Benefits of dimension reduction in penalized regression methods for high dimensional grouped data: a case study in low sample size

Soufiane Ajana, Niyazi Acar, Lionel Brétillon, Boris Hejblum, Hélène Jacqmin-Gadda, Cecile Delcourt

Benefits of dimension reduction in penalized regression methods for high dimensional grouped data: a case study in low sample size

Soufiane Ajana[1] *, Niyazi Acar[2], Lionel Bretillon[2],Boris Hejblum[4,5], Hélène Jacqmin-Gadda[6], Cécile Delcourt[1] for the BLISAR Study Group

[1] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team LEHA, UMR 1219, F-33000 Bordeaux, France
[2] Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRA, Université Bourgogne Franche-Comté, Dijon, France
[3] Laboratory for Biology, Imaging, and Engineering of Corneal Grafts, EA2521, Faculty of Medicine, University Jean Monnet, Saint-Etienne, France
[4] Univ. Bordeaux, ISPED, Inserm Bordeaux Population Health Research Center 1219, Inria SISTM, F-33000 Bordeaux, France
[5] Vaccine Research Institute (VRI), Hôpital Henri Mondor, Créteil, France
[6] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team Biostatistics, UMR 1219, F-33000 Bordeaux, France
* To whom correspondence should be addressed

Abstract:

Motivation: In some prediction analyses, predictors have a natural grouping structure and selecting predictors accounting for this additional information could be more effective for predicting the outcome accurately. Moreover, in a high dimension low sample size (HDLSS) framework, obtaining a good predictive model becomes very challenging. The objective of this work was to investigate the benefits of dimension reduction in penalized regression methods, in terms of prediction performance and variable selection consistency, in HDLSS data. Using two real datasets, we compared the performances of lasso, elastic net, group lasso (gLasso), sparse group lasso (sgLasso), sparse partial least squares (sPLS), group partial least squares (gPLS) and sparse group partial least squares (sgPLS).

Results: Considering dimension reduction in penalized regression methods improved the prediction accuracy. The sgPLS reached the lowest prediction error while consistently selecting a few predictors from a single group.
Availability and implementation: R codes for the prediction methods are freely available at https://github.com/SoufianeAjana/Blisar
Contact: Soufiane.ajana@u-bordeaux.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1. Introduction:

High dimensional data have become of increasing importance in the biological domain. Data generated by high-throughput technologies allow to measure up to millions of features at once (Clarke *et al.*, 2008). A new type of information is thus generated, commonly known as "omics" data. Selecting a few predictors associated with a biological or clinical outcome among such high dimensional data is a challenging task (Filzmoser *et al.*, 2012). Traditional approaches usually fail because of intrinsic multicollinearity among the very large number of potential predictors (James *et al.*, 2017). The concepts of sparsity and penalization have shifted from being exclusively used by statisticians to becoming commonly used techniques by biologists and clinicians. Note that sparsity here does not refer to techniques dealing with sparse data but instead refers to models having a few non zero parameters (Hastie *et al.*, 2015). In high dimensional data, the presence of predictors with very small contributions to predictive power is likely. Keeping these predictors in the model may generate noise, leading to overfitting and lowering the prediction performance when the true vector of parameters is sparse (Geron, 2017).

When the aim is to reach a compromise between model interpretation (i.e., parsimonious model) and prediction performance, many approaches have been proposed in the literature.

Genuer et al. proposed VSURF, a variable selection approach based on random forests (Genuer *et al.*, 2010). Other nonlinear methods also performing variable selection such as support vector machines (Zhang *et al.*, 2016) or boosting (Xu *et al.*, 2014) were also widely discussed in the literature. However, since we position ourselves in a high dimension low sample size (HDLSS) framework, such complex models would tend to overfit our data while linear models proved to be more generalizable (Boucher *et al.*, 2015). For instance, penalized linear regression methods allow for variable selection by penalizing the size of the estimated parameters. Particularly, the lasso method (Tibshirani, 1994) shrinks the regression coefficients towards zero and estimates some of them to exactly zero. However, in some situations, when the predictors are highly correlated for example, the lasso fails to select the most relevant ones. A generalized version of the lasso, known as elastic net (Zou and Hastie, 2005), tackles this issue by giving highly correlated predictors similar regression coefficients, up to a change of signs if negatively correlated. Alternatively, one can also handle high correlations among predictors by incorporating dimension reduction in penalized regression methods. Particularly, sparse partial least squares (sPLS) (Lê Cao *et al.*, 2008; Chung and Keles, 2010) seeks sparse latent components (i.e., linear combinations of the original predictors) that are highly correlated with the outcome and a have high variance (Hastie *et al.*, 2009). These kind of approaches have been successfully applied in many domains (Bastien *et al.*, 2015; Lê Cao *et al.*, 2009).

In some applications, predictors have a natural grouping structure and selecting predictors clustered into groups could be more effective for predicting accurately the outcome than considering single predictors. For instance, in the BLISAR study (presented in section 3), our objective was to predict retinal omega 3 (n-3) polyunsaturated fatty acids (PUFA) levels from circulating biomarkers measured in blood samples using gas chromatography (GC) and liquid chromatography coupled to electroSpray ionization tandem mass spectrometry (LCMS) techniques (Berdeaux *et al.*, 2010; Acar *et al.*, 2012). We measured these circulating biomarkers from several blood compartments using different methods that structured them into 5 groups (Supplementary Figure S2). A prediction model for retinal n-3 PUFA including predictors from a few groups would make the interpretation of the model easier and its use cheaper, by lowering the number of biological analyses to perform. Indeed, since each analysis results in a spectrum allowing for the concomitant measurement of a large number of biomarkers, the number of

biomarkers measured in one compartment has little impact on the cost while adding a compartment increases a lot the cost.

Over the years, some authors proposed extensions to the previously presented statistical methods to take into account the grouping structure of high dimensional predictors as shown in Supplementary Figure S1. Yuan and Lin proposed the group lasso (gLasso)(Yuan and Lin, 2006) which selects or discards an entire group of predictors ("all-in-all-out" fashion). To achieve a bi-level sparsity, a recent method known as sparse group lasso (sgLasso)(Simon *et al.*, 2012) performs variable selection at the group level but also within each relevant group (Friedman *et al.*, 2010). In the same line, sPLS was also extended to group PLS (gPLS) and sparse group PLS (sgPLS) (Liquet *et al.*, 2016). We will refer to sPLS based approaches as dimension reduction methods in the rest of this article.

Surprisingly, the benefits of dimension reduction in penalized regression approaches were never investigated when the predictors are structured into groups.

The objective of this article is to compare the prediction performances and the variable selection consistency of 7 methods in high-dimensional settings while accounting or not for the group and high correlation structures. The present study aspires to lend insights into best practices when such methods are required.

The rest of this article is organized as follows. In section 2, we give an overview of penalized regression methods (lasso, gLasso, sgLasso, elastic net) and dimension reduction approaches (sPLS, gPLS, sgPLS). In section 3, we present the BLISAR study and we compare these methods on this real dataset in terms of variable selection frequency and prediction accuracy, using a repeated double cross-validation scheme. Main results are confirmed on a second data set described in the Supplementary Material. Finally, we summarize and discuss some perspectives in section 4.

2. Methods:

In regression settings, we commonly use the linear regression model to predict a real-valued response Y from a set of predictors X. When predictors have a natural grouping structure, we can write the linear regression model as:

$$Y = \sum_{l=1}^{L} X_l \beta_l + \varepsilon \quad (1)$$

Where Y is the $n \times 1$ response vector of n observations, $X = (X_1,...,X_L)$ is the $n \times p$ matrix of predictors and $X_l = (X_{l1},...,X_{lp_l})$ is the $n \times p_l$ matrix of p_l predictors in group l such that $l = 1,...,L$ and $p = \sum_{l=1}^{L} p_l$, $\beta = (\beta_1^T,...,\beta_L^T)^T$ is the $p \times 1$ vector of parameters to estimate and $\beta_l = (\beta_{l1},...,\beta_{lp_l})^T$ is the $p_l \times 1$ vector of parameters associated to the l^{th} group. The random error vector ε $(n \times 1)$ is a mean-zero with constant variance σ^2 normally distributed variable. We also assume that the outcome is centered, and therefore, no intercept is included in the model.

A famous estimator for such a model is the ordinary least squares estimator (OLS) obtained by minimizing the residual sum of squares (RSS).

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

However, in high dimensional settings ($n << p$), two issues are to be considered: collinearity of predictors and the signal to noise ratio (i.e., sparsity of the true vector of parameters). Collinearity is a property relative to the presence of redundancy/correlation among some predictors. In this case, rank$(X) < p$ and $X^T X$ becomes singular (Næs and Mevik, 2001; Tropp and Wright, 2010). In such settings, direct application of traditional variable selection methods (such as stepwise subset selection) may result in lack of stability, high computational effort or both. In this case, there is no unique $\hat{\beta}$ to minimize the RSS (Strang, 2016) and we need to regularize the estimation process. Even in low dimensional settings ($p \leq n$), predictors can be highly correlated and we may still need some regularization. The signal to noise ratio is relative to the concept of parsimonious models (only a few predictors associated with the response among a large set of available predictors). Penalized regression and dimension reduction methods are two approaches based on these concepts. The former technique makes a prior assumption that a few predictors are individually related to the outcome. The latter approach is based on the assumption that a few latent variables (also called underlying components) contribute to the observed covariance between the predictors and the outcome.

2.1. Penalized regression methods

Penalized regression methods investigated in this work (namely lasso, gLasso, sgLasso and elastic net) perform the estimation of parameters and variable selection simultaneously. Indeed, a penalty term, controlling the size of β, is added to the RSS in the optimization problem in order to reduce the variance and thus stabilize the OLS estimates. When the predictors are structured into groups, the optimization problem to solve becomes:

$$argmin_{\beta}\left\{\|Y - \sum_{l=1}^{L} X_l\beta_l\|_2^2 + \lambda\left(\alpha\|\beta\|_1 + (1-\alpha)\sum_{l=1}^{L}\sqrt{p_l}\|\beta_l\|_2\right)\right\} \qquad (2)$$

where, $\alpha \in [0,1]$ is a tuning parameter which controls the combination between the L1 and L2 penalties and $\lambda \geq 0$ is a tuning parameter determining the sparsity of the solution by controlling the bias-variance tradeoff. Larger values of λ lead to a sparser vector of the estimated parameters $\hat{\beta}$. The $\sqrt{p_l}$ term accounts for the group size. Moreover, note that when $L = p$, all groups are composed of one predictor.

Otherwise, when there is no *a priori* group assumption, the optimization problem simplifies to:

$$argmin_{\beta}\{\|Y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2)\} \qquad (3)$$

Lasso:

The lasso (Tibshirani, 1994) is a shrinkage method imposing an L1 penalty ($\alpha = 1$ in (2) or (3)). The non-differentiability of the L1 penalty at 0 allows an automatic variable selection (by shrinking some of the coefficients to exactly 0). Indeed, when λ is sufficiently large, the lasso produces a sparse solution. However, the lasso suffers from some major limitations: (i) when $n < p$, the number of selected predictors is bounded by the sample size and (ii) in case of highly correlated predictors, the lasso fails to perform grouped selection and selects instead only one variable from the entire group of correlated predictors.

Elastic net:

The elastic net (Zou and Hastie, 2005) is a combination of the L2 and L1 penalties ($0 < \alpha < 1$ in (3)) and can be considered as a generalization of the lasso. The L2 penalty (squared) allows

the elastic net to account for high collinearity and to select highly correlated predictors (e.g., genes located close on the same chromosome) by giving them similar weights, up to a change of signs if negatively correlated. Moreover, the L1 penalty gives the elastic net the sparse property of the lasso. Finally, the number of selected predictors is not bounded by the sample size as in lasso.

Group Lasso:

To consider the inherent interconnections inside a natural group of predictors, the gLasso (Yuan and Lin, 2006) mimics the lasso selection procedure but at the group level ($\alpha = 0$ in (2)). Indeed, the L2 penalty (not squared) is non-differentiable at the origin, setting groups of coefficients to exactly 0. In contrast, the elastic net performs grouped selection of highly correlated predictors when the group information is unknown *a priori* (Zeng *et al.*, 2017). It is worth mentioning that the gLasso is equivalent to the lasso if the size of each group is 1. However, the gLasso is not able to discriminate signal from noise inside a group since it either selects or discards the whole group of predictors. Moreover, the gLasso performs better than lasso when data are truly structured into groups (Huang *et al.*, 2009).

Sparse group Lasso:

A further refinement of the gLasso is the sgLasso (Simon *et al.*, 2012), which is a convex combination of the gLasso and the lasso penalties ($0 < \alpha < 1$ in (2)). Indeed, the sgLasso performs a bi-level selection by combining two nested penalties. The L2 penalty allows for group selection by taking into account the prior group information and the L1 penalty performs within-group selection and produces more parsimonious and more interpretable models. Thus, the sgLasso identifies important groups and discards irrelevant predictors inside each relevant group simultaneously.

2.2. Dimension reduction methods

The aforementioned penalized regression methods assume that some predictors contribute individually to the prediction of the outcome. By contrast, dimension reduction approaches assume that only a few latent variables inform the model. For example, each latent variable $T \in R^n$ of our predictors matrix X ($n \times p$) is constructed as a linear combination of the original predictors with their weight coefficients stored in a loading vector $u \in R^p$ such that $T = Xu$. Dimension reduction methods investigated in this work (namely sPLS, gPLS and sgPLS) are designed to relate a matrix of predictors X to a matrix of responses Y ($n \times q$) by maximizing the covariance of their projections onto orthogonal latent variables (also called latent scores). In the present study, we focus on the case of a univariate response ($q = 1$) and one latent dimension as explained in section 3.2. Under such conditions, the maximization criterion can be written as:

$$\sum_{l=1}^{L}\{\boldsymbol{cov}(\boldsymbol{X_l u_l}, \boldsymbol{Y}) - \boldsymbol{\lambda}(\alpha\|u\|_1 + (1-\alpha)\sqrt{p_l}\|u_l\|_2)\} \qquad (4)$$

where u_l is the estimated loading vector associated to the l^{th} group. $\lambda \geq 0$ is a tuning parameter which determines the amount of penalization, while $\alpha \in [0,1]$ controls the trade-off between the L1 and L2 penalties. Larger values of λ lead to a sparser vector of the estimated loadings. The $\sqrt{p_l}$ term accounts for the group size.

Sparse PLS:

The sPLS (Lê Cao *et al.*, 2008) aims at combining variable selection and dimension reduction in a one-step procedure. Indeed, the sPLS performs variable selection to obtain sparse loading

vectors by imposing an L1 penalty ($\alpha = 1$ in (4)). This means that only a few original predictors will contribute to each latent variable. Moreover, the sPLS is especially well suited for highly correlated predictors since it considers a contribution from all the relevant predictors when constructing a latent variable. However, if we can structure the data into groups, the sPLS cannot take into account this additional information.

Group PLS:

Inspired by the gLasso approach, when the underlying model exhibits a grouping structure, the gPLS (Liquet *et al.*, 2016) aims to select only a few relevant groups of X which are related to Y by imposing an L2 penalty ($\alpha = 0$ in (4)). In gPLS, each latent score is constructed as a linear combination of all the predictors inside the selected groups. However, as gLasso, the gPLS is not able to select the most predictive predictors inside each relevant group.

Sparse group PLS:

The sgPLS (Liquet *et al.*, 2016) is performed by combining the L1 and the L2 penalties ($0 < \alpha < 1$ in (4)). When the objective is to construct latent scores while achieving sparsity at both the group and the individual levels, the sgPLS can be a good alternative to gPLS. Indeed, as sgLasso, the sgPLS is capable of discriminating important predictors from unimportant ones within each selected group.

3. Design of the comparative study

3.1. Real data:

The general aim of the BLISAR study is to identify and validate new circulating biomarkers of lipid status that are relevant for retinal aging. In this application, our objective is to predict retinal n-3 PUFA concentrations from circulating biomarkers in post-mortem samples from human donors.

Samples of retina, plasma and red blood cells were collected from human donors free of retinal diseases according to previously published procedures (Berdeaux *et al.*, 2010; Acar *et al.*, 2012). Retinal n-3 PUFA status was measured using GC. Circulating biomarkers were obtained from 5 sets of analyses (Supplementary Figure S2): GC applied to lipids from total plasma (PL), cholesteryl esters (CE), phosphatidylcholines (PC), and red blood cells (GR). Finally, structural analyses of red blood cells were performed by LCMS as detailed previously (Berdeaux *et al.*, 2010; Acar *et al.*, 2012). Therefore, the analyses were performed on N=46 subjects and 332 predictors.

3.2. Repeated double cross-validation scheme:

We compared the prediction performances of the regression methods on the BLISAR dataset *via* a repeated double cross-validation scheme. Estimating both the tuning parameters and prediction errors by using a single cross-validation would lead to an overly optimistic estimate of the error rate value (Smit *et al.*, 2007). As an alternative, we designed a double cross-validation scheme to limit overfitting by performing model selection in the internal loop and model assessment in the external loop (Supplementary Figure S3) (Ambroise and McLachlan, 2002; Baumann and Baumann, 2014).

For all the compared methods, we estimated the tuning parameters in a data-driven fashion. Concerning dimension reduction methods, we considered one latent dimension as we are predicting a univariate outcome and to facilitate interpretation of the model. Furthermore, cross-

validation can fail to correctly estimate the optimal number of latent variables when the ratio of sample size to predictors is very low (Rendall *et al.*, 2017), as in our case. Moreover, the choice of the PLS dimension remains an open research question as mentioned by several authors (Boulesteix, 2004; Lê Cao *et al.*, 2008).

Our double cross-validation algorithm is the following (Supplementary Figure S3):

1. outer cross-validation cycle: randomly split the entire dataset into training (outer train) and test (outer test) sets using 10-fold cross validation (to reduce the sampling dependence and thus better estimate the prediction performance as well as its variability).
2. inner cross-validation cycle: the outer train portion is used to estimate the optimal tuning parameters using a leave-one-out cross-validation (due to our low sample size) and a grid search over the parameters space.
3. using the optimal tuning parameters selected at step 2, estimate the model on the whole outer train set.
4. predict the outcome values in the outer test set and compute the criteria for evaluating the quality of prediction.

As recommended, we repeated the double cross-validation procedure 100 times with different random splits into outer train and outer test sets in order to estimate the variance of the prediction performances (Molinaro *et al.*, 2005; Martinez *et al.*, 2011; Garcia *et al.*, 2014). Additionally, Filzmoser et al. reported that repeated double cross-validation is well suited for small data sets (Filzmoser *et al.*, 2012).

We used the CRAN R package SGL to train and to test the lasso, the gLasso and the sgLasso. We fitted the sPLS, the gPLS and the sgPLS to our data *via* the R package sgPLS which relies heavily on the package mixOmics. Concerning the elastic net, we used the package glmnet.

3.3. Model evaluation criteria

Root mean squared error of prediction (RMSEP):

The RMSEP is frequently used to assess the performance of regressions (Mevik and Cederkvist, 2004; Ivanescu *et al.*, 2016). In the present study, we calculated the RMSEP through cross-validation for both model selection and model assessment by averaging the squared prediction errors of the test sets:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{n_{test}}}$$

Where n_{test} is the test sample size, y_i (respectively $\hat{y}_i$) is the observed (respectively predicted) value of the outcomes for the i^{th} individual. Lower values of RMSEP are associated with better performances.

Goodness of fit (R^2):

We calculated the coefficient of determination (R^2) as the square of Pearson correlation coefficient (Feng *et al.*, 2012) between observed and predicted outcome values in the test set. This coefficient evaluates the prediction performance and thus was also used to compare our models. It is noteworthy that the R^2, calculated *via* cross-validation on the test data, assesses the quality of predictions on independent sets (Acharjee *et al.*, 2013; Rendall *et al.*, 2017).

4. Results:

We show the prediction performance of each method according to the RMSEP and the R^2 in Table 1. We also reported the number of predictors selected in at least 60% of the samples. The sgPLS model had the lowest prediction error (RMSEP=2.27), while selecting only one group (CE) and only 7 predictors inside that group. Interestingly, sPLS had a behavior very close to that of sgPLS although it does not consider explicitly the grouping structure. Indeed, sPLS had a RMSEP of 2.32 and selected only 8 predictors: 7 lipids from the CE group (identical to those selected by sgPLS) and 1 lipid from the PL group. In comparison, gPLS had a somewhat higher RMSEP (2.43) and selected only the CE group, but retained all the 32 predictors of this group.

The Supplementary Figure S4.A of the Venn diagram displays the intersection between predictors selected by the 3 dimension reduction methods.

Table 1. Comparison of the multivariable regression methods for 10 random divisions with 100 runs (N=46, p=332)

Method	Test data R^2 (SD)	Test data RMSEP (SD)	Number of selected predictors*	Selected groups*
Lasso	0.14 (0.05)	2.73 (0.14)	4	CE, PC
sgLasso	0.20 (0.05)	2.72 (0.16)	143	CE, PC, LCMS
gLasso	0.21 (0.05)	2.69 (0.15)	285	CE, PC, PL, LCMS
Elastic net	0.18 (0.05)	2.65 (0.12)	23	CE, PC, PL, LCMS
sPLS	0.36 (0.03)	2.32 (0.05)	8	CE, PL
gPLS	0.30 (0.03)	2.43 (0.05)	32	CE
sgPLS	**0.38 (0.02)**	**2.27 (0.04)**	**7**	CE

***** In at least 60% of the samples

Without dimension reduction, penalized regression methods exhibited higher RMSEP and lower R^2. The lasso had the highest RMSEP and the lowest R^2, while selecting only 4 predictors (from CE and PC groups). The sgLasso and the gLasso models performed similarly to lasso in terms of RMSEP but retained more groups and many more predictors. Notably, gLasso selected 4 groups out of 5 while sgLasso selected 3 groups. The groups selected by gLasso included those selected by lasso and sgLasso. The sgLasso selected many predictors inside each group, reaching a total of 143 predictors. The elastic net obtained similar prediction performances as gLasso and sgLasso but selected only 23 predictors from 4 groups. Intersections between selected predictors from the 4 penalized regression methods without dimension reduction are displayed in Supplementary Figure S4.B. Interestingly, 3 out of the 4 predictors commonly selected by penalized regression methods were included in the 7 ones commonly selected by dimension reduction methods.

As a result of the repeated double cross-validation scheme, we observed a variability in the tuning parameters estimates and thus in the trained models. Therefore, we also reported the selection frequency of the predictors by each method (see Supplementary Figures S5-11). Indeed, the most relevant predictors tend to be more often selected during the model training. We observed that sgPLS and gPLS were the most stable methods in terms of variables selection frequency. These two methods systematically retained the most frequently selected predictors (say over 60% of the times) across the random different splits and over the 100 runs. These findings were still valid even if we slightly lowered the threshold. Nevertheless, sgPLS had the advantage of consistently selecting fewer predictors compared to gPLS. Furthermore, standard

deviations of RMSEP and R^2 values were lower for sPLS, gPLS and sgPLS compared to penalized regression methods without dimension reduction (Table 1).

To further investigate the variance of the prediction accuracy obtained by each method over the 100 runs, we considered sgPLS as a benchmark. For each of the other methods and for each run, we computed the difference between their RMSEP and that of sgPLS. Supplementary Figure S12 displays the boxplots of these differences and shows that sgPLS outperformed the other methods for all runs (except for sPLS). As mentioned before, although not as good as sgPLS, sPLS performed similarly to sgPLS in terms of prediction accuracy. Of note, the R^2 criterion showed similar results (Supplementary figure S13).

We also compared the performances of these seven methods on another real dataset (DALIA trial) (Lévy et al., 2014; Liquet et al., 2016). The results are presented in the Supplementary Material. Again, sgPLS reached the best performances in terms of prediction accuracy while consistently selecting only a few relevant predictors from a single group. To summarize, adding dimension reduction exhibited a net and robust benefit for penalized regression methods in terms of prediction performances and stability of variable selection.

5. Discussion

In this study, we compared the prediction performances of several regression methods based on their RMSEP and R^2 for HDLSS data with a group structure of the predictors. All the compared approaches performed variable selection. The penalized regression methods performed better when combined with dimension reduction. Particularly, lasso had the worst prediction performance. In contrast, sgPLS reached the lowest (resp. the highest) RMSEP (resp. R^2) and found almost systematically a better predictive model than the other approaches. Interestingly, sPLS behaved similarly to sgPLS in terms of prediction performances.

In terms of variable selection, at the group level, sgPLS selected predictors from only one group (CE) compared to the sPLS which selected predictors form two different groups (CE and PL). Since selecting fewer groups would help to diminish the related costs, we retained sgPLS as the best approach. The fact that sgPLS selected consistently only a few predictors (7) suggests that our signal is relatively scarce with only a few relevant predictors.

From a biological point of view, since CE was the only group selected by sgPLS (and gPLS), one can suggest that prediction of retinal n-3 PUFA concentrations may rely on this analysis only, thereby much simplifying the analytical work. This observation is consistent with some of our preliminary findings, showing that retinal n-3 PUFA correlated strongly with n-3 in CE of the underlying vascular structure (retinal pigment epithelium/choroid) (Bretillon *et al.*, 2008). To our knowledge, this is the first study showing the benefits of dimension reduction in penalized regression methods while accounting for the grouping structure.

All the compared methods in the present study had a common objective: predict the outcome while dealing with different levels of collinearity and sparsity by discarding irrelevant predictors. There is, however, no guarantee that these kinds of approaches will always give the best results in all situations. The best prediction method will usually depend on the nature and the underlying structure of the data at hand, which cannot be known beforehand. If the data contains numerous noise predictors that can be discarded, then sparse methods may yield high prediction performances. Otherwise, if the true model is not parsimonious (many predictors driving the response), it is likely that a method using a linear combination of all the predictors, like PLS or ridge regression, would yield better prediction performances than sparse methods. In the biological domain, the number of subjects is often small and the measured quantities are generally highly correlated. If one has prior information about the group structure of the data and aims at selecting fewer groups then sgPLS seems to be a good approach.

Some of the considered models did not achieve good prediction performances. This could be due firstly to the linearity assumption of all the compared models. The true relationship between the outcome and the predictors may be nonlinear. However, we could not apply more complex methods (e.g., neural network or support vector machines) because of our small sample size. Such approaches would tend to overfit our data and predict with less accuracy (Boucher *et al.*, 2015). In contrast, linear models tend to be more generalizable and may outperform nonlinear approaches in case of a small training sample size or sparse data (Hastie *et al.*, 2001). Secondly, some of the applied methods may not be consistent in terms of variable selection, which could lower their prediction performances. Particularly, lasso shrinks each regression coefficient by the same amount. Thus, it heavily penalizes large coefficients and could lead to inconsistent model selection (Zou, 2006). As gLasso and sgLasso are built on lasso, they may suffer from similar problems (Fang *et al.*, 2015) and may also tend to select irrelevant predictors in the model. Additionally, when the data is structured into few groups and when each group contains more predictors than observations, the sgLasso is not expected to perform well in terms of variable selection (Simon *et al.*, 2012).

In contrast, adaptive lasso (Zou, 2006), adaptive gLasso (WEI and HUANG, 2010) and adaptive sgLasso (Fang *et al.*, 2015) remedy these shortcomings by using adaptive weights for penalizing different regression coefficients. Thus, the adaptive alternatives to lasso, gLasso and sgLasso are selection consistent. However, the adaptive methods' performances depend on the initial estimator used in their initial selection step. Therefore, there is a high risk of missing important predictors with an inappropriate initial estimator (Benner *et al.*, 2010). Thirdly, it is also possible that the true active set of predictors was not included as input. Indeed, it is very likely that the concentrations of circulating n-3 PUFA measured in blood samples are not sufficient to predict the retinal concentrations of n-3 PUFA with a high accuracy.

Some other techniques not investigated in the present work could also be good alternatives to reach better generalization performances. Stacked generalization, also called stacking or blending, consists in combining the predictions obtained by several models to form a final set of predictions (Wolpert, 1992). This approach was successfully applied in many domains, especially in machine learning challenges (e.g. Netflix challenge) (Sill *et al.*, 2009) but makes interpretation of the selected associations more challenging. The gain in prediction performance is often not worth the complexity of the final model. Furthermore, interesting interpretation properties could also be reached via orthogonal projection to latent structures (OPLS) method which removes variation from the predictors matrix that is not correlated to the outcome (Trygg and Wold, 2002; Féraud *et al.*, 2017). In particular, OPLS modeling of a univariate outcome requires only one predictive component. However, sparse generalizations of OPLS taking into account the group structure of the data are not implemented yet and could be investigated in future work.

In conclusion, one objective of this study was to assess the benefits of dimension reduction in penalized linear regression approaches for small sample size with high dimensional group structured predictors. The other objective was to lend insights into best practices when such methods are needed. Adding dimension reduction while considering both group structure and high correlations allowed to select the most biologically relevant group of predictors and to improve the prediction performance.

Funding:
This work was supported by grants from Agence Nationale de la Recherche [ANR-14-CE12-0020-01 BLISAR]; the Conseil Régional Bourgogne, Franche-Comté [PARI grant]; the FEDER (European Funding for Regional Economical Development); and the Fondation de France/Fondation de l'œil.
Acknowledgements:
BLISAR Study Group:
Niyazi Acar[1], Soufiane Ajana[2], Olivier Berdeaux[1], Sylvain Bouton[3], Lionel Bretillon[1], Alain Bron[1,4], Benjamin Buaud[5], Stéphanie Cabaret[1], Audrey Cougnard-Grégoire[2], Catherine Creuzot-Garcher[1,4], Cécile Delcourt[2], Marie-Noelle Delyfer[2,6], Catherine Féart-Couret[2], Valérie Febvret[1], Stéphane Grégoire[1], Zhiguo He[7], Jean-François Korobelnik[2,6], Lucy Martine[1], Bénédicte Merle[2], Carole Vaysse[5],
[1] Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRA, Université Bourgogne Franche-Comté, Dijon, France,
[2] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team LEHA, UMR 1219, F-33000 Bordeaux, France,
[3] Laboratoires Théa, Clermont-Ferrand, France
[4] Department of Ophthalmology, University Hospital, Dijon, France
[5] ITERG - Equipe Nutrition Métabolisme & Santé, Bordeaux, France.
[6] CHU de Bordeaux, Service d'Ophtalmologie, Bordeaux, F-33000, France
[7] Laboratory for Biology, Imaging, and Engineering of Corneal Grafts, EA2521, Faculty of Medicine, University Jean Monnet, Saint-Etienne, France,

6. Bibliography:

Acar,N. *et al.* (2012) Lipid Composition of the Human Eye: Are Red Blood Cells a Good Mirror of Retinal and Optic Nerve Fatty Acids? *PLoS ONE*, **7**.

Acharjee,A. *et al.* (2013) Comparison of Regularized Regression Methods for ~Omics Data. *MetabolomicsOpen Access*, **3**.

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci.*, **99**, 6562–6566.

Bastien,P. *et al.* (2015) Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics*, **31**, 397–404.

Baumann,D. and Baumann,K. (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J. Cheminformatics*, **6**, 47.

Benner,A. *et al.* (2010) High-Dimensional Cox Models: The Choice of Penalty as Part of the Model Building Process. *Biom. J.*, **52**, 50–69.

Berdeaux,O. *et al.* (2010) Identification and quantification of phosphatidylcholines containing very-long-chain polyunsaturated fatty acid in bovine and human retina using liquid chromatography/tandem mass spectrometry. *J. Chromatogr. A*, **1217**, 7738–7748.

Boucher,T.F. *et al.* (2015) A study of machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy. *Spectrochim. Acta Part B At. Spectrosc.*, **107**, 1–10.

Boulesteix,A.-L. (2004) PLS dimension reduction for classification with microarray data. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article33.

Bretillon,L. *et al.* (2008) Lipid and fatty acid profile of the retina, retinal pigment epithelium/choroid, and the lacrimal gland, and associations with adipose tissue fatty acids in human subjects. *Exp. Eye Res.*, **87**, 521–528.

Chung,D. and Keles,S. (2010) Sparse Partial Least Squares Classification for High Dimensional Data. *Stat. Appl. Genet. Mol. Biol.*, **9**.

Clarke,R. *et al.* (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer*, **8**, 37–49.
Fang,K. *et al.* (2015) Bi-level variable selection via adaptive sparse group Lasso. *J. Stat. Comput. Simul.*, **85**, 2750–2760.
Feng,Z.Z. *et al.* (2012) The LASSO and sparse least square regression methods for SNP selection in predicting quantitative traits. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **9**, 629–636.
Féraud,B. *et al.* (2017) Combining strong sparsity and competitive predictive power with the L-sOPLS approach for biomarker discovery in metabolomics. *Metabolomics*, **13**, 130.
Filzmoser,P. *et al.* (2012) Review of sparse methods in regression and classification with application to chemometrics. *J. Chemom.*, **26**, 42–51.
Friedman,J. *et al.* (2010) A note on the group lasso and a sparse group lasso. *ArXiv10010736 Math Stat*.
Garcia,T.P. *et al.* (2014) Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinforma. Oxf. Engl.*, **30**, 831–837.
Genuer,R. *et al.* (2010) Variable selection using random forests. *Pattern Recognit. Lett.*, **31**, 2225–2236.
Geron,A. (2017) Hands-On Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems O'Reilly Media, Inc, USA, Beijing Boston Farnham Sebastopol Tokyo.
Hastie,T. *et al.* (2015) Statistical Learning with Sparsity: The Lasso and Generalizations 1 edition. Chapman and Hall/CRC, Boca Raton.
Hastie,T. *et al.* (2001) The Elements of Statistical Learning - Data Mining, Inference New York, NY, USA.
Hastie,T. *et al.* (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction 5e ed. Springer-Verlag New York Inc., New York.
Huang,J. *et al.* (2009) Learning with Structured Sparsity. *ArXiv09033002 Math Stat*.
Ivanescu,A.E. *et al.* (2016) The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research. *Int. J. Obes. 2005*, **40**, 887–894.
James,G. *et al.* (2017) An Introduction to Statistical Learning: with Applications in R 1st ed. 2013, Corr. 7th printing 2017 edition. Springer, New York Heidelberg Dordrecht London.
Lê Cao,K.-A. *et al.* (2008) A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 35.
Lê Cao,K.-A. *et al.* (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, **25**, 2855–2856.
Lévy,Y. *et al.* (2014) Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load: Clinical immunology. *Eur. J. Immunol.*, **44**, 2802–2810.
Liquet,B. *et al.* (2016) Group and sparse group partial least square approaches applied in genomics context. *Bioinforma. Oxf. Engl.*, **32**, 35–42.
Martinez,J.G. *et al.* (2011) Empirical Performance of Cross-Validation With Oracle Methods in a Genomics Context. *Am. Stat.*, **65**, 223–228.
Mevik,B.-H. and Cederkvist,H.R. (2004) Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *J. Chemom.*, **18**, 422–429.
Molinaro,A.M. *et al.* (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301–3307.

Næs,T. and Mevik,B.-H. (2001) Understanding the collinearity problem in regression and discriminant analysis. *J. Chemom.*, **15**, 413–426.

Poignard,B. (2018) Asymptotic theory of the adaptive Sparse Group Lasso. *Ann. Inst. Stat. Math.*

Rendall,R. *et al.* (2017) Advanced predictive methods for wine age prediction: Part I – A comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods. *Talanta*, **171**, 341–350.

Sill,J. *et al.* (2009) Feature-Weighted Linear Stacking. *ArXiv09110460 Cs*.

Simon,N. *et al.* (2012) A Sparse-Group Lasso. *J. Comput. Graph. Stat.*

Smit,S. *et al.* (2007) Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta*, **592**, 210–217.

Strang,G. (2016) Introduction to Linear Algebra, Fifth Edition Fifth Edition edition. Wellesley-Cambridge Press, Wellesley, MA.

Tibshirani,R. (1994) Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Tropp,J.A. and Wright,S.J. (2010) Computational Methods for Sparse Solution of Linear Inverse Problems. *Proc. IEEE*, **98**, 948–958.

Trygg,J. and Wold,S. (2002) Orthogonal projections to latent structures (O-PLS). *J. Chemom.*, **16**, 119–128.

WEI,F. and HUANG,J. (2010) Consistent group selection in high-dimensional linear regression. *Bernoulli Off. J. Bernoulli Soc. Math. Stat. Probab.*, **16**, 1369–1384.

Wolpert,D.H. (1992) Stacked generalization. *Neural Netw.*, **5**, 241–259.

Xu,Z. *et al.* (2014) Gradient Boosted Feature Selection. In, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14. ACM, New York, NY, USA, pp. 522–531.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**, 49–67.

Zeng,B. *et al.* (2017) A link-free sparse group variable selection method for single-index model. *J. Appl. Stat.*, **44**, 2388–2400.

Zhang,X. *et al.* (2016) Variable Selection for Support Vector Machines in Moderately High Dimensions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **78**, 53–76.

Zou,H. (2006) The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.*, **101**, 1418–1429.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 301–320.