Scientific Research and Essays Vol. 6(20), pp. 4369-4377, 19 September, 2011
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE10.1147
ISSN 1992-2248 ©2011 Academic Journals

Full Length Research Paper

Compared application of the new OPLS-DA statistical model versus partial least squares regression to manage large numbers of variables in an injury case-control study

Homayoun Sadeghi-Bazargani[1,2]*, Shrikant I. Bangdiwala[2,3], Kazem Mohammad[4], Hemmat Maghsoudi[5] and Reza Mohammadi[2]

[1]Nueroscience Research Center, Statistics and Epidemiology Department, Faculty of Health and Nutrition, Tabriz University of medical sciences, Tabriz Iran.
[2]PHS Department, Karolinska Institute, Stockholm, Sweden.
[3]Department of Biostatistics, University of North Carolina at Chapel Hill, USA.
[4]Epidemiology and Biostatistics Department, Tehran University of Medical Sciences, Tehran, Iran.
[5]Injury Epidemiology and Prevention Research Center, Surgery Department, Tabriz University of medical sciences, Tabriz, Iran.

Accepted 7 July, 2011

The use of modern statistical methodology to overcome the known pitfalls of classical regression models in the analysis of large numbers of highly correlated data, has increased considerably in recent years. Statisticians in the field of chemometrics and OMICS research have developed a new method called Orthogonal projections to latent structures (OPLS). In comparison with the regular partial least squares (PLS) regression, OPLS provides a simpler method with the additional advantage that the orthogonal variation can be analyzed separately. Use of the OPLS model has spread to fields other than its origin but it is not yet applied to the field of epidemiology, which is a wide field of research. In public health and clinical research, there are situations in which large numbers of correlated variables need to be modeled. The authors successfully applied OPLS-DA to model large numbers of variables in a case-control study and compared it with discriminant analysis done by partial least squares regression. **Prior to fitting the models, the dataset was split into two parts: a training set and a prediction set. Models fitted on the training dataset were later tested for validity in the prediction dataset. The OPLS-DA was compared with PLS-DA for model fitness, diagnostics and model interpretability. Both models suited the data but OPLS-DA was preferable. The authors encourage the use of these methods to increase study power and statistical validity in epidemiology and similar settings in which large numbers of correlated variables need to be modeled.**

Key words: Partial least squares regression, orthogonal projections to latent structures, logistic regression, multicollinearity, injury epidemiology, burns.

INTRODUCTION

The traditional regression models in classical statistics have been shown to get problematic when there are large numbers of variables and a small sample size. Multicolinearity and missing values make the situation even more complex. Multicolinearity not only increases standard errors of regression coefficients and decreases power, but also makes it difficult to separate individual effects of predictor variables, making the regression coefficients less reliable (Dohoo et al., 1997b). Thus such limitations may lead to either bias or loss of power in testing hypotheses. Methods have been developed to

*Corresponding author. E-mail: homayoun.sadeghi@gmail.com. Tel: +984113373741, +46738732855.

manage this problem. The available partial least squares (PLS) regression is a known method of analysis to statisticians in many fields. It attenuates the above-mentioned problems but PLS also suffers some limitations such as interpretability problems, multi-component results and biased coefficients in some situations leading to a higher risk of overlooking real correlations (Eriksson et al., 2006a; Richard and Cramer, 1993). Recently a newer statistical method has been introduced namely orthogonal projections to latent structures (OPLS). It is a modification of the NIPALS PLS algorithm. This method was first proposed in 2002 for chemistry studies (Trygg and Wold, 2002c). The OPLS either can be seen as a pure preprocessing method to remove systematic orthogonal variation or can become an integrated part of regular PLS modeling to provide a simpler method with the additional advantage that the orthogonal variation can be analyzed separately. Later extensions of OPLS gave rise to OPLS-DA in 2005 thus making it appropriate for use for discriminant analysis along with prediction purposes (Bylesjo et al., 2006a).

In injury research, we may encounter large numbers of correlated variables, which usually have insufficient sample size to study predictors of interest. These can be defined through a wide range of variables categorized as environmental, human related and object related variables. Supervised statistical modeling techniques like PLS and OPLS can be a good option to increase power as well as statistical validity in managing multivariate situations in injury epidemiology and possibly other similar fields of epidemiology. PLS has been used by only a few researchers in injury risk assessment until quite recently (Cadieux et al., 2006; Eriksson et al., 2009b; Sowa et al., 2006b) while application of the new OPLS-DA modeling technique is novel in the field of injury and public health epidemiology. In this article, we explain the PLS-DA and OPLS-DA statistical techniques and discuss the possibility of their application in injury risk assessment, combined with or as an alternative to classical statistical methods.

METHODS

Source data

The data used to apply and compare the statistical models is taken from an ongoing case-control study on determinants of unintentional burn injuries. Cases were enrolled from a regional burn center, which is the referral burn center in north-west of Iran. Controls were enrolled from a referral pediatric university hospital. Control selection was done in a way to ensure common source population for cases and controls (Wacholder et al., 1992). Considering age related variety in injury patterns, only children under the age of 14 were enrolled.The original dataset included 396 observations. The case-control study was approved by the responsible regional committee of ethics.

Modeling process and model diagnostics

Data were collected from a project database designed in Microsoft Access format and imported into a SIMCA P+ version 12 statistical software packages, it is an appropriate software package for supervised modeling techniques (Umetrics AB, SE-90719, Umea, Sweden). OPLS is implemented in SIMCA P+ such that the method is available under the standard PCA and PLS modeling framework. It was split into two parts before starting the modeling procedures. The training dataset data contained 316 observations and the prediction set contained 80 observations. All variables, including combined variables along with their constituting variables, were entered into the model. Categorical variables after being changed into dummy variables were entered into the model like other dichotomous and continuous measures. The modeling process is presented in Figure 1.

Prior to fitting supervised models a preliminary principal component analysis was done for data overview, detecting outliers and groups among the observations. Model goodness of fit was assessed using R^2; however as this is an inflationary measure and rapidly approached unity as the model complexity increases, Q^2 was used to assess model predictability. To provide a measure of statistical significance for the predictive power in cross-validation, response permutation was used. In this process, the X-data are left intact while Y-data are permuted to appear in a different order. The model is then fitted to the permuted Y-data and by using cross-validationR^2Y and Q^2Y are computed for the derived model. Leverage was assessed using Hottelings $T^{2.}$.Variable influence was assessed using VIP (variable importance in the projection) measures.Residual distribution graph and observed vs. Predicted graphs were also plotted both through analysis. Model significance testing was done using CV-ANOVA methodology (Eriksson et al., 2008).

PLS and PLS-DA

This method was first presented by Wold in 1975 to model in complicated datasets in terms of chains of matrices and was later modified by other researchers (Wold et al., 2001). The method is well described in the literature so we have cited only a few references and will focus on some core facts regarding this method which are essential for later understanding of O-PLS (Richard and Cramer, 1993; Wold et al., 2001). In common with principal component analysis(PCA) and to a lesser amount factor analysis, PLS also looks into the internal relationships in the matrix of variables and those cases combining the characteristics of single variables in new definitions of factors or components; but in contrast to them a main objective in PLS is to predict outcome related variables from possible predictors. This is done by linking X and Y matrices. This characteristic of the PLS method which is also present in O-PLS makes them both more effective than PCA and FA, for which reason they are referred to as supervised methods. In other words, in supervised methods variables are projected into new coordinating systems similar to PCA but their aim is to maximize the covariance between outcome and predictor variables instead of trying to explain as much variance inside the matrix as possible. If we consider one matrix of possible predictor variables (model X) and one matrix of outcome variables (model Y), PLS tries to model X and Y, and at the same time to predict Y from X (Eriksson et al., 2006c).

In this study the authors have tried to demonstrate this advantage using a case-control two-class scenario in which the first class is observations regarding burns cases and the second class is observations of non-burned control subjects. Class sizes were equal in this study.

OPLS and OPLS-DA

Orthogonal Projections to Latent Structures (OPLS) are a linear

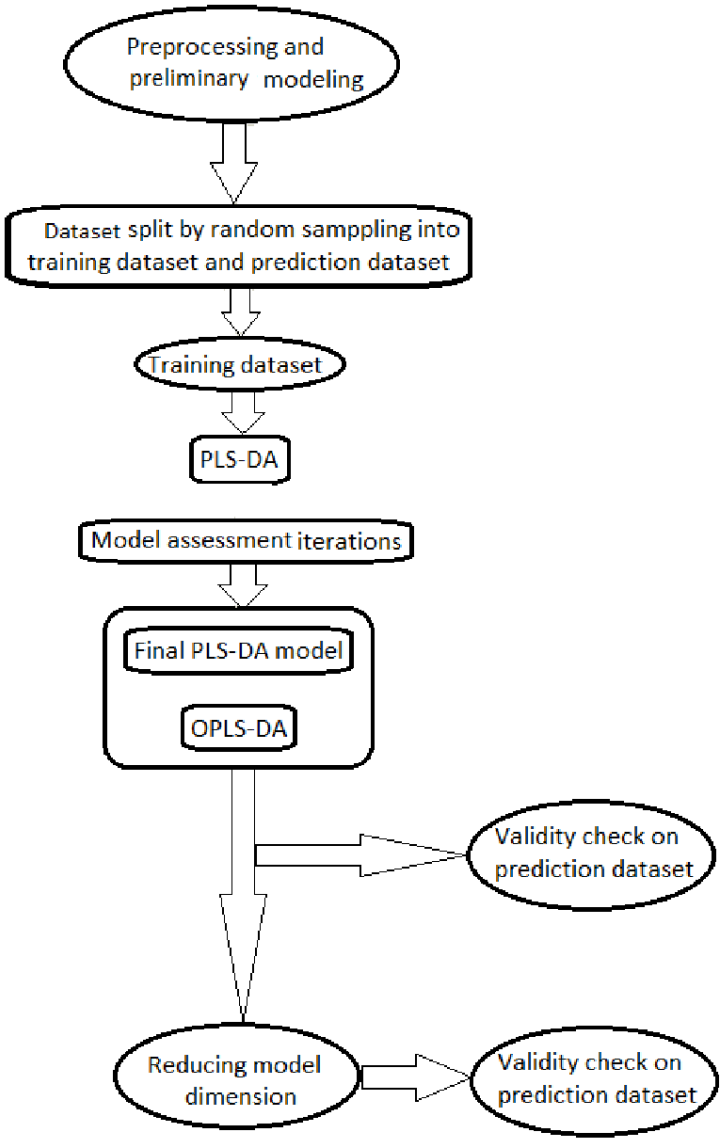

Figure 1. Algorithmic presentation of the modeling process.

regression method that has been employed successfully for prediction modeling in various biological and biochemical applications. OPLS is a modification of the usual PLS method which filters out variation that is not directly related to the response. The result is more transparent models which are easier to interpret (Trygg and Wold, 2002b).The OPLS method was first proposed in 2002 and is a modification of the original NIPALS PLS algorithm. OPLS-DA was later discussed in 2006 and considering the advantages of Orthogonal PLS over simple PLS, applying an OPLS discriminant analysis will keep the known advantages of OPLS modeling in field of discriminant analysis (Bylesj et al., 2006).

RESULTS

Descriptive

Mean age of the participants was 4.6 (SD = 3.5) years. Of all 396 participants 236 (59.6%) were males. Home was the main place of injuries in 84% of the cases. Scalds and flame burns were the major types of burn injuries. Mean total body surface area burnt was 11.6%.

Training-set models

Both the PLS-DA and OPLS-DA models were fitted into a training dataset, which contained 247±2 variables and 316 observations measured on two groups. The least possible count of components in the PLS-DA model were three components. Only one predictive component in the OPLS-DA model was captured excluding two orthogonal components in X. Residual distribution in both models was generally normal but the first component in the PLS

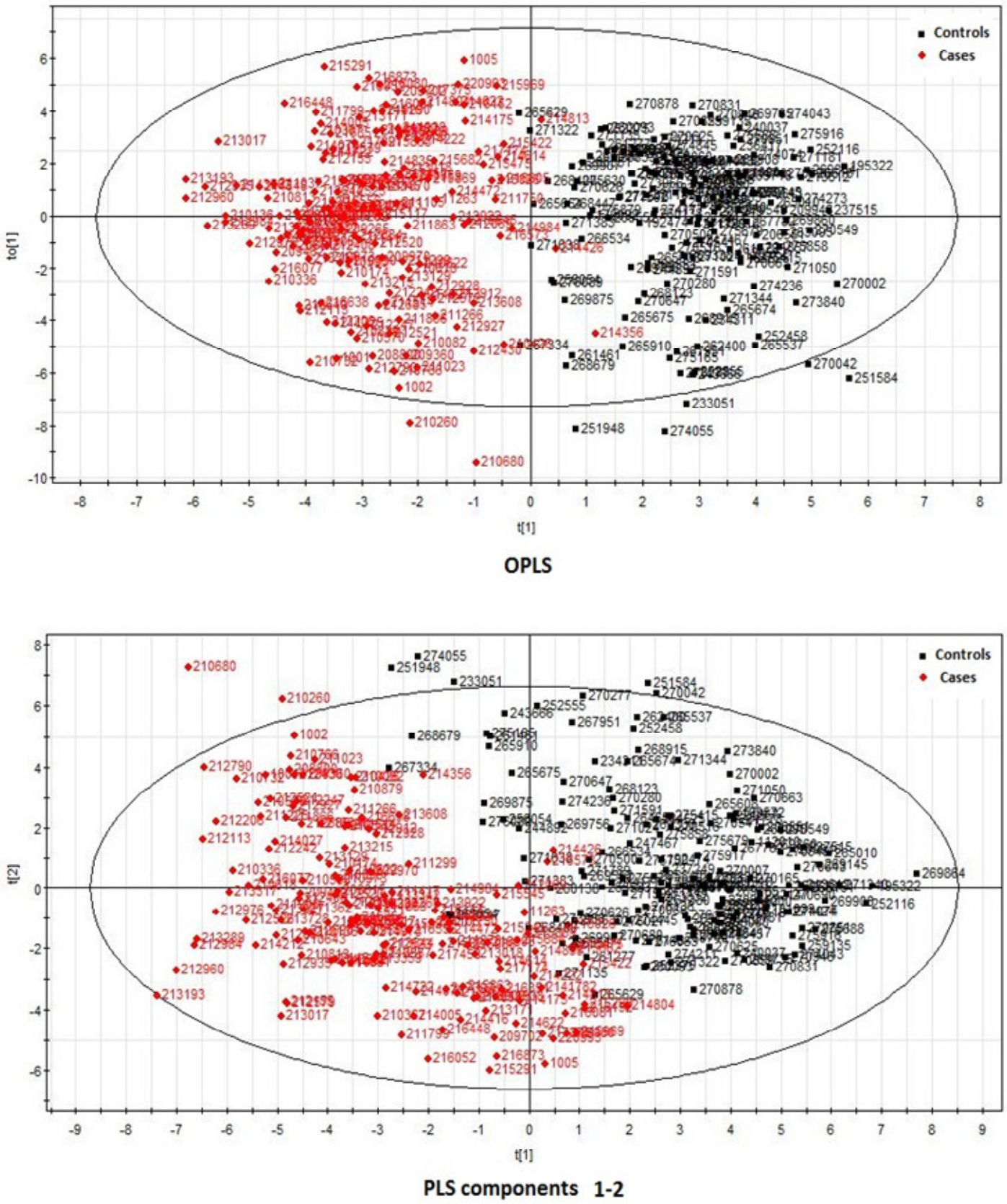

Figure 2. Score plots of the PLS and OPLS models fitted to training dataset to discriminate burned cases from controls. Red color: Burned cases; Black color: Controls.

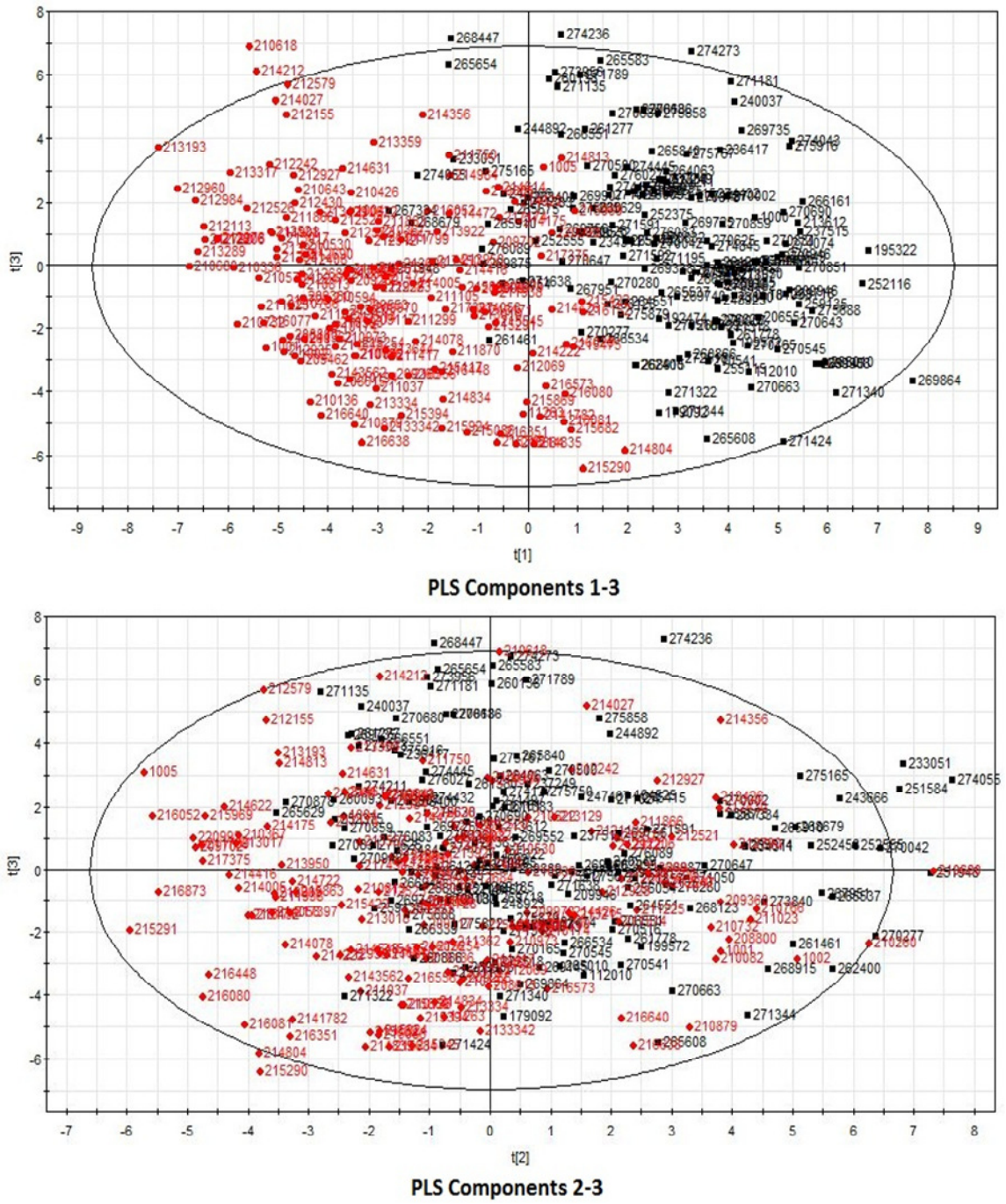

Figure 2. Contd.

model deviated slightly more from normality than the others. The primary scatter plot of the scores revealed good discrimination of the cases from controls both in the PLS-DA and OPLS-DA models. However, few outliers were present which could be excluded. This plot shows how the modeled observations in X space are situated with respect to each other. The observations are colored according to their class as cases and controls. In this type of graph, observations that lie close to each other are more similar than observations that lie relatively distant from each other. The orthogonal components have a subscript o, e.g. to1 for the first orthogonal component in X. As shown in the score plots of the two OPLS and PLS models fitted into training data, obtaining one predictive component in OPLS as the only component related to Y, makes the interpretation of the model easier than the PLS model (Figure 2). The rotation effect in OPLS is also evident from the graphs. Figure 3

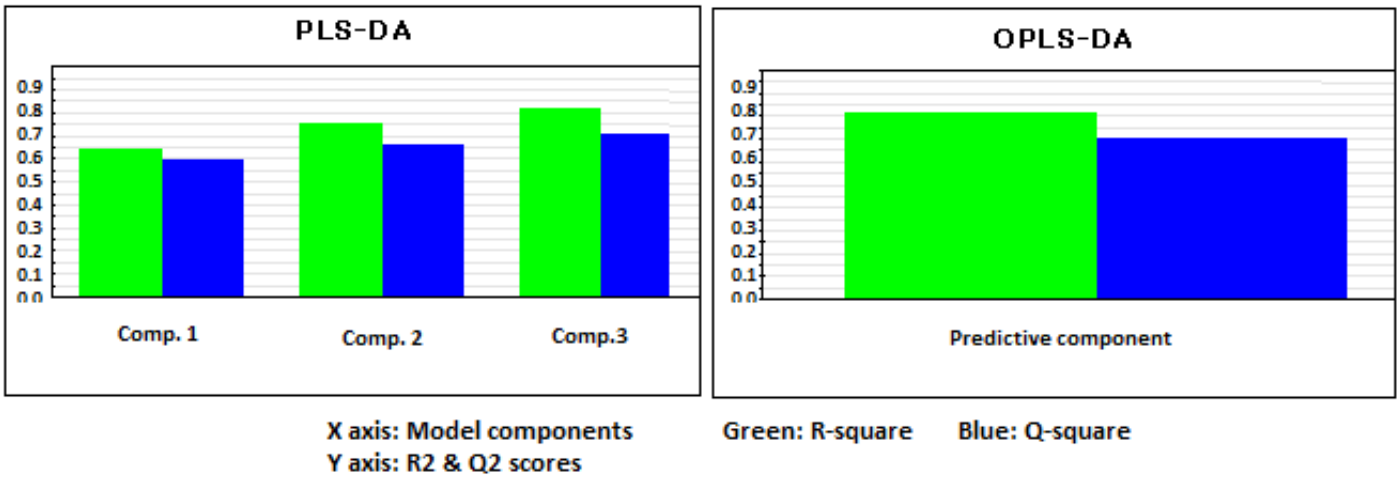

Figure 3. Model overview plots compared for PLS-DA and OPLS-DA.

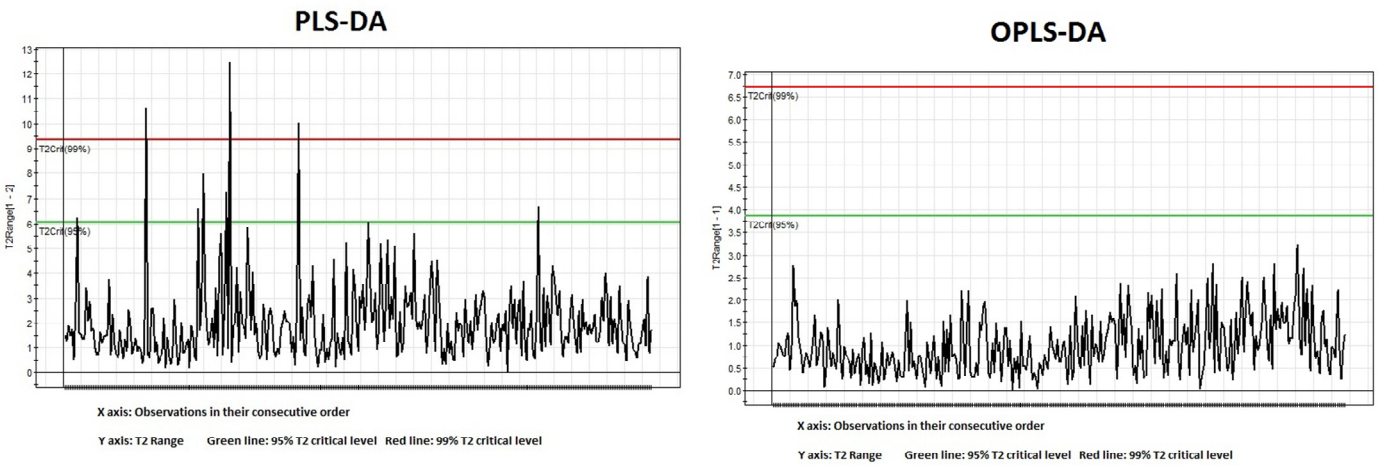

Figure 4. Hotelling's T2Range plot compared between PLS-DA & OPLS-DA.

compares goodness of fit for the components in models.The R2Y cumulated for all three components of PLS as well as R2Y for OPLS model were above 0.8 Q2Y for components in the PLS and OPLS models were both above 0.7.Cross-validation of the two components in PLS-DA with 20 permutations showed a much more reliable Q^2 than R^2. Both models had an acceptable situation assessing Hotelling's T^2 range plot but like other diagnostics, OPLS-DA had a better distribution regarding Hotelling's T^2 measure (Figure 4).

To discriminate burns cases from controls, it was found that up to 87 variables had regression coefficients significantly different from zero at 95% confidence level for the OPLS model. 107 variables for the first PLS component , 74 variables for the second PLS component and 65 variables for the third PLS component were found to be statistically different from zero. This was the main area where OPLS-DA presented its higher interpretability and reliability. PLS-DA results showed substantial variability in types of significant predictors and magnitude of the coefficients. This makes it a bit difficult in PLS modeling to draw more reliable conclusions than OPLS-DA.

Our aim in this article was not to discuss the predictors of burn injuries, something which the authors are planning to do in a separate paper after the full dataset is available. Nevertheless we would like to present here some primarily identified predictors in the OPLS-DA model which are nonspecifically presented and are as follows: using different types of cooking-heating appliance predictors; safety knowledge; risky behaviors; house structure; some other housing characteristics; economic status; caregiver and child psychological measurements; previous accident history; and age. One

more finding regarding the advantages of these modeling methods was that it was possible to model ADHD (attention deficit hyperactivity disorder) scale measurements, simultaneously entering the total score with its single questions and sub-domain scores. This is usually not possible in classical regression methods due to high correlation inside scale measures. The total ADHD score had borderline non-significant effect on the likelihood of burn injuries, but those questions that measured hyperactivity as part of the disorder were highly significant in this regard.

Prediction-set model validation

After finalizing the models on the training dataset, acquired model parameters were applied on the prediction dataset to assess model validity on new data. The prediction dataset included nearly 20% of the original dataset (80 out of 396 observations). Good discrimination was acquired after applying the model parameters to the prediction set data. The misclassification proportion was calculated to be less than 4% for both models (Fisher's prob. < 0.001). To assess the effect of reducing model parameters on prediction power, the authors selected only 40 variables based on coefficient magnitude which were taken equally from both effect directions (negative and positive). Dummy based variables which couldn't be excluded were also kept in the model. After fitting new PLS-DA and OPLS-DA models in the training datasets, the total R2Y measures did not decrease lower than 0.7 and Q^2 measures didn't decrease lower than 0.64 in any of the models. Two components were captured in the new reduced PLS-DA model. Model diagnostics despite not being as good as the original models were also acceptable. The new model parameters were also applied to the prediction dataset showing an increased misclassification proportion up to 15% on average for the models (Fisher's prob.< 0.001). So the authors found that an 80% decrease in the number of parameters only leads to a 10 % increased misclassification.

DISCUSSION

Like other fields of injury risk assessment, in studies concerning burns, methodological difficulties and properties of the data must be considered when choosing statistical methods. Just as in other case-control studies, logistic regression has often been a popular method to analyze injury case-control studies. Using maximum likelihood estimation methods, logistic regression has some advantages over linear regression analysis; but like other classical regression methods, it suffers some limitations like meeting independence of X-variables, exactness of X-variables and random distribution of errors. Power and missing data are two other areas of concern for classical regression models (Eriksson et al., 2006b). The limitations of this method when applied in studies involving large numbers of correlated variables, necessitates use of newer alternatives or complementary methods.

There is a high variety of possible predictors of burn injuries. These predictors can be categorized as human related predictors, environmental predictors and object (mainly heating-cooking appliance) related predictors (Walender et al., 2000). This usually leads to large numbers of variables being studied in injury research. The number of predictors will also need to be expanded while creating dummy variables and interactions. In the present study, a total of 247 variables needed to be modeled. A similar situation is expected in other fields of injury epidemiology. For example, regarding fall injuries, as many as 400 variables have been suggested to be of interest (Masud and Morris, 2001). The first problem arising from large numbers of variables using classic regression analysis methods, especially the logistic regression, is the power requisite. This gets more sophisticated if the number of variables exceeds the number of observations. Although estimation method options (conditional vs. unconditional maximum likelihood) in logistic regression are of help, the correlation among predictor variables that gets more likely with large numbers of variables is also something that must be handled properly. Multicollinearity is a problem when large numbers of variables are of possible interest with respect to the dependent variable (Dohoo et al. 1997a). The existence of multicollinearity inflates the variances of the parameter estimates. That may result, particularly for small and moderate sample sizes, in lack of statistical significance of individual independent variables while the overall model may be strongly significant. Multicollinearity may also result in wrong signs and magnitudes of regression coefficient estimates, and consequently in incorrect conclusions about the relationships between independent and dependent variables.

Sometimes even with very parsimonious models with few independent variables in the model, the existence of high association between two variables can be a problem. For example, if using a traditional heater is highly associated with using a single-burner gas stove and both are associated with getting burned, the statistician may be obliged to keep only one of the variables as a predictor in the model and miss the advantage of keeping the other one. Statistical methods based on latent variables may be a solution. A well-known method is PLS regression, which produces a set of the original predictor variables, a latent variable that are correlated with each other and predict the outcome variable (Eriksson et al., 2006b). So the correlation of predictor variables which may seem to be bothersome in ordinary regression methods, becomes a useful source of information about groups of variables (Henningsson et

al., 2001; Richard and Cramer, 1993). PLS has been used in injury research but not for risk assessment purposes in all cases. Cadieux used PLS to investigate validation of an instrument designed to conduct an Occupational Health and Safety (OHS) self-diagnosis, using workers' observations of tangible facts and actions in the workplace (Cadieux et al., 2006). PLS has also been used in injury risk assessment (Eriksson et al., 2009a; Sowa et al., 2006a), however, it has not turned into a popular method in injury epidemiology while older latent variable based methods have been used more than PLS. This may be partly due to the fact that it is not widely known by injury epidemiologists or due to its limitations. However PLS have many advantages over classical regression methods and also over traditional component- based methods. Independence of X-variables is not a must for PLS, it doesn't suffer from multicollinearity and limitations in the number of variables, it can cope with noise in both X and Y and moderate amounts of missing data in both X and Y. Also in contrast to classical regression methods, in PLS modeling the statistician is neither obliged to exclude one category of dummy variables nor to include a computed variable along with the original variables. For example one can include husband's income and wife's income along with total couple income in the same model or as in present study, one may prefer to include the raw total ADHD score, adjusted total ADHD score, hyperactivity subscale, attention deficit subscale and all the single questions assessing ADHD in the same model. This cannot be done in classical regression methods because it is considered to be a source of multicollinearity. However, there seem to be some main concerns regarding the use of the PLS model:

1. Interpretability problems due to the number of components in the model
2. Arbitrary methodology to define number of components
3. Difficulties due to the necessity of preprocessing and the effect of influential values
4. Many readers are not familiar with the concept of latent variable modeling
5. Insufficient handling of orthogonal variation in data

The first and the last of these concerns, especially in OMIC's research, carry higher importance and have drawn attention in recent research to find amendments (Bylesjo et al., 2006b; Goicoeches and Olivieri, 2001; Trygg and Wold, 2002a; Trygg and Wold, 2003; Wold et al., 1998). Wold et al. were the first to introduce orthogonal signal correction (OSC) to remove systematic variation from the matrix X that is unrelated (orthogonal) to the response matrix Y (Wold et al., 1998). OSC results in improved model interpretability but the main problem with the OSC method concerns the risk of over fitting the estimated OSC components and ensuring that they do not significantly alter the predictive power (Eriksson et al., 2006a). OPLS, which was developed later, was an improved model in this regard (Bylesjo et al., 2006c; Trygg and Wold, 2002a). We have described OPLS and OPLS-DA earlier, but to compare PLS and OPLS procedures it must be stated that, some of the difficulties in interpretation of regular PLS model parameters are due to the fact that PLS components are usually not the principal components of the current X matrix. Difficulty in interpreting PLS models increases in proportion to the amount of Y-orthogonal variation present in X. This problem is resolved in OPLS by removing the Y-orthogonal variation in X. In our study three PLS components were captured while one predictive component explained the variations in the OPLS model and the structured noise in the X matrix, which was nearly 10%. As the regression vector in OPLS is equivalent to the first principal component of the existing X-matrix, OPLS is expected to yield both good predictions and good interpretability. This can be more noticeable for data with more structured noise leading to a higher number of PLS components. Compared to some OMICS research or chemometrics data, 10 % is not considered a substantially large amount of structured noise and only three components were captured in our PLS model, however even this amount of interpretability improvement can lead to a preference for choosing the OPLS model. Other than a preference for OPLS over PLS in this study, the overall advantages of supervised modeling (including PLS and OPLS) compared to principal component analysis in latent variable based methods as well as the advantages of these models over classical regression methods, recommends the OPLS model as a suitable alternative method for analysis in injury epidemiology. This is the first time that OPLS models have been used in injury epidemiology and investigation into the further applications of these models in burn, fall and traffic injury risk assessment is strongly recommended. Considering the characteristics of these methods and the general popularity of classical regression methods, OPLS and PLS modeling may also be used as data reduction tools in the variable selection phase of logistic and linear regression methods. PLS based variable selection is preferable to some other selection methods like Lasso or stepwise regression particularly when the error variance is large in dataset or when the model fitness is relatively low. Also, PLS method seems to be insensitive to noise while the others seem to be sensitive.

Conclusion

Both the PLS-DA and OPLS-DA models were successfully fitted and yielded good diagnostics. The OPLS-DA proved to be preferable to the PLS-DA model in that it had better interpretability than PLS-DA. The authors encourage injury epidemiologists and statisticians to use and assess the applicability of these

methods in the analysis of injury data, in order to increase study power and statistical validity. Other epidemiological studies engaged with large numbers of correlated variables may also benefit from these methods.However, one major concern with these methods preventing them to help in drawing stronger conclusions is that they are quite new and not well tested in different settings of epidemiology.

REFERENCES

Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J (2006). OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification, J. Chemometr., 20(8-10): 341-351.

Cadieux J, Roy M, Desmarais L (2006). A preliminary validation of a new measure of occupational health and safety, J Safety Res., 37(4): 413-419.

Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D (1997a). An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies, Prev.Vet. Med., 29(3): 221-239.

Eriksson L, Johansson E, Wold N, Trygg J, Wikstrom C, Wold S (2006a). Multi- and Megavariate data analysis: Advanced applications and method extensions, 1st edn, Umetrics AB, Umea.

Eriksson L, Johansson E, Wold N, Trygg J, Wikstrom C, Wold S (2006c). PLS," in Multi- and Megavariate data analysis: Advanced applications and method extensions, Umetrics AB, Umea, pp. 63-101.

Eriksson L, Johansson E, Wold N, Trygg J, Wikstrom C, Wold S (2006b). Multi- and Megavariate data analysis: Basic principals and application Umetrics AB, pp. 7-19: 39-101.

Eriksson L, Trygg J, Wold S (2008). CV-ANOVA for significance testing of PLS and OPLS-« models, J. Chemometr., 22(11-12): 594-600.

Eriksson S, Lundquist A, Gustafson Y, Lundin-Olsson L (2009a), Comparison of three statistical methods for analysis of fall predictors in people with dementia: negative binomial regression (NBR), regression tree (RT), and partial least squares regression (PLSR), Arch. Gerontol. Geriatr., 49(3) : 383-389.

Goicoeches H, Olivieri A (2001). "A comparison of orthogonal signal correction and net analyte preprocessing methods, Chemometrics Intell. Lab. Syst., 56: 73-81.

Henningsson M, Sundbom E, Armelius BA, Erdberg P (2001). PLS model building: a multivariate approach to personality test data, Scand. J. Psychol., 42(5): 399-409.

Masud T, Morris RO (2001), Epidemiology of falls, Age Ageing, l. 30 Suppl., 4: 3-7.

Richard D, Cramer I (1993). "Partial least squares (PLS): Its strengths and limitations", Perspectives in drug discovery and design, 1: 269-278.

Sowa MG, Leonardi L, Payette JR, Cross KM, Gomez M, Fish JS (2006a). Classification of burn injuries using near-infrared spectroscopy, J. Biomed. Opt., 11(5): 054002.

Trygg J, Wold S (2002a). Orthogonal projections to latent structures. Journal of Chemometrics 16: 119-128.

Trygg J, Wold S (2003). O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter, J. Chemometr., 17(1): 53-64.

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992). Selection of controls in case-control studies. I. Principles, Am. J. Epidemiol., 135(9): 1019-1028.

Walender G, Svanstrom L, Ekman R (2000). Safety promotion and introduction Kristianstads boktryckeri AB.

Wold S, Antti H, Lindgren F, Ohman J (1998). Orthogonal signal correction of near-infrared spectra., Chemometrics Intell.Lab.Syst, 44: 175-185.

Wold S, Sjorn M, Erikson L (2001). PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab, Syst., 58: 109-130.