Research

Oikos 118: 681–690, 2009
doi: 10.1111/j.1600-0706.2008.16881.x,

Subject Editor: Tim Benton. Accepted 17 November 2008

Partial least squares regression as an alternative to current regression methods used in ecology

Luis M. Carrascal, Ismael Galván and Oscar Gordo

L. M. Carrascal (lmcarrascal@mncn.csic.es), I. Galván and O. Gordo, Museo Nacional de Ciencias Naturales, CSIC, C/ José Gutiérrez Abascal 2, ES–28006 Madrid, Spain.

This paper briefly presents the aims, requirements and results of partial least squares regression analysis (PLSR), and its potential utility in ecological studies. This statistical technique is particularly well suited to analyzing a large array of related predictor variables (i.e. not truly independent), with a sample size not large enough compared to the number of independent variables, and in cases in which an attempt is made to approach complex phenomena or syndromes that must be defined as a combination of several variables obtained independently. A simulation experiment is carried out to compare this technique with multiple regression (MR) and with a combination of principal component analysis and multiple regression (PCA + MR), varying the number of predictor variables and sample sizes. PLSR models explained a similar amount of variance to those results obtained by MR and PCA + MR. However, PLSR was more reliable than other techniques when identifying relevant variables and their magnitudes of influence, especially in cases of small sample size and low tolerance. Finally, we present one example of PLSR to illustrate its application and interpretation in ecology.

Although the experimental method in scientific research, based on manipulative experiments, is the cornerstone of modern science, it requires solid hypotheses for testing. These hypotheses are often the consequence of abstract reasoning or inspired thought, though in some disciplines, such as ecology, they are mainly derived from patterns emerging from previous observations. When analyzing patterns, the variation in the response variables is established as a function of several predictors. These patterns of covariation are usually established by means of regression techniques. Multiple regression analysis is the most widespread statistical tool for this purpose. The relationships between the response (i.e. dependent) and the predictor (i.e. independent or explanatory) variables are measured with standardized regression coefficients and are interpreted as partial effects influencing the variability in the response variable. From these patterns of covariation, probable cause-effect interactions may be inferred (Quinn and Dunham 1983, James and McCulloch 1985, Hairston 1989).

The classical regression approach poses four main problems when analyzing ecological data. First, ecological phenomena, such as geographical variation of species richness, habitat preferences, ecomorphological relationships or colour patterns, are usually described by a large array of variables (e.g. different orographic, climatic and landscape descriptors, several habitat structure and floristic composition measurements, morphometric dimensions of bones and muscles, or reflectance measured at different intervals within the wavelength spectrum). Second, these multivariate descriptors are in many instances non-independent, as they are usually ordered in environmental gradients, habitat configurations or morphological syndromes (i.e. collineal patterns). This introduces complex interactions and redundancies. Moreover, the addition of the sum of squares of the partial effects of the predictors (ΣSS_i of each variable i) does not add up to the sum of squares of the whole model (SS_{model}). Third, if sample size (e.g. the number of different study plots or individuals) is not large enough compared to the number of predictor variables, the ability of regression analysis to find a significant effect is reduced. This fact consequently inflates type II errors (i.e. accepting a false null hypothesis). This limitation may be paramount in regression analyses when the magnitude of the effects is low (i.e. small R^2), the redundancy among predictors is high (i.e. low tolerance) and the degrees of freedom for each regression term are small (note that DF = sample size – number of response and predictor variables). On the other hand, a multiple regression analysis can not be carried out when the number of predictor variables equals the number of sample units, because the degrees of freedom equal zero. Finally, the classical regression approach does not address the analytical situation in which more than one response variable is considered together. This is the case when an attempt is made to approach complex phenomena such as body condition or health, which can be measured in several ways and must be defined as a combination of the several variables obtained independently (e.g. immune response, fat reserves, parasite infestation, etc.).

In spite of these difficulties, researchers often take shortcuts to overcome these problems. One shortcut is to

remove some variables, selecting redundant variables or those having non-significant effects on the response variable. In this way, the balance between the number of sample units and the number of explanatory variables is improved and type II errors decrease due to the avoidance of low degrees of freedom. Unfortunately, the process of selecting variables is generally not shown in published articles, and consequently this procedure becomes an unknown practice in scientific research. Another shortcut is to reduce the multidimensionality in the predictor or response variables by multivariate reduction methods, such as principal component analysis (PCA) or multidimensional scaling. In this accepted practice, the researcher carries out one or several PCAs, and the factor scores of components considered relevant or 'significant' are retained and included in subsequent multiple regression analyses. With this approach, the balance between the number of samples and the number of variables (now multivariate components) is improved, the redundancy among predictors does not exist and problems inherent to the variance partitioning are solved (i.e. $SS_{model} = \Sigma SS_i$). Nevertheless, the obtained components maximize the covariation among the predictor variables independently of the among-sample variation in the response variable. Therefore, components derived from any multivariate reduction technique do not pursue the maximization of the variance explained in the response variable subjected to study. This fact can yield patterns or syndromes within the explanatory variables making little or no biological sense and consequently hinders the interpretation of results. Furthermore, the application of any multivariate technique to the regression approach lengthens analytical time.

Apart from these caveats, one question emerges when facing the previously mentioned regression problems specific to many ecological studies dealing with patterns: is there any statistical tool directly oriented to treat, as a whole, multiple predictor variables often themselves related, which maximizes the explained variability in one or more response variables, when working with modest sample sizes? The answer is partial least squares regression analysis (PLSR hereafter), a little known statistical tool in ecological research but widely used in other scientific disciplines.

The use of PLSR in analytical chemistry began in the early 1980s and has increased steadily since then. In contrast, the use of PLSR in ecological studies began recently, in the late 1990s, though there has not been any increase in its use in recent years compared to other fields (Escabias et al. 2007). The intensive use of this statistical technique in other scientific fields, such as chemistry, is undoubtedly related to the mathematical properties and benefits associated with PLSR, which is especially appropriate in dealing with a large number of explanatory variables in comparison with the number of observations, and in cases of severe multicolinearity (Mevik and Wehrens 2007). While these types of data are quite common in ecological work, the use of PLSR remains surprisingly infrequent.

The lack of knowledge about PLSR may be a constraint in ecologists' ability to analyze data and to elucidate underlying patterns. In this paper, we highlight the utility of PLSR and propose this method as an alternative to the majority of multivariate techniques that are currently used in all fields of ecology. We believe that there is a great potential for the use of PLSR in ecological studies that has been overlooked until now. The applicability of this method will likely be higher for inductive approaches in which the aim is the definition of patterns of variation among large sets of usually related variables.

This paper has three goals related to the lack of information and scholarly practice of PLSR. The first is to briefly present the aims, requirements and types of results obtained from PLSR. The second goal is to carry out a simulation experiment varying the number of predictors and sample sizes in order to compare this technique with multiple regression and a combination of regression and multivariate methods, which are more commonly used in ecology. Finally, we present an example of PLSR to illustrate its application and interpretation in ecological studies.

The partial least squares regression

The partial least squares regression (PLSR) was developed by Wold in the late 1960s for econometrics (Wold 1975) and then introduced as a tool to analyze data from chemical applications in the late 1970s (Geladi and Kowalski 1986, Martens et al. 1986, Mevik and Wehrens 2007). An introduction and a statistical overview of PLSR can be found in Geladi and Kowalski (1986), Frank and Friedman (1993), Wold et al. (2001), Tobias (2003) and Abdi (2007). This technique is an extension of multiple regression analysis in which the effects of linear combinations of several predictors on a response variable (or multiple response variables) are analyzed. Associations are established with latent factors extracted from predictor variables that maximize the explained variance in the dependent variables. These latent factors are defined as linear combinations constructed between predictor and response variables, such that the original multidimensionality is reduced to a lower number of orthogonal factors to detect the structure in the relationships between predictor variables and between these latent factors and the response variables. The extracted factors account for successively lower proportions of original variance (Hubert and Branden 2003, Tobias 2003, Maestre 2004).

PLSR is especially useful when (1) the number of predictor variables is similar to or higher than the number of observations (i.e. overfitting) and/or (2) predictors are highly correlated (i.e. there is strong collinearity). The first situation constitutes a limitation because regression coefficients cannot be calculated. The second situation produces erratic signs in regression coefficients, thus increasing the difficulty of interpreting the linear regression equation. These properties should render this technique an essential analytical tool for consideration in the design of any scientific study in which the number of sample units collected is low with respect to the number of measured explanatory variables. The application fields of PLSR also cover situations in which there are more than one response variable, thus serving as an alternative to MANOVA designs. In these cases where multiple response variables are used, PLSR creates other latent factors from the linear

combination of the original response variables that act as synthetic response variables.

PLSR is implemented in the statistical packages most widely used by ecologists, such as Statistica (StatSoft 2001), SAS/STAT (SAS 2001), SPSS (<www.spss.com>); Statgraphics (<http://www.statgraphics.com/>), MatLab (Andersson and Bro 2000), or as an add-in for Microsoft Excel spreadsheets (XL-Stat, <www.xlstat.com>). Hence, it is a readily available statistical tool.

A comparative analysis of PLSR with other statistical approaches

By comparison of the explanatory capacity of the models (i.e. R^2) and the consistency of the results according to parameter estimates and their significance (i.e. coefficients and p-values) given different scenarios of sample sizes and numbers of predictor variables, we attempt to illustrate the pros and cons of PLSR against two usual alternative approaches: multiple regression (MR) and principal component analysis followed by a multiple regression with the obtained components (PCA + MR). Explanatory capacity is usually the most important goal of ecological modelling, however the type, sign and consistency of relationships between the response and explanatory variables are equally important, because they are the basis for our interpretations of numerical outputs of statistical software packages and consequently of the establishment of functional links with a biological meaning among variables.

Results obtained by each statistical method were checked against a 'true model' (hereafter TM). The TM has known statistical properties generated ad hoc by simulation processes, and thus it is the 'omniscient truth' of relationships and effects structuring the data. The TM was a data matrix of 5000 sample units and 21 variables with different levels of association among them (one variable served as a response variable, while the remaining 20 were used as continuous predictors). By a randomization process, several data subsets of different sample sizes were chosen to measure the consistency of the results obtained in different sampling trials, and thus to test the robustness of the three compared statistical methods.

The data matrix

Data were generated using the randomization routines of Pop Tools 2.7 (<http://www.cse.csiro.au/poptools/>). Twenty predictor variables of different means and standard deviations were built showing diverse levels of relationships among them. All predictor variables had a normal distribution. Absolute values of correlations between pairs of predictor variables ranged between –0.987 and 0.987. The tolerance for each variable ranged between 0.788 (i.e. highly independent of the remaining 19 predictor variables) and 0.024 (i.e. highly redundant with the informative content of the other variables). The total sample size randomized was 5000 sample units.

After this first randomization process, a multiple regression equation of known parameters was built to predict the response variable. The equation included both positive and negative regression coefficients showing a varying degree of association with the predicted response variable, from very intense to nearly null influence. After this first step, a random variable following a normal distribution was added to the predictions of the previous multiple regression model, thus establishing the final response variable.

Three different groups of predictor variables were selected from this matrix of 20 predictors: (1) all variables, (2) eight highly related (i.e. redundant) predictor variables, and (3) eight variables with a low level of relationship among them (i.e. relatively independent with small pairwise correlations and high tolerances). These three sets of predictor variables configure two different scenarios of many vs few variables, and highly vs scarcely related.

Random selection of samples

Four different scenarios were defined according to sample size: 15, 30, 60 and 120 sample units. These are typical ranges of sample size managed by ecologists. Therefore, 12 experimental situations were simulated (three variable sets × four sample sizes). For each of these experimental situations, 20 random extractions without replacement of 15, 30, 60 or 120 samples were made within the all 5000 sample units (shuffle design – all sampling trials obtained different sample units). This design mimics the reality of ecological sampling in which a researcher would have sampled the same phenomenon on 20 different occasions, obtaining at every occasion different records from the same population or within the same environmental gradients.

Generation and evaluation of statistical models

Partial least square regressions (PLSR)

We retained only 'significant' components, which we defined as those explaining more than 5% of original variance in the response variable. Two components were always retained in all simulations. We gathered from each simulation the explanatory capacity (R^2) of each component as well as the weight of each predictor within each component, which helped us to understand the latent factors defined by each component. The sum of the R^2 of the two significant components gave us the total explanatory capacity of the PLSR models. The correlations between the weights for each simulation and those of the TM model using the 5000 sample units were obtained as an estimate of the consistency of patterns obtained in each sample.

We also evaluated the consistency between the results in each simulation by counting the coincidences of 'significant' (i.e. with square weights > 0.05) and 'non-significant' (square weights < 0.05) variables in each component of each trial in comparison with the TM obtained for the whole sample of 5000 sample units. This fact allowed us to calculate a percentage of correctly assigned 'significant' and 'non-significant' weights for each set of explanatory variables and sample sizes. These coincidences are hereafter referred to as 'hits' for the sake of brevity. The percentage of hits among 'significant' variables in our results that should be 'significant' according to the TM (called 'positive hits') is inversely related to the type I error (to reject the null hypothesis when in fact it is true), whereas the percentage of hits among non-significant variables in our results that

should be non-significant according to the TM (called negative hits) is inversely related to the type II error (to accept the null hypothesis when in fact it is false).

Multiple regression (MR)

We performed saturated models and gathered for each simulation the standardized estimates of the coefficients for each predictor (β) and its significance (p-value), along with the explanatory capacity of the model (adjusted R^2). We then calculated the correlation between the β's from each simulation with β's from the whole sample of 5000 sample units (TM) to obtain a measure of the degree of consistency in the effects (magnitude and sign) of each explanatory variable on the response variable. Correlations were averaged for each sample size in each set of variables.

Consistency of the effects of the explanatory variables was also evaluated by means of the 'hits' procedure (coincidence among significant and non-significant predictor variables in each trial compared to the 'true model', TM). The percentage of correctly assigned significant and non-significant predictor variables was calculated for each sample size and set of variables.

'Best subsets' is another way to select the most relevant variables in our set of predictors (Johnson and Omland 2004, Whittingham et al. 2006). Ecologists are usually faced with the existence of several highly probable models. This fact makes conclusions difficult as each model may represent a compelling hypothesis. However, there is only one true relationship between explanatory and response variables (in our case the TM). In our case, we calculated corrected Akaike information criterion (AICc; Burnham and Anderson 2002) values for models with eight explanatory variables. We could not perform this with the set of 20 variables because the number of possible models ($2^{20}-1=1048575$) exceeds computing capacity of our statistical software. Models were ordered according the AICc value (corrected for small sample sizes) and we counted the total number of models with an increase in the AICc (ΔAICc) value lower than 2 with respect to the best ranked model. Only those models with $\Delta AICc < 2$ have a substantial level of empirical support (i.e. those models must be considered because all of them are highly probable; Burnham and Anderson 2002). The higher the number of highly probable models, the higher the number of compelling hypotheses and thus greater difficulty in ascertaining reality. Hence, the amount of highly probable models can also be considered an approximation of the consistency of results.

Principal components analyses+mMultiple regressions (PCA+RM)

The application of a PCA prior to MR requires greater time and effort as more statistical analyses must be preformed and the identity of new variables (i.e. factors) obtained from the original set of explanatory variables must be interpreted. We retained only those factors with an eigenvalue equal to or higher than one. Factor loadings of explanatory variables within each component are used to ascertain the meaning of each factor. The meaning of components is usually defined only by those variables with loadings higher than a certain threshold (0.7 chosen arbitrarily in this paper). We correlated factor loadings of components obtained for the whole 5000 sample units (TM) with those obtained for each trial PCA model to test the consistency of the components' meaning. In a second step, we carried out a MR with the components obtained in the PCA as predictor variables and we noted the adjusted R^2 of the saturated model.

Consistency of results among the different statistical approaches

Explanatory capacity (R^2) of the statistical approaches

The first component of PLSR simulated models explained about 85% of variability in those trials with 20 and 8 poorly-related variables. However, in the case of eight highly-related variables, this proportion diminished to 63%. The final explanatory ability of the PLSR models considering the first two components was very similar among the three sets of variables, explaining more than 83% of the original variance in the response variable (Fig. 1). These patterns illustrate a typical situation in ecologcial research, in which the addition of new explanatory variables to models is often at the expense of higher levels of collinearity among them, but with a modest gain of explained variance. In this way, PLSR is unaffected by this phenomenon because it detects the main multivariate

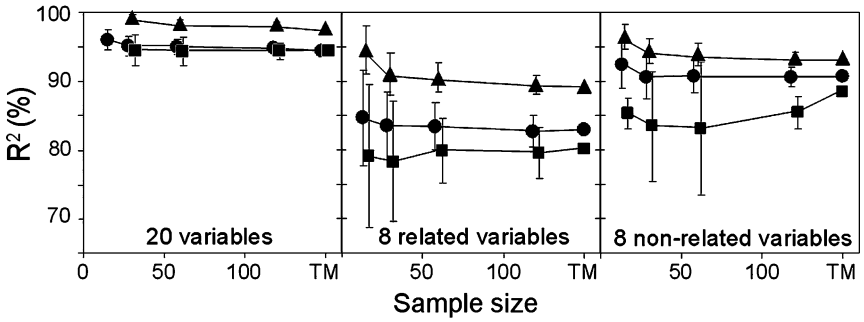

Figure 1. Variation of the explanatory capacity (R^2) of the PLSR (dots), MR (triangles) and PCA+MR (squares) under different scenarios of sample size and sets of explanatory variables. True values of R^2 for the whole dataset are shown as TM ('true model' or true pattern with the whole sample of 5000 sample units; see text). Vertical bars are SD.

syndromes or latent factors that maximize the variance explained in the response variable.

The effect of sample size was weak in the three different scenarios of explanatory variables. Results were quite similar to those obtained for the 'true model' (TM), even with just 15 cases. Furthermore, R^2 values of the different PLSR models were slightly affected by the stochasticity involved in the random selection of sample units (see SD bars in Fig. 1). Therefore, PLSR gave reliable results even in the least desirable situations of low sample sizes and large numbers of variables.

MR and PCA+RM are also able to explain a large amount of variability in the response variable (i.e. R^2 is always very high; Fig. 1), however they cannot be applied in analytical scenarios of more predictor variables (20) than sample units (15). The variability of PCA+MR results was much higher than in the other approaches (see SD bars in Fig. 1). High collinearity among predictors (i.e. models with eight highly-related variables) also reduced the explanatory capacity achieved by the models, with PCA+MR again being the most negatively affected method. Changes in sample size did not severely affect the explanatory ability, although MR was the approach that tended to more frequently overestimate the true effect of the explanatory variables with low sample sizes. Therefore, PCA+MR was the worst method, while MR was the best.

To summarise, based on the models' explanatory capacity (i.e. R^2), the PLS regression and multiple regression accounts for an amount of variance very similar to the expected one ('true model'). The combination of principal component analysis and multiple regression produced less confident results (lower explained variance and results more variable among different trials) and was the least repeatable and least robust against reductions in the number of explanatory variables or increases in their collinearity.

Meaning of PLSR components and MR and PCA patterns

The identity of the first component of PLSR models was extremely robust to variations in sample size. Correlations among weights obtained in each PLSR trial and those of the TM were always higher than 0.9 and reached values close to 1 with sample sizes of 60 or 120 sample units (Fig. 2). Therefore, the pattern found in the predictor variables that

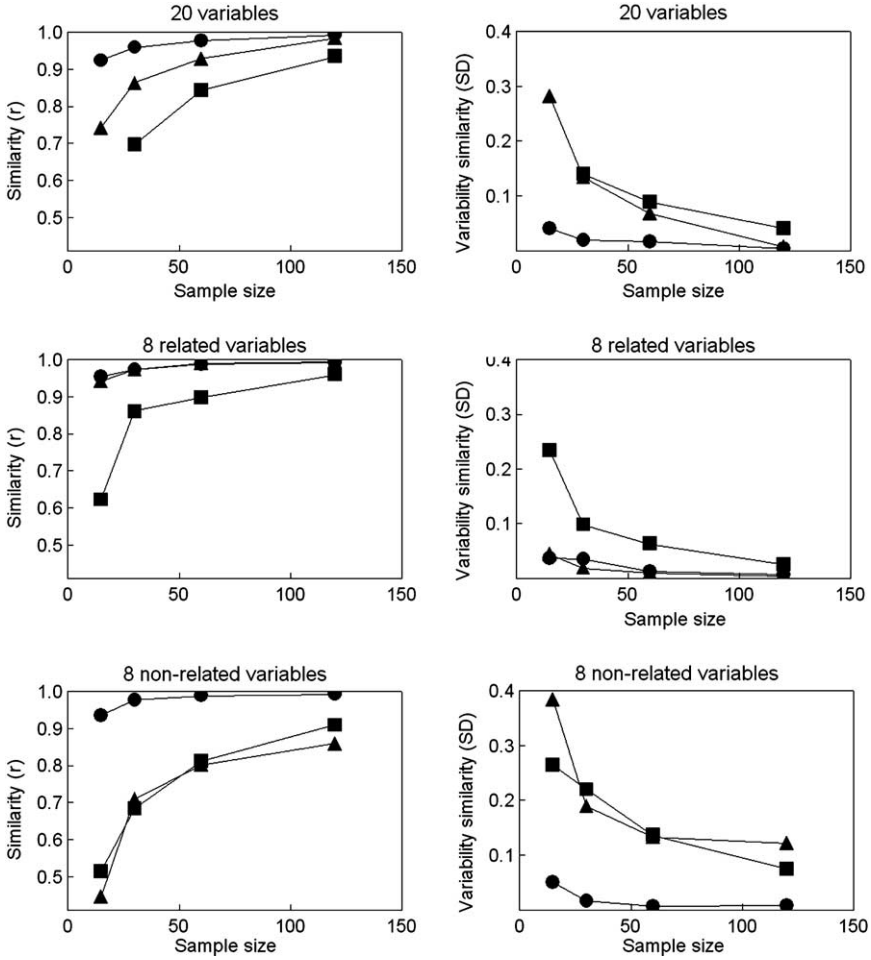

Figure 2. Consistency of results between simulations and the 'true model' (i.e. true pattern with the whole sample of 5000 sample units). Similarity is measured by means of the Pearson correlation (r) coefficients (20 trials per point) between weights of the two first components of PLSR models or β's of all variables in MR and those values corresponding to the TM for different sample sizes and sets of variables. Right graphs represent standard deviation values for these average correlation coefficients. Dots are values for the weights of the first component of the PLSR, triangles for the second component of the PLSR and squares for the βs of the MR.

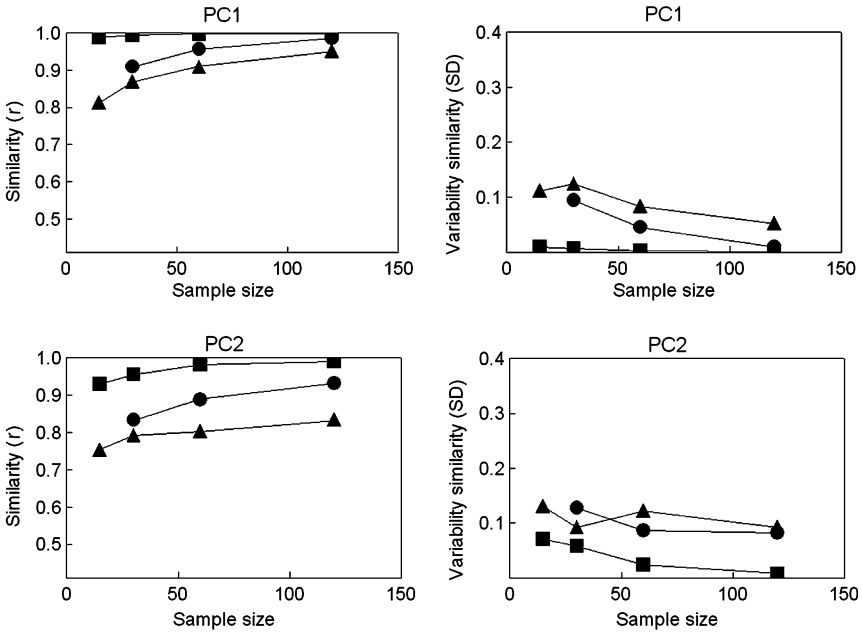

Figure 3. Consistency in the meaning of components of PCAs. Left graphs represent average Pearson correlation coefficients between factor loadings of simulations and those factor loadings corresponding to the 'true model' (TM; i.e. true pattern with the whole sample of 5000 sample units). Right graphs represent standard deviation values for these average correlation coefficients. There are 20 trials per point. Dots are values for the set of 20 variables, triangles for eight non-related variables and squares for eight related variables.

maximizes the variability explained in the response variable has ever the same meaning.

The second component of the PLSR model was less similar to that expected in the TM, although similarity figures are very high in working with 30 or more sample units and 20 predictor variables or a subset of eight highly-related variables (Fig. 2). Only in the case of eight non-related predictor variables was the similarity low, especially with a sample size lower than 30. It should be noted that the second component in the sets of 20 variables and eight less-related variables had very low relevance, since it accounted for a minimal proportion of the total explained variance (R^2) of the response variable (about 5–10%).

Correlation values between the standardized regression coefficients (β) from MR trials and those obtained in the whole sample of 5000 sample units (TM) are low and very dependent on working sample sizes (Fig. 2). Estimates of β's reach reliability only with large samples ($n = 120$). Although MR shows slightly higher R^2 than PLSR (Fig. 1), the influence of the predictor variables is less repeatable. This fact is a concern because it implies that conclusions from statistical analyses will depend strongly on the inherent stochastiticy involved in the sampling process, especially when working with small sample sizes.

Principal components analysis carried out with the whole 5000 sample units provided only two components with an eigenvalue higher than 1 in the sets of eight predictors, while the full set of 20 predictors provided four. For this reason we limited our consistency analyses to the first two components of the PCA. The average correlation between factor loadings of PCA models in trials with those of the TM were high in general, which stresses that the meaning of components was consistent even with low sample sizes (Fig. 3). Consistencies were especially high when variables were strongly collinear. The first component was more consistent than the second, similar to that found in PLSR.

The probability of correctly rejecting the false null hypothesis and thus accepting the alternative true hypothesis, (i.e. power of the test) can be assessed considering the results shown in Fig. 2 and 3. The similarity between the true patterns of relationship of the predictor and response variables was always higher, and the variability in this similarity was always lower, in the PLSR compared to MR for any sample size, number of variables involved or degree of relationship among predictor variables. This is especially evident with the first component of the PLSR which accounts for a large proportion of the explained variance in the response variable. Similarly, higher 'power' is obtained in the PLSR when compared to PCA + MR, except when working with eight related variables, where both methods provided patterns very similar to the 'true model'. Nevertheless, the explained variance in the response variable was always higher in the PLSR than in the PCA + MR method (Fig. 1).

'Hits' in PLSR and MR

Inconsistencies found previously by standardized regression coefficients (β) in MR are even more evident by assessing the number of 'hits' in the correct assignment of significant and non-significant effects of predictors. Furthermore, the effect of sample size is stronger in MR than in PLSR. In most cases with small sample sizes, MR is unable to identify significant relationships between predictors and the response variable when there are, in fact, significant

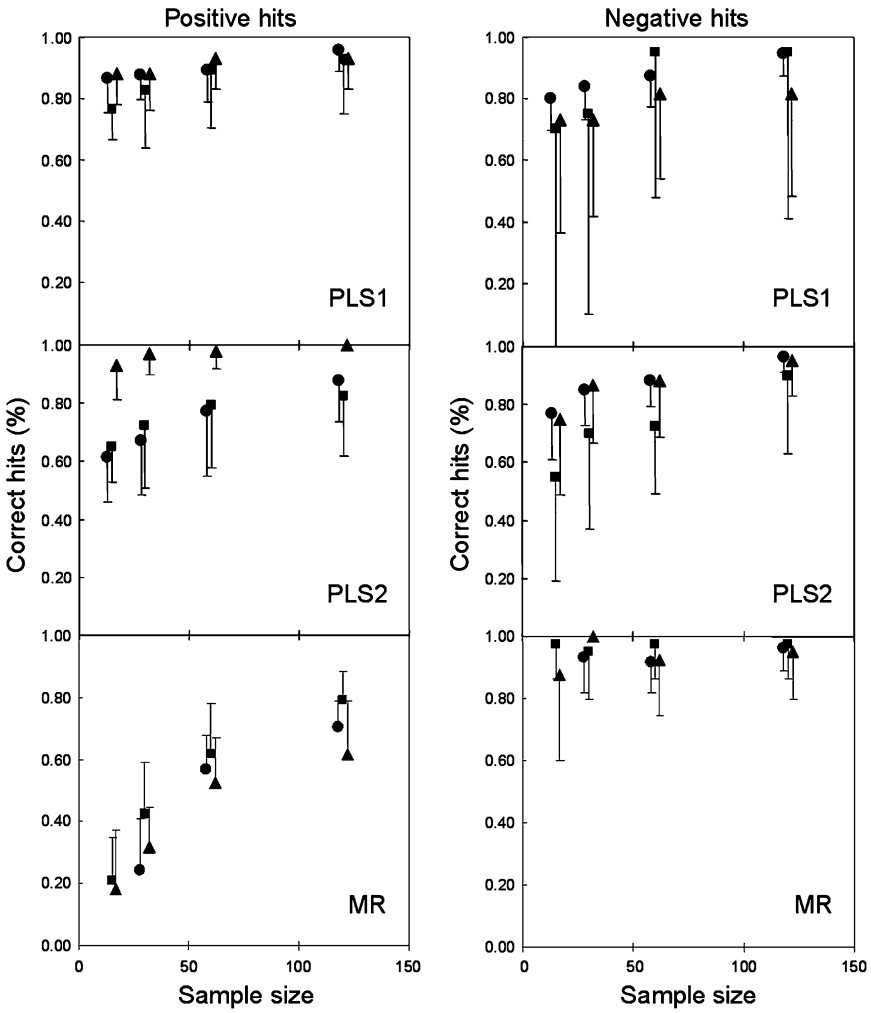

Figure 4. Percentage of correct 'hits' in the assignment of significance or non-significance of results according to the TM ('true model' or true pattern with the whole sample of 5000 sample units; see text). Positive 'hits' refer to significant variables in the TM that were correctly assigned as significant in models. Negative 'hits' refer to non-significant variables in the TM that were correctly assigned as non-significant in models. PLS1: results for the first component of the PLSR; PLS2: results for the second PLSR component; MR: multiple regression results. Dots: set of 20 variables; triangles: eight related variables; squares: eight non-related variables. Vertical bars show standard deviations.

relationships present (Fig. 4). Even with a large sample size, such as 120, percentages of correct positive significant effects in MR are lower than those observed in the PLSR. Nevertheless, MR showed a high level of agreement with truly non-significant effects under a large variety of analytical scenarios.

Those variables with a 'significant' weight within the first PLSR component of the TM are correctly assigned as significant in PLSR simulations in more than 85% of cases (Fig. 4). This percentage reaches values around 93–96% with a sample size of 120 data points. In the case of negative 'hits' (truly 'non-significant' effects identified as 'non-significant'), percentages are only slightly lower for the first component of PLSR models (Fig. 4), especially with high collinearity of predictors.

In the case of the second component of PLSR models, correct assignments are fewer than in the first component, but remain high in most cases, especially identifying true 'significant' effects working with eight highly related variables.

Hence, MR models are unable to detect significant effects of explanatory variables more often than PLSR, especially in cases with small sample sizes and high levels of collinearity. PLSR arises again as strongly robust to variations in sample size and strongly shielded against both type I and type II errors.

Best MR models according to AICc

Ideally, there exists a single best model according to AICc as there is only one true relationship between the response and explanatory variables. However, MR yields up to 13 highly probable models when analysing datasets with eight predictor variables (mean = 4.11, and SD = 2.59 for all the array of possible analytical scenarios). The number of highly probable models increased with sample size and was not affected by the degree of collinearity of predictors.

In summary, partial least squares regression analysis provides similar results to those obtained with multiple regression, or a combination of principal components

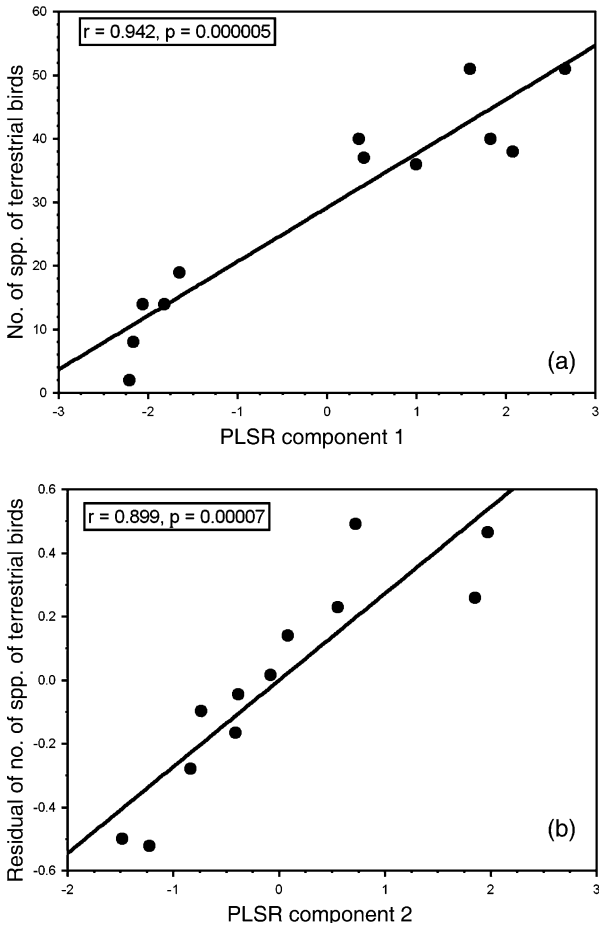

Figure 5. Scatterplots of the PLSR analyzing the species richness of terrestrial bird species in the Canary and Selvagem Islands. Sample size is all 12 islands larger than 1 km^2 in the archipelago. (a) relationship between the species richness and the position of each island in component 1 of the PLSR. (b) residual variation in species richness, after removing the effect of component 1 in (a), with the second component. For the meaning of each component see Table 1.

analysis plus multiple regression, according to the amount of explained variance. However, PLSR produces more stable results with regard to the identification of the relevant variables and their magnitudes of influence independent of the sample size in the analyses, a situation in which other regression approaches fail. In addition, the probability of correctly rejecting the false null hypothesis, and thus accepting the alternative true hypothesis (i.e. power of the test), was higher in the PLSR analysis. PLSR may simplify some statistical analyses in working with ecological data, with the additional advantage that it can be applied to cases in which sample sizes are equal to or lower than the number of predictor variables. The latter makes PLSR an excellent alternative in studies in ecology, where the relationships between response and predictor variables are often complex due to the high redundancy and interactions among groups of predictor variables. This is especially the case in ecological designs in which the aim is to test hypotheses when predictor variables, though not high in number, interact and/or cancel each other.

An example of application of PLSR

Recent examples of the use of PLSR can be found in some areas of ecology, such as microbial ecology (Stepanauskas et al. 2003, Allen et al. 2005), paleolimnology (Zhang et al. 2007), limnology (Larocque et al. 2006, Karle et al. 2007, Sobek et al. 2007), soil ecology (Ekblad et al. 2005), ecotoxicology (Sonesten 2003, Spanos et al. 2008), environmental effects on biodiversity (Maestre 2004, Davis et al. 2007, Palomino and Carrascal 2007), palaeoclimatological reconstructions and biogeography (Seppa et al. 2004), large scale influence of climate (Bergant et al. 2006, Finsinger et al. 2007), biodiversity mapping (Schmidtlein 2005), definition of ecological indicators (Amand et al. 2004, Potapova et al. 2004), community ecology (ter Braak and Schaffers 2004, Carrascal and Alonso 2006), modelling of phenology (Gordo et al. 2008) and ecomorphology (Hoffsten 2004). Results of Zhang-yu et al. (2007) also demonstrate that PLSR is more effective than stepwise MR or regression analyses with PCA in relating hyperspectral leaf reflectance in rice *Oryza sativa* crops to the disease severity of the fungus *Bipolaris oryzae*.

As an illustrative example of PLSR application, we analyzed differences in the number of breeding landbird species among the Canary and Selvagem Islands using data on area, distance to mainland, maximum altitude and age of each island, average structural complexity of habitats on each island, and within-island habitat diversity (Carrascal and Palomino 2002). In this example, the number of independent or predictor variables is six, while the sample size (number of islands) is twelve. Table 1 shows the results of MR and PLSR.

The multiple regression model (MR) provides a very significant result ($F_{6,5} = 41.96$, $p = 0.0004$) accounting for 98.1% of among-island variability in species richness. Paradoxically, none of the predictor variables reached significance at $\alpha = 0.05$ in the multiple regression. Therefore, an extremely explanatory and significant model is obtained but which effects are significant remains unknown. This bizarre result is due to a doubly undesirable phenomenon in ecological research: sample size cannot be enlarged by increasing sampling effort (all the islands in the archipelago were studied), and all the relevant predictor variables are highly correlated. This last concern is well illustrated by the very low tolerance reached by predictor variables (Table 1).

Results of the PLSR analysis provide two significant components explaining 97.8% of the original variance in the response variable (Table 1). The amount of variance explained is very similar to that obtained by MR. The first component accounts for a major proportion of the explained variance, while the second component accounts for a marginal, but significant, 9.1%. The meaning of the components can be interpreted considering the weights attained by the variables. The addition of the squares of the weights within each component sums to one, so the contribution of each predictor variable to the meaning of each component can be easily estimated. Component 1 mainly associates island area to the maximum altitude and the landscape diversity of each island. This means that these predictor variables cannot be seen as independent variables, but they comprise an 'island syndrome' affecting bird species richness. These three variables alone retain

Table 1. Results of the multiple regression analysis (MR) and the partial least squares regression analysis (PLSR) carried out with the number of terrestrial bird species breeding in the Canary and Selvagem Islands (response variable) and six predictor variables describing the large scale characteristics of 12 islands. For original data, see Carrascal and Palomino (2002). Tolerance: 1 minus the squared multiple correlation of that variable with all other independent variables in the regression equation (higher values denote more independent variables). Beta: standardized multiple regression coefficient. w COMP 1 and 2: weights of each variable in the first and second PLSR components. R^2: proportion of the variance in the response variable accounted for by the multiple regression analysis or each component of the PLSR. All predictor variables were included log-transformed in the models. PLSR weights whose squares are larger than 0.2 are shown in bold type.

	Tolerance	MR		PLSR	
		Beta	p	w COMP1	w COMP2
Island area	0.042	0.379	0.270	**0.553**	0.429
Maximum altitude	0.042	0.379	0.268	**0.530**	0.173
Distance from mainland	0.310	−0.180	0.169	0.079	**–0.751**
Island age	0.268	0.149	0.270	0.334	0.074
Average habitat complexity	0.235	0.013	0.922	0.253	**–0.465**
Landscape diversity	0.137	0.223	0.244	**0.481**	0.025
R^2		0.981		0.888	0.091
p		<0.0001		<0.0001	<0.0001

81.8% of the information content of the first component ($0.553^2+0.530^2+0.481^2=0.818$). The correlation between species richness and the position of the 12 islands in the first component of the PLSR is shown in Fig. 5a ($r=0.942$, $p<0.001$). The second PLSR component works on the residual variation not explained by the first component (i.e. $1-0.888=0.112$). Figure 5b depicts the relationship between the residual variation in species richness and the second component of the PLSR ($r=0.899$, $p<0.001$). Therefore, the original variance explained by the second PLSR component can be estimated as the proportion of the residual variance (after subtracting the first component) accounted for by the second component: $0.112 \times 0.899^2 = 0.091$. The information content of the second component is negatively associated with distance of the islands to mainland and the average habitat complexity of the landscape ($-0.751^2+-0.465^2=0.780$), identifying islands more distant from the dry African mainland as wetter and with more developed vegetation, defining another 'island syndrome'. Therefore, the among-islands variation in species richness in this archipelago can be disclosed as different environmental syndromes resulting from combinations of non-independent predictors in several components. The meaning and the true explanatory magnitude of each of these PLSR components can be then easily estimated thanks to the weights of the original predictor variables.

References

Abdi, H. 2007. Partial least square regression (PLS regression). – In: Salkind, N. J. (ed.), Encyclopedia of measurement and statistics. Sage.

Allen, A. E. et al. 2005. Influence of nitrate availability on the distribution and abundance of heterotrophic bacterial nitrate assimilation genes in the Barents Sea during summer. – Aquat. Microbiol. Ecol. 39: 247–255.

Amand, M. et al. 2004. A step toward the definition of ecological indicators of the impact of fishing on the fish assemblage of the Abore reef reserve (New Caledonia). – Aquat. Living Resour. 17: 139–149.

Andersson, C. A. and Bro, R. 2000. The N-way toolbox for MATLAB. – Chemometr. Intell. Lab. 52: 1–4.

Bergant, K. et al. 2006. Uncertainties in modelling of climate change impact in future: an example of onion thrips (*Thrips tabaci* Lindeman) in Slovenia. – Ecol. Modell. 194: 244–255.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference. – Springer.

Carrascal, L. M. and Palomino, D. 2002. Factors affecting bird species richness in Selvagem and Canary Islands. – Ardeola 49: 211–221.

Carrascal, L. M. and Alonso, C. L. 2006. Habitat use under latent predation risk. A case study with wintering forest birds. – Oikos 112: 51–62.

Davis, J. D. et al. 2007. Local and landscape effects on the butterfly community in fragmented Midwest USA prairie habitats. – Landscape Ecol. 9: 1341–1354.

Ekblad, A. et al. 2005. Forest soil respiration rate and delta C-13 is regulated by recent above ground weather conditions. – Oecologia 143: 136–142.

Escabias, M. et al. 2007. Functional PLS logit regression model. – Comput. Stat. Data Anal. 51: 4891–4902.

Finsinger, W. et al. 2007. Modern pollen assemblages as climate indicators in southern Europe. – Global Ecol. Biogeogr. 16: 567–582.

Frank, I. E. and Friedman, J. H. 1993. A statistical view of some chemometrics regression tools. – Technometrics 35: 109–135.

Geladi, P. and Kowalski, B. R. 1986. Partial least-squares regression – a tutorial. – Anal. Chim. Acta. 185: 1–17.

Gordo, O. et al. 2008. Geographic variation in onset of singing among populations of two migratory birds. – Acta Oecol. 34: 50–64.

Hairston, N. G. 1989. Ecological experiments. Purpose, design and execution. – Cambridge Univ. Press.

Hoffsten, P. O. 2004. Site-occupancy in relation to flight-morphology in caddisflies. – Freshwater Biol. 49: 810–817.

Hubert, M. and Branden, K. V. 2003. Robust methods for partial least squares regression. – J. Chemometr. 17: 537–549.

James, F. C. and McCulloch, C. E. 1985. Data analysis and the design of experiments in ornithology. – Curr. Ornithol. 2: 1–63.

Johnson, J. B. and Omland, K. S. 2004. Model selection in ecology and evolution. – Trends Ecol. Evol. 19: 101–108.

Karle, I. M. et al. 2007. Biogeochemical response of an intact coastal sediment to organic matter input: a multivariate approach. – Mar. Ecol. Progr. Ser. 342: 15–25.

Larocque, I. et al. 2006. Factors influencing the distribution of chironomids in lakes distributed along a latitudinal gradient in northwestern Quebec, Canada. – Can. J. Fish. Aquat. Sci. 63: 1286–1297.

Maestre, F. T. 2004. On the importance of patch attributes, environmental factors and past human impacts as determinants of perennial plant species richness and diversity in Mediterranean semiarid steppes. – Div. Distr. 10: 21–29.

Martens, H. et al. 1986. Partial least-squares regression on design variables as an alternative to analysis of variance. – Anal. Chim. Acta. 191: 133–148.

Mevik, B. H. and Wehrens, R. 2007. The pls package: principal components and partial least squares regression in R. – J. Stat. Software 18: 1–24.

Palomino, D. and Carrascal, L. M. 2007. Habitat associations of a raptor community in a mosaic landscape of central Spain under urban development. – Landscape Urban Plan. 83: 268–274.
Potapova, M. G. et al. 2004. Quantifying species indicator values for trophic diatom indices: a comparison of approaches. – Hydrobiologia 517: 25–41.
Quinn, J. F. and Dunham, A. E. 1983. On hypothesis testing in ecology and evolution. – Am. Nat. 122: 602–617.
SAS 2001. SAS/STAT user's guide, ver. 8·01. – SAS Inst.
Schmidtlein, S. 2005. Imaging spectroscopy as a tool for mapping Ellenberg indicator values. – J. Appl. Ecol. 42: 966–974.
Seppa, H. et al. 2004. A modern pollen climate calibration set from northern Europe: developing and testing a tool for palaeoclimatological reconstructions. – J. Biogeogr. 31: 251–267.
Sobek, S. et al. 2007. Patterns and regulation of dissolved organic carbon: an analysis of 7500 widely distributed lakes. – Limnol. Oceanogr. 52: 1208–1219.
Sonesten, L. 2003. Catchment area composition and water chemistry heavily affects mercury levels in perch (*Perca fluviatilis* L.) in circumneutral lakes. – Water Air Soil Pollut. 144: 117–139.
Spanos, T. et al. 2008. Environmetrics to evaluate marine environment quality. – Environ. Monit. Assess. 143: 215–225.
StatSoft 2001. Statistica (data analysis software system), ver. 6. <www.statsoft.com>.
Stepanauskas, R. et al. 2003. Covariance of bacterioplankton composition and environmental variables in a temperate delta system. – Aquat. Microbiol. Ecol. 31: 85–98.
ter Braak, C. J. F. and Schaffers, A. P. 2004. Co-correspondence analysis: a new ordination method to relate two community compositions. – Ecology 85: 834–846.
Tobias, R. D. 2003. An Introduction to partial least squares regression. <http://www.ats.ucla.edu/stat/sas/library/pls.pdf>.
Whittingham, M. J. et al. 2006. Why do we still use stepwise modelling in ecology and behaviour? – J. Anim. Ecol. 75: 1182–1189.
Wold, H. 1975. Soft modelling by latent variables; the nonlinear iterative partial least squares approach. – In: Gani, J. (ed.), Perspectives in probability and statistics. Papers in honour of M. S. Barlett. Academic Press, pp. 117–142.
Wold, S. et al. 2001. PLS regression: a basic tool of chemometrics. – Chemometr. Intell. Lab. 58: 109–130.
Zhang, E. et al. 2007. A chironomid-based salinity inference model from lakes on the Tibetan Plateau – J. Paleolimnol. 38: 477–491.
Zhang-yu, L. et al. 2007. Characterizing and estimating rice brown spot disease severity using stepwise regression, principal component regression and partial least-square regression. – J. Zhejiang Univ. Sci. B 8: 738–744.