Made available by Hasselt University Library in https://documentserver.uhasselt.be

Treatment of missing values for multivariate statistical analysis of gel-based proteomics data

Peer-reviewed author version

Pedreschi, Romina; Hertog, Maarten L. A. T. M.; Carpentier, Sebastien C.; Lammertyn, Jeroen; ROBBEN, Johan; NOBEN, Jean-Paul; Panis, B.; Swennen, R. & Nicolai, B.M. (2008) Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. In: PROTEOMICS, 8(7). p. 1371-1383.

DOI: 10.1002/pmic.200700975
Handle: http://hdl.handle.net/1942/8262

Treatment of missing values for multivariate statistical analysis of gel-based proteomics data

Romina Pedreschi[1*], Maarten L.A.T.M. Hertog[1], Sebastien C. Carpentier[2], Jeroen Lammertyn[1], Johan Robben[3], Jean-Paul Noben[3], Bart Panis[2], Rony Swenen[2] and Bart M. Nicolaï[1]

[1]BIOSYST-MeBioS Division, Katholieke Universiteit Leuven, Belgium

[2]Division of Crop Biotechnics, Katholieke Universiteit Leuven, Belgium

[3]Biomedical Research Institute, Hasselt University and School of Life Sciences, Transnationale Universiteit Limburg, Diepenbeek, Belgium

*CORRESPONDING AUTHOR: MS Romina Pedreschi, BIOSYST-MeBioS Division, Katholieke Universiteit Leuven. Willem de Croylaan 42, Leuven (Heverlee) B-3001. Belgium.

E-mail: Romina.PedreschiPlasencia@biw.kuleuven.be

Phone: + 32 16 32 23 76, Fax: +32 16 32 29 55

Abbreviations: PCA, principal component analysis; PLS-DA, partial least square discriminant analysis; BPCA, Bayesian principal component analysis; KNN, k-nearest neighbor; asinh, inverse hyperbolic sine; IS, internal standard;. VIP, variable importance plot; EM, expectation-maximization; MI, Multiple imputation

Key words: missing value, statistics, DIGE, post run staining, preprocessing

Abstract

The presence of missing values in gel-based proteomics data represents a real challenge if an objective statistical analysis is pursued. Different methods to handle missing values were evaluated and their influence is discussed on the selection of important proteins through multivariate techniques. The evaluated methods consisted of directly dealing with them during the multivariate analysis with the NIPALS algorithm or imputing them by using either k-nearest neighbor or Bayesian principal component analysis before carrying out the multivariate analysis. These techniques were applied to data obtained from gels stained with classical post running dyes and from DIGE gels. Before applying the multivariate techniques, the normality and homoscedasticity assumptions on which parametric tests are based on were tested in order to perform a sound statistical analysis. From the three tested methods to handle missing values in our datasets, BPCA imputation of missing values showed to be the most consistent method.

1. Introduction

Two dimensional electrophoresis (2-DE) requires proper data analysis techniques to avoid misleading conclusions. The use of post run protein stains for quantitative analysis is currently being questioned due to its limited power in terms of dynamic range, sensitivity and variability [1]. The improved power of the DIGE approach arises from the use of an internal standard [2] which is used to calculate a standardized abundance of each spot and to match the spots across the gels. The classical post run dyes are however still useful as long as the technical variance is kept low and the number of replicates is high enough.

The use of appropriate statistical tools to interpret the data is a must, either with classical dyes or with DIGE. The simplest statistical analysis commonly involves pairwise comparison using parametric or non parametric tests while more complicated statistical analysis involves the use of multivariate statistics and multiple comparison tests [3-5]. Before applying a statistical test, its assumptions need to be fulfilled and some data pre-processing might be required depending on the experimental data. If a parametric test is used, for every protein, normality and homoscedasticity should be tested. For a small number of replicates (3-6 in most proteomics studies), the Shapiro-Wilk test is the most reliable test for non-normality [6].

It has been shown that low intensity spots exhibit a smaller variance between replicate gels as compared to high intensity spots [7]. As the data should be homoscedastic (show equal variances), some form of data transformation (log, asinh, square root) is required [8]. Another important issue is that samples should be independent to prevent false positives [9].

Proteomics data always contain missing values; being a spot detected in the reference or master gel but not in the sample gel. The main causes for the occurrence of missing values are (i) spots below a threshold or detection limit; (ii) mismatches caused by distortions in the protein pattern (iii) absent spots due to bad transfer from the first to the second dimension or (iv) truly absent spots from the samples. Two-dimensional data can have around 50% missing values [10-12]. However, there are no straightforward rules how to deal with missing values.

It has been demonstrated that the deletion of variables containing missing values assumes that the number of missing values is relatively small and completely at random [13]. But, in gel-based proteomics, the number of missing data is often considerable and not at random but for instance correlated to the staining procedure or the mean volume percent of the matched spots [7]. If variables with missing values, are just discarded or ignored a substantial bias can be introduced because information is simply lost. One other possibility is filling the missing values with zeroes or some lower threshold value. When a missing value is the result of a spot being below the detection limit, a threshold or zero value can be justifiable. However, whenever a value is missing due to mismatching, this would lead to wrong interpretation of the results [10, 13]. Several methods have been suggested to impute missing values such as: the row average method, *k*-nearest neighbor (KNN), singular value decomposition (SVD) impute algorithm [14-15], Bayesian Principal Component Analysis (BPCA) missing value estimation method [16] and the Maximum likelihood algorithm [12].

Multivariate statistical packages such as Unscrambler (CAMO, Trondheim, Norway), Decyder EDA (GE Healthcare, Upsula, Sweden) and SIMCA-P (Unimetrics

AB, Sweden) can deal with missing values during multivariate analysis (PCA, PLS-DA) avoiding the need to impute them. They rely on the NIPALS algorithm to set the residuals for the missing values to zero during the calculations of the principal components or latent variables. This flexibility for the user to perform the analysis when missing data is present can represent a serious problem if the amount of missing data is substantial. Morover, the amount of missing data that is considered to be substantial to distort the results is debatable.

Currently, the all against all matching approach introduced by some image analysis packages (e.g Progenesis Same Spots), theoretically generates complete datasets suitable for multivariate statistical analysis after proper data standardization. However, technical issues intrinsically associated with 2-DE and image analysis such as: gel distortions, missing spots due to bad transfer from first to second dimension, incorrect spot merging or splitting, are ignored introducing 'misleading' values that generate bias [17]. Considering all the possibilities available we believe it is crucial to be aware of the importance of how missing values are faced. Whatever approach is taken in the end, must consider the structure of the data and a compromise should be found between a sound statistical and biological interpretation of the data.

Multivariate statistics have a key role to play in 'systems biology' because much more information can be extracted than by a simple univariate test. Therefore there is an urgent need to handle missing values in an accurate way to draw realistic conclusions. When univariate statistical tests are performed (e.g t-test) it might be argued that missing values can be ignored analyzing only the available data. The reduced number of replicates due to missing values would result in a reduced power.. In both univariate or

multivariate statistical analysis, missing values represent a problem. The univariate statistical analysis in presence of missing data is out of the scope of this study. This study focuses on different techniques to handle missing values for multivariate statistical analysis and the subsequent possible impact on the interpretation of the results.

2. Materials and Methods

2.1 Proteomics data

This manuscript focuses on the statistical data analysis using proteomics datasets from pear and banana as case studies. Technical details for pear and banana proteomics can be found in respectively [3] and [18]. For this reason the experimental background of these datasets is only described in summary. The pear dataset contains data from six independent biological replicate samples for each of four treatments (different storage gas conditions). Proteins were visualized by silver staining [19]. Image analysis was performed with the Image Master 2-D Platinum software 6.0 (GE Healthcare). Spots were detected without spot editing and quantified as percentage volume.

The banana data set contains data from three replicate gels for each of four treatments (different sample dates; 2, 4, 8 and 14 days). Samples were labeled using the fluorescent Cyanine dyes developed for DIGE (GE Healthcare) according to the manufacturer's recommendations. In order to anticipate any dye specific effect, the samples were labeled at random with Cy3 and Cy5 and randomized over the gels. The internal standard was a mixture of all analyzed samples and was labeled with Cy2. Labeled proteins were visualized using a Typhoon™ imager (GE Healthcare) and the gels were analyzed using the Decyder EDA software.

The data pre-processing with DIGE occurs automatically in the DECYDER™ software: the data is normalized using a ratiometric approach and a $\log_{10}$ transformation is used on the standard abundance to stabilize the variance.

2.2 Handling of missing values

For the datasets presented in this paper, three methods to handle missing values were tested which consisted of two imputation techniques preceding the multivariate analysis (KNN and BPCA) and simply dealing with the missing values during the multivariate analysis (referred to as 'NIPALS').

k-Nearest neighbor (KNN)

The KNN method assumes a relationship between spot volume patterns of groups of proteins. The KNN method selects spots showing spot volume patterns similar to the spot of interest for which to impute missing values [15]. A weighted average of values from the *k* most similar spots is used as an estimate for the missing value under concern. The contribution of each spot is weighted by its similarity determined as the Euclidean distance. The optimum number of *k*-neighbors has to be determined empirically. The KNN imputation procedure was implemented in Matlab (The MathWorks, Inc., Natick, MA, USA) by Jörsten et al. [20] and applied in this manuscript using k=20.

Bayesian Principal Component Analysis (BPCA)

In BPCA the missing values are estimated from the known spot volumes using principal component regression (PCR). The principal components are estimated

simultaneously with the regression coefficients of the PCR model using a variational Bayes algorithm. After convergence of the algorithm missing values are imputed. The BPCA imputation procedure was implemented in Matlab (The MathWorks, Inc., Natick, MA, USA) by Oba et al. [16]. BPCA consists of three processes as described above: (i) principal component regression, (ii) Bayesian estimation and (iii) Expectation-maximization (EM) like repetitive algorithm. For a detailed explanation refer to [16].

Nonlinear Estimation by Iterative Partial Least Squares (NIPALS)

Both Unscrambler and Decyder EDA softwares are able to perform multivariate analysis in the presence of missing data using the NIPALS algorithm. In every iteration, during calculation of the principal components or latent variables, the residuals for the missing elements in the least square function are set to zero or the missing values are replaced by their minimum distance projections onto the current estimate of the loading and score vector [21]. This method is generally used in chemometrics and proteomics [22] and is tolerant to small amounts of missing data (up to 5-20 %).

2.3 Multivariate analysis

Principal Component Analysis (PCA) (unsupervised)

PCA forms new variables (principal components) that are linear combinations of the original ones thus capturing the essential data patterns of the original data in a reduced form. PCA is useful to examine datasets with multicollinearity (e.g. proteins that act in concert with other proteins) and to get insight into certain patterns or trends [23-24]. The score plots obtained show the distribution of the objects (gels) and their

185 configuration allowing the identification of outliers through the Hotelling T^2 ellipse. The
186 Loading plots obtained show the relationship between the different variables and their
187 distribution. The further a variable is from the origin, the more influential is the variable
188 for explaining relationships in the dataset. The distances along the first components are
189 more important because the first principal components explain more of the variation in
190 the dataset.
191
192 *Partial least squares (PLS) discriminant analysis (DA) (supervised)*
193 PLS is a bilinear regression model to create prediction models of one or several
194 responses from a set of factors [23]. PLS-DA will construct latent variables in such a way
195 that a maximum separation is obtained among them. PLS-DA can be useful in addition to
196 PCA to correlate variation in a dataset with class membership [24] and to select important
197 variables involved in class distinction. As in PCA, score and loading plots are obtained
198 and can be interpreted in the same way as in PCA. In addition, plots for variable
199 importance (VIP), model coefficients, residuals, distances to model plots and validation
200 plots are obtained [25].
201
202 *VIP Procedure*
203 The Variable Importance Plot (VIP) identifies those variables that are important
204 for explaining the variance in the model response [24]. The VIP coefficient of a protein is
205 calculated as a weighed sum of the squared correlations between the PLS-DA
206 components and the original variable. The weights correspond to the percentage variation
207 explained by the PLS-DA component in the model. The number of terms in the sum

208 depends on the number of PLS-DA components found to be significant in distinguishing
209 the classes. Care must be taken when excluding variables from the model. If many
210 important variables are excluded, important explanatory information may be lost as well
211 [25]. For more details about PLS and the VIP procedure one is referred to Norden et al
212 [26].
213
214 **2.4 Performance of handling missing values**
215 The performance of handling missing values was tested on a subset of the DIGE
216 dataset referred to as 'complete DIGE' dataset containing 542 proteins matched across all
217 the gels without missing values. The experimental set-up is described in Figure 1A. From
218 this 'complete DIGE' dataset thirty percent of the data was randomly removed. Using this
219 dataset with artificially induced missing values, the various methods for handling missing
220 values described above were tested. Since the underlying normality and equal variance
221 assumptions are supposed to be met with DIGE data after Decyder analysis [8],
222 transformation of the data was not required.
223 The multivariate data analysis involved PLS-DA analysis to discriminate the
224 individual gels according to similar protein expression profiles. Cross-validation was
225 applied to test the performance of the models since the number of observations is too
226 small to validate the models on an independent test set. The VIP procedure was used to
227 identify the 50 most important proteins describing the difference in protein expression
228 profiles. These selected proteins were compared between the different approaches of
229 handling missing values using the 'complete DIGE' dataset as a reference. This
230 procedure, starting from the induction of random missing values, was repeated 10 times

to evaluate its consistency. A method is considered to be 'consistent' if by repeating several times (10 in this particular case), the obtained proteins are the same as the 'real ones' (obtained when no missing values are present). PLS-DA and VIP analyses were performed using The Unscrambler Version 9.1 (CAMO A/S, Trondheim, Norway).

2.5 Impact of missing values handling techniques on VIP selection using DIGE data

To test the impact of different missing values handling techniques on the final VIP selection, the original incomplete DIGE data (covering 1462 proteins, containing missing values) was used. As the normality and equal variance assumptions were assumed to be met, transformation of the data was not required. Missing values were handled either during the multivariate analysis (NIPALS) or by imputing them on beforehand using either the KNN or BPCA method. PLS-DA and the VIP procedure were used to build models able to explain the variance in the dataset. The followed procedure is described in Figure 1B.

2.6 Impact of missing values handling techniques on VIP selection using classical dyes data

Normality was checked with the Shapiro and Wilk test. To meet the equal variance assumption, different transformations were tested: no transformation, a logarithmic (log), inverse hyperbolic sine (asinh) and square root transformation. Handling missing values during the multivariate analysis (NIPALS) was compared to imputing them on beforehand using either the KNN or BPCA method. If for a particular protein in one of the treatments all replicates presented missing values but were clearly

present in the other treatments, they were treated as threshold values. Before performing PLS-DA and the VIP procedure to select the fifty most important proteins involved in class distinction, PCA outlier detection through the Hotelling T^2 ellipse was performed.

3. Results

3.1. Matching of the data and estimation of missing values

The percentage of missing values in either the DIGE or classical dyes datasets was 24 % and 29 % respectively (Table 1). Despite the use of an internal standard and the co-detection algorithm with the DIGE, the individual gels still need to be matched resulting in substantial amounts of missing values (Table 1A). The total number of spots fully matched across all samples of the DIGE dataset was 542.

3.2. Performance of handling missing values

The 'complete DIGE' dataset (542 proteins) was used to evaluate the performance of different methods to handle missing values after random removal of 30% of the data (Figure 2). Based on the score plots, none of the methods clearly outperformed the others in terms of quality of the separation (Figure 3). The score plots are a useful visualization tool to inspect if the real variance from the 'complete dataset' is being masked or not by the tested methods to handle missing values in the derived datasets with artificially induced missing values. Particularly, since we have the 'complete dataset' a direct comparison can be made. However, looking at the proteins involved in the classification, quantitative differences are observed. Depending on how missing values were handled, in average only 34% to 63% of the selected proteins were identical to the fifty selected

proteins obtained from the 'complete DIGE' dataset (Figure 4a). The number of imputed missing values in these fifty selected proteins for all the methods tested did not differ extensively. In addition, the BPCA imputed data seems to be closer to the original data (Figure 4b) as compared to the KNN imputed data. The calculated correlation coefficients for the real data *vs* BPCA imputed data and real data *vs* KNN imputed data were 0.85 and 0.65, respectively. These coefficients clearly show that BPCA provides more accurate estimates of the missing values than KNN. The selection of proteins for the KNN also varied extensively during the ten simulations (34% ± 17%, Figure 4a). From these results, BPCA showed to be the most consistent method in terms of selecting those proteins that would have been selected if there were no missing values in the dataset.

3.3 Impact of missing values handling techniques on VIP selection using DIGE data

Depending on how missing values were handled different selections of 50 proteins were obtained for the original incomplete DIGE data (covering 1462 proteins, containing 24 % missing values). Between KNN and BPCA 30 out of the 50 selected proteins were the same. When the missing data was handled during the multivariate analysis (NIPALS), only one out of the fifty proteins was the same when compared to the BPCA method which in the previous section was shown to perform best (Figure 5a).

Most of the proteins selected based on the BPCA imputed data contained no missing values while the proteins selected when missing values were handled during the multivariate analysis (NIPALS) contained large numbers of missing values (Figure 5b). The score plots and explained variances do not differ significantly for the BPCA and KNN methods (Figure 6b and c). But when missing data was handled during the

multivariate analysis (NIPALS), the variance within each group seems to be artificially reduced (Figure 6a) which was not observed with the 'complete DIGE' dataset (Figure 3). By handling missing data during the multivariate analysis or prior application of BPCA and KNN, PLS-DA was able to explain 83%, 86% and 84% of the total variance when only the 50 most important proteins were kept although the final selection of these proteins clearly differed (Figure 5a).

3.4 Impact of missing values handling techniques on VIP selection using classical dyes data

According to the Shapiro and Wilk test, approximately 5% of the spots failed normality. Applying different transformations did not reduce this percentage but mainly stabilized the variances (data not shown). The log transformation improved homoscedasticity since the standard deviation was no longer correlated with the mean percentage spot volume. Thus, the log transformation was applied for further processing. In average the fifty selected proteins obtained by handling the missing values during the multivariate analysis (NIPALS) contained in average 8 missing values out of 24 values while after prior application of BPCA and KNN the fifty selected proteins contained only 6 missing values (Figure 7b). In addition, the score plots obtained after the treatment of missing values and the final selection of the 50 most important proteins according to the VIP procedure and amount of explained variance are shown in Figure 8.

4. Discussion

Missing values are often present in classical stained and DIGE gels and must be treated appropriately. In general, less intense spots are more susceptible to be missing; nonetheless, these proteins might represent an important class responsible for regulation and signaling [10, 12]. The introduction of more and more sensitive mass spectrometric techniques, allow the identification of this low abundant class of proteins. In addition, currently many diagnostic studies rely on data mining techniques to assign samples to a certain group, thus the low abundant fraction proteins is essential [17]. Discarding such proteins, otherwise, would result in enormous loss of valuable biological information. The BPCA method showed to be the most consistent in terms of selecting most of the proteins that would have been selected if there were no missing values in the data while KNN tended to distort the structure of the original data (Figure 4b). This was confirmed with the calculated correlations coefficients.

When evaluating the three methods to handle missing values on the original DIGE dataset (1462 variables, 24% missing values), the fifty most important proteins selected with PLS-DA by handling the missing values during multivariate analysis was completely different from the results obtained after imputation by BPCA or KNN (Figure 5a). An explanation for this is that missing values for proteomics data are not just the result of completely random events. This can be clearly seen in Figure 2 in which the distribution of missing values is plotted for the artificial dataset based on the ‘complete DIGE’ dataset and for the original incomplete DIGE dataset. By just discarding the missing dimensions, Eisen et al. [27] found cluster of genes with many missing values when carrying out a cluster analysis on gene expression profiles. This finding was caused

346 by ignoring the missing values which is similar to assume that the expression levels are
347 the same within an experimental group. Statistically spoken, it means that the distance
348 between vectors with missing values tends to be smaller than the distance without
349 missing values. When there are too many missing values present during multivariate
350 analysis the score estimation error increases as the loading vector approaches the missing
351 variable axis. Since influential variables will have large weights in the loading vector, the
352 score estimation error will increase as well. The presence of missing data in the
353 multivariate analysis thus caused a bias towards the selection of proteins containing 60%
354 missing values (Figure 5b). It has been shown that NIPALS tends to cause loss of
355 robustness as the amount of missing values increases to 20% [22] compared to other
356 algorithms such as BPCA [16] or Multiple imputation (MI) [28]. It is worth to mention
357 here that not only the total amount of missing data in the dataset (24%) is important but
358 how it is distributed among the different proteins. For instance, in the‘incomplete DIGE’
359 dataset, 27% of the total number of proteins containing missing values showed to have
360 missing values equal or higher than 50%

361 From the original datasets false positives or negatives cannot be recognized, but
362 imputation of missing values by BPCA is more appropriate than just handling them
363 during the multivariate analysis. In contrast to the BPCA method that includes maximum
364 likelihood estimation, the other two methods do not take into account the uncertainty
365 associated with the prediction of the missing values. In addition the maximum likelihood
366 algorithm does not assume the existence of missing values completely at random across
367 all the observations but only at random within one or more subgroups (e.g., missing more
368 among low abundant proteins than high abundant proteins, but within this low abundant

category they are missing at random) which is an advantage. However, the total uncertainty associated with the prediction is not included and some other features such as the dependency of missing values on the characteristics (e.g. abundance, hydrophobicity, etc) of the proteins might be disregarded.

For the classical dyes dataset, the normality and equal variance assumptions were tested before performing the statistical analysis. The use of different transformations to stabilize the variance has been described before for proteomics data [5, 7, 11, 29]. For the classical dyes dataset it was shown that applying a log transformation is only needed to stabilize the variance but not to turn the data normal as 95 % of the data was already normally distributed regardless the transformation applied. For the different ways to handle missing data in the classical dyes dataset, 60% homology in terms of the same selected 50 most important proteins remains (Figure 7a). It has been shown in a previous study with gene expression data by Bras and Menezes [30] that PLS based imputation methods performed better when the correlation structure of the data is weak (e.g non time series experiments), as this experiment. However, with all the datasets tested (time series, non-time series and mixed experiments) BPCA in most of the cases outperformed the PLS based estimation methods. The fact that the three of them yielded more or less the same results is encouraging in terms of robustness for a biological interpretation of the data, given that a choice has to be taken. Some examples of how the imputation methods are affecting the inclusion of particular proteins in the final VIP selection for the ‘pear dataset’ are given in Figure 9. All these proteins were visually inspected and confirmed as real spots. The figure shows both the imputed and original non-missing observations. In case of BPCA and KNN imputed data the VIP selection is based on the combination of

the original non-missing observations with the respective imputed values. In case of the NIPALS data set, the VIP selection is based on the original non-missing observations only. A typical protein included in all final VIP selections after each of the three methods used to deal with missing data (Figure 9A) showed imputed values similar to the original non-missing spot volumes, suggesting accurate imputations. The protein selected by the three methods showed to be involved in a physiological disorder in pears which confirms what was found in our previous study [3]. Whenever a protein was not selected after one missing values handling method but was selected by the remaining two missing values handling methods (Figure 9B-D) this was due to the fact that the imputed values were clearly different from each other and the original non-missing values. However, one needs to be careful in interpreting data of individual proteins (an implicit univariate approach) as the selected proteins were identified within their original multivariate context.

One possible argument, for the disagreement in performance of the NIPALS algorithm between this dataset and the 'incomplete DIGE dataset' might be related to the total percentage of individual proteins containing huge amounts of missing data. Even when this classical dyes dataset presents a higher total amount of missing values (29%) than the 'incomplete DIGE' dataset (24%), the classical dyes dataset only presented 13% of the total proteins containing missing values with 50% or more missing values. This feature leads to a better performance of the NIPALS algorithm for this particular dataset. It might be argued that a 'preliminary filtering' of proteins, in terms of the maximum amount of missing values allowed within each protein would be good practice but would still be subjective in where to set the maximum.

5. Conclusions

Data pre-processing steps have a large impact on the final selection of the most important proteins when using multivariate statistical tools such as PLS and VIP and heavily rely on how missing values are treated. There is no absolute truth in terms of which is the most appropriate way to deal with missing data, however, from the ones studied, BPCA gave the best result.

We recommend: (1) not to discard proteins containing missing values from the start, (2) estimate the amount of missing values in the dataset and within each individual protein, (3) based on the amount of missing values make a choice to impute missing values with an appropriate available method (we recommend BPCA in our case), (4) go back to the gels to check whether those selected proteins are real spots and not just artifacts or threshold values.

6. Acknowledgments

We would like to thank Dr. Rebecka Jörsten and Dr. Ming Ouyang (University of Rutgers, USA) and Dr. Shigeyuki Oba (Nara Institute of Science and Technology, Japan) for kindly providing us with the KNN and BPCA Matlab codes. We would like to thank Dr. Natasha Karp (University of Cambridge) for her useful comments on this paper. This research has been carried out in the framework of EU COST action 924. R. Pedreschi extends the acknowledgement to the International Relations Office of the K.U.Leuven (IRO Scholarship). Dr. S.C.Carpentier is supported by a postdoctoral fellowship of the K.U. Leuven.

438

439 **7. References**

440

441 [1] Miller, I., Crawford, J., Gianazza, E. P. Protein stains for proteomic applications:
442 which, when and why? *Proteomics* 2006, 6, 5385-5408.

443

444 [2] Tonge, R., Shaw, J., Middleton, B., Rowlinson, R. et al., Validation and development
445 of fluorescence two-dimensional differential gel electrophoresis proteomics technology.
446 *Proteomics* 2001, 1, 377-396.

447

448 [3] Pedreschi, R., Vanstreels, E., Carpentier, S.; Hertog, M. et al., Proteomic analysis of
449 core breakdown disorder in Conference pears (*Pyrus communis* L.). *Proteomics* 2007, 7,
450 2083-2089.

451

452 [4] Fuji, K., Kondo, T., Yokoo, H., Yamada, T. et al., Protein expression pattern
453 distinguishes different lymphoid neoplasms. *Proteomics* 2005, 5, 4274-4286.

454

455 [5] Tuomainen, M., Nunan, N., Lehesranta, S., Tervahauta, A. et al., Multivariate analysis
456 of protein profiles of metal hyperaccumulator *Thlaspi caerulescens* accessions.
457 *Proteomics* 2006, 6, 3696-3706.

458

459 [6] Shapiro, S., Wilk, M. An analysis of variance test for normality (complete samples).
460 *Biometrika* 1965, 52, 591-611.

461

462 [7] Grove, H., Hollung, K., Uhlen, A., Martens, H. et al., Challenges related to analysis of
463 protein spot volume from two-dimensional gel electrophoresis as revealed by replicate
464 gels. *J. Proteome Res* 2006, 5, 3399-3410.

465

466 [8] Karp, N., Lilley, K. Maximising sensitivity for detecting changes in protein
467 expression: experimental design using minimal CyDyes. *Proteomics* 2005, 5, 3105-3115.

468

469 [9] Karp, N., McCormick, P.S., Russell, M.R., Lilley, K.S. Experimental and statistical
470 considerations to avoid false conclusions in proteomic studies using differential in-gel
471 electrophoresis. *Mol. Cell. Proteomics* 2007, 6, 1354-1364.

472

473 [10] Wood, J., White, I., Cutler, P. A likelihood-based approach to defining statistical
474 significance in proteomic analysis where missing data cannot be disregarded. *Signal*
475 *Process* 2004, 84, 1777-1788.

476

477 [11] Jung, K., Gannoun, A., Sitek, B., Meyer, H. et al., Analysis of dynamic protein
478 expression data. *REVSTAT-Statist.J* 2005, 3, 99-111.

479

480 [12] Krogh, M., Fernandez, C., Teilum, M., Bengtsson, S. et al., A probabilistic treatment
481 of the missing spot problem in 2D gel electrophoresis experiments. *J. Proteome. Res*
482 2007, *6*, 3335-3343.

483

[13] Little, R.J., Rubin, D.B. Statistical analysis with missing data. John Wiley&Sons, New York, USA. 1987.

[14] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P. et al., Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001, 17, 520-525.

[15] Jung, K., Ganooun, A., Sitek, B., Apostolov, O. et al., Statistical evaluation of methods for the analysis of dynamic protein expression data from a tumor study. *REVSTAT-Statist J* 2006, 4, 67-80.

[16] Oba, Shigeyuki., Sato, Masa-aki., Takemasa, I., Monden, M. et al., A Bayesian missing values estimation method for gene expression profile data. *Bioinformatics* 2003, 19, 2088-2096.

[17] Karp, N., Feret, R., Rubtsov, D., Lilley, K. Comparison of DIGE and post-stained gel electrophoresis with both Traditional and SameSpots analysis for quantitative proteomics. Proteomics 2007, in press.

[18] Carpentier, S., Witters, E., Laukens, K., Van Onckelen, H. et al., Banana (Musa spp.) as a model to study the meristem proteome: Acclimation to osmotic stress. *Proteomics* 2007, 7, 92-105.

[19] Blum, H., Beier, H., Gross, H. Improved silver staining of plant proteins, RNA and

507 DNA in polyacrylamide gels. *Electrophoresis* 1987, 8, 93-99.
508
509 [20] Jörsten, R., Wang, H., Welsh, W., Ouyang, M. DNA microarray data imputation
510 and significance analysis of differential expression *Bioinformatics* 2005, 21, 4155-4161.
511
512 [21] Nelson, P., Taylor, P.A., MacGregor, J.F. Missing data methods in PCA and PLS:
513 score calculations with incomplete observations. *Chem Intel Lab Syst* 1996, 35, 45-65.
514
515 [22] Grung, B., Manne R. Missing values in principal component analysis.
516 *Chemom.Intel.l Lab. Syst* 1998, 42,125-139.
517
518 [23] Wold, S. Principal Component Analysis. *Chemoms. Intell. Lab Syst* 1987, 2, 37-52.
519
520 [24] Karp, N., Griffin, J., Lilley, K. Application of partial least squares discriminant
521 analysis to two-dimensional difference gel studies in expression proteomics. *Proteomics*.
522 2005, 5, 81-90.
523
524 [25] Danvind, J. PLS prediction as a tool for modeling wood properties. *Holz als Roh-und*
525 *Werstoff* 2002, 60, 130-140.
526
527 [26] Norden, B., Broberg, P., Lindberg, C., Plymoth, A. Analysis and understanding of
528 high-dimensionality data by means of multivariate data analysis. *Chem Biodivers* 2005, 2,
529 1487-1494.

[27] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 1998, 95, 14863-14869.

[28] Allison, P. Multiple imputation for missing data. *Sociological Methods & Research* 2000, 28, 301-309.

[29] Hunt, S., Thomas, M., Sebastian, L., Pedersen, S. et al., Optimal Replication and the Importance of Experimental Design for Gel-Based Quantitative Proteomics. *J. Proteome Res* 2005, 4, 809-819.

[30] Brás, L.P., Menezes, J.C. Dealing with gene expression missing data. *IEE Proc-Syst. Biol* 2006, 153, 105-119.

Table 1A. Matching results for incomplete DIGE dataset

Gel (treatments: cy3, cy5, cy2)	*Detected spots*	*% spots matched to master gel 3*
Gel 1	1601	75
Gel 2	1532	67
Gel 3	1692	100
Gel 4	1412	67
Gel 5	1548	78
Gel 6	1256	69

Table 1B. Matching results for classical dyes dataset

Treatments	*Average detected spots (n=6)*	*% average matched spots to reference gel (n=6)*
Condition 1	733	63
Condition 2	520	64
Condition 3	622	69
Condition 4	609	63

Figure Legends

Figure 1. Flow chart detailing the procedure followed for (A) testing the performance of handling missing values using the 'Complete DIGE dataset' composed of 542 totally matched proteins, (B) testing the impact of missing values handling techniques on VIP selection using the 'incomplete datasets': DIGE and classical dyes. The asterisk indicates that missing values were not imputed during preprocessing but were handled during the multivariate analysis through the NIPALS algorithm.

Figure 2. Distribution of missing values for (a) random removal in 'complete DIGE' dataset (test dataset, containing 542 proteins matched across all gels) and (b) incomplete DIGE dataset (containing 1462 proteins).

Figure 3. PLS-DA score plots for the (a) 'complete DIGE' dataset (542 proteins matched across all gels), (b) after random removal of 30% of the data and treated with, Unscrambler (NIPALS algorithm) or imputed with (c) BPCA and (d) KNN.

Figure 4. (a) Number of important proteins selected through VIP 50* after random removal of 30% of the data in the 'complete DIGE' dataset and treated with the different options to handle missing values, (b) 'complete DIGE' dataset versus imputed data with BPCA or KNN. VIP 50* is defined as the fifty most important proteins selected by PLS-DA and VIP analysis.

Figure 5. (a) Venn diagrams showing the overlap of the selected proteins through PLS-DA and VIP 50* (b) Percentage of proteins from the 50 selected as a function of the number of missing values for the incomplete DIGE dataset. VIP 50* is defined as the fifty most important proteins selected by PLS-DA and VIP analysis. The maximum number of missing values in this dataset would be 10 out of 12 because of the DIGE set up (3 dye approach).

Figure 6. Score plots (PLS) after the VIP 50* procedure for the incomplete DIGE dataset, (a) missing values handled during the calculations NIPALS (b) BPCA imputed, (c) KNN imputed. VIP 50* is defined as the fifty most important proteins selected by PLS-DA and VIP analysis.

Figure 7. (a) Venn diagrams showing the overlap of the selected proteins through PLS-DA and VIP 50*, (b) Percentage of proteins from the 50 selected as a function of the number of missing values for the incomplete classical dyes dataset. VIP 50* is defined as the fifty most important proteins selected by PLS-DA and VIP analysis. The maximum number of missing values in this dataset would be 23 out of 24 for this dataset.

Figure 8. Score plots (PLS-DA) after the VIP 50* procedure for the classical dyes dataset (563 proteins containing 29% missing values), (a) missing values ignored during the calculations (b) BPCA imputed, (c) KNN imputed. VIP 50* is defined as the fifty most important proteins selected by PLS-DA and VIP analysis.

Figure 9. Observed and imputed spot volume values for 4 selected proteins (plot A-D) from the 'classical dyes dataset'. Treatments (1-4) stand for the different storage conditions used. The open symbols represent the imputed values using either BPCA (◊) or KNN (Δ) imputation. The closed symbols (●) represent the original non-missing observations making up the NIPALS dataset. Plot A, 'None differs' shows data for a protein (439) that was included in the VIP selection for all three missing values handling methods (either imputed during preprocessing, by BPCA or KNN imputation, or handled during the multivariate analysis through the NIPALS algorithm). The other plots (B-D) show data for proteins (respectively 401, 589 and 348) that were NOT selected after the missing values handling method referred to in the heading of the plot, but were selected by the other two missing values handling methods.

Figure 1

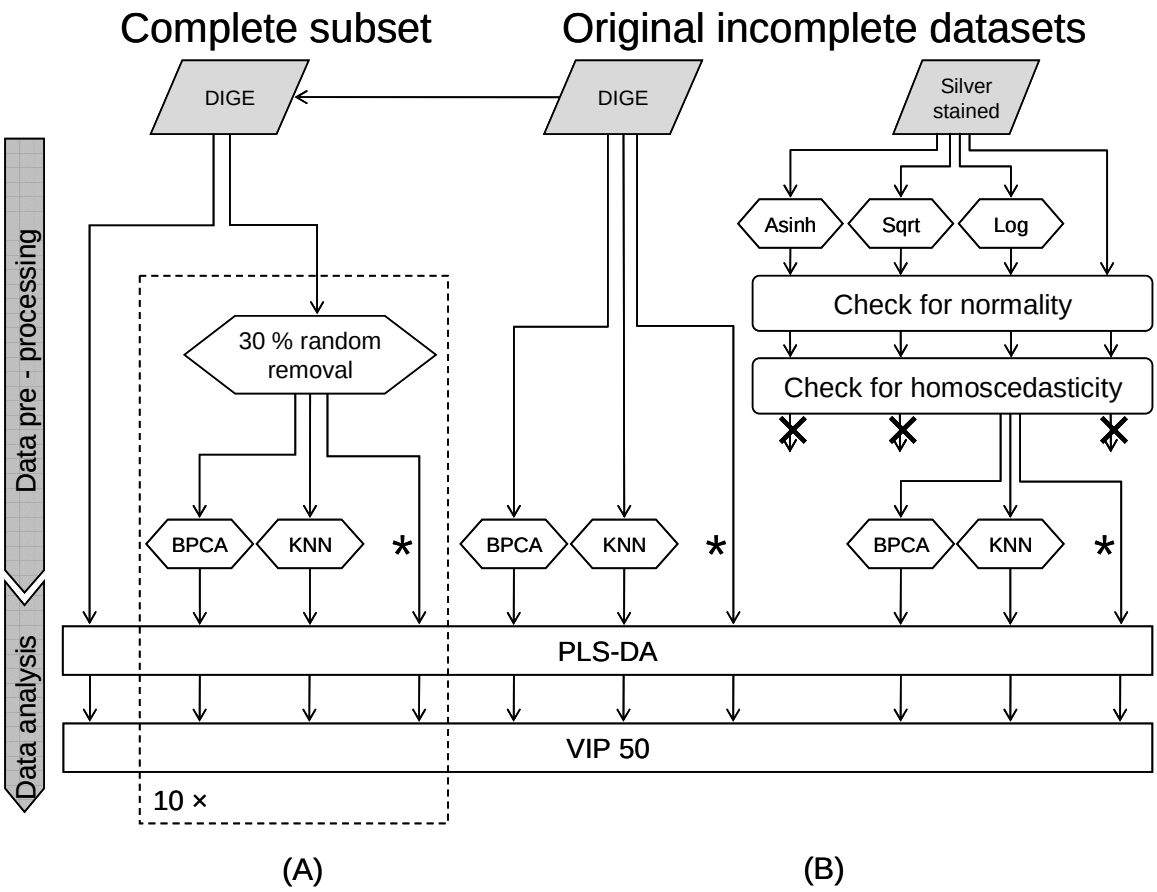

Complete subset
Original incomplete datasets
DIGE
DIGE
Silver stained
Asinh
Sqrt
Log
Check for normality
Check for homoscedasticity
30 % random removal
BPCA
KNN
*
BPCA
KNN
*
BPCA
KNN
*
PLS-DA
VIP 50
10 ×
Data pre - processing
Data analysis
(A)
(B)

Figure 2

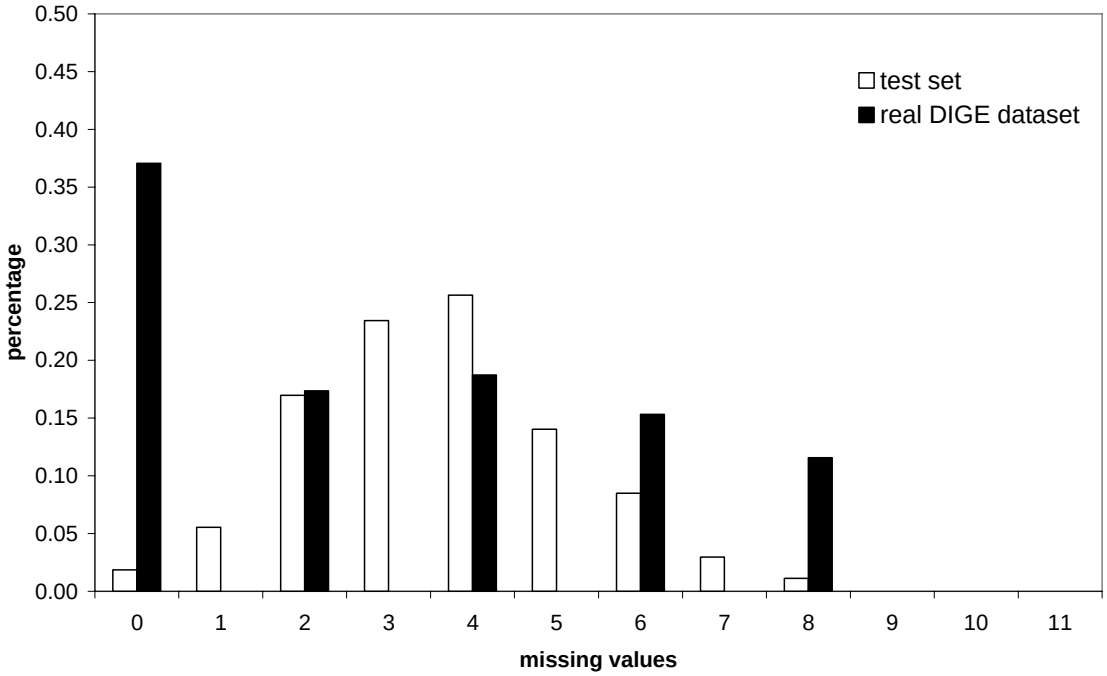

0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
percentage
test set
real DIGE dataset
0
1
2
3
4
5
6
7
8
9
10
11
missing values

Figure 3

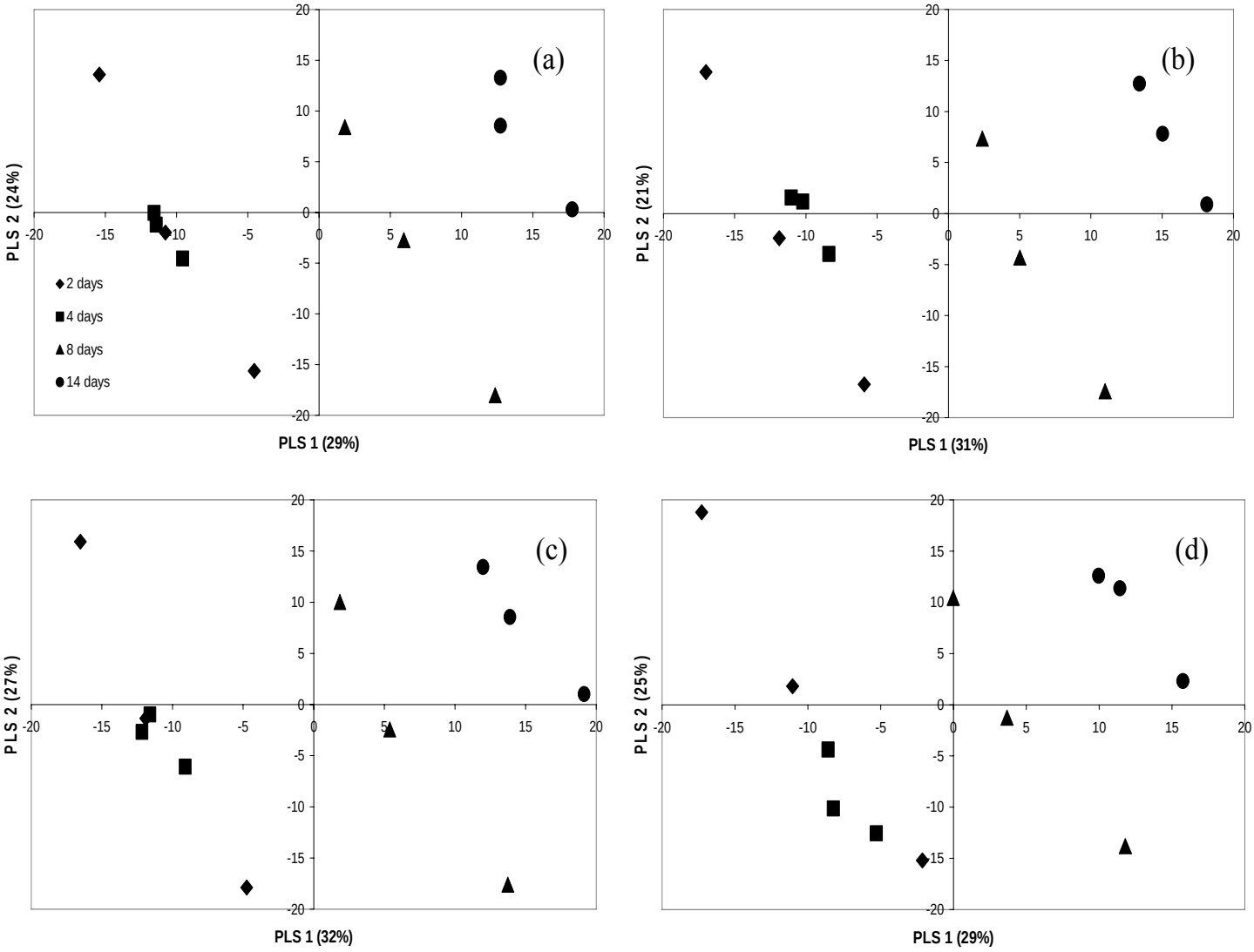

(a)
PLS 2 (24%)
PLS 1 (29%)
2 days
4 days
8 days
14 days
(b)
PLS 2 (21%)
PLS 1 (31%)
(c)
PLS 2 (27%)
PLS 1 (32%)
(d)
PLS 2 (25%)
PLS 1 (29%)

Figure 4

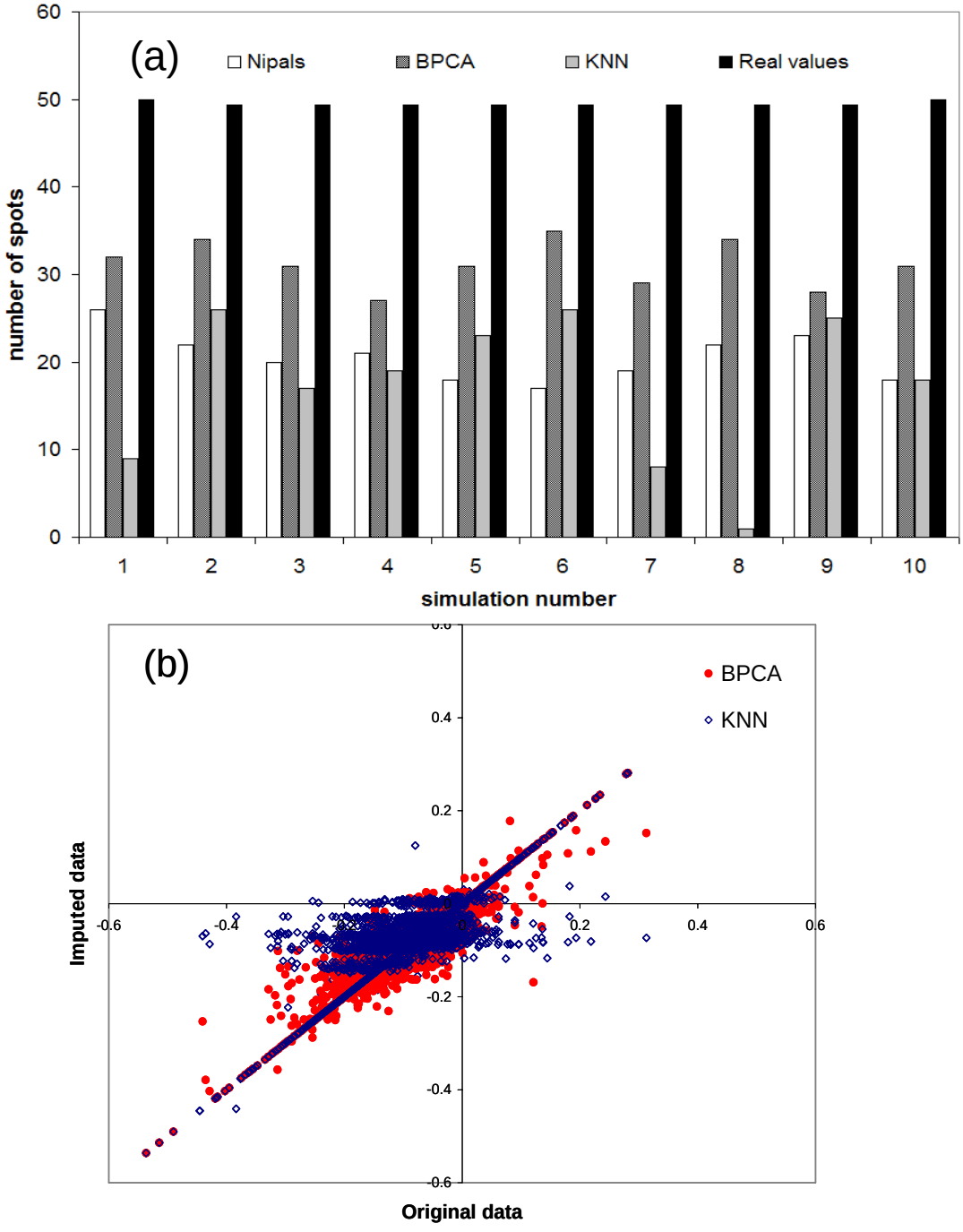
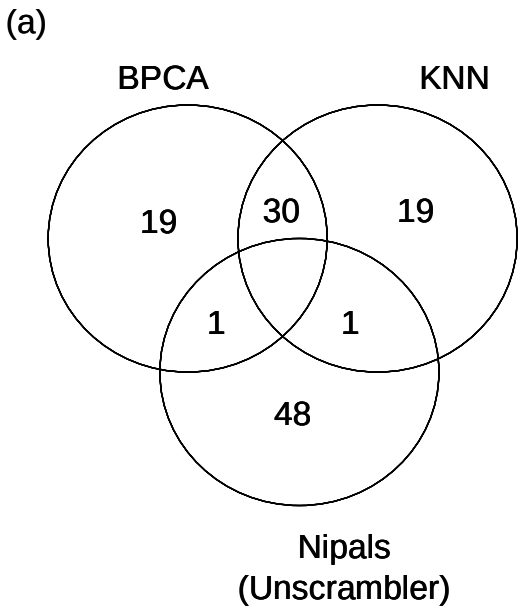

(a)
Nipals
BPCA
KNN
Real values
number of spots
simulation number
(b)
BPCA
KNN
Imputed data
Original data

Figure 5


(a)
BPCA
KNN
19
30
19
1
1
48
Nipals
(Unscrambler)

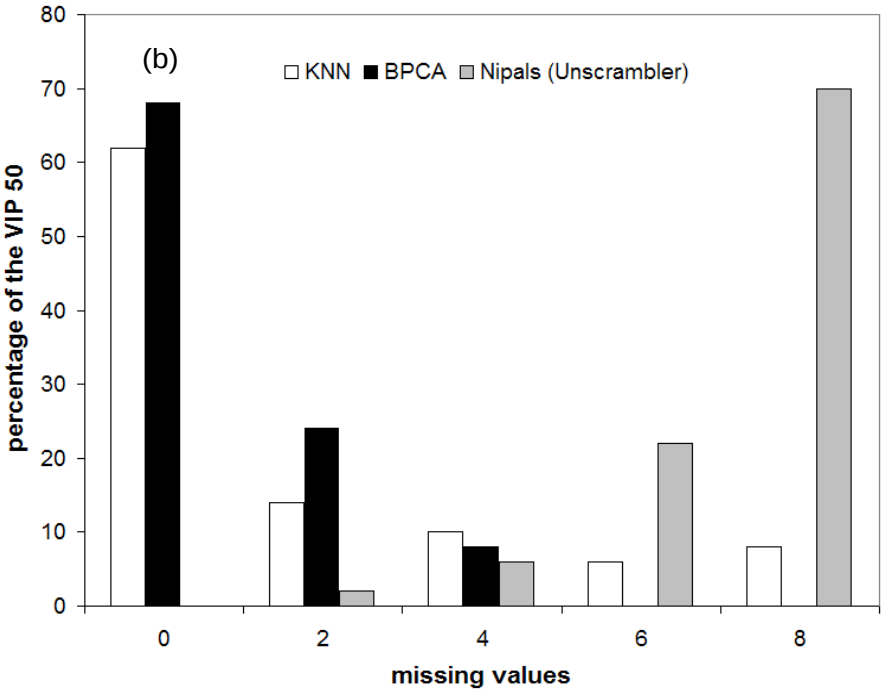

(b)
KNN
BPCA
Nipals (Unscrambler)
percentage of the VIP 50
missing values
0
10
20
30
40
50
60
70
80
0
2
4
6
8

Figure 6

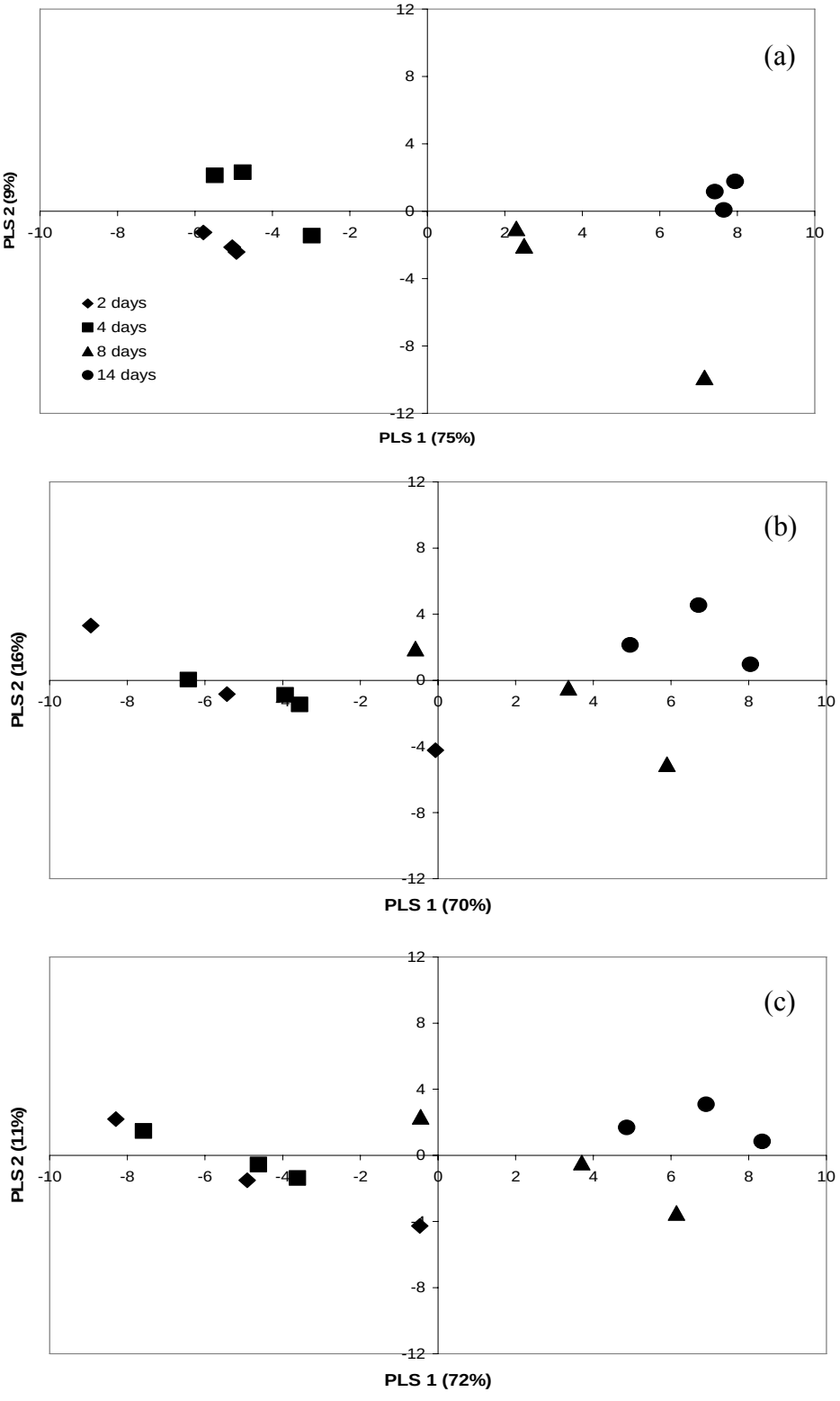

(a)
PLS 2 (9%)
PLS 1 (75%)
2 days
4 days
8 days
14 days
(b)
PLS 2 (16%)
PLS 1 (70%)
(c)
PLS 2 (11%)
PLS 1 (72%)

Figure 7

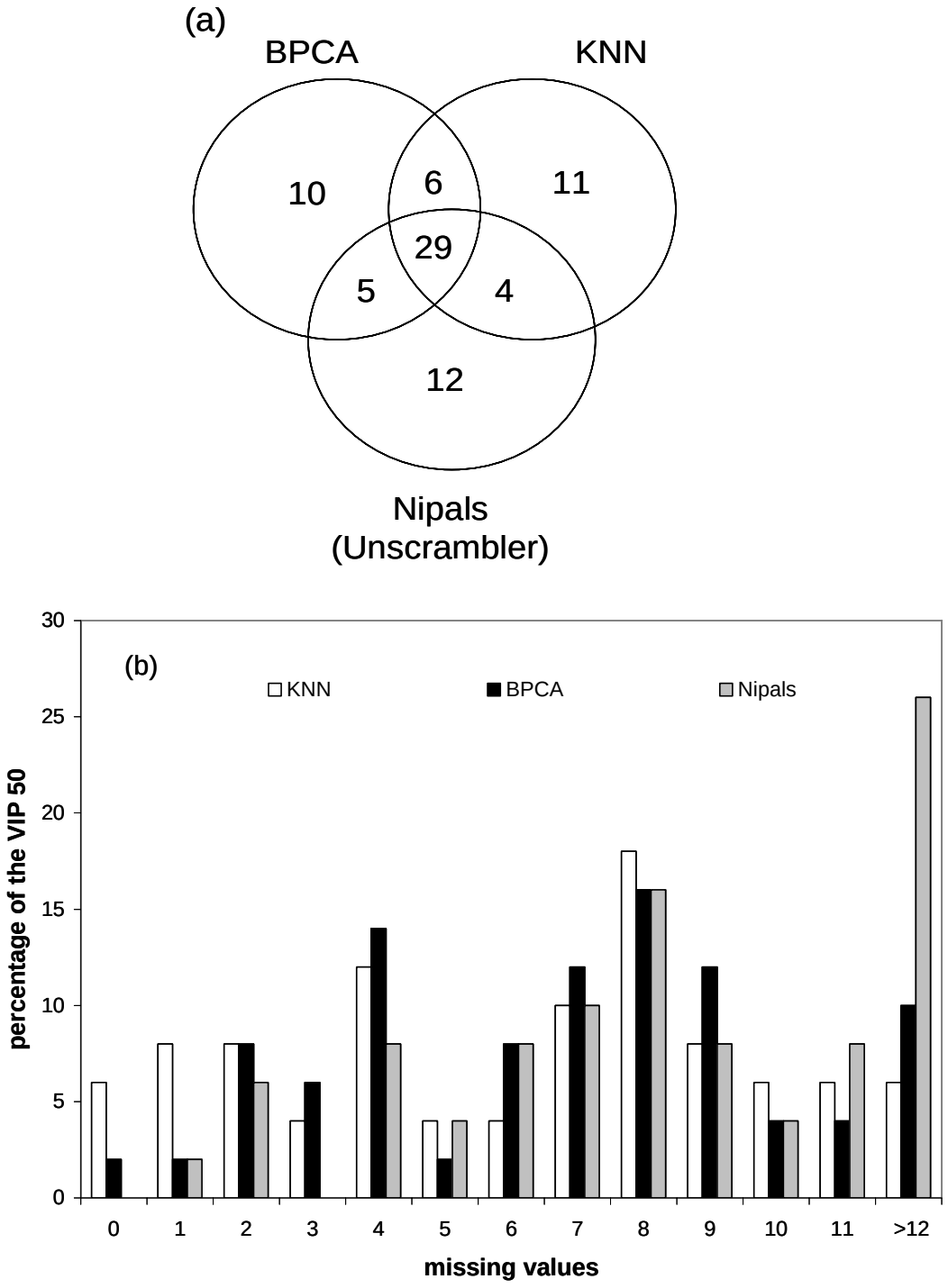

(a)
BPCA
KNN
10
6
11
29
5
4
12
Nipals
(Unscrambler)
(b)
KNN
BPCA
Nipals
percentage of the VIP 50
missing values
0
5
10
15
20
25
30
0
1
2
3
4
5
6
7
8
9
10
11
>12

Figure 8

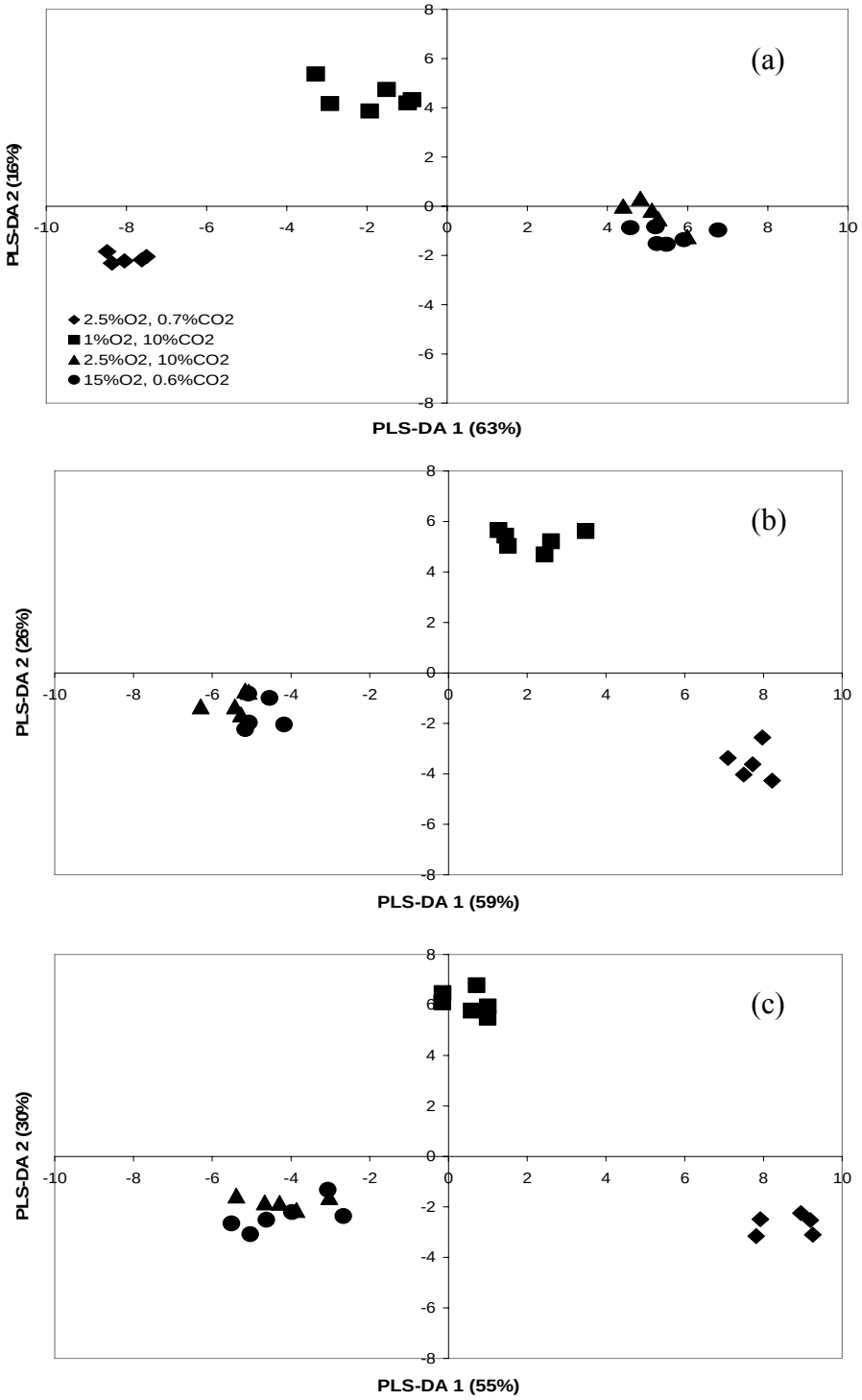
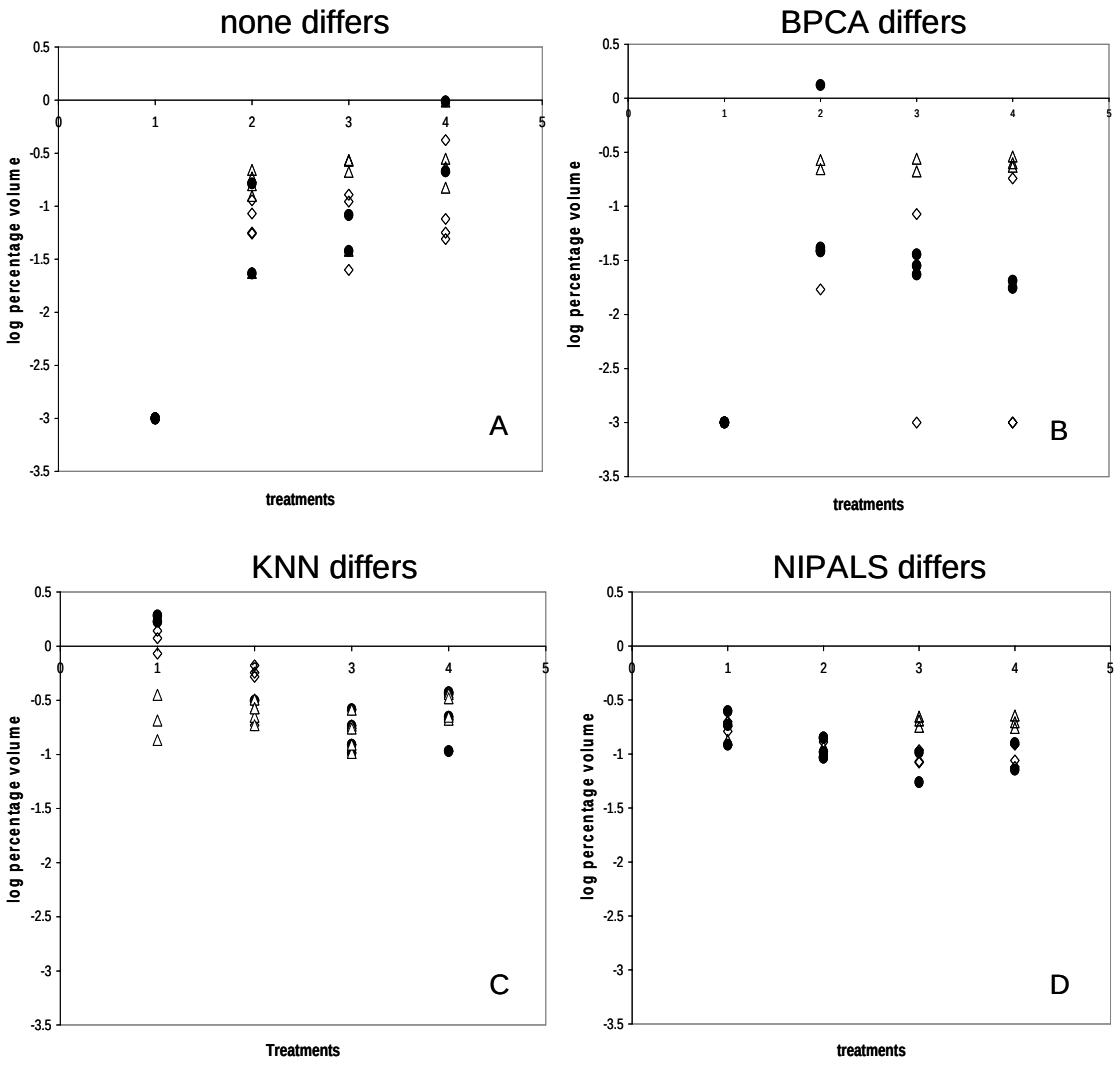

(a)
PLS-DA 2 (16%)
PLS-DA 1 (63%)
2.5%O2, 0.7%CO2
1%O2, 10%CO2
2.5%O2, 10%CO2
15%O2, 0.6%CO2
(b)
PLS-DA 2 (26%)
PLS-DA 1 (59%)
(c)
PLS-DA 2 (30%)
PLS-DA 1 (55%)

Figure 9


none differs
BPCA differs
KNN differs
NIPALS differs
log percentage volume
treatments
Treatments
A
B
C
D