ryan TP. (2009) Modern regression methods. 2nd Edition. New York: Wiley.

Ryan M, Mason-Parker E, Tate WP, Abraham WC, Williams JM (2011) Rapidly induced gene networks following induction of long term potentiation at perforant synapses in vivo. Hippocampus 21:541-553.

Seber GAF (1984) Multivariate observations. New York: Wiley.

Smith PF (2012) Statistical analysis in pharmacology is not always BO. Trends Pharmacol Sci 33(11):565-566.

Smith PF (2017) A guerilla guide to common problems in 'neurostatistics': essential statistical topics in neuroscience. J Undergrad Neurosci Ed 16(1):R1-R12.

Smith PF, Ganesh S, Liu P (2013a) A comparison of random forest regression and multiple linear regression for prediction in neuroscience. J Neurosci Meth 220:85-91.

Smith PF, Haslett SJ, Zheng Y (2013b) A multivariate statistical and data mining analysis of spatial memory-related behaviour following bilateral vestibular deafferentation in the rat. Behav Brain Res 246:15-23.

Staude B, Rotter S, Grün S. (2010) CuBIC: cumulant based inference of higher-order correlations in massively parallel spike trains. J Comput Neurosci 29(1-2):327-350.

Stevens JP (2009) Applied multivariate statistics for the social sciences. 5th Edition., Hillsdale, NJ: Lawrence Erlbaum.

Tabachnick BG, Fidell LS (2007) Using multivariate statistics. 5th Edition. Boston, MA: Pearson Education Inc.

Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE (2005) Regression methods in statistics: linear, logistic, survival and repeated measures models. New York: Springer.

West SG, Aiken LS (2005) Multiple linear regression. In Encyclopedia of statistics in behavioral science (Everitt BS, Howell DC, eds) pp 1333-1338. Chichester, NH: John Wiley and Sons.

Williams GJ (2009) Rattle: a data mining GUI for R. The R Journal ½: 45-55.

Williams GJ (2011) Data mining with Rattle and R. New York: Springer.

Wilson MD (2008) Support vector machines. In Encyclopedia of ecology, pp 3431-3437. New York: Elsevier.

Zheng Y, Cheung I, Smith PF (2012) Performance in anxiety and spatial memory tests following bilateral vestibular loss in the rat and effects of anxiolytic and anxiogenic drugs. Behav Brain Res 235: 21-29.

Zheng Y, Wilson G, Stiles L, Smith PF (2013) Glutamate receptor subunit and calmodulin kinase II expression in the rat hippocampus, with and without T maze experience, following bilateral vestibular deafferentation. PLoS ONE 8(2):e54527. doi:10.1371/journal.pone.0054527, pp. 1-10.

Received September 25, 2017; revised February 26, 2018; accepted March 08, 2018.

Address correspondence to: Professor Paul Smith, Department of Pharmacology and Toxicology, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand. Email: paul.smith@otago.ac.nz

www.funjournal.org