Communications for Statistical Applications and Methods
2015, Vol. 22, No. 6, 575–587

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.6.575
Print ISSN 2287-7843 / Online ISSN 2383-4757

Multivariate Procedure for Variable Selection and Classification of High Dimensional Heterogeneous Data

Tahir Mehmood[1,a,b] , Zahid Rasheed[a]

[a]Statistics, Department of Basic Sciences, Riphah International University, Pakistan;
[b]Biostatistics, Department of Chemistry, Biotechnology and Food Sciences,
Norwegian University of Life Sciences, Norway

Abstract

The development in data collection techniques results in high dimensional data sets, where discrimination is an important and commonly encountered problem that are crucial to resolve when high dimensional data is heterogeneous (non-common variance covariance structure for classes). An example of this is to classify microbial habitat preferences based on codon/bi-codon usage. Habitat preference is important to study for evolutionary genetic relationships and may help industry produce specific enzymes. Most classification procedures assume homogeneity (common variance covariance structure for all classes), which is not guaranteed in most high dimensional data sets. We have introduced regularized elimination in partial least square coupled with QDA (rePLS-QDA) for the parsimonious variable selection and classification of high dimensional heterogeneous data sets based on recently introduced regularized elimination for variable selection in partial least square (rePLS) and heterogeneous classification procedure quadratic discriminant analysis (QDA). A comparison of proposed and existing methods is conducted over the simulated data set; in addition, the proposed procedure is implemented to classify microbial habitat preferences by their codon/bi-codon usage. Five bacterial habitats (*Aquatic*, *Host Associated*, *Multiple*, *Specialized* and *Terrestrial*) are modeled. The classification accuracy of each habitat is satisfactory and ranges from 89.1% to 100% on test data. Interesting codon/bi-codons usage, their mutual interactions influential for respective habitat preference are identified. The proposed method also produced results that concurred with known biological characteristics that will help researchers better understand divergence of species.

Keywords: partial least squares, classification, variable selection, parsimonious model, high dimensional data sets, identification, multi collinearity, microbial

1. Introduction

Tremendous advances in technology has made it possible to sample observations based on a huge number of genetic and ecological variables. It is much easier to generate gigantic sets of raw data, establish relations and provide a biological understanding (Bachvarov *et al.*, 2008).

Huge sets of variables are typically used as explanatory variables and are assumed with a potential impact on classification variables. Most biological studies result in large numbers of variables p compared the number of samples n. In such situation logistic regression or other traditional classification methods like linear discriminant analysis (LDA) (Barker and Rayens, 2003; Lachenbruch and Goldstein, 1979) or quadratic discriminant analysis (QDA) (Hastie *et al.*, 2009; Lachenbruch

[1] Corresponding author: Biostatistics, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Oslo, Norway. E-mail: tahime@gmail.com

Published 30 November 2015 / journal homepage: http://csam.or.kr

and Goldstein, 1979) face a multi collinearity and identification problem (Wold *et al.*, 1984). In this instance, multivariate approach, like partial least square (PLS), is natural to use. This regression provides the solution in the 'large p small n' situation, see (Martens and Næs, 1989). PLS in its original form is a regression method; however, but has been extensively used for classification analysis (Alsberg *et al.*, 1998; Liland *et al.*, 2013; Mehmood *et al.*, 2011a, 2011b, 2012a, 2012b, 2012c, 2014; Wold *et al.*, 1984). In general, if we have C classes, one of the accepted procedure is to first convert the response into C binary responses, and then fit each binary response with an explanatory variable. The predicted class is determined based on the *sign* of predicted PLS response (Martens and Næs, 1989; Sæbø *et al.*, 2008).

An alternative approach for the classification of high dimensional data set is to couple PLS scores with LDA, which was first used by Lindgren *et al.* (1994), and later studied in (Boulesteix, 2004; Chun and Keleş, 2010; Lindgren *et al.*, 1994; Mehmood *et al.*, 2011b; Nguyen and Rocke, 2002a). The extensive comparison study performed by Boulesteix (2004), which included many classification methods, employing PLS as a dimension reduction method and using the PLS components scores as predictors in LDA ranges among the best classification procedures for all the eight cancer data sets. PLS scores corresponding to an optimum PLS model are normally very few compared to the total number of explanatory variables and more over are assumed to be orthogonal. Subsequently, orthogonal PLS components as an input for the LDA handles the problem of multi collinearity and identification, and results the satisfactory classification accuracy. LDA is assumed to provide best possible accuracy given that its assumptions are satisfied, which are the homogeneity i.e. common variance covariance structure and multivariate normal distribution. For high dimensional data sets multivariate normal distribution is mostly conserved, while the common variance covariance structure for each class is in question (Hastie *et al.*, 2009). The non-common variance covariance structure motivates to couple PLS scores with QDA first introduced by Nguyen and Rocke (2002a, 2002b) and later studied in (Boulesteix, 2004).

Extracting the parsimonious model (the smallest number of variables which explains the modeled relation better while keeping the satisfactory classification accuracy) is also required with the classification accuracy in high dimensional data sets. This motivates researchers to create variable selection strategies with several possibilities. For instance soft threshold based shrinkage in PLS (stPLS) (Sæbø *et al.*, 2008), regularized stepwise elimination procedure for variable selection in PLS (rePLS) (Mehmood *et al.*, 2011a). For classification purpose, stPLS uses the *sign* of predicted response based approach, while rePLS merges the PLS with LDA (rePLS-LDA). Motivated by a recently introduced variable selection procedure regularized elimination in partial least square (rePLS) (Mehmood *et al.*, 2011a) and heterogeneous classification procedure QDA (Hastie *et al.*, 2009), we have introduced regularized elimination in partial least square coupled with QDA (rePLS-QDA) for the parsimonious variable selection and classification of high dimensional heterogeneous data sets, while the heterogeneity means the non-common variance covariance structure in classes. The accuracy of rePLS-QDA has been compared with PLS-LDA, PLS-QDA, rePLS-LDA and stPLS over the simulated data. A comparison of rePLS-QDA has been done with PLS-QDA and stPLS on real data since the considered real biological data sets appears heterogeneous.

2. Methods

2.1. Partial least squares

We have considered a classification problem where every object belongs to one of two possible classes, as indicated by the $n \times 1$ class label vector $\boldsymbol{y}$. We fit C models if response has C classes then we create

C numeric response vector $\boldsymbol{y}$ by coding as +1's (if the sample is from respective class) and −1's (if the sample is not from respective). The association between $\boldsymbol{y}$ and the $n \times p$ predictor matrix $\boldsymbol{X}$ is assumed to be explained by the linear model $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ where $\boldsymbol{\beta}$ are the $p \times 1$ vector of regression coefficients. The purpose of variable selection is to find a column subset of $\boldsymbol{X}$ capable of satisfactory explaining the variations in $\boldsymbol{y}$.

From a modeling perspective, ordinary least square fitting is no option when $n < p$. PLS resolves this by searching for a small set of components, 'latent vectors', that performs a simultaneous decomposition of $\boldsymbol{X}$ and $\boldsymbol{y}$ with the constraint that these components explain much of the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$.

Initially the variables are centered into $\boldsymbol{X}_0 = \boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}'$ and $\boldsymbol{y}_0 = \boldsymbol{y} - 1\bar{y}$. Let A be the number of components to be extracted. Then for $a = 1, 2, \ldots, A$ the algorithm runs:

1. Compute the loading weights by

$$\boldsymbol{w}_a = \boldsymbol{X}'_{a-1}\boldsymbol{y}_{a-1}.$$

The weights define the direction in the space spanned by $\boldsymbol{X}_{a-1}$ of maximum covariance with $\boldsymbol{y}_{a-1}$. Normalize to loading weights to have length equal to 1 by

$$\boldsymbol{w}_a \leftarrow \frac{\boldsymbol{w}_a}{\|\boldsymbol{w}_a\|}.$$

2. Compute the score vector $\boldsymbol{t}_a$ by

$$\boldsymbol{t}_a = \boldsymbol{X}_{a-1}\boldsymbol{w}_a.$$

3. Compute the X-loadings $\boldsymbol{p}_a$ by regressing the variables in $\boldsymbol{X}_{a-1}$ on the score vector:

$$\boldsymbol{p}_a = \boldsymbol{X}'_{a-1}\frac{\boldsymbol{t}_a}{\boldsymbol{t}'_a\boldsymbol{t}_a}.$$

Similarly compute the Y-loading q_a by

$$q_a = \boldsymbol{y}'_{a-1}\frac{\boldsymbol{t}_a}{\boldsymbol{t}'_a\boldsymbol{t}_a}.$$

4. Deflate $\boldsymbol{X}_{a-1}$ and $\boldsymbol{y}_{a-1}$ by subtracting the contribution of $\boldsymbol{t}_a$:

$$\boldsymbol{X}_a = \boldsymbol{X}_{a-1} - \boldsymbol{t}_a\boldsymbol{p}'_a,$$
$$\boldsymbol{y}_a = \boldsymbol{y}_{a-1} - \boldsymbol{t}_a q_a.$$

5. If $a < A$ return to 1.

Let the loading weights, scores and loadings computed at each step of the algorithm be stored in matrices/vectors $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_A]$, $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_A]$, $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_A]$ and $\boldsymbol{q} = [q_1, q_2, \ldots, q_A]$. Then the PLS estimators for the regression coefficients for the linear model are found by $\hat{\boldsymbol{\beta}} = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}\boldsymbol{q}$, which indicates the involvement of respective variables in the PLS model.

2.2. Regularized elimination procedure for variable selection in PLS

Recently, a stepwise regularized variable elimination procedure for variable selection (Mehmood *et al.*, 2011a) is proposed for parsimonious model fitting. The procedure starts with the split of the training data into test and training subsets. For each split, a stepwise procedure is adopted to select the variables. Stable variables that are being extracted by stepwise elimination from all splits of the data are consequently selected. The proposed algorithm requires a ranking of variables in $\boldsymbol{X}$, which is accomplished by variable importance in PLS projections (VIP) (Eriksson *et al.*, 2001).

VIP for the variable j is defined according to (Eriksson *et al.*, 2001) as

$$v_j = \sqrt{p \sum_{a=1}^{a^*} \left[\left(\boldsymbol{p}_{2a}^2 \boldsymbol{t}_a' \boldsymbol{t}_a \right) \left(\frac{\boldsymbol{w}_{aj}}{||\boldsymbol{w}_a||} \right)^2 \right] / \sum_{a=1}^{a^*} \left(\boldsymbol{p}_{2a}^2 \boldsymbol{t}_a' \boldsymbol{t}_a \right)},$$

where $a = 1, 2, \ldots, A$, $\boldsymbol{w}_{aj}$ is the loading weight for variable j using a components and $\boldsymbol{t}_a$, $\boldsymbol{w}_a$ and $\boldsymbol{p}_{2a}$ are CPPLS scores, loading weights and y-loadings respectively corresponding to the a^{th} component. Gosselin *et al.* (2010) explains the main difference between the regression coefficient β_j and v_j. The v_j weights the contribution of each variable according to the variance explained by each PLS component, i.e. $\boldsymbol{p}_{2a}^2 \boldsymbol{t}_a' \boldsymbol{t}_a$ where $(\boldsymbol{w}_{aj}/||\boldsymbol{w}_a||)^2$ represents the importance of the j^{th} variable. Variable j can be eliminated if $v_j < u$ for some user-defined threshold $u \in [0, \infty)$. It is generally accepted that a variable should be selected if $v_j > 1$, see (Eriksson *et al.*, 2001).

The algorithm is: Let $\boldsymbol{Z}_0 = \boldsymbol{X}$ and let VIP_j be the variable importance for variable j.

1) For iteration g run $\boldsymbol{y}$ and $\boldsymbol{Z}_g$ through cross-validated PLS. The matrix $\boldsymbol{Z}_g$ has p_g columns, and we get the same number of criterion values, sorted in ascending order as $\text{VIP}_{(1)}, \ldots, \text{VIP}_{(p_g)}$.

2) There are M criterion values below (above for criterion q_j) the cutoff u. If $M = 0$, terminate the algorithm here.

3) Else, let $N = \lceil fM \rceil$ for some fraction $f \in \langle 0, 1]$. Eliminate the variables corresponding to the N most extreme criterion values.

4) If there are still more than one variable left, let $\boldsymbol{Z}_{g+1}$ contain these variables, and return to 1).

The fraction f determines the 'step length' of the elimination algorithm, where an f close to 0 will only eliminate a few variables in every iteration. We used $f = 1$ and $u = 1$ as reported in (Eriksson *et al.*, 2001; Mehmood *et al.*, 2011a). The implemented iterative procedure allows for a marginal decrease in model discrimination performance that can significantly decrease the number of selected variables, which in turn improve the interpret ability of the model noticeably, see Mehmood *et al.* (2011a).

2.3. Classification of heterogeneous data sets by using QDA with rePLS

The procedure rePLS results in $\boldsymbol{X}_{selected}$ containing selected variables only. Since the rePLS scores S are normally distributed and usually initial few component explain the most of the variation of the original data. In presence of heterogeneity, we assume the density $f_C(s)$, which presents rePLS scores S in class C follows multivariate normal distribution with class mean μ_C and Σ_C is the variance covariance for class C that maximizes the discriminant function

$$\delta_C(x) = -\frac{1}{2} \log |\Sigma_C| - \frac{1}{2}(x - \mu_C)' \Sigma_C^{-1} (x - \mu_C) + \log \pi_C. \tag{2.1}$$

Hence by employing rePLS as a dimension reduction and variable selection method and using the rePLS components in each iteration as predictors in a QDA, classification of heterogeneous data can be achieved.

2.4. Reference methods

2.4.1. PLS-LDA and PLS-QDA

We can couple PLS with LDA or QDA which results is PLS-LDA and PLS-QDA respectively since PLS is regression procedure, for the discrimination. In PLS-LDA setting (Boulesteix, 2004; Lindgren *et al.*, 1994), we assume the density $f_C(s)$, which presents PLS scores S in class C follows multivariate normal distribution with class mean μ_C and pooled with in class covariance matrix Σ, while in QDA-PLS (Boulesteix, 2004; Nguyen and Rocke, 2002a) setting we assume the heterogeneous variance covariance structure Σ_C over PLS scores.

2.4.2. Soft-Threshold PLS (stPLS)

Sæbø *et al.* (2008) introduced a soft-thresholding step in PLS algorithm (stPLS) based on ideas from the nearest shrunken centroid method (Tibshirani *et al.*, 2003). The stPLS approach is more or less identical to the Sparse-PLS presented independently by Lê Cao *et al.* (2008). At each step of the sequential stPLS algorithm the PLS loading weights $\boldsymbol{w}$ are modified as:

1. Scaling:

   $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k / \max_j |w_{k,j}|$, for $j = 1, \ldots, p$ and $k = 1, \ldots, A$, where A is the number of PLS components.

2. Soft-thresholding:

   $w_{k,j} \leftarrow \text{sign}(w_{k,j})(|w_{k,j}| - \delta)_+$, for $j = 1, \ldots, p$ and some $\delta \in [0, 1\rangle$. For any real number a, here $(a)_+$ means $\max(0, a)$.

3. Normalizing:

   $\boldsymbol{w}_k \leftarrow \frac{\boldsymbol{w}_k}{||\boldsymbol{w}_k||}$.

The shrinkage $\delta \in [0, 1)$ sets the degree of thresholding, i.e. a larger δ gives a smaller selected set of variables. Cross validation is used to define this threshold. stPLS uses the sign(β_{stPLS}) for the classification of new samples in to class '+1' or into class '−1' (Sæbø *et al.*, 2008).

3. Simulation Study

3.1. Data simulation

In order to make comparison of stPLS, rePLS-LDA and rePLS-QDA, data for each class was simulated from a known model $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$, where $\boldsymbol{y}$ takes -1 if respective sample is from class 1, and $+1$ if respective sample is from class 2. The p-vectors of variables $\boldsymbol{x}$ was assumed multivariate normally distributed with mean-vector $\mu = 0$ and covariance matrix Σ.

$$\boldsymbol{X} \sim \text{MVN}(0, \Sigma)$$

The variances of all variables were set equal to 1, hence Σ is also a correlation matrix. Further, groups of correlated $\boldsymbol{x}$ variables were constructed by imposing a block diagonal structure on Σ with L blocks i.e.

$$\Sigma_C = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_L \end{bmatrix}.$$

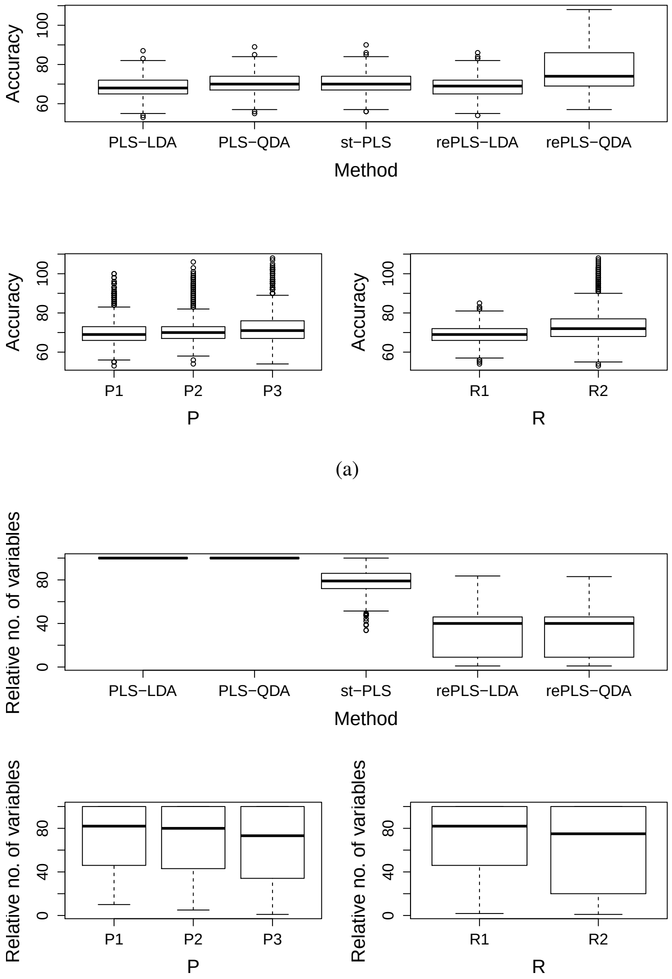
3.2. Simulation based results

For simulation study, three levels of number of variables $P = (50, 100, 500)$ with total of $L = 10$ groups of equal number of variables were considered. We have assumed different correlation structures for both classes since we have considered a two class problem and our objective is to make the comparison when each class follows day different variances covariance structure (Σ_C). We have used uniform Σ_{Cl} by $R = [0.3, 0.3, 0.2, 0, 0, 0, 0, 0, 0, 0]$ for class 1 and by $R = ([0.6, 0.9, 0, 0, 0, 0, 0, 0, 0, 0]$, $[-0.5, -0.3, 0, 0, 0, 0, 0, 0, 0, 0])$ for class 2. This indicates Σ_C has 2 levels, where the second level of R creates the more heterogeneous data for class 2 compared to the first level of R. Two levels of Σ_C together with 3 levels of P results in total 6 different data sets. To have stable estimates of the fitted models and their comparison 100 runs were used.

For each run, we have simulated training and test data set each of size $N = 100$, where training data set was used to train the PLS methods while test data set was used to evaluate methods. The classification accuracy over the test data and relative number of selected variables (= (number of selected variables/total number of variables) × 100) were extracted for each run. Through F-test, we found PLS methods (p-value < 0.001) and Σ_k structures (p-value < 0.001) are significantly factors which explain the variation in the accuracy of PLS methods. Similarly, PlS methods, number of variables and Σ_k structures were all found significant (p-value < 0.001) in explaining the variation in a relative number of selected variables. Figure 1 presents the distributions of accuracy and relative number of selected variables for PlS methods, number of variables and Σ_k structures. In terms of accuracy, we found rePLS-QDA outperforms the PLS-QDA, rePLS-LDA and stPLS. Further, as the variance structure of class 2 get more different from the variance structure of class 1 the accuracy of PLS based approaches increased. Of interest is that rePLS-LDA and rePLS-QDA are both found to have relatively small number of variables compared to stPLS as shown in lower panel of Figure 1. A relatively small number of selected variables were found with $P = 500$ and $R = [-0.5, -0.3, 0, 0, 0, 0, 0, 0, 0, 0]$.

4. Application

An application of the classification procedure is to find the preferred habitat based on relevant codons/ bi-codon associated with a certain microbial. A huge amount of genetic divergence in microbial is observed. Genomic data can be used to characterize different microbial communities occupying different environmental niches (Tringe *et al.*, 2005). There are many factors that cause this divergence, and habitat preference is one among them (Hübner *et al.*, 2013). Habitat preference is also important to study for evolutionary genetic relationship and may help a potential industry to produce specific enzymes. Production organisms need to accurately portray and selection pressure under which the microbes evolve with required needs (Jensen *et al.*, 2012). Microbial habitat preference is mostly effected by ecological conditions like depth of water, season, temperature, desert and soil condition (Handelsman, 2004). Genome from same species may have different habitat preferences, which could be the result of competitive elimination on a very small scale and ecological conditions (Watson *et al.*,



(a)

(b)

Figure 1: *The distribution of (a) accuracy measured over the test data and (b) relative number of selected variables is presented for different levels of PLS methods, number of variables (P) and Σ_k structures (R).*

2004). Microbial communities in ecosystems are influenced by habitat preferences and are considered important sources of genetic variation with microbes classified based on habitat type (Singh *et al.*, 2006).

Microbial codon usage can be classified between different bacterial habitat and may eventually help researchers understand genetic divergences of microbial species. The overall codon usage is affected by the selection of amino acids and codon bias within the redundant amino acids. Codons are triplets of nucleotides in coding genes and messenger RNA that codons translate genetic information into specific proteins. There are 20 amino acids are singly coded by 1, 2, 4 or 6 different codons (excluding the three stop codons there are 61 codons). However, the different codons encoding individual amino acids are not selectively correspondent because corresponding tRNAs differ in abundance, allowing for selection on codon usage. Codon usage is a pointer of the force shaping genome evolution in prokaryotes (Mehmood *et al.*, 2011a; Mehmood and Snipen, 2013), reflection of life style (Hanes *et al.*, 2009) and organisms within similar habitat often have similar codon usage pattern in their genomes (Chen *et al.*, 2007). Higher order codon frequencies, e.g. di-codons are considered important with respect to joint effects of codons, like synergistic effect (Nguyen *et al.*, 2009).

4.1. Habitat discrimination data

The genomic sequence data used to train the model was divided into two groups, Positives and Negatives. Positives contained microbial coding sequence having respective habitat preference, and Negatives consist of random coding sequences.

4.1.1. Positives

We have considered 445 microbial genomes, their genomic sequence and the respective habitat information were obtained from NCBI Genome Projects (`http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi`). The response variable in our data set is habitat preference. There are in total 5 habitat preferences included in our data set, namely *Aquatic*, *Host Associated*, *Multiple*, *Specialized* and *Terrestrial*. For each genome, genes were predicted by gene prediction software Prodigal (Hyatt *et al.*, 2010), these genes are considered as a set of Positives.

4.1.2. Negatives

We only consider two-class problems, i.e. for some fixed 'habitat A', we only classify genomes as either 'habitat A', or 'not habitat A'. Thus, we have 5 different responses of $-1/+1$ outcome, considering one at a time. Negatives sequences must be contrasted against sequences those are not belonging to respective habitat. While constructing Negatives AT/GC content should be preserved, so we have permuted the position of DNA alphabets in predicted genes i.e. of Positive sequence. Hence we get the Negatives sequences having the equal length and number of Positive sequences.

4.2. Data splitting

The genome of one strain from each habitat category was randomly selected from which Positive and Negative sequences were extracted. We used a cross validation type approach where the data sets containing Positives and Negatives were randomly divided into 10 equally sized subsets, with one of these subsets taking the role of test data while the other 9 remaining subsets are considered as training data.

4.3. Results and discussion

For identification of codon variations that distinguishes different habitat preference of microbes, 5 models, representing each habitat preference were considered separately. The number of genomes, average GC-content, average GC-variation, average genome size in mega bases (MB) and average growth temperature (C) for each habitat (Table 1).

For the discriminating the habitat preference of microbial based on codon/di-codon usage, PLS-QDA, stPLS, rePLS-LDA, and rePLS-QDA was used. All habitant preference discriminative models we found the PLS scores are following the multivariate normal distribution but the common variance

Table 1: An overview of the habitats used in the current study along with number of genomes, average GC-content, average GC-variation, average genome size in mega bases (MB) and average growth temperature (C)

Habitat	Number of genomes	GC-content	GC variation	Genome size (MB)	Growth temperature (C)
Aquatic	105	0.51	0.031	3.56	38.8
Host Associated	155	0.44	0.029	2.61	34.4
Multiple	142	0.53	0.030	4.12	30.7
Specialized	56	0.48	0.030	2.42	59.7
Terrestrial	27	0.58	0.029	5.01	31.9

Table 2: Number of components, number of selected variables, discrimination performance on test data by using PLS-QDA, rePLS-QDA and stPLS are presented

Habitat	Number of components			Number of selected variables		Discriminatory performance (%)		
	PLS-QDA	rePLS-QDA	stPLS	rePLS-QDA	stPLS	PLS-QDA	rePLS-QDA	stPLS
Aquatic	3	2	2	56	62	96.7	98.4	92.1
Host Associated	2	8	2	19	65	96.8	98.9	94.7
Multiple	4	8	8	32	62	96.4	100.0	94.2
Specialized	2	8	3	28	62	94.1	100.0	91.2
Terrestrial	2	9	2	20	64	94.4	89.1	94.4

PLS = partial least square; QDA = quadratic discriminant analysis; rePLS = regularized elimination in PLS; stPLS = soft-thresholding PLS.

covariance over the classes is not satisfied. For *Aquatic* preference discrimination, it appears PLS scores follows multivariate normal distribution (Mardia's multivariate normality test statistics = 6.30, p-value = 0.177) and PLS scores have non-constant variance over the discriminating classes (Box M's test = 1001.9, p-value < 0.01), and this trend holds for each discriminating habitat. In such situation, rePLS-LDA is not recommended, while rePLS-QDA, PLS-QDA and stPLS are natural to use to classify samples as +1 or −1. For model estimation, a tenfold cross validation was used. The number of components presenting the complexity of each model, number of selected variables, discrimination accuracy on test data by using PLS-QDA, rePLS-QDA, rePLS-LDA and stPLS (Table 2). These results indicates both rePLS-QDA performs the best in discriminating the most of habitat preferences of microbial by using codon/di-codon usage. It appears that, the *Terrestrial* habitat is in general more difficult to classify, simply because there are more diversity inside the group (Hättenschwiler *et al.*, 2011). This group also has the smallest number of genomes and adds a sampling bias. Number of components indicates PLS-QDA has on the average simplest level of complexity, while both rePLS-QDA and stPLS also have relatively higher but similar model complexity level on the average. Since PLS-QDA is the classification method only, while rePLS-QDA and stPLS are variable selection and classification method as well, it appears rePLS-QDA is more parsimonious compared to stPLS, since it selects comparatively less number of influential variables and results in comparatively better classification accuracy in all cases. This is because, stepwise iterative parsimonious variable elimination procedure was implemented to find codon/di-codon usage which improves the model interpretation as well as eliminates some useful redundancy from the model and use a small number of variables to discriminate the habitat (Norgaard *et al.*, 2000), and that is why consistency of selected variables, as utilized, retains the discriminating performance (Mehmood *et al.*, 2011a). Hence for rest of the analysis we have focused over the rePLS-QDA and have chosen the habitat *Aquatic* for a detailed illustration of the method, while results for all habitat preferences are also provided in supplementary material. Figure 2 shows the correlation biplot for *Aquatic* over the first two PLS components. The correlation biplot shows for each codon/bi-codon their contribution to the two dimensions i.e. underlying phenomena (loadings), and for each genome (sample) their relative position in two dimensional

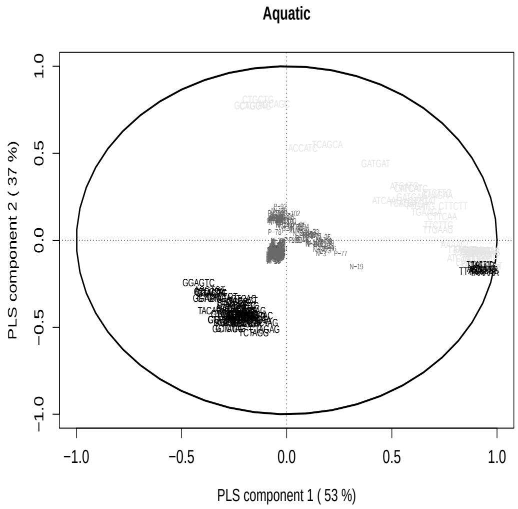


Figure 2: *The biplot for* Aquatic *is presented. Influential codon/di-codon influential for respective habitat (positive PLS regression coefficients) are labeled by their names in red color and influential codon/bi-codon influential codon/bi-codon for random sequences (negative PLS regression coefficients) are labeled by their names in green color. Further genomes (samples) are indicated by gray color.*

space (scores).

Given that the habitat having reasonable discriminative performance, specific genetic variations are expected (Costello *et al.*, 2009). These genetic variations are expressed as selection of codon/bi-codon. Influential codon/bi-codon having positive PLS regression coefficients i.e. influential for respective habitat are labeled by their names in red color and influential codon/bi-codon having negative PLS regression coefficients i.e. influential codon/bi-codon for random sequences are labeled by their names in green color. This identifies the influential codon/bi-codon for *Aquatic* as grouped in the same direction while the bi-codon 'AGCAGC' acts in different direction. This bi-codon is translated into amino acid Serine. Further genomes (samples) are indicated by gray color and the correlation loadings indicate no outlier in the samples for this analysis.

All of the selected variables are bi-codon and provide additional support for the interaction of genetic information is highly important for explaining variations in habitat (Botzman and Margalit, 2011; Lejeusne and Chevaldonné, 2006). From influential codon/bi-codon, we have considered the only bi-codon having positive PLS regression coefficient for rest of the analysis i.e. those are responsible of explaining variation in respective habitat. The influential bi-codon interactions are plotted in bipartite plot, which presents the network of influential codons for *Aquatic* in Figure 3. We marked the codon as most interactive codon if it has at least three linkages, and found two most interactive codons as 'TGA', and 'AAA' which codes for stop codon and amino acid K(Lys) respectively. Codon bias is also obvious to detect here, since there are many more than one codons that translate into unique amino acids (Figure 3).

5. Conclusion

We have proposed PLS based approaches called regularized backward elimination algorithm in PLS coupled with quadratic discriminant analysis (rePLS-QDA) for the variable selection and classifica-

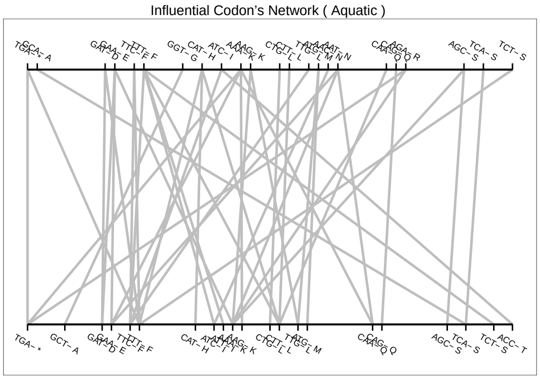


Figure 3: *The bipartite plots showing the network of influential codons for* Aquatic *is presented.*

tion of heterogeneous high dimensional data sets. We obtained a huge reduction in the number of selected variables with acceptable classification accuracy. Proposed method out performs the PLS-LDA, PLS-QDA, rePLS-LDA and stPLS over the simulated data. Proposed methods were also successfully applied for habitat classification based on codon/bi-codon usages of microbes. We obtain habitat models with superior interpretation; however, any type of genome-wide association study may potentially benefit from the use of a multivariate selection.

References

Alsberg, B. K., Kell, D. B. and Goodacre, R. (1998). Variable selection in discriminant partial least-squares analysis, *Analytical Chemistry*, **70**, 4126–4133.

Bachvarov, B., Kirilov, K. and Ivanov, I. (2008). Codon usage in prokaryotes, *Biotechnology & Biotechnological Equipment*, **22**, 669–682.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination, *Journal of Chemometrics*, **17**, 166–173.

Botzman, M. and Margalit, H. (2011).Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles, *Genome Biol*, **12**, R109.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data, *Statistical Applications in Genetics and Molecular Biology*, **3**, 1–30.

Chen, R., Yan, H., Zhao, K. N., Martinac, B. and Liu, G. B. (2007). Comprehensive analysis of prokaryotic mechanosensation genes: Their characteristics in codon usage, *DNA Sequence*, **18**, 269–278.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 3–25.

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I. and Knight, R. (2009). Bacterial community variation in human body habitats across space and time, *Science*, **326**, 1694–1697.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (2001). *Multi-and Megavariate Data Analysis*, Umetrics Academy, Umeå.

Gosselin, R., Rodrigue, D. and Duchesne, C. (2010). A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, *Chemometrics and Intelligent Laboratory Systems*, **100**. 12–21.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms, *Microbiology and Molecular Biology Reviews*, **68**, 669–685.

Hanes, A., Raymer, M. L., Doom, T. E. and Krane, D. E. (2009). A comparision of codon usage trends in prokaryotes, In *Proceedings of Ohio Collaborative Conference on Bioinformatics (OC-CBIO'09)*, Cleveland, OH, 83–86.

Hastie, T., Tibshirani, R. and Friedman, J. (2009).*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Hättenschwiler, S., Fromin, N. and Barantal, S. (2011). Functional diversity of terrestrial microbial decomposers and their substrates, *Comptes Rendus Biologies*, **334**, 393–402.

Hübner, S., Rashkovetsky, E., Kim, Y. B., Oh, J. H., Michalak, K., Weiner, D., Korol, A. B. Nevo, E. and Michalak, P. (2013). Genome differentiation of Drosophila melanogaster from a microclimate contrast in Evolution Canyon, Israel, In *Proceedings of the National Academy of Sciences*, **110**, 21059–21064.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, **11**, 119.

Jensen, D. B., Vesth, T. C., Hallin, P. F., Pedersen, A. G. and Ussery, D. W. (2012). Bayesian prediction of bacterial growth temperature range based on genome sequences, *BMC Genomics*, **13**(Suppl 7), S3.

Lachenbruch, P. A. and Goldstein, M. (1979). Discriminant analysis, *Biometrics*, **35**, 69–85.

Lê Cao, K. A., Rossouw, D., Robert-Granié, C. and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data, *Statistical Applications in Genetics and Molecular Biology*, **7**, 1–32.

Lejeusne, C. and Chevaldonné, P. (2006). Brooding crustaceans in a highly fragmented habitat: the genetic structure of Mediterranean marine cave-dwelling mysid populations, *Molecular Ecology*, **15**, 4123–4140.

Liland, K. H., Høy, M., Martens, H. and Sæbø, S. (2013). Distribution based truncation for variable selection in subspace methods for multivariate regression, *Chemometrics and Intelligent Laboratory Systems*, **122**, 103–111.

Lindgren, F., Geladi, P., Rännar, S. and Wold, S. (1994). Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms, *Journal of Chemometrics*, **8**, 349–363.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*, Wiley & Sons, New York.

Mehmood, T., Bohlin, J., Kristoffersen, A. B., Sæbø, S., Warringer, J. and Snipen, L. (2012b). Exploration of multivariate analysis in microbial coding sequence modeling, *BMC Bioinformatics*, **13**, 97.

Mehmood, T., Bohlin, J. and Snipen, L. (2014). A partial least squares based procedure for upstream sequence classification in prokaryotes., *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **12**, 560–567.

Mehmood, T., Liland, K. H., Snipen, L. and Sæbø, S. (2012a). A review of variable selection methods in partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, **118**, 62–69.

Mehmood, T., Martens, H., Sæbø, S., Warringer, J. and Snipen, L. (2011a).A partial least squares based algorithm for parsimonious variable selection, *Algorithms for Molecular Biology*, **6**, 27.

Mehmood, T., Martens, H. and Sæbø, S., Warringer, J. and Snipen, L. (2011b). Mining for genotype-phenotype relations in Saccharomyces using partial least squares, *BMC Bioinformatics*, **12**, 318.

Mehmood, T. and Snipen, L. (2013). Clustered variable selection by regularized elimination in PLS.

In H. Abdi, et al. (Eds.), *New Perspectives in Partial Least Squares and Related Methods* (pp. 95–105), Springer, New York.

Mehmood, T., Warringer, J., Snipen, L. and Sæbø, S. (2012c). Improving stability and understandability of genotype-phenotype mapping in Saccharomyces using regularized variable selection in L-PLS regression, *BMC Bioinformatics*, **13**, 327.

Nguyen, D. V. and Rocke, D. M. (2002a). Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, **18**, 39–50.

Nguyen, D. V. and Rocke, D. M. (2002b). Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics*, **18**, 1216–1226.

Nguyen, M. N., Ma, J., Fogel, G. B. and Rajapakse, J. C. (2009). Di-codon usage for gene classification. In V. Kadirkamanathan, et al. (Eds.), *Pattern Recognition in Bioinformatics* (pp. 211–221), Springer Berlin, Heidelberg.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L. and Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, *Applied Spectroscopy*, **54**, 413–419.

Sæbø, S., Almøy, T., Aarøe, J. and Aastveit, A. H. (2008). ST-PLS: a multi-dimensional nearest shrunken centroid type classifier via PLS, *Journal of Chemometrics*, **22**, 54–62.

Singh, B. K., Nazaries, L., Munro, S., Anderson, I. C. and Campbell, C. D. (2006). Use of multiplex terminal restriction fragment length polymorphism for rapid and simultaneous analysis of different components of the soil microbial community, *Applied and Environmental Microbiology*, **72**, 7278–7285.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science*, **18**, 104–117.

Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P. and Rubin, E. M. (2005). Comparative metagenomics of microbial communities, *Science*, **308**, 554–557.

Watson, J. E., Whittaker, R. J. and Dawson, T. P. (2004). Avifaunal responses to habitat fragmentation in the threatened littoral forests of south-eastern Madagascar, *Journal of Biogeography*, **31**, 1791–1807.

Wold, S., Ruhe, A., Wold, H. and Dunn, III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM Journal on Scientific and Statistical Computing*, **5**, 735–743.

Received May 23, 2015; Revised August 26, 2015; Accepted September 23, 2015